United States
Department of
Agriculture

Statistical
Reporting
Service

**Statistical
Research
Division**

**Staff Report
No. AGES811007**

Washington, D.C.

November 1981

Farmers' Attitudes Toward Crop and Livestock Surveys

A Collection of Papers Related to the Analysis of the Survey of Dakota Farmers and Ranchers

Ron Fecso
Robert D. Tortora

FARMERS' ATTITUDES TOWARD CROP AND LIVESTOCK SURVEYS

A Collection of Papers Related to the Analysis
of the Survey of Dakota Farmers and Ranchers

by

Ron Fecso and

Robert D. Tortora

Staff Report AGES811007
Research Division
Statistical Reporting Service

U.S. Department of Agriculture, Washington, D.C.

November 1981

FARMERS' ATTITUDES TOWARD CROP AND LIVESTOCK SURVEYS: A Collection of Papers Related to the Analysis of the Survey of Dakota Farmers and Ranchers; by Ron Fecso and Robert D. Tortora, Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., SRS Staff Report No. AGES811007. November 1981.

ABSTRACT

This report contains studies related to the Dakota survey of farmers and ranchers. Profile analysis is used to illustrate characteristics related to survey participation. Survey participation is modeled. Organizational influences are explored and multivariate analysis from a complex design is studied.

Key words: Survey response, respondent burden, model selection, principal component analysis, variable elimination.

Contents

SUMMARY

- Respondents who reportedly agree to participate on C&L surveys, as a group, have a more favorable attitude toward our surveys, use more information, find it easier to use and are more aware of the various users of C&L data.

- Production costs and demand information were reported to be necessary more often than other items.

- Survey contact when control data is reliable has a beneficial effect on response rate. Removal from a survey has a negative effect on some farmers' attitudes toward our program.

- Empirical data indicates that pps sampling, inversely with burden, can increase response rates.

- Education, perception of organizational influences and number of sources of information used are positively related to reported participation on C&L surveys.

- Different procedures which were used to prepare data from a single-stage, disproportionate, stratified survey for use in multivariate analysis are shown to affect the results.

FARMERS' ATTITUDES TOWARD CROP AND LIVESTOCK SURVEYS: A Collection of Papers Related to the Analysis of the Survey of Dakota Farmers and Ranchers

By Ron Fecso and Robert D. Tortora*

INTRODUCTION

The data used in these papers were obtained from a survey of farmers and ranchers in North and South Dakota which was conducted by the National Opinion Research Center (NORC) in cooperation with USDA. The primary purpose of this survey was to gather information concerning farmers' and ranchers' understanding and attitudes towards crop and livestock surveys and to improve USDA's understanding of their data needs, concerns, and motivation to participate in surveys.

A disproportionate stratified sample of farmers and ranchers was drawn from the list frames in North and South Dakota. In order to conduct various methodological studies two versions of the questionnaire were developed. There were several identical items on the two questionnaires, but each questionnaire explored some different areas, allowing measurement of the effects of question wording and ordering.

These papers represent an extension of the analysis done by NORC in their 1979 report No. 128, "Dakota Farmers and Ranchers Evaluate Crop and Livestock Surveys" by Jones, Sheatsley, and Stinchcombe. Included in this report are some analysis techniques which were not utilized in the NORC study. These analyses include profile and principal component analyses, some areas where the study can be beneficial to agricultural surveying methodology, such as organizational influences and respondent burden reduction, and some technical aspects about design effects on analysis techniques. The papers should be of interest to a general audience since they present insight into the factors underlying the farmer's decision not to participate in surveys. The second, third and especially fifth paper also present topics of interest to the technically oriented researcher.

* Ron Fecso is a Mathematical Statistician in the Sample Survey Research Branch. Robert D. Tortora is Chief of the Sample Survey Research Branch.

A PROFILE OF REPORTED PARTICIPATION

by
Ron Fecso

Introduction

The techniques of profile analysis are used to illustrate characteristics and opinions related to a respondent's decision to participate in agricultural surveys. 1/ The questions asked in the National Opinion Research Center (NORC) survey are first grouped by subject matter, and then profiles of the survey respondents who reported they participate in agricultural surveys are compared with profiles for the respondents who reported that they do not respond to agricultural surveys. 2/

Study Variables

The data used in the analysis consist only of responses from the version I NORC questionnaire, because there is no adequate method to compare the participation rate questions between the two versions of the questionnaire (a detailed discussion of the comparison problems appears later in the paper). Further, only those respondents indicating prior crop and livestock (C&L) survey contact (Q25 - yes) were analyzed, representing 88 percent of the version I data file.

The responses were divided into four groups based on the replies to the subjective question on participation:

"When you are asked to participate in a crop or livestock survey, do you almost always agree to participate, do you agree most of the time, only some of the time, or do you hardly ever agree to participate?"

The groups formed were:

Response	Number	Percent
Almost always agree	97	13
Agree most of the time	162	22
Agree only some of the time	195	27
Hardly ever agree	282	38
Total	736	100

A small number of "don't know" responses were removed from the analysis.

1/ Donald F. Morrison. Multivariate Statistical Methods (New York: McGraw-Hill Book Company, 1967).

2/ D. Jones, P. Sheatsley and A. Stinchcombe. Dakota Farmers and Ranchers Evaluate Crop and Livestock Surveys (Chicago, Illinois, National Opinion Research Center Report No. 128, 1979).

Each of these respondents replied to a series of questions concerning their attitudes toward government, C&L data, and agricultural surveys. Also, nonattitude questions were asked regarding age, education, and operation size.

A profile analysis is used to explore the differences between the responses to these questions for the groups above. In order to glean the most information in a profile analysis, the response scales should be commensurate. To establish this condition the analysis will be presented in two parts: a profile of the attitude questions and a review of the group differences for the nonattitude questions.

The attitude questions were scaled to approximately a [0,1] range. The responses to each question were arranged on the scale so that the most favorable attitudes toward C&L data or surveys were given the highest value, and the least favorable responses received the lowest value. Neutral answers were given a 0.5 response value, making the average response for yes/no questions differ slightly, but not significantly, from the ordinary percentage summary. 3/

Analysis

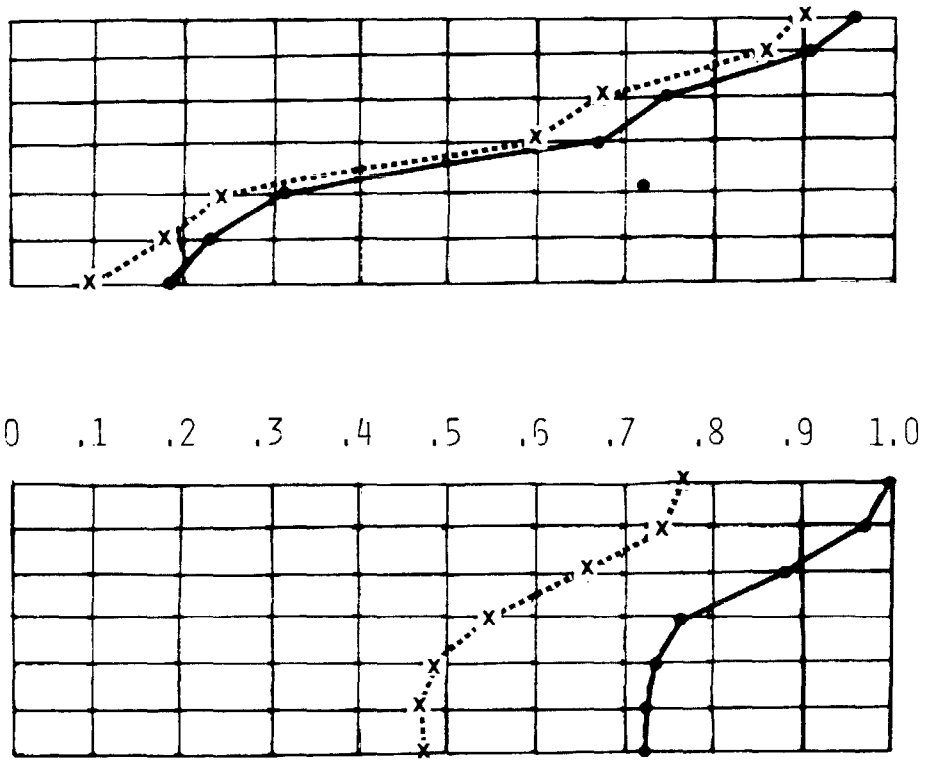
The first analysis of the data compared the mean scores on the attitude questions among the four groups based on the responses to the participation question. The mean profiles were dissimilar, that is, there was response by group interaction. Individual hypothesis tests then showed two prominent features. First, all statistically different response means show that the group which reportedly agrees to participate had a more favorable attitude on other questions. Second, logically grouped questions, such as accuracy questions and important decision factor questions, even when not quite significantly different individually, showed similar profiles and different response levels.

Very little profile difference existed between the "almost always agree" and the "agree most of the time" groups. These groups were, therefore, combined for the remaining analysis. Although the two least agreeable groups had some differences, they were combined for ease of presentation. These groups will be called the high and low participation groups and their members will be called participants and nonparticipants, respectively. Figures 1 through 4 depict 58 variables which showed significant level differences for individual items and/or item groupings. When interpreting the profile graphs, note that the mean responses by participation group are the plotted points, and the means within each participation group are connected within logical question groupings. Basically, a profile line to the right of another, indicates a group

3/ Since the data is subjectively scaled and since profile analysis is used as an exploratory technique, most numerical calculations have been omitted. The report centers on interpretation and presentation of the mean profiles of the responses.

FIGURE 1--PROFILE OF PARTICIPATION

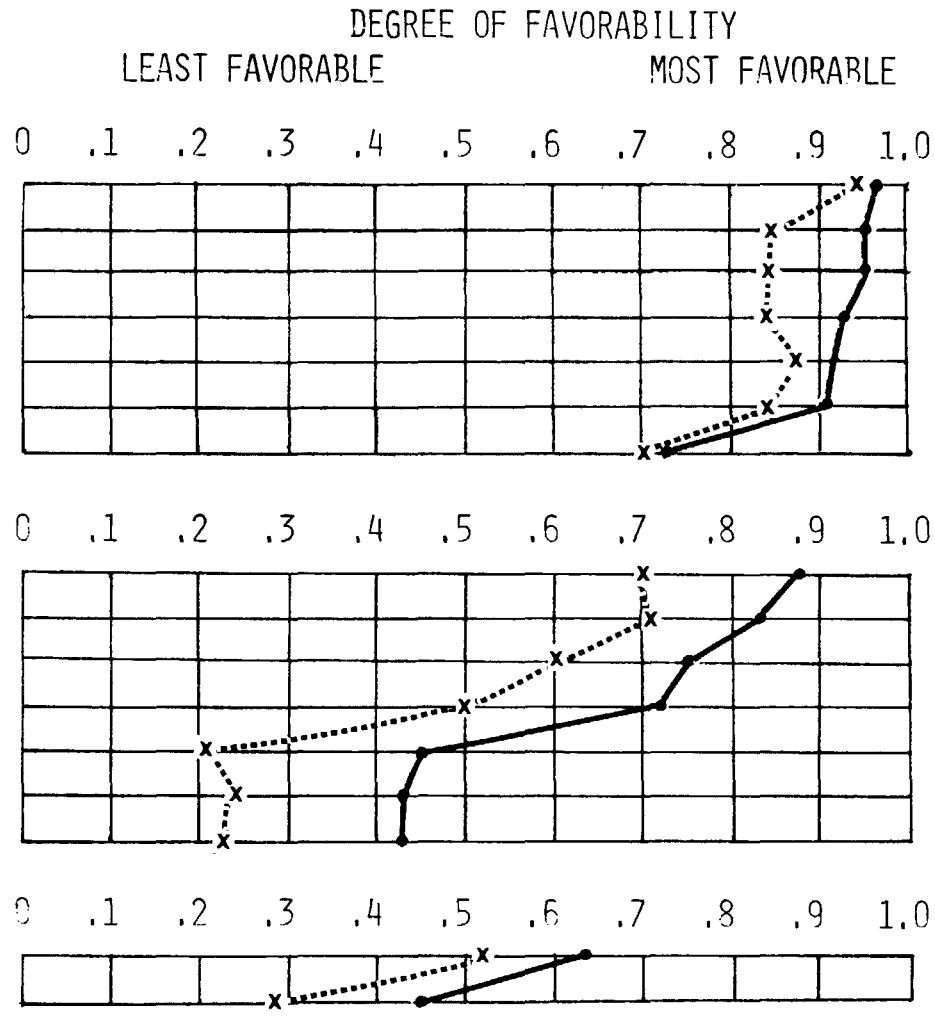
LINE #				DEGREE OF FAVORABILITY: LEAST FAVORABLE ... MOST FAVORABLE
	INFORMATION			0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0
1	HAS AN INFO SOURCE	(7) [1]	% [2]	
2	KNOWS SOURCE DATA BASE	(9)	%	
3	DISCUSS AGRI INFO	(13)	%	
4	USE MAGS - JOURNALS	(7)	%	
5	USE EXT OR ASCS AGENT	(7)	%	
6	USE USDA PUBS	(7)	%	
7	USE C & L REPORTS	(7)	%	
	USEFULNESS			0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0
8	PRODUCTION COSTS USEFUL	(11F)	#	
9	CROP DEMAND USEFUL	(11G)	#	
10	CROP PRICE USEFUL	(11C)	#	
11	STOCK TO BE SOLD USEFUL	(11E)	#	
12	CROP ESTIMATES USEFUL	(11B)	#	
13	PLANTING INTENTIONS USEFUL	(11A)	#	
14	# OF STOCK USEFUL	(11D)	#	



1/ Numbers in parentheses indicate the associated question number in version I of the questionnaire.

2/ Type of scale coded for the responses: % - percent of respondents, # - index value.

FIGURE 2--PROFILE OF PARTICIPATION



LINE #			
	KNOWN DATA USERS		
15	GRAIN BUYERS	(14) 1/	%
16	STOCK BUYERS	(14)	%
17	BANKERS	(14)	%
18	STORAGE OPERATORS	(14)	%
19	FOOD PROCESSORS	(14)	%
20	FARM ORGS	(14)	%
21	LOCAL OR STATE GOV'T	(14)	%
	USE HELPS FARMERS		
22	FARM ORGS	(14c)	%
23	BANKERS	(14c)	%
24	STORAGE OPERATORS	(14c)	%
25	LOCAL OR STATE GOV'T	(14c)	%
26	STOCK BUYERS	(14c)	%
27	FOOD PROCESSORS	(14c)	%
28	GRAIN BUYERS	(14c)	%
	CONFIDENTIALITY		
29	RELEASE INFO - PRIVATE	(38)	%
30	RELEASE INFO - GOV'T	(39)	%

1/ Numbers in parentheses indicate the associated question number in version I of the questionnaire.

FIGURE 3--PROFILE OF PARTICIPATION

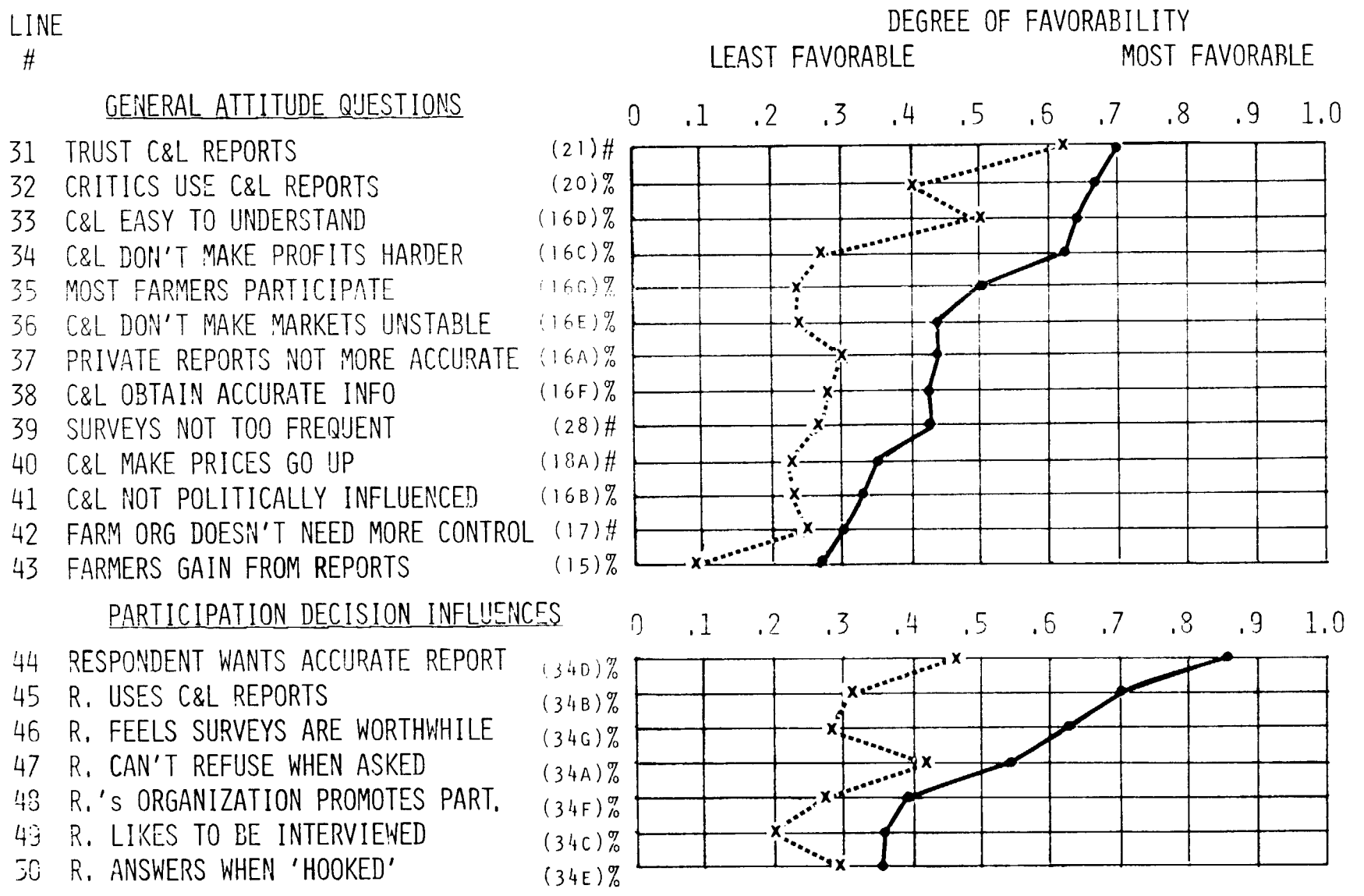
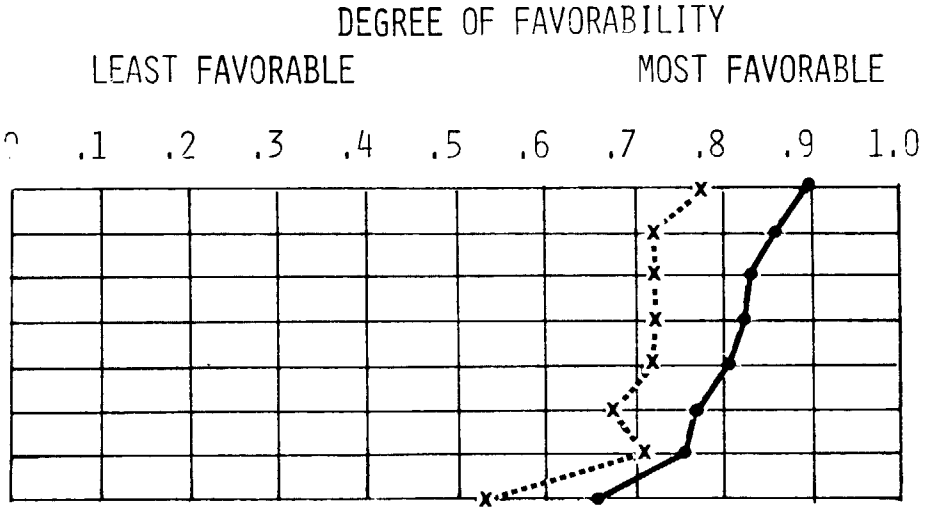


FIGURE 4--PROFILE OF PARTICIPATION

LINE #	RESPONDENTS ABILITY TO BE ACCURATE		
51	KIND/# STOCK	(37D)	%
52	ACRES PER CROP	(37B)	%
53	OFF FARM INCOME	(37H)	%
54	CROP YIELDS HARVESTED	(37C)	%
55	SUPPLY COSTS	(37F)	%
56	TOTAL FARM INCOME	(37G)	%
57	STOCK PRICES RECEIVED	(37E)	%
58	ACRES TO BE PLANTED	(37A)	%



which has a more favorable attitude about C&L statistical reporting. The reportedly high participation group is the profile on the far right of each graph.

The first seven questions (figure 1, lines 1 through 7) are concerned with information sources used. The profiles are similar, but the high participation group has more knowledge of, and uses more information sources. There is a difference between information sources. USDA sources were reportedly used less than mass media, which may, in part, be attributed to the need for a more thorough explanation than provided in C&L reports (see line 33). Also, since mass media outlets such as magazines and news programs supply the information and would be used anyway, there may be little need to go directly to C&L reports.

The types of information considered useful (lines 8 through 14) are again similar or parallel but with a considerable difference in the level of usefulness. The larger number of information sources and the increased ease in understanding the data (line 33) are probably major factors in the high level of usefulness reported by the high participation group. In this question grouping, production costs and demand information were reported to be necessary more than the other items.

Participant's knowledge of who uses the data (lines 15 through 21) lacks parallelism. Both participation groups have a nearly equal awareness of the use of C&L data by grain buyers and local and State governments. Overall, the high participation group is more aware of the various users of C&L data. One might speculate that the use of C&L reports by grain buyers and food processors during contract negotiations makes the use by these parties more readily remembered by the low participation group members. Neither participation group, especially those indicating lower participation, feels that the use of C&L data by the various buyers is very helpful to farmers (lines 22, 23, and 26).

The general attitude (lines 31 through 43) of participants is considerably more favorable toward C&L reports than the nonparticipants'. A value of 0.5 would indicate an even split between favorable and unfavorable responses to these questions. With the exception of trust and ease of understanding, the general attitude toward C&L reports was not very favorable in either participation group, but the level of favorability among participants was much higher than that of the group which seems prone to refuse to answer surveys.

In the area of factors which were reported to influence the decision to participate (lines 44 through 50), the nonparticipants, as might be expected, claimed to be less influenced by the feelings which they were questioned about. Here the lack of parallelism is very interesting. Participants consider the worth, use and accuracy of reports to be the most worthwhile reason to participate in surveys. Nonparticipants, on the other hand, were influenced more than expected by "can't refuse when asked" or "answers when hooked."

The ability to be accurate about agricultural information (lines 51 through 58) had a similar pattern with some level difference. The difference may be only that participants have tried, or are more willing to try, to answer surveys.

Looking at the overall profile, the typical nonparticipant appears less able or inclined to utilize the agricultural economic system as it exists. Those who indicated frequent refusal did not use the data and recognized it only when it was pointed out as a tool used against them. This may be a person whose decisions are more visceral than logical, and the key to the occasional successful survey response may just be an "ok, what the heck, they've got me hooked anyway" attitude. The profile for nonattitude questions reflects this type of participant. Being slightly less educated and younger, this person may find it difficult to utilize the market information and may be somewhat more reactionary. The group is socially distinct in the area of communications. They have more telephone party lines, were included on fewer surveys, and received (or remembered receiving) fewer presurvey letters, which may indicate a lower farm income. Yet, they had a smaller number of nonfarm jobs which also may indicate an inability or reluctance to change during shifting economic conditions.

PROFILE FOR NON-ATTITUDE TYPE QUESTIONS

NONPARTICIPANT GROUP

Slightly younger
Slightly less educated
Received (or remembered) fewer presurvey letters
Have more party lines
Included in fewer surveys
Fewer have nonfarm jobs

Conclusions

Recognizable differences are apparent between the group which participates in agricultural surveys and the group which does not. Although this analysis makes no reflection on the ability to change the responsiveness of individuals, it presents the areas in which further work in public relations may help reduce nonresponse to surveys.

Appendix

Comparisons Between Participation Rate Questions

The NORC split ballot design contained certain questions with wording differences between questionnaires. For example, the question that asked for the respondent's frequency of participation was written in two forms. The two forms proved difficult to analyze as a combined data set. One form asked the respondent:

> When asked to participate, do you
> Almost always agree,
> Agree most of the time,
> Agree only sometimes, or
> Hardly ever agree.

Thus, participation is measured subjectively. The second form of the question asked how many survey requests the respondent had received and how many replies were given. Thus, a numerical participation rate was available.

The comparison problems start with the respondent's interpretation of the subjective categories. The profile analysis showed little difference between groups who answered "almost always agree" and "agree most of the time." Likewise, little difference appeared between the two groups indicating the least likelihood of participation. The lack of distinction between these responses makes it difficult to associate a range of numerical participation to each response. For example, a respondent who participates 80 percent of the time may respond "almost always" or might say "most of the time." Thus, there is no strict linear ordering of the subjective responses.

A related problem concerns the number of survey requests used for the base of the numerical participation rate. When there were only a few chances for participation, the actual rate of participation did not reflect the rate that may have occurred with a large number of requests. For example, with only one request, the numerical rate is 0 or 100 percent, neither of which would be expected to hold in most cases when more survey activity is requested. Therefore, classification of a numerical rate into subjective categories based on few requested surveys would undoubtedly result in many errors.

An analysis of participation which combined both versions would require some very questionable assumptions about class equivalencies; therefore, only the observations from version I are used in the profile analysis.

A NOTE ON THE USE OF UNEQUAL PROBABILITY SAMPLING TO REDUCE RESPONDENT BURDEN

by
Ron Fecso

Introduction

Tortora proposed the method of unequal probability sampling as an active research item in 1977. 1/ 2/ At that time, it was shown "that it is possible to reduce the expected burden of larger farm operators at practically no loss in sampling efficiency by using a probability of selection computed inversely to the operators' burden.". The result was theoretically encouraging, and the paper implied that the method had "the potential of lowering" the refusal rate. To make a methodological change of this scope, a considerable program benefit must be shown. Reducing burden in itself would not necessarily help our program, but if burden reduction can reduce nonresponse without other adverse affects, then the method deserves renewed attention. Some subjective evidence, based on the version I questionnaire, indicates that unequal probability sampling could have a beneficial effect on crop and livestock (C&L) surveys.

Study Variables

Two questions were used to develop five groups reflecting various degrees of numerical burden (number of times asked to participate in surveys). The first question (Q25) asked the respondent: "Have you ever been asked to participate in a crop or livestock survey, either by mail, or on the phone, or in person?" Those answering "no" became the profile group "A." This group had the least amount of numerical participation burden. The respondents answering "yes" to the first question were divided into four groups based on the following question (Q27):

> "During the past 12 months ... how often were you asked to participate in a crop or livestock survey? Were you asked more than 10 times, from 5 to 10 times, or were you not asked at all during the past 12 months?"

The responses and group labels were:

0 times	Group B
1 to 4 times	Group C
5 to 10 times	Group D
More than 10 times	Group E

A profile analysis of all the attitude questions asked in the survey revealed that related questions, such as the different sources of

1/ Tortora, Robert D., "Reducing Respondent Burden for Repeated Samples", Agricultural Economics Research, Vol. 30, No. 5 (1977), 41-44.

2/ _____ and K. N. Crank, "The Use of Unequal Probability Sampling to Reduce Respondent Burden", ESCS Staff Report. (1978).

information or trust and confidentiality questions, increased or decreased in a similar pattern among the members of each of the five profile groups. Six questions were picked for this presentation because they reflected the attitude expressed for the related questions without deleting any questions which would be contradictory to the ideas presented in this paper. The survey questions and response codes follow (N = 836 for each of the questions):

SUBJECTIVE PARTICIPATION (Q26)
When you are asked to participate in a crop or livestock survey, do you:

	Percent
(3) Almost always agree to participate	12
(2) Agree most of the time	19
(1) Agree only some of the time	23
(0) Don't know or not asked the question 3/	12
(-1) Hardly ever agree	34

REQUESTS LAST YEAR (Q29)
During those 12 months--between March 1977 and February-- were you asked to participate in crop and livestock surveys?

	Percent
(1) More often than the previous year	8
(0) No change, don't know or not asked	70
(-1) Less often	22

C&L PUBLICATIONS (Q7)
Where do you get your information about things like livestock numbers, acres planted to various crops, and forecasts of yields?

	Percent
(1) Mentioned Crop and Livestock Reporting Service	12
(0) Didn't mention C&L Reporting Service	88

TRUST (Q21)
How often do you think you can trust the results of Government crop and livestock surveys?

	Percent
(3) Almost always	1
(2) Most of the time	20
(1) Only some of the time	60
(0) Don't Know	1
(-1) Hardly ever	18

FARM ORGANIZATION (Q41)
Are you a member of any farm organization or commodity association organized to represent the interests of farmers or ranchers?

	Percent
(1) Yes	60
(-1) No	40

3/ These respondents were removed from further analysis.

POLITICAL INFLUENCE (Q16B)

Do you agree or disagree with the statement "Government crop and livestock reports are not influenced by politics?"

	Percent
(1) Agree	24
(0) Don't know, no opinion	5
(-1) Disagree	71

Analysis

The data in the table titled "Profile of Burden from Survey Requests" displays the group means of the six questions chosen as representative of the survey along with the average crop acreage for the operations in each group. The groups are in order of increasing burden, with group A having no burden and profile group E indicating the largest number of survey requests. Interestingly, and somewhat contrary to what might be hypothesized, the group willingness to participate (subjective participation) tended to increase considerably as numerical burden increased. Numerical burden is related to the size of the operation: large diversified operations have more chances to be sampled. Additional analysis showed that, on the average, the respondents with larger farm operations were more educated and showed a greater inclination to utilize C&L information. The percentages of respondents who said they were influenced to participate in crop and livestock surveys by their desire for accuracy in these reports had an increasing pattern similar to the subjective participation question. This implies that the inclination to use the data is related to the decision to participate in surveys.

Another characteristic is related to increased participation. It appears that the survey requests themselves may help increase the response rate. This is implied by the following group comparisons. Groups A and C are very similar with three exceptions--having been asked to participate, farm size, and farm organization membership. Although not presented in the table, age, education, trust, and most other important group characteristics included in the survey were also similar for groups A and C. Groups B and C are very similar in size, education, and age, but differ considerably in the index of previous requests compared to the last year, and in the subjective participation index (0.44 and 0.76 respectively). Group B had survey experience but none in the last year. Group C indicated an increased amount of survey activity in the last year. The difference in mean scores across groups for subjective participation and requests last year tends to indicate that some initial survey contact has a sort of educational effect about what we do, and thus may encourage participation. Being removed from the surveys (group B) has a counter-effect which is implied by the reduced scores for subjective participation and the overall confidence in the agricultural statistical program for C&L items as indicated by the trust and political influence scores. The B group may be slightly uncharacteristic when comparing means because it may contain some refusals which were purged from nonprobability survey activity. These refusals would be expected to have lower participation and trust. Group B would also be expected to contain proportionately more respondents whose initial contact concerned an item in which they were a "small" operator.

These patterns are consistent with NORC's conclusions about burden. Basically, the center feels that the farmer is not so much concerned about the survey length or number of requests, but becomes upset when asked to spend time answering questions about items in which he has little interest.

This analysis suggests the following hypotheses:

A. Increased survey requests (within a reasonable amount) can have a beneficial effect on the response rate.
B. Some survey contact is desirable when control data is reliable.
C. Once surveyed, not being in a survey for a year or more has a negative effect on some farmers' attitudes about our program, and these feelings can result in a reduction in the future response rate.

The following sampling scheme will help relate these hypotheses to unequal probability sampling. Assign a burden to each operator based on the frequency of contacts over some fixed and predetermined time interval. For operators in strata based on a large amount of survey item presence, say large cattle operators, sample in the usual (pps, equal probability, etc.) manner. For the zero and possible small size strata, sample with probability inversely proportional to assigned burden.

Logic indicates that this sampling scheme would result in:

1. No change for the strata of large operators.
2. A slightly lower average burden for those on any survey.
3. More operators being included on at least one survey.
4. Little loss of efficiency in the small operator strata, provided there is no correlation between the present survey item and any previous survey items. If there is, it is a topic for further research.
5. Operators with a large burden can be "expected" to receive fewer survey contacts about items which are of lesser interest to them.
6. Response rates increasing if the hypothesis is true.

Conclusion

This analysis provided encouraging support of the usefulness of unequal probability sampling. Further studies should be planned to estimate more precisely the response increases which could be attributable to spreading the survey burden. With the record-keeping abilities of the List Frame's Sample Select System, the procedure should not be difficult to implement if it proves to be as beneficial as this analysis suggests.

PROFILE OF BURDEN FROM SURVEY REQUESTS

Mean values for the coded responses to the six attitude questions are presented by profile group. The coding of the responses was shown previously. Note that larger mean values denote a more favorable attitude.

Question:	Had the Respondent Even Been Asked to Participate?				
	Never Asked	YES			
Response:	To Participate	Reported Number of Survey Requests Between March 1977 and February 1978			
		0	1-4	5-10	More than 10
Profile Group	A	B	C	D	E
Group Size	87	137	423	136	41
Average Crop Acres	488	671	696	796	981
Subjective Participation	1/	.44	.76	.66	1.54
Requests last year	1/2/	1/3/	-.26	.01	-.02
C&L Publications	.10	.07	.12	.18	.22
Trust	.92	.66	.88	.96	1.07
Farm Organization	-.26	.23	.19	.40	.51
Political Influence	-.45	-.59	-.47	-.47	-.22

1/ Question was not asked for this group of respondents.
2/ Implied 0.
3/ Must be between 0 and -1.

STANDARD ERRORS OF THE MEANS FOR THE PROFILE DATA
USING THE UNWEIGHTED SIMPLE RANDOM SAMPLE FORMULA

	Profile Group				
	A	B	C	D	E
Subjective participation	N/A	.118	.073	.126	.229
Requests last year	N/A	N/A	.029	.058	.082
C&L publications	.033	.022	.016	.033	.065
Trust	.094	.088	.048	.075	.128
Farm organizations	.104	.084	.048	.078	.136
Political influence	.095	.067	.042	.073	.150

N/A = not Applicable

MODELING SURVEY PARTICIPATION IN NORTH AND SOUTH DAKOTA

by
Robert D. Tortora

Introduction

What factors account for participation in crop and livestock (C&L) surveys? Are they variables that SRS can influence? The data collected by NORC in North and South Dakota allows us to model survey participation. Although answers to the above questions cannot be specifically obtained we can gain insight into the variables that predict survey participation. Conditional on stratum membership and number of survey requests, it is found that respondent educational level accounts for the largest increase in the multiple correlation coefficient.

The following sections discuss the variables that were used in the analysis, the method of model selection, and an analysis of the amount of variation accounted for by the variables in the model.

Study Variables

Data used in this paper were collected from version II of the NORC questionnaire. This version asked the respondent the number of C&L survey requests received during the year prior to the NORC interview and the number of times the respondent agreed to participate. The latter is used as the dependent variable ,Y, to model survey participation. The variables used in this analysis are different than those used by Jones, Sheatsley, and Stinchcombe (1979). The differences arise because indexes were developed for more of the questions in order to reduce the number of independent variables for modeling. Groups of like questions were combined, and a total of 38 explanatory variables was originally considered. They included nine dummy variables to account for stratum membership. Using a method based on principal component analysis to eliminate redundant variables, 1 dummy variable and 17 independent variables were deleted. 1/

The dependent variable in this paper is the number of times the respondent reported participating in C&L surveys during the year prior to the NORC interview. Of 38 independent variables, 8 dummy variables (to account for stratum membership) and 12 variables from the questionnaires were retained for analysis. A description of the 12 variables from the questionnaire follows:

1/ Tortora, Robert D. "The Effect of a Disproportionate, Stratified Design on Principal Component Analysis Used for Variable Elimination," Farmers' Attitudes Toward Crop and Livestock Surveys: A Collection of Papers Related to the Analysis of the Survey of Dakota Farmers and Ranchers. Ed. Ron Fecso and Robert D. Tortora. U.S. Department of Agriculture, (1981).

Variable X_1 is the number of times the respondent was asked to participate in C&L surveys during the year prior to the NORC interview (Q25). Only those respondents who indicated that they had been asked to participate at least once during that year were included in this analysis.

Variables X_2 and X_3 deal with the respondent's crop characteristics. Variable X_2 is the total cropland acreage (Q45) in the farm operation. Variable X_3 is the total number of (main) crops (Q46) mentioned by the respondent.

Variables X_4 and X_5 deal with usefulness of C&L reports. Variable X_4 allows the respondent to describe the usefulness of C&L reports in managing the operation (Q11) as very useful, somewhat useful, not at all useful, or don't know (X_4 = 2, 1, -1, 0, respectively). Variable X_5 is an index that describes the usefulness to the respondent of the C&L reports at the county, State, National, and foreign country levels (Q12). The appendix describes X_5 in detail.

Variable X_6 measures the encouragement of the respondent's farm organization(s) for survey participation (Q48). A plus 1 is given for each organization that the respondent perceived as encouraging participation; a minus 1 stands for each organization that was perceived as neutral or for which the respondent was unaware of the organization's stand.

Variable X_7 is an index that measures the number of reasons a respondent stated for participating in C&L surveys. The higher the value of X_7, the more reasons a respondent has for survey participation. X_7 is built using parts of Q35. For each part of Q35, a plus 1 is added to the index if the respondent felt the part was important to survey participation, or a minus 1 is added to the index if the respondent felt the part was not important to participation. Nothing is added to the index for a part of Q35 if the respondent felt it was not applicable to survey participation.

Variables X_8 and X_9 generally measure the impact of C&L reports on the farmer or rancher and the agricultural community. X_8 is the number of groups that use C&L reports to hurt farmers and ranchers (Q15). Variable X_9 is the sum of nine responses about C&L reports. A plus 1 is added to the index if a response is favorable, or a minus 1 is added to the index if the response is not favorable (Q16).

Variable X_{10} measures the respondent's feelings towards the accuracy of data reported in Government C&L surveys by fellow farmers and ranchers (Q17). The respondent had 5 choices: (1) almost all of the time, (2) most of the time, (3) some of the time, (4) hardly ever, and (5) don't know, with corresponding values for X_{10} of 4, 3, 2, 1, and 0, respectively.

Variable X_{11} measures the number of sources of information the respondent used to help make managerial decisions. The more sources mentioned, the higher the value of X_{11} (Q77).

Finally, X_{12} is the educational level of the respondent. The higher the value, the higher the level (Q44).

The Model

This section is concerned with developing the form of the model used to predict survey participation. The term "form" means the general statistical description of the model that adequately fits the data in terms of an analysis of the residuals. Thus, one is not concerned with minimum variance unbiased estimation of the model parameters; but is concerned with uncovering the relationships between regressand and regressors.

NORC originally used survey participation rate, say $Y' = Y/X_1$, as the dependent variable. However, in the multiple linear regression,

$$Y' = a + B_2X_2 + \ldots + B_{12}X_{12} + e,$$

an analysis of residuals for predicted participation rate, resulted in a systematic departure from the fitted equation. The departure was noticed by examining the plots of residuals versus Y'.2/ A multiplicative model,

$$Y' = a(X_2)^{b2}\ldots(X_{12})^{b12}\ e,$$

was also examined, but the same problem occurred.

The specific reason for this problem was not isolated but after some reflection, the usefulness of the survey participation rate as the regressand becomes questionable. Notice that Y' takes values in the closed interval [0, 1]. Thus, a farm operator who was asked to participate in one survey and cooperated has the same value (1) for Y' as the farm operator who was asked to participate in 10 surveys and reportedly always cooperated. The same situation occurs for noncooperators. A farm operator who was asked to participate once and reportedly did not participate gets the same value (0) for Y' as the farm operator who said he did not comply with 10 survey requests. In addition, the two points, 0 and 1 have a high frequency of occurrence making Y' look bimodal. So, the model was developed by including the number of survey requests (X_1) as a regressor. The model is:

$$y = a + b_1x_1 + \ldots + b_{20}x_{20} + e.$$

Which makes the 1 request, 1 participation and 10 requests, 10 participations distinguishable.

2/ Draper, N. and H. Smith. Applied Regression Analysis (New York: John Wiley & Sons, Inc., 1967, p. 90).

A plot, Y' versus the residuals, shows that the variance is not constant, indicating the necessity to transform the data or to use weighted least squares. A square root or log transformation did not alleviate the problem. So, a weighted least squares analysis with weights $(Y-16)^{-2}$ was conducted, producing an acceptable distribution of residuals and a satisfactory plot of residuals versus predicted values. 3/

Analysis

The prediction of survey participation given stratum membership as defined by the dummy variables and given the total number of survey requests, X_1, is now discussed. Denote this subset of variables by S.

The tempting question to ask in this type of analysis is, "what are the important variables in the model?" Unfortunately, this nice English question is statistically vacuous. We can determine which variable accounts for the largest increase in R^2 given S. By examining the R^2 for all possible subsets of regression variables, the relationships between the regressand and regressor variables can be displayed.

Using all 20 variables in the model, an $R^2 = 0.55$ is obtained. The interest is in what combination of variables accounts for a large part of this R^2, say 90 percent, or a multiple correlation coefficient (mcc) of 0.50. The mcc associated with subset S is 0.34, or approximately 62 percent of the total R^2. Given S, the largest increase in R comes from including the respondent's educational level (X_{12}) in the model. The mcc is 0.39, or 71 percent, of the total R^2. The two-variable model that includes educational level and organizational influence (X_6) accounts for 76 percent of the total R^2. The best three-variable model, given S, has the variables X_6, X_{11}, and X_{12}, accounting for 87 percent of R^2. Recall that X_{11} measures the number of sources of information the respondent uses to help make managerial decisions. Finally, when the variable X_8 (number of groups that uses C&L reports against farmers and ranchers) is included 93 percent of R^2 is reached. From this point, no increase of 1 percent or more is found, thus further analysis is discontinued. The tabulation below summarizes the above results.

At the risk of oversimplifying the multivariate nature of the problem some univariate statistics stating the direction of the relationship between the regressand and regressors are presented. Survey participation increases, as educational level increases. Thirty-seven percent of respondents with less than a high school education said they participated in surveys when asked. However, 43 percent of those respondents with at least a high school education said they always participate when asked.

3/ Daniel, C., and F. Wood. Fitting Equation to Data (New York: Wiley Interscience, 1971 pp 27-28, ch.9).

Variables Accounting for the Largest Increase in MCC Given the Subset S.

Variable(s)	MCC	Percent of $R^2 = 0.55$
Subset S	0.34	62
X_{12}	0.39	71
X_6, X_{12}	0.42	76
X_6, X_{11}, X_{12}	0.48	87
X_6, X_8, X_{11}, X_{12}	0.51	93

These two percentages are significantly different at the alpha = 0.003 level. On the other hand, 39 percent of the respondents with less than a high school education always refuse when asked, and 38 percent with at least a high school education said they never participated. In addition, the perception of organizational influence, and the number of sources of information used by farmers and ranchers, are also positively related to participation. Interestingly, during the survey period, the Farm Bureau and Stockgrowers Association had resolutions to do away with Government C&L surveys. As expected, the number of groups that farmers and ranchers perceived as using C&L reports to hurt farmers and ranchers is negatively related to survey participation.

Summary

A weighted least squares analysis to predict survey participation found that the educational level, perception of organizational influence, number of sources of information used by farmers and ranchers, and number of groups that use C&L reports against farmers and ranchers in North Dakota and South Dakota account for 93 percent of the total mcc, given stratum membership and number of survey requests.

The implications of this analysis on survey participation and SRS public relations programs can be detailed but with some caution. Models built on individuals do not necessarily apply to entire populations. Many models have been developed that indicate increasing a variable X would cause a favorable value in the dependent variable Y. Yet, when funds are spent to increase X, an unfavorable value of Y appears because some underlying variable was not included in the analysis and seriously affected Y. With this warning, survey participation may increase as:

1. Educational level increases,
2. Farm and ranch organizations actively back C&L reports,
3. Farmers and ranchers rely on more sources of C&L information for managing their operation, and
4. Farmers and ranchers perceive that more groups in the agricultural community use reports to help them.

Appendix

Variable X_5 is developed from Q12 of version II of the questionnaire. A scale is developed as follows: the respondent can decide which, if any, combination of C&L reports at the county (C), State (S), National (N), or foreign level (F) are most useful. In addition, the respondents could have replied that they didn't know which level of report was useful. Thus, there are 16 values of X_5 since there are

$$\sum_{i=0}^{4} \binom{4}{i} = 16$$

arrangements of 4 levels of C&L reports, where

$$\binom{4}{i} = \frac{4!}{(4-i)!\,i!} .$$

The arrangements of the levels of corresponding values of X_5 are:

no level mentioned	0
C	1
S	2
C, S	3
N	4
C, N	5
S, N	6
C, S, N	7
F	8
C, F	9
S, F	10
C, S, F	11
N, F	12
C, N, F	13
S, N, F	14
C, S, N, F	15

ORGANIZATIONAL INFLUENCES ON DAKOTA FARMERS

by
Ron Fecso

Introduction

The National Opinion Research Center's survey of Dakota farmers and ranchers provided data on farm organization membership, the respondents' opinions of the organization's policy concerning participation in crop and livestock (C&L) surveys and the frequency of participation of the respondent. About 60 percent of the respondents report that they belonged to one or more farm organizations. This paper presents some unweighted cross tabulations from questionnaire version I which asked the respondent for a subjective evaluation of participation in C&L surveys.

Analysis

The first table presents data concerning the perceived encouragement to participate (or not to participate) in C&L surveys that members of various farm organizations reported. The data are grouped by State and organization. The three organizations mentioned most often, National Farm Organization (NFO), Farm Bureau, and Farmers Union, are listed individually. Miscellaneous livestock organizations are grouped under "livestock." Overall about half the members of organizations felt that they were encouraged to participate in crop and livestock surveys by some organization. It should also be noted that there were more North Dakota respondents who belonged to a farm organization, yet the perceived encouragement as a percentage of those belonging to the group was about the same for each state. There is a significant difference in perceived encouragement between some of the organizations. Only about 20 percent of the NFO or livestock organization members felt that their group encouraged participation, while other groups had encouragement rates generally above 50 percent.

Table II lists the column proportions for a cross tab of the respondents perceived organizational encouragement versus the reply for the subjective participation questions. Although no cause and effect relationship can be established from this table, the possibility that organizations can impact the members response rate is not disputed by the data. It should also be noted that the respondent may assume that their attitude is also that of their organization. About half the members of groups which were perceived as discouraging participation responded "hardly ever." Although the data is combined none of the questioned members of the perceived discouraging groups responded "almost always agree." About half of the members in the organizational groups which were perceived as encouraging participation responded "most of the time" or "almost always." Only 20 percent stated that they "hardly ever" agreed to respond. These relationships hold true when examining the actual position of the organization as defined by the respective Dakota State Statistical Offices (Table 1). The organizations which were in support of the Crop and Livestock (C&L) program were perceived as encouraging participation by over 50% of the respondents, while discouraging organizations were perceived as encouraging participation by only

Table I--Dakota Farmers Responses to the Question:

"As far as you know, does the ORGANIZATION encourage or discourage participation in government crop and livestock surveys?" (Asked only of respondents who said they were members of the organization.)

Name or Type of Organization	Encourage	Discourage	Neutral	Total	Actual Position1/
ND NFO	5	6	6	17	D
SD NFO	2	9	6	17	N
	7 (21%)2/	15 (44%)	12 (35%)	34	
ND Livestock	1	3	7	11	N
SD Livestock	6	7	11	24	D
	7 (20%)	10 (29%)	18 (51%)	35	
ND Farm Bureau	24	8	15	47	S
SD Farm Bureau	12	5	9	26	D
	36 (49%)	13 (18%)	24 (33%)	73	
ND Farmers Union	71	13	38	122	S
SD Farmers Union	33	7	26	66	N
	104 (55%)	20 (11%)	64 (34%)	188	
Others ND & SD	52 (58%)	9 (10%)	29 (32%)	90	3/
Total Mentions	206 (49%)	67 (17%)	147 (34%)	420	

1/ Actual position of the organization during the survey period as evaluated by the ND and SD State Statistical office.
D - Does not support the C&L program or officially against it.
N - Neutral
S - Supports the C&L program

2/ row percentage

3/ The organizations vary in the amount of support. In general this group consists of neutral and passive supporters of the C&L program.

The data presented is unweighted and is not intended for use in point estimation. See the first papers for discussions on weighting.

34% of the respondents. Unfortunately, the members of groups which were perceived to have had neutral or unknown attitudes about participation reported very poor subjective participation rates. Since the perception of organizational encouragement may have a positive effect on response and considering that the majority (93 + 173 = 266 out of 354) of the members belonged to perceived uncommitted or neutral groups, it may be inferred that public relations directed toward these organizations could be a worthwhile endeavor along with continued efforts to reverse the policy of groups which discourage participation.

Table II--Participation and Organizational Influences
Column Percentages

When asked to participant respondent claims to:	Does the organization encourage or discourage participation in government crop and livestock surveys?			
	Encourage	Neutral	Don't know	Discourage
Hardly ever agree	19	40	43	49
Agree only some of the time	31	22	26	33
Agree most of the time, or almost always agree	50	38	31	18
Total	100	100	100	100
Column N =	154	93	173	33

Total N = 453 differs from table I because table I excludes the don't know respondent while table II excludes respondents who were members of organizations, but were not asked the subjective participation question.

THE EFFECT OF A DISPROPORTIONATE, STRATIFIED DESIGN ON PRINCIPAL COMPONENT ANALYSIS USED FOR VARIABLE ELIMINATION

by
Robert D. Tortora

Introduction

Data from a sample survey that is primarily designed for descriptive statistics are often used for multivariate analyses. Typically, the population parameters are estimated by these descriptive statistics. The survey design can be complex, that is, not self-weighting. The observations must be appropriately weighted in order to obtain unbiased estimates of the parameters. Methods of adjusting the survey data which make the design self-weighting and allow ease of computation have been discussed by various authors, including Kish (1965) and Murthy (1967). 1/ 2/ These discussions have been limited to the problem of parameter estimation. However, the issue is also of concern in multivariate analysis.

Beddington and Smith have illustrated the problem of estimating the correlation matrix for complex sample designs. 3/ Their results indicate that proportional allocation leads to little or no impact on the multivariate analysis. However, it is often the case that the analyst has data from a disproportionate design and must develop a model or uncover relationships for the entire population either by choice or by force (insufficient sample size per stratum for the number of explanatory variables). A model over all strata must be developed. Thus, the data analyst must select a procedure on which to base the analysis which should not bear on the final results. Various procedures are available to develop the model. They include (P1) ignore the disproportionate design and analyze unweighted data, (P2) reweight the data (Jones, Sheatsley and Stinchcombe, 1979) and proceed as if using a simple random sample, (P3) introduce dummy variables (d.v.'s) to account for stratum membership (Draper and Smith, 1966), (P4) random elimination, (P5) random duplication, and (P6) random elimination and duplication of the data to obtain a self-weighting design (Kish, 1965).

Because the use of P4, P5, and P6 are dependent on the particular sample eliminated and/or duplicated, only the first three procedures will be considered. Thus, only procedures that avoid replication of the results receive attention.

1/ Kish, L. Survey Sampling (New York: John Wiley & Sons, 1965).

2/ Murthy, M.N. Sampling Theory and Methods (Calcutta: Statistical Publishing Society, 1967).

3/ Beddington, A. and T.M F. Smith. "The Effect of Survey Design on Multivariate Analysis," Model Fitting (Ed. O'Muircheurtuigh and Payne: New York: John Wiley & Sons, 1977).

The impact of the first two procedures, using P3 as the standard is measured in the sequel. Comparisons will be made on actual survey data to study the effects of P1, P2, and P3 on discarding variables using a method based on principal component analysis, since one is often concerned with developing a model where a reduced number of variables account for most of the variation in the data. The desire is to obtain a model with only the pertinent variables. It would be unfortunate if the variables returned were in the model because of improper weighting (or absence of weighting) and not because they account for a large part of the variation.

The following assumptions are made:

1) The data are the result of a single stage, disproportionate, stratified, survey design,

2) There are insufficient observations within each stratum to conduct a separate analysis by stratum, and

3) The d.v. approach is the standard since it produces an "average" multiple regression over the strata. 4/

The Data

The data from version II of the questionnaire were used for analysis in this paper. This version allowed the respondents to describe their past numerical participation rates (number of times responded to surveys divided by number of times asked to respond) during the previous year. Only those respondents who indicated that they had been asked to participate in at least one survey during the year prior to interview were included in the data set. The total weighted sample size of 630 was disproportionately allocated to 10 strata. The sample size was adequate for parameter estimation within each stratum but not large enough to permit multivariate analysis in each stratum without subjectively eliminating variables.

Nineteen variables (table 1) were considered for procedures P1 and P2. However, for P3, nine additional dummy variables were added to account for the 10 strata in the sample design. The variables can be classified into the following categories:

(1) Six background information variables such as total number of cattle and total cropland acres,
(2) Thirteen Crop & Livestock Evaluation (C&LE) variables such as source of agricultural information, usefulness of agricultural statistics, attitudes about confidentiality of survey data, and
(3) For procedure P3, the nine dummy variables.

4/ Kendall, M. Multivariate Analysis (New York: Hefner Press, 1975).

Table 1--
Variable Descriptions

	Variable Number	Description
Background Information Variables	1	Age of farm operator
	2	Education of farm operator
	3	Total acres of cropland
	4	Total number of cattle
	5	Total number of pigs
	6	Total number of crops
Crop & Livestock Evaluation Variables	7	USDA divulge data to private company
	8	USDA divulge data to another gov't agency
	9	Number of sources of farm information
	10	Influence to farm organization on participation
	11	Impact of C&L reports
	12	Use of C&L reports by others aiding farmers
	13	Capability of Crop and Livestock reports to harm farmers
	14	Number of groups that use C&L reports to harm farmers
	15	Why farmers and ranchers participate in surveys
	16	Usefulness of C&L reports for farm management
	17	Who benefits most from C&L Livestock reports
	18	Accuracy of C&L reports
	19	Geographic use of C&L reports

The variables are seperated into these categories because the first relates farm and farm operator characteristics and is, in a sense, given. They cannot be affected by any programs, say, to improve survey participation rates. On the other hand, changes in the C&LE variables have the possibility of impact on survey participation. For example, if the confidentiality variable accounts for a large part of the variation, it may be possible to improve the interview introduction and also initiate a public relations program to increase awareness of confidentiality with the hope of improving survey response rates. This second category represents the variables the analyst is often concerned with detecting, since their importance can cause changes in management and fiscal policy towards improving survey participation.

Unweighted and Reweighted Data

Unweighted data are usually used when conducting a multivariate analysis and when the data comes from a proportional allocation. However, the design may be disproportionate and the analysis conducted on un-

weighted data. If the variables associated with the model are dependent on stratum membership the under- or over-representation of certain subpopulations may affect the outcome of the analysis.

On the other hand, it is natural for the analyst to consider re-weighting the data in attempting to avoid this under- or over- representation problem. For the purpose of this paper we will use the method of reweighting presented in Jones, et. al. (1979). Procedure P2 uses a method developed by Kish (p. 420, 1965) to measure the increase in variance caused by disproportionate allocation when proportionate allocation is optimum. Under the constraint that the reweighted sample size is equal to the raw sample size n, the relative efficiency of the sample is computed using

$$nV^2 = \Sigma W_h k'_h (1 - f/k'_h)$$

where $W_h = N_h/N$, the stratum weight k'_h equals the initial element weight, and f equals n/N, the overall sampling fraction. For the data described in section 2, the relative efficiency is 0.8143 or just over 80 percent of that of a proportionate sample of equal size. Final weights, those values attached to the data to reweight it, are the products of the initial weights and the relative efficiency of the sample. These values are summarized as folows:

Stratum	Initial weights	Final weights
1	0.094	0.076
2	1.352	1.101
3	0.313	0.255
4	1.453	1.183
5	1.080	0.874
6	0.338	0.275
7	0.079	0.064
8	1.350	1.099
9	1.280	1.042
10	1.024	0.834

A more detailed description of this procedure can be found in Jones, et. al. (1979).

Notice the use of weighted data that produces unbiased estimates over the entire population is purposely omitted since these initial weights are close to the weights used in P2.

P2 allows for slightly easier computation of estimates of population parameters since it avoids computing estimates for each stratum and then combines these into an estimate for the population. Thus, P2 allows for the use of statistical software packages in which the data is assumed to come from a simple random sample. The standard errors of estimates cannot be calculated using the algorithms in the package. The design effect must be calculated in order to compute these estimates of variability.

The reweighted data may be used for the elimination of redundant variables if a multivariate analysis is being conducted. Does the reweighting have an impact on the final variables retained for further analysis?

The Dummy Variable Approach

The dummy variable or pseudo-variable approach is useful for modeling when some of the independent variables are discrete rather than continuous. Draper and Smith (1966) use this technique in regression analysis to account for data that occurs at two or more distinct levels. These variables then take account of the fact that separate deterministic effects are produced on these different levels. For K levels, K-1 dummy variables are required. For example, suppose we have three strata from which responses have been obtained. Then two dummy variables, Z_1 and Z_2 say, are required to account for the strata. They are:

$$\begin{aligned}(Z_1, Z_2) &= (1,0) \text{ for stratum 1} \\ &= (0,1) \text{ for stratum 2} \\ &= (0,0) \text{ for stratum 3.}\end{aligned}$$

Kendall (1975) has shown that these dummy variables produce a regression line. The slope was the weighted average of the lines, had regressions been calculated for each stratum. Thus, as Beddington and Smith recommend, it would be appropriate to conduct the analysis by stratum. Unfortunately, there is often an insufficient parameter to sample size relationship to conduct such an analysis (10 observations per independent variable). 5/ Therefore, the use of dummy variables presents a viable alternative in this situation.

Variable Elimination and Principal Component Analysis (PCA)

Variable elimination is important to the data analyst because redundant or colinear variables are removed. Variables are often present that complicate the analysis yet do not provide additional knowledge. Thus, by eliminating these extraneous variables, efficiencies are realized with a consolidated measurement instrument and with fewer variables to be analyzed, particularly as future investigations are conducted. The variable elimination technique used in this paper has been studied by Jolliffe using 587 artificial data sets (1972) and 4 real data sets (1973). 6/ 7/ It was found to perform as well as, or better than, various other methods of variable elimination.

5/ Kendall, M. Personal communication. 1978.

6/ Jolliffe, I. T. "Discarding variables in a principal component analysis I: Artificial data," Applied Statistics, Vol 21 (1972), pp. 160-173.

7/ Jolliffe, I. T. "Discarding Variables in a principal component analysis II: Real Data", Applied Statistics, Vol. 22 (1973), pp. 21-31.

A Principal Component analysis (PCA) is performed on all p variables, and the eigenvalues inspected. If p' eigenvalues are less than 0.7 (a value determined empirically) the corresponding eigenvectors are considered in turn, starting with the eigenvector associated with the smallest eigenvalue and so on until all eigenvectors with corresponding eigenvalues less than 0.7 have been considered. One variable is then associated with each of the p' eigenvectors, namely the variable which has the largest coefficient in the eigenvector under consideration and which has not already been associated with a previously considered component. The p' variables associated with the p' eigenvectors are then eliminated. The remaining p - p' variables are retained for further analysis. In order to compare principal components for the full and reduced sets of data, the product moment correlations between the full and reduced set of data are computed (Jolliffe, 1973).

Suppose the entire set of data contains n observations measured on k variables $x_1, x_2, \ldots, x_k$. All analysis is done on the correlation matrix and the sample correlation r_{ij} between each pair of variables (x_i, x_j) is computed.

Any principal component is a linear combination of the variables in the set. For the entire set of p variables, it can be written as:

$$y = a_1x_1 + \ldots + a_px_p,$$

where the a_j's are constant. For the reduced set of p-p' variables it can be written as :

$$z = b_1x_1 + \ldots + b_px_p,$$

where the b_i's are constant, but here all p' constants corresponding to eliminated variables are zero.

Using the n observations for y and z the correlation coefficient between them can be calculated--call it r. If the first k components are of interest for the full data set, then the similarity between components for the entire and the reduced data sets are defined by:

$$Q = \left(\sum_{i=1}^{k} q_i r(i)\right) / \left(\sum_{i=1}^{k} q_i\right),$$

where r(i) is the maximum value of r between the i^{th} component for the full set of data and any component for the reduced set and q_i is the proportion of the total variation accounted for by the i^{th} component in the entire data set. So the similarity between components and the weights are proportional to the amount of variation explained by the first few components of the entire data set.

Comparison of the Three Procedures on Variable Elimination

A PCA was conducted for each procedure. Thirteen variables were retained for the unweighted data, the PCA on the reweighted data retained nine variables, and the PCA, when the 28 variables were included,

retained 19 variables. The following tabulation gives the variables retained by category of variable.

Variables Retained by Category

Procedure	Background Information Variables	Crop & Livestock Evaluation Variables	D.V.
P1	3 4 5 6	8 9 10 11 12 16 17 18 19	
P2	4 5 7	9 10 11 12 15 16 18	
P3	2 3 6	9 10 11 14 15 16 18 19	1-7 9

Comparing P1 and P3, it is apparent that P1 retained two of the background information variables that P3 retained, but adds two unnecessary background information variables. On the other hand, P2 had no matches with P3 for background information variables. Six of the nine C&LE variables retained by P1 matched with P3. P1 retained three variables that are not in P3 and also two variables (14 and 15) were missed by P1. Procedure P2 also had six variables matching with P3; it added two (7 and 15) unnecessary C&LE variables and missed two variables (14 and 19). Notice that d.v. eight was eliminated by the PCA. This combines strata 8 and 10, the small-scale cattle operations in North Dakota. Note that the retention of variables 4 and 5 in P1 and P2 may be the result of what otherwise accounted for stratum membership.

The following tabulations show the similarity between all variables and the reduced set of variables by procedure. Nine components were used for comparison since P2 retained the fewest (nine) variables.

Measure of similarity, r, Q, between components for all variables and reduced set of variables by procedure

	P1	P2	P3
r_1	0.827	0.430	0.891
r_2	.998	.352	.960
r_3	.998	.525	.984
r_4	.879	.531	.837
r_5	.872	.390	.903
r_6	.912	.735	.586
r_7	.986	.164	.859
r_8	.837	.973	.996
r_9	.788	.230	.731
Q	.893	.503	.868

Procedures P1 and P3 were nearly equivalent with a weighted average of correlations of 0.893 and 0.868, respectively. P2 fell sharply below P1 and P3 with a weighted average of 0.503. Examination of the individual correlations for P2 indicates that the correlations for P1 and P3 were about twice as large as the correlations for P2 in six of the nine components.

In summary, P1 matches eight, adds five unnecessarily, and missed three variables when compared to P3. The components retained by P2 were not as similar to the full data set as the components retained by P1 and P3.

Summary

The effect of three procedures to prepare survey data for analysis were examined for a method of variable elimination based on principal component analysis. The data was obtained from a single-stage, disproportionate, stratified design, and the analysis was conducted on (P1) unweighted data, (P2) reweighted data, and (P3) additional dummy variables to account for stratum membership (the standard).

P1 came closest to matching the variables retained in P3, but it also added the most extraneous variables. A high similarity existed between the complete and reduced data sets for P1 and P3 while P2 retained little similarity. Thus, the procedure selected can potentially affect the results. The procedure that prepares the data caused variables to be retained or eliminated without sufficient statistical justification.

U.S. GOVERNMENT PRINTING OFFICE: 1981-360-931:407