COMPARISON STUDY ON THE QUALITY OF FINANCIAL DATA COLLECTED THROUGH PERSONAL AND TELEPHONE INTERVIEWS

Pierre Caron, Pierre Lavallée, Statistics Canada
Pierre Caron, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A-0T6

Key Words: interview mode, non-response, data quality

1. INTRODUCTION

In business surveys, it is common practice to ask detailed financial information about the responding units. Such interviews tend to be time consuming and intrusive. These are some of the reasons why in many instances the collection of such data is conducted in a face to face interview with the respondent. The Farm Financial Survey (FFS) conducted every two years at Statistics Canada is one example of such a survey. It collects detailed financial information on variables such as assets, liabilities, expenses, income and non-farm finances for agricultural operations in Canada. A typical interview lasts in the neighborhood of 30 minutes. The sample, which covers all provinces, usually includes about 12,000 farming operations. Because of the high cost of collection through personal interviews, the future of the survey was uncertain after the 1996 occasion. In 1996, it is worth noting that the unit cost of one personal interview was roughly 10 times the unit cost of one telephone interview (note that a small number of telephone interviews are usually conducted for farms in remote areas). Some options that were available were not considered practical. Examples of such include decreasing the sample size to cut costs or performing the survey at less frequent intervals. Given that the sample size is already at a minimum to obtain quality estimates by province and farm type, and that going more than two years between survey occasions is undesirable, both these options were rejected. Another possibility was the use of telephone interviews for data collection which would reduce costs considerably, but could possibly affect the quality of the data.

After the collection of the 1996 FFS, it was decided to conduct a test survey to determine if a change to telephone interviews for data collection would have a negative effect on the quality of the data. The collection for this test was done in December of 1996. The major concerns with switching to telephone interviews were the following:

1 - Would the quality of the data collected by telephone be as good?

2 - Would respondents agree to supply detailed financial information over the phone and would they give accurate figures?

3 - Would the length of the interview and the complexity of some of the concepts be a problem for a telephone interview?

4 - Would the non-response rates increase significantly, knowing in advance that telephone interviews tend to have sightly higher refusal rates than personal interviews?

In section 2, we describe in detail the test survey and present the basis for our analysis. In sections 3 and 4 we discuss the steps in constructing a frame and the sample selection. We follow this with a brief description of the data collection in section 5. Finally we present the theoretical foundation and the results of our analysis in section 6. We conclude with a brief summary and outline of what was done for the 1998 FFS, which went to the field in March of 1998.

2. DESCRIPTION OF THE TEST

The test survey was designed in a way that would allow us to test the two modes of data collection against one another for many quality indicators. This included some standard comparisons such as response rates, refusal rates, average length of interview, partial non-response rates, etc. A second more in-depth analysis of the data tried to evaluate the quality of the actual data reported by the respondents. This was accomplished by comparing the respondents' individual answers to benchmark values coming from taxation records. This comparison could only be done on the variables that were common to both sources of data. If one of the collection modes were superior to the other, its reported data would be closer to the benchmarks than for the other mode. The distances from the "true" values were analyzed using a non-parametric statistical test. For this comparison to be valid, both sources of data had to cover the same time period. We therefore asked respondents to report for the same reference periods as they did for their taxes. In most cases, this corresponded to the calendar year.

The sample itself covered most provinces in Canada so that regional differences could be detected. Once an overall sample size had been determined (based on precision desired and costs), it was divided equally into three groups for data collection. The first consisted of personal interviews (group P), the second comprised telephone interviews from the regional offices (group T) and the third involved telephone interviews with the possibility of conversion to a personal interview upon the respondents' request (group T/P). For this last group, the telephone interviews were conducted from the interviewers' homes and the respondents were offered personal interviews only as a last resort to avoid a refusal. This conversion rate was an important statistic in deciding whether or not telephone interviews are justifiable for FFS.

Note that there was no interest in using the data to estimate population means and totals. The data was used strictly for data quality analysis purposes.

Before establishing a sampling plan and proceeding with the sample selection, we had to build a list frame from which the units could be selected. This posed some constraints for this survey, as discussed in the following section.

3. FRAME CONSTRUCTION

Although the purposes of the test were quite different than those of the regular FFS, the steps in selecting a sample remained roughly the same, the first one being the creation of a list frame from which a sample could be drawn. The starting point was the list frame that was used for the 1996 FFS. The target population for FFS includes all agricultural operations with the following exclusions: institutional farms, community pastures, farms on Indian reserves, multi-holding companies and operations with less than $2,000 in sales from agricultural activities,.

In addition to satisfying the usual FFS criteria, the following conditions had to be satisfied to be eligible for the test:

1 - Taxation data had to be available for the operations since part of the analysis depended on it. Note that an agriculture operation can be matched to either a T1 (unincorporated) or a T2 (incorporated) record. The majority of farms in Canada are businesses that are not incorporated, and therefore file a T1 tax report with Revenue Canada. Taxation data was available for roughly 70,000 operations in Canada through the Tax Data Program.

2 - For operations that are not incorporated (T1), we had to make sure that a detailed balance sheet was supplied when they filed their income tax since most of the variables used in the analysis come from the balance sheet. This condition was automatically satisfied for the T2's, as they are required by law to supply this information.

3 - We had to have a link between the list frame and the tax data frame. A yearly statistical record linkage between these two sources is done at Statistics Canada in the context of the Whole Farm Data Program. For the purpose of the study, the 1995 linkage was not completed in time and we therefore had to use the 1994 links. Furthermore, in an attempt to keep only the strongest links, we used only links that were one to one between the frame and the tax data. This eliminated operators with multiple operations who may file only one income tax form for all their operations.

4 - Finally, operations that participated in the 1996 FFS were also excluded since the response burden of having them respond twice to the same survey for the same reference period was not justifiable.

The initial FFS frame included over 170,000 records. When all cleaning up was completed, the frame for our consisted of only 5,870 agricultural operations, of which 1,900 were not incorporated. The remaining operations were incorporated.

4. SAMPLE SIZE AND SAMPLE SELECTION

One consequence of using such a small frame was that the Atlantic region had to be excluded from the study because of very low counts. Even if we had sampled all units, we would not have had a sufficient number of units to perform an adequate analysis.

Minimal sample sizes were determined for each province in a way that statistically significant differences between the collection modes could be detected with a confidence level of 95%. It was determined that an overall sample of 2,400 operations would yield accurate results at the provincial level. The actual distribution of the provincial sample by interview type, province and operation type are given in tables 1 and 2 respectively.

Table 1: Provincial Sample Distribution

	Interview Type			
Province	T	T/P	P	Total
Québec	177	177	177	531
Ontario	136	134	133	403
Manitoba	98	93	107	298
Sask.	134	114	98	346
Alberta	126	139	135	400
BC	134	134	133	401
Canada.	805	791	783	2,379

Table 2: Sample Distribution by type of Operation

	Type of Operation		
Province	T1	T2	Total
Québec	375	156	531
Ontario	153	250	403
Manitoba	72	226	298
Sask.	32	314	346
Alberta	51	349	400
BC	72	329	401
Canada	755	1,624	2,379

Note that this allocation by type of operation does not reflect the true population since most agricultural operations are not incorporated. The T1's are therefore under represented. This is a direct consequence of most T1's not supplying a balance sheet with their income tax. This also explains why the sample size for Québec was slightly higher than for other provinces; Québec had a much higher rate of T1's with a balance sheet than other provinces and was therefore over-sampled.

The actual selection of the 2,400 operations did not result from a truly random process. In order to keep the costs of collection to a minimum, the sample for personal interviews was selected using postal codes and input from the regional offices to minimize interviewer traveling. Clusters of various sizes were then formed. This also explains why sample sizes between the 3 groups are sometimes slightly different within a province. For the telephone interviews, we did not have such a constraint. The fact that this is not truly a probability sample does not create any serious problems since we are not using the data to obtain population estimates. The analysis is performed on the actual measured data. This implies a model-based approach which does not refer to the population per se but rather to the random process which dictates the values that the individual population records assume. Despite this fact, some efforts were made to ensure that the sample covered each province adequately for each type of interview.

5. DATA COLLECTION AND PROCESSING

Data collection was spread over a period of one month. In order to replicate the actual FFS survey as much as possible, each respondent, regardless of the interview mode, was sent an introductory letter and a small publication on agricultural statistics. The latter usually serves as an incentive for encouraging response and has proven to work well for that purpose in the past. In addition, respondents to be contacted by telephone were sent a copy of the questionnaire. Again, this was done to replicate the real life situation of what would happen if FFS became a telephone survey.

Telephone interviews with no possibility of conversion to a personal interview were conducted from the regional offices whereas those which offered that option were conducted from interviewers' homes. Both used a paper and pencil interview. It is fair to assume that the actual quality of data collected by telephone could be improved by using computer assisted telephone interviewing (CATI). If FFS did decide to switch to telephone interviewing, such an application would be developed.

Data capture was performed using a modified version of the 1996 FFS data capture system. This used a generalized data collection and capture system developed at Statistics Canada (DC2).

After data collection, a subset of the edits used in the 1996 FFS were used to verify the data. The idea was to identify errors and correct them when possible or leave them as missing values otherwise. We did not want to use imputation for missing values because the analysis is so dependent on the comparison between micro level data. Imputed values could create some large discrepancies between the two sources of data which could lead to false

conclusions. Both micro and macro-editing were used in identifying erroneous values.

6. RESULTS

6.1 QUALITY INDICATORS

As we mentioned in our introduction, two separate analyses were performed on the collected data. First, overall quality indicators were derived for each interview mode. These indicators include non-response rates, refusal rates, no-contact rates, average length of interview, and average number of missing cells per questionnaire. They are summarized in the tables that follow.

Table 3: Refusal rates by interview type

	Interview Type			
Province	T	T/P	P	Total
Québec	6.9%	5.0%	6.9%	6.3%
Ontario	9.7%	4.5%	9.0%	7.7%
Manitoba	10.3%	16.5%	11.4%	12.6%
Sask.	16.7%	18.0%	19.6%	17.9%
Alberta	16.7%	15.2%	14.1%	15.3%
BC	20.0%	17.4%	12.4%	16.6%
Canada	13.1%	12.0%	11.7%	12.3%

Table 4: No contact rates by interview type

	Interview Type			
Province	T	T/P	P	Total
Québec	1.1%	2.8%	5.2%	3.0%
Ontario	3.0%	2.2%	1.5%	2.2%
Manitoba	2.0%	7.7%	1.9%	3.8%
Sask.	5.3%	0.9%	7.2%	4.4%
Alberta	4.8%	4.4%	3.7%	4.2%
BC	3.9%	0.8%	3.9%	2.8%
Canada	3.3%	2.9%	3.9%	3.4%

Although refusal and no contact rates varied slightly from one interview mode to the other, we cannot conclude that one collection mode gives better response rates than the other. Results vary by province and it is interesting to note that the no-contact rate at the Canada level is highest for personal interviews. For the refusal rate, we note bigger differences between the provinces than between the interview modes. The western provinces show significantly higher non-response rates (no-contacts + refusals) then the eastern provinces, and this independent from the collection method.

Table 5: Average Number of missing cells per questionnaire (excluding non-respondents)

Average Number of missing Cells			
Tel.	Tel./Pers.	Pers.	Total
1.2	1.2	0.9	1.1

Table 5 is meant to give an idea of the partial non-response for the various interview modes. In all three cases, the average was about 1 missing cell per questionnaire. Such missing cells would usually be the result of the respondent refusing to answer a question or simply not knowing the answer to a question.

Table 6: Average length of Interview

Average Length of Interview			
Tel.	Tel./Pers.	Pers.	Total
27 min.	35 min.	50 min.	36 min.

Table 6 shows the average duration of an interview for the three collection modes. By far, personal interviews took the longest on average with 50 minutes (this excludes travel time for the interviewer). One reason that may explain this difference is that respondents from telephone interviews received a copy of the questionnaire prior to being contacted and a number of them had already completed the questionnaire (they were asked to complete as much of the questionnaire as they could before being contacted). This saved time for the telephone interviews. The difference of 7 minutes on average between the two types of telephone interviews cannot be so easily explained, although other factors come into play. For instance, the first group of telephone interviews were conducted from the regional offices whereas the second group (with possible conversion to personal interview) were conducted from interviewer homes.

When a respondent had completed a FFS questionnaire,

he was asked to share his data with Agriculture an Agri-Food Canada. If he refused the data sharing agreement, he would eventually be treated as a refusal. We found that 3% of the respondents refused the data sharing agreement, regardless of the interview mode, which corresponds to the rate usually observed for FFS.

One last statistic of key interest was the conversion rate from telephone to personal interview for respondents who were given the option (group T/P). Out of 791 such respondents, only 14 had to be converted to a personal interview. This represents 1.8% of the respondents. This figure represents an upper bound since there is no guarantee that all 14 would have refused the telephone interview (3 of the 14 cases were never contacted by telephone).

None of the indicators studied so far have proven to show that telephone interviews would not be a viable solution for the FFS.

6.2 COMPARISON TO TAX DATA

The second part of the analysis dealt with the comparison of the survey data to the tax data, assuming the latter to be correct. This analysis was possible for close to thirty variables which are available on both sources of data. The results are based on the notion that if one collection mode is superior to the other, we would expect the distribution of the differences between the two sources (for one particular variable) to be more clustered around the value 0.

Although some classical parametric analysis such as analysis of variance were performed on the data, we found that the skewness and the magnitude of the data being analyzed did not perform well for such analyses. In particular, such tests are not robust in the presence of outlying values. Although many efforts were made to remove true erroneous values from the data, such values likely still existed in the final data. In addition to the usual reasons for errors (capture errors, etc.), there was an additional source here in that the linkage between the tax data and the frame (see section 3) is a probabilistic one that is likely to contain some accepted false links. We did not have a quality indicator for the validity of the links.

Because of these reasons, we decided to base our analysis on a non-parametric test which is robust and makes no prior assumption about the distribution of the data. The non-parametric test we used is called the *Run Test*. It allows to test for equality of the distribution functions of two random variables.

Let X_i be the benchmark (taxation) value for a specific question for respondent i.

Let Y_i be the reported value on the FFS questionnaire for the same question for the same respondent.

Let $D_i = Y_i - X_i$.

The distributions we want to compare are the D_i for the personal interviews (D_{ip}) and the telephone interviews (D_{it}). If there is no significant differences between the interview modes, the distributions should be similar. Therefore we wish to test $H_0 : F(D_{it}) = F(D_{ip})$.

We need to first define the concept of runs. Suppose we have n_1 observations from one distribution (say D_{it}) and n_2 observations from another distribution (say D_{ip}). The combination of the two sets of observations into one collection of n_1+n_2, placed in ascending order could yield an arrangement like

t t t p p t p t p p t t

where t denotes an observation coming D_{it} and p an observation from D_{ip}. Each underlined group represents one run. In the example above, there would be 7 runs.

Let R be the random variable for the number of runs in the combined ordered sample. Under H_0 all permutations of the n_1 observations of D_{it} and the n_2 observations of D_{ip} have equal probability. A test based on the number of runs can easily be derived for testing H_0. A small number of observed runs usually leads to the rejection of H_0. That is, the critical region is of the form $r < c$, where the constant c is determined using the p.d.f. of R to yield the desired significance level. One advantage of the run test is that it is sensitive both to differences in location and differences in spread of the two distributions.

Now, when n_1 and n_2 are large, R can be approximated by a $N(\mu,\sigma^2)$ where

$$\mu = E(R) = \frac{2n_1n_2}{n_1+n_2} + 1 \quad \textit{and} \quad \sigma^2 = \frac{(\mu-1)(\mu-2)}{n_1+n_2-1}$$

It follows that $Z = (R - \mu) / \sigma \sim$ approx. $N(0,1)$. The critical region for testing $H_0 : F(D_{it}) = F(D_{ip})$ is of the form $z < - z(\alpha)$ where $z(\alpha)$ is the $100(1-\alpha)$ upper percentile of the $N(0,1)$ distribution.

This test was performed for each variable common to both sources of data both at the national and provincial levels using α=0.05. This added up to over 400 run tests.

Out of all of them, H_0 could only be rejected 35 times, and in most of these cases it was by a very narrow margin.

7. CONCLUSION

Two separate comparative studies were performed on the two collection modes and neither one proved to show that the quality of data collected through telephone interviews would be of lesser quality then that collected through personal interviews. Based on the results of this study, FFS 1998 proceeded last March with telephone interviews exclusively. The change in interview mode generated more savings than what was actually needed, allowing the difference to be used to increase the sample size by fifty percent (from 12,000 to 18,000). This increase was necessary to improve the quality of the estimates for some of the smaller domains of interest. In addition, a CATI application was developed for the survey. This should further increase the quality of the data and save time and money as most of the data capture and much of the verification can be conducted directly in the field during the interviews. Early results from the 1998 FFS show that the response rates have indeed stayed at the same level they were in the past.

REFERENCES

Caron, P. (1996), *Traitement des données pour l'enquête financière sur les fermes de 1996 (EFF 1996)*, Internal Document, Statistics Canada, Business Survey Methods Division.

Hamel, N. (1996), *1996 Farm Financial Survey: Design Documentation*, Internal Document, Statistics Canada, Business Survey Methods Division.

Hogg, R., and Tanis, E. (1983). *Probability and Statistical Inference*, 2nd Edition. New York: Macmillan

Lavallée, P. (1996), *Comparaison de la qualité entre les entrevues personnelles et téléphoniques pour l'EFF*, Internal Document, Statistics Canada, Business Survey Methods Division.