

A Study of Nonrespondents in the Canadian Vehicle Survey

Martin Beaulieu

Business Survey Methods Division, Statistics Canada,
100 Tunney's Pasture Driveway, section 17-P, Ottawa, ON, Canada, K1A 0T6
(martinj.beaulieu@statcan.ca)

Abstract

While many efforts are put into following up respondents, nonresponse remains an important issue in the Canadian Vehicle Survey (CVS). The low response rates may lead to biased estimates if the response mechanism is non-ignorable. The sponsors of the survey, Transport Canada and Natural Resources Canada are interested in knowing whether a nonresponse bias exists and, if so, its direction and magnitude. A study of nonrespondents was conducted in early 2006 with three main objectives: (i) estimate the nonresponse bias, (ii) determine the reasons why nonrespondents did not complete the survey questionnaire in order to improve the data collection procedures, and finally (iii) collect nonrespondents' characteristics in order to improve the nonresponse treatment. This paper will describe the methodology of the CVS study of nonrespondents and then present the analysis of the results.

Keywords: nonresponse bias, respondent follow-up, travel surveys, characteristics of nonrespondents.

1. Introduction

The Canadian Vehicle Survey (CVS) was developed in 1999 at the request of Transport Canada (TC). The goal of the survey is to provide quarterly and annual estimates of vehicle-kilometres (distance traveled by vehicles on roads) and passenger-kilometres (sum of the distances traveled by individual passengers, including the driver) by characteristics of: vehicles, users, trips, time of day and fuel purchased. The results are the main source of road vehicle use information for researchers and interested members of the public. Prior to 2004, the survey was sponsored by TC. Since then, the survey has been co-sponsored by TC and Natural Resources Canada (NRCan). They plan to combine the CVS data with other available data to improve road safety, monitor fuel consumption and deal with the impact of vehicle usage on the environment.

While much effort is put into the follow-up of respondents, the response rates for the CVS are generally between 55 % and 60 %. Such response rates pose a high risk of nonresponse bias. In the case of the CVS, the direction and the magnitude of this potential bias are unknown. Some would say that vehicle owners using their vehicle often and travelling long distances would not take time to fill in their log after each trip. While others would say that those who do not use their vehicle much do not respond as they feel their data is not relevant for the purpose of the survey. To answer these questions, a study of nonrespondents was conducted in early 2006 with three main objectives: (i) estimate the nonresponse bias, (ii) determine the reasons why nonrespondents did not complete the survey questionnaire, and finally (iii) collect characteristics of the nonrespondents in order to improve the nonresponse treatment.

After a short overview of the CVS in Section 2, this paper will describe the measures taken to deal with nonresponse in the CVS in Section 3, followed by the description of the methodology of the study of nonrespondents in Section 4. The analysis of the results will be discussed in Section 5, followed by a brief description of other studies related to response/nonresponse issues in Section 6. Finally, a conclusion and some future challenges are presented in Section 7.

2. CVS Overview

The target population of the CVS includes all on-road vehicles registered in Canada, except special equipment (such as street cleaners or snowploughs), motorcycles and, since 2004, buses. The survey frame consists of the vehicle registration lists provided quarterly by the provincial and territorial governments.

The population is stratified by jurisdiction (13 provinces/territories), vehicle type (light vehicles and two types of heavy vehicles) and age ("old" or "new"). The stratification by age is performed

once a year, prior to the selection of the Quarter 1 sample. The same stratification by age applies for all quarters in the same year. It uses data from previous years and identifies a splitting year which minimizes the variance for the vehicle-kilometres estimate within each stratum. The year identified is the last year of the “old” group. This stratification process results in 78 strata.

The CVS uses a two-stage design. At the first stage, a stratified systematic sample of vehicles is selected. The sample size allocation is done proportionately to the cubic root of the population size of each stratum and then the vehicles are selected systematically by postal code, to assure good provincial coverage. The target sample size is 5,375 vehicles in the provinces and 2,800 vehicles in the territories. At the second stage, a start date included in the reference quarter is randomly selected for each vehicle. The start date is the first of a cluster of consecutive days for which the driver of the selected vehicle is asked to report trips.

Data collection for the provincial component of the survey consists of two steps. The first step is a computer assisted telephone interview (CATI) with the owners of the sampled vehicles. This interview is used to collect some general information on the usage of the vehicle as well as to ask the respondent to complete a trip log specific to the type of vehicle. The trip log is then mailed out as a second data collection step. On this trip log, the respondent reports every trip made with the selected vehicle during a specific reporting period (20 trips for light vehicles and 7 days worth of trips for heavy vehicles). For the purposes of the CVS, a new trip occurs each time the driver or a passenger gets in or out of the vehicle. Respondents are also required to report two fuel fill-ups in the fuel supplement (also referred as the fuel log) at the end of the log. If respondents cannot be contacted by phone at the first step, the trip log is mailed out with a short questionnaire to collect some of the information usually collected during the CATI. The territorial component of the survey consists of two short questionnaires. One is mailed to the respondents at the beginning of the quarter and the other is mailed at the end of the quarter. The first questionnaire asks respondents to record the odometer reading at the beginning of the first day of the quarter. All those returning the first questionnaire are mailed a second questionnaire asking them to record the odometer reading at

the beginning of the first day of the next quarter. These two odometer readings allow the calculation of the distance the vehicle was driven during the quarter.

Nonresponse treatment consists of reweighting in the case of total nonresponse and imputation for partial nonresponse. The edit and imputation process is performed with a complex system which uses many imputation methods (deterministic, donor, regression models etc.). A detailed description can be found in Landry (2005).

The estimation strategy was built to use the information from both the trip log and the fuel log in an optimal way. Since respondents have to report odometer readings at the time of a fuel purchase, the fuel log can be used to estimate the vehicle-kilometres. To produce these estimates, it was decided to use the log (trip log or fuel log) with the longest reporting period (in terms of days). In order to keep consistency between the vehicle-km and the passenger-km estimates for cases where the fuel log is used, an adjustment factor is used to estimate the passenger-km estimates and the vehicle-km estimates by trip characteristics as the passenger information trip characteristics are not on the fuel log. The estimation strategy is described in detail in Beaulieu (2005).

3. Nonresponse Issue

Nonresponse has been an issue for the CVS since its beginning in 1999. Response rates from the provinces to the CATI are usually around 60% and return rates of the trip log are between 35% and 40%, with very little variation from one quarter to another. Many measures have been put in place throughout the years in order to prevent nonresponse and to measure its impact.

3.1 Nonresponse Prevention

With the response rates observed in the CVS, it is crucial to put in all of the necessary effort to reduce nonresponse. One key element to avoid nonresponse is to have quality contact information on the survey frame. The CVS uses a database identifying changes in vehicles ownership to update the addresses of the owners of the sampled vehicles.

The collection mode can also have an impact on response rates. A study that was conducted in

2000 by the CVS team compared a dry mail-out of the trip log to a pre-contact telephone interview prior to sending the trip log. This study showed that response rates in the group that directly received the trip log by mail were 50% lower than the ones in the group with a CATI as a pre-contact. Following this study it was obvious that despite its high cost, the pre-contact using CATI should be kept.

Intense and relatively costly follow-up procedures were also introduced to prevent nonresponse. Following the initial CATI and the log mail-out, a phone call is made to the vehicle owners on the assigned date when they have to start reporting trips. The goal of this phone call is to remind them to fill out the log and that they can contact Statistics Canada toll-free at any time if they have questions. A reminder letter is also sent on the first week. If no response is obtained after nine weeks, a short questionnaire is sent to the vehicle owner. Finally, a phone follow-up is made on the 10th week if a response has still not been received.

Finally, always looking for new ways to improve the response rate, a study on the use of indirect incentives was conducted during the second and third quarters of 2005. The samples for these two quarters were split in three groups. The first group received a mechanical pencil with the log, the second group received a mechanical pencil and a key-chain along with the log and the third group, the control group, did not receive any incentives. The idea of the mechanical pencil came from focus groups, where some participants mentioned that they always kept the log in the car, but did not always have something with which to write to write. In those cases, they filled in the log later (or simply omitted to fill in the log after a trip) which led to errors or inconsistencies. The response rate for the mechanical pencil group was 2% higher than the control group, but this was not statistically significant using Fisher's Exact Test.

3.2 Measurement of the Impact of Nonresponse

After nonresponse treatment (see Section 2) and estimation, the variance due to nonresponse is estimated using the System for Estimation of Variance due to Nonresponse and Imputation (SEVANI) developed by Beaumont and Mitchell (2002). In order to better inform users on the real precision of the estimates, the published

coefficients of variation take into account both the sampling variance and the variance due to nonresponse and imputation.

Despite many efforts to avoid nonresponse and a meticulous treatment of nonresponse, such low response rates lead to a high risk of bias due to nonresponse. Therefore, a follow-up study of nonrespondents was very important to estimate the direction and the magnitude of this potential bias.

4. Methodology of the CVS Study of Nonrespondents

The CVS study of nonrespondents was conducted with three main objectives: (i) determine if the estimates of vehicle-km produced are biased and if so, what is the direction and the magnitude of that bias; (ii) determine the reasons for nonresponse in order to improve the data collection methods and (iii) obtain some characteristics of the nonrespondents, which could be used for nonresponse treatment.

4.1 Design of the Study

Nonrespondents from the third quarter of 2005 from the provinces were the target population for this study. Two types of units were defined as nonrespondents: (i) vehicle owners who did not respond to the CATI and did not return the trip log and (ii) vehicle owners who responded to the CATI but did not return the trip log. Even if the last group is usually considered as respondents in the regular survey (under the condition that some key variables are reported by the respondent during the CATI), they were considered as nonrespondents for the purposes of this study since the key data used to produce the estimates are obtained via the trip log. The population of nonrespondents was formed using these two types of units.

Nonrespondents were divided in four strata using two stratification variables: vehicle type (light or heavy) and whether the unit has already been contacted by phone (yes or no, yes meaning that we actually spoke to someone for the regular survey). This last variable allowed using different sampling fraction for the 'yes' stratum and the 'no' stratum, in order to have a more efficient allocation of the sample size with the 'yes' group getting a larger part of the sample. The units that were part of the 'no' group had a

smaller sample size as we had never heard from these units despite intense follow-up procedures (see Section 3.1). A sample of 1,700 was systematically drawn among the 2,750 nonrespondents, which were previously sorted by postal code within their stratum in order to assure a good coverage across all Canada's ten provinces.

4.2 Data Collection Strategy

Data collection took place between January 23 and March 31, 2006. The data was collected by telephone interviews using a paper questionnaire. No CATI application was developed for this study due to time constraints. The interviewers assigned to this study were experienced and most of them had already worked on the CVS prior to the study. Clear instructions were given to the interviewers to make sure calls would be spread among all days of the week, at different times of the day. Before the start of data collection, new contact information research was done for the sampled vehicles in order to update the contact information used when they were in the regular CVS.

4.3 Questionnaire

Three versions of the questionnaire were developed – one for the households, one for the businesses and one for the vehicle owners who were never contacted, as no information was available to determine whether these vehicles were owned by households or businesses. All versions were short, the household questionnaire being the longest with 8 questions. Each question was directly related to one of the survey objectives. The following were the questions on the household version of the questionnaire.

1. A few months ago, we sent you a log to complete for the Canadian Vehicle Survey. The purpose of the log was to collect information on the use of the Make Model Year. Our records show that we did not receive the log back. We would like to know what was the main reason why the log was not returned.
2. What was the total distance driven with this vehicle in the last 7 days?
3. How many trips were made with this vehicle yesterday? For the purpose of this survey, a new trip occurs each time the driver or a passenger gets in or out of the vehicle. (For example, if a

man leaves his house with his wife, drops her off at her workplace and then goes to his own workplace, this counts as 2 trips, one from the house to his wife's workplace and one from his wife's workplace to his own workplace)

4. Is there a main driver for this vehicle?
5. Is the main driver of this vehicle male or female?
6. What is the age group of the main driver of this vehicle?
7. Is this vehicle the main vehicle in your household?
8. How many vehicles does your household have?

The first question was intended to address the second objective of the survey, which was to determine the reasons for nonresponse. This question was asked first as an introduction for the respondent to the survey, and was written in a way that the respondent does not feel accused of anything because he did not respond to the trip log. The goal of the second question was to estimate the bias of the key variable, vehicle-km. The results of that question would be compared with the results of the same question obtained via the CATI. The other questions were to address the third objective of the study which was to obtain characteristics of the nonrespondents. These results would also be compared to those obtained with the CATI. The business questionnaire had the same question as the household questionnaire, with Questions 3, 7 and 8 removed. The questionnaire for those who were never contacted was the same as the business questionnaire, with Question 1 removed, since they never received the trip log.

5. Analysis of the Results

The overall response rate obtained for this study was 70%, which was beyond expectations. The response rates by type of questionnaire were 85.7% for businesses, 65.1% for households and 51% for those who were never contacted before. As expected, the response rate among the never contacted group was lower than the two other types of questionnaire. It is also important to note that even though the response rate for businesses was very high, not all data from these businesses could be used for the bias estimation,

as the person answering the phone did not always know the distance driven by a specific vehicle in the last seven days. Even with that problem, the number of respondents was large enough to obtain a statistically significant conclusion.

5.1 Estimation of the Nonresponse Bias

The bias is defined as the difference between the expectation of an estimator \hat{Y} and the population total Y . In the case of the CVS, the population total of vehicle-km driven during a quarter is not available. The next best thing is to calculate an estimate using respondents and nonrespondents to stand for the population total. Let us define \hat{Y}_1 , the vehicle-km estimate obtained using respondents to both the CATI and the trip log, and \hat{Y}_2 , the vehicle-km estimate obtained using respondents to both the CATI and the trip log and 2005-Q3 nonrespondents who responded to this study's questionnaire. A t-test was made under the null hypothesis that there is no bias, and the alternative that the bias is different than 0, since we do not know the direction of the bias, if any.

H_0 : Bias=0 = E [\hat{Y}_1] - Y = 0;
Replaced in this study by $\hat{Y}_1 - \hat{Y}_2 = 0$;

H_1 : $\hat{Y}_1 - \hat{Y}_2 \neq 0$

The t-test used was:

$$t_{v,\alpha} = \frac{|\hat{Y}_1 - \hat{Y}_2|}{\sqrt{\hat{Var}(\hat{Y}_1 - \hat{Y}_2)}};$$

Where:

$$\hat{Var}(\hat{Y}_1 - \hat{Y}_2) = \hat{Var}(\hat{Y}_1) + \hat{Var}(\hat{Y}_2) - 2 \text{cov}(\hat{Y}_1, \hat{Y}_2).$$

The level of significance α used was 5%. We reject H_0 if $t > 1.96$.

The variables used to produce \hat{Y}_1 and \hat{Y}_2 were obtained via the question "What was the total distance driven with this vehicle in the last 7 days?" since this question was asked to the respondents via the CATI and to the nonrespondents through the nonrespondents study. Furthermore, this is the only question related to the vehicle-km asked to both respondents and nonrespondents.

The result obtained was $\hat{Y}_1 - \hat{Y}_2 = 79,217,437$ vehicle-km. This amount represents about 2 % of

\hat{Y}_1 . The value of t obtained is 2.10, which means that the difference is statistically significant. Thus, the published estimates of CVS are biased and overestimate the total for the population as nonrespondents tend to drive less vehicle-km than respondents.

The respondents used to compute \hat{Y}_1 and \hat{Y}_2 were the respondents from 2006-Q1. These were used rather than those of 2005-Q3 since the data collection of the study of nonrespondents took place between January and March of 2006, and previous studies have shown seasonality in the vehicle-km estimates. Respondents from 2006-Q1 were used to compare distances travelled by respondents and nonrespondents during the same period and, in the same way, avoid confusion between a bias and simply the effect of seasonality. The difference between respondents and nonrespondents is assumed to be similar in each quarter.

5.2 Reasons for Nonresponse

The second objective of the study was to determine the reasons for nonresponse. Interviewers had a list of possible answers for that question printed on the questionnaire so they could immediately code the answer provided by the respondent. This list was not read to the respondents as it could have influenced their answer.

The most frequent answers from light vehicle owners were:

- Do not remember / forgot to fill or return the log (26%)
- Log too long (22 %)
- Do not use vehicle much (10%)
- and for heavy vehicle owners:
- Do not remember / forgot to fill or return the log (25%)
- Too many drivers (16 %)
- Log too long (14 %)
- Vehicle was not being used at all (12 %)

It was expected, although not wished, to get "Do not remember / forgot to fill or return the log" as the most frequent response for both vehicle types. Not much more can be done about it, as intense and costly follow-up procedures are already in place for the CVS (see Section 3.1).

It was not a surprise to see "Log too long" as one of the main reasons for nonresponse for both vehicle types since all the CVS team, sponsors,

subject-matter and methodology teams, realize it is a survey with a high response burden for respondents and this is the main comment obtained from respondents quarter in, quarter out.

The second most popular reason for heavy vehicles, “Too many drivers” raises an issue for businesses. When one specific vehicle of their fleet is selected in the sample, it can be hard for them to keep track of which driver has to fill the log on each day, even if they have the best intentions. Some drivers forget to fill it, do not leave it in the vehicle or simply refuse to fill it. The CVS team will have to think of new ways to solve this problem.

Finally, answers “Do not use vehicle much / Vehicle was not being used at all” is not really a concern since these proportions (10% and 12%) are about the same as those observed among the respondents. However, it is important to make it clear to the respondents that even if they do not use the vehicle much or if they do not use it at all, their response is still important.

5.3 Nonrespondents Characteristics

The third objective of that survey was to find characteristics of nonrespondents which could be used for nonresponse treatment. To do so, logistic regressions were performed, in order to predict response using characteristics such as the age and the sex of the main driver, the number of vehicles owned by the household and the number of trips made on the day before the interview as independent variables.

Due to great variability in the answers obtained, only one model converged. This model included the age and the sex of the main driver as explanatory variables. Results show that women seem to respond less than men to the CVS, and that the two extremities of age groups (24 and less and 65 and over) also seem to respond less than other age groups. It is important to note that due to the very small number of observations for some of these groups, these results should be used with caution.

Exploratory analyses were also performed. The only result to notice is the fact that 77% of the respondents to the study of nonrespondents made 2 trips or less the day before the interview, while this percentage drops to 54% for respondents to the regular CVS.

6. Other Response/Nonresponse Related Studies

In conjunction with the study of nonrespondents, some tests were performed between different categories of respondents in order to find new avenues for the treatment of nonresponse. In particular, a comparison was made of the units that responded only to the CATI and those who responded to both the CATI and the log. Respondents who require a high number of attempts before being reached (also referred to as ‘high-effort’ respondents) were compared to those who require a small number of attempts (‘low-effort’). These two studies were performed using 2005-Q3 data.

6.1 Respondents to the CATI only vs. Respondents to both the CATI and the Trip Log

These types of respondents have been compared by type of vehicle. Like for the estimation of the nonresponse bias, the question “What was the total distance driven with this vehicle in the last 7 days?” from the CATI was used, since it is the only source of information available for all units. For the heavy vehicles, units who responded to both the CATI and the trip log reported on average 23.3% more vehicle-km for that question than the units who responded only to the CATI. This goes the same way for the estimation of the nonresponse bias, as those who return the log drive more than those who do not. For light vehicles, the two types of respondents reported on average about the same amount of vehicle-km. Hence, we conclude that there is no difference.

6.2 ‘High-effort’ Respondents vs. ‘Low-Effort’ Respondents

Many surveys, for example the Canadian Labour Force Survey (CLFS) (see Beaumont 2005), use variables from the data collection process to handle unit nonresponse. It was shown for that survey that units who required a large number of attempts before being reached and interviewed have more characteristics in common with the nonrespondents than with respondents who required a small number of attempts.

For the CVS, the study of nonrespondents was a good opportunity to verify whether variables from the data collection process could provide useful auxiliary information. In particular, the

number of attempts needed to reach the respondent, would be a good source of information for the nonresponse treatment especially for the key variable vehicle-km. For this study, 'high-effort' respondents were defined as those who required more than five calls for the interview to be completed. Again, the question on the total distance driven over the last seven days was used. The results show that 'high-effort' respondents report on average 17% more vehicle-km than 'low-effort' respondents. This means that 'high-effort' respondents could not be used for nonresponse treatment of the CVS as they tend to drive more than 'low-effort' respondents as opposed to nonrespondents who tend to drive less.

One limitation of that study should be mentioned. The number of calls used in the definition of 'high-effort' is a lot smaller than what is usually used. For example, in the CLFS, respondents who required more than 15 attempts were defined as 'high-effort'. Such a high number of attempts could not be used in the CVS as not enough units would have been in the 'high-effort' category to draw any conclusions.

7. Conclusion and Future Challenges

Despite intense follow-up procedures and many efforts to prevent nonresponse, low response rates lead to a nonresponse bias in the CVS estimates. This study has shown that nonrespondents tend to drive less than the respondents, which means that the published data overestimates the population total by 2 %. The study also showed that the main reasons for nonresponse were the fact that people forget to fill or return the log and the length of the trip log. It also raised the issue of businesses who own vehicles that can be driven by many drivers, which makes it difficult for them to keep track of where they are with the log. Finally, it also showed that the sex and the age of the main driver has an influence on the probability of response.

The next step will be to look for adjustments for the unit nonresponse treatment in order to eliminate, or at least reduce the nonresponse bias. It was shown that 'high-effort' respondents could not be used for the correction of total nonresponse in the CVS. The fact that very little information is available on the survey frame makes the research for useable auxiliary information difficult. Collection procedures of

the regular CVS should also be looked at in order to explain why the study of nonrespondents obtained a 70% response rate while the regular survey obtains around 60% from quarter to quarter.

References

- Landry, S. (2005). "Editing and Imputation Strategy for a Fuel Consumption Supplement to the Canadian Vehicle Survey", 2005 Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, JSM 2005, Minneapolis, MN.
- Beaulieu, M. (2005). "The Estimation Methodology of the Redesigned Canadian Vehicle Survey", 2005 Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, JSM 2005, Minneapolis, MN.
- Beaumont, J.-F. and Mitchell, C. (2002). "The System for Estimation of Variance due to Nonresponse and Imputation (SEVANI), Proceedings of Statistics Canada Symposium 2002, Statistics Canada.
- Beaumont, J.-F. (2005). "On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment", *Survey Methodology*, 31, 227-231.