COLLECTING VEHICLE USE DATA - THE CANADIAN VEHICLE SURVEY EXPERIENCE

Adam Wronski, Statistics Canada,
11-O, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada, wronada@statcan.ca

Key Words: Non-sampling errors, Transportation statistics, Vehicle usage

1. INTRODUCTION

Until recently Canadian transport activity statistics were inadequate due to the lack of any routine measurement of road vehicle activity. While road vehicles dominate passenger travel and freight traffic, no measures of total vehicle-kilometres, passenger-kilometres or tonne-kilometres were available.

The Canadian Vehicle Survey (CVS), which began data collection in the first quarter of 1999, was developed at the request of Transport Canada to fill this data gap. It is designed to collect information about the usage of road motor vehicles registered in Canada. The CVS target population includes all on-road vehicles registered in Canada. The CVS frame is created from vehicle registration files at the beginning of each quarter. A sample of vehicles is drawn each quarter. Data collection is a mix of Computer Assisted Telephone Interviews (CATI) and mail-out / mail-back questionnaires. The CATI interview is used to verify mailing information and obtain various characteristics of the vehicle. A seven-day trip log is used to gather detailed vehicle usage patterns. The log includes questions on kilometres driven, number of passengers, vehicle characteristics, individual trip purpose and times, driver and passenger demographics and fuel purchases.

The design of this survey is a unique solution. Because this is voluntary survey it poses a series of original challenges. This paper would like to focus on the distinctive characteristics of the CVS, show their impact and present research set off by the challenges. First, this paper contains a brief description of the survey design concentrating on the issues specific to the CVS. In the second part it shows some quality issues that are of concern to a survey methodologist. Next, it presents research studies undertaken to address the quality issues and their results. The paper ends with conclusions.

2. THE SURVEY DESIGN

2.1. General description

The target population of CVS includes all registered motor vehicles in Canada except trailers, motorcycles, off road vehicles (e.g., snowmobiles, dune buggies, amphibious vehicles) and special equipment (e.g., cranes, street cleaners, snow ploughs and backhoes).

The survey main goal is to produce Canada-wide annual and quarterly estimates of vehicle-kilometres and passenger-kilometres for all in scope vehicles and by vehicle type. Where possible the provincial estimates are calculated. The survey provides annual and quarterly estimates broken down by types of vehicles, age and sex of driver and time of day.

The results are the prime source of road vehicle use information for researchers, policy analysts and interested members of the public.

At the time of design it was anticipated that a high burden would affect negatively the response rates. It takes seven days to fill the log and it requires some discipline on the part of the respondent.

2.2. Survey frame and sample design

The frame is derived from the 13 jurisdictions (10 provinces and 3 territories) vehicle registration files, which are provided four times in a year for CVS.

Since the registration files are different in each jurisdiction, each file is processed and data extracted so that it is possible to merge it with the data from other files. The processing includes identification of in-scope vehicles, expired registrations and duplicates. It is especially challenging to establish to which jurisdiction the vehicle should be assigned in the presence of valid multiple records for one vehicle.

Sometimes the registration files arrive late, are created early or do not arrive at all. This obviously is a potential source of coverage errors.

The vehicles are divided into four vehicle types. Buses are identified first. The remaining vehicles are then divided, using GVW (Gross Vehicle Weight), into three weight types: below 4500kg, 4500kg or more and less than 15000kg, and 15000kg and over, called light

vehicles, trucks and "class 8" trucks respectively.

The available funds allow a total sample size of about 5000 units in provinces and 2500 units in territories per quarter. A two-stage sample design is used. Every quarter, three months before reference period, a sample of vehicles is drawn (first stage) and each of the drawn vehicles is randomly assigned a cluster of seven days to report on (second stage).

The strata are constructed using jurisdiction, the vehicle type and vehicle age. All vehicles of one type within one jurisdiction are divided into two age strata (old and new) using vehicle model year. The boundary between newer and older vehicles is found by minimizing the variance for the estimate of vehicle-kilometres at the province by type level. Every year the boundary is readjusted.

In the first stage power allocation (cube root) applied to the number of vehicles in the province determine the number of units sampled from each province. Then the number of units sampled from each vehicle type within each province is again determined using by the cube root allocation rule. The cube root rule was established arbitrarily on the client request to allow basic estimates for smaller jurisdictions. The optimal allocation between the two vehicle age strata is found empirically based on the latest available year of data. The allocation is subject to a minimum sample size in each stratum. Then the sample of vehicles is drawn from the registration list systematically by postal code. This is to ensure that large companies do not get a large number of selected vehicles in one quarter and to ensure a variety of vehicles in smaller provinces. Another measure used to limit response burden is that any selected vehicle stays in sample for one quarter and out of sample for the three following quarters.

In the second stage, each selected vehicle is randomly assigned a seven-day period to report vehicle usage. For each stratum, the first reporting day is uniformly spread (systematic assignment) over the quarter to ensure a uniform number of responses over time. In this way approximately the same number of logs starts each day of the week. The sample selection for the territories is less complex due to minimal data requirements.

2.3. Data collection

Survey collection began on February 1, 1999. Only eight vehicle registration lists were received on time to be included in the sample. Starting October 1, 1999, vehicles from all provinces and territories were included in the survey.

The provincial component of the survey data collection consists of two stages. The first stage is a telephone interview (CATI) with the registered owners of the sampled vehicles. This interview is used to collect some general information on the vehicle, e.g., to verify its current status, as well as to ask the respondent to complete a trip log. The trip log is then mailed out. If respondents cannot be contacted by phone, the trip log is mailed out with a short questionnaire to collect some of the information normally collected during the CATI interview.

A seven-day trip log, specific to each vehicle type, is used to gather detailed vehicle usage patterns. The log includes questions on kilometres driven, vehicle body and fuel type, trip purpose, length and time, number of passengers, driver and passenger demographics and fuel purchases and information about carrying dangerous goods. Vehicles over 4500kg are asked additional questions about truck configuration (e.g., type of trailer).

The territorial component of the survey collection consists of two postcards, one that is mailed to the respondents at the beginning of the quarter and the second that is mailed at the end of the quarter. The first postcard asks respondents to record the odometer reading at the beginning of the first day of the quarter and answer questions about vehicle characteristics. All those returning the first postcards are mailed a second postcard 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.

Because this is a voluntary survey and the log takes seven days to complete, every effort is made to ensure a good response rate and to prevent response errors. First, all addresses of the owners of the sampled vehicles are updated from a federal database that tracks vehicle ownership changes. Then the phone numbers are searched using computerized phonebooks and manual research. Next, for each selected unit, the CATI procedure starts about two weeks before the assigned period. During this interview a respondent approximation of kilometres driven in the week preceding the interview is also collected. If the respondent does not agree to the log an abbreviated version of the questionnaire is offered that collects odometer readings only. The second CATI contact takes place on the first day of the assigned seven-day reporting period. The respondent is asked to start filling the log and to raise any questions he may have about the log. The third CATI contact is a follow-up to remind respondents to mail back the log and to obtain data from nonrespondents by offering them an abbreviated version of the questionnaire.

To increase the response rate for vehicle, whose owners are not contacted by CATI within one week, the starting date is moved by seven days. Up to four such delays are allowed. Respondents not contacted during the four weeks or the respondents for which the phone number has not been found are mailed the log. The respondents that forgot to start the log are urged to start it a week or two weeks later.

2.4. Data processing and estimation

The processing of the data starts with the consolidation of the collected information from the three CATI interviews, abbreviated questionnaires and vehicle type specific logs. All the information is merged into a database that contains two parts: the information about the vehicle and the data about the vehicle usage (trips during the reporting period).

Next, edits are applied and inconsistent data are flagged. This is followed by imputation. During this process all erroneous data are replaced and all trip partial responses are imputed. A response is considered partial when kilometres traveled for seven days based on the odometer readings or the respondent's approximation of kilometres from the CATI interview are available.

The imputation algorithm tries to derive missing items using other information from the same respondent first, then the nearest neighbor is employed as a donor based on the kilometres driven, geographical and vehicle characteristics, main driver's sex and age, day of the week and time of day. An extensive set of post-imputation reports is created to monitor the results of the process. At the end every selected vehicle has seven days of trips, i.e., seven vehicle-days. The trips over two days are split at midnight for the estimation purpose.

Tests on vehicle-day observations (F and Waller - Duncan multiple comparison tests) have shown that the usage pattern is significantly different for days of work and holidays. Therefore, in the provincial component the vehicle-days are poststratified into non-working (holidays) and working days. Calibration to the number of vehicle-days in each poststratum is used:

$$\sum_{i \in s_h} \sum_{k \in s_i^{(W)}} a_i a_k g_k^{(W)} = N_h M^{(W)}$$

$$\sum_{i \in s_h} \sum_{k \in s_i^{(H)}} a_i a_k g_k^{(H)} = N_h M^{(H)}$$

In this equation N_h is the number of vehicles in stratum h calculated as an average of the counts from the registration lists at the beginning and at the end of the reference period. $M^{(W)}$ and $M^{(H)}$ are the numbers of days in the working-days and holidays poststrata during the reference period. The g_k's are the calibration factors. The second-stage weight for the k^{th} vehicle-day is $a_k = M_i / m_i$, where M_i is the number of days within the reference period and m_i is the number of days within the reference period with the data available. The first-stage weight a_i is based on four possible categories of response within stratum h: n_{hr} – in-scope respondents (includes imputed records), n_{hs} – nonrespondents for which it was established that they are in-scope, n_{ho} – out of scope respondents, and vehicles for which there is no information whatsoever. For the i^{th} respondent, $a_i = N_h / (n_{ho} + n_{hr} + n_{hs}) \times (n_{hr} + n_{hs}) / n_{hr}$, for nonrespondents $a_i = 0$ and for the out of scope respondents $a_i = N_h / (n_{ho} + n_{hr} + n_{hs})$. The reweighting assumes that the statistical characteristics of in-scope respondents are representative for the nonrespondents for whom it was established that they are in-scope. Vehicles with no information have the same statistical characteristics as all the vehicles for which it was established that they are in- or out-of-scope.

This calibration ensures that every vehicle-day reported within a stratum contributes equally to the estimates. The estimates of counts for vehicle characteristics and all estimates for territories are based on a simple one-stage design.

The CVS is also attempting to incorporate imputation rates into its calculations of precision of the estimates. Simulation is used each quarter to take into account the impact of imputation. Using the data from respondents as the population, an initial sample is selected, and then sub-samples are selected to represent respondents and non-respondents within the initial sample. Then the nonrespondents are imputed using the respondents and the CVS imputation strategy. The estimated c.v.'s for the initial sample and for the initial sample with imputed nonrespondents are computed. This procedure allows us to visualize the change of the c.v. depending on the amount of imputation. A linear regression model is then used to quantify the dependence of the c.v. on the imputation rate and the c.v. when imputing for nonresponse.

3. THE CHALLENGES

The quarterly results were available to the public starting from the fourth quarter of 1999. Most of the estimates were very close to the educated guesses of experts. This may cause overconfidence in the survey estimates that may be biased due to non-sampling errors. There is no administrative or statistical data to compare the CVS results with. Thus there are no

indications about the biases. The non-sampling errors can strongly affect the estimates and today they constitute the major challenge for the CVS methodologists.

The primary source of all the difficulties is the low response rate. For the year 2000 only about 38% of vehicle-days reported contain all the necessary items without or with only minor defects (full response). Since the current collection procedure is very elaborate and expensive (over 50% of the survey total cost) and there cannot be more follow-ups or respondent incentives, no recipe has been found to improve the response rate.

The difficulty is even larger due to the fact that for almost half of the days (during the reference period) of the full response the vehicles were not used. Another 30% of all vehicle days reported requires extensive imputation (partial response), because the reported data are inconsistent or only the respondent approximation of kilometres driven in the week preceding the CATI interview is available. For most estimates involving the individual trip characteristics the imputation rate (the imputed part of the estimate) is from 40% to 60%. The vehicle counts and estimates of kilometres traveled at the stratum level have imputation rates from 0% to 6%. Thus it is critical that the imputation system does not create bias.

A potential source of imputation bias is the respondent approximation of kilometres driven in the week preceding the CATI interview. This information is used extensively during the imputation process as a proxy for the kilometres driven during the week. So far there have been no comparative studies of respondent approximations (recall data) and odometer readings to assess its quality. Since a specific week (not the average week) is collected it is assumed only that the respondent approximation follows the same distribution as the odometer readings. The respondent approximations are corrected by the ratio of the respondent approximation mean to the mean of the odometer read distance for the vehicles with both measures on hand. When calculating these means the vehicle type, province, driver demographics are used (if enough such respondents) to base the correction on similar vehicles. But what is the impact of this procedure on the bias?

4. THE RESEARCH AND ITS RESULTS

Several additional studies were done to answer these challenges.

4.1. Collection methodology study

The study goal was to look for cheaper collection alternatives to the current collection scenario. For the second quarter of 2000 instead of the regular sample of approximately 5,000 vehicles, a sample of 8,000 was drawn. This sample was split randomly into two separate samples according to the following rules: each of the two samples should have approximately the same size within each stratum and all the units having the same 6 digit postal code were grouped into one of the samples. These rules were introduced to avoid repeated contacts of respondents having multiple vehicles selected.

Then for one of the samples the data were collected using the current CVS collection methodology. The second sample did not undergo the CATI part of the regular collection procedure. The questionnaires were sent directly by mail to the respondents without any pre-contact and there was no CATI follow-up.

For the purpose of the study, the information collected from both samples was edited and imputed separately. Next, for each scenario, the estimation was done and the results were compared.

The response rate from the mail only collection is less than half of the response rate from the regular collection (about 20% of the selected sample gave all requested information and another 10% provided partial responses). Such a small response rate would likely generate biased estimates.

The analysis of responses found that if CATI is not used the respondents' profile changes in the following way:

- There are fewer responses indicating that the vehicle is not in use. From the current method sample, 24.3% of the dates have a value of zero for daily kilometres, while from the mail-out, only 20.9% of the dates have zero daily kilometres.
- The current method is more likely to collect responses from owners of vehicles that are used a lot.
- The absences of CATI yields much lower estimates of trip kilometres among young drivers (under 25).

The use of CATI allows compensation (at least partially) for the under-representation of heavily used vehicles, the young respondents and vehicles not in use. For mail-only collection there is less information to use in the imputation.

Moreover, any increase of the sample size would generate higher response burden in small provinces among truck and bus owners. There was already some

indication that some respondents are unhappy about the number of the questionnaires received from CVS. Thus enlarging the sample would not yield a proportionally larger number of responses.

Thus the conclusion of the study was that the mail-out only collection could not be used without negatively affecting quality.

4.2. Analysis of imputation effects

Some comparison of imputed and non-imputed data was done. It was discovered that imputed trips tend to be longer than the reported ones and vehicles for which there is have only respondent approximation travel more. The differences are shown on Figure 1 and Table 1.

Figure 1. Trip length distribution by collection method.

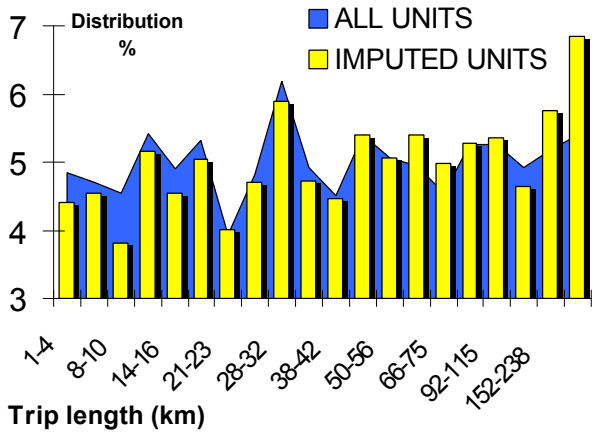


Table 1. Average distance traveled by the type of imputation.

Imputation type	In a day	In one trip
None (full response)	59 km	30 km
Partial (partial response)	63 km	25 km
Total (based on respondent approximation)	82 km	42 km

Table 2 below lists the domains of interest for which the imputation procedure has the biggest effect on the average trip distance traveled.

Table 2. The domains with large differences of average daily distance traveled for reported and imputed data.

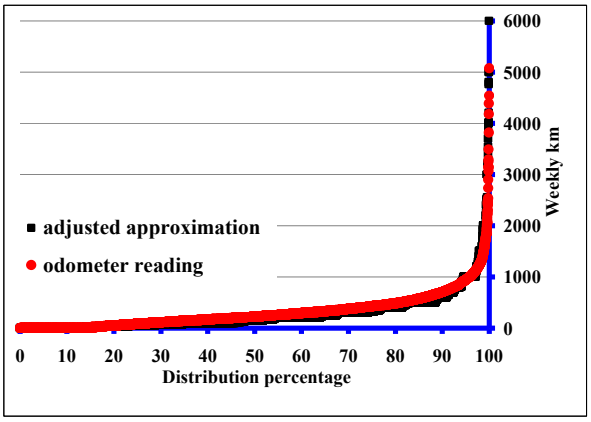
Domain of interest	Average distance	
	Reported	Imputed
Male driver	25 km	33 km
Driver younger than 20 years	21 km	29 km
Driver older than 85 years	35 km	14 km
Trips between 12am and 6am	33 km	95 km

Thus the study confirmed that imputed data could be significantly different than the reported one.

4.3. The respondent approximation study.

Since the respondent approximation of the distance driven last week is extensively used in the imputation process it is necessary to verify the hypothesis that the respondent approximations distribution can substitute for the odometer readings distribution. The distributions were compared for respondents that provided both, by the vehicle type (the data from 2000 was used). The correlation between them for each type was higher then 0.989. Figure 2 presents the comparison of distributions for vehicles with GVW less than 4500kg in Ontario. Given that the average approximation and odometer reading can differ up to 20%, the ratio of averages is applied to adjust the approximation before its use in the imputation process (see 3.). For other vehicle types and provinces tested so far the distribution behaviors exhibited are similar. During imputation the adjustment is calculated and applied at the level of imputation class and thus a further study at that level is necessary.

Figure 2. Comparison of the odometer readings and respondent approximation distributions for light vehicles in Ontario.



4.4. "Air Care" study.

The "Air Care" program carried out in a region of British Columbia provided us with their database of annual odometer readings. Annually, the program tests "light use" vehicles that are on the road for over two years. This enabled us to compare an estimate based on "Air Care" data with the 1999 CVS estimate for the same area (based on 388 responses). The "Air Care" program area encompasses about 6% of the CVS population, i.e., about 1.04 million vehicles.

It was established that the "Air Care" database contains a high rate of odometer reading errors. There were many negative differences and extremely large positive differences between two consecutive annual readings.

Figures 3 and 4 show the distribution of the "Air Care" odometer reading differences and its detail displaying unusual number of observations in the vicinity of 100 thousand-kilometres per year. For illustration, on figure 4, the number of vehicles associated with the observations that are black can be considered inaccurate due to, for example, an addition or omission of a digit during the odometer reading could be a potential source of such error.

Thus, it had to be decided which "Air Care" observations was an error. Based on the expert advice the estimates were calculated for cut-offs from 80,000km to 100,000km. The estimates depending on the availability of 1998, 1999 and 2000 odometer readings were also calculated.

Figure 3. The distribution of the "Air Care" odometer reading differences adjusted to represent 365 days.

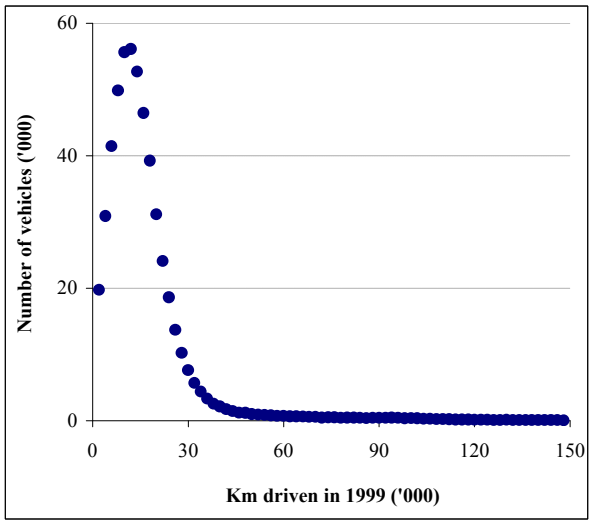
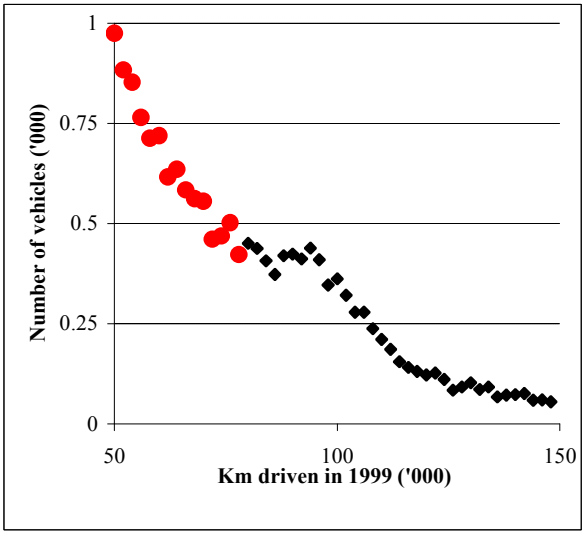


Figure 4. The detail of the "Air Care distribution.



Depending on the choice of the "Air Care" observations considered in error the estimates of the distance driven range from 15.9 to 17.7 billions kilometres. The number of observations they are based on varies from 531 thousand to 895 thousand vehicles. The c.v.'s are less than 0.3%. The CVS estimate is 15.9 billions kilometres, with a c.v. of 6.3%. Thus the "Air Care" estimates differ from 0% to 12% from the CVS estimates.

5. CONCLUSIONS

The CVS is a new survey with a unique methodology. The evaluation studies did not expose any major problems. But it requires more research to improve the response and imputation rates within survey human and financial resources constraints. Also, to be able to carry out more comprehensive comparisons new administrative data initiatives (mostly from provinces) are necessary.

There are still issues that need monitoring and further studies:

- impact of the nonresponse on the quality of estimates, e.g., correlation of heavy-driving with the nonresponse;
- impact of the imputation on the quality of estimates and
- the use of respondent approximations during the imputation process.

And finally, even if the survey estimates look very convincing, it is also necessary to diligently warn users that high nonresponse and imputation rate may affect their quality.

REFERENCES

Briant, N. (2001), Estimation de la distance parcourue par les véhicules du Programme administratif "Air Care", technical report, Statistics Canada.

Christoff, W, Wronski A, (2000-2001), Canadian Vehicle Survey, Statistics Canada publication, http://www.statcan.ca/english/freepub/53F0004XIE/free.htm).

Matthews, S. (2000), Canadian Vehicle Survey - Summary of Response and Imputation Distribution Study Findings, technical report, Statistics Canada.

Xiao, P. (2000), Canadian Vehicle Survey - Summary Q2 2000 Study Estimates and Response Rates comparison, technical report Statistics Canada.

Wronski A. (2000), Canadian Vehicle Survey (CVS) – Quarter 2, 2000 Collection Study, technical report Statistics Canada.