The Estimation Methodology of the Redesigned Canadian Vehicle Survey

Martin Beaulieu, Business Survey Methods Division, Statistics Canada,
R.H. Coats Building 17-P, Ottawa, ON, Canada, K1A 0T6
(martinj.beaulieu@statcan.ca)

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.

The CVS underwent a major redesign in 2004 to better meet the clients' needs. The main objective of the redesign was to add a fuel supplement to the trip log at the request of both TC and NRCan to better monitor fuel consumption. In order to keep the response burden as low as possible, changes were made to some survey concepts, including the reporting period.

After a short overview of the CVS in section 2, this paper will describe the main impacts of the redesign in section 3, followed by the description of the new estimation strategy, including changes made to the estimation of variance due to nonresponse and imputation, in section 4. Some post-redesign observations are discussed in section 5 and, in section 6, some future projects are presented.

2. Survey Overview

The CVS is a voluntary vehicle-based survey that provides quarterly and annual estimates of road vehicle activity (vehicle-kilometres and passenger-kilometres). The target population is registered on-road vehicles 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 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 sample for quarter 1. 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 results in 78 strata.

The CVS follows 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 vehicles are selected systematically by postal code. 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 quarter of reference 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. If respondents cannot be contacted by phone, 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).

3. CVS Redesign

In the context of the ratification of the Kyoto protocol on the reduction of greenhouse gas emission, the two survey sponsors were interested in getting more detailed information on fuel consumption by on-road vehicles registered in Canada. At the same time, TC was interested in getting more details on every trip, such as the number of passengers and their age at every moment, to help them deal with road safety issues. This led to a major redesign of the survey in 2004.

3.1 Trip Definition

One of the most important changes made in the redesign was the modification of the trip definition for light vehicles. Prior to 2004, the respondent had to report a new trip if (i) the purpose of the trip changed; (ii) after a 30 minute or more stop; or (iii) after a change in driver. After the redesign, the respondent is asked to report a new trip every time someone gets in or out of the vehicle. This new definition allows TC to have the exact itinerary of the driver for each trip, which meets the objective of getting more trip details in order to deal with road safety issues.

For example, let's assume that the driver of the selected vehicle leaves home in the morning to go to work and has to drop off his/her children, one to daycare and one to school. The first trip is from home to daycare. Leaving daycare, since there is one passenger less in the car, the driver has to report a new trip. The driver has to report a third trip leaving school to go to work for the same reason. Prior to the redesign, since there was no change in the trip purpose (go to work), no stop that lasted more than 30 minutes and no change in driver, this was only one trip. With the new trip definition, the driver would report three trips.

3.2 Fuel Consumption

Another main objective of the redesign was to better monitor fuel consumption. The addition to the trip log, for all vehicle types, of a supplement on fuel consumption achieved this goal. In this supplement, the respondent reports the odometer reading at the time of the purchase, the type of fuel purchased, the quantity purchased, the price per unit (litre or U.S. gallon) and the total amount spent for each of the two fuel tank fill-ups required (or each of the five partial fill-ups if the purchases are not fill-ups). This information combined with the trip log allows the estimation of a fuel consumption ratio for each vehicle and allows to derive a quantity of fuel consumed for each trip.

3.3 Reporting Period

One potential problem raised by the addition of the fuel supplement was the significant increase in an already high response burden. The new trip definition also increased the number of trips to report. In order to compensate for the higher response burden, it was decided to change the reporting period for light vehicles from seven days worth of trips to 20 trips, regardless of how many days it takes to report 20 trips. This new reporting period would help reduce the response burden, especially for respondents who make many trips each day. This new approach also has the advantage of making sure the response burden would be the same for every unit in the sample. With the old approach, where the respondents had to report every trip made during the seven-day reporting period, the response burden was higher for the respondents who were using their vehicle a lot (some had to report up to 40 trips). Keeping response burden at the same level for everyone is also a protection against nonresponse bias.

When the 20^{th} trip occurs, the respondent can stop reporting trips. Since the second-stage weight is calculated in terms of days (see section 4), the end of the last day is imputed using other days reported for the same vehicle, in order to have a set of complete days.

For the fuel supplement (or fuel log), the respondent is asked to report two fuel tank fill-ups or five partial fill-ups. If less than two fill-ups are reported when the 20^{th} trip occurs, the respondents has to keep carrying the log in the car until a second fill-up is reported.

There are two different reporting periods for each vehicle, one for the trip log and one for the fuel log. For example, if the starting date randomly assigned for a vehicle is March 1^{st}, then the trips and fuel purchases could have the distribution showed in table 1.

Table 1 – Example of trips and fuel purchases distribution

Date	*March 1st*	*March 2nd*	*March 3rd*	*March 4th*	*March 5th*	*March 6th*	*Total*
Number of trips	4	5	3	6	2	-	20
Fuel purchases		1				1	2

In this example, the trip log would have a reporting period of five days, since the 20th trip occurs on March 5th. On the date of March 5th, only one fuel fill-up has been recorded since the start of the reporting period, which means the respondent kept carrying the log until a second fill-up. In that example the second fill-up occurred on March 6th, which means a reporting period of six days for the fuel log. As it will be explained in section 4, an optimal use of these two reporting periods is made to produce more precise estimates.

3.4 Sample Sizes

The new reporting period raised another concern: a potential reduction of the sample size in terms of vehicle-days. Some studies on data from previous years showed that the 20th trip occurred on average on the fifth day of the reporting period. If this applies with the new reporting period and the new trip definition, a reduction of two days for each vehicle would be observed in average, compared to the previous seven-day reporting period, which would mean a decrease of ~ 10,000 vehicle-days in the sample for each quarter. In order to compensate, in was decided to increase the vehicle sample size for all provinces from 5,000 vehicles per quarter to 5,375 vehicles per quarter.

4. Estimation Strategy

The changes made to the survey with the redesign presented some challenges that affected the estimation strategy. Since both logs provide odometer readings, the data collected from both the trip log and the fuel log can be used to produce vehicle-kilometres estimates. In fact, the trip log and the fuel log represent two sources of data with overlapping but different reporting periods to estimate vehicle-kilometres totals. This means each vehicle has two different second-stage weights. The main challenge was to use the two sources of data and their different second-stage weights in an optimal way to produce one single set of estimates. Furthermore, users from TC and NRCan must be able to reproduce the published estimates using micro data files provided by Statistics Canada.

4.1 Weights Calculation

The first-stage weight for a vehicle i in stratum h is defined as:

$$w_{1i} = \frac{N_{hi}}{n_{hi}}$$

where N_{hi} is the population size of the stratum h and n_{hi} is the sample size of stratum h.

The second stage weights for the vehicle i on the day j are defined as:

$$w_{2ij,trip} = \frac{M}{m_{i,trip}} \quad ;$$

$$w_{2ij,\,fuel} = \frac{M}{m_{i,\,fuel}}$$

where M is the total number of days in the quarter (90, 91 or 92 days depending on the quarter), $m_{i,\,trip}$ is the number of days reported by vehicle i on the trip log and $m_{i,\,fuel}$ is the number of days reported by vehicle i on the fuel log. Both $w_{2ij,trip}$ and $w_{2ij,\,fuel}$ are adjusted later on in the process as the vehicle-days are post-stratified into non-working days (weekends, holidays) and working days.

4.2 Estimation of Vehicle-kilometres

The goal of the strategy is to produce estimates of vehicle-kilometres that are as accurate as possible. In order to reduce the variance, the log (trip or fuel) with the longest reporting period should be used for the estimation of vehicle-kilometres. In the CVS case, it is not always the same log that has the longest reporting period. In the example described in section 3.3, the fuel log had the longest reporting period (six days vs. five days for the trip log). For another vehicle it could be the other way around if the second fuel fill-up occurs before the 20th trip.

It was decided to use an estimator which takes into consideration the longest reporting period, by using a contribution factor α_i applied to an estimate obtained with trip log data and a factor $1-\alpha_i$ applied to an estimate obtained using data from the fuel log. The factor α_i is equal to 1 if the trip log has the longest reporting period. Otherwise, if the fuel log has the longest period, α_i is equal to 0. This leads to the following estimator of Y, the quarterly total of vehicle-kilometres in the population:

$$\hat{Y} = \sum_{i \in s_1} w_{1i} \left[\alpha_i \hat{Y}_{i,trip} + (1-\alpha_i) \hat{Y}_{i,\,fuel} \right]$$

where $\alpha_i = 1$ if $m_{i\,trip} \geq m_{i\,fuel}$

0 if $m_{i\,trip} < m_{i\,fuel}$

$$\hat{Y}_{i,trip} = \sum_{j \in s_2} w_{2\,ij,trip}\, y_{ij,trip}$$

$$\hat{Y}_{i,\,fuel} = \sum_{j \in s_2} w_{2\,i,\,fuel}\, y_{ij,\,fuel}$$

$y_{ij,trip}$ is the total distance driven by vehicle i on the day j reported on the trip log and $y_{ij,\,fuel}$ is the total distance driven by vehicle i on the day j reported on the fuel log.

Even though respondents do not provide odometer readings every day of the fuel log reporting period, vehicle-kilometres for each day are derived using the total distance driven reported on the fuel log in order to apply post-stratification.

4.3 Estimation of Passenger-kilometres

Passenger-kilometres is another key variable of the CVS. It is calculated for each trip and is defined as the number of passengers (including the driver) multiplied by the distance driven. Passenger-kilometres estimates cannot be calculated in the same way as vehicle-kilometres estimates since passengers' information is only available on the trip log. On the other hand, the distance traveled reported on fuel log has to be taken into consideration to maintain the consistency between the vehicle-kilometres and passenger-kilometres estimates. In the case that the fuel log is used to calculate the vehicle-kilometres estimate for a vehicle, it could happen that the passenger-kilometres estimate would be smaller than the vehicle-kilometres estimate. For example, if the driver is alone for each trip reported for vehicle i on the trip log, then the passenger-kilometres estimate ($\hat{Z}_i$) for that vehicle would be equal to $\hat{Y}_{i,trip}$, the vehicle-kilometres estimate using the trip log for vehicle i. Let's assume that for this respondent, $\hat{Y}_{i,\,fuel} > \hat{Y}_{i,trip}$ and the fuel log reporting period is longer than the trip log reporting period, which means that $\hat{Y}_{i,\,fuel}$ is used to produce $\hat{Y}_i$. In that case, $\hat{Z}_i$ would be smaller than $\hat{Y}_i$. Since the driver is considered as a passenger, it is impossible to have a passenger-kilometres estimate smaller than the vehicle-kilometres estimate.

To avoid this kind of inconsistency, an adjustment factor Ω_i, calculated at the vehicle level, is applied to the passenger-kilometres estimate using trip log data. If the fuel log is used to produce the vehicle-kilometres estimate, this adjustment factor is the ratio of the vehicle-kilometres estimate using fuel log data over the vehicle-kilometres estimate using trip log data. Otherwise the adjustment factor is one, since both the vehicle-kilometres estimate and the passenger-kilometres estimate use the same source of data.

$\Omega_i = 1$ if $\alpha_i = 1$;

$$\Omega_i = \frac{\hat{Y}_{i,\,fuel}}{\hat{Y}_{i,trip}} \text{ if } \alpha_i = 0.$$

In the example shown in section 3.3, if we assume that a total of 100 vehicle-kilometres were reported for five days on the trip log and 180 vehicle-kilometres were reported for six days on the fuel log, then the quarterly vehicle-kilometres estimates for that vehicle would be:

$$\hat{Y}_{i,trip} = \sum_{j \in s_2} w_{2\,ij,trip}\, y_{ij,trip} = \frac{91}{5} * 100 = 1820$$

vehicle-kilometres

$$\hat{Y}_{i,\,fuel} = \sum_{j \in s_2} w_{2\,ij,\,fuel}\, y_{ij,\,fuel} = \frac{91}{6} * 180 = 2730$$

vehicle-kilometres

Since $\hat{Y}_{i,\,fuel} > \hat{Y}_{i,trip}$, $\alpha_i = 0$, which means

$$\Omega_i = \frac{\hat{Y}_{i,\,fuel}}{\hat{Y}_{i,trip}} = \frac{2730}{1820} = 1.5.$$

The adjustment factor Ω_i is included in the following estimator for passenger-kilometres, $\hat{Z}$,

$$\hat{Z} = \sum_{i \in s_1} W_{1i} \Omega_i \sum_{j \in s_2} W_{2\,ij,trip} z_{ij,trip}$$

where $z_{ij,trip}$ is the total of passenger-kilometres for vehicle *i* on day *j*.

4.4 Sampling Variance

The estimation of sampling variance for vehicle-kilometres uses the common two-stage design form of calculation:

$$V(\hat{Y}) = V_1[E_2(\hat{Y})] + E_1[V_2(\hat{Y})]$$

The estimator for the variance of passenger-kilometres, $\hat{Z}$, has to consider the adjustment factor Ω_i in its calculation. Since the adjustment factor is calculated at the vehicle level, it can be considered as a constant for the second-stage component (V_2) of the variance and squared in the variance calculation. This leads to a variance estimator of the form:

$$V(\hat{Z}) = V_1[E_2(\hat{Z})] + E_1[\Omega_i^2 V_2(\hat{Z})].$$

4.5 Variance due to nonresponse and imputation

With response rates between 50% and 60% depending on the quarter, it is essential to add a variance component due to nonresponse and imputation to the total variance. In fact, studies have shown that a proportion of 40% of the total variance is due to nonresponse and imputation for the CVS. Prior to the redesign, the variance due to imputation was estimated using a regression model which had for independent variables the response rate and the coefficient of variation calculated with the sampling variance. Since methods have evolved substantially in the domain of variance due to imputation estimation in the recent years, another challenge of the new estimation strategy was to improve the estimation of variance due to nonresponse.

Starting in 2004, the SEVANI system (Beaumont and Mitchell, 2002) will be used to estimate the variance due to nonresponse and imputation. The main advantages of the method used in the SEVANI system are that it uses exact calculation, as opposed to replicate methods like Bootstrap, and it includes a component for reweighting, which is the method used for the treatment of total nonresponse in the CVS. The only constraint resulting from the use of SEVANI is the fact that this system does not support the many imputation methods used by the CVS all at once. The solution was to assume the use of only one method of imputation, donor imputation, the method most commonly used in the CVS.

The quality indicators that appear in the CVS publication (Statistics Canada, 2005) consider both the sampling variance and the variance due to nonresponse and imputation.

5. Post-Redesign Observations

After the production of four quarters of the CVS since the changes due to the redesign were put into effect, some of the main impacts of the redesign can be highlighted.

A small drop in the response rates can be observed. This is probably due to the increase in the response burden with the addition of the fuel supplement and the new trip definition.

On the other hand, the sample size has increased in terms of vehicle-days. Three reasons can explain this unanticipated increase. First, the increase of the vehicle sample size is an obvious reason. Then, the use of the longest reporting period between the trip log and the fuel log helped gain supplementary vehicle-days when the fuel log had the longest reporting period. Finally, the new reporting period of 20 trips instead of seven days for light vehicles increased rather than decreased the number of vehicle-days since, on average, it takes between eight and nine days to report 20 trips. Despite this observation, the goal of reducing the response burden is met for respondents who make many trips each day. For these respondents, it takes on average three or four days to report 20 trips.

6. Future work

With the redesign well in place, some other projects were started in order to keep improving the survey. Two projects will address the issue of nonresponse which is still a concern for the CVS.

6.1 Study on the use of incentives

First a study on the use of incentives is currently running. For quarters 2 and 3 of 2005 the vehicle sample has been split in three groups in order to evaluate whether or not the use of incentives would increase the response rates. Group 1 receives a mechanical pencil attached to the log; group 2 receives the mechanical pencil and a key chain which serves as a reminder and finally the third group

receives only the log without any incentives. The choice of these incentives was made after the analysis of focus groups reports. These focus groups were held in the early 90's after the National Private Vehicle Use Survey (NaPVUS) was conducted. When the participants were asked why they did not respond to the NaPVUS or what could help them fill the log, the most common answer was that they often kept the log in the car but did not have anything with which they could write. This caused errors in the data reported as the respondents did not exactly remember the information asked for every trip. Moreover, since it had missing information, some of them gave up and did not even send the log back.

The exact Fisher test will be used to detect if there are significant differences between any of the groups. With the sample sizes available for these quarters and the expected responses rates, a 2.2% difference in the response rates should be detected as significant. The results of this study should be available early in 2006.

6.2 Nonrespondent study

Another issue the CVS has to deal with is nonresponse bias. It is planned to conduct a study among the nonrespondents in early 2006. The goals of the study would be to (i) evaluate the direction and the extent of the nonresponse bias; (ii) improve data collection methods and (iii) improve the edit and imputation methods. In order to evaluate whether or not their characteristics are different from those of the respondents, nonrespondents will be contacted by phone and asked a very small number of questions such as why they did not respond, how many kilometres have they driven during a certain amount of time etc.

7. Conclusion

After a major redesign, the CVS better meets the clients' needs with the addition of the fuel supplement. The new trip definition also helped achieving this goal, since it allows the clients to have the information they need on the itinerary of each trip in order to deal with some road safety issues. The CVS now uses in an optimal way two sources of data available the estimate the key variable of the survey, vehicle-kilometres. The estimation of variance due to nonresponse and imputation has also been improved with the use of the system SEVANI. It is now important to keep working towards better response rates or at least to have a thorough understanding of the nonresponse bias in order to continue improving the quality of the CVS.

8. References

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.

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.

Statistics Canada (2005). "The Canadian Vehicle Survey – First quarter 2004", catalogue no 53F0004XIE.