EDITING AND IMPUTATION STRATEGY FOR A FUEL CONSUMPTION SUPPLEMENT TO THE CANADIAN VEHICLE SURVEY

Sébastien Landry, Statistics Canada
Business Surveys Methods Division, R.H. Coats Building, 17th floor, Ottawa, ON, Canada, K1A 0T6
sebastien.landry@statcan.ca

ABSTRACT

The Kyoto Protocol, whose objective is to reduce greenhouse gas emissions, has generated increased interest in measuring fuel consumption. It is in this context that a fuel supplement was added to the Canadian Vehicle Survey (CVS), which is conducted by Statistics Canada. The CVS measures travel activity in terms of vehicle-kilometres and passenger-kilometres on roads by most of the registered Canadian motor vehicles (excluding buses, motorcycles and other small vehicles). The objective of the fuel supplement to CVS is to inform federal agencies such as Transport Canada and Natural Resources Canada, both co-sponsors of the survey, as well as the general populace, of the quantity of fuel consumed and the fuel efficiency of motor vehicles registered in Canada. This paper will focus on the edit and imputation strategy that was developed to impute the variables introduced by the fuel supplement. A major element of the strategy involves the creation of a fuel consumption model. This regression model will be used to impute several variables of the fuel supplement (such as odometer reading and quantity of fuel purchased) and to calculate the amount of fuel consumed each day of the reporting period (using the distance driven during the day).

Key words: travel surveys, imputation, regression model, fuel consumption

1. INTRODUCTION

The Canadian Vehicle Survey (CVS), which was conducted for the first time in 1999, went through a redesign for reference year 2004. The objectives of the redesign were to provide more detailed information about trips made by vehicles and to calculate estimates of fuel consumption. A fuel supplement was added in the questionnaire as an answer to the last objective. This fuel supplement, which collects information on every fuel purchase made for the vehicle during the reporting period, will also be used to improve the estimates of distance travelled by vehicles. This paper will focus on the edit and imputation process for the fuel supplement.

An overview of the CVS and its questionnaire will be provided in section 2. The edit and imputation process for the fuel supplement will be described in sections 3 and 4. Section 5 will show how fuel consumption is calculated, which is the main objective of the fuel supplement. Some imputation results and a conclusion will be presented in sections 6 and 7.

2. OVERVIEW OF CVS

The Canadian Vehicle Survey is a quarterly survey which is sponsored by Transport Canada and, since 2004, Natural Resources Canada. The first main objective of the CVS is to measure road vehicle activity across Canada. The variables of interest for this objective are vehicle-kilometres (distance travelled by vehicles on roads) and passenger-kilometres (total distance travelled by all passengers, including drivers, in road vehicles). For example, a vehicle with three people in it which travels 10 kilometres will contribute to 10 vehicle-kilometres and 30 passenger-kilometres. The second main objective of the CVS, which was added for reference year 2004, is to measure and monitor fuel consumption for road vehicles in Canada.

The target population for the CVS contains all motor vehicles, except buses (buses were included in the survey prior to 2004), motorcycles, off road vehicles (e.g., snowmobiles, dune buggies, amphibious vehicles) and special equipment (e.g., cranes, street cleaners, snowploughs and backhoes), registered in Canada anytime during the survey reference period, that have not been scrapped or salvaged. The survey frame consists of the motor vehicle registration files provided by all 10 provincial and all 3 territorial governments.

The CVS uses a stratified two-stage sampling design. At the first stage, a sample of vehicles is chosen. All vehicles from the survey frame are stratified according to their province/territory, their vehicle type (Light vehicles: less than 4.5 tonnes; Heavy vehicles: between 4.5 and 15 tonnes; or Class 8 trucks: 15 tonnes and more) and their age (all combinations of province/vehicle type are separated in two age groups: newer vehicles and older vehicles). The sum of the sample sizes of all provinces should equal 21,500 vehicles per year (5,375 per quarter). For the territories, the total sample size should be 10,800 vehicles per year (2,700 per quarter). At the second stage, a starting date within the quarter is randomly selected and the respondent has to fill in information about every trip and fuel purchase made until 20 trips and two fill-ups (or 5 fuel purchases) have been made, with a maximal reporting period of 28 days.

Data collection goes as follows: a Computer Assisted Telephone Interview (CATI) is conducted with the vehicle's owner. The owner is asked to provide general information about the vehicle (vehicle type, fuel type used, odometer reading, vehicle maintenance questions, etc.). If the respondent agrees, a log is then sent so that information about the vehicle's trips and fuel purchases can be provided. The log is divided into two parts: a trip log, where the respondent provides general information about the vehicle (body type, fuel type used, etc.) and reports the first 20 trips made by the vehicle during the reporting period, and the fuel log (fuel supplement), where the respondent declares two fill-ups or the first five fuel purchases made for the vehicle during the reporting period. Note that a new trip is reported for light vehicles every time the driver gets in the vehicle or a passenger gets in or out of the vehicle. A new trip is reported for heavy vehicles[1] (greater than 4.5 tonnes) if any of the following happens:

- a stop of more than 30 minutes
- a change of driver
- a change of purpose or use
- a change in the truck configuration
- a change in the status of the load from loaded to unloaded or the reverse

[1] From now on, Class 8 trucks will be included in Heavy vehicles.

There are two different types of trip logs: one for light vehicles and one for heavy vehicles. In the trip log, the respondent has to provide information for every reported trip, as shown in Table 1 (an "X" indicates that the selected variable is required).

Table 1: Variables to be reported in the trip log

Variable	Light vehicles	Heavy vehicles
Start and stop dates and times	X	X
Start and stop odometer readings	X	X
Origin and destination	X	
Trip purpose		X
Number and age group of passengers	X	
Number of passengers at the start and the end of the trip		X
Sex and age group of the driver	X	X
Fuel purchase made during the trip	X	X
Distance travelled on roads with posted speed limit of 80km/h or more	X	X
Truck configuration		X
Dangerous goods		X

In the fuel log, the respondent has to declare the following information for every reported fuel purchase:

- date of the purchase
- odometer reading at the time of the purchase
- fuel type purchased
- quantity of fuel purchased
- amount spent
- price per unit of the fuel
- fill-up indicator to tell if the fuel tank was full after the purchase

To increase the number of responses, a follow-up process has been established. Respondents are contacted by phone on the first day of the reporting period to remind them to fill out and return the questionnaire and to let them know they can call Statistics Canada anytime at a toll-free number to ask questions. A reminder is sent by mail one week after the beginning of the reporting period. If the questionnaire has not been sent back after nine weeks, a short

questionnaire is sent, and then a phone follow-up is performed one week later.

For more information about the survey, please consult *Canadian Vehicle Survey – First Quarter 2004* (Statistics Canada, 2005) or Beaulieu (2005).

Reweighting is used to handle total non-response. As for partial non-response, imputation is used to correct inconsistencies or missing values. The next sections focus on the imputation process.

3. EDIT PROCESS FOR THE FUEL LOG

When the questionnaires are returned, it is likely that they will not have been filled out perfectly, so an edit process is needed to catch errors and determine which variables need to be imputed.

The process first verifies, for every variable, if there are any missing or invalid values. Consistency edits are then performed to ensure the data collected in the fuel log is consistent with the data from the trip log. An example of such an edit is the comparison of the fuel type purchased in the fuel log with the vehicle's fuel type reported at the beginning of the trip log[2]. If they are different, the value from the fuel log will need to be imputed.

Other edits are also performed. Some variables might have out of range values; these are sent to imputation. For example, the quantity purchased should never be greater than the vehicle's fuel tank capacity or a certain quantity (180 litres for light vehicles, 600 litres for heavy vehicles). In the same way, the amount spent should never be greater than $180 for light vehicles and $500 for heavy vehicles. Also, the price per unit (in litres or US gallons) should stand between 30¢ and 300¢ (for 2004).

Finally, another consistency edit is performed. The process verifies if the fuel purchase equation was respected for the purchase. The fuel purchase equation is:

Amount spent = Quantity purchased * Price per unit

If the fuel equation is not satisfied for the purchase, this means that at least one out of the three variables will need imputation.

Now that the data from the fuel log has been edited, the imputation process can begin.

4. IMPUTATION PROCESS FOR THE FUEL LOG

The imputation process is used to handle partial non-response, i.e. reported fuel purchases with missing or invalid variables. This section is separated into subsections where each imputation method used in the fuel log imputation process is presented.

- **Deterministic imputation using the fuel purchase equation**

The fuel purchase equation shown previously can be used to impute the amount spent, the quantity purchased or the price per unit. If two of the three variables do not require imputation, the equation uses them to derive the value of the third variable.

- **Deterministic imputation using the fuel tank capacity**

The maximal quantity of fuel that the fuel tank can contain is useful in the process to impute the fill-up indicator, the quantity type (to determine if the quantity purchased was in litres or in US gallons) and the quantity purchased (the fuel tank capacity, in this case, is used as an upper bound over which the imputed quantity purchase cannot go).

Unfortunately, a database containing the fuel tank capacities for all vehicle makes and models could not be found in time and Vehicle Identification Number decoding could not provide this information either. In order to obtain those fuel tank capacities, Carguide magazines were referenced. Each year, the magazine produces a list detailing vehicle characteristics (including fuel tank capacity) for most light vehicles on the market. Using this source, the fuel tank capacities were manually entered into the database. Hopefully an electronic database containing fuel tank capacities of all new vehicles will be available soon so that there will no longer be a need to manually enter this information into the database.

[2] At this point of the process, the trip log imputation has been completed, so the value from the trip log, reported or imputed, is considered as the true value for the vehicle.

- **Deterministic imputation using the trip log**

The objective here is to use trip log information to impute fuel log variables. For example, if the reported trip during which a fuel purchase was made can be found (using the purchase date or the odometer reading), the purchase date or the odometer reading can then be imputed using the information of the matched trip (we assume then that the fuel purchase was made at the end of the trip).

The trip log can also be used to impute the type of fuel purchased. Since the fuel type of the vehicle was already asked at the beginning of the trip log, this information can be used to impute a value on the fuel log. If the vehicle's fuel type is not gasoline, imputing is easy since there is a one-to-one match. Otherwise, one of the gasoline types (regular, mid-grade, premium, ethanol blend) will be selected with probability proportional to the quantity purchased from each gasoline type.

- **Various imputation methods using the fuel log**

Information from other fuel purchases, coming either from the same vehicle or different ones, can also be used to impute variables from the fuel log. Imputation methods that are used in the process include mean and ratio imputation, which are used to impute the date of purchase and the odometer reading, and nearest neighbour imputation, which is used to impute the price per unit.

- **Imputation using the fuel consumption regression model**

The fuel consumption model is the last, but also the most important imputation method used in this process. The primary objective of the model is to be able to predict the amount of fuel consumed by a vehicle based on the distance travelled and other characteristics that will be described later.

A study was previously made by the Automobile mobility data compendium (see Bonin, 2002) which attempted to determine the most suitable regression model to be used as a fuel consumption model and which dependent variables should be included in such a model. Most of their recommendations were retained for the CVS fuel consumption model. One of them was the use of a log-linear regression model which will predict the logarithm of the quantity of fuel consumed using a linear combination of dependant variables, as summarized in the following formula.

$$LN(\text{Fuel consumption}) = \beta_1 X + \beta_0 + \varepsilon$$

The log-linear model was chosen over the linear model because it provided the best fit for the CVS data.

Note that there are two different fuel consumption models applied to the CVS: one for light vehicles, the other for heavy vehicles. This was necessary due to the different fuel consumption patterns these two vehicle types have.

The following table gives a list of the dependant variables included in the fuel consumption models. An "X" indicates that the selected variable is included in the model.

Table 2: Dependant variables included in the fuel consumption models

Variable	Model for Light vehicles	Model for Heavy vehicles
LN(Distance)	X	X
3 or 4 cylinders	X	
5 or 6 cylinders	X	
3 to 6 cylinders		X
Manual transmission	X	X
Body type	X	X
Business use	X	
Rural indicator	X	
Over 15 tonnes		X

The first and most important dependant variable included in both models is the logarithm of distance travelled by the vehicle. It is obvious that the longer the distance a vehicle travels, the more fuel it will consume.

All remaining dependant variables are binary variables indicating if the vehicle possesses the vehicle characteristic. The first three indicator variables are related to the number of cylinders in the motor of the vehicle which is used as an approximation for vehicle weight. For the Light model, there are three categories: 3 or 4 cylinders, 5 or 6 cylinders, and More than 6 cylinders. Only two binary variables (3 to 4

cylinders and 5 to 6 cylinders) were created to avoid multicolineatiry problems. For the Heavy model, there are two categories: 3 to 6 cylinders, and More than 6 cylinders, which was not included in the model to avoid multicolinearity problems. The manual transmission variable is self-explanatory. The Body type group consists of different variables reflecting the answers to the trip log question about the vehicle's body type. The Business use variable indicates if the vehicle is used for commercial purposes. The Rural indicator tells if the owner of the vehicle lives in a rural area. Finally, the Over 15 tonnes indicator separates the heaviest heavy vehicles from the other heavy vehicles.

To create the necessary data to estimate the parameters β of the model, every reporting period between two fill-ups coming from a vehicle not needing imputation is used. Since the quantity of fuel purchased between two fill-ups corresponds exactly to the quantity of fuel consumed between these fill-ups, that quantity, the distance travelled between the two fill-ups (calculated using the odometer reading) and the vehicle characteristics can be entered in the regression.

The fuel consumption model can now be used in many parts of the imputation process. It can be used for the imputation of the quantity purchased (using the distance travelled between two fill-ups and the model), the fill-up indicator and the odometer reading (using the quantity of fuel purchased between two fill-ups, the distance travelled can be derived through the model and the odometer reading can be calculated from that distance and the odometer reading from another fill-up).

The fuel log, after going through all these imputation methods, can now be considered complete. The next step uses this fuel log to calculate fuel consumption.

5. CALCULATING FUEL CONSUMPTION

To determine the quantity of fuel consumed by a vehicle during the reference quarter, the fuel consumption rate has to be calculated. The vehicle's fuel consumption rate in litres per 100 kilometres is calculated using the distance travelled and the quantity of fuel consumed during the reporting period.

If two fill-ups or more were made for a vehicle, which is the ideal scenario, the distance travelled during the reporting period (fuel distance) will be the distance travelled between the first and the last fill-up made by the vehicle. If there was less than two fill-ups made, but at least two fuel purchases, the fuel distance for the vehicle will be the distance travelled between the first and the last fuel purchase. Otherwise, the distance travelled between the first and the last trip reported in the trip log will be used as the fuel distance.

If two fill-ups or more were made for a vehicle, the quantity of fuel consumed during the reporting period (fuel consumed) can be calculated by summing the quantity of fuel purchased between the first and the last fill-up made for the vehicle. Otherwise, the fuel consumption model will be used with the fuel distance previously calculated to derive the fuel consumed.

The fuel consumption rate (in L/100 km) for a vehicle can now be calculated using the following formula:

$$\text{Rate} = (\text{Fuel consumed} / \text{Fuel distance}) * 100$$

To calculate the quantity of fuel consumed by a vehicle during a trip, the following formula can be used:

$$\text{Fuel consumed} = (\text{Rate} * \text{Trip distance}) / 100$$

As previously stated, the fuel consumption model is mainly used to predict a quantity of fuel consumed given a certain distance travelled. However, when it comes to fuel consumption rates, the model can produce really extreme rate values when a very short or a very large distance is entered in the regression. So, an outlier detection module, based on the Hidiroglou-Berthelot method (1986), was built to identify these extreme fuel consumption rates. The outliers were then imputed to a value more suitable for their respective vehicle types.

6. IMPUTATION RESULTS

The first imputation table presents the fuel consumption model's coefficients of determination (R^2) for the reference year 2004 to see how well the CVS data fits the selected model.

Table 3: Coefficient of determination of fuel consumption model

Quarter	Coefficient of determination (R^2)	
	Light vehicles	Heavy vehicles
2004 Q1	80.0%	85.1%
2004 Q2	75.2%	82.6%
2004 Q3	72.4%	84.9%
2004 Q4	73.7%	82.2%

The coefficients of determination are quite good, varying from 72.4% to 85.1%, which means that the models have a good fit with the data. This is desirable since the fuel consumption model is used in many parts of the imputation process (as shown previously).

The next table presents the percentage of respondents who provided at least two fill-up, the minimum number of fill-ups that is necessary for a vehicle to calculate its fuel consumption rate using only reported data. These respondents were also used to estimate the parameters of the fuel consumption model, providing they did not need imputation.

Table 4: Percentage of respondents who provided two fill-ups or more for their vehicle

Quarter	Respondents with two fill-ups or more
2004 Q1	16.0%
2004 Q2	14.4%
2004 Q3	13.7%
2004 Q4	14.7%

This table shows that, for a majority of respondents, the fuel consumption model was needed to calculate the fuel consumption rate, which reflects the importance of a good fuel consumption model.

The last table presents imputation rates for key variables from the trip log.

Table 5: Imputation rates for key variables from the trip log

Variable	Imputation Rate for 2004 (Quarter)			
	Q1	Q2	Q3	Q4
Odometer reading	4.7%	0.5%	1.4%	1.4%
Quantity purchased	10.8%	12.6%	11.0%	13.9%
Price per unit	25.1%	26.1%	23.6%	25.2%
Fill-up indicator	12.1%	14.7%	11.3%	12.4%

The imputation rates per variables were stable throughout the year 2004. The odometer reading was the best answered variable, needing very little imputation. The price per unit, which needed imputation around a quarter of the times, was the variable needing the most imputation.

7. CONCLUSION

The fuel supplement to the Canadian Vehicle Survey was implemented in 2004 so that better estimates of fuel consumption and vehicle-kilometres could be provided to the sponsors of the survey, Transport Canada and Natural Resources Canada. Therefore it was important to develop and implement a good edit and imputation process. Various auxiliary information (trips from trip log, fuel tank capacity) and imputation methods (deterministic imputation, mean imputation, ratio imputation, nearest neighbour imputation, regression imputation) were used in the process, but the key element was the fuel consumption regression model, which was used to impute many variables from the fuel log and to calculate fuel consumption rates. The whole process contributed to the improvement of the quality of both the estimates and the microdata files provided to the sponsors.

ACKNOWLEDGEMENTS

The author would like to thank Martin Beaulieu, Danny Finch, François Gagnon, Jennifer Taylor and Julie Trépanier for their contribution to this project.

REFERENCES

Beaulieu, M. (2005). "The Estimation Methodology of the Redesigned Canadian Vehicle Survey", 2005 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM].

Bonin, S. (2002). "Proposed approach for the imputation of data from incomplete fuel purchase diaries of the National Private Vehicle Use Survey", Report no. N02-10fa, Automobile mobility data compendium, Université Laval, Quebec City.

Hidiroglou, M.A. and Berthelot, J.-M. (1986). "Statistical Editing and Imputation for Periodic Business Surveys", Survey Methodology, Vol. 12, pp. 73-83.

Statistics Canada (2005). “Canadian Vehicle Survey - First Quarter 2004”, Catalogue no. 53F0004XIE.