Canadian Vehicle Survey - Data Accuracy

While considerable effort is put forth to ensure that a high standard is maintained throughout all survey operations, the resulting estimates are inevitably subject to a certain degree of error. The total survey error is defined as the difference between the survey estimate and the true value for the population, at which the survey estimate aims. The total survey error consists of two types of errors: sampling and non-sampling errors.

Sampling error

When a sample is selected from a population, estimates based on the sample data may not be exactly the same as what would be obtained from a census of that population. The two results will likely differ since only data for sampled units are used. In the case of a census, there is no sampling error.

The difference between the estimates from a sample survey and a census conducted under the same conditions is referred to as the sampling error of a survey estimate. Factors such as the sample size, the sample design, the variability of the population characteristic under study and the estimation method affect the sampling error. If the population is very heterogeneous like the population of registered motor vehicles, a large sample size is needed to obtain reliable estimates.

The sampling error is measured by a statistical quantity called the standard error. This quantity reflects the expected variability of the survey estimate of a particular population characteristic if repeated sampling is carried out. The true value of the standard error is, of course, not known but can be estimated from the sample. The estimated standard error is used in terms of a relative measure called the coefficient of variation (or CV). This measure is simply the estimated standard error expressed as a percentage of the value of the survey estimate. Therefore, a smaller CV indicates better reliability of the estimate.

Non-sampling errors

The sampling error is only one component of the total survey error. All other errors arising from all phases of a survey are called non-sampling errors. As the sample size becomes closer to the population size, the sampling error component of the total survey error is expected to decrease. However, this is not necessarily true for the nonsampling error component. For example, this type of error can arise when a respondent provides incorrect information or does not answer certain questions, when a unit in the population of interest is omitted or covered more than once, when a unit that is out-of-scope for the survey is included by mistake or when errors occur in data processing, such as coding and capture errors.

Some non-sampling errors will cancel over a large number of observations, but systematically occurring errors (i.e. those that do not tend to cancel) will contribute to a bias in the estimates. For example, in the case of the CVS, if individuals that use their vehicles more than an average person consistently tend not to respond to the survey, then the resulting estimate of the total vehicle-kilometres will be below the true population total. Any such biases are not reflected in the estimates of standard error.

The non-sampling error as a whole is only one part of the total survey error but its contribution may be important. To minimize the effect of this type of error, a quality assurance program is carried out for each survey. For instance, follow-ups of nonrespondents can be conducted to obtain information from the total nonrespondents or to complete partially unanswered questionnaires for questions that are deemed essential. Various quality assurance procedures can be exercised at the data capture step. The data editing procedures can identify some inconsistencies in the data structure and the imputation procedures can then correct the identified inconsistencies.

In general, non-sampling errors are difficult to quantify. Special studies must be conducted to estimate them. However, certain measures such as response and imputation rates are easily obtained and can be used as indicators of the non-sampling errors. Different types of non-sampling errors are discussed below.

Coverage errors

Coverage errors arise when the survey population does not adequately cover the population of interest. As a result, certain units belonging to the population of interest are either excluded (undercoverage), or counted more than once (overcoverage). In addition, out of scope units may be present in the survey population (overcoverage).

The following sources of coverage errors for the CVS were observed:

- Errors in the classification variables of the survey may result in either under- or overcoverage of the registered vehicles.
- The sample is drawn from the list created three months prior to the beginning of the reference period. Thus the vehicles registered after the list was created and before the end of the reference period cannot be drawn into the sample.
- A vehicle list from any jurisdiction that was not created on time or did not arrive at all results in even larger undercoverage since an older list has to be used for sampling.
- A vehicle list created early causes overcoverage.
- A vehicle that has been scrapped or salvaged and remained on the list causes overcoverage.

- The survey population (see "Data quality, concepts and methodology" section of the Statistics Canada – Catalogue no. 53F0004X) can contain vehicles with the same Vehicle Identification Number (VIN), for example, when a vehicle is on the registration file of more than one jurisdiction. Since every vehicle has a unique VIN, this is likely to cause some overcoverage and consequently overestimation.
- A vehicle that was registered and subsequently unregistered between two consecutive registration lists causes undercoverage.

Thus the CVS is subject to some degree of under and over coverage. The estimation procedure is designed to compensate for the part of the under- and over coverage that has been determined.

Since we assume that the respondent is right (unless we have hard evidence to the contrary), the corrections at the estimation stage are mostly based on the respondent statements.

Response errors

Response errors occur when a respondent provides incorrect information due to a misinterpretation of the survey questions or due to a lack of correct information, or when a respondent is reluctant to disclose the correct information. Large response errors are likely to be caught during editing. However, others may simply go through undetected.

Few response errors were discovered during editing of the data.

Nonresponse errors

Nonresponse errors can occur when a respondent does not respond at all (total nonresponse) or responds only to some questions (partial nonresponse). These errors can have a serious effect if the nonrespondents are systematically different in survey characteristics from the respondents and/or the nonresponse rate is high. See the response rate tables in "Data quality, concepts and methodology" section of the Statistics Canada – Catalogue no. 53F0004X.

Processing errors

Apart from coverage, response and nonresponse errors described above, errors that occur during the processing of the data constitute another component of the non-sampling error. Processing errors can arise in data capture, coding, transcription, editing, imputation, outlier detection and treatment, and other types of data handling.

A coding error occurs when a field is coded erroneously because of a misinterpretation of the coding procedures or a bad judgment. A data capture

error occurs when the data are misinterpreted or keyed incorrectly. For example, an odometer reading of 53467 could be keyed as 54367.

Once data are coded and captured, they are subject to editing and imputation of missing or erroneous values. The quality of the data used in the estimation depends on the amount of imputation and the difference between the imputed and the true, but unknown, values. The imputation system could result in bias of the estimates. This can happen due to wrong assumptions or due to inability to impute. For example, in the CVS, it is impossible to detect, for vehicles that travel only a small distance during the reported period, fuel purchases that are missing or entered in error.

Quality indicators

Response rates by province and vehicle type vary usually between 50 and 80%.

The c.v. response rate and the relative imputation rate should be considered simultaneously to make an assessment of the reliability of an estimate. To assist the user in evaluating the potential effect of nonresponse, imputation and sampling error, a quality indicator accompanies every estimate. The quality indicator takes into account simultaneously the c.v. and the relative imputation rate.

Quality Symbol	C.V. equivalent	Explanation of estimate quality
A	Less than 5%	Excellent
B	5% to 10%	Very good
C	10% to 15%	Good
D	15to 20%	Acceptable
E	20% to 35%	Use with caution
F	35% or more	Too unreliable to be published

Using these ratings, at the national level, the estimates were judged generally of sufficient quality for publication. The provincial estimates are of a lesser quality in most cases.