Extract from the report

Methodology of the Canadian Vehicle Use Study

By Patrick Boily, Ph.D.

1. Objectives

The aim of the Canadian Vehicle Use Study (CVUS) is to measure various vehicle-related quantities (such as vehicle-km traveled, passenger-km traveled, fuel consumption, speed, fuel consumption ratio, etc) at different national, provincial and rural/urban levels, and to provide estimates of these quantities to the public, analysts and policy makers.

1.1 Canadian Vehicle Survey (CVS)

The Canadian Vehicle Survey (CVS) was conducted by Statistics Canada under contract to Transport Canada and Natural Resources Canada between 1999 and 2009. The quarterly survey employed a two-stage sample design: a sample of vehicles was selected and then a period of travel within the quarter was selected for each vehicle. Vehicles were grouped into three categories: light vehicles (passenger cars and light trucks/vans) and two types of heavy vehicles, based on the gross vehicle weight. A paper questionnaire was then mailed out to the owners of the selected vehicles, requesting that they record the number of trips, distance driven, and fuel consumption during the observation period.

The CVS was hampered by low participant response rates over its existence caused in large part by the burdensome paper collection methods. The quality of the estimates was also weakened by significant errors in the way in which the on-road vehicle fleet was classified.

1.2 Canadian Vehicle Use Study (CVUS)

As a result, Transport Canada decided to conduct a revised Canadian Vehicle Use Survey (CVUS), with improved methods. This includes the use of electronic data loggers to reduce reporting burden, introduction of a more robust vehicle decoder to increase the accuracy of the in-scope fleet, and a modified sampling design that includes the addition of additional strata to enhance the ability to carry out more detailed analyses of motor vehicle use.

[…]

7. Editing and Imputation

Ultimately, we would like to reach a quantitative understanding of some characteristic x for all vehicles in the population. The true population parameters (the mean μ, the variance σ^2, the quantiles q^α) for x remain unknown, but they can be estimated by judiciously selecting units from the population, observing a value of x for these units and using statistical sampling theory. This will be the topic of section 8.

However, before we can start doing so, we must first clarify what is meant by **characteristic**, **unit** and **observation**.

The **basic characteristics** of a vehicle's activity for a given day are:

a. nTrips – the number of trips;
b. VKT – vehicle-kilometres of travel or distance traveled by each vehicle in km;
c. PKT – passenger-kilometres (the product of VKT and the number of individuals in the vehicle);
d. Use – the number of hours for which the engine was turned on;
e. UseNI – the number of hours for which the engine was turned on and not idling;
f. Fuel – the fuel consumed in litres.

Vehicles present themselves as natural sampling units since we have access to a good sampling frame consisting of registered vehicles.[1] In the CVUS, the characteristics of interest are thus observed for sampled vehicles. We would like to define "observation" in such a way as to ensure that a single observation corresponds to each vehicle.

At the rawest level, an observation for x consists of a measurement of x for a specific vehicle over an interval of roughly one second. Over such a small interval of time, it seems safe to assume that the observation is quite precise

[1] Households or drivers could also have been used as units, but it is harder to get quality sampling frames in that case.

Purpose	Driver Gender	Driver Age	Occupancy	Trip Length	Type of Day
00_NONE	00_UNKNOWN	00_UNKNOWN	01_DRV_ALONE	00_IDLE	01_WORKDAY
01_WORK/BUSINESS	01_FEMALE	01_15-24	02_DRV_WITH_1_PASS	01_(0,5]	02_WEEKEND
02_SCHOOL/DAYCARE	02_MALE	02_25-44	03_DRV_WITH_2+_PASS	02_(5,10]	
03_SHOP/APP/ERRAND		03_45-64		03_(10,15]	
04_LEISURE/FAMILY/FRIENDS		04_65+		04_(15,20]	
05_COMMUNITY_SERVICE				05_(20,30]	
				06_(30,50]	
				07_(50,100]	
				08_100+	

Table 1 – Possible values of the trip identifiers.

(barring a possible malfunction of the recording equipment). Such precision comes at a price, however: 3 hours of travelling corresponds to roughly 10,800 observations. For a large number of vehicles, studied for weeks, the total size of the observations becomes prohibitive. The obvious solution is to consider an average: for instance, a vehicle travelling 200 km over 10,000 seconds travels at the average speed of 0.02 km/sec.

Such a small scale can be useful (we might want to determine which proportion of a trip was undertaken at speeds between two given thresholds, for instance; more on this later). Yet the scale of these observations leaves something to be desired: a conversion table can easily show that the average speed of a vehicle which travelled 12.1 meters in a second is 43.56 km/h, but the two quantities do not have the same power of invocation.

The permeating nature of the periodic day/night cycle in human affairs suggests that aggregating the raw observations at the daily level will provide a good balance between preciseness and ease of interpretation, certainly for the basic characteristics, for both actual study days and active days of observation. The next four sub-sections tackle this process of aggregation; the problem of transforming daily observations into a single observation for a given vehicle is described in the last sub-section.

7.1 Importing and Editing Data

Data are first collated at the trip level: each record consists of:

a. *trip and vehicle identifiers*: vehicle id, trip id, logger id;
b. *stratum identifiers*: province, vehicle type, vehicle age, forward sortation area;
c. *trip parameters*: trip year, trip month, trip day, trip start time, trip end time;
d. *trip identifiers*: purpose, driver age and gender, number of occupants, trip length, type of day;
e. *basic trip characteristics*: VKT, PKT, Use, UseNI, Fuel, and
f. *basic sub-trip characteristics:* VKT, PKT, Use, UseNI and Fuel, by cross-tabulations of engine temperature, vehicle speed and period of day.

The SAS code which collates the data is found in **Importing and Editing Data.sas**, Converting Raw CVUS Data.epg.

The allowed values of the trip identifiers are shown in Table 1.Within the study period for each vehicle, days for which it is not in use (**non-active days**) are added to the dataset, under the assumption that all basic trip and sub-trip characteristics take on the value 0 on these days.

7.2 Creating Daily Summaries

A problem appears for the first and last days of the study period: as we do not know exactly when the electronic logger has been installed (or uninstalled), we cannot *a priori* assume that the basic trip characteristics recorded on these days are complete. For instance, if the logger is installed at 10am on a Monday, any driving occurring before 10am will not be recorded. As such, the first and last days of a vehicle study should not be weighed in the same manner as the other (regular) days.

By convention, the **daily weight for vehicle** i on regular day is w_{reg}^{i}=1 (because a full day's worth of observations on these days is actually worth… one regular day of observations). To determine the daily weights of the first and last days, we proceed as follows.

For any vehicle i, let $b_{\min}^{i}$ (resp. $b_{\max}^{i}$) be the earliest start time (resp. latest end time) amongst all trips by that vehicle (as a fraction of a single day). The *base driving day* for vehicle i is the interval $[b_{\min}^{i}, b_{\max}^{i}] \subseteq [0,1]$.[2] Let α_i (resp. ω_i) be the earliest trip start time on the first day (resp. the latest trip end time on the last day) of the study. The daily weight w_{first}^{i} (resp. w_{last}^{i}) is the proportion of the base driving day occurring after α_i on the first day (resp. before ω_i on the last day), that is

$$w_{\text{first}}^{i} = \frac{b_{\max}^{i} - \alpha_i}{b_{\max}^{i} - b_{\min}^{i}} \quad \text{and} \quad w_{\text{last}}^{i} = \frac{\omega_i - b_{\min}^{i}}{b_{\max}^{i} - b_{\min}^{i}}.$$

For instance, if, amongst all trips, the earliest start time is 0.3 and the latest end time is 0.9, and if the earliest start time on the first day is 0.5 and the latest end time on the last day is 0.6, then

$$w_{\text{first}}^{i} = \frac{0.9 - 0.5}{0.9 - 0.3} = \frac{2}{3} \quad \text{and} \quad w_{\text{last}}^{i} = \frac{0.6 - 0.3}{0.9 - 0.3} = \frac{1}{2},$$

meaning that the observations on the first day are actually worth 2/3 regular days of observations and those on the last day are worth 1/2 regular days of observations.

The SAS code which adds the non-active days and computes the daily weights is found in **Creating Daily Summaries.sas**, Converting Raw CVUS Data.epg.

The observations then aggregated at the day-level, along trip identifiers: each record consists of:

a. *vehicle identifier*: vehicle id;
b. *stratum identifiers*: province, vehicle type, vehicle age, forward sortation area;
c. *travel parameters*: year, quarter, month, day, numerical date, weekday, active day flag;
d. *trip identifiers*: purpose, driver age and gender, number of occupants, trip length, type of day;
e. *basic characteristics*: daily weight, nTrips, VKT, PKT, Use, UseNI, Fuel, and
f. *basic sub-trip characteristics:* VKT, PKT, Use, UseNI and Fuel, by cross-tabulations of engine temperature, vehicle speed and period of day

For instance, the observations could be those shown in Table 2. Note the presence of non-active days (those rows for which the number of trips is 0), as well the daily weights on the first and last days for a given vehicle.

7.3 Rural / Urban Classification

The classification of a vehicle as belonging to either an urban or rural setting can be done with the Forward Sortation Area (FSA) portion of the postal code found in the registration file.[3]

The easiest way to do so is to use a system which is already in place: Canada Post defines an FSA as **rural** if the digit in the second position is a "0", and as **urban** otherwise. There are some issues with this approach, however.

[2] In practice, we allow for the possibility $b_{\max}^{i} > 1$: a trip which start on a given calendar day but ends on the following day has to be classified as occurring on a single day. We arbitrarily declare that the entire trip has taken place on the start date. In that case, the latest end time would actually be $1 +$ length of the trip in the early morning of the *second* day.

[3] For privacy reasons, the full address is not available before a vehicle has been selected.

v id	prov	type	age	fsa	year	qtr	month	day	date	week day	day type	active day flag	purpose cd	daily weight	nTrips	VKT	PKT	Use / 24	UseNI / 24	Fuel
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	14	18853	Sun	WeekEnd	1	0	0.963	1	0	0	0.0006	0	0.046
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	14	18853	Sun	WeekEnd	1	1	0.963	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	14	18853	Sun	WeekEnd	1	2	0.963	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	14	18853	Sun	WeekEnd	1	3	0.963	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	14	18853	Sun	WeekEnd	1	4	0.963	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	14	18853	Sun	WeekEnd	1	5	0.963	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	15	18854	Mon	WorkDay	1	0	1	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	15	18854	Mon	WorkDay	1	1	1	1	6.266	6.266	0.0083	0.0065	0.798
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	15	18854	Mon	WorkDay	1	2	1	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	15	18854	Mon	WorkDay	1	3	1	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	15	18854	Mon	WorkDay	1	4	1	7	299.131	588.919	0.2124	0.1981	27.610
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	15	18854	Mon	WorkDay	1	5	1	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	16	18855	Tue	WorkDay	0	0	1	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	16	18855	Tue	WorkDay	0	1	1	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	16	18855	Tue	WorkDay	0	2	1	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	16	18855	Tue	WorkDay	0	3	1	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	16	18855	Tue	WorkDay	0	4	1	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	8	16	18855	Tue	WorkDay	0	5	1	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮			⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	35_ON	02_LT	02_NEW	K8N	2011	3	9	4	18874	Sun	WeekEnd	1	0	0.698	2	44.076	44.076	0.0323	0.0302	3.954
2	35_ON	02_LT	02_NEW	K8N	2011	3	9	4	18874	Sun	WeekEnd	1	1	0.698	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	9	4	18874	Sun	WeekEnd	1	2	0.698	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	9	4	18874	Sun	WeekEnd	1	3	0.698	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	9	4	18874	Sun	WeekEnd	1	4	0.698	0	0	0	0	0	0
2	35_ON	02_LT	02_NEW	K8N	2011	3	9	4	18874	Sun	WeekEnd	1	5	0.698	0	0	0	0	0	0
3	35_ON	01_PC	02_NEW	K1C	2011	3	8	15	18854	Mon	WorkDay	1	0	0.985	0	0	0	0	0	0
3	35_ON	01_PC	02_NEW	K1C	2011	3	8	15	18854	Mon	WorkDay	1	1	0.985	0	0	0	0	0	0
3	35_ON	01_PC	02_NEW	K1C	2011	3	8	15	18854	Mon	WorkDay	1	2	0.985	0	0	0	0	0	0
3	35_ON	01_PC	02_NEW	K1C	2011	3	8	15	18854	Mon	WorkDay	1	3	0.985	1	11.058	11.058	0.0158	0.0098	1.394
3	35_ON	01_PC	02_NEW	K1C	2011	3	8	15	18854	Mon	WorkDay	1	4	0.985	2	15.022	30.044	0.0219	0.0144	1.759
3	35_ON	01_PC	02_NEW	K1C	2011	3	8	15	18854	Mon	WorkDay	1	5	0.985	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮			⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2 – Summarized data at the daily level (in order to make the table more readable, purpose is the only trip identifier retained and the basic sub-trip characteristics are not shown).

For instance, New Brunswick has recently changed its FSA codes so that none of the province's sortation areas will be classified as rural, in spite of the obvious fact that New Brunswick is not entirely made up of urban areas.

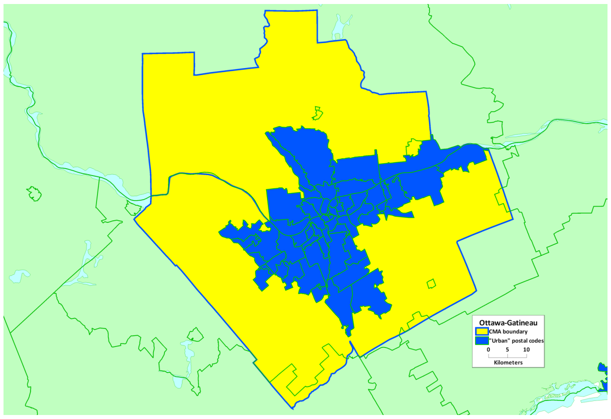
Furthermore, FSA codes may change fairly frequently, according to some arbitrary (at least, with respect to CVUS aims) internal logic at Canada Post. There is a chance that after such a change, a vehicle which would have been considered rural one day would suddenly be considered urban the next yet be used in the same area in both instances. Lastly, as we are not privy to the internal logic that allows the classification of FSAs, there remains the possibility that what would be considered rural in one jurisdiction may prove to be urban in another, cancelling any effort to provide estimates across jurisdictions.[4]

An eventual solution to this conundrum is to use population and population density data in order to classify the FSA: any FSA with a given population above a certain threshold and with a population density above a certain threshold is considered "urban", all other FSA are considered "rural".

This has the obvious advantage of being a uniform definition across all jurisdictions and sub-regions, and it avoids the pitfalls of random FSA classification changes by Canada Post.

Another solution involves manually selecting those FSA that intersect the boundaries of Census Metropolitan Areas (CMA) or some other municipal regroupings. The map on the following page showing the FSAs overlayed over the CMA

[4] Note: the Northwestern Territories and Nunavut share FSAs starting with the character "X": should this change at some point, the program **Importing and Editing Data.sas**, Converting Raw CVUS Data.epg will need to be edited to reflect this.



Map 1 – **Map of the Ottawa-Gatineau CMAs with overlapping FSAs.**

boundaries for Ottawa illustrate some of the problems associated with this approach: the overlap of FSA with the CMAs boundary is not exact.. For the time being, we use Canada Post's classification, as the thresholds mentioned in the approach above will depend on how many (and which) jurisdictions join the CVUS line-up.

The SAS code which adds the urban/rural classification to the tables is found in **Rural / Urban Classification.sas**, Converting Raw CVUS Data.epg. In Table 2, both vehicles 2 and 3 would be classified as URBAN since their FSA are K**8**N and K**1**C, respectively.

7.4 Basic and Derived Characteristics

Strictly speaking, a derived characteristic is a characteristic which is obtained by multiplying or dividing two or more basic characteristics. As such, PKT could be considered a derived characteristic, as it is obtained by multiplying VKT and the number of passengers; however, for the purposes of the CVUS, where the number of passengers is not a basic characteristics, it is classified as a basic characteristic.

The **derived characteristics** of a vehicle for a given day are ratios of basic trip characteristics. Some of these derived characteristics are more commonly recognizable under their common names: distance per hour of use is simply the **average vehicle speed**, whereas distance per litre consumed is the **average fuel consumption ratio** (after an appropriate re-scaling).

The following convention will be used to facilitate the reading of this document: the derived characteristic obtained by dividing the basic characteristic "a" by the basic characteristic "b" will be denoted by "a_b".

Each derived characteristic has an **associated daily characteristic weight**, which is simply the denominator in the computation of the ratio (which would be "b", above). In the event that the computation of a derived characteristic involves a division by 0 (i.e., if the associated weight is 0), we set the derived characteristic to 0. For instance, if on a given day the engine was started but the vehicle was not driven, the daily fuel consumption per km travelled is set to 0.

From 6 basic characteristics, 30 core derived characteristics can be built. They are presented in Table 3. At the subtrip-level, each of the variables is treated as a basic characteristic. There is nothing to stop us from creating derived characteristics for these variables, but it might not be practical to do so, due to their sheer quantity.

The SAS code which computes the basic and derived daily characteristics is found in **Basic and Derived Characteristics.sas**, Converting Raw CVUS Data.epg.

7.5 Vehicle Observations, Accuracy, Precision and Measurement Error

Following the previous sub-sections, let us assume that for a given vehicle j we have a series of i_j daily observations of the characteristic $x_{j,1}, \dots, x_{j,i_j}$, with accompanying weights $w_{j,1}, \dots, w_{j,i_j}$ $(\neq 0)$ and daily weights $v_{j,1}, \dots, v_{j,i_j}$.[5]

[5] For basic characteristics, the daily weights and accompanying weights are identical; for derived characteristics, they may not be.

Ratio of Column to Row	nTrips	VKT (km)	PKT (km)	Use (hr)	UseNI (hr)	Fuel (L)
nTrips		distance per trip (km)	passenger km per trip (km)	hours per trip (h)	non-idling hours per trip (h)	fuel consumption per trip (L)
VKT (km)	trips per km travelled (km^{-1})		passenger km per km travelled	hours per km travelled (h/km)	non-idling hours per km travelled (h/km)	fuel consumption per km travelled (L/km)
PKT (km)	trips per passenger km travelled (km^{-1})	distance per passenger km		hours per passenger km (h/km)	non-idling hours per passenger km (h/km)	fuel consumption per passenger km (L/km)
Use (h)	trips per hour of use (h^{-1})	distance per hour of use (km/h)	passenger km per hour of use (km/h)		ratio of non-idling use to use	fuel consumption per hour of use (L/h)
UseNI (h)	trips per hour of non-idling use (h^{-1})	distance per hour of non-idling use (km/h)	passenger km per hour of non-idling use (km/h)	ratio of use to non-idling use		fuel consumption per non-idling hour of use (L/h)
Fuel (L)	trips per litre consumed (L^{-1})	distance per litre consumed (km/L)	passenger km per litre consumed (km/L)	hours per litre consumed (h/L)	non-idling hours per litre consumed (h/L)	

Table 3 – Derived characteristics.

Write $z_j = \sum_{k=1}^{i_j} w_{j,k}$, $\xi_j = \sum_{k=1}^{i_j} w_{j,k}^2$, $\varphi_j = \sum_{k=1}^{i_j} w_{j,k}\, x_{j,k}$, $\zeta_j = \sum_{k=1}^{i_j} w_{j,k}\, x_{j,k}^2$ and $d_j = \sum_{k=1}^{i_j} v_{j,k}$. The weighted sample mean of the daily observations is

$$y_j = \frac{1}{z_j}\varphi_j.$$

The (weighted) sample variance of the observations is

$$s_j^2 = \frac{z_j}{z_j^2 - \xi_j} \sum_{k=1}^{i_j} w_{j,k} \left(x_{j,k} - y_j\right)^2 = \frac{z_j}{z_j^2 - \xi_j}\left(\zeta_j - z_j y_j^2\right).$$

Obviously, this is only well-defined for vehicles and characteristics for which $z_j^2 \neq \xi_j$.[6] We use the sample mean as the observation (or measurement) of the characteristic x for vehicle j.

Clearly, the number of observations affects the accuracy (how close the estimate is to the true value) and the precision (how small the variance of the estimate is) of the sample mean as an estimate of the true mean. If daily observations are available for every day in the time period of interest (a quarter, say), we can be reasonably certain that the sample mean is both very accurate and very precise: in fact, the sample mean is the true mean of x for vehicle j.[7] At the other extreme, if we only have one daily observation to work with, we have no way to determine the accuracy and precision of

[6] When some of the weights are not integers, z_j is a generalization of the number of observations in the computation of the sample mean, while the term $\frac{z_j^2 - \xi_j}{z_j}$ is a generalization of the degrees of freedom in the computation of the unbiased sample variance.

[7] If we assume that all other measurement errors are nil.

the sample mean (in this case, the lone daily observation) as an estimate of the true mean: it is possible that the sample mean could match the true mean, but we would not have enough information to qualify (let alone quantify) that statement.

If n daily observations of the characteristic x for vehicle j, each with weight $w_{j,k} = 1$, are drawn independently from an infinite population following a distribution $\mathcal{M}_j$ with mean μ_j and σ_j^2, then the accuracy of the sample mean y_j is measured by $A_j = y_j - \mu_j$, while its precision is measured by its variance

$$V(y_j) \approx \frac{\sigma_j^2}{n},$$

for large n. The **Central Limit Theorem** (CLT) guarantees that $A_j, e_j^2 \to 0$ as $n \to \infty$. In practice, however, the number of daily observations is limited by the number of available days: the variance must include a **finite population correction** factor $1 - \frac{n}{N}$ (FPC).

This can be generalized to our situation as follows. Let N be the number of days on which observations can be made. If i_j daily observations of the characteristic x for vehicle j, with accompanying weights $w_{j,k}$ and daily weights $v_{j,k}$, for $k = 1, \ldots, i_j$, are drawn independently and without replacement from a finite population following a distribution $\mathcal{M}_j$ with estimated mean $\hat{\mu}_j$ and estimated variance s_j^2, then the precision of the sample mean $y_j = \hat{\mu}_j$ is estimated by

$$e_j^2 = \begin{cases} \frac{s_j^2}{d_j}\left(1 - \frac{d_j}{N}\right), & \text{if } d_j < N \\ 0, & \text{otherwise} \end{cases}$$

Strictly speaking, the assumption of independence is not satisfied as they necessarily occur on consecutive days and are thus likely to be positively correlated at some level. However, over a long collection period, and perhaps due to the nature of the presumed dimorphism of driving behaviour between weekends and weekdays, it can be hoped that the assumption holds approximately. Note that for basic characteristics with integer weights equal to their daily weights, this does indeed collapse to the classical result.

A measure of accuracy is not provided as the only estimate of the true mean μ_j is the sample mean y_j itself, leading to $\hat{A}_j = 0$, no matter the sample size. Furthermore, accuracy is more easily affected by faulty or misused equipment than precision: constantly overshooting or undershooting the true daily observations by the same additive factor, for instance, would introduce a bias in the accuracy, but not in the precision.

As such, the observation of the characteristic x for a given vehicle j consists of the **mean** y_j, the **vehicle-characteristic weight** z_j and the **within-vehicle error** e_j^2. Thus, for each vehicle, there are 12 basic trip characteristics (days, active days) + 30 derived trip characteristics = 42 trip characteristics.

The basic sub-trip characteristics are simply the basic trip characteristics (except for nTrips), tabulated across 4 engine temperature categories (COLD: less than 80°C, WARM: 80°C to 100°C, HOT: more than 100°C, UNK: unknown), 6 period of the day (before morning traffic, during morning traffic, between morning and afternoon traffic, during afternoon traffic, after afternon traffic, overnight) and 10 instantaneous speed categories (idle, 0 km/h to 5 km/h, 5 km/h to 10 km/h, 10 km/h to 20 km/h, 20km/h to 30 km/h, 30 km/h to 50 km/h, 50 km/h to 80 km/h, 80 km/h to 100 km/h, 100 km/h to 120 km/h, more than 120 km/h). There are thus

$$4 + 6 + 10 + 4 \cdot 6 + 4 \cdot 10 + 6 \cdot 10 + 4 \cdot 6 \cdot 10 = 384$$

basic sub-trip characteristics for each of the 5 basic trip characteristics, hence 1920 basic sub-trip characteristics in total.

The edited dataset with which the analysis is conducted would then take on the form shown in Table 4.

The SAS code which computes the vehicle observation, weight and precision is found in **Vehicle Observations, Accuracy, Precision and Measurement Error.sas**, Converting Raw CVUS Data.epg

8. Estimation and Data Analysis

As was previously the case, we seek a quantitative understanding of some characteristic x for all vehicles in the population, in particular through the estimation of the true population parameters (the mean μ, the variance σ^2, etc).

8.1 Vehicle Observations at the Stratum Level

For the given characteristic x, let us assume that m vehicles have been sampled in a given stratum with overall population M. Thus we have a series of observations $(y_1, z_1, e_1^2), \dots, (y_m, z_m, e_m^2)$, as described in section 7.

Write $z = \sum_{j=1}^m z_j$, $\xi = \sum_{j=1}^m z_j^2$, $\varphi = \sum_{j=1}^m z_j\, y_j$, $\zeta = \sum_{j=1}^m z_j\, y_j^2$ and $\delta = \sum_{j=1}^m z_j\, e_j^2$. The estimate of the mean of x in the stratum is given by the (weighted) sample mean of the observations y_j:

$$\bar{y} = \frac{1}{z}\varphi.$$

The estimate for the variance in x in the stratum is slightly more complex: with perfect precision for each observation, only the (weighted) sample variance in y_j between the sampled vehicles contributes to the variance:

$$\hat{V}_b = \frac{z}{z^2 - \xi}\sum_{j=1}^{m} z_j\left(y_j - \bar{y}\right)^2 = \frac{z}{z^2 - \xi}\left(\zeta - z\bar{y}^2\right).$$

This **between-vehicle** contribution does not tell the whole variance-story, however, as each of the measurements y_j comes with a measure e_j^2 of its own **within-vehicle** uncertainty:

$$\hat{V}_w = \frac{z}{z^2 - \xi}\sum_{j=1}^{m} z_j\, e_j^2 = \frac{z}{z^2 - \xi}\delta.$$

It is reasonable to further assume that precision errors are independent of one another from vehicle to vehicle. The total (weighted) sample variance of the observations over the stratum is then estimated by

$$s_Y^2 = \hat{V}_b + \hat{V}_w = \frac{z}{z^2 - \xi}\left(\zeta - z\bar{y}^2 + \delta\right)$$

In order to provide an estimate for $s_{\bar{Y}}^2$ (the sample variance of the mean $\bar{y}$ over the stratum), keep in mind that both the number of sampled vehicles and the precision of their respective estimate affect the accuracy and precision of the sample mean $\bar{y}$ as an estimate of the true mean for the characteristic x at the stratum level.

Following the Central Limit Theorem argument presented in section 7.5, the stratum variance for the sample mean $\bar{y}$ in the stratum is estimated by

$$s_{\bar{Y}}^2 \approx \frac{\hat{V}_b}{m}\left(1 - \frac{m}{M}\right) + \frac{\hat{V}_w}{m} \approx \frac{s_Y^2}{m}\left(1 - \frac{m}{M}\right), \text{ when } m \ll M.$$

v id	prov	type	age	urbrural	fsa	week day	day type	purpose cd	data type	nTrips daily wt	nTrips daily	nTrips daily e2	VKT daily wt	VKT daily	VKT daily e2		Fuel UseNI wt	Fuel UseNI	Fuel UseNI e2
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	999999_ALL	999	16	21.618	3.806	0.271	21.618	60.620	279.483	...	19.803	6.856	0.134
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	999999_ALL	999	16	24.572	6.743	0.353	24.572	72.417	45.269	...	34.718	4.791	0.009
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	999999_ALL	0	17	21.618	3.436	0.261	21.618	46.493	174.464	...	14.893	7.209	0.204
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	999999_ALL	1	17	21.618	0.046	0.002	21.618	0.290	0.065	...	0.000	0.000	0.000
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	999999_ALL	2	17	21.618	0.000	0.000	21.618	0.000	0.000	...	0.000	0.000	0.000
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	999999_ALL	3	17	21.618	0.000	0.000	21.618	0.000	0.000	...	0.000	0.000	0.000
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	999999_ALL	4	17	21.618	0.324	0.081	21.618	13.837	148.002	...	0.000	0.000	0.000
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	999999_ALL	5	17	21.618	0.000	0.000	21.618	0.000	0.000	...	0.000	0.000	0.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	999999_ALL	0	17	24.572	0.857	0.032	24.572	2.561	0.733	...	1.424	5.592	0.840
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	999999_ALL	1	17	24.572	3.744	0.204	24.572	43.541	30.654	...	19.568	5.064	0.014
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	999999_ALL	2	17	24.572	0.000	0.000	24.572	0.000	0.000	...	0.000	0.000	0.000
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	999999_ALL	3	17	24.572	0.615	0.078	24.572	3.947	4.402	...	2.744	4.021	0.063
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	999999_ALL	4	17	24.572	1.486	0.108	24.572	21.948	32.880	...	10.753	4.393	0.038
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	999999_ALL	5	17	24.572	0.041	0.001	24.572	0.420	0.131	...	0.000	0.000	0.000
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	01_WorkDay	999	20	15.000	3.933	0.414	15.000	59.516	455.729	...	14.077	6.488	0.175
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	02_WeekEnd	999	20	6.618	3.516	0.927	6.618	63.124	810.123	...	5.726	7.762	0.582
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	01_WorkDay	999	20	17.053	6.454	0.599	17.053	69.336	68.974	...	22.296	4.936	0.011
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	02_WeekEnd	999	20	7.519	7.399	0.807	7.519	79.404	147.010	...	12.422	4.530	0.037
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	01_WorkDay	0	21	15.000	3.400	0.395	15.000	39.156	219.846	...	9.166	6.864	0.360
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	02_WeekEnd	0	21	6.618	3.516	0.927	6.618	63.124	810.123	...	5.726	7.762	0.582
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	01_WorkDay	5	21	15.000	0.000	0.000	15.000	0.000	0.000	...	0.000	0.000	0.000
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	99_ALL	02_WeekEnd	5	21	6.618	0.000	0.000	6.618	0.000	0.000	...	0.000	0.000	0.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	01_WorkDay	0	21	17.053	0.707	0.043	17.053	3.069	1.335	...	1.198	5.773	1.082
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	02_WeekEnd	0	21	7.519	1.197	0.123	7.519	1.408	1.098	...	0.227	4.637	1.577
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	01_WorkDay	5	21	17.053	0.000	0.000	17.053	0.000	0.000	...	0.000	0.000	0.000
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	99_ALL	02_WeekEnd	5	21	7.519	0.133	0.013	7.519	1.374	1.388	...	0.000	0.000	0.000
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	01_Mon	999999_ALL	999	24	3.000	6.667	0.370	3.000	112.679	7738.282	...	5.886	5.510	0.187
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	07_Sun	999999_ALL	999	24	3.618	2.009	0.158	3.618	14.719	63.106	...	1.325	4.989	0.270
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	01_Mon	999999_ALL	999	24	4.000	6.750	5.797	4.000	89.292	689.010	...	6.580	4.771	0.016
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	07_Sun	999999_ALL	999	24	4.000	6.500	1.688	4.000	71.602	329.549	...	6.091	4.129	0.010
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	01_Mon	999999_ALL	0	25	3.000	4.000	3.333	3.000	10.880	25.424	...	0.976	4.124	0.005
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	02_Tue	999999_ALL	0	25	3.000	1.667	1.204	3.000	29.891	744.269	...	1.378	6.158	1.026
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	06_Sat	999999_ALL	5	25	3.000	0.000	0.000	3.000	0.000	0.000	...	0.000	0.000	0.000
2	35_ON	02_LT	02_NEW	02_URBAN	K8N	07_Sun	999999_ALL	5	25	3.618	0.000	0.000	3.618	0.000	0.000	...	0.000	0.000	0.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	01_Mon	999999_ALL	0	25	4.000	0.500	0.063	4.000	0.008	0.000	...	0.007	5.616	29.448
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	02_Tue	999999_ALL	0	25	4.000	0.750	0.172	4.000	3.270	3.388	...	0.332	4.410	0.043
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	06_Sat	999999_ALL	5	25	3.519	0.284	0.062	3.519	2.935	6.619	...	0.000	0.000	0.000
210	35_ON	01_PC	01_NEWEST	02_URBAN	K2M	07_Sun	999999_ALL	5	25	4.000	0.000	0.000	4.000	0.000	0.000	...	0.000	0.000	0.000

Table 4 – Vehicle observations, at each level (in order to make the table more readable, purpose is the only trip identifier retained and the basic sub-trip characteristics are not shown).

When observations are available for each of the stratum vehicles, the precision of the sample mean as an estimate of the true mean is precisely that of the individual observations, which explains the finite population correction term in the "between" component of $s_{\overline{Y}}^2$. There is no such factor for the "within" component since its uncertainty goes to 0 with the number of sampling days, not with the number of sampled vehicles. However, the FPC is approximately equal to 1 when $m \ll M$, and $s_{\overline{Y}}^2$ can be assumed to take the classical form in our case.

As such, in the l^{th} stratum, the characteristic x is described by the **stratum mean** $\overline{x}_l = \overline{y}$, the **estimated variance of the stratum mean** $s_{\overline{X}_l}^2 = s_{\overline{Y}}^2$ and the **stratum weight** $M_l = M$.

In each stratum, the **coefficient of variation** $\mathrm{cv}(\mu_l)$ is obtained by dividing the standard deviation of the stratum mean by the mean:

$$\mathrm{cv}(\mu_l) = \frac{\sigma_{\mu_l}}{\mu_l} \approx \frac{s_{\overline{X}_l}}{\overline{x}_l}.$$

Confidence intervals (CI) are then easy to compute: an $(1-\alpha)\%$ confidence interval for μ_l is approximated by

$$CI_{(1-\alpha)}(\mu_l) = \overline{x}_l \pm z_\alpha \overline{x}_l \widehat{\mathrm{cv}}(\mu_l),$$

where z_α represents the $(1-\frac{\alpha}{2})^{\text{th}}$ percentile of the standard normal distribution.

8.2 Combining the Strata

For the given characteristic x, let us assume that vehicles are selected in k strata: thus we have a series of stratum statistics $\left(\overline{x}_1, s^2_{\overline{X}_1}, M_1\right), \ldots, \left(\overline{x}_k, s^2_{\overline{X}_k}, M_k\right)$, as described in the preceding section.

Write $M = \sum_{l=1}^{k} M_l$, $\phi = \sum_{l=1}^{k} M_l \overline{x}_l$ and $\tau = \sum_{l=1}^{k} M_l^2\, s^2_{\overline{X}_l}$. The estimate of the true mean of x over all strata is given by the (weighted) sample mean of the stratum means $\overline{x}_l$:

$$\overline{x} = \frac{1}{M}\phi.$$

The estimate for the variance in x over all strata is then simply obtained using the formulas of stratified sampling:

$$s^2_{\overline{X}} = \frac{1}{M^2}\sum_{l=1}^{k} M_l^2\, s^2_{\overline{X}_l} = \frac{1}{M^2}\tau.$$

Appendices

The first appendix (4 pages) contains the (unofficial) analysis results for Ontario during the first quarter of 2012.

The second appendix (27 pages) contains a description of how to use the results.