Contents

1 Survey of Quantitative Methods **2**
1.1 Principles of Data Collection 3
1.1.1 Questionnaire Design 6
1.1.2 Automated Data Collection 7
1.1.3 Statistical Survey Sampling 15
1.1.4 Case Study: Canada Vehicle Use Study 25

List of Figures

1 `robots.txt` file for a random webpage 10
2 Inspecting a webpage elements using Chrome's *Developer Tools* 13
3 The sampling model 16
4 Schematics of sampling designs 21

List of Tables

1 Survey of Quantitative Methods

The bread and butter of quantitative consulting is the ability to apply quantitative methods to business problems in order to obtain actionable insight. Clearly, it is impossible (and perhaps inadvisable, in a more general sense) for any given individual to have expertise in every field of mathematics, statistics, and computer science.

We believe that the best consulting framework is reached when a small team of consultants possesses expertise in 2 or 3 areas, as well as a decent understanding of related disciplines, and a passing knowledge in a variety of other domains: this includes keeping up with trends, implementing knowledge redundancies on the team, being conversant in non-expertise areas, and knowing where to find detailed information (online, in books, or through external resources).

In this section, we present an introduction for 9 "domains" of quantitative analysis:

- survey sampling and data collection;
- data processing;
- data visualisation;
- statistical methods;
- queueing models;
- data science and machine learning;
- simulations;
- optimisation, and
- trend extraction and forecasting;

Strictly speaking, the domains are not free of overlaps. Large swaths of data science and time series analysis methods are quite simply statistical in nature, and it's not unusual to view optimisation methods and queueing models as sub-disciplines of operations research. Other topics could also have been included (such as Bayesian data analysis or signal processing, to name but two), and might find their way into a second edition of this book.

Our treatment of these topics, by design, is brief and incomplete. Each module is directed at students who have a background in other quantitative methods, but not necessarily in the topic under consideration. Our goal is to provide a quick "reference map" of the field, together with a general idea of its challenges and common traps, in order to highlight opportunities for application in a consulting context. These subsections are emphatically NOT meant as comprehensive surveys: they focus on the basics and talking points; perhaps more importantly, a copious number of references are also provided.

We will start by introducing a number of motivating problems, which, for the most part, we have encountered in our own practices. Some of these examples are reported on in more details in subsequent sections, accompanied with (partial) deliverables in the form of charts, case study write-ups, report extract, etc.).

As a final note, we would like to stress the following: it is **IMPERATIVE** that quantitative consultants remember that acceptable business solutions are not always optimal theoretical solutions. Rigour, while encouraged, often must take a backseat to applicability. This lesson can be difficult to accept, and has been the downfall of many a promising candidate.

1.1 Principles of Data Collection

Fisher's Maxim

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

– R.A. Fisher, Presidential Address to the *First Indian Statistical Congress*, 1938

Data analysis tools and techniques work in conjunction with collected data. The type of data that needs to be collected to carry out such analyses, as well as the priority placed on the collection of quality data relative to other demands, will dictate the choice of data collection strategies. The manner in which the resulting outputs of these analyses are used for decision support will, in turn, influence appropriate data presentation strategies and system functionality (we will revisit this aspect in a later section).

Although consultants should always endeavour to work with **representative** and **unbiased data**, there will be times when the available data is flawed and not easily repaired. Quantitative consultants have a professional responsibility to explore the data, looking for potential fatal flaws **prior** to the start of the analysis and to inform their client of any findings that could halt, skew, or simply hinder the analytical process or its applicability to the situation at hand.

Unless a clause has specifically been put in the contract to allow a graceful exit at this point, you will have to proceed with the analysis, flaws and all. It is **EXTREMELY IMPORTANT** that you do not simply sweep these flaws under the carpet. Address them repeatedly in your meetings with the clients, and make sure that the analysis results you present or report on include an appropriate *caveat*.

Data Collection System Consultants might also be called upon to provide suggestions to evaluate or fix the data collection system. The following items could help with that task.

- **Data Validity**: the system must collect the data in such a way that data validity is ensured during initial collection. In particular, data must be collected in a way that ensures sufficient accuracy and precision of the data, relative to its intended use.
- **Data Granularity, Scale of Data**: the system must collect the data at a level of granularity appropriate for future analysis.
- **Data Coverage**: the system must collect data that comprehensively, rather than only partially or unevenly, represents the objects of interest. As well, the system must collect and store the required data over a sufficient amount of time, and at the required intervals, to support data analyses that require data spanning a certain duration.
- **Data Storage**: the system must have the functionality to store the types and amount of data required for a particular analysis.
- **Data Accessibility**: the system must provide access to the data relevant for a particular analysis, in a format that is appropriate for this analysis.
- **Computational/Analytic Functionality**: the system must have the ability to carry out the computations required by relevant data analysis techniques.
- **Reporting, Dashboard, Visualization**: the system must be able to present the results of the data analysis in a meaningful, usable and responsive fashion.

A number of different overarching strategies for data collection can be employed. Each of these different strategies will be more or less appropriate under certain data collection circumstances, and will result in different system functional requirements. In this section we will focus on survey sampling, questionnaire design, and automated data collection.

Formulating the Problem The **objectives** drive all other aspects of quantitative analysis. With a **question** (or questions) in mind, an investigator can start the process that leads to **model selection**. With potential models in tow, the next step is to consider what **variates** (fields, variables) are needed, the **number** of observations required to achieve a pre-determined **precision**, and how to best go about **collecting**, **storing** and **accessing** the data.

Another important aspect of the problem is to determine whether the questions are being asked of the data in and of **itself**, or whether the data is used as a **stand-in for a larger population**. In the later case, there are other technical issues to incorporate into the analysis in order to be able to obtain generalizable results.

Questions do more than just drive the other aspects of data analysis – they also drive the development of quantitative methods. They come in all flavours and their variability and breadth make attempts to answer them challenging: no single approach can work for all of them, or even for a majority of them, which leads to the discovery of better methods, which are in turn applicable to new situations, and so on, and so on.

Not every question is answerable, of course, but a large proportion of them may be answerable partially or completely; quantitative methods can provide insights, estimates and ranges for possible answers, and they can point the way towards possible implementations of the solutions.

As an example, consider the following questions.

- Is cancer incidence higher for second-hand smokers than it is for smoke-free individuals?
- Using past fatal collision data and economic indicators, can we predict future fatal collision rates given a specific national unemployment rate?
- What effect would moving a central office to a new location have on average employee commuting time?
- Is a clinical agent effective in the treatment against acne?
- Can we predict when border-crossing traffic is likely to be higher than usual, in order to appropriately schedule staff rotations?
- Can personalized offers be provided to past clients to increase the likelihood of them becoming repeat customers?
- Has employee productivity increased since the company introduced mandatory language training?
- Is there a link between early marijuana use and heavy drug use later in life?
- How do selfies from over the world differ in everything from mood to mouth gape to head tilt?

How can such questions be answered? In many instances, the next step requires obtaining data.

Data Types Data has **attributes** and **properties**. Fields are classified as **response**, **auxiliary**, **demographic** or **classification** variables; they can be **quantitative** or **qualitative**; **categorical**, **ordinal** or **continuous**; **text-based** or **numerical**.

Data is **collected** through experiments, interviews, censuses, surveys, sensors, scraped from the Internet, etc.

Collection methods are not always sophisticated, but new technologies usually improves the process in many ways (while introducing new issues and challenges): modern data collection can occur over one pass, in batches, or continuously.

How does one decide which data collection method to use? The type of question to answer obviously has an effect, as do the required precision, cost and timeliness. Statistics Canada's *Survey Methods and Practices* provides a wealth of information on probabilistic sampling and questionnaire design, which remain relevant in this day of big (and real-time) data.

A full discussion on the various collection methods currently available is outside the scope of this section, but the importance of this step cannot be overstated: without a well-designed plan to collect meaningful data, and without safeguards to identify flaws (and possible fixes) as the data comes in, subsequent steps are likely to prove a waste of time and resources.

As an illustration of the potential effect that data collection can have on the final analysis results, contrast the two following ways to collect similar data on two different populations.

Yes. I Mean No. ... I Think.

The Government of Québec has made public its proposal to negotiate a new agreement with the rest of Canada, based on the equality of nations; this agreement would enable Québec to acquire the exclusive power to make its laws, levy its taxes and establish relations abroad — in other words, sovereignty — and at the same time to maintain with Canada an economic association including a common currency; any change in political status resulting from these negotiations will only be implemented with popular approval through another referendum; on these terms, do you give the Government of Québec the mandate to negotiate the proposed agreement between Québec and Canada?

– 1980 Québec sovereignty referendum question

Do You Think They Learned Something From 1980?

Should Scotland be an independent country?

– 2014 Scotland independence referendum question

The end result was the same in both instances, but arguments can be made that the Scottish 'No' was a much clearer 'No' than the Québec 'No' of 34 years earlier – in spite of the smaller 2014 victory margin (55.3%-44.7%, as opposed to 59.6%-40.4%).

Data Storage and Access Data **storage** is also strongly linked with the data collection process, in which decisions need to be made to reflect how the data is being collected (one pass, batch, continuously), the volume of data that is being collected, and the type of access and processing that will be required (how fast, how much, by whom).

Stored data may go **stale** (e.g. people move, addresses no longer accurate), so it may be necessary to implement regular updating collection procedures.

Until very recently, the story of data analysis has been written for small datasets: useful collection techniques yielded data that could, for the most part, be stored on personal computers or on small servers. The advent of Big Data has introduced new challenges *vis-à-vis* the collection, capture, access, storage, analysis and visualisation of datasets; some effective solutions have been proposed and implemented, and intriguing new approaches are on the way (such as DNA storing, to name but one). We shall not discuss those challenges in detail, but be aware of their existence.

1.1.1 Questionnaire Design

A Modern Paradox

People resist a census, but give them a profile page and they'll spend all day telling you who they are.

– Max Berry, *Lexicon*, 2013

A **questionnaire** is a sequence of questions designed to obtain information on a subject from a respondent. Design principles vary according to the subject matter and the mode of data collection, but we strongly encourage pre-testing a variety of questionnaires.

Questionnaire Design Basics In general, questionnaires should:

- be as brief as possible, with no wasted questions;
- be accompanied by clear and concise instructions;
- keep the respondent's interests in mind;
- emphasise confidentiality;
- be serious and courteous in tone;
- be free of mistakes and laid out attractively;
- be worded clearly;
- be designed to be accurately answered, and
- be ordered attentively.

Question Types The basic questionnaire unit is, of course, the **question**, which comes in two flavours:

- **closed**, with a fixed number of pre-determined mutually exclusive and collectively exhaustive answer choices (and should always include an "Other (Please Specify)" category to counteract the loss of expressiveness of such questions), and
- **open**, which serves to identify common response choices to be used as closed question choices in subsequent questionnaires.

Wording Considerations It is well known that the wording of the questions can influence a questionnaire's responses [5]; please keep the following **wording considerations** in mind when designing a questionnaire:

- avoid **abbreviations** and **jargon** ("Does your organization use any TTWQ practices?");
- do not use words and terminology that are **too complex** ("How often have you been defenestrated?" vs. "How often have you been thrown out of a window?");
- specify the **frame of reference** ("What is your income?" vs. "What was your household's total income form all sources before taxes and deductions in 2017?");
- make the question as **specific** as possible ("How much fuel did your moving company use during the last year?" vs. "How much did your moving company spend on fuel during the last year?");
- ensure that the questions can be answered by all respondents;
- avoid **double-barrelled** questions ("Do you plan to leave your car at home and take the bus to work during the coming year?"), and
- avoid leading questions (see the always excellent [7] for a not-so-facetious example).

Question Order The order of the questions is just as important as the wording. Questionnaires should be designed to **flow smoothly** and **follow a logical sequence** (logical to the respondent, that is).

1. Start with an **introduction** which provides the title, subject, and purpose of the survey.
2. Request **cooperation** and explain the importance of the survey and how the results will be used.
3. Indicate the degree of **confidentiality** and provide a deadline and a contact address.
4. Open with a series of **easy** and **interesting questions** to establish the respondent's confidence.
5. Group similar questions under a **common heading**.
6. Introduce **sensitive topics** once trust and confidence are likely to have developed.
7. Allow some space and/or time for **additional comments**.
8. **Thank** the respondent for their participation.

A lot more has been written about questionnaire design (see [3], for instance). It can be surprisingly easy to get lost in the jungle and spend way too much time on the "perfect" design; remember that without a sound sampling plan, whatever data is collected may not prove up to the task of drawing the actionable insights that the client is really interested in seeing answered.

1.1.2 Automated Data Collection

One Man's Trash...

It's been said that the streets of the Web are paved with data that cannot wait to be collected, but you'd be surprised at how much trash there is out there.

– Patrick Boily, 2018

The way we **share**, **collect**, and **publish** data has changed over the past few years due to the ubiquity of the *World Wide Web*. **Private businesses**, **governments**, and **individual users** are posting and sharing all kinds of data and information. At every moment, new channels generate vast amounts of data.

There was a time in the recent past where both scarcity and inaccessibility of data was a problem for researchers and decision-makers. That is **emphatically** not the case anymore.

Data abundance carries its own set of problems, however, in the form of

- tangled masses of data, and
- traditional data collection methods and classical (small) data analysis techniques not being up to the task anymore.

The growth and increasing popularity and power of **open source software**, such as *R* and *Python*, for which the source code can be inspected, modified, and enhanced by anyone, makes program-based automated data collection quite appealing.

One note of warning, however: time marches on and packages become **obsolete** in the blink of an eye. If the analyst is unable (or unwilling) to **maintain their extraction/analysis code** and to **monitor the sites** from which the data is extracted from, the choice of software will not make much of a difference.

So why bother with automated data collection? Common considerations include:

- the sparsity of financial resources;
- the lack of time or desire to collect data manually;
- the desire to work with up-to-date, high-quality data-rich sources, and
- the need to document the analytical process from beginning (data collection) to end (publication).

Manual collection, on the other hand, tends to be cumbersome and prone to error; non-reproducible processes are also subject to heightened risks of "death by boredom", whereas program-based solutions are typically more reliable, reproducible, time-efficient, and produce datasets of higher quality (this assumes, of course, that coherently presented data exists in the first place).

Automated Data Collection Checklist That being said, **web scraping** or **statistical text processing** is not always recommended. As a start, it is possible that no online and freely available source of data meets the client's needs, in which case an approach based on survey sampling is probably indicated.

If most of the answers to the following questions are positive, then an automated approach may be the right choice.

- Is there a need to repeat the task from time to time (e.g. to update a database)?
- Is there a need for others to be able to replicate the data collection process?
- Are online sources of data frequently used?

- Is the task non-trivial in terms of scope and complexity?
- If the task can be done manually, are the financial resources required to let others do the work lacking?
- Is the will to automate the process by means of programming there?

The objective is simple: automatic data collection should yield a collection of unstructured or unsorted datasets, at a reasonable cost.

Ethical Considerations We now turn our attention to a burning question for consultants and analysts alike: is all the freely available data on the Internet ACTUALLY freely available?

A **spider** is a program that grazes or crawls the web rapidly, looking for information. It jumps from one page to another, grabbing the entire page content. **Scraping** is taking specific information from specific websites (which is the goal): how are these different?

> "Scraping inherently involves **copying**, and therefore one of the most obvious claims against scrapers is copyright infringement." [8]

So what can be done to minimize the risk?

- Work as transparently as possible;
- Document data sources at all time;
- Give credit to those who originally collected and published the data;
- If you did not collect the information, you probably need permission to reproduce it, and, more importantly,
- Don't do anything illegal.

A number of cases have shown that the courts have not yet found their footing in this matter (see *eBay* vs. *Bidder's Edge*, *Associated Press* vs. *Meltwater*, *Facebook* vs. *Pete Warden*, *United States* vs. *Aaron Swartz*, for instance [9]). There are legal issues that we are not qualified to discuss, but in general, it seems as though larger companies/organisations usually emerge victorious from such battles.

Part of the difficulty is that it is not clear which scraping actions are illegal and which are legal. There are rough guidelines: re-publishing content for commercial purposes is considered more problematic than downloading pages for research/analysis, say. A site's `robots.txt` (Robots Exclusion Protocol) file tells scrapers what information on the site may be harvested with the publisher's consent – heed that file (see Figure 1 for an example).

Perhaps more importantly, **be friendly!** Not everything that can be scraped needs to be scraped. Scraping programs should 1) behave "nicely"; 2) provide useful data, and 3) be efficient, in that order. When in doubt, contact the data provider to see if they will grant access to the databases or files.

```
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
# Files
Disallow: /CHANGELOG.txt
Disallow: /cron.php
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /INSTALL.sqlite.txt
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /LICENSE.txt
Disallow: /MAINTAINERS.txt
Disallow: /update.php
```

Figure 1: `robots.txt` file for **cqads.carleton.ca**.

Finally, note the importance of following the **Scraping Do's and Don't's**:

1. **stay identifiable**;
2. **reduce traffic** – accept compressed files, check that a file has been changed before accessing it again, retrieve only parts of a file;
3. **do not bother server with multiple requests** – many requests per second can bring smaller server downs, webmasters may block you if your scraper is too greedy (a few requests per second is fine), and
4. **write efficient and polite scrapers** – there is no reason to scrape pages daily or to repeat the same task over and over, select specific resources and leave the rest untouched.

Web Data Quality Data quality issues are inescapable. It is not rare for clients to have spent thousands of dollars on data collection (automatic or manual) and to respond to the news that the data is flawed or otherwise unusable with: "well, it's the best data we have, so find a way to use it."

These issues can be side-stepped to some extent if consultants get involved in the project during or prior to the data collection stage, asking questions such as

- What type of data is best-suited to answer the client's question(s)?
- Is the available data of sufficiently high quality to answer the client's question(s)?
- Is the available information systematically flawed?

Web data can be **first-hand** information (a tweet or a news article), or **second-hand** (copied from an offline source or scraped from some online location, which may make it difficult to retrace). **Cross-referencing** is a standard practice when dealing with secondary data.

Data quality also depends on its **use(s)** and **purpose(s)**. For example, a sample of tweets collected on a random day could be used to analyse the use of a hashtags or the gender-specific use of words, but that dataset might not prove as useful if it had been collected on the day of the 2018 U.S. Presidential Election to predict the election outcomes (due to **collection bias**).

An example might help to illustrate some the pitfalls and challenges. Let's say that a client is interested in finding out what people think of a new potato peeler using a standard telephone survey. Such an approach has a number of pitfalls:

- **unrepresentative sample** – the selected sample might not represent the intended population;
- **systematic non-response** – people who don't like phone surveys might be less (or more) likely to dislike the new potato peeler;
- **coverage error** – people without a landline can't be reached, say, and
- **measurement error** – are the survey questions providing suitable info for the problem at hand?

Traditional solutions to these problems require the use of survey sampling (more on this later), questionnaire design (see previous section), omnibus surveys, reward systems, audits, etc. These solutions can be **costly**, **time-consuming**, and **ineffective**.

Proxies – indicators that are strongly related to the product's popularity without measuring it directly, could be useful. If **popularity** is defined as large groups of people preferring a potato peeler over another one, then sales statistics on a commercial website may provide a proxy for popularity.

Rankings on `Amazon.ca` (or a similar website) could, in fact, paint a more comprehensive portrait of the potato peeler market than would a traditional survey. It would suffice, then, to build a scraper that is compatible with Amazon's **application program interface** (API) to gather the appropriate data.

Of course, there are potential issues with this approach as well:

- **representativeness** of the **listed products** – Are all potato peelers listed? If not, is it because that website doesn't sell them? Is there some other reason?
- **representativeness** of the **customers** – Are there specific groups buying/not-buying online products? Are there specific groups buying from specific sites? Are there specific groups leaving/not-leaving reviews?
- **truthfulness** of customers and **reliability** of reviews – how can we distinguish between paid (fake) reviews and real reviews?

Web scraping is usually well-suited for collecting data on products (such as the aforementioned potato-peeler), but there are numerous questions for which it is substantially more difficult to imagine where data could be found online: what data could you collect online to measure the popularity of a government policy, say?

Web Technologies 101 Online data can be found in **text**, **tables**, **lists**, **links**, and other structures, but the way data is presented in browsers is not necessarily how it is stored in HTML/XML. Furthermore, when web pages are **dynamic**, there is a "cost" associated with automated collection. Consequently, a basic knowledge of the web and web-related techs and documents is crucial. Information is readily available online (see references) and in [8, 9].

There are three areas of importance for data collection on the web:

- technologies for **content dissemination** (HTTP, HTML/XML, JSON, plain text, etc.);
- technologies for **information extraction** (R, Python XPath, JSon parsers, Beautiful Soup, Selenium, regexps, etc.), and
- technologies for **data storage** (R, Python, SQL, binary formats, plain text formats, etc.).

Webpage content itself comes into three main categories: Hypertext Markup Language (HTML; used for web content and code), Cascading Style Sheets (CSS; used for webpage style), and JavaScript (js; used for interactivity with the webpage). HTML is, in some sense, the most fundamental; understanding the tree structure of HTML documents, for instance, will go a long way towards helping consultants get full use of the **scraping toolbox**.

Scraping Toolbox Our experience has shown that a number of tools can facilitate the automated data extraction process, including: *Developer Tools*, *XPath*, *Beautiful Soup*, *Selenium*, and *regular expressions*.

Developer Tools show the correspondence between the HTML code for a page and the rendered version seen in the browser (see Figure 2 for an example).

Unlike "View Source", Developer Tools show the *dynamic* version of the HTML content (i.e. the HTML is shown with any changes made by JavaScript since the page was first received).

Inspecting a page's various elements and discovering where they reside in the HTML file is **crucial** to efficient web scraping:

- **Firefox** – right click page → Inspect Element
- **Safari** – Safari → Preferences *to* Advanced → Show Develop Menu in Menu Bar, then Develop → Show Web Inspector
- **Chrome** – right click page → Inspect

XPath is a query (domain-specific) language which is used to select specific pieces of information from marked-up documents such as HTML, XML, or variants such as SVG, RSS. Before this can be done, the information stored in a marked-up document needs to be converted (or **parsed**) into a format suitable for processing and statistical analysis; this is implemented in the R package XML, for instance.

The process is simple; it involves

1. specifying the data of interest;
2. locating it in a specific document, and
3. tailoring a query to the document to extract the desired info.

XPath queries require both a **path** and a **document** to search; paths consist of hierarchical addressing mechanism (succession of nodes, separated by forward slashes ("/"), while a query takes the form `xpathSApply(doc,path)`.

For instance, `xpathSApply(parsed_doc,''/html/body/div/p/i'')` would find all `<i>` tags found inside a `<p>` tag, itself found inside a `<div>` tag in the `body` of the `html` file of `parsed_doc`.

Consult [8] for a substantially heftier introduction.

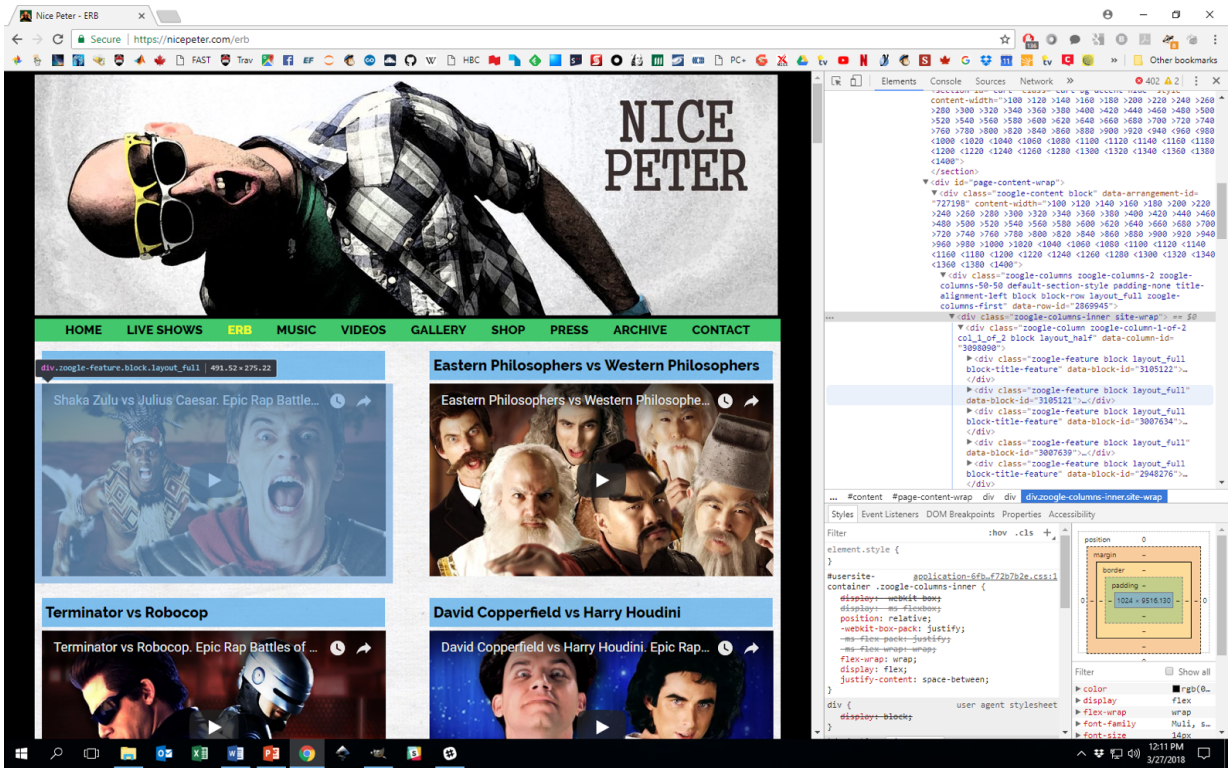


Figure 2: Inspecting **nicepeter.com/erb**'s elements using Chrome's *Developer Tools*.

Regular Expressions can be used to achieve the main web scraping objective, which is to extract relevant information from reams of data.

Among this mostly unstructured data lurk **systematic elements**, which can be used to help the automation process, especially if quantitative methods are eventually going to be applied to the scraped data.

Systematic structures include numbers, names (countries, etc.), addresses (mailing, e-mailing, URLs, etc.), specific character strings, etc.

Regular expressions (regexps) are abstract sequences of strings that match concrete recurring patterns in text; they allow for the systematic extraction of the information components from plain text, HTML, and XML.

Some examples that illustrate the main concepts are shown in one of the *Jupyter Notebooks*

Beautiful Soup is a Python library that helps extract data out of HTML and XML files. It parses HTML files, even if they're broken.

Beautiful Soup does not simply convert bad HTML to good X/HTML; it allows a user to fully inspect the (proper) HTML structure it produces, in a programmatical fashion.

When Beautiful Soup has finished its work on an HTML file, the resulting *soup* is an API for **traversing**, **searching**, and **reading** the document's elements. In essence, it provides **idiomatic** ways of navigating, searching, and modifying the parse tree of the HTML file, which can save a fair amount of time.

For instance, `soup.find_all('a')` would find and output all `<a ...> ... </a>` tag pairs (with attributes and content) in the `soup`, whereas

```
for link in soup.find_all('a'):
print(link.get('href')
```

would output the URLs found in the same tag pairs.

The Beautiful Soup documentation is quite explicit [14].

Selenium is a Python tool used to automate web browser interactions. It is used primarily for testing purposes, but it has data extraction uses as well.

Mainly, it allows the user to open a browser and to act as a human being would:

- clicking buttons;
- entering information in forms;
- searching for specific information on a page, etc.

Selenium requires a driver to interface with the chosen browser. Firefox, for example, uses `geckodriver`.

Other supported browsers have their own drivers (see [15–18]).

Selenium automatically controls a complete browser, including rendering the web documents and running JavaScript. This is useful for pages with a lot of dynamic content that isn't in the base HTML.

Selenium can program actions like "click on this button", or "type this text", to provide access to the dynamic HTML of the current state of the page, not unlike what happens in *Developer Tools* (but now the process can be fully automated).

More information can be found in [12, 13].

Let us end this section by providing a short summary of the **automated data collection decision process** [8, 9], as seen by quantitative consultants.

1. **Know exactly what kind of information the client needs**, either **specific** (e.g. GDP of all OECD countries for last 10 years, sales of top 10 tea brands in 2017, etc.) or **vague** (people's opinion on tea brand X, etc.)

2. **Find out if there are any web data sources that could provide direct or indirect information on the client's problem.** That is easier to achieve for specific facts (a tea store's webpage will provide information about teas that are currently in demand) than it is for vague facts. Tweets and social media platforms may contain opinion trends; commercial platforms can provide information on product satisfaction.

3. **Develop a theory of the data generation process when looking into potential data sources.** When was the data generated? When was it uploaded to the Web? Who uploaded the data? Are there any potential areas that are not covered, consistent, or accurate? How often is the data updated?

4. **Balance the advantages and disadvantages of potential data sources.** Validate the quality of data used – are there other independent sources that provide similar information against which to crosscheck? Can original source of secondary data be identified?

5. **Make a data collection decision**. Choose the data sources that seem most suitable, and document reasons for this decision. Collect data from several sources to validate the final choice.

1.1.3 Statistical Survey Sampling

You Can't Say It's Not True

The latest survey shows that 3 out of 4 people make up 75% of the world's population.

– David Letterman (attributed)

While the *World Wide Web* does contain troves of data, web scraping does not address the question of data validity: will the extracted data be **useful** as an analytical component? Will it suffice to provide the quantitative answers that the client is seeking?

A **survey** (a fair amount of information for this section is taken from [1, 6]) is any activity that collects information about characteristics of interest:

- in an **organized** and **methodical** manner;
- from some or all **units** of a population;
- using **well-defined** concepts, methods, and procedures, and
- compiles such information into a **meaningful** summary form.

A **census** is a survey where information is collected from all units of a population, whereas a **sample survey** uses only a fraction of the units.

Sampling Model When survey sampling is done properly, we may be able to use various statistical methods to make inferences about the **target population** by sampling a (comparatively) small number of units in the **study population**. The relationship between the various populations (**target**, **study**, **respondent**) and samples (**sample**, **intended**, **achieved**) is illustrated in Figure 3.

Deciding Factors In some instances, information about the **entire** population is required in order to solve the client's problem, whereas in others it is not necessary. How does one determine which type of survey must be conducted to collect data? The answer depends on multiple factors:

- the type of question that needs to be answered;
- the required precision;
- the cost of surveying a unit;
- the time required to survey a unit;
- size of the population under investigation, and
- the prevalence of the attributes of interest.

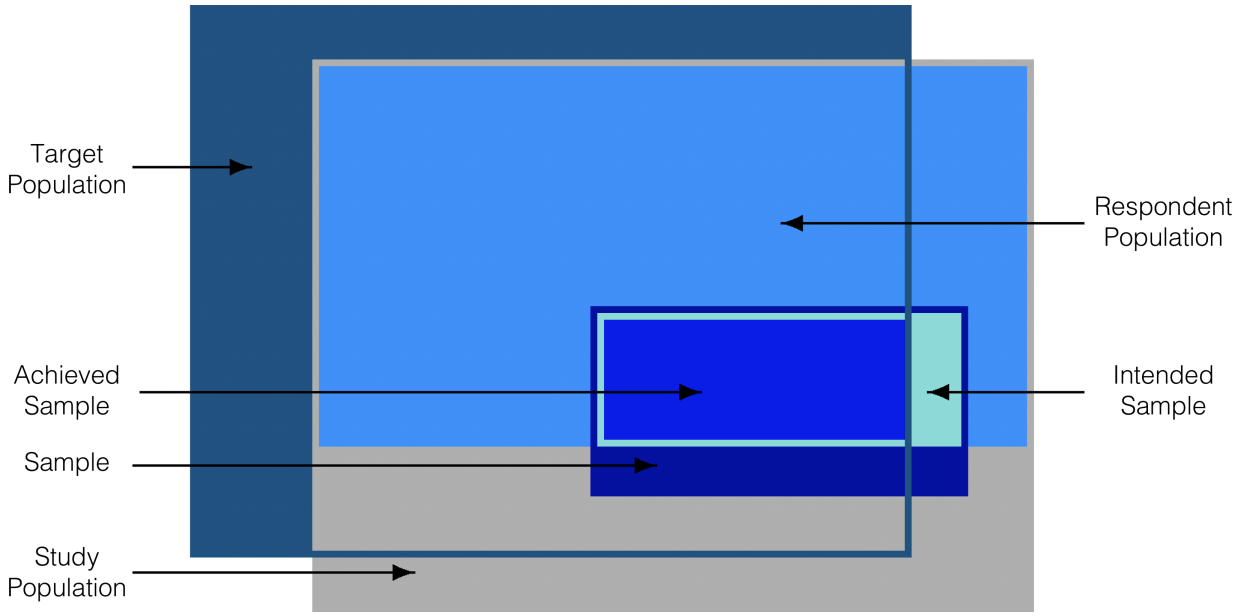


Figure 3: Various populations and samples in the sampling model.

Once a choice has been made, each survey typically follows the same **general steps**:

1. statement of objective
2. selection of survey frame
3. sampling design
4. questionnaire design
5. data collection
6. data capture and coding
7. data processing and imputation
8. estimation
9. data analysis
10. dissemination
11. documentation

The process is not always linear, in that preliminary planning and data collection may guide the implementation (selection of a frame and of a sampling design, questionnaire design), but there is a definite movement from objective to dissemination.

Survey Frames The **frame** provides the means of **identifying** and **contacting** the units of the study population. It is generally costly to create and to maintain (in fact, there are organisations and companies that specialise in building and/or selling such frames). Useful frames contain:

- identification data,
- contact data,
- classification data,
- maintenance data, and
- linkage data.

The ideal frame must minimize the risk of **undercoverage** or **overcoverage**, as well as the number of **duplications** and **misclassifications** (although some issues that arise can be fixed at the data processing stage).

Unless the selected frame is **relevant** (which is to say, it corresponds, and permits accessibility to, the target population), **accurate** (the information it contains is valid), **timely** (it is up-to-date), and **competitively priced**, the statistical sampling approach is contraindicated.

Survey Error One of the strengths of statistical sampling is in its ability to provide estimates of various quantities of interest in the target population, and to provide some control over the **total error** (TE) of the estimates. The TE of an estimate is the amount by which it differs from the true value for the target population:

$$\text{Total Error} = \text{Measurement Error} + \text{Sampling Error} + \text{Nonresponse Error} + \text{Coverage Error},$$

where the

- **coverage error** is due to differences in the study and target populations;
- **non-response error** is due to differences in the respondent and study populations;
- **sampling error** is due to differences in the achieved sample and the respondent population;
- **measurement error** is due to true value in the achieved sample not assessed correctly.

If we let

- $\overline{x}$ be the computed attribute value in the achieved sample;
- $\overline{x}_{\text{true}}$ be the true attribute value in the achieved sample under perfect measurement;
- x_{resp} be the attribute value in the respondent population;
- x_{study} be the attribute value in the study population, and
- x_{tar} be the attribute value in the target population,

then $\text{Total Error} = \overline{x} - x_{\text{tar}} = (\overline{x} - \overline{x}_{\text{true}}) + (\overline{x}_{\text{true}} - x_{\text{resp}}) + (x_{\text{resp}} - x_{\text{study}}) + (x_{\text{study}} - x_{\text{tar}})$.

In an ideal scenario, Total Error $= 0$. In practice, there are two main contributions to Total Error: **sampling errors** (which we will discuss shortly) and **nonsampling errors**, which include every contribution to survey error which is not due to the choice of sampling scheme. The latter can be controlled, to some extent:

- **coverage error** can be minimized by selecting a high quality, up-to-date survey frame;
- **non-response error** can be minimized by careful choice of the data collection mode and questionnaire design, and by using "call-backs" and "follow-ups";
- **measurement error** can be minimized by careful questionnaire design, pre-testing of the measurement apparatus, and cross-validation of answers.

In practice, these suggestions are perhaps less useful than one could hope in modern times: survey frames based on landline telephones are quickly becoming irrelevant in light of an increasingly large and younger population who eschew such phones, for instance, while response rates for surveys that are not mandated by law are surprisingly low. This explains, in part, the impetus towards automated data collection and the use of **non-probabilistic sampling** methods.

Modes of Data Collection How is data traditionally captured, then? There are **paper-based** approaches, **computer-assisted** approaches, and a suite of other modes.

- **Self-administered questionnaires** are used when the survey requires detailed information to allow the units to consult personal records (which reduces measurement errors), they are useful to measure responses to sensitive issues as they provide an extra layer of privacy, and are typically not as costly as other collection modes, but they tend to be associated with high non-response rate since there is less pressure to respond.
- **Interviewer-assisted questionnaires** use well-trained interviewers to increase the response rate and overall quality of the data.
  Face-to-face **personal interviews** achieve the highest response rates, but they are costly (both in training and in salaries). Furthermore, the interviewer may be required to visit any selected respondents many times before contact is established.
  Telephone interviews, on the other hand produce "reasonable" response rates at a reasonable cost and they are safer for the interviewers, but they are limited in length due to respondent phone fatigue. With random dialing, 4-6 minutes of the interviewer's time is spent in out-of-scope numbers for each completed interview.
- **computer-assisted interviews** combine data collection and data capture, which saves valuable time, but the drawback is that not every sampling unit may have access to a computer/data recorder (although this is becomine less prevalent).
  All paper-based modes have a computer-assisted equivalent: **computer-assisted self-interview** (CASI), **computer-assisted interview** (CAI), **computer-assisted telephone interview** (CATI), and **Computer-assisted personal interview** (CAPI).
- Unobtrusive direct observation
- Diaries to be filled (paper or electronic)
- Omnibus surveys
- Email, Internet, and social media (see Section 1.1.2)

Non-Probabilistic Sampling There exists a number of methods to select sampling units from the target population that use subjective, non-random approaches (NPS). These methods tend to be **quick**, **relatively inexpensive** and **convenient** in that a survey frame is not needed. NPS methods are ideal for **exploratory analysis** and **survey development**.

Unfortunately, they are sometimes used **instead** of probabilistic sampling designs, which is problematic; the associated selection bias makes NPS methods **unsound** when it comes to **inferences**, as they cannot be used to provide **reliable estimates of the sampling error** (the only component of Total Error on which the analysts has direct control). Automated data collection often fall squarely in the NPS camp, for instance. While we can still analyse data collected with a NPS approach, we **may not generalise the results** to the target population (except in rare, census-like situations).

NPS methods include

- **Haphazard** sampling, also known as 'man on the street' sampling; it assumes that the population is homogeneous, but the selection remains subject to interviewer biases and the availability of units;

- **Volunteer** sampling in which the respondents are self-selected; there is a large selection bias since the silent majority does not usually volunteer; this method is often imposed upon analysts due to ethical considerations; it is also used for focus groups or qualitative testing;
- **Judgement** sampling is based on the analysts' ideas of the target population composition and behaviour (sometimes using a prior study); the units are selected by population experts, but inaccurate preconceptions can introduce large biases in the study;
- **Quota** sampling is very common (and is used in exit polling to this day in spite of the infamous "Dewey Defeats Truman" debacle of 1948 [19]); sampling continues until a specific number of units have been selected for various sub-populations; it is preferable to other NPS methods because of inclusion of sub-populations, but it ignores non-response bias;
- **Modified** sampling starts out using probability sampling (more on this later), but turns to quota sampling in its last stage, in part as a reaction to high non-response rates;
- **Snowball** sampling asks sampled units to recruit other units among their acquaintances; this NPS approach may help locate hidden populations, but it biased in favour of units with larger social circles and units that are charming enough to convince their acquaintances to participate.

There are contexts where NPS methods might fit a client's need (and that remains their decision to make, ultimately), but the consultant MUST still inform the client of the drawbacks, and present some probabilistic alternatives.

Probabilistic Sampling The inability to make sound inferences in NPS contexts is a monumental strike against their use. While probabilistic sample designs are usually **more difficult and expensive** to set-up (due to the need for a quality survey frame), and take **longer** to complete, they provide **reliable estimates** for the attribute of interest and the sampling error, paving the way for small samples being used to draw inferences about larger target populations (in theory, at least; the non-sampling error components can still affect results and generalisation).

We shall take a deeper look at traditional probability sample designs such as **simple random**, **stratified random**, and **systematic**, – **cluster**, **probability proportional to size**, **replicated**, **multi-stage** and **multi-phase** variants also exist (see [1, 6] for details).

Let us start with some basic mathematical concepts. Consider a finite population $\mathscr{U} = \{u_1, \ldots, u_N\}$. The **mean** and **variance** of the population are given by

$$\mu = \frac{1}{N}\sum_{j=1}^{N} u_j \quad \text{and} \quad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j - \mu)^2, \quad \text{respectively.}$$

If $\mathscr{Y} = \{y_1, \ldots, y_n\}$ is a sample of $\mathscr{U}$, the **sample mean** and **sample variance** (also known as the **empirical mean** and **empirical variance**) are given by

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2, \quad \text{respectively.}$$

Let $X_1, \ldots, X_n$ be random variables, $b_1, \ldots, b_n \in \mathbb{R}$, and E, V, and Cov be the **expectation**, **variance** and **covariance** operators, respectively. Recall that

$$\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) &= \sum_{i=1}^{n} b_i \,\mathrm{E}(X_i) \\
\mathrm{V}\left(\sum_{i=1}^{n} b_i X_i\right) &= \sum_{i=1}^{n} b_i^2 \,\mathrm{V}(X_i) + \sum_{1 \le i \ne j}^{n} b_i b_j \,\mathrm{Cov}(X_i, X_j) \\
\mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j) \\
\mathrm{V}(X_i) &= \mathrm{Cov}(X_i, X_i) = \mathrm{E}\left(X_i^2\right) - \mathrm{E}^2(X_i).
\end{aligned}$$

The **bias** in an error component is the average of that error component if the survey is repeated many times independently under the same conditions. The **variability** in an error component is the extent to which that component would vary about its average value in the ideal scenario described above. The **mean square error** of an error component is a measure of the size of the error component:

$$\begin{aligned}
\mathrm{MSE}(\hat{\beta}) &= \mathrm{E}\left((\hat{\beta} - \beta)^2\right) = \mathrm{E}\left((\hat{\beta} - \mathrm{E}(\hat{\beta}) + \mathrm{E}(\hat{\beta}) - \beta)^2\right) \\
&= \mathrm{V}(\hat{\beta}) + \left(\mathrm{E}(\hat{\beta}) - \beta\right)^2 = \mathrm{V}(\hat{\beta}) + \mathrm{Bias}^2(\hat{\beta})
\end{aligned}$$

where $\hat{\beta}$ is an estimate of β. Incidentally, the unusual denominator in the sample variance insures that it is an unbiased estimator of the population variance.

Finally, if the estimate is unbiased, then an approximate **95% confidence interval** (95% CI) for β is given by

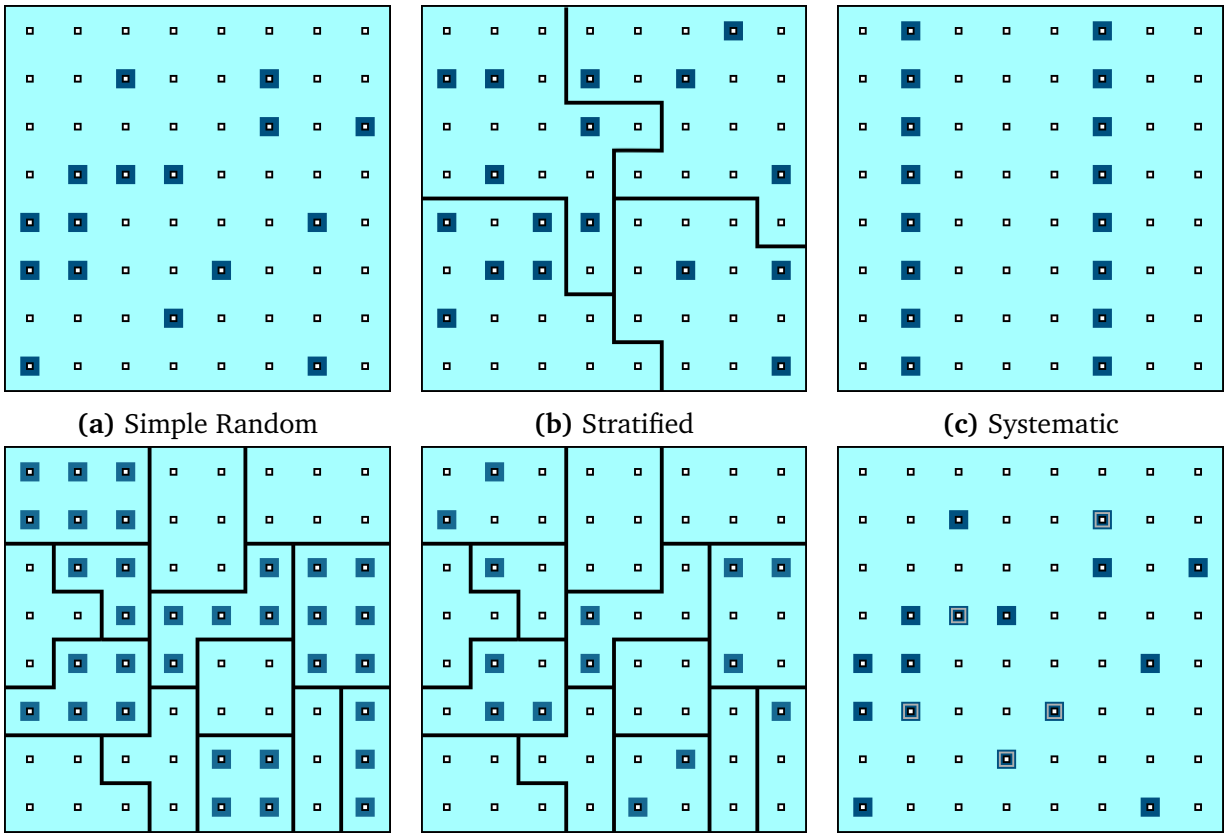
$$\hat{\beta} \pm 2\sqrt{\hat{\mathrm{V}}(\hat{\beta})},$$

where $\hat{\mathrm{V}}(\hat{\beta})$ is a sampling design-specific estimate of $\mathrm{V}(\hat{\beta})$.

In what follows, we discuss a number of sampling designs and present some of their advantages and disadvantages. We also show how to compute estimates for various population attributes (mean, total, proportion, ratio, difference, regression) and how to estimate the corresponding 95% CI. Finally, we briefly discuss how to compute sample sizes for a given **error bound** (an upper limit on the radius of the desired 95% CI), and how to determine the **sample allocation** (how many units to be sampled in various sub-population groups), for designs where it is appropriate to do so.

In all instances, the target population consists of N measurements/units, $\mathscr{U} = \{u_1, \ldots, u_N\}$, and the true population mean, population variance, population total, and population proportion for the variable of interest are μ, σ^2, τ, and p, respectively. The sample is a subset of the target population, $\mathscr{Y} = \{y_1, \ldots, y_n\} \subseteq \mathscr{U}$ from which we estimate the respective population attributes *via* $\overline{y}$, s^2, $\hat{\tau}$, and $\hat{p}$.

For a given characteristic, we define δ_i as 1 or 0 depending on whether the corresponding sample unit y_i possesses the characteristic in question or not. Lastly, we set the error bound to $B = 2\sqrt{\hat{V}} > 0$.



(a) Simple Random **(b)** Stratified **(c)** Systematic

(d) Cluster **(e)** Multi-Stage **(f)** Multi-Phase

Figure 4: Schematics of sampling designs.

In **Simple Random Sampling** (SRS), n units are selected randomly from the survey frame, as in Figure 4a. It is by far the easiest sampling design to implement, and estimates for the resulting sampling errors are well known and easy to compute, which leads to SRS often being used at a later stage in the sampling process. Another advantage is that SRS does not require auxiliary information, which can be useful with more economical survey frames.

This can backfire however, as SRS makes no use of such information even when it is available. There is also no guarantee that the sample will be representative of the population. Note as well that SRS may be costly if the sample is widely spread out, geographically.

The SRS estimators are

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\tau} = N\overline{y}, \quad \text{and} \quad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} \delta_i$$

with respective variances

$$\mathrm{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right), \quad \mathrm{V}(\hat{\tau}) = N^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right), \quad \text{and} \quad \mathrm{V}(\hat{p}) = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right).$$

The 95% CI is approximated by substituting the true variance σ^2 by the unbiased estimator $\frac{n-1}{n}s^2$:

$$\hat{\mathrm{V}}(\overline{y}) = \frac{s^2}{n}\left(1-\frac{n}{N}\right), \quad \hat{\mathrm{V}}(\hat{\tau}) = N^2 \cdot \frac{s^2}{n}\left(1-\frac{n}{N}\right), \quad \text{and} \quad \hat{\mathrm{V}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right).$$

Finally, the sample size required to achieve an upper error bound B are

$$n_{\overline{y}} = \frac{4N\sigma^2}{(N-1)B^2+4\sigma^2}, \quad n_{\hat{\tau}} = \frac{4N^3\sigma^2}{(N-1)B^2+4N^2\sigma^2}, \quad \text{and} \quad n_{\hat{p}} = \frac{4Np(1-p)}{(N-1)B^2+4p(1-p)},$$

where σ^2 and p have been previously estimated (perhaps as part of a prior survey).

In **Stratified Random Sampling** (StS), $n = n_1 + \cdots + n_k$ units are selected randomly from the survey frame by first establishing k natural strata (such as provinces, or age groups), and selecting n_j units from the N_j units in stratum j, with $\overline{y}_j$ and $\hat{p}_j$ the SRS estimators in stratum j, $j = 1, \ldots, k$. An illustration is provided in Figure 4b.

StS may produce a smaller bound on the error of estimation than would be produced by a SRS of the same size, particularly if measurements within a strata are homogeneous, and it may be less expensive to implement if the elements are stratified into convenient groupings. Another added benefits is that it may provide parameter estimates for sub-populations that coincide with the strata (see Section 1.1.4 for an application). There are no major disadvantage to this sample design, except for the fact that there might not be natural ways to stratify the frame (in the sense that each stratum might not be homogeneous in its units), in which case StS is roughly equivalent to SRS.

The StS estimators are

$$\overline{y}_{\mathrm{st}} = \sum_{j=1}^{k} \frac{N_j}{N}\overline{y}_j, \quad \hat{\tau}_{\mathrm{st}} = N\overline{y}_{\mathrm{st}}, \quad \text{and} \quad \hat{p}_{\mathrm{st}} = \sum_{j=1}^{k} \frac{N_j}{N}\hat{p}_j,$$

with approximate variances given by

$$\hat{\mathrm{V}}(\overline{y}_{\mathrm{st}}) = \frac{1}{N^2}\sum_{j=1}^{k} N_i^2 \hat{\mathrm{V}}\left(\overline{y}_j\right), \quad \hat{\mathrm{V}}(\hat{\tau}_{\mathrm{st}}) = N^2\hat{\mathrm{V}}(\overline{y}_{\mathrm{st}}), \quad \text{and} \quad \hat{\mathrm{V}}(\hat{p}_{\mathrm{st}}) = \frac{1}{N^2}\sum_{j=1}^{k} N_i^2 \hat{\mathrm{V}}\left(\overline{p}_j\right).$$

In the StS design, the sample determination question is two-fold: what size n should the sample have, and how should they be allocated to each stratum (n_j, $j = 1, \ldots, k$).

We can select n based on **cost** considerations or on error bound considerations. Let c_0 be the fixed survey operation costs (**overhead**), c_j be the **cost per response** in stratum j (which may need to include the cost of trying to reach non-respondents), and C be the **total cost** of conducting the survey. The sample size n that minimises $\hat{\mathrm{V}}(\overline{y}_{\mathrm{st}})$ subject to $C = c_0 + \sum_{j=1}^{k} c_j n_j$ and $n = \sum_{j=1}^{k} n_j$ is

$$n_{\mathrm{st},C} = (C - c_0)\frac{\sum_{j=1}^{k} \frac{N_j\sigma_j}{\sqrt{c_j}}}{\sum_{j=1}^{k} N_j\sigma_j\sqrt{c_j}}.$$

In the **general optimum allocation scheme**, the sampling weights (by strata) are

$$w_j = \frac{n_j}{n} = \frac{N_j \sigma_j c_j^{-1/2}}{\sum_{\ell=1}^{k} N_\ell \sigma_\ell c_\ell^{-1/2}}.$$

In the **Neyman allocation scheme**, we assume that the cost per response is identical in each stratum, whence

$$w_{j,N} = \frac{n_j}{n} = \frac{N_j \sigma_j}{\sum_{\ell=1}^{k} N_\ell \sigma_\ell},$$

while in the **proportional allocation scheme** we further assume that $\sigma_j = \sigma$ for all j, so that

$$w_{j,P} = \frac{n_j}{n} = \frac{N_j}{N}.$$

Other allocation schemes are also sometimes selected, such as the **square root proportional scheme** which fixes

$$w_{j,S} = \frac{N_j^{1/2}}{\sum_{\ell=1}^{k} N_\ell^{1/2}}$$

in order to insure that smaller strata (e.g. provinces with smaller populations, say) are allocated enough observations to produce sub-population estimates.

Note that while budgetary considerations need to be considered in practice, the preceding approach does not allow prescribed error bounds, which could prove problematic. The sample size required to achieve an upper error bound B are

$$n_{\text{st},\overline{y}} = \frac{4\sum_{j=1}^{k} \frac{N_j \sigma_j^2}{w_j}}{N^2 B^2 + 4\sum_{j=1}^{k} N_j \sigma_j^2}, \quad n_{\text{st},\hat{\tau}} = \frac{4N^2 \sum_{j=1}^{k} \frac{N_j \sigma_j^2}{w_j}}{N^2 B^2 + 4\sum_{j=1}^{k} N_j \sigma_j^2}, \quad \text{and} \quad n_{\text{st},\hat{p}} = \frac{4\sum_{j=1}^{k} \frac{N_j p_j (1-p_j)}{w_j}}{N^2 B^2 + 4\sum_{j=1}^{k} N_j p_j (1-p_j)},$$

where σ_j^2 and p_j have been previously estimated, and a specific allocation scheme $\{w_j\}$ has already been selected. In **Systematic Sampling** (SyS), n units are selected randomly from the survey frame by first (randomly) selecting a unit y_1 among the first $k = \lfloor \frac{N}{n} \rfloor$ units in the frame and s systematically adding every subsequent k^{th} unit to the sample. An illustration is provided in Figure 4c.

SyS is typically appropriate when the frame is already **sorted** along the characteristic of interest in which case it provides greater information by unit cost than SRS. It is simpler to implement than SRS since only one random number is required, and like SRS, it does not require auxiliary frame information. Depending on the sample size and on how the frame is sorted, SyS can produce a sample that is more widely spread (and thus perhaps more representative) than SRS, which may help eliminate other sources of bias.

On the other hand, it can introduce bias when the pattern used for the systematic sample coincides with a pattern in the population, and it makes no use of auxiliary frame information even if such information exists. Furthermore, any advantage in precision over SRS disappears if the frame is randomly ordered. Embarrassingly, SyS may lead to a variable sample size if n does not

evenly divide N; perhaps more importantly, SyS does not allow for an unbiased estimator of the sampling variance.

For all practical purposes, SyS behaves like SRS for a random population. In that case, the SRS variance formula may provide a decent approximation.

If the frame is **ordered** along the characteristic of interest, each SyS sample will contain some of the smaller values as well as some of the larger values, which would not necessarily be the case in a general SRS sample. This implies that the SyS estimators will have smaller variances than the corresponding SRS estimators, so that the use of the SRS variance formula produces an overestimate of the true sampling error in that case.

In a similar vein, a population is **periodic** if the frame is **periodic** along the characteristic of interest, a SyS sample that hits both the peaks and valleys of a cyclical trend will bring the method more in line with SRS and allow the use of the SRS variance formula as a reasonable approximation. To avoid the problem of underestimating the variation, consider changing the random starting point several times.

If n divides N evenly, then systematic sampling can be viewed as grouping the population into $k = N/n$ strata, and selecting one unit from each stratum. The difference between SyS and StS is that only the first unit is picked randomly in SyS – all other samples are automatically selected based on the position of the first choice.

One can also view SyS as a one-stage cluster sampling (see the next sub-section), where a primary sampling unit is defined as one of the $k = N/n$ possible systematic samples. An SRS of one unit can then be drawn from these k primary sampling units. The SyS sample will consist of all of the items in the selected primary sample.

The SyS estimators are computed exactly as the corresponding SRS estimators; their variances are given by

$$\mathrm{V}(\overline{y}_{\mathrm{sys}}) = \frac{\sigma^2}{n}[1+(n-1)\rho], \quad \mathrm{V}(\hat{\tau}_{\mathrm{sys}}) = N^2\,\mathrm{V}(\overline{y}_{\mathrm{sys}}), \quad \text{and} \quad \mathrm{V}(\hat{p}_{\mathrm{sys}}) = \frac{p(1-p)}{n}[1+(n-1)\rho],$$

where ρ is the **intra-cluster correlation** (which is typically impossible to compute exactly).

Other sampling schemes tend to be substantially more complicated (in the sense that the estimators and variance estimates are harder to derive), but the conceptual ideas behind those sampling schemes are still pretty straightforward; if required, in-depth details can be found in [6].

Cluster Sampling (ClS), for instance is typically used when the data collection cost increases with the "distance" separating the element. The population is separated in clusters, and an SRS of clusters is selected – all units within a selected clusters are retained in the sample (see Figure 4d for an illustration). As an example, to sample individuals in the population without a population frame (which might be hard to come by), it might be easier to obtain a dwelling frame and to start by sampling dwellings (which are the population **clusters**), and then to select all individuals in the sampled dwellings. ClS surveys are usually less expensive and less time-consuming to conduct than SRS, and they can be used to show "regional" variations, but they will be wasteful if the cluster sizes are too large, and biased if only a few clusters are sampled.

1.1.4 Case Study: Canada Vehicle Use Study

The **Canadian Vehicle Survey** (CVS) was sponsored by *Transport Canada* (TC) 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 duration ($\approx 20\%$), 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 due to mistakes in the *Vehicle Identification Number* (VIN) decoding code.

As a result, TC decided to conduct a pilot **Canadian Vehicle Use Study** (CVUS) to validate (or invalidate) the CVS methodology and results. Improvements included

- the use of **electronic data loggers** to reduce reporting burden;
- the introduction of a more **robust** VIN decoder to increase the accuracy of the in-scope fleet, and
- a **modified sampling design** that incorporated additional strata to enhance the ability to carry out more detailed analyses of motor vehicle use.

The pilot study was carried out in the 4th quarter of 2010 on $n = 1011$ light vehicles, selected via **simple random sampling** (SRS) from a list of vehicles registered with the *Ministry of Transportation of Ontario* (MTO) having an address whose *Forward Sortation Area* (FSA) code was associated with Ottawa and surrounding Ontario areas.

In order to evaluate the performance of the pilot CVUS, *vehicle-km traveled* (VKT) tallies were compared against corresponding CVS observations for the 4th quarter of 2009 ($n = 1016$).

The pilot CVUS was found to have a smaller number of observations at low VKT values than the CVS, whereas that trend was reversed at medium VKT values. The empirical means also seemed substantially different, at $\overline{x}_{\text{CVUS}} = 16,716$ km/year vs. $\overline{x}_{\text{CVS}} = 14,237$ km/year, although the high standard deviations $s_{\text{CVUS}} = 11,616$ km/year vs. $s_{\text{CVS}} = 13,844$ km/year made for inconclusive point comparisons.

Perhaps more importantly, the proportion of non-active vehicle in the fleet was much higher for 2009 in the CVS (8.7%) than it was for 2010 in the pilot CVUS (2.1%), and the distribution ranges are quite dissimilar: down to 79,500 km/year in 2010 from 112,500 km/year in 2009.

In any event, a **Kolmogorov-Smirnov test** rejected the null hypothesis that the two samples were drawn from the same distribution at the 99.9% significance level.

The CVS project management team steadfastly refused to update their survey in the face of this evidence, which gave TC the impetus to go ahead with a full-fledge CVUS survey.

We present an extract from a report entitled “Methodology of the Canadian Vehicle Use Study", it contains the following sections (more information on the CVUS can be found in [20]):

1. Objectives
..... 1.1 Canadian Vehicle Survey (CVS)
..... 1.2 Canadian Vehicle Use Study (CVUS)
7. Editing and Imputing
..... 7.1 Importing and Editing Data
..... 7.2 Creating Daily Summaries
..... 7.3 Rural/Urban Classification
..... 7.4 Basic and Derived Characteristics
..... 7.5 Vehicle Observations, Accuracy, Precision, and Measurement Error
8. Estimation and Data Analysis
..... 8.1 Vehicle Information at the Stratum Level
..... 8.2 Combining the Strata
Appendix A: Results for Ontario, Q1, 2012

References

[1] Farrell, P., *STAT 4502 Survey Sampling Course Package*, Fall 2008
[2] Lessler, J. and Kalsbeek, W. [1992], *Nonsampling Errors in Surveys*, Wiley, New York
[3] Oppenheim, N. [1992], *Questionnaire Design, Interviewing, and Attitude Measurement*, St. Martin’s Press
[4] Hidiroglou, M., Drew, J. and Gray, G. [1993], “A Framework for Measuring and Reducing non-response in Surveys,” *Survey Methodology*, v.19, n.1, pp.81-94
[5] Gower, A. [1994], “Questionnaire Design for Business Surveys,” *Survey Methodology*, v.20, n.2, pp.125-136
[6] *Survey Methods and Practices*, Statistics Canada, Catalogue no.12-587-X
[7] **Sir Humphrey’s Primer on Leading Questions**, *Yes, Prime Minister*, S01, E02, BBC, 1986.
[8] Munzert, S., Rubba, C., Meissner, P., Nyhuis, D. [2015], *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, Wiley
[9] Mitchell, R. [2015], *Web Scraping with Python*, O’Reilly.
[10] XPath introduction, **https://www.w3schools.com/xml/xpath_intro.asp**
[11] Wikipedia article on XML/HTML, **https://en.wikipedia.org/wiki/XHTML**
[12] Taracha, R. [2017], **Introduction to Web Scraping Using Selenium**.
[13] Selenium documentation, **https://pypi.python.org/pypi/selenium**
[14] Beautiful Soup documentation, **https://www.crummy.com/software/BeautifulSoup/bs4/doc**
[15] *Chrome* driver: **https://sites.google.com/a/chromium.org/chromedriver/downloads**
[16] *Edge* driver: **https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/**
[17] *Firefox* driver: **https://github.com/mozilla/geckodriver/releases**
[18] *Safari* driver: **https://webkit.org/blog/6900/webdriver-support-in-safari-10/**
[19] DeTurck’s, D., **Case Study 2: the 1948 Presidential Election**, retrieved on 12 July 2018.
[20] Allie, E. [2014], Canadian Vehicle Use Study: Electronic Data Collection, in *Proceedings of Statistics Canada Symposium 2014*, Beyond traditional survey taking: adapting to a changing world

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^i_{reg}=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 $b^i_{\min}$ (resp. $b^i_{\max}$) 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 $[b^i_{\min}, b^i_{\max}] \subseteq [0,1]$.[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^i_{first} (resp. w^i_{last}) 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

$$w^i_{\text{first}} = \frac{b^i_{\max} - \alpha_i}{b^i_{\max} - b^i_{\min}} \quad \text{and} \quad w^i_{\text{last}} = \frac{\omega_i - b^i_{\min}}{b^i_{\max} - b^i_{\min}}.$$

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

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

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 $b^i_{\max} > 1$: 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.

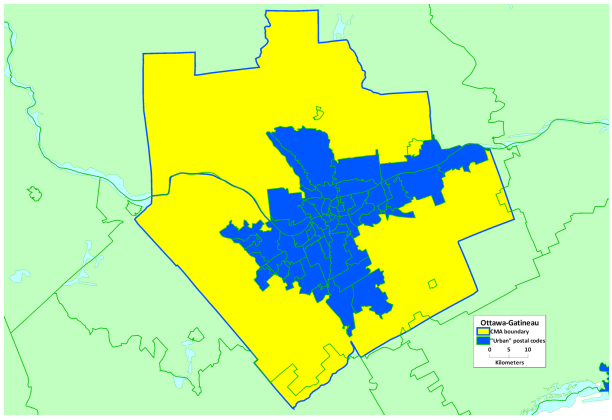
Furthermore, FSA codes may change fairly frequently, according to some arbitrary (at least, with respect to CVUS aims) internal logic at Canada Post. There is a chance that after such a change, a vehicle which would have been considered rural one day would suddenly be considered urban the next yet be 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 consump-tion 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 $(y_1, z_1, e_1^2), \ldots, (y_m, z_m, e_m^2)$, 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:

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

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:

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

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

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

In order to provide an estimate for $s_{\overline{Y}}^2$ (the sample variance of the mean $\overline{y}$ 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 $\overline{y}$ 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 $\overline{y}$ in the stratum is estimated by

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

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 $s^2_{\overline{Y}}$. 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 $s^2_{\overline{Y}}$ 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** $s^2_{\overline{X}_l} = s^2_{\overline{Y}}$ 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.

Ontario – 1st Quarter, 2012

Canadian VEHICLE USE Study — Powering Informed Decisions

Trip Characteristics

		Fleet Size	Sample Size	Average Number of Study Days	Average Number of Active Days	Daily Number of Trips	Daily Vehicle km Traveled	Daily Passenger km Traveled	Daily Fuel Consumption (L)	Daily Driving Time (h)	Fuel Consumption Ratio (L/100km)	Idling Ratio	Average Vehicle Occupancy	Average Speed (km/h)	Average Trip Length (km)	Average Trip Duration (min)
VEHICLE TYPE AND AGE (NRCan & TC)																
Ontario		**7,176,462**	**873**	**21.8**	**18.8**	**4.7** [a]	**46.9** [a]	**75.8** [a]	**5.4** [a]	**1.09** [a]	**11.3** [b]	**22.8%** [a]	**1.6** [a]	**42.9** [a]	**9.8** [a]	**13.7** [a]
	0 TO 8	4,538,722	571	22.1	19.1	5.1 [a]	50.2 [a]	82.3 [a]	5.7 [a]	1.17 [a]	11.3 [a]	22.8% [a]	1.7 [a]	43.0 [a]	9.9 [a]	13.8 [a]
	9+	2,637,740	302	21.4	18.3	4.3 [a]	41.2 [a]	64.6 [b]	4.8 [a]	0.96 [a]	11.4 [d]	22.8% [a]	1.6 [a]	42.7 [a]	9.6 [a]	13.5 [a]
	PRE '96	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
PASSENGER CAR		3,869,086	445	21.3	18.7	4.6 [a]	46.4 [a]	71.3 [a]	4.3 [a]	1.05 [a]	9.2 [a]	21.6% [a]	1.5 [a]	44.0 [a]	10.0 [a]	13.7 [a]
MINIVAN		823,659	116	21.7	19.1	5.2 [b]	48.7 [b]	97.8 [b]	6.8 [b]	1.24 [b]	13.4 [b]	25.6% [a]	2.0 [a]	39.3 [a]	9.4 [b]	14.4 [b]
PICKUP/CARGO		985,411	114	23.5	18.0	4.6 [b]	45.9 [b]	69.7 [b]	7.5 [b]	1.02 [b]	16.1 [f]	23.5% [a]	1.5 [a]	44.0 [a]	9.8 [b]	13.2 [a]
SUV		1,498,306	198	22.1	19.5	5.0 [a]	47.7 [b]	79.4 [b]	6.0 [a]	1.16 [a]	12.5 [a]	24.0% [a]	1.7 [a]	41.4 [a]	9.5 [a]	13.8 [a]
PASSENGER CAR	0 TO 8	2,321,197	279	21.6	19.1	4.8 [a]	49.5 [a]	77.9 [b]	4.5 [a]	1.12 [a]	9.0 [a]	21.6% [a]	1.6 [a]	44.3 [a]	10.3 [a]	13.9 [a]
	9+	1,547,889	166	21.0	18.0	4.3 [a]	41.7 [b]	61.5 [b]	4.0 [b]	0.95 [b]	9.4 [a]	21.6% [a]	1.5 [a]	43.7 [a]	9.7 [b]	13.4 [a]
	PRE '96	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
MINIVAN	0 TO 8	454,040	64	21.6	19.1	5.6 [b]	48.8 [c]	95.6 [b]	7.1 [c]	1.28 [c]	14.0 [a]	25.9% [a]	2.0 [a]	38.1 [a]	8.7 [b]	13.7 [b]
	9+	369,619	52	21.8	19.1	4.6 [b]	48.6 [c]	100.5 [d]	6.4 [b]	1.19 [b]	12.5 [c]	25.3% [b]	2.1 [b]	40.7 [b]	10.4 [c]	15.3 [b]
	PRE '96	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
PICKUP/CARGO	0 TO 8	610,279	72	25.5	19.0	5.2 [b]	55.0 [b]	87.6 [c]	8.9 [b]	1.20 [b]	16.1 [a]	22.5% [a]	1.6 [a]	45.4 [a]	10.5 [b]	13.8 [b]
	9+	375,132	42	20.2	16.3	3.6 [b]	31.1 [c]	40.7 [d]	5.3 [c]	0.75 [c]	16.2 [f]	25.2% [b]	1.3 [a]	41.8 [a]	8.6 [c]	12.4 [b]
	PRE '96	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
SUV	0 TO 8	1,153,206	156	21.6	19.1	5.1 [a]	49.5 [b]	83.4 [b]	6.0 [b]	1.20 [a]	12.1 [a]	24.3% [a]	1.7 [a]	41.2 [a]	9.6 [b]	13.9 [a]
	9+	345,100	42	23.8	20.8	4.5 [b]	42.1 [c]	65.9 [c]	5.9 [c]	1.00 [b]	13.8 [a]	22.7% [b]	1.6 [a]	41.9 [a]	9.4 [b]	13.4 [b]
	PRE '96	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
VEHICLE TYPE AND AGE (EC)																
Ontario		**7,176,462**	**873**	**21.8**	**18.8**	**4.7** [a]	**46.9** [a]	**75.8** [a]	**5.4** [a]	**1.09** [a]	**11.3** [b]	**22.8%** [a]	**1.6** [a]	**42.9** [a]	**9.8** [a]	**13.7** [a]
	0 TO 3	1,812,892	198	21.5	19.4	5.3 [a]	56.2 [b]	91.7 [b]	6.3 [a]	1.28 [a]	11.1 [a]	23.4% [a]	1.6 [a]	43.7 [a]	10.6 [a]	14.5 [a]
	4 TO 8	2,725,830	373	22.5	18.9	4.9 [a]	46.2 [a]	76.1 [b]	5.3 [a]	1.09 [a]	11.4 [a]	22.5% [a]	1.7 [a]	42.5 [a]	9.5 [a]	13.4 [a]
	9+	2,637,740	302	21.4	18.3	4.3 [a]	41.2 [a]	64.6 [b]	4.8 [a]	0.96 [a]	11.4 [d]	22.8% [a]	1.6 [a]	42.7 [a]	9.6 [a]	13.5 [a]
	OLD	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
	V.OLD	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
PASSENGER CAR		3,869,086	445	21.3	18.7	4.6 [a]	46.4 [a]	71.3 [a]	4.3 [a]	1.05 [a]	9.2 [a]	21.6% [a]	1.5 [a]	44.0 [a]	10.0 [a]	13.7 [a]
LIGHT TRUCK		3,307,376	428	22.4	19.0	4.9 [a]	47.4 [a]	81.1 [a]	6.6 [a]	1.14 [a]	13.8 [c]	24.2% [a]	1.7 [a]	41.7 [a]	9.6 [a]	13.8 [a]
PASSENGER CAR	0 TO 3	884,661	85	22.0	20.0	5.2 [a]	54.4 [b]	85.8 [b]	4.8 [b]	1.24 [b]	8.8 [a]	22.8% [a]	1.6 [a]	43.7 [a]	10.5 [b]	14.4 [a]
	4 TO 8	1,436,536	194	21.3	18.6	4.6 [a]	46.5 [b]	73.0 [b]	4.3 [b]	1.04 [a]	9.2 [a]	20.8% [a]	1.6 [a]	44.6 [a]	10.1 [b]	13.6 [a]
	9+	1,547,889	166	21.0	18.0	4.3 [a]	41.7 [b]	61.5 [b]	4.0 [b]	0.95 [b]	9.4 [a]	21.6% [a]	1.5 [a]	43.7 [a]	9.7 [b]	13.4 [a]
	OLD	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
	V.OLD	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
LIGHT TRUCK	0 TO 3	928,231	113	21.1	18.7	5.4 [a]	57.9 [b]	97.4 [b]	7.8 [b]	1.32 [b]	13.2 [a]	23.9% [a]	1.7 [a]	43.8 [a]	10.7 [b]	14.7 [a]
	4 TO 8	1,289,294	179	23.8	19.3	5.2 [a]	45.8 [b]	79.7 [b]	6.5 [b]	1.14 [b]	13.9 [a]	24.3% [a]	1.7 [a]	40.3 [a]	8.9 [a]	13.2 [a]
	9+	1,089,851	136	21.9	18.7	4.2 [a]	40.5 [b]	69.0 [c]	5.9 [b]	0.98 [b]	14.2 [f]	24.4% [a]	1.7 [a]	41.5 [a]	9.5 [b]	13.7 [a]
	OLD	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
	V.OLD	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
VEHICLE TYPE AND AGE																
Ontario		**7,176,462**	**873**	**21.8**	**18.8**	**4.7** [a]	**46.9** [a]	**75.8** [a]	**5.4** [a]	**1.09** [a]	**11.3** [b]	**22.8%** [a]	**1.6** [a]	**42.9** [a]	**9.8** [a]	**13.7** [a]
	0 TO 8	4,538,722	571	22.1	19.1	5.1 [a]	50.2 [a]	82.3 [a]	5.7 [a]	1.17 [a]	11.3 [a]	22.8% [a]	1.7 [a]	43.0 [a]	9.9 [a]	13.8 [a]
	9+	2,637,740	302	21.4	18.3	4.3 [a]	41.2 [a]	64.6 [b]	4.8 [a]	0.96 [a]	11.4 [d]	22.8% [a]	1.6 [a]	42.7 [a]	9.6 [a]	13.5 [a]
	PRE '96	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
PASSENGER CAR		3,869,086	445	21.3	18.7	4.6 [a]	46.4 [a]	71.3 [a]	4.3 [a]	1.05 [a]	9.2 [a]	21.6% [a]	1.5 [a]	44.0 [a]	10.0 [a]	13.7 [a]
LIGHT TRUCK		3,307,376	428	22.4	19.0	4.9 [a]	47.4 [a]	81.1 [a]	6.6 [a]	1.14 [a]	13.8 [c]	24.2% [a]	1.7 [a]	41.7 [a]	9.6 [a]	13.8 [a]
PASSENGER CAR	0 TO 8	2,321,197	279	21.6	19.1	4.8 [a]	49.5 [a]	77.9 [b]	4.5 [a]	1.12 [a]	9.0 [a]	21.6% [a]	1.6 [a]	44.3 [a]	10.3 [a]	13.9 [a]
	9+	1,547,889	166	21.0	18.0	4.3 [a]	41.7 [b]	61.5 [b]	4.0 [b]	0.95 [b]	9.4 [a]	21.6% [a]	1.5 [a]	43.7 [a]	9.7 [b]	13.4 [a]
	PRE '96	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]
LIGHT TRUCK	0 TO 8	2,217,525	292	22.7	19.1	5.3 [a]	50.9 [a]	87.1 [a]	7.0 [a]	1.22 [a]	13.6 [a]	24.2% [a]	1.7 [a]	41.7 [a]	9.6 [a]	13.8 [a]
	9+	1,089,851	136	21.9	18.7	4.2 [a]	40.5 [b]	69.0 [c]	5.9 [b]	0.98 [b]	14.2 [f]	24.4% [a]	1.7 [a]	41.5 [a]	9.5 [b]	13.7 [a]
	PRE '96	0	0	0.0	0.0	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.00 [f]	0.0 [f]	0.0% [f]	0.0 [f]	0.0 [f]	0.0 [f]	0.0 [f]

Ontario – 1st Quarter, 2012

Canadian VEHICLE USE Study — Powering Informed Decisions

Sub-Trip Characteristics

		Fleet Size	Sample Size	Average Number of Study Days	Average Number of Active Days	Daily Vehicle km Traveled	Daily Passenger km Traveled	Daily Fuel Consumption (L)	Daily Non-Idling Time (h)	Daily Idling Time (h)
VEHICLE SPEED	**Ontario**	**7,176,462**	**873**	**21.8**	**18.8**	**46.9** [a]	**75.8** [a]	**5.4** [a]	**0.84** [a]	**0.24** [a]
	IDLING					0.0 [a]	0.0 [a]	0.4 [a]	0.00 [c]	0.24 [a]
	1 km/h TO 24 km/h					2.1 [a]	3.4 [a]	0.6 [a]	0.18 [a]	0.00 [a]
	25 km/h TO 49 km/h					7.6 [a]	11.9 [a]	1.0 [a]	0.20 [a]	0.00 [a]
	50 km/h TO 79 km/h					15.2 [a]	24.0 [a]	1.4 [a]	0.24 [a]	0.00 [a]
	80 km/h TO 99 km/h					10.3 [a]	16.5 [a]	0.9 [a]	0.12 [a]	0.00 [a]
	100 km/h TO 119 km/h					9.7 [b]	16.7 [b]	0.9 [b]	0.09 [b]	0.00 [a]
	120+ km/h					2.0 [c]	3.4 [c]	0.2 [c]	0.02 [c]	0.00 [a]
IDLING TYPE	**Ontario**	**7,176,462**	**873**	**21.8**	**18.8**	**46.9** [a]	**75.8** [a]	**5.4** [a]	**0.84** [a]	**0.24** [a]
	NOT IDLING					46.9 [a]	75.8 [a]	5.0 [a]	0.84 [a]	0.00 [a]
	IDLING DURING TRIP					0.0 [a]	0.0 [a]	0.2 [a]	0.00 [a]	0.15 [a]
	TRIP START IDLING					0.0 [a]	0.0 [a]	0.1 [a]	0.00 [a]	0.07 [a]
	TRIP END IDLING					0.0 [a]	0.0 [a]	0.0 [a]	0.00 [a]	0.02 [b]
TIME OF DRIVING	**Ontario**	**7,176,462**	**873**	**21.8**	**18.8**	**46.9** [a]	**75.8** [a]	**5.4** [a]	**0.84** [a]	**0.24** [a]
	EARLY (06:00-08:59)					7.3 [a]	10.1 [a]	0.8 [a]	0.13 [a]	0.04 [a]
	MORNING (09:00-11:59)					7.1 [a]	11.8 [a]	0.9 [a]	0.13 [a]	0.04 [a]
	MIDDAY (12:00-14:59)					9.0 [a]	15.3 [a]	1.1 [a]	0.17 [a]	0.05 [a]
	AFTERNOON (15:00-17:59)					11.6 [a]	18.3 [a]	1.3 [a]	0.21 [a]	0.06 [a]
	EVENING (18:00-20:59)					7.1 [a]	12.3 [b]	0.8 [a]	0.12 [a]	0.03 [a]
	NIGHT (21:00-05:59)					4.8 [b]	8.0 [b]	0.5 [b]	0.08 [b]	0.02 [b]
ENGINE TEMP.	**Ontario**	**7,176,462**	**873**	**21.8**	**18.8**	**46.9** [a]	**75.8** [a]	**5.4** [a]	**0.84** [a]	**0.24** [a]
	COLD (< 50°C)					1.5 [a]	2.2 [a]	0.4 [a]	0.05 [a]	0.05 [a]
	WARM (50°C to 80°C)					7.2 [a]	11.0 [a]	1.0 [a]	0.16 [a]	0.06 [a]
	HOT (> 80°C)					38.0 [a]	62.6 [a]	4.0 [a]	0.63 [a]	0.13 [a]
	NO DATA					0.1 [f]	0.1 [f]	0.0 [f]	0.00 [f]	0.00 [f]

Quality of Estimates (cv)
a: less than 5% (excellent)
b: between 5% and 10% (good)
c: between 10% and 15% (acceptable)
d: between 15% and 20% (use with caution)
e: between 20% and 35% (unreliable)
f: more than 35% (unusable)

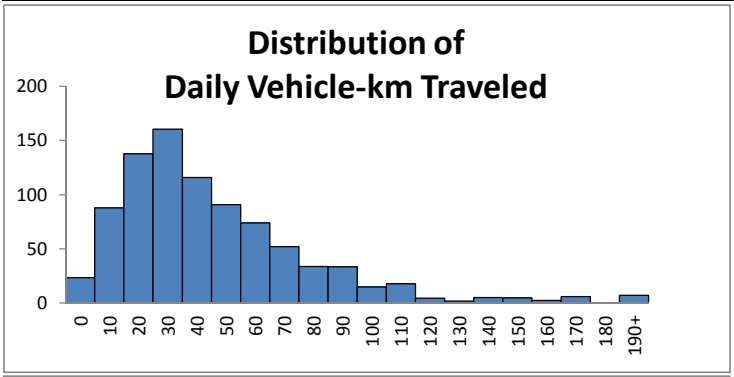
Vehicle Age
0 TO 3: 3 years old and younger
4 TO 8: between 4 and 8 years old
9+: 9 years old and older with model year post-1995
OLD: model year between 1981 and 1995
V.OLD: model year pre-1981

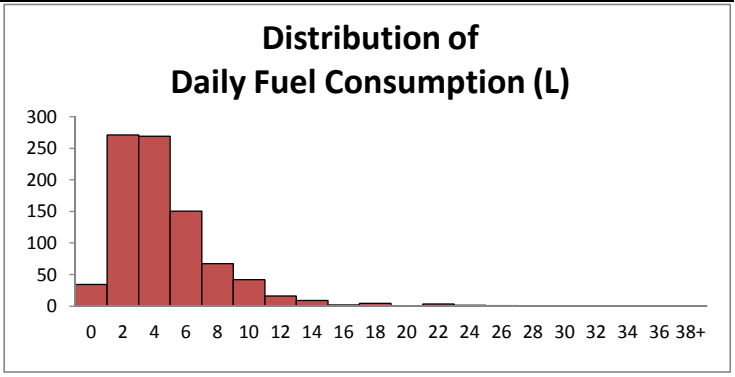
Notes on Driver Age and Gender
The estimates provided in the DRIVER AGE and GENDER category are VEHICLE characteristics, not DRIVER characteristics. Without further information on the distribution of drivers in a given jurisdiction (by AGE and GENDER), the estimates of the basic characteristics (nTrips, VKT, PKT, Fuel, Use, UseNI) cannot be used to predict the average driving behaviour of various combinations of DRIVER AGE and GENDER for that jurisdiction.

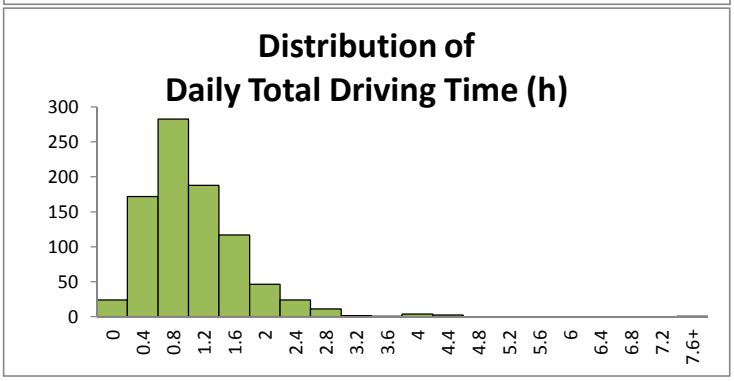
Values in columns may not add up or average (weighted) exactly to the corresponding column header due to to round off errors.

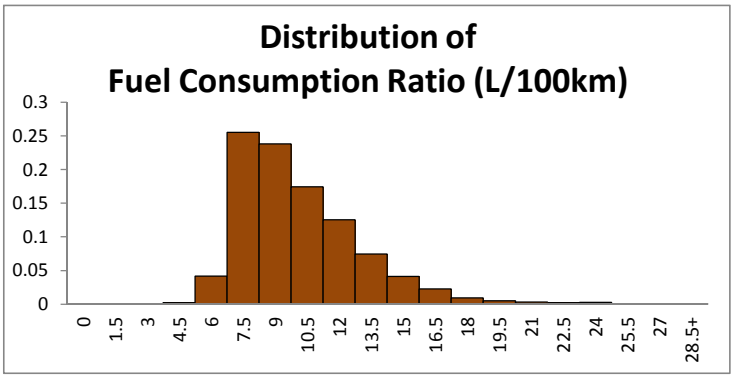
Ontario – 1st Quarter, 2012

Histograms





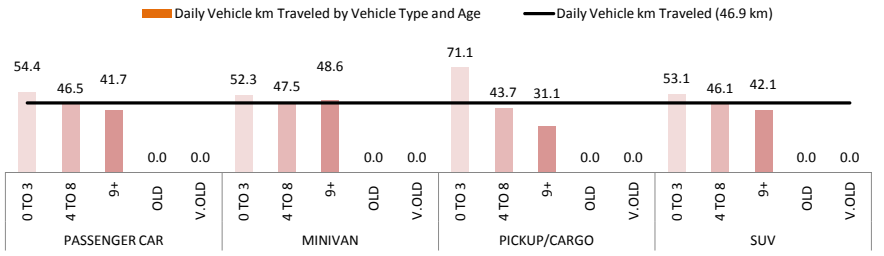




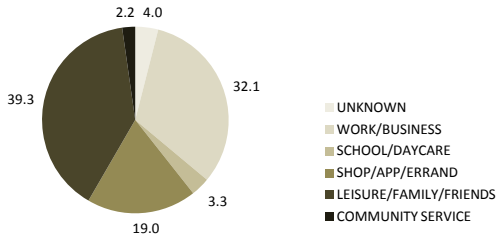
Ontario – 1st Quarter, 2012

Trip Characteristics

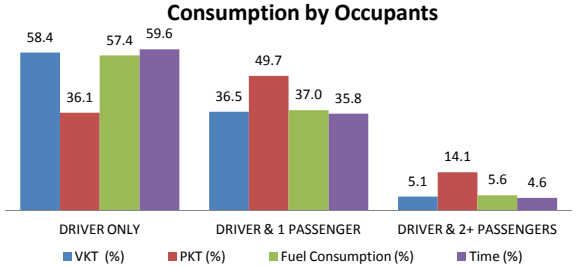
Daily Vehicle km Traveled by Vehicle Type and Age



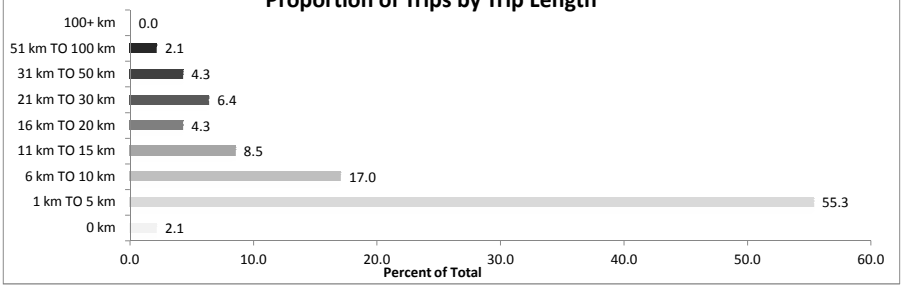
Passenger km Traveled by Purpose (%)



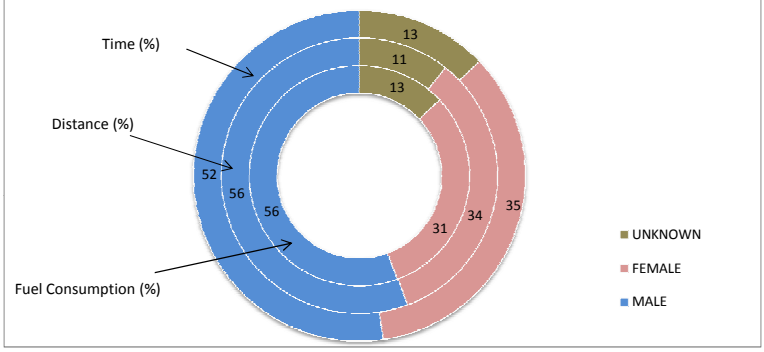
Distance, Passenger km Traveled, Time and Fuel Consumption by Occupants



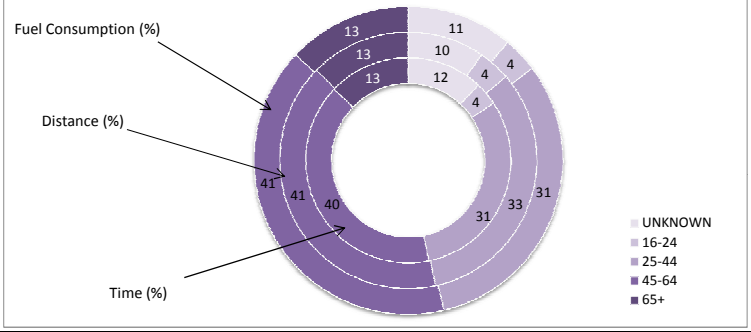
Proportion of Trips by Trip Length



Time, Distance and Fuel Consumption by Gender

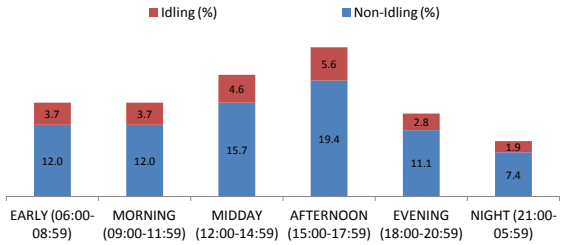


Fuel Consumption, Distance and Time by Age Category

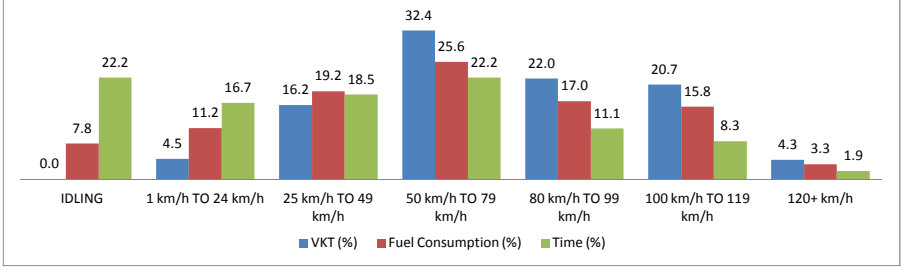


Sub-Trip Characteristics

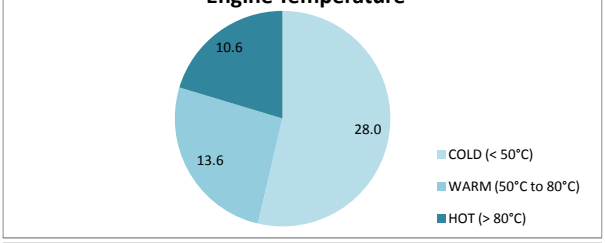
Driving Time by Idling and Non-Idling



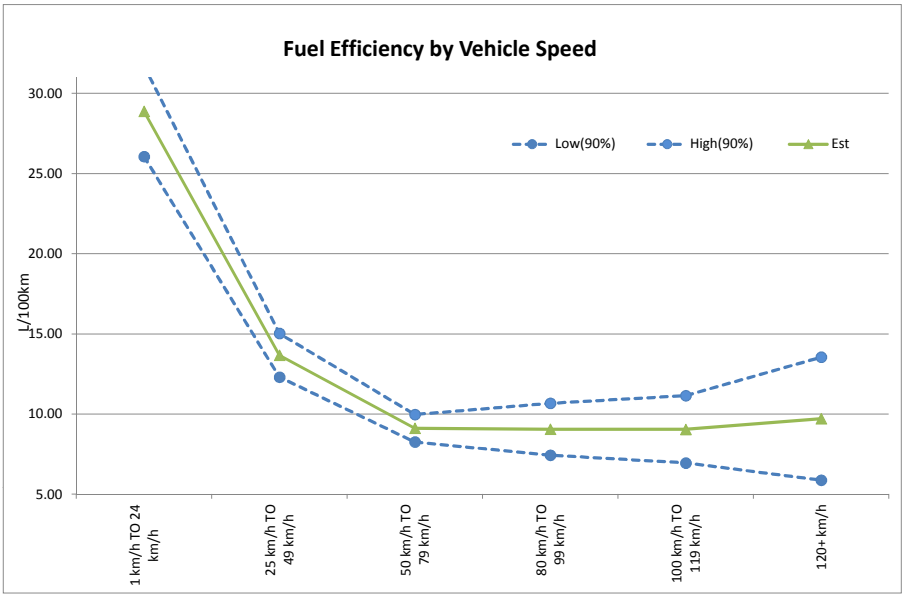
Distance, Fuel Consumption and Time by Speed



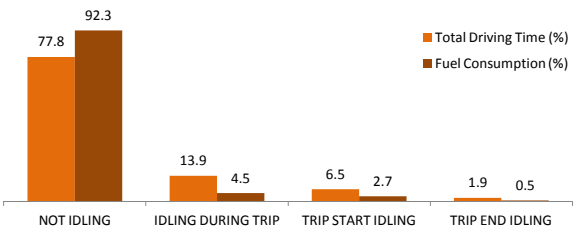
Fuel Efficiency (L/100km) by Engine Temperature



Fuel Efficiency by Vehicle Speed



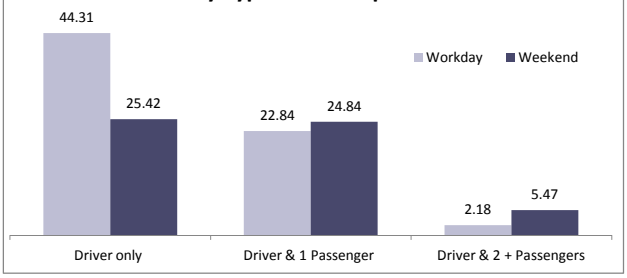
Time and Fuel Consumption by Idling Type



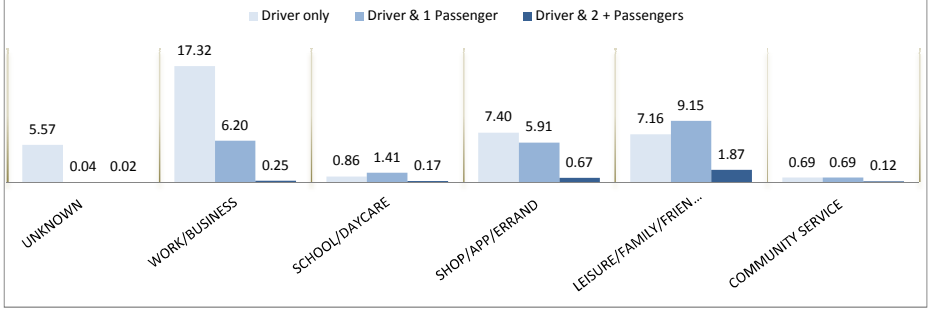
Ontario – 1st Quarter, 2012

Mixed Characteristics

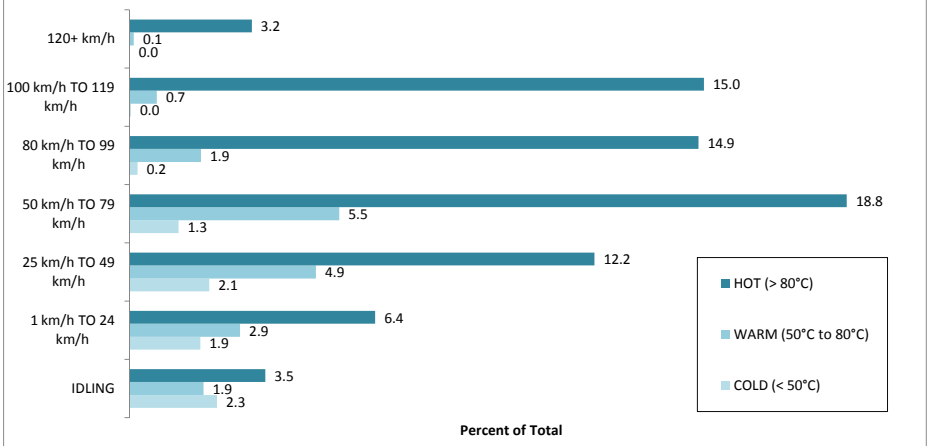
Driving Time (min) by Day Type and Occupants



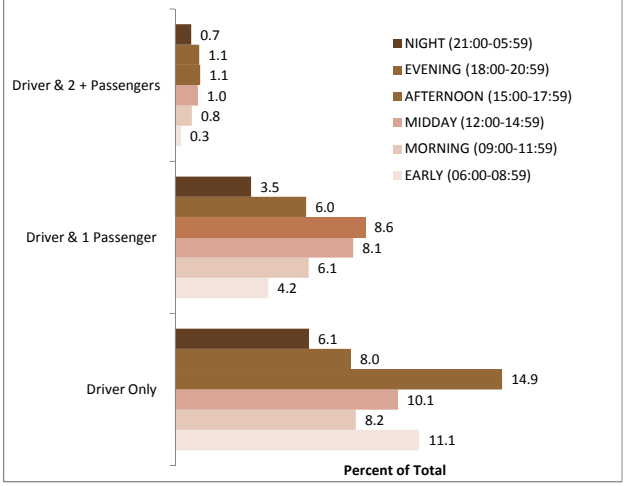
Time Driven (min) by Purpose and Occupants



Fuel Consumed by Speed and Engine Temperature



Distance by Occupants by Time of Driving



Fuel Consumption by Idling and Time of Driving