Data Science for Social Good 2018:
Transportation Vehicle Modelling for Policy Analysis - Final Report

November 20th, 2018

Research Fellows: **Nimrah Anwar**[1], **Andy Hong**[1], **Mia Kramer**[1], and **Kevin Wong**[1]
Project Leads: **Maxwell Sykes**[2] and **Aaron Licker**[3]

[1]Student at the University of British Columbia
[2]Climate and Energy Manager, City of Surrey, Sustainability Office
[3]Principal Consultant, Licker Geospatial Consulting Co.

Project Collaborators:

Project Mentors and Sponsors:

1. Introduction

1.1. Data Science for Social Good

The Data Science for Social Good (DSSG) Program at the University of British Columbia is a summer program designed to provide an interdisciplinary data science research experience for undergraduate and graduate UBC students. This program—hosted by the Data Science Institute at UBC in partnership with the eScience Institute, University of Washington—brings together students from diverse backgrounds to work closely with data providers and domain experts on focussed, collaborative projects that have the potential to benefit society.

1.2. Data Science for Transportation Emissions

Data science, that is to say novel combinations of divergent data sources, applications of data analysis and advanced methods of forecasting and reporting has been well applied in the area of transport and more specifically, transportation emissions modeling and forecasting. Through a considered application of data modeling, forecasting and visualization tools it is possible to answer critical questions with regards to the current state of transportation emissions and their potential forecast in the years to come. It is our hope that this project starts the conversation in this area.

1.3. Transportation Emissions and The City of Surrey

With the anticipated growth in population and increased activity in business sectors around Surrey, it is expected that transportation emissions will continue to grow. To reduce emissions growth and steadily decarbonize its transportation system, the City will need to take policy, planning, and program action to shift people from vehicles to transit and other modes, and to shift vehicles and fleets to electric and other low- and zero-emission transportation technologies.

To accomplish this, the City needs a greater understanding of its current and forecasted transportation system as well as a transportation energy and emissions model that can be used to test potential policies and develop a path to the City's 2050 GHG emissions reduction target. A better understanding derived from trends in transportation and vehicle registration data can enhance practices for climate planning, which includes land use, transportation, and infrastructure.

1.4. Project Goals and Objectives

The primary objective of this project is to develop and analyze a spatial passenger vehicle baseline and business-as-usual (BAU) forecast to 2050 that City staff can use for further modelling and policy analysis. This will be supported by:

- Context-specific research into the changing nature of transportation and passenger vehicle ownership in the City of Surrey (Vehicle Stock Insights); and
- The development of a functional tool that will allow for policy analysis and forecasting with regards to transportation energy demand and resulting greenhouse gas emissions (Policy Analysis Tools).

1.5. Project Deliverables

The primary deliverable for this project is the development of a tailpipe emissions forecasting model with structured inputs for emerging technologies, demographic drivers, local and senior government policies, and the ability to take transportation demand projections from the EMME model currently being developed by the City.

A secondary deliverable for this project is the development of a repeatable methodology for structuring and formatting vehicle registration data from BC's provincial automobile insurance authority (ICBC) into standard inputs for the primary deliverable

A final supporting deliverable for this project is an exploratory analysis of vehicle ownership trends in Surrey and a discussion of their validity for further modeling and analysis.

2. Methods and Approach

2.1. General Approach

The general approach to developing the critical deliverables for this project involved a five-step process which gradually refined and merged disparate datasets together into a final coherent product. Each step was executed by one primary

team member with the support of the remainder of team and generally followed a concurrent order. The steps are as follows:

(1) **Geographic Rebasing of data** – Preparing all spatial data points to be available at common geographic levels
(2) **Vehicle stock classification** – classifying vehicle registration data and appending tailpipe emissions profiles to each appropriate record in the data universe
(3) **Exploratory analysis** – visually and analytically inspecting the data to understand trends in the information and to develop a more in-depth understanding of vehicle stock profiles in the City
(4) **Vehicle stock forecasting** – building on the exploratory analysis to develop models that can with a reasonable degree of confidence predict future vehicle stocks
(5) **Tool development –** creation of a graphic interface that summarizes the information developed in the steps above

2.2. Summary of Results

2.2.1. Geographic Re-Basing

In transportation planning, Traffic Analysis Zone (TAZ) systems are often used as the base geographic unit of analysis. TAZ systems are typically drafted by transportation engineers and planners based on geographic divisions, the road system and the desired level of detail at each of the regions. TAZ systems are incompatible with any geographic system defined in the Canadian Census or the Canada Post Postal Code system, which are also incompatible with each another. Accordingly, our team was required to develop a method to re-base all of our geographic data such that a common unit of analysis could be used for all work moving forward. Ultimately, we selected the TAZ system of polygons and re-based the ICBC data as well as census data to these units.

2.2.2. Vehicle Stock Classification

Our team received three datasets that contained thousands of records from three eras (2006,2011,2016) of ICBC vehicle registration data for the City of Surrey. To protect privacy, vehicle data were anonymized by ICBC to the postal code level and all personal information was removed from the records. From these data, our team used a mixed-methods approach to vehicle stock classification such that every single vehicle in the data universe received a tailpipe emissions profile appropriate to the vehicles' make, model and year of construction. Overall, we were able to match 93% of vehicles to a known emissions profile from the US EPA. This emissions profile was then used in further analysis as discussed below.

2.2.3. Exploratory Analysis

To provide data-driven insights in City of Surrey's vehicle stock for climate and energy planning, we conducted exploratory analysis on three main attributes of our dataset: vehicle stock, vehicle age, and vehicle weight. Insights to where our passenger vehicle stock was increasing per capita, getting progressively older, getting heavier, and the converse, may reveal trends and identify spatial clusters vulnerable to climate policy. In our exploratory analysis, and to align with current transportation planning efforts underway in the City of Surrey, we aimed to geospatially visualize these attributes in each Traffic Analysis Zone (TAZ) unit within Surrey. As a secondary aim, analysis results may be used to verify assumptions for development of the GHG policy analysis tool.

Passenger vehicle stock was found to be increasing at a greater rate of change greater than commercial vehicle stock. Commercial vehicle stock also exhibited complex behaviors when normalized to total employment in each TAZ. Further analysis of commercial vehicles was discontinued in this report, but initial findings are discussed in Appendix 5.5.1. Green vehicle adoption, Hybrid and Electric vehicles, was observed to be early in its stages sharing at the highest 1.4% and 0.1% respectively of the 2016 vehicle registry. Lack of green vehicle data suggested re-focus on passenger vehicle stock analysis as defined in the initial project scope.

Through qualitative inspection, our results also identified clusters where vehicle ownership is, on average, increasing per capita, getting older, and getting heavier in the City of Surrey. Further analysis is required to contextualize the overlapping clusters and understand the interplay with various other demographic variables.

Finally, unique identifiers to each vehicle owner was identified as a critical data gap required to statistically model and rigorously test these hypotheses.

2.2.4. Forecasting

A key component of the study was to estimate future vehicle stocks in Surrey given the three data points of historic information. Using a business as usual forecasting approach our team developed and tested numerous models at various levels of fitness. Throughout the fitting process, a total of 23 class-based stock models at the community level were fitted. These models were developed based on 10 City-level and 78 community-level models predicting total vehicle counts of all classes, as well as 31 city-level models predicting vehicle counts per vehicle class. City-level models and total stock models are simple and provide valuable insights into which predictor variables should be considered for the final model, as well as the distribution that should be used in the regression. TAZ-based models were also attempted; however, these models displayed too much "noise" in comparison to the community-level models, with non-satisfactory model fit and diagnostics. Conversely, city-wide models are too broad in coverage and fitted with very little data (n = 3). Thus, we believe that a community level model should be adopted at the end. Ultimately our team selected a model with log-normal regression which appeared to be the most satisfying. Forecast information ultimately was fed into the visualization tool which is discussed below.

2.2.5. Tool Development

A critical aspect of data analysis is making the results usable, and actionable in the case of policy design. So, a tool that would allow policy makers in the City of Surrey to view information collected about the city—including information from ICBC that had never been released before—was deemed critical. In addition, viewing the results of different changes to vehicle stock on GHG emissions is an essential tool for the City of Surrey to reduce its GHG output by 50% relative to 2016 levels.

The design had two main functions:

1. Policy Testing: comparison of different vehicle stock targets.
2. Variable Visualization: visualization of a single variable (from census data, ICBC data, or City of Surrey Data) over the City of Surrey, with the basic geographic region being the TAZ.

Due to time constraints, the second tool could not be completed. In the current state, the framework for visualization of any of the variables is in place (but mostly untested), although the following likely useful features are missing:

- Automatic colour schemes for diverging variables
- Highlight tooltips with information about each TAZ

There are functions defined for measuring correlation, however due to the fact that the data was not available until the final day this could not be tested.

The first tool is largely complete. While there are usability improvements that could be made, its primary function works. The tool allows the user to create different "Policies" - sets of vehicle stock targets. There is an overall stock number, as well as percentage composition for several vehicle classes. The policies can be saved, deactivated, activated, and deleted. They have names and an area to store notes about the policy. The VKT model from the City of Surrey is used along with EPA fuel efficiency and emissions data is then used to produce a GHG output for each active policy, along with the outputs from the BAU model. All results are scaled relative to 2016 levels, which is normalized to 1.

3. Specific Project Components

3.0. Geographic Re-Basing

3.0.1. Introduction & Approach

In transportation planning, Traffic Analysis Zone (TAZ) systems are often used as the base geographic unit of analysis. TAZ systems are typically drafted by transportation engineers and planners based on geographic divisions, the road system and the desired level of detail at each of the regions. TAZ systems are incompatible with any geographic system defined in the Canadian Census or the Canada Post Postal Code system, which are also incompatible with each another. In the ICBC data provided to this team, each vehicle entry is tied to a postal code, and no other geographic identifiers are available. Census data, on the other hand, would be vital in the analysis to understand changes in the vehicle stock. Transportation data, such as travel demand and generalized travel costs, are usually supplied at the TAZ level. Without bringing these datasets to a common geographic unit, it is nearly impossible to carry out any further analysis.

In order to analyze vehicle stock by TAZ, a scheme is needed to convert postal code to TAZ. Ultimately, due to the lack of high quality, open-source postal code data, it was decided that each postal code would be approximated by a point before being classified into a specific TAZ. The geocoding of all postal codes contained within the ICBC data was carried out through Google Maps at first, then the postal codes that could not be geocoded by Google Maps were run through Geocoder.ca. Direct google search is the last resort in case both Google Mapes and Geocoder.ca could not identify a postal code. Perhaps not surprisingly, those mysterious postal codes all show up on realtor websites as housing units in newly developed communities.

Geocoding on Google Maps and Geocoder.ca is scripted in Python by making use of their Python APIs. Manual adjustments have to be made for some postal codes that lie near the border of the City of Surrey or at erroneous locations (such as at the ocean). Some postal codes were determined not to belong to the City of Surrey, and their corresponding vehicle entries were dropped from future analysis.

In summary, the steps to rebase postal codes are as follows:

1. Approximating each postal code as a point, geocode all postal codes using the Google Maps Geocoding API.
2. For postal codes that could not be geocoded by Google Maps, geocode them using Geocoder.ca. Afterwards, most postal codes would have been geocoded.
3. Google search is used to manually find the remaining postal codes. There are typically fewer than 5 postal codes in each year's dataset that require manual searching.
4. Manually check geocoding results in GIS software. Adjust points for postal codes that lie at the boundary of the city and remove postal codes that lie outside of city. Only very few postal codes exhibit these issues.

Re-basing census data to TAZ level is a much more challenging task. To begin with, each census dataset is enormous in size, and not all columns (also known as census vectors) need to be kept. As a result, census vectors need to be manually filtered based on the following criteria:

- That these census vectors are relevant to vehicle stock and transportation modelling in general
- That these census vectors are consistent across all years (unless they are new transportation census vectors that get added throughout the years)

Once the needed census vectors for each dataset are filtered, the corresponding data are extracted at the dissemination area (DA) level, which is the most refined geographic level of the census. Census data is extracted using the cancensus library in R. TAZ and DA form a "many-to-many" relationship, i.e.: one TAZ may correspond to multiple DA, and vice versa. Specific algorithms are available to interpolate and re-base geo-data, including the most readily available algorithm - areal interpolation (implemented by the "tongfen" package in R). However, they can only be applied to data which are of the type of summations. Any statistical or rate measures, such as median income and employment rate, would be calculated either through relevant summation data, or by computing the area-weighted averaging of DA data at each TAZ as a last resort. Furthermore, the area of each TAZ needs to be adjusted such that sparser and larger TAZ would not be erroneously distributed too much population and too many housing units. This is done by only extracting areas within 250-m buffers of Surrey's primary and secondary road network. Primary and secondary road network tend to indicate where people live and work, as opposed to highways and collectors. The 250-m buffering parameter is determined using trial-and-error. It minimizes the extract area in sparse TAZs while ensuring that all TAZs are represented by some areas.

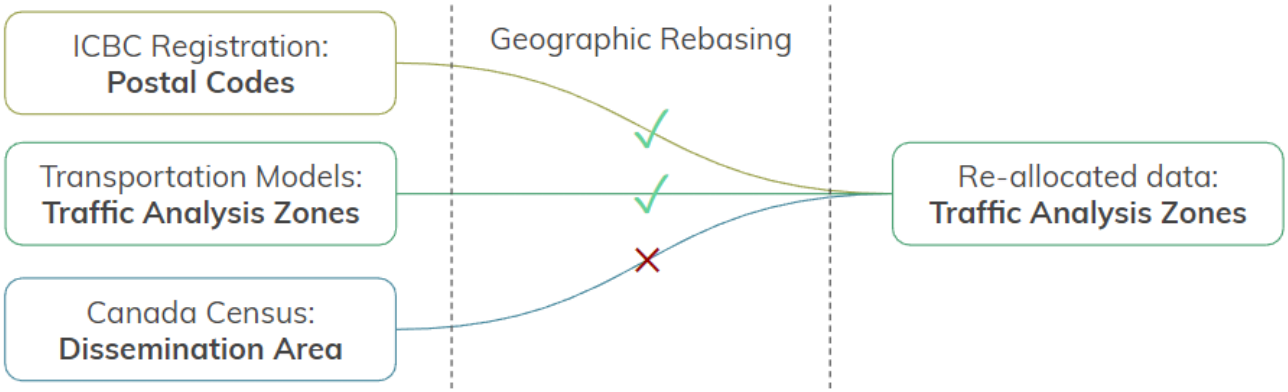


Figure 1-Rebasing - General Approach

After areal interpolation, summation-type variables are scaled to correct for underestimating Surrey's population and housing units by the census program (More details are given in the next section). Census summation-type variables are divided into population-variables (e.g.: number of people aged 65 and above) and housing-variables (e.g.: number of households that rent). These variables are then scaled by either the ratio of total population in Surrey estimates to total population in census or the same ratio but for housing units, correspondingly.

Now that TAZ-level census data are available at the original census data specification, the final step would be to standardize all census variables across all datasets. This requires renaming variables, carrying out arithmetic operations to variables that do not match completely across the years, and calculating mean, median and rate measures as mentioned in the last paragraph. In particular, the algorithm to calculate the median from grouped frequency table is provided online by Statistics Canada1. Because census data specifications changed over the years with many minor twists to variable definitions, extensive procedural code has been written to address all the special cases.

To summarize, the steps to rebase census data are as follows:

1. Manually filter census vectors that meet the criteria mentioned above for each census dataset
2. Extract areas within 250-m buffers of Surrey's primary and secondary road network for areal interpolation
3. Run areal interpolation (implemented by the "tongfen" package in R) on census summation-type variables with areas adjusted as in step 2. Estimate some census value-type variables using area-weighted averaging
4. Scale summation-type variables to match Surrey's own population and employment estimates.
5. Develop a standardized list of census variables with common variable names and definitions.
6. Rename original census variables that have direct equivalents under the standardized variable definitions and store them to the standardized dataset

7. Calculate most value-type variables (mean, median and rate) based on summation-type variables and store them to the standardized dataset
8. Extensive procedural coding to account for special-case variables within all census datasets that are different from the standardized variable definitions due to census specification changes

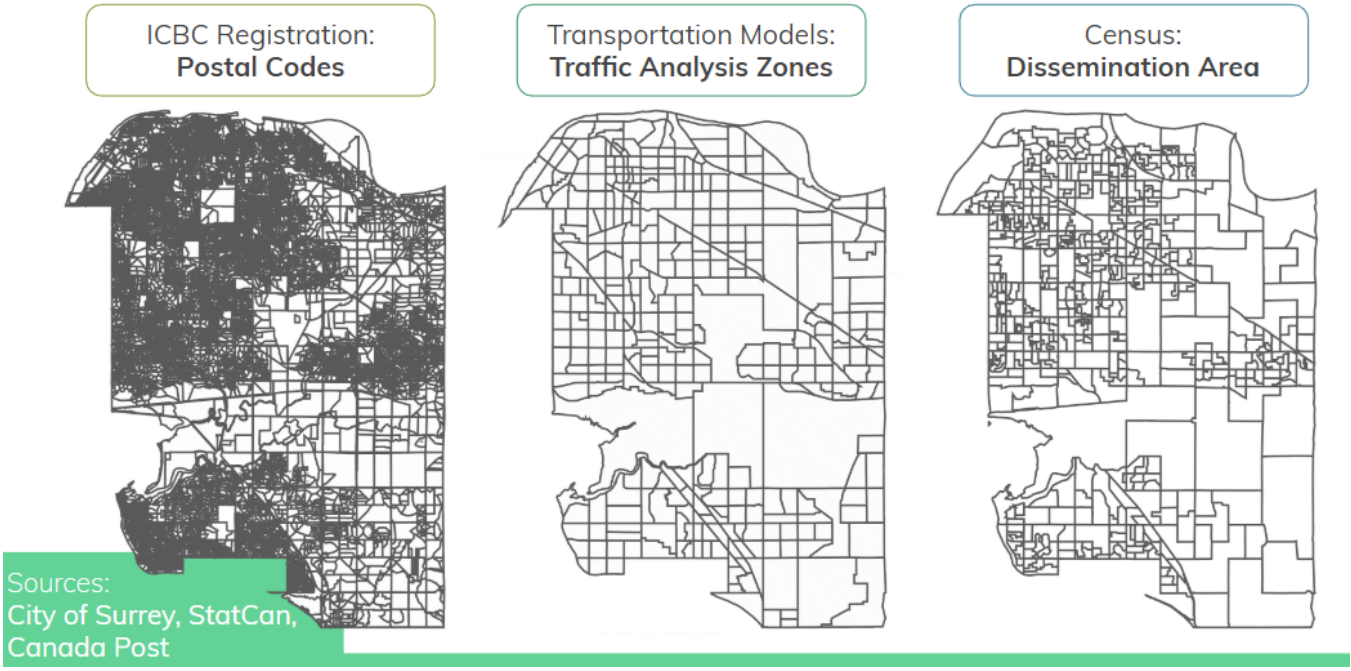


Figure 2-Data Sources for Rebasing

3.0.2. Data Sources Used

The Canadian Census Program is carried out every five years, with the most recent three census programs completed in 2006, 2011 and 2016. ICBC vehicle stock datasets have also been provided coinciding with those years. Transportation planning models typically have their scenarios aligned with census years as well. As a result, these years (2006, 2011, 2016) serve as reference years for analysis.

- *ICBC Vehicle Stock Data* - ICBC Vehicle Stock Data of cars registered in the City of Surrey in the years 2006, 2011 and 2016 have been kindly provided to by the City of Surrey, which have obtained these data through ICBC. The data contains a great deal of information; however, geographic identifiers of each vehicle are of the most concern in this section. Each vehicle is tied to their postal code and nothing else, and no unique identifier is given to each vehicle such that vehicles cannot be traced across the years.
- *Google Maps Geocoding API* - Google Maps Geocoding API is used to geocode most of the postal codes, thereby making use Google Maps data.
- *Geocoder.ca API* -Geocoder.ca API is used to geocode the remaining postal not geocoded by Google Maps, thereby making use Geocoder.ca data.
- *Other Website with Postal Code Data* -As a last resort, Google search is used to determine the location of postal codes that are neither stored in Google Maps database nor Geocoder.ca database. These postal codes, which are typically located in newly developed neighbourhoods as of 2016-2018, are all identified through house listings on realtor websites.
- *Canada Census Program Data* -Data of the Canadian Census program in 2006, 2011 and 2016, collected and compiled by Statistics Canada are gathered using the cancensus library in R.
- *Other Statistics Canada Data* -Besides Canadian Census data, annual average CPI data for the Greater Vancouver region in 2005, 2010 and 2015 (for which individual and household income data were collected)

- *Surrey Open Data - Road Network Profile* -Road network file of City of Surrey, available for free on Surrey Open Data, is used to adjust area of TAZs to account for sparser TAZs unfairly receiving higher population and housing unites in areal interpolation.

3.0.3. Sources of Error

Multiple sources of error can be identified, either within the datasets or due to operations carried out on these datasets.

- The Surrey vehicle stock dataset is administered and processed by staff at ICBC. There could well be errors when these data are recorded or filtered for Surrey only vehicles.
- Estimating postal code catchment area with the centroid of a point may lead to some vehicles being distributed to surrounding TAZs, for postal codes with catchment areas that cross multiple TAZs.
- It is assumed that postal code catchment areas do not change over-time, thereby allowing the use of current Google Maps / Geocoder.ca to geocode postal code from 2006 and 2011 datasets. It should be reasonable to assume that no address has its postal codes changed overtime; however, the development and densification of an area may lead to new postal codes being created, thus shrinking the catchment area of pre-existing codes.
- Google Maps and Geocoder data may not be fully accurate, particular for postal codes that correspond to sparser areas or new development in the City of Surrey.
- While road network is a decent representation of population distribution, some areas (particularly Agricultural Land Reserves) may contain roads where nobody live or work nearby.
- Values reported by the Canadian census may also contain inaccuracy. In particular, population of City of Surrey may be underestimated in recent census programs due to the vast availability of "hidden housing" (e.g.: laneway houses, basement suites).
- The entire process of converting DA-based data to TAZ-based data uses relatively simple techniques (areal interpolation, population / housing unit-scaling, area-weighted averages), which may reduce accuracy of results after the transformation.

3.0.4. Findings

Out of 9819 unique postal codes that exist in 2006, 2011 and 2016 ICBC dataset, 9486 of them are geolocated in Google Maps, 326 are geolocated in Geocoder.ca, 4 points have to be manually searched through Google, and 2 have to be manually moved as a result of obvious errors in Google Maps / Geocoder.ca data (point at ocean, for instance). 5 postal codes have to be removed since they do not physically lie within the City of Surrey.

As for census data, after all the steps described in section 3.0.0 relevant to census processing are carried out, three standardized census table are developed. Each table contains 369 variables that are common across all three time points (2006, 2011 and 2016). Note that the tables may contain cells with NA in one or more of the following circumstances:

- If the data are missing in the original census data extract obtained from the cancensus library.
- If abnormal values of value-type data have been obtained in divisions or median calculations due to irregularities the data, particularly after areal transformation. For instance, after areal interpolation, a sparsely-populated TAZ may be estimated to have 0.2 people living in households and 0.002 households, leading to an estimate of average household size of 100 people. To cope with such issues, maximum and minimum limits have been established for these kinds of variables based on perceived reasonableness, and any value-type estimates that lie outside of the range of expected values are assigned as NA.
- If a standardized variable cannot be easily computed based on data from one original census dataset. For instance, for the standardized census variable "households with 4 people", an exact match of variables exists for the 2006 and 2016 census table. However, the closest corresponding vector in the 2011 census

table is “households with 4 to 5 people”, and the former cannot be easily computed from the latter. In such cases, the variables are left as NA.

- If a standardized variable is relevant to transportation and have been only added in recent census (i.e.: 2011, 2016). Despite missing some timepoints, one may still want to make use of the data available for cross-sectional insights. Thus, the data for those years that do not contain these relatively new transportation variables would contain NAs.

3.0.5. Validation

Validation of the geo re-basing results is primarily carried out through data visualization and visual inspections.

As for geographic re-basing of postal code field of the ICBC data, validation is carried out through creating spatial and other visualizations of Surrey vehicle stock. Those visualizations can be found in section 3.1. In summary, the resulting visualizations show very reasonable patterns, especially when comparing to the population and housing units’ visualizations created using data estimated by the City of Surrey.

As for census data, histograms have been created for certain census variables (such as median household income and employment rate) to check that results are reasonable when these values are being computed during the geographic re-basing process. Preliminary visualization work that is similar to the work done for the ICBC data has been carried out for a few census variables. Due to time constraints and little applications of the census data, no extensive validation work has been carried out on the final dataset. More work should be carried out to check the results of this part.

3.0.6. Next Steps

Additional validation should be carried out for the standardized census data to ensure its integrity and soundness before using it for further analysis. In addition, algorithms that improve accuracy of geographic re-basing of census data (rather than using relatively simple approach including areal interpolation and area-weighted average) should be further investigated.

3.1. Vehicle Stock Classification

3.1.1. Introduction & Approach

The United Nations Intergovernmental Panel for Climate Change lays down the methodology to approximating greenhouse gas emissions (GHGs) from cities. The two-step approach first aims to approximate the total volume of a different fuels consumed within the city by the transportation sector. In the second step, the fuel volume is converted to GHG emissions by multiplying with **emission factors**. To calculate fuel volume consumed by vehicles, we need three parameters:

1. Vehicle Stock of the City
2. Fuel Consumption of each vehicle
3. Kilometers travelled by each vehicle (VKT)

One of the key deliverables of this project, was to predict the future GHG emissions so that effective policies may be designed to reduce future mobile combustion related emissions. However, it is not feasible to model current and future GHG emissions with respect to each of different types of vehicles driven in Surrey. Therefore, we decided to adopt a classification scheme that simplified the process of building a robust GHG predictive emission model.

The Vehicle Stock Classification adopted for this project was taken from FuelEconomy.gov which is a website maintained by the U.S. Department of Energy using the data provided by the U.S. Environmental Protection Agency (EPA). Fuel Economy aims to provide accurate fuel consumption information to consumers.

The reason why this particular classification scheme was adopted was that it offered a fairly detailed classification that fit the requirements set by the City of Surrey. Additionally, we were able to source all the fuel

consumption data required for GHG emissions from the same source. Since much of the regulations regarding vehicle emissions set in Canada are already adopted from EPA, adopting a classification from them helps with standardization and streamlining our tool with other tools offered by EPA.

The final classification scheme is shown in Figure 1 below:

Cars		
Class	**Passenger & Cargo Volume (cu. ft)**	
Two Seaters	Any	
Sedans		
Mini-compact	< 85	
Subcompact	85 - 99	
Compact	100 - 109	
Midsize	110 - 119	
Large	120 - more	
Station Wagons		
Small	< 130	
Mid-size	130 – 159	
Large	160 or more	
Trucks		
Pickup Trucks	Through MY[1] 2007	As of MY[1] 2008
Standard	4,500 to 8,500 lbs.	6,000 to 8,500 lbs.
Vans	Through MY[1] 2010	As of MY[1] 2011
Passenger	<8,500 lbs.	<10,000 lbs.
Cargo	<8,500 lbs.	
Minivans	<8,500 lbs.	
SUVs	Through MY[1] 2010	As of MY[1] 2011
	<8,500 lbs.	<10,000 lbs.
Special Purpose Vehicles	Through MY[1] 2010	As of MY[1] 2011
	<8,500 lbs.	<8,500 lbs. or <10,000 lbs., depending on configuration

Figure 3. Table showing final Vehicle Stock Classification Scheme.

NOTE: 1 MY = Model Year

Fueleconomy.gov. (2018). *Fuel Economy Web Services*. [online] Available at: https://www.fueleconomy.gov/feg/ws/index.shtml#emissions

We decided some of the classes were too small for our purposes, and so they were combined in the following manner:

- Small Pickups and Standard Pickups -> Pickups
- Passenger Vans and Cargo Vans -> Vans
- Small, Midsize and Large Station wagons -> Station wagons

This bought the total number of vehicle classes carried forward to **12** that would later assist with our exploratory and regression analysis.

To classify Surrey's vehicle stock, we developed a process that matched car names based on make (brand), model name and model year. The vehicle stock was matched to the cars whose data was made available by the EPA. The flowchart shown in Figure 2 below gives a simplified process flow of how the two datasets were matched and the resulting outputs.

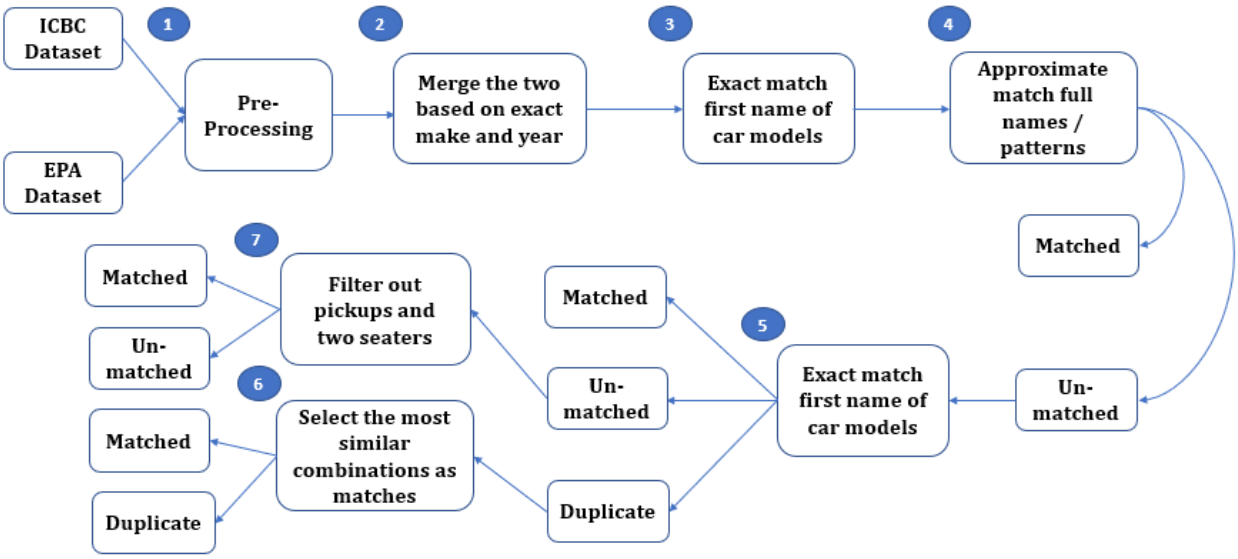


Figure 4. Flowchart showing the simplified process for matching datasets and the resultant outputs.

- **Step 1** – Both the datasets were pre-processed to remove spaces and symbols. The ICBC dataset was filtered to include only passenger vehicles and commercial – pickup trucks (explained in further detail in section 3.1.2). Make (brand) and model names were also standardized at this stage.

- **Step 2** - The two datasets were merged on the **exact** make and model year of the car. Since EPA has data for cars starting from 1984, cars older than this could not be matched. Additionally, some of the cars not matched here due to misspelled make names or unavailable data.

- **Step 3** – The first names of model names were matched to reduce the size of the merged dataset and improve the performance of the script.

- **Step 4** - The merged data contained all the potential matches for each car based on make and model year. To then match model names, approximate string matching was used where shorter strings (patterns) were matched with longer strings for each row. However, this approach only was not entirely effective as a significant amount of data at this stage remained unmatched.

- **Step 5** - The problem of unmatched data was largely due model naming inconsistencies between ICBC and EPA. To address this, only the first name (or part) of car model were matched and this significantly increased the proportion of matched cars. However, this also introduced the problem of duplicates. E.g. A Ford **Explorer Sport Trac XLT** is matched to Ford **Explorer Sport Trac** and Ford **Explorer Sport**.

- **Step 6** – To address the problem of duplicates as mentioned in step 5, we used the longest common substring distance (LCS) to pick the most similar matches while disregarding the other matches. E.g. In the above example, Ford **Explorer Sport Trac XLT** would be matched to Ford **Explorer Sport Trac** (as it is most similar) where as Ford **Explorer Sport** would be discarded.

- **Step 7** - To reduce the amount of unmatched data, ICBC's own classification was used to classify some of the unmatched vehicles with a reasonable degree of confidence. E.g. EPA classifies all two-seater cars in one class regardless of the volume. Therefore, all two-seater cars were filtered out in the ICBC stock and were classified as "Two Seaters". Then, they were assigned average fuel efficiencies of matched two-seater vehicles. Similar process was applied to pickup trucks.

The resulting schema for vehicle classification is presented below in the following figures:

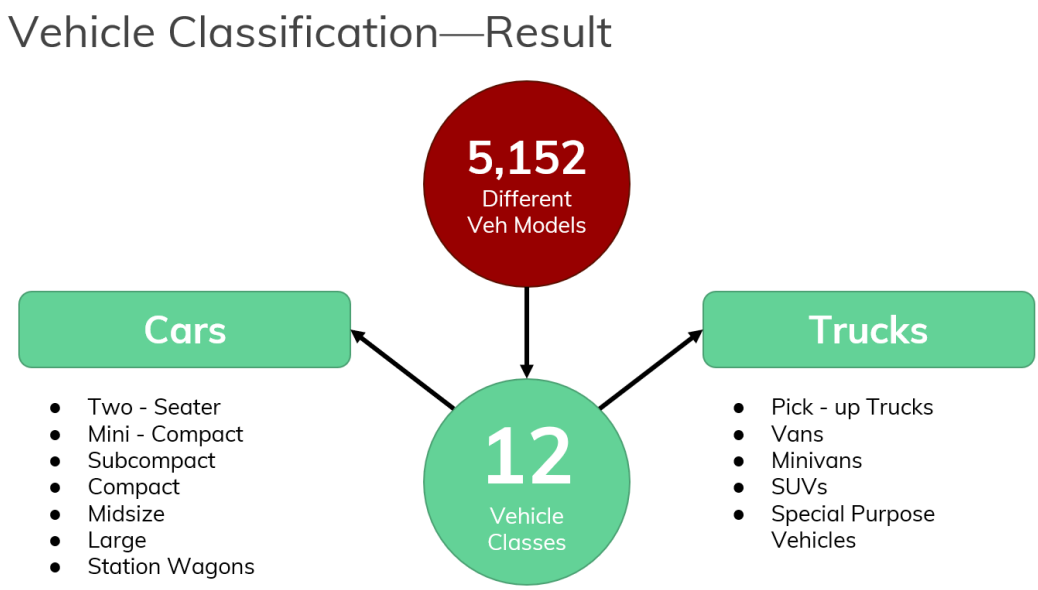


Figure 5- ultimate vehicle classification schema

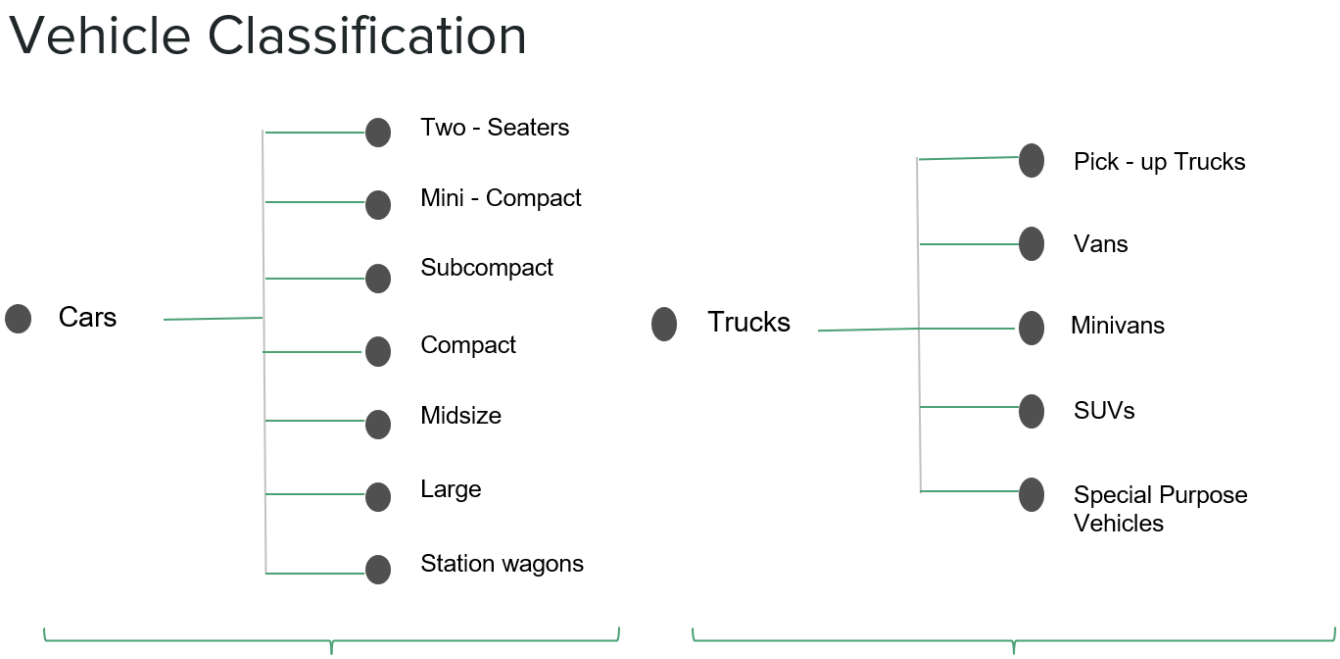


Figure 6-Vehicle classification

3.1.2. Data Sources Used

The vehicle stock in the city of Surrey was provided to by ICBC for the years 2006, 2011 and 2016. The classification and the resulting fuel consumption data was taken from FuelEconomy.gov. The emission factors used to convert the volume of fuel consumed to GHG emissions were taken from the yearly National Inventory Report published by Environment Canada.

Due to the findings of the exploratory analysis, it was initially decided that the project would only focus on passenger vehicles, and accordingly, the ICBC vehicle stock was filtered to include just types of vehicle. Later in the process we discovered that all pickup trucks are classified as commercial vehicles by ICBC. Since pickup trucks make up a significant portion of Surrey's vehicle stock, it was decided to include commercial – pickup trucks in our analysis.

3.1.3. Sources of Error

The main source of error was due to mainly model name inconsistencies between the ICBC Stock dataset and the EPA dataset. ICBC often uses abbreviations for model names e.g. a "Dodge Grand Caravan" is listed as a "CARA". As one can imagine, this caused significant problems when it came to string matching. ICBC does not officially define these abbreviations but they seem to be a general practice among ICBC brokers.

To solve this issue (to some extent), a list of the most common unmatched cars was developed to find model names that had been abbreviated and corrected them. This helped to significantly increase the number of matched vehicles but unfortunately, because of time limitations, it was not possible to locate and rename all abbreviated model names in the ICBC stock.

Additionally, we had some duplicate cars in our final output as we lack enough information to classify them. E.g. A 1991 BMW 525I Touring is matched to what looks like two identical cars that belong to different classes (see below).

MAKE	MODEL_YEAR	MODEL	EPA_MODEL	CLASS
BMW	1991	525I 525I TOURING	525I	Compact Cars
BMW	1991	525I 525I TOURING	525I	Midsize Cars

In the EPA dataset, one of the 525I is classified as a compact car whereas the other is classified as midsize. This is because BMW released an automatic and manual transmission for the 1991 BMW 525I that differ in cargo volumes and so belong to different classes. Now without knowing the transmission of the car in ICBC vehicle stock (or its cargo volume), it is not possible to assign the car to either of the two classes with any confidence. Hence, the problem of duplicate cars.

3.1.4. Findings

The diagram below shows the percentage of matched, unmatched and duplicate data in the final output. Our current process currently matches an average of 93.5 % of the total passenger (and pickup) vehicle stock. On the issue of unmatched and duplicate data, clearly unmatched data is a more significant problem. However, we had enough matched data to satisfactorily build our regression model on what the future vehicle stock of Surrey will look like. Additionally, the classification and fuel consumption ratios for all the matched cars along with their respective VKTs helped develop our GHGs emissions tool.

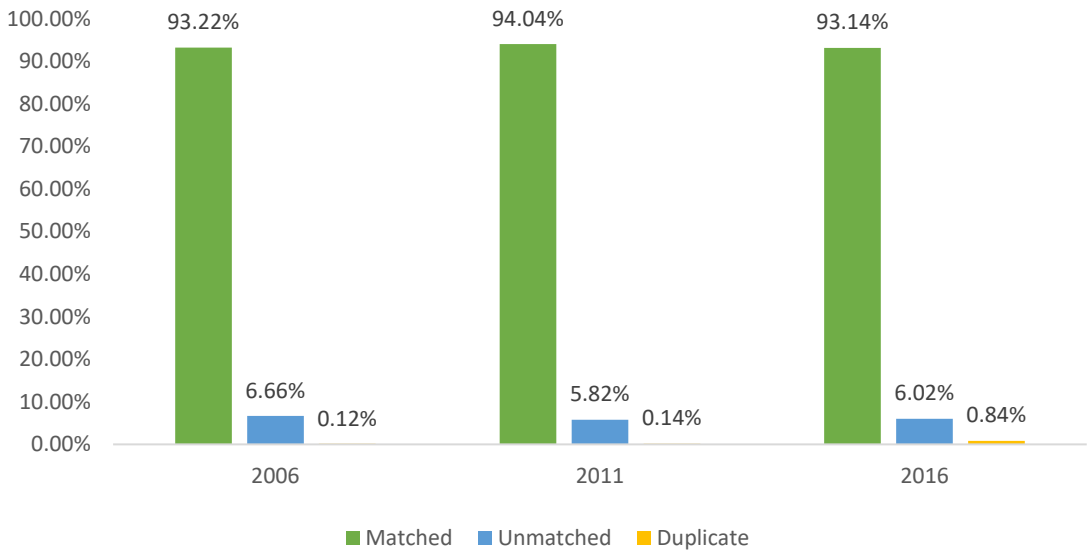


Figure 7. Percentage of matched, unmatched and duplicated data in the final output.

Another insight from this exercise, was to see how Surrey's vehicle stock has evolved over the years. The figure below shows the distribution of vehicles over the years 2006, 2011 and 2016 for the 12 vehicle classes.

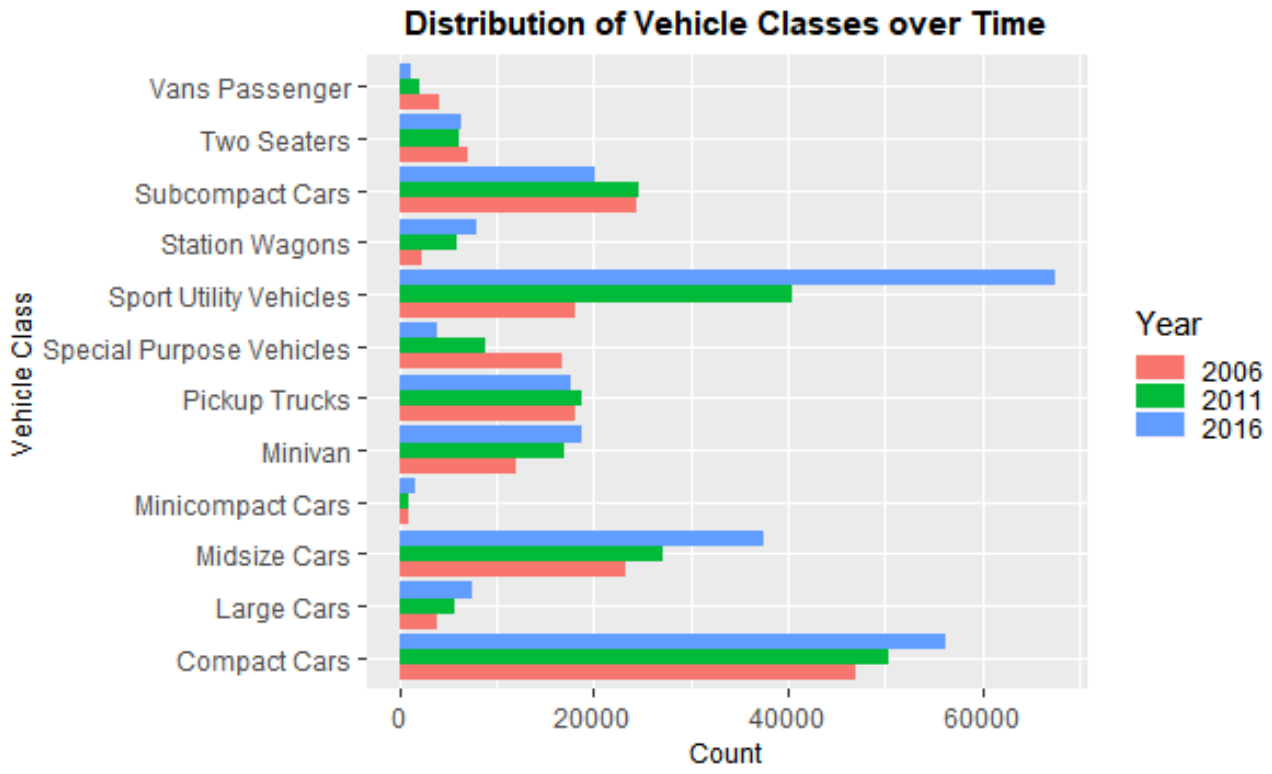


Figure 8. Distribution of vehicles over 2006, 2011, and 2016 for 12 vehicle classes.

Based on the above figure, it's very clear to see the general increasing trend in vehicle counts for most of the vehicle classes. The trend is especially visible for Sport Utility Vehicles (SUVs) which will be worrying for the City of Surrey. SUVs are larger passenger vehicles and so consume a lot more fuel and emit a lot for GHGs and therefore their increasing popularity should a cause for concern. The City of Surrey should look into enacting policies where people would prefer to purchase smaller or hybrid cars than go for an SUV. On the other end of the spectrum, we can see a reduction in Passenger Vans, Subcompact Cars and Special Purpose Vehicles counts. Also, pickup trucks count has stayed relatively stable over the years.

As City of Surrey population expands, it is important to implement policy that will favor increasing the amount of smaller and more efficient vehicles to meet future GHG emission targets. With fuel prices in Metro Vancouver being among the highest in North America and increasing popularity of electric vehicles, it would be interesting to see how the vehicle distribution of Surrey will evolve. We hope that our work, will play a part (albeit a small one) in helping City of Surrey guide its transportation sector towards a cleaner and more efficient future.

3.1.5. Validation

The first part of the process, we matched the make **exactly** between the two datasets whereas for model name matching, we compared string length in each row and used the shorter string as a "pattern" to be detected in the longer string. So, we can be confident that matches found at this stage are true matches.

In the second part of the process, we exact matched make (brand) names and only the first part of model names. E.g. A Ford Freestar Sport was matched to Ford Freestar Cargo Van and Freestar Wagon. Since both the matches are minivans, Ford Freestar Sport was classified as a minivan and the fuel consumption was averaged for both matches. The matches found in the second part are obviously associated with some degree of error. But due to time constraints and the difficulty of the problem, we could not quantify the error.

3.1.6. Next Steps

The obvious next steps are to address the issue of unmatched and duplicate data. For unmatched data, one option would be to continue the work we have done and locate all the abbreviated model names and rename them. The other option would be to try some machine learning algorithms capable of matching abbreviated car model names. Further work needs to be done to quantify the error introduces in our data by first-name matching (discussed in more detail in the previous section). It would also be interesting to see how the model performs for a new and larger ICBC dataset.

There is also the need to address the issue of duplicate cars where we lack enough information to confidently classify a vehicle. One suggestion that we looked into was of assigning weights. E.g. if a car can be classified as both compact or midsize, we attach equal weights to both classes. However, further research and discussion needs to be done on how best to address this problem.

3.2. Exploratory Analysis

3.2.1. Introduction & Approach

The City of Surrey seeks to gain a greater understanding of its transportation system to better inform climate policy, planning, and program action as the City continues to grow. With special access to proprietary datasets, such as the ICBC vehicle registry and Surrey demographics, we hope to elucidate trends and gain data-driven insights in the past and current states of vehicle ownership. Furthermore, our analysis focused on passenger vehicle ownership as defined in the initial project scope.

Our approach to exploratory analysis focuses on three main areas of inquiry: vehicle stock, vehicle age, and vehicle weight. Examples of central research questions may be as follows:

1. Where is passenger vehicle stock increasing per capita?
2. Where is passenger vehicle stock getting progressively younger or older?
3. Where is passenger vehicle stock getting heavier?

Various Surrey communities, such as those at a Regional Town Centre-level, exhibit different characteristics geospatially. Our exploratory analysis incorporates spatial components in visualizing these vehicle stock attributes on the Surrey map. In understanding areas where vehicle stock is increasing, aging, getting heavier, and the converse, our work aims to inform future climate planning.

In this section of the report, our exploratory analysis aims to provide context-specific research into the changing nature of transportation in the City of Surrey. As a secondary aim, insights and trends revealed in the vehicle stock may verify assumptions in the development of a GHG policy analysis tool.

3.2.2. Data Sources

In the exploratory analysis, data sources included:

- ICBC Vehicle Registry in 2006, 2011, 2016
- City of Surrey's Population Estimations
- City of Surrey's Employment Estimations

3.2.3. Sources of Error

Error from Geographic Rebasing

The ICBC Vehicle Registry dataset contained vehicle attributes (ex. make, model, net weight) that were connected to vehicle ownership at the Postal Code level. To be consistent with other project outputs and conventional transportation planning, Traffic Analysis Zones (TAZ) were used as the geospatial unit in visualizations and analysis. To convert between Postal Code and TAZ units required geographic rebasing which was completed at an earlier stage in the project.

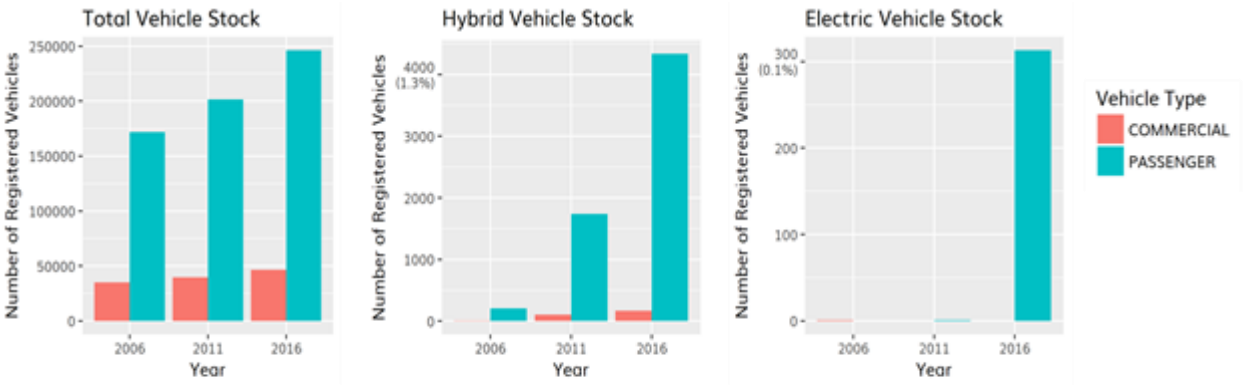
Defining Commercial and Passenger Vehicle Labels

Our exploratory analysis queried two main vehicle groups: Commercial and Passenger vehicles. In this section of the report, these labels were used according to Vehicle Type Description in the original ICBC registration. It is critical to note that our later work in defining broader vehicle classes identified differences in Commercial and Passenger vehicle classification according to their Use Category. Further investigation is required to understand the differences in this definition and its implication in downstream analysis

3.2.4. Findings

Total Vehicle Stock Counts in Relation to Green Vehicle Ownership

Total vehicle stock counts in 2006, 2011, and 2016 were visualized along with Green vehicle categories (Hybrid and Electric), as shown in Figure 5. We observed that the rate of change of vehicle stock counts were greater for passenger vehicles than commercial. Total vehicle stock was composed of passenger vehicles as a majority. Green vehicle adoption was also showed a trend to increase between 2006 and 2016; however, still early in its stages. In 2016, highest Hybrid and Electric vehicle counts were observed in our sample but only shared a relatively low 1.3% and 0.1% of the total vehicle stock. It should be noted that passenger, commercial, hybrid, and electric vehicle classification was defined by the original ICBC dataset.

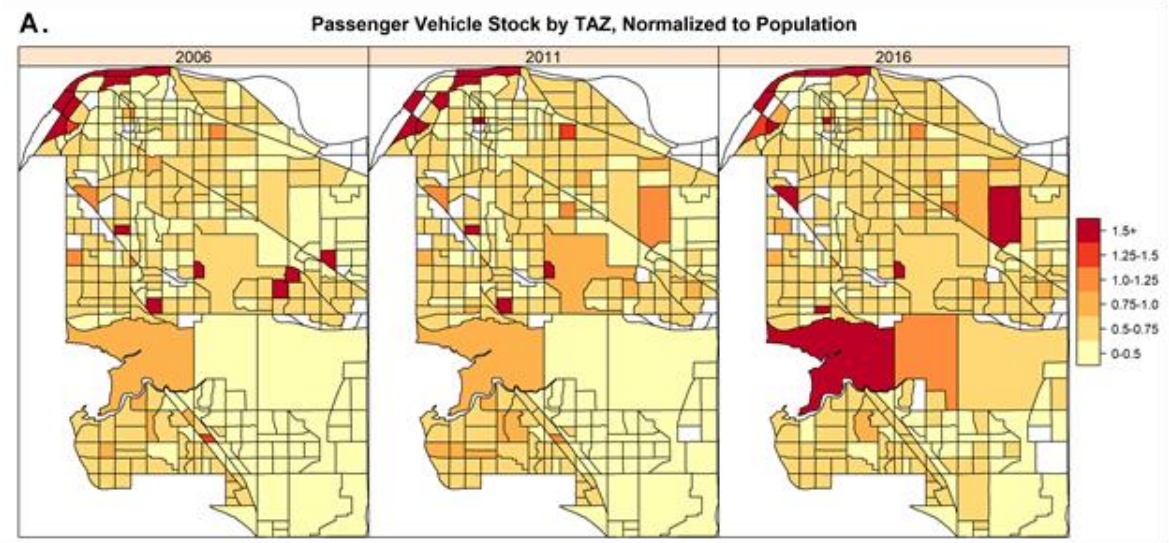


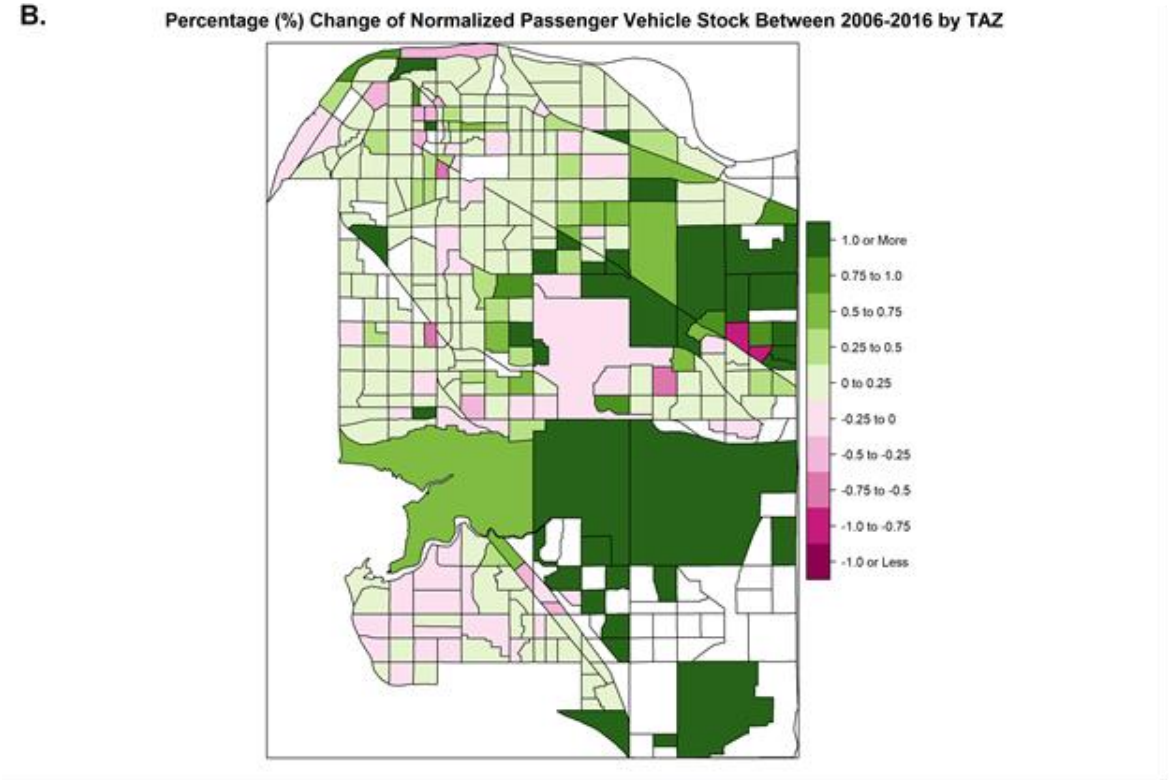
Passenger vehicles exhibit a greater rate of change and accounts for more of the total vehicle stock than Commercial vehicles. Green vehicle adoption, Hybrid and Electric, are still in early stages with max counts of approximately 4000 and 300 respectively sharing 1.3% and 0.1% of the total vehicle stock in 2016.

Figure 9. Visualizing Vehicle Stock Counts in 2006, 2011, and 2016

Spatially Visualizing Changes in Passenger Vehicles per Capita Between 2006 and 2011

To analyze passenger vehicle counts, we normalized values by the residence population in each TAZ for each year of data, as shown in Figure 6a. The normalization scheme was chosen to better depict the relationship between passenger vehicle counts and the likely end user group. Analyzing vehicle per capita is advantageous in spatial visualizations, as shown in Figure 6b, where areas of change may be interpreted as changes in vehicle ownership behavior opposed to the consequence of changes in population. Our visualizations revealed large spatial clusters of increasing passenger vehicles per capita. Passenger vehicle ownership is concluded to increase by visual inspection. Further research is required to contextualize with other external changes between 2006 and 2016.



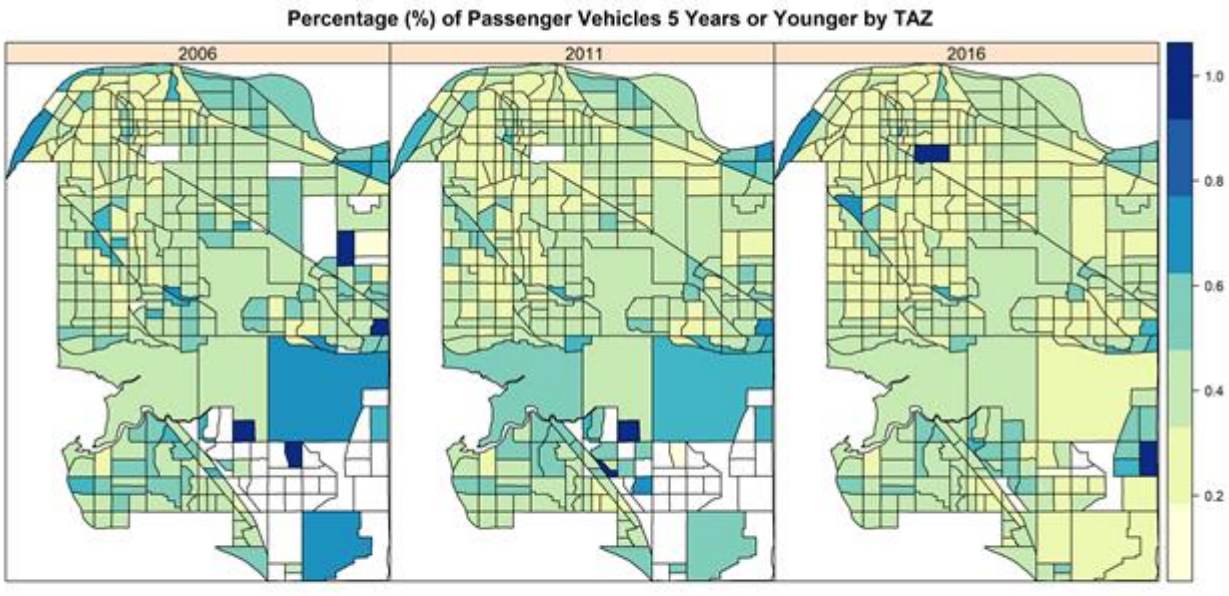


A. Spatially Visualizing Passenger Vehicles Per Capita in Each Year by TAZ. Passenger vehicle stock was normalized to population residing in each TAZ, obtaining passenger vehicle per capita. Color bar indicates vehicles per capita. *B. Percentage (%) Change of Normalized Passenger Vehicle Stock Between 2006 and 2016.* Color bar represents decimal (percentage converted) increase (green) and decrease (pink). Blank TAZ indicate unavailable data in passenger vehicle count or residential population.

Figure 10. Passenger Vehicle Stock

Identifying Spatial Clusters of Younger Vehicle Stock

Percentage of younger passenger vehicles were visualized in each TAZ, as shown in Figure 7. Younger vehicles were defined to be no more than five years of age, where vehicle age was defined as the difference between the dataset year and model year of the vehicle. On the two endpoints of the color bar, yellow indicates lower percentages of younger vehicles and implies larger proportions of older vehicles. Conversely, higher percentages (blue) also implies smaller proportions of older vehicles older than 5 years. Visually, there is a decreasing trend of younger vehicles as observed from increasing lighter yellow clusters on the map. Although there are few TAZ exceptions, we generally concluded that vehicles are getting older.



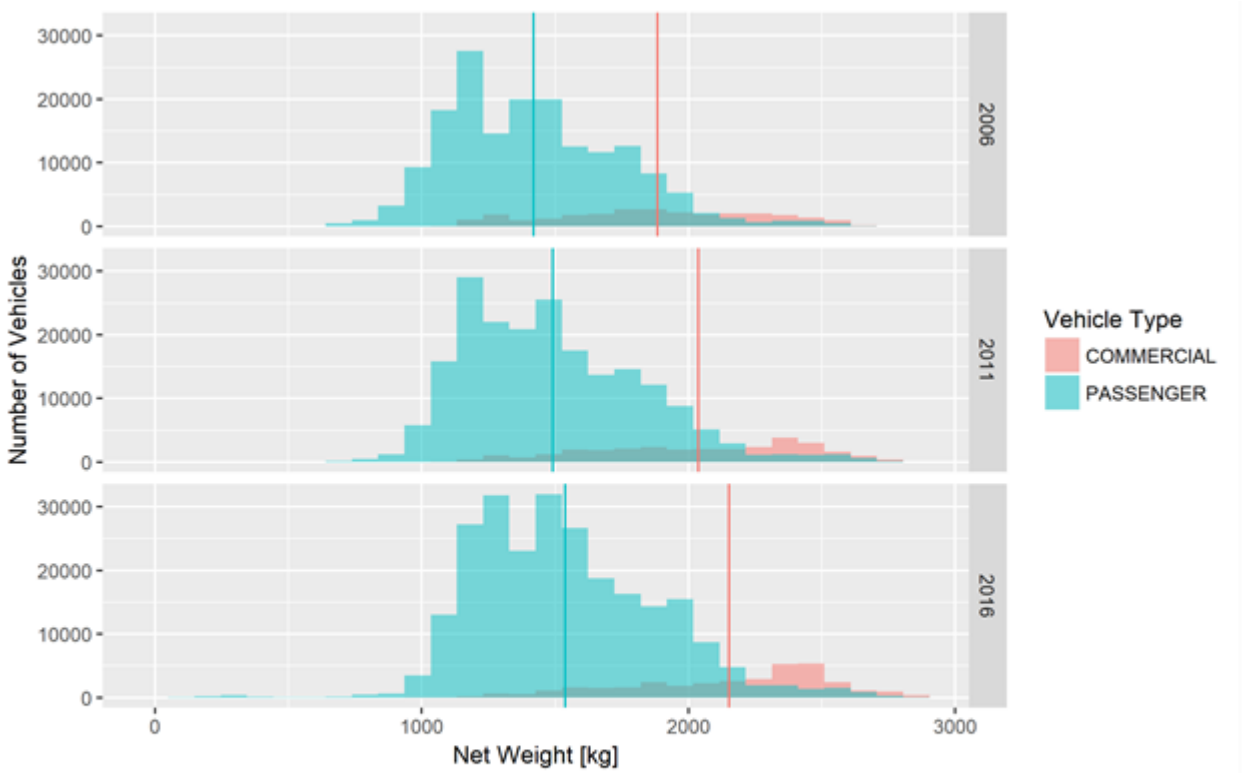
Younger Vehicles were defined to be of 5 years or younger. Color bar represents decimal (percentage converted) gradient between 100% (blue) and 0% (yellow) younger vehicles. Blank TAZ indicate unavailable data in each dataset year.

Figure 11. Percentage of Younger Vehicles in each TAZ

Investigating Changes in Vehicle Net Weight

Distribution of vehicle net weight for passenger and commercial vehicles were visualized by a histogram in Figure 8. Visually, we observed that passenger vehicle net weight showed an increasing trend from 2006 to 2016. Commercial vehicle stock was also observed to be distributed at higher net weights compared to passenger vehicles.

To identify areas with changing mean net weight of passenger vehicles, percentage change of mean net weight was spatially visualized in Figure 9. We observed larger spatial clusters of increasing (purple) mean net weight with less areas of decreasing (orange) mean net weight. As the lighter shade clusters were interpreted with caution, our spatial visualization is consistent with our observation from the histogram that vehicles are generally increasing in net weight. Further statistical modelling is required to test the hypothesis that vehicle net weight is significantly changing with time.



Vertical lines of corresponding colors represent the mean net weight of each group. 30 bins used in histogram. Visually, histograms showed a relative increase of mean net weight from 2006 to 2016 for both passenger and commercial vehicles. Distribution of Commercial vehicle net weight was generally found to be higher than Passenger vehicles.

Figure 12. Histogram of Vehicle Net Weight.

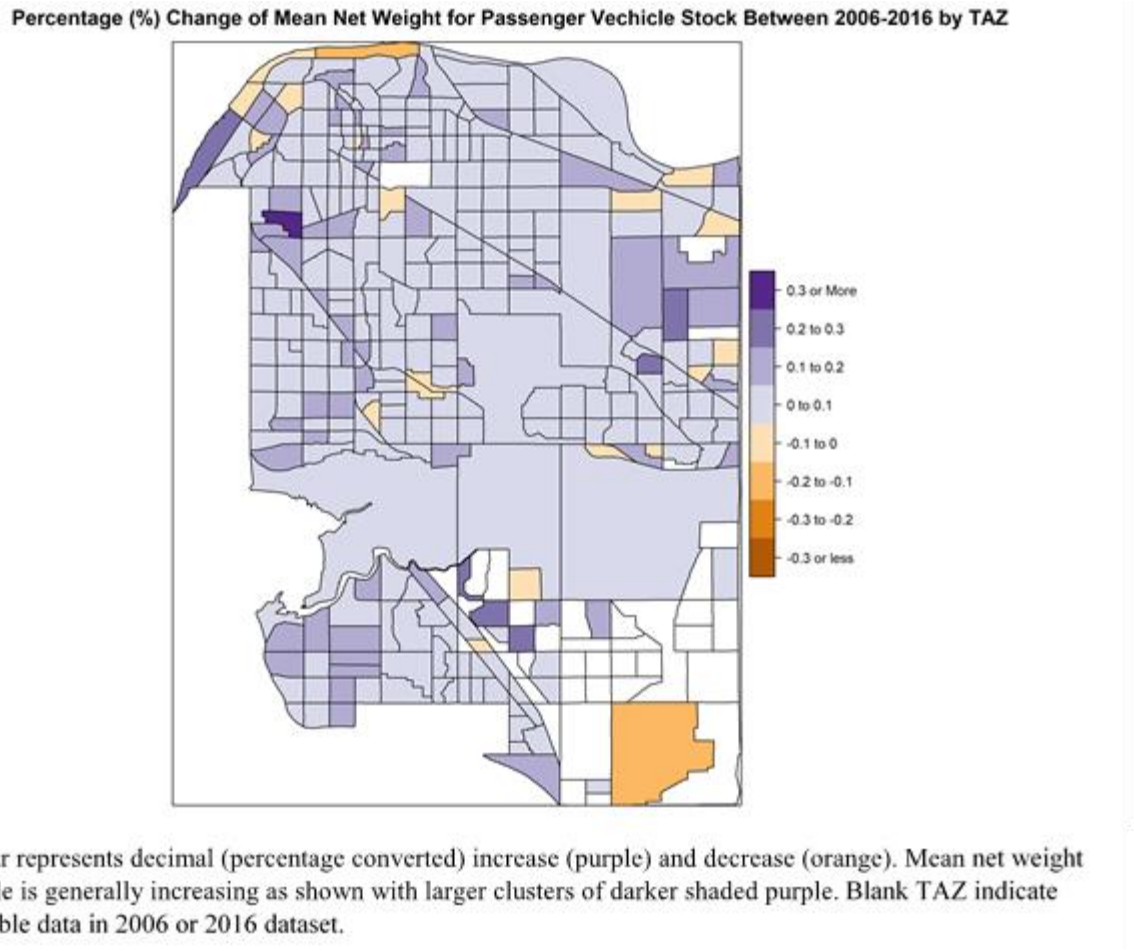


Figure 13. Spatially Visualizing Percentage Change of Average Net Weight Between 2006 and 2016

Summarizing Changes in Vehicle Attributes from 2006 to 2016

To summarize our findings, we focused on three main attributes of our vehicle stock: Vehicle per Capita, Vehicle Weight, and Vehicle Age. In Figure 10, we summarized the changes in the attributes by grouping our spatial units according to their changes, whether the attribute has increased or decreased between 2006 and 2016. The summary table further provides a link between the geospatial visualizations of the three main vehicle attributes noted in the previous sections. We observed that most spatial units had increasing Vehicles per Capita, increasing Vehicle Weight, and is also getting older. This group was observed at 52% of TAZ Count and accounted for 69% of the residential population. The large spatial weight and population that this group holds may point towards a vulnerable group to green policy. Conversely, a group of decreasing Vehicles per Capita, decreasing Vehicle Weight, and is getting younger had little weight in this summary.

Vehicle per Capita	Vehicle Weight	Vehicle Age	TAZ Count	% of TAZ Count	% of TAZ by Pop. (in 2016)
↑ Vehicles Per Capita	↑ Weight	Older	195	52.14%	69.43%
		Younger	43	11.50%	11.92%
	↓ Weight	Older	5	1.34%	0.28%
		Younger	3	0.80%	0.99%
↓ Vehicles Per Capita	↑ Weight	Older	45	12.03%	9.25%
		Younger	11	2.94%	2.97%
	↓ Weight	Older	4	1.07%	0.23%
		Younger	2	0.53%	0.34%

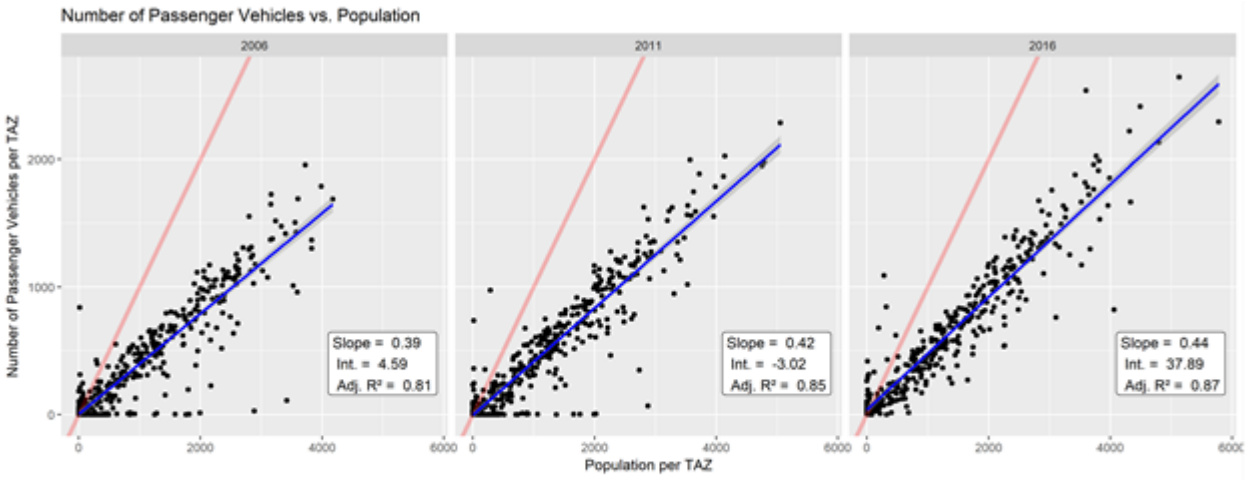
TAZs were grouped by the increase or decrease of 3 main vehicle attributes: Vehicle per Capita, Vehicle Weight, and Vehicle Age. TAZ Count refers to the number of spatial units within each group. % of TAZ Count is the proportional weight the group accounts for in terms of TAZ while % TAZ by Pop. refers to the amount of population accounted for within each group. Note that TAZ with unavailable information was accounted for but not grouped.

Figure 14. Table Summary of the Changing Vehicle Attributes Within Each TAZ Between 2006 and 2016.

3.2.5. Validation

Validating Vehicles per Capita Normalization Scheme

To check that our vehicle per capita normalization scheme was reasonable, we visualized the number of passenger vehicles to the residential population in each TAZ in Figure 11. Passenger vehicle counts were found to increase with population linearly by visual inspection and a fitted simple regression model. As there seems to be a clear relationship between passenger vehicles and population, this suggests that normalizing passenger vehicles to residential population was reasonable.



Red line represents a reference line y=x. Blue line represents a simple regression model and certain outputs in bottom right square. We observed that passenger vehicles increased with population, suggesting a reasonable normalization scheme.

Figure 15. Number of Passenger Vehicles vs. Population in each TAZ.

3.2.6. Next Steps

In our exploratory analysis, we were able to conclude trends on the vehicle stock, age, and weight by visual inspection. To answer our initial research questions, we found that vehicle stock is generally increasing, getting older, and getting heavier. These three attributes are suggested to give insight in areas where it is exceptionally bad for GHG emissions from passenger vehicles. Subsequently, these areas may also be vulnerable and have the most impact from policy with a deeper insight. Further research is required to contextualize the overlap between these three areas of query. Critical next steps may include (1) characterizing the behaviors that drive the trends of these three attributes and (2) to obtain more data for rigorous statistical modelling and hypothesis testing:

Characterizing Stock, Age, and Weight Trends with Demographic Variables

To further extend our exploratory analysis, contextualizing our findings with demographic variables would be key. Due to the constraint of time and difficulty to extract census data to a TAZ level, we were unsuccessful to provide additional insights with demographics. Next steps may be to query socio-economic factors to find correlations to increasing, aging, and heavier vehicle profiles.

Hypothesis Testing and Statistical Modelling: Obtaining Unique Identifiers

Our findings work to visualize and leverage the abundance of data that was available for analysis. As we were able to identify trends through visual inspection, more rigorous next steps may be to verify our observations through hypothesis testing with inferential models. We encountered a data gap where unique identifiers were not made available across ICBC datasets. This was found to be a critical feature as vehicle ownership, as shown in our findings, often lasted more than 5 year, the time interval of data that was available. As a result, each dataset contained a large proportion of retained vehicles where our models failed to reliably hypothesis test with overdispersion. Overdispersion occurs when the variability of the dataset is higher than the variance of a theoretical model. If unique identifiers were available, Mixed-Effects modelling may be considered to resolve this issue.

3.3. Forecasting (Kevin)

3.3.1. Introduction & Approach

Business-As-Usual (BAU) vehicle stock models, which assume that Surrey's future vehicle growth trend will be similar to the past (during the period of 2006-2016), have been developed to provide baseline forecasts of Surrey vehicle profile assuming the status quo with regards to current local and senior government policy.

These models were developed at three geographic levels: A City-wide level, a community-level and a Traffic Analysis Zone (TAZ) level. In addition to be useful for creating and evaluating the size of future vehicle stocks, these forecasts can be also used to estimate future GHG emissions, given adequate information on vehicle kilometers travelled (vkt) of these vehicles, their fuel consumption (included in the ICBC dataset) and appropriate emission factors.

For each geographic unit, two categories of vehicle stock models were developed, with one type predicting total vehicle counts within each geographic unit, and the other predicting vehicle counts per vehicle class (as defined for the vehicle stock classification scheme) While the latter type of model would yield a more interesting vehicle stock profile since it considers the trends and variations of each vehicle class, the former type of model was also developed for comparisons of results obtained from the vehicle class-based model.

With the exception of a few simple models, most BAU vehicle stock models were fitted as Linear Mixed Effects Models (LMEM), as opposed to being standard linear regression models or time-series models. The fitting was carried out using the gam function of the mgcv package in R. In linear mixed effects models, effects of any independent variables on the dependent variable can be separated into fixed and random effects, where the coefficient of a random effects is a random variable that vary across another categorical variable. For instance, consider the following basic model with m independent variables:

$$\log(Cij)=(0+i,0+j,0)+(1+i,1+j,1)y1+(2+i,2+j,2)y2+...+ij$$

where

Cij = vehicle count in class i and geographic unit j

yk= independent variable k (k from 1 to m)

k = fixed-effect coefficient for independent variable k (k from 1 to m) or the intercept (where k = 0)

i,k= random-effect coefficient for independent variable k (k from 1 to m) or the intercept (where k = 0) and an observation in vehicle class i

j,k= random-effect coefficient for independent variable k (k from 1 to m) or the intercept (where k = 0) and an observation in geographic unit j

ij= Residual of the model

Comparing to a standard linear regression model, the model above offers the advantage of taking into account how group variables (such as vehicle class, geographic unit) may play a role in the relationship between dependent and the independent variables. This allows the model to more accurately capture the intrinsic relationships between the dependent, independent and categorical variables. Furthermore, the use of random effects to model group effects reduces the complexity of the model by "hiding" group and interaction effects with random effects. Here, this works because these BAU modes are supposed to be mainly used for forecasts, and there is little interest in the coefficients of these models, which, as will be shown, cannot and will not be interpreted in any reasonable fashion.

On the other hand, time-series regression models were not adopted due to the size of our data-set. If such models were to be fitted, one class-based model would be needed for the combination of each geographic unit and vehicle class. Since only the 2006, 2011 and 2016 ICBC data has been made available, each model would be limited to

having 3 observations, which is undoubtedly insufficient for any model fitting purposes. The data have too short of a time span to for building reliable forecast models.

Data visualization took place prior to the modelling process, which led to the production of some histograms and linear plots, giving a sense of the potential variable transformation and distribution that should be used in regression. The model fitting process mainly consists of building models with different combinations of independent variables, transformation of the dependent variable and model distributions.

Each model is evaluated for its fit and standard regression diagnostics, which include the quantile-quantile plot (q-q plot) of the deviance residuals, plot of residuals versus predicted values, histogram of the residuals and plot of response versus fitted variable. Fit and diagnostics are then used to inform whether a model should be retained, further modified, or eliminated from consideration. Between models of different specifications, the following statistics are used to compare their performance:

- Adjusted Coefficient of Determination (R2) - A goodness-of-fit metric that measures the percentage of variance in the dependent variable explained by the independent variables, adjusted to penalize for high number of independent variables
- Deviance explained - A goodness-of-fit measure that generalizes R2 to models fitted by maximum likelihood
- Akaike Information Criteria (AIC) - A goodness-of-fit measure that is similar to the adjusted R2, but for likelihood-based models (which include LMEM). Its rewards models with good fit and penalizes models for adding "bad" explanatory variables. Unfortunately, AIC cannot be used to compare models with a different form of dependent variables (vehicle stock) or model distributions.

Moreover, "backcasting" has also been carried out to validate model results and to ensure that selected models fit pre-existing data adequately well. This will be further described in section 3.3.5. Cross-validation techniques are also applied to compare model performance.

We should note that models in this section are fitted by iteratively modifying the model specification to improve their fit and diagnostics. While this strategy allows the building of a good-performing prediction models, it effectively voids any validity and usefulness of the model coefficients. None of the coefficients from such a prediction model can be relied upon for the purpose of hypothesis testing or interpreting correlations. Therefore, the only plausible use of such models must be to produce vehicle stock predictions.

In addition to BAU vehicle stock models, attempts were made to construct a regression model to understand the relationship between new vehicle stock (defined as vehicles of less than 5 years of age) and census demographic variables. However, after a few initial unsuccessful attempts of fitting models via backward selection, it became obvious that this would be a very challenging effort as a stand-alone task, and it was not be completed in time due to this team's tight schedule. As data have been prepared to complete this effort, we strongly suggest that this task be completed at a later date.

3.3.2. Data Sources

3.3.2.1. Surrey Population and Units Data

The City of Surrey prepared for this project population and housing unit estimates between 2001 and 2017, as well as forecasts between 2018 and 2046. These estimates and forecasts were also broken down into housing types (apartment, single family housing, etc.), However, it is preferred to use total population and employment as the explanatory variables due to the small sample size of the data, as using population or employment variables broken down into housing types may lead to overfitting.

3.3.2.2. Census Data

The attempt to fit a new vehicle stock model made use of the census data from Statistics Canada which was standardized by our team to TAZ level, which is described in section 3.0 in further detail.

3.3.3. Sources of Error

There are two main potential sources of error in the modelling work described in this section:

1. The models heavily rely on Surrey's population and employment estimates and forecasts, thus, any issues with those estimates would also lead to inaccuracies of the models; and
2. The modestly-sized ICBC dataset may lead to issues in the model specification. For example, the variance may actually fit another distribution better, or different transformations of the dependent and/or independent variables should have been used. In any case, the short time space of the ICBC data may keep certain trends in vehicle stock secret.

3.3.4. Findings

Throughout the fitting process, a total of 23 class-based stock models at the community level were fitted. These models were developed based on 10 City-level and 78 community-level models predicting total vehicle counts of all classes, as well as 31 city-level models predicting vehicle counts per vehicle class. City-level models and total stock models are simple and provide valuable insights into which predictor variables should be considered for the final model, as well as the distribution that should be used in the regression. TAZ-based models were also attempted; however, these models displayed too much "noise" in comparison to the community-level models, with non-satisfactory model fit and diagnostics. Conversely, city-wide models are too broad in coverage and fitted with very little data (n = 3). Thus, we believe that a community level model should be adopted at the end.

Ultimately, three models were selected out of all 23 class-based community-level stock models for final evaluations. For vehicle class i and community j, the three models can be written as :

log(Countsi,j)=0 +0,i +0,j +1log(Unitsj)+(2,i +2,j)Unitsj+ ij,
where the model residual ij follows a normal distribution

log(Countsi,j)=0 +0,i +0,j +(1+ 2,i)Unitsj+ij,
where the model residual ij follows a log-gamma distribution

log(Countsi,j)=0 +0,i +0,j +(1+ 2,i)Unitsj+ ij,
where the model residual ij follows a log-normal distribution

Note: i refers to the index of a vehicle class and j refers to the index of a community

As one can see, the three models have very similar formulation, with community units or its logarithmic transformation as the main independent variables. They also include random group effects due to the community or vehicle class. The biggest difference of these models lies within the residual distribution.

The second model was the first to be eliminated due to the apparent heteroscedasticity illustrated in its residual vs. fitted plot, as shown below. Residuals are getting smaller as the predicted value increases, which makes the second model undesirable.

The first model is very similar to the third model, both in terms of its model specification, fit statistics as well as diagnostics. However, when looking at the backcasting results of the first model, it appears that the model underestimates 2016 vehicle stock by 10%, which was higher than the inaccuracy of the log-normal model. This leads to the concern that whether this model, with only one logarithmic transformation of the dependent variable, overlooks some other non-linear effects in the relationship. Furthermore, the Cross Validation Mean Square Error (CV-MSE), which is calculated through a technique (cross validation) that assesses how the model would fit independent data-sets, is slightly higher for the first model than for the third model. (The second and the third model has similar CV-MSE). Overall, the third model with log-normal regression appear to be the most satisfying. Its summary is as follows: where the model residual ij follows a log-normal distribution

Model formula: log(Countsi,j)=0 +0,i +0,j +(1+ 2,i)Unitsj+ ij

Fit via log-normal regression

Adj-R2 = 0.915; Deviance Explained = 92.3 %; n = 273

Using this model, the city-wide vehicle stock can be predicted as in the following plot:

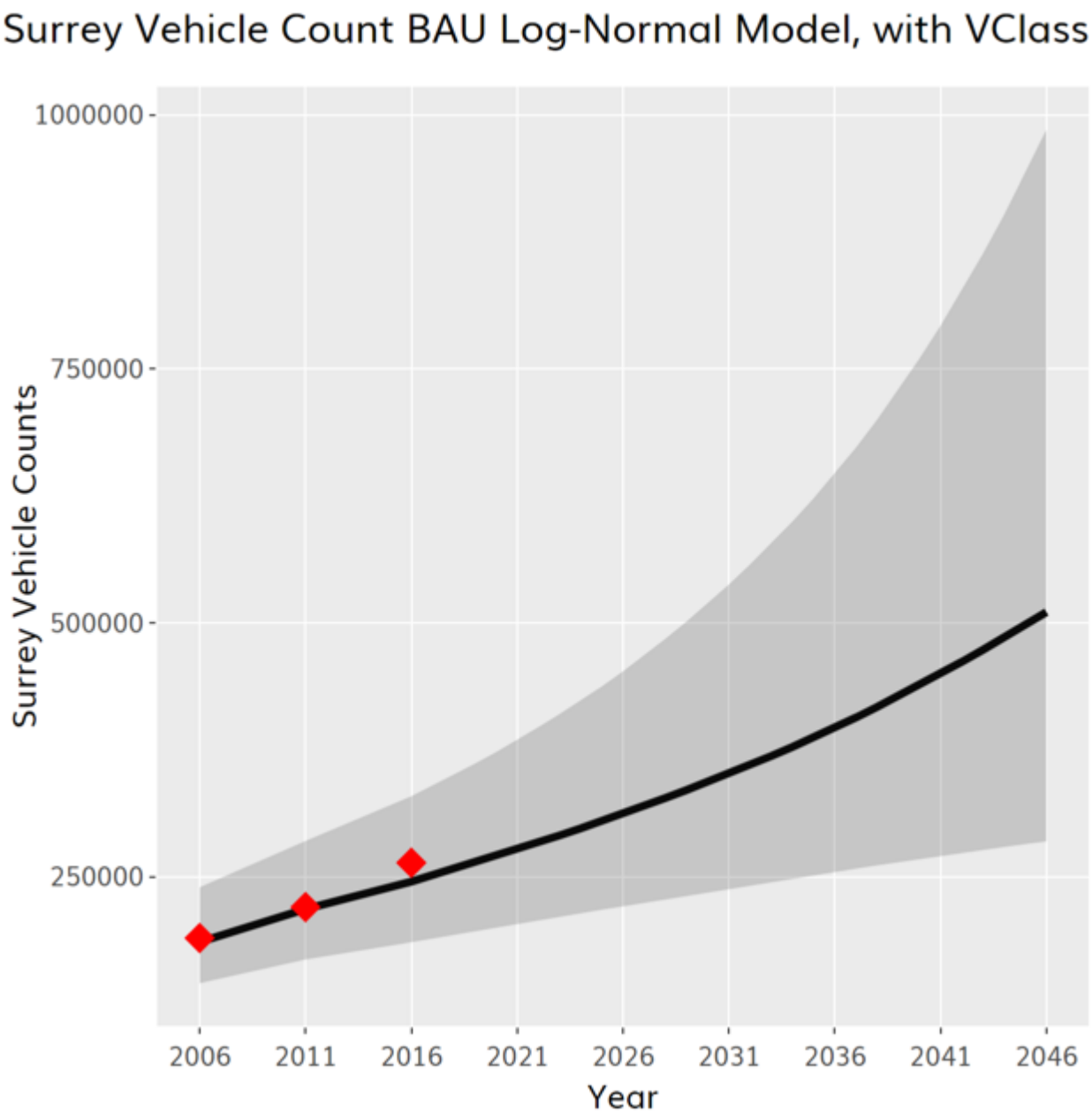


Figure 16 – Model 3 Forecasting Results

Converting the city-wide forecasts to per-capita ones:

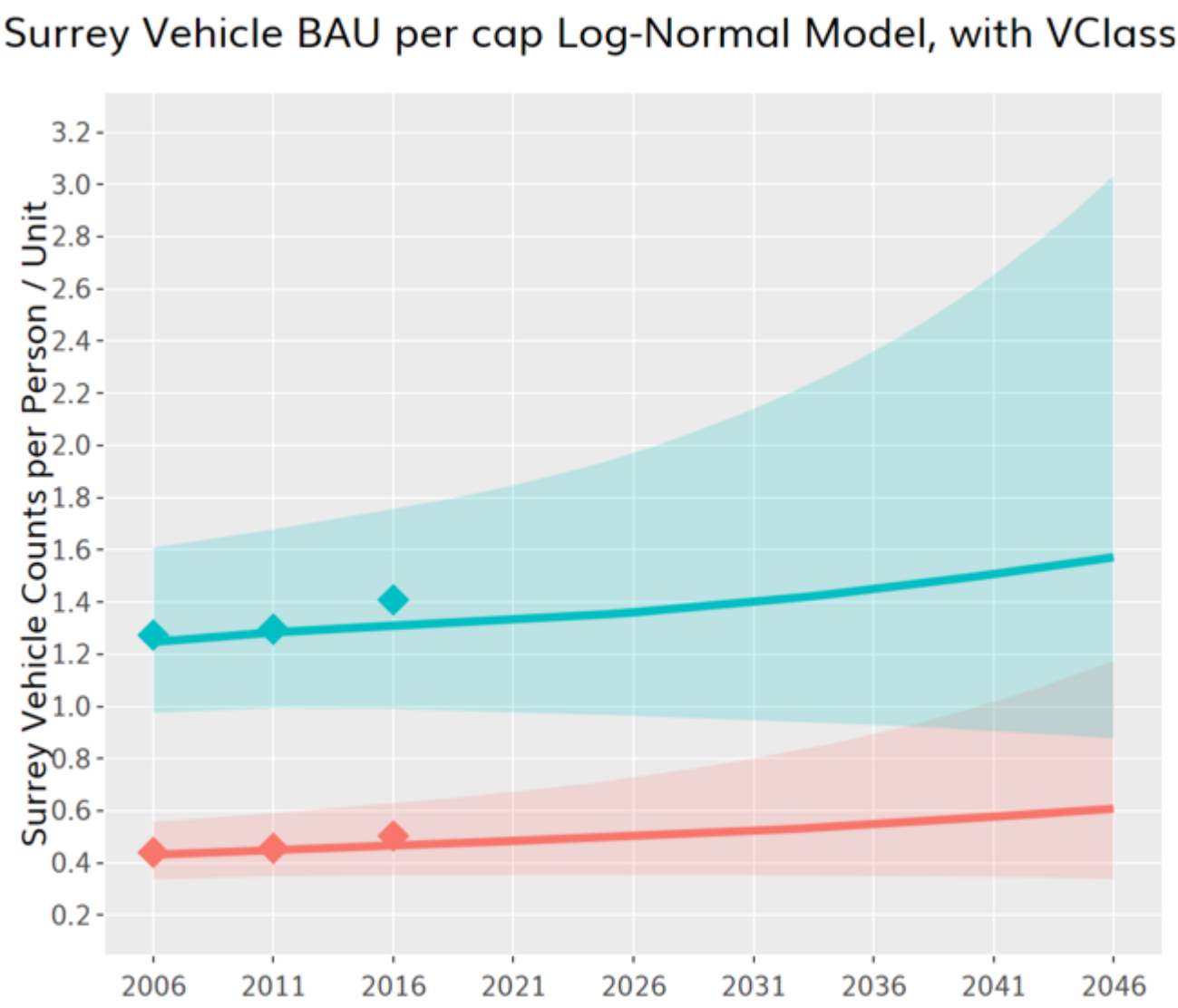


Figure 17 – Per capita Log Normal Model Results

In the plots above, all bands indicate 95% prediction intervals, and the diamond-shaped points are actual vehicle counts (total or per capita) obtained from raw ICBC data. One thing that stands out is the large prediction intervals as it gets further into the future. This should not come as a surprise. For one thing, the future is naturally full of unpredictable variables and uncertainties. Thus, any educated assumption would have a challenging time describing the future accurately. Additionally, the limited data size further increases the difficulty for models to estimate the trends underlying Surrey's vehicle stock. As a result, the large prediction interval for predictions more than 20 years from now is unavoidable. One should be extra cautious or outright avoid using predictions beyond 2031/2036, and ensure that the 95% prediction intervals are reported whenever these predictions are used.

An additional issue that requires special attention is how the effects of a few special vehicle classes may get buried within the general trend. As the log-linear LMEM separates the intercept and slope for the housing unit variable into fixed effects (common to all vehicle class) and random effects (only for specific classes), if majority of the classes share a similar trend, then "outlier" classes that follow a different trend may have their trend overshadowed by the "general trend", particularly due to the small data-set problem. Take the vehicle class

"Subcompact Cars" as an example. It is believed that ICBC began to define some vehicles that would have fallen in the SPV class in 2006 into other classes in 2011 and 2016. As a result, the vehicle counts show a decreasing trend for this category, while forecasts from models, including the log-linear LMEM shows an increasing trend:

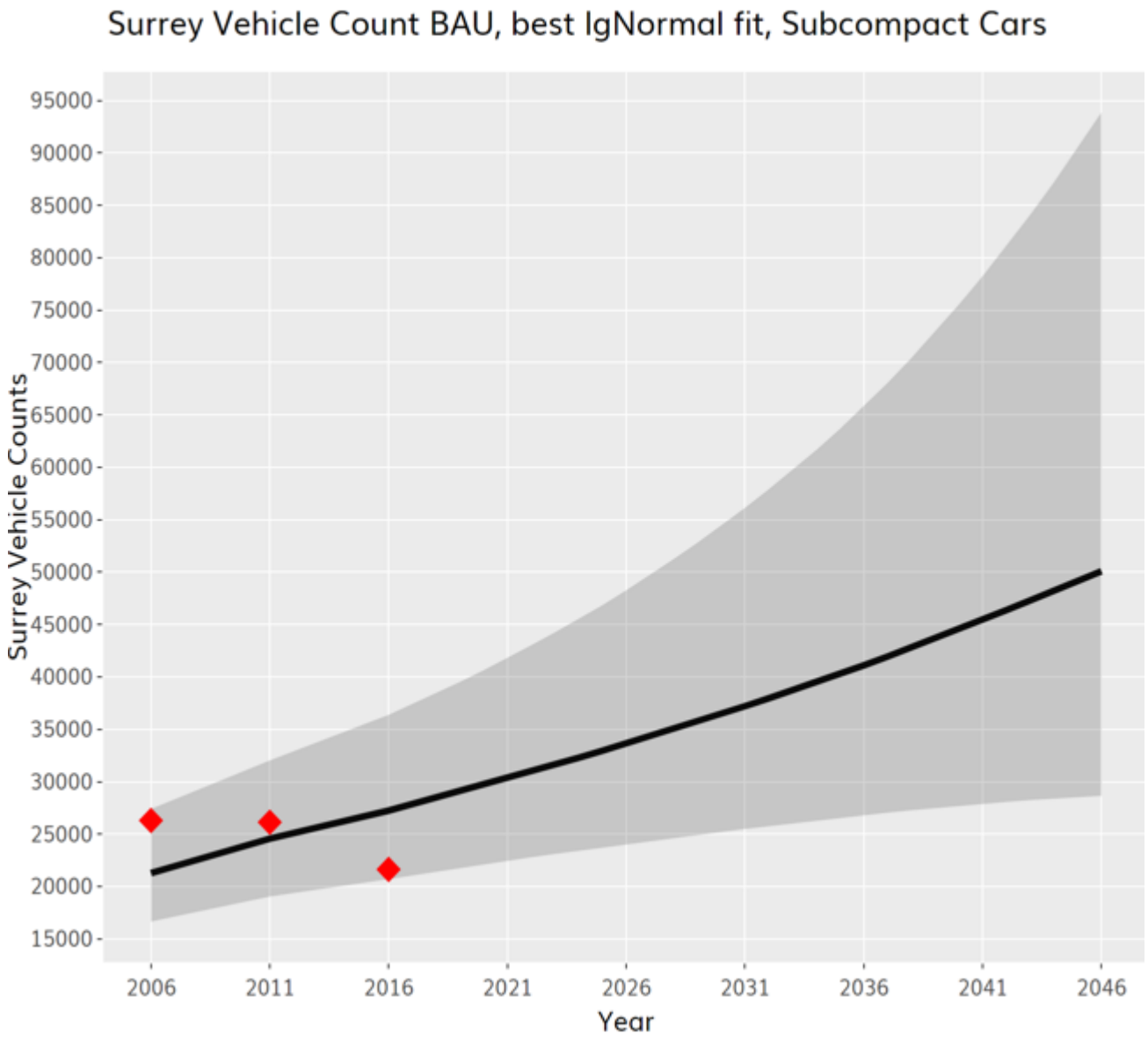


Figure 18 – Example of Decreasing data with Increasing Trend

Given the modest ICBC data-set size in this project, it is certainly hard to detect trends of all 15 vehicle classes accurately in 1 model. It is recommended that further refining of vehicle classes and merging certain vehicle classes be considered to improve accuracy of class-based predictions.

3.3.5. Validation

As mentioned above, validations of our main model are carried out through model diagnostics and backcasting. The model diagnostics show the following:

The model does have a residual vs. predicted plot that displays slight heteroscedasticity and a residuals histogram that is not perfectly normal. That being said, the diagnostic plots are all reasonably acceptable and no other models fitted have been found to circumvent the problems shown in diagnostics of this model. As for backcasting, the predicted values and actual vehicle counts at the city level are quite close (within 5-7%).

3.3.6. Next Steps

Going forward, it will be ideal if there are more ICBC vehicle stock data made available to modelling vehicle stock. This may occur naturally as time passes by, or as ICBC releases more vehicle stock data in the past. Regardless, the limited dataset size has been a key roadblock in the modelling work, and it will be very useful if this can be addressed. Furthermore, more considerations should be given to merging certain vehicle classes in the meantime, to make up for some vehicle class trends being "covered up" by the general trend due to the size of the data-set.

A more challenging task that remains incomplete involves finding the relationship between demographic, transportation and other variables and the vehicle stock life-cycle. Unfortunately, this was not possible to execute due to the following issues:

- Time constraints prohibited moving forward with this complicated task;
- No transportation data were made available (network skims / travel costs, mode share, etc.) to provide additional variables for modelling; and
- No unique identifiers of the vehicles were provided with the ICBC data, making it extremely difficult to trace vehicles across years.
- Had we overcome time constraints and data gaps, it is theoretically possible to develop stronger more detailed models that would allow for a more in-depth discussion of the link between vehicle stocks and other variables, such as stock-flow models or Markov models.

3.4. Tool (Mia)

3.4.1. Introduction & Approach

A major plus for data-driven policy making is easy and intuitive access to data and the results of models and simulations, with no programming experience required. As such, we planned to create such an interface.

The creation of this interface allows people in the City of Surrey who have no data analysis or programming experience to quickly and easily make use of the results we have gotten so far. While it is not entirely complete, it has a framework that can be extended into a working tool that could provide insights for policy makers in the city.

The idea behind the tool is that it would have two major components, to match the two major components of our project. These are as follows:

3.4.1.1. Variable Visualization Tool

Given the unprecedented access to ICBC data, allowing people to explore this data and possible correlations with other data are essential to understanding vehicle stock in the City of Surrey. The goal, essentially, is to find which variables correlate well and to ask questions that imply correlation. For example, we might ask if owning larger vehicles correlates with income, or family size? Intuitively, we might say yes, but this tool would allow numeric validation (or invalidation) of these intuitions.

Currently, the only functioning part of the app is the display of Surrey's 2016 total population estimate (Figure 15). However, this demonstrates that the single variable visualization is functional. The correlation mode works in theory, but hasn't been tested as there were no other completed remapped variables for testing.

Methods for computing correlation metrics are in place, but have not been tested due to there only being one variable available.

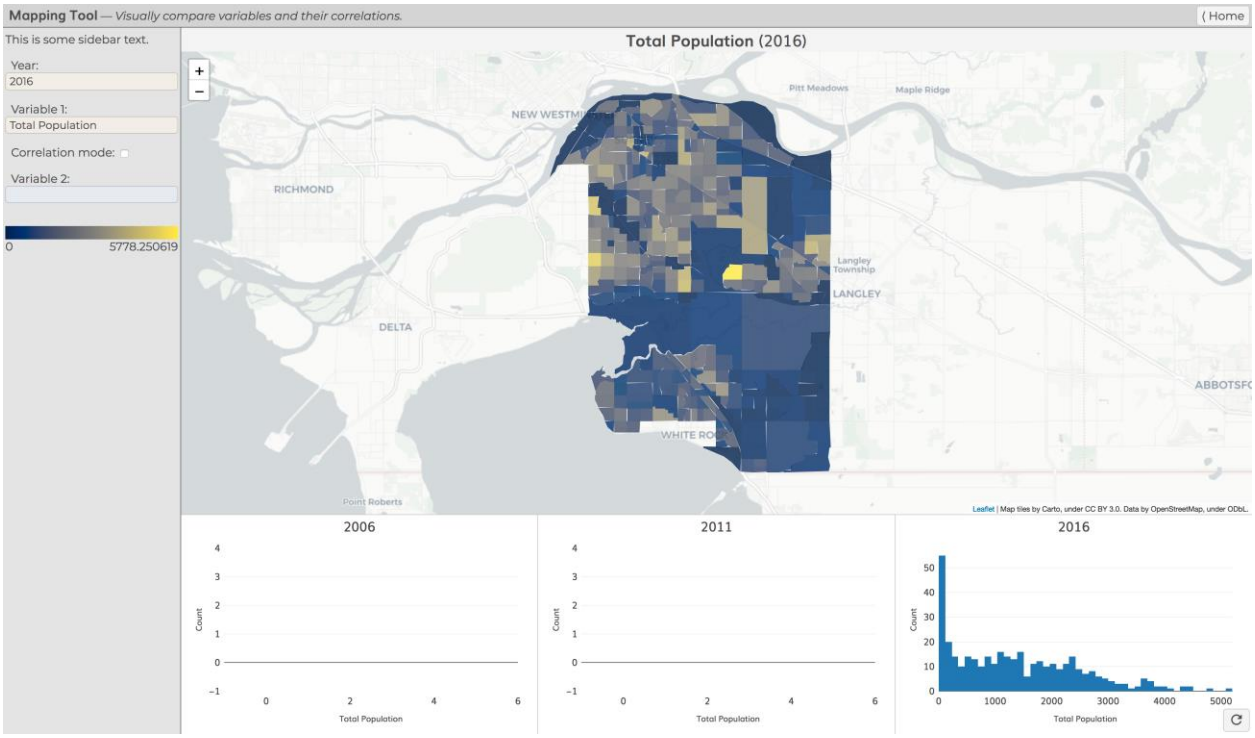


Figure 19. Screenshot of the functioning portion of the map tool.

In the image above, on the left can be seen the controls: selecting the year that is displayed on the map, the first variable, whether we are comparing, and the second variable. If correlation is off, the second variable can not be chosen and the first will be displayed. If it is on, both can be chosen and their correlation will be displayed on the map. For a single variable, histograms are displayed on the bottom for each year. For two variables, a scatterplot is shown.

Once data is fully prepared for the app, it should be converted to the following two formats:

- GeoJSON. For each TAZ, the variable should be stored as a Property of each TAZ, which should be stored as a Feature.
- JSON (custom schema). This flatter format is described below.

Two formats are used because this allows us to load data only once for both the map and the histograms/scatterplots. Loading the data twice makes computation much faster. Switching a variable means simply switching a key, and the reactive nature of Vue.js makes updating the plots very simple, and doesn't replace any data structures.

The format is as follows:

```
{

"variable_key": {
        "2006": [1, 2, 3, …],
        "2010": [4, 5, 6, …],
        "2016": [5, 6, 7, …]
},
"other_key": {
…
}

}
```

Note that all the arrays are the same length; the number of TAZs there are. The variable keys may not be the same between the two sets. The variables are stored in a JavaScript class:

```
class Variable {

constructor(code, feature, name, data=null, nomap=null) {
        this.code = code;
        this.name = name;
        this.feature = feature;
        this.data = data;
        this.nomap = nomap;
}

        }
```

The code member is the variable code, used for lookup and for the flat (non-geographic) data storage. The name field is the user-facing name. The feature is the name used in the geographic data. The data and nomap fields are where data is stored after loading.

3.4.1.2. Policy Analysis Tool

This tool is largely completed. The basic idea is that the user creates Policies, targets for vehicle stock in a particular year. The format is displayed in Table 2:

Table 2: Demonstration of the Policy format in the policy analysis tool.

	2020	**2030**	**2040**	**2050**
Total Stock	123456	234567	345678	456789
Vehicle class 1	25%	30%	40%	45%
Vehicle class 2	5%	6%	7%	8%
.	.	.	.	.

The total stock is given in number of cars, and the composition of cars is given in percentages. There is a sum of percentages displayed to help the user sum percentages to unity.

Each policy is also given a name and a space for notes. Policies can be saved and deleted.

In the interface (Figure 16), there are two lists of policies. There is a large list in the centre of the interface: these are the *active* policies. These policies will have their GHG emissions computed and

compared when a simulation is run. On the left can be seen a less detailed list of policies: these are *inactive*. Policies can be activated and deactivated. Inactive policies can still be edited and saved (Figure 17).

The lock icon next to the name of a policy displays it save state. If the icon is red and unlocked, the policy has never been saved. If it is locked and green, it is saved and un- modified. If it is locked and yellow, it has been saved but has been modified.

On the right are simulation parameters. There are two options; the emissions factor used and the interpolation method. The first has no effect on the output, since all outputs are scaled relative to the 2016 levels. For completeness, there are five interpolation options: nearest, zero, linear, quadratic, and cubic. It defaults to quadratic, and this seems to work quite well. To my knowledge, there are no particular advantages to any, other than that second and third order interpolation produces smoother results, which are at the very least more visually appealing.

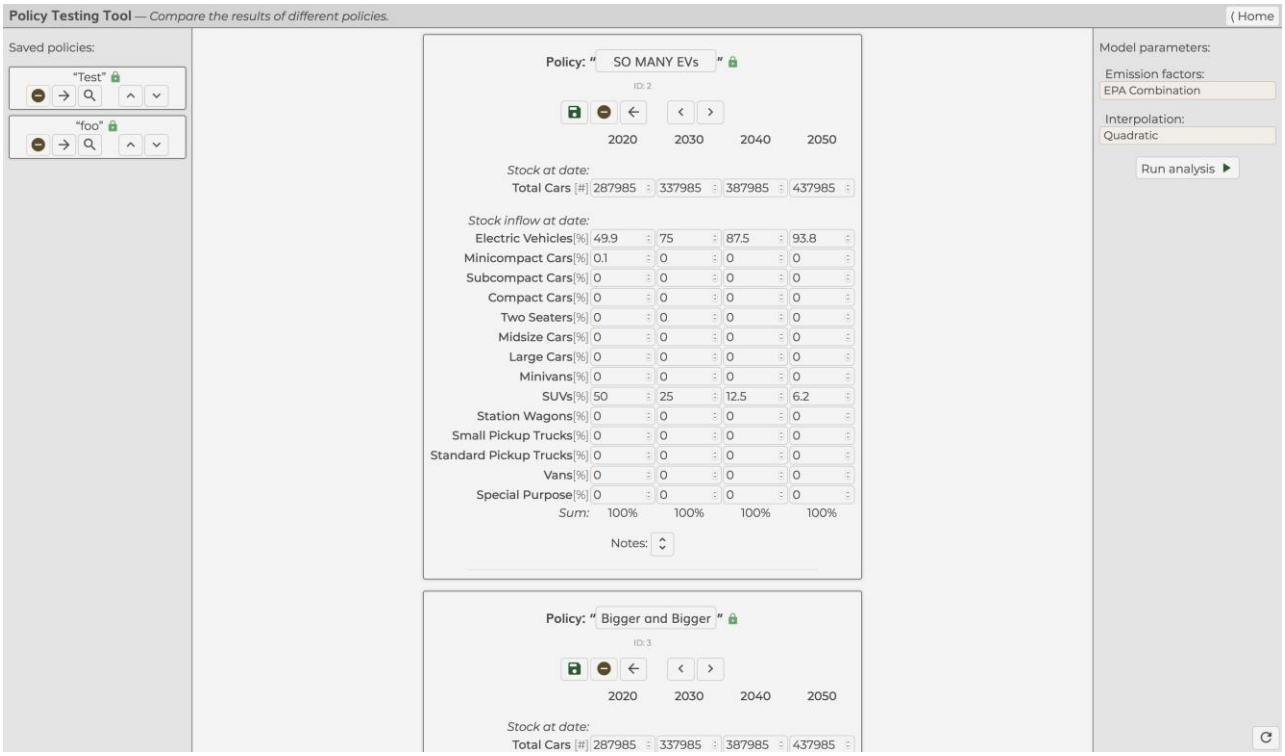


Figure 20. A screenshot of the policy tool, when editing policies.

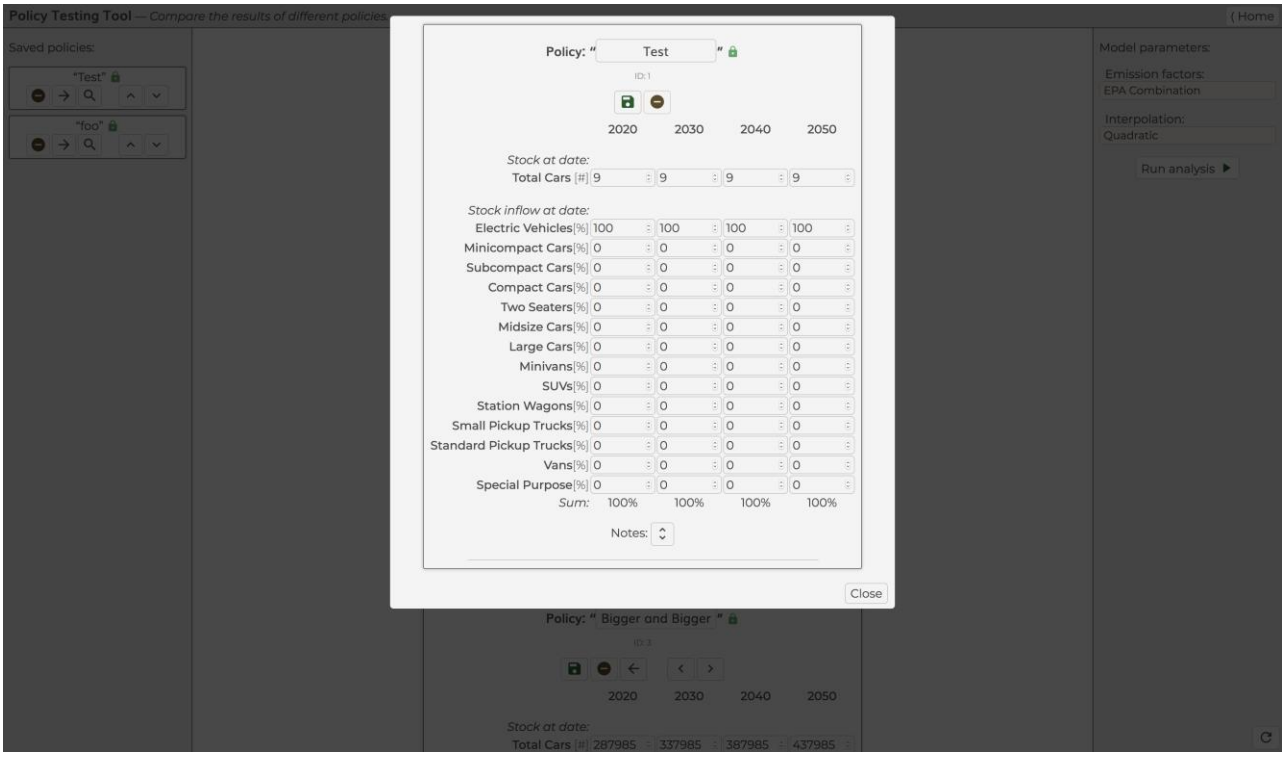


Figure 21. A screenshot of the policy tool, when editing a saved policy. This is useful mostly for comparisons, as saved policies do not have a full view.

After running the analysis, the results are displayed with the BAU forecast as a reference, as well as BAU $\pm 2\sigma$, the 95% confidence interval. This is shown in Figure 17.

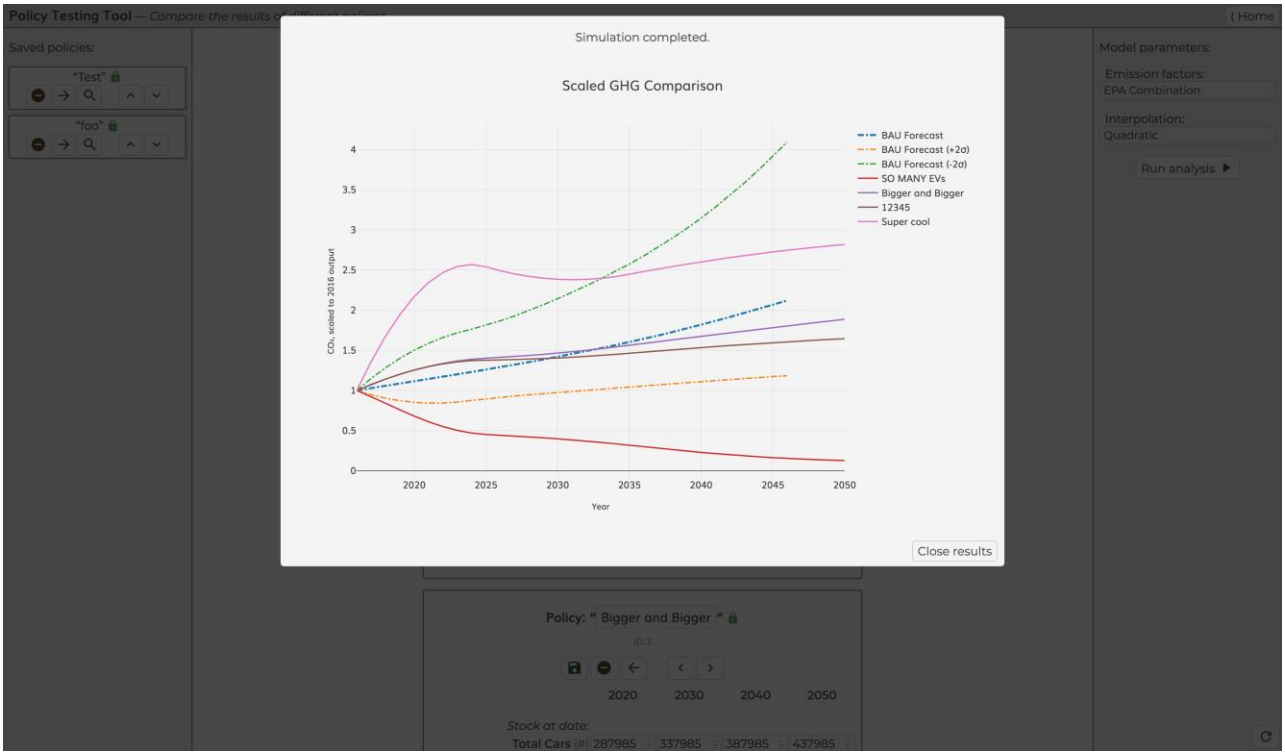


Figure 22. Sample results, displayed in the policy tool.

3.4.1.3. GHG Calculations

In order to calculate the greenhouse gas emissions, the following basic method is used:

$$\frac{\text{litres}}{\text{kilometer}} \cdot \frac{CO_2}{\text{litre}} \cdot \text{kilometres driven} \cdot \text{number of cars} = \text{total } CO_2$$

This calculation is done per vehicle class, per Policy. The results are then scaled to the 2016 results, which are the same for every Policy. This is done simply by dividing total CO_2 in every year by CO_2 in 2016. The GHG emissions in 2016 are defined to be 1, as a relative metric does not overstate the accuracy of the GHG point estimate. This is a compound effect of:

- Lack of correct VKT model
- Lack of increasing emissions from aging vehicles and changing emissions factors
- Errors in BAU model fitting

Following is the exact process used, in pseudocode. The variables are listed, then the process.

- ids: 1D array of Policy uniqueIDs.
- stocks: 2D array of vehicle stock counts. First dimension is Policy, second is stock count.
- compositions: 3D array of vehicle stock composition. First dimension is Policy, second is vehicle class, third is year.
- lpk: 2D array of vehicle efficiencies. First dimension is vehicle class, second is year.
- emi: 2D array of emissions factors. First dimension is vehicle class, second is year.
- VKT_SUM: sum of VKT from every TAZ

```
interpolate vehicle composition and stock over years
ghg := allocate 3D array (Policy, stock, year)

for i in [0, number of policies):
    ghg[i,:,:] = composition[i,:,:] * VKT_SUM * lpk * emi * stocks[i]

ghg_sum = 2D array, result of summation of ghg over second dimension
return ghg_sum
```

Note that * here indicates elementwise multiplication.

As vehicle stock is interpolated across all years, if one wanted to change the emission factor for future years (a simple approximation to improving emissions standards), this would be a relatively simple adjustment. In fact, they could rather easily be added for every year. However, if one wanted to stack vehicle age and apply different emissions standards based off of that, it would require completely rewriting the GHG calculation routine.

3.4.2. Data Sources

The map tool uses data exclusively from the portion of the project for remapping data to the TAZ level. All validation that applies to that data is the same here. For the correlation, the Pearson correlation coefficient was chosen. This function has not been checked for completeness. A complete list of variables that were planned is listed in Table 3, below.

The policy analysis tool uses various sources.

1. A VKT estimate from the City of Surrey. However, the desired VKT (for each vehicle class, for a whole day) is not available. Currently only an estimate for a particular time of day is available.
2. Vehicle stock forecast and current vehicle stock, internally created.
3. Vehicle emissions factors, from the EPA.

Table 3: List of planned variables for the map tool.

Data Source	Variable Name
City of Surrey	Population
	Employment
	Private Dwellings

Census Canada	Median age Total private dwellings Total private dwellings—single Total private dwellings—apartment Total private dwellings—other Median private household size Average private household size Average census family size Total number of families Median population income Median household income Percentage of people in low-income Total people in low-income Percentage households spending ≥30% of income on shelter costs Total households spending ≥30% of income on shelter costs Median value of dwellings [2006 $] Average value of dwellings [2006 $] Average value of rented dwellings [2006 $] Average value of owned dwellings [2006 $] Percentage of population with postsecondary Total population with postsecondary Labour force participation rate Employment rate Work trip modeshare—vehicle driver Work trip modeshare—vehicle passenger Work trip modeshare—public transit Work trip modeshare—walking Work trip modeshare—biking Work trip modeshare—other Commute destination share—within census subdivision Commute destination share—different census subdivision Commute destination share—different censusdivision
Translink	Number of commutes < 15min Number of commutes ≥ 15min, < 60min Number of commutes ≥ 60min

3.4.3. Architecture

The tool uses a client-server architecture, on a local machine (though there is no reason why it wouldn't work over the internet). The frontend is written using web technologies, and the backend is written in Python.

The frontend is written using HTML + SCSS + JavaScript. The JavaScript framework, "Vue", is used in order to make interface design easier. In addition, the JavaScript libraries, "Plotly" and "Leaflet", are used, both for data visualization. No data is stored in the frontend, and the GHG calculations are not done here. However, the calculation of correlations is done here. Since the tool uses the JavaScript template syntax, it will **not** work in Internet Explorer. It should work in all other major browsers.

The backend is written with Python (version 3.6+) using the libraries, "Flask","numpy", "scipy", and "sqlalchemy".

Communication with HTTP is done using the Flask web server framework. It provides several URLs on the localhost (port 5000 by default) to the frontend for communication. It can be accessed at http://localhost:5000/.

Exact installation steps are included in the README.mdfile in the source code.

3.4.4. Next Steps

The following improvements are suggested for the policy analysis tool:

- Full documentation of functions and limitations of tool
- Integration of improved VKT data
- User testing period
- Interface to input changing emissions factors
- Listing of the emissions assumptions
- Geography VKT use (currently only a sum is used)
- Option to fill a Policy from the BAU forecast, for comparison
- Visualization of the stock composition of Policies
- Reporting for each policy the breakdown of which vehicle classes are contributing how much CO_2.

The following improvements are suggested for the map tool:

- Integrate the remaining remapped data
- Test and use the correlation view
- Remove variables that are uninteresting or redundant.

4. Summary

4.1. Summary of Successes and Relevance to Social Good

In this introductory project, we were able accomplish basic goals of classifying, re-basing, analysing, forecasting and visualizing vehicle stock registration data to develop a more fulsome understanding of transportation energy and emissions in Surrey. Succinctly we:

- Developed and successfully applied a comprehensive classification schema to historic vehicle registration data (93% success rate);
- Rebased three disparate datasets into a common basis of understanding for exploratory analysis;
- Developed key vehicle stock insights especially with regards to passenger vehicle registration;

- Completed a preliminary effort to develop a business-as-usual vehicle forecast for the by vehicle type City; and
- Commenced the process to develop a data visualization for Surrey's ultimate use.

4.2. Summary of Shortcomings

Notwithstanding the successes above, the project did suffer from several shortcomings that should be addressed in the future. Notably:

- We did not 100% classify all vehicles into standard EPA vehicle classes and extensive validation has not been completed on the current completed classification.
- Rebasing of Census information was not completed in time for project completion and, as such, these data sets were not incorporated into the final visualization tool or into the exploratory analysis
- Rebasing could not be completed through an automated process and as Census specification and vector names varies across the years. As such, standardization required extensive "manual" adjustments.
- Exploratory analysis did not incorporate the demographic variables as discussed above
- Hypothesis testing for exploratory analysis could not be completed in a fulsome manner due to a lack of unique vehicles ids between years
- Vehicle stock forecasting suffered from a paucity of historic data to build forecasts
- There was additionally, a need to account for geographic effects w/o necessarily interpreting them and there was no effort to account for effects of vehicle class
- Due to a lack of available information, omitted variables may have affected accuracy of the forecast
- For all regression models, there is a need for future values of independent variables
- The policy analysis and visualization tool were hamstrung do a lack of input information and the complexity required to build and deploy this tool.

4.3. Critical Next Steps

- Future efforts for classification should achieve a higher success rate through a more elegant classification routine such as NLP or machine learning
- Future iterations of this project should allocate more effort into data standardisation such that all input data can be incorporated into final products
- Smaller sets of census data should be used to expedite the rebasing process
- Vehicle stock ids should be included with future releases of ICBC data
- A vehicle aging model should be developed to better understand vehicle ownership dynamics
- Regression Modelling of Vehicle Stock with Demographic Variables
- More Advanced Modelling of Vehicle Stock (Markov Chain, Stock-Flow)
- Validate and Redevelop Vehicle Stock Model with More Data
- Compute GHG Emissions based on Vehicle Stock Forecasts once Transportation Demand Data is Made Available
- With regards to the policy analysis tool the following is recommended:
  - Full documentation of functions and limitations of tool
  - Integration of improved VKT data
  - User testing period
  - Interface to input changing emissions factors
  - Listing of the emissions assumptions
  - Geography VKT use (currently only a sum is used)
  - Option to fill a Policy from the BAU forecast, for comparison
  - Visualization of the stock composition of Policies
  - Reporting for each policy the breakdown of which vehicle classes are contributing how much CO_2.

5. Appendices

5.1. Dead Ends- Exploratory Analysis

Querying the Commercial Vehicle Stock

Although it was not part of the initial project scope, commercial vehicle stock was also investigated on the side. Trends in commercial vehicle stock were not obvious when applied the same analysis with passenger vehicles. We attempted to normalize commercial vehicles by employment within each TAZ. As shown in Figure 19, spatial visualization of the normalized values shows a complex relationship with no obvious spatial trends. Distribution of normalized commercial vehicle stock, in Figure 20**,** also indicated substantial weight of higher normalized values. With such a large dynamic range in normalized values, this may suggest that various vehicle ownership behaviors exist in the commercial vehicle stock.

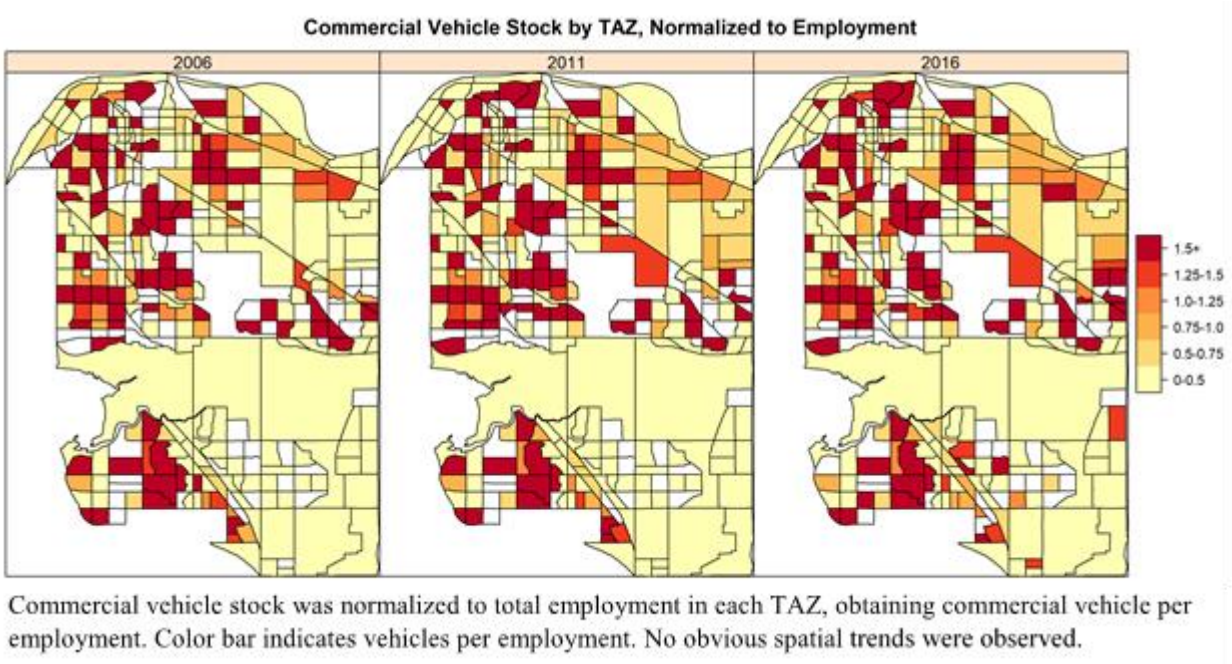
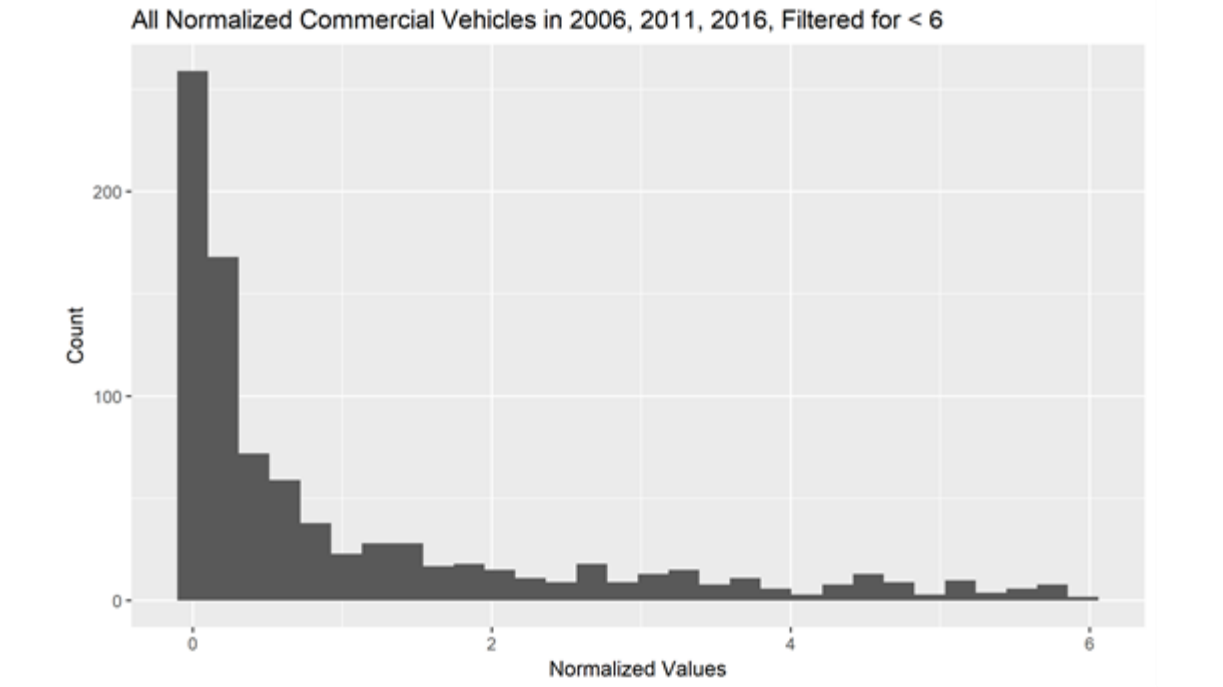


Figure 23. Spatially Visualizing Commercial Vehicles Per Capita in Each Year by TAZ.



30 bins used in histogram. Values greater than 6 were qualitatively filtered out to observe dynamics at relatively lower normalized values. Substantial distribution weight at higher normalized values were observed.

Figure 24. Histogram of Normalized Commercial Vehicle Values.

A scatterplot in Figure 21 of Commercial Vehicles vs. Total Employment in each TAZ further suggests that commercial vehicle stock may not be appropriately normalized by employment. Similar results were found even after normalizing by commercial and industrial employment independently. Next steps may be to query various commercial business types, such as acquiring data on commercial licenses, and cluster by various vehicle ownership behaviors

To work around the data gap of commercial vehicle behaviors, we applied high dimensional clustering on the original ICBC dataset for commercial vehicles only. We aimed to output data-driven clusters to hopefully identify more obvious trends between commercial vehicle stock and employment. Since the original ICBC dataset contained various data types, including continuous and categorical variables, clustering by Gower distance was used.[1] As shown in Figure 20, a Silhouette Width analysis of optimal cluster number was observed to be highest at k=2 clusters. At k=2 clusters, suggests that our clustering failed and no meaningful clusters would contribute to finding many commercial vehicle ownership behaviors. Due to the constraint of time, further analysis was not conducted.

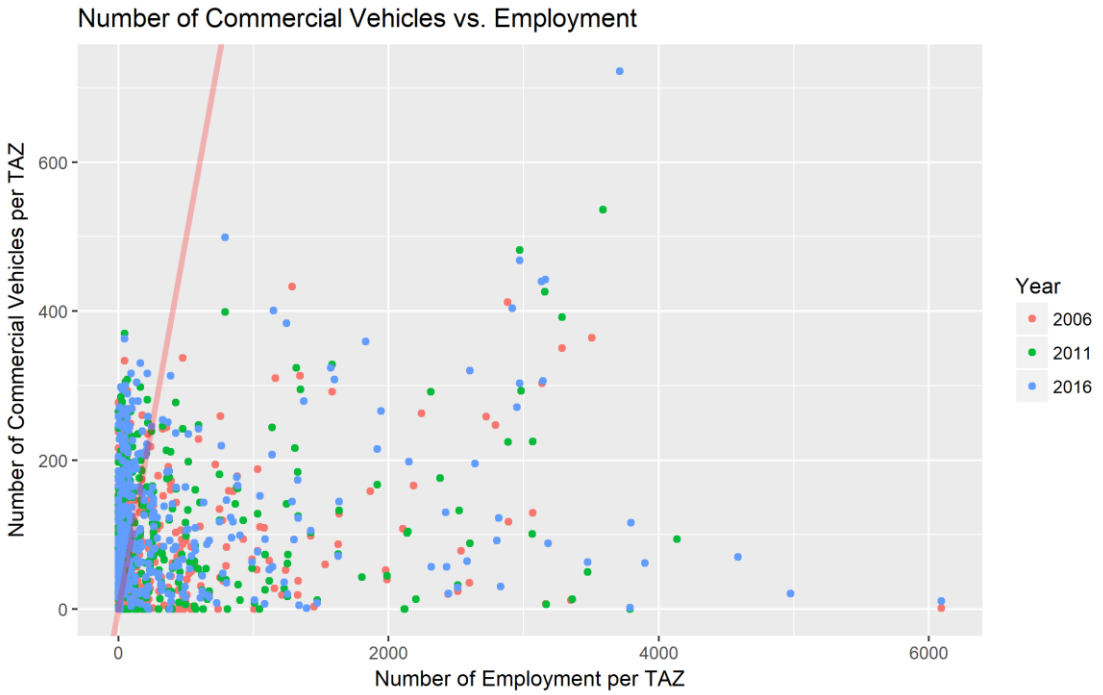
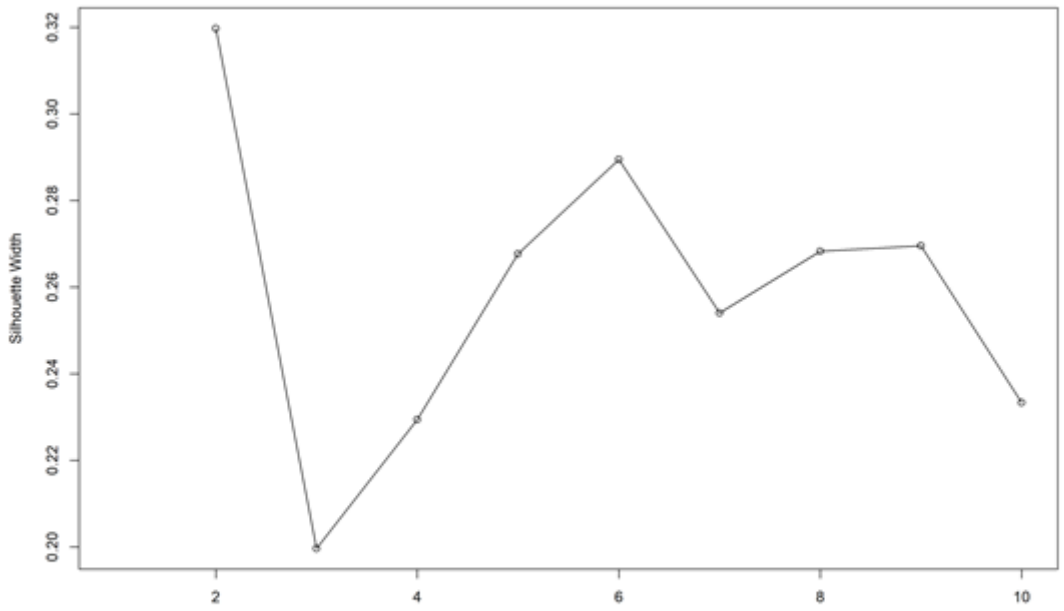


Figure 25. Number of Commercial vs Employment vehicles. No trends were observed. Visualization indicates that normalizing by commercial vehicles by employment may be inappropriate.



Optimal cluster was found to be for k=2. No meaningful clusters were found. Clustering analysis was unsuccessful.

Figure 26. Silhouette Width Analysis for Optimal Gower Distance Clustering for k=2 to k=10.

Developing a Vehicle Aging Model

It is important to model a unique vehicle stock dynamic that vehicles enter our system, either new to market or inflow from external sources, and remain in the system to age (5+ years) before retirement. To further elucidate the aging dynamics, we attempted to model survivors by a Generalized Logistic Mixed-Effects Model (GLMM); however, due to lack of unique identifiers, development was not pursued. We also attempted to identify duplicate vehicles by model year and vehicle class to simplify our assumptions and trace survivability. Initial analysis suggested that vehicle inflow may exceed vehicle outflow in various model years and vehicle classes that may lead to oversimplification and error of any model results. Due to time constraints, further analysis and development was not pursued.

References

[1]Wicked Good Data – r. "Clustering Mixed Data Types in R". Article published in Jun 21 2016. Accessed on Aug 30 2018. Online URL: https://www.r-bloggers.com/clustering-mixed-data-types-in-r/