ESTIMATING GREENHOUSE GAS EMISSIONS OF ELECTRIC DELIVERY TRUCKS

by
MERT ÖZÇELİK

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabancı University
July 2024

ABSTRACT

ESTIMATING GREENHOUSE GAS EMISSIONS OF ELECTRIC DELIVERY TRUCKS

MERT ÖZÇELİK

INDUSTRIAL ENGINEERING M.Sc. THESIS, JULY 2024

Thesis Advisor: Asst. Prof. Tuğçe Yüksel

Thesis Co-advisor: Asst. Prof. Sinan Yıldırım

Keywords: vehicle electrification, battery electric delivery trucks, greenhouse gas emissions, machine learning, simulation

In this study, we investigate the regional differences in emission benefits of battery electric delivery truck electrification. In this regard, we build a simulation framework to quantify the regional differences in the use phase emissions across the United States. A vital part of our framework is the machine learning model to predict the unit energy consumption of a battery electric delivery truck which is based on real world driving data. Using our framework, we perform two case studies to quantify the effect of ambient temperature and driving profile on the use phase emissions, respectively. In the first case study, we observe that our machine learning model can capture the increase in energy consumption at low temperatures quite well, however more data is needed to predict high temperature effects. As expected, the emissions are lower in regions where electricity production is cleaner. In the second case study, we observe that our framework can differentiate between the energy consumption under aggressive and gentle driving profiles.

ÖZET

ELEKTRİKLİ TESLİMAT KAMYONLARININ SERA GAZI SALIMLARININ TAHMİN EDİLMESİ

MERT ÖZÇELİK

ENDÜSTRİ MÜHENDİSLİĞİ YÜKSEK LİSANS TEZİ, TEMMUZ 2024

Tez Danışmanı: Dr. Öğr. Üyesi Tuğçe Yüksel

Tez Eş Danışmanı: Dr. Öğr. Üyesi Sinan Yıldırım

Anahtar Kelimeler: araç elektrifikasyonu, elektrikli teslimat kamyonları, sera gazı salımları, makine öğrenmesi, benzetim

Bu çalışmada, elektrikli teslimat kamyonu elektrifikasyonunun sağlayacağı emisyon azaltımındaki bölgesel farklılıklar incelenmiştir. Bu kapsamda, Amerika Birleşik Devletleri genelinde elektrikli araç şarjından kaynaklı salımlardaki bölgesel farklılıkların niceliksellleştirilmesi için bir benzetim çerçevesi geliştirilmiştir. Çerçevenin çok önemli bir parçası, elektrikli teslimat kamyonunun birim enerji tüketimini tahmin etmek için geliştirdiğimiz gerçek sürüş verilerine dayalı makine öğrenmesi modelidir. Bu çerçeve kullanılarak, hava sıcaklığının ve sürüş profilinin şarj kaynaklı salımlar üzerindeki etkisinin niceliksellleştirilmesi için iki vaka çalışması gerçekleştirilmiştir. İlk vaka çalışmasında, makine öğrenmesi modelinin düşük sıcaklıklardaki enerji tüketimi artışını yakalayabildiği, yüksek sıcaklıklarda doğru tahminler yapabilmek için ise daha fazla veriye ihtiyaç olduğu görülmüştür. Beklendiği üzere, elektriğin yenilenebilir kaynaklardan üretildiği bölgelerdeki salımlar daha düşüktür. İkinci vaka çalışmasında, çerçevenin agresif ve yumuşak sürüş profilleri altındaki enerji tüketimini ayırt edebildiği görülmüştür.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisors Asst. Prof. Tuğçe Yüksel and Asst. Prof. Sinan Yıldırım for their invaluable guidance. Also, I would like to thank my friend Khosro Parviz Naderi Varandi for his genuine help about LaTeX and Python. Besides, I would like to thank my loved ones for their endless support throughout this journey.

Further, I would like to thank my jury members Prof. Bülent Çatay, Assoc. Prof. Öznur Taştan, and Assoc. Prof. Mustafa Gökçe Baydoğan for accepting to be in my thesis jury and for devoting their precious time to evaluate my study.

Last but not least, I would like to thank Robert Prohaska for providing the NREL Fleet DNA data which is at the core of this study.

Dedication page
To all women who illuminate their paths with science...

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES xi

1. INTRODUCTION 1

2. DATA ANALYSIS AND TRIP IDENTIFICATION 5

2.1. Source Data 5

2.2. Data Pre-processing and Feature Extraction 9

2.2.1. Initial Trip Identification Algorithm 9

2.2.2. Modified Trip Identification Algorithm 13

2.2.3. Feature Extraction 17

2.3. Descriptive Statistics of the Post-Processed Trips 23

3. MACHINE LEARNING MODEL FOR UNIT ENERGY CONSUMPTION 26

4. SIMULATION 34

4.1. Effect of Ambient Temperature on Use Phase Emissions 35

4.1.1. Results 40

4.2. Effect of Driving Profile on Use Phase Emissions 43

5. CONCLUSION 45

BIBLIOGRAPHY 47

APPENDIX A 52

LIST OF TABLES

Table 1.1. Summary of studies about the energy consumption modelling of delivery vehicles 3

Table 2.1. Smith-Newton Delivery Truck Vehicle Specifications 6
Table 2.2. Relevant Parameters for the Study 6
Table 2.3. States the NREL Fleet DNA Data was Collected from along with the Collection Period and Truck Identification Numbers 8
Table 2.4. Trip Features 17
Table 2.5. Post-Processing 23

Table 3.1. Predictors of Random Forest 27
Table 3.2. Predictors of LASSO 28
Table 3.3. Range of each Predictor in the Train and Test Datasets 29
Table 3.4. Hyperparameter Values for LASSO 30
Table 3.5. Set of Possible Hyperparameter Values for Random Forest 31
Table 3.6. The Best Hyperparameter Combination for Random Forest 32
Table 3.7. Gini Importance Values for Random Forest 32
Table 3.8. Summary of Error Statistics for the Machine Learning Models . 33

Table 4.1. Synthetic Data for Case Study 2 44
Table 4.2. Average annual energy consumption and emissions for aggressive and gentle driving profiles 44

Table A.1. Whole Set of Parameters for the NREL Fleet DNA Data 52
Table A.2. Regression Coefficients for LASSO 53
Table A.3. Marginal Emission Factors for FRCC [kg-CO_2/MWh] 54
Table A.4. Marginal Emission Factors for NPCC [kg-CO_2/MWh] 55
Table A.5. Marginal Emission Factors for MRO [kg-CO_2/MWh] 56
Table A.6. Marginal Emission Factors for RFC [kg-CO_2/MWh] 57
Table A.7. Marginal Emission Factors for SERC [kg-CO_2/MWh] 58
Table A.8. Marginal Emission Factors for SPP [kg-CO_2/MWh] 59
Table A.9. Marginal Emission Factors for TRE [kg-CO_2/MWh] 60

Table A.10.Marginal Emission Factors for WECC [kg-CO_2/MWh] 61

LIST OF FIGURES

Figure 2.1. Distribution of Ambient Temperature for each State based on National Solar Radiation Database (NSRDB) data.................... 9
Figure 2.2. Example of a trip with contradictory characteristics........... 13
Figure 2.3. Distribution of Average Speed for Selected Vehicles............ 14
Figure 2.4. Examples of spikes in latitude, longitude, and state of charge measurements.. 21
Figure 2.5. Average monthly energy consumption for each state. The black line labelled "ALL" shows the energy consumption based on all states. 24
Figure 2.6. Distributions of daily number of trips, duration per trip, average ambient temperature per trip, unit energy consumption per trip, and distance travelled per trip.. 25

Figure 3.1. Average Mean Squared Error versus Cross Validated Lambda Value for LASSO 31

Figure 4.1. Illustration of the simulation framework 34
Figure 4.2. Average annual energy consumption for each county in the US. The energy consumption values are reported in Wh/km, and they are averages of 30 Monte Carlo runs..................................... 41
Figure 4.3. Average annual use phase emissions for each county in the US. The emissions values are reported in g-CO_2/km, and they are averages of 30 Monte Carlo runs..................................... 41
Figure 4.4. Unit Energy Consumption versus average ambient temperature based on the Random Forest model 43

1. INTRODUCTION

Global warming is an ongoing environmental problem: July 2023 has been the hottest month ever recorded (NASA, 2023). Global warming is caused by the accumulation of greenhouse gases in the atmosphere. Fossil fuels account for 75% of greenhouse gas (GHG) emissions worldwide (UN, 2024), and the transportation sector is one of the largest sources of fossil fuel emissions (How et al., 2020). As the global warming continues to become a severer concern, strategies to reduce the greenhouse gas emissions from the transportation sector are coming into prominence. At this point, electrification of transportation may be a viable strategy (Chen and Fan, 2013). Battery electric vehicles (BEVs) attract significant research attention mainly because they emit zero tailpipe emissions (Pan et al., 2023). If the GHG emissions from the generation of electricity which is used for charging can also be minimized, BEVs may be a promising alternative to internal combustion engine vehicles.

Electrification of light-duty passenger vehicles has been studied extensively in the literature. Many scholars have investigated the emission reduction potential of light-duty passenger vehicles (Zivin et al., 2014; Yang et al., 2014; Alvarez et al., 2014; Maia et al., 2015; Li et al., 2015; Zhang and Yao, 2015; Tamayao et al., 2015; Onat et al., 2015; Yuksel and Michalek, 2015; Archsmith et al., 2015; Yuksel et al., 2016; Fiori et al., 2016; Woo et al., 2017; Genikomsakis and Mitrentsis, 2017; Fiori et al., 2018; Jiménez et al., 2018; Wu et al., 2019; Vepsäläinen et al., 2019; Sun et al., 2019; Fernández et al., 2019). Although the research on the electrification of light-duty vehicles continues (Al-Wreikat et al., 2021; Ahmed et al., 2022; Perugu et al., 2023; Pan et al., 2023; Hull et al., 2023), reduction of emissions in the other segments of road transport is also necessary. Medium and heavy duty vehicles make up 30% of the road transport emissions, however their electrification has been relatively slower and there is an increasing shift of focus towards this segment (EVO, 2024; IEA, 2024). In this segment, electrification of medium-duty delivery vehicles have showed an emerging potential, due to their increasing amount of use with online shopping (Woody et al., 2021), and relatively smaller size compared to other commercial vehicles which makes them easier and cheaper to electrify (Woody et al., 2022a). In addition, the medium-duty delivery vehicles are usually used in predetermined routes which provides more flexibility to optimize operational details which can help with range anxiety and charging issues. Several companies have started testing and operating electric vehicles for their delivery operations (Woody et al., 2021). These trends point out the importance of understanding the implications and benefits of electrifying the delivery operations.

One of the important aspects of assessing the emission benefits of BEVs is to predict the energy consumption accurately. A precise prediction of energy consumption will not only help alleviating the so-called range anxiety issue, but it will also allow for a more accurate forecast of the GHG emissions. Previous studies have shown that the energy consumption is sensitive to many factors such as driving profile (Karabasoglu and Michalek, 2013), road profile (Prohaska et al., 2016), and ambient temperature (Yuksel and Michalek, 2015). As the energy consumption varies regionally depending on these factors, the emission benefits which will be obtained from BEVs may also vary regionally. Therefore, it is of utmost importance to take into account the effects of these factors while modelling the energy consumption of a BEV. Among these factors, ambient temperature plays a critical role. For example, Yuksel and Michalek (2015) has shown that the driving range of a BEV may decrease by up to 40% on too cold winter and hot summer days. Thus, it is important to consider the effect of ambient temperature while assessing the emission benefits of electrified medium-duty delivery vehicles as well.

Energy consumption modelling of BEVs has been studied extensively in the literature. However, most of the studies have focused on passenger vehicles; the studies on delivery vehicles are sparse. In Table 1.1, we categorize the existing studies about the energy consumption modelling of delivery vehicles based on the energy consumption modelling approach, whether or not they consider the effect of ambient temperature on the energy consumption, and whether or not they analyze the regional variation in GHG emissions. The energy consumption modelling approach is either constant, physics-based, or data-driven. If the energy consumption approach is constant, a constant value is assumed for the unit energy consumption which might be based on metrics such as depth of discharge (DOD) or laboratory test results such as drive cycles. A physics-based model predicts the energy consumption based on physical laws and vehicle dynamics (Maia et al., 2015). A data-driven model is generally based on data analysis techniques or statistical relationships, and it does not necessarily consider the physical relations between the input variables and the energy consumption (Pan et al., 2023). It may be constructed based on machine learning models. The 'Temperature?' and 'Emissions benefits?' columns of the table indicate whether or not the study considers the effect of ambient temperature on the energy consumption and it discusses the regional variation in the emission benefits, respectively.

Table 1.1 Summary of studies about the energy consumption modelling of delivery vehicles

Study	**Energy Consumption Modelling**	**Temperature?**	**Emissions benefits?**
Lee and Thomas, 2017	Data-driven		✓
Giordano et al., 2018	Constant		✓
Yang et al., 2018	Physics-based		✓
Marmiroli et al., 2020	Physics-based		✓
Woody et al., 2021	Constant		✓
Burnham et al., 2021	Physics-based	✓	✓
Woody et al., 2022c	Physics-based	✓	✓
Woody et al., 2022b	Physics-based	✓	✓
Woody et al., 2022a	Constant		✓
O'Connell et al., 2023	Constant		✓
Qiu et al., 2023	Data-driven	✓	
This study	Data-driven	✓	✓

To the best of authors knowledge, the studies which model the energy consumption of a medium-duty delivery vehicle using a data-driven approach and considering the effect of ambient temperature are sparse. In fact, the study by Qiu et al. (2023) is the only such study in Table 1.1. However, they do not analyze the emission benefits of the delivery trucks. In this study, we develop a data-driven model to predict the unit energy consumption of a medium-duty delivery vehicle and consider the effect of ambient temperature. In addition, we quantify the emission benefits of delivery vehicle electrification and discuss the regional variation in the emissions.

The contributions of this study to the literature are as follows:

- We develop a data-driven model to predict the energy consumption of a medium-duty delivery vehicle which accounts for factors such as ambient temperature, average speed, acceleration, and road profile.
- We quantify the emission benefits of medium-duty delivery vehicle electrification using simulation, and analyze the effect of ambient temperature on the emission benefits, and
- We capture the regional variability in the emissions through random trip generation and simulation of energy consumption.

The rest of this thesis is organized as follows: In Chapter 2, we introduce the source data, and devise an algorithm to identify trips from the source data . In Chapter 3, we develop a machine learning model for unit energy consumption. In Chapter 4, we build a simulation framework to quantify the regional variability in greenhouse gas emissions, and present two case studies. In Chapter 5, we summarize our conclusions.

2. DATA ANALYSIS AND TRIP IDENTIFICATION

Our purpose is to construct an empirical model to predict the unit energy consumption of a medium-duty battery electric delivery truck. As our source data, we utilize the National Renewable Energy Laboratory's (NREL) Fleet DNA data (NREL, 2017a) which consists of GPS- and sensor-based measurements collected from medium-duty battery electric delivery trucks. We aim to make the energy consumption predicition at an aggregate level, so we devise an algorithm to identify trips of the trucks from the NREL Fleet DNA data. Based on the set of trips, we develop a Random Forest model for unit energy consumption prediction. The rest of this chapter is organized as follows: In Section 2.1, we provide detailed information about the NREL Fleet DNA data. In Section 2.2, we perform data pre-processing, feature extraction, and post-processing. Finally, in Section 2.3, we present some descriptive statistics about the post-processed set of trips.

2.1 Source Data

The NREL Fleet DNA data was collected from 200 Smith-Newton trucks. Some of the vehicle specifications for Smith-Newton trucks are presented in Table 2.1 (NREL, 2017b). The data was collected over a period of around two years, from 2013 to 2015, corresponding to 738 operating days. However, not all trucks operate on all days, and different trucks may have a different number of operating days. The measurements in the NREL Fleet DNA data were recorded in a second-by-second manner, and there are measurements for 54 different parameters. NREL (Prohaska et al., 2016) tested the Smith-Newton trucks on road to

- compare battery electric trucks to internal combustion engine trucks in terms of energy consumption, charging requirements and cost, and hence learn more about the energy savings and emission benefits potential of electric trucks in commercial delivery operations.

- obtain data about the battery performance and charging of electric trucks, and possibly understand which charging strategies may be better in terms of emission reduction

Table 2.1 Smith-Newton Delivery Truck Vehicle Specifications

Specification	**Value/Range**
Curb weight	$9,700$-$10,200$ [lbs]
Payload	$12,324$-$16,200$ [lbs]
Advertised range	< 241 [km]
Electric top speed	50 [mph]
Battery capacity	80 [kWh]

The measurements in the NREL Fleet DNA data were stored daily. For a particular operating day of a truck, we have records for the entire day. In other words, the records include not only the times when the truck is moving but also the times when the truck is parked or being charged. As we aim to predict the energy consumption resulting from the driving of the truck, we first need to identify the trips (i.e., times at which the truck is actually moving) from the NREL Fleet DNA data. For this purpose, we devise a trip identification algorithm which is presented in Section 2.2. Out of the 54 parameters in the NREL Fleet DNA data, only 10 of them are relevant for this study. The relevant parameters are described in Table 2.2. The parameter names in Table 2.2 are the same as in the NREL Fleet DNA data. Whole set of parameters for the NREL Fleet DNA data is given in Table A.1.

Table 2.2 Relevant Parameters for the Study

Parameter Name	**Description**	**Symbol**	**Unit**
'BMU_Mode_SYS'	Battery management mode	bmu	-
'Battery_Current_SYS'	Battery current	I	[A]
'Battery_Voltage_SYS'	Battery voltage	V	[V]
'GPS_Altitude'	Altitude	h	[m]
'GPS_Latitude'	Latitude	lat	[decimal degrees]
'GPS_Longitude'	Longitude	long	[decimal degrees]
'GPS_Speed'	Speed	v	[mph]
'RD_Ambient_Temp_degC'	Ambient temperature	T	[degC]
'SOC_SYS'	State of charge	soc	[%]
'Timestamp'	Timestamp	t	[s]

Below are some remarks about Table 2.2:

- All of the parameters are data vectors. All vectors give corresponding values at each time step the measurements are taken, with a resolution of one second in general.
- The battery management mode does not have a unit, and it has three possible values: 0, 1, and 2. These values denote that the motor is off, the battery is being charged, and the motor is on, respectively.
- The battery current shows the amount of current drawn into/from the battery. There is an important detail about the sign convention: Contrary to the common practice, the positive current represents charging, and negative current represents discharging.
- The latitude and longitude are measured up to 13 decimal places.
- The state of charge (SOC) is measured with 1% resolution, and it is recorded with no decimal places.

The NREL Fleet DNA data was collected from seven states in total. The states the data was collected from, collection period for each state, and the trucks operating at each state are presented in Table 2.3. The collection periods are provided by season and year combinations. The Winter season includes the months from November to March, the Summer season includes the monts from May to September, and the transition season (Trans) consists of April and September. We observe that all states except for California have data collected for all possible season and year combinations.

Table 2.3 States the NREL Fleet DNA Data was Collected from along with the Collection Period and Truck Identification Numbers

State	**Collection Period**	**Truck IDs**
California	Winter'13, Trans'13, Summer'13 Winter'14, Trans'14, Summer'14	113, 163, 214, 217, 282, 384, 385, 387, 396, 431, 432, 437
Illinois	Winter'13, Trans'13, Summer'13 Winter'14, Trans'14, Summer'14 Winter'15	429, 430
Maryland	Winter'13, Trans'13, Summer'13 Winter'14, Trans'14, Summer'14 Winter'15	330
New Jersey	Winter'13, Trans'13, Summer'13 Winter'14, Trans'14, Summer'14 Winter'15	414, 439, 441
New York	Winter'13, Trans'13, Summer'13 Winter'14, Trans'14, Summer'14 Winter'15	106, 167, 173, 426
Virginia	Winter'13, Trans'13, Summer'13 Winter'14, Trans'14, Summer'14 Winter'15	434
Washington	Winter'13, Trans'13, Summer'13 Winter'14, Trans'14, Summer'14 Winter'15	109, 150, 207, 215, 218, 275, 418

In Figure 2.1, we provide the distribution of ambient temperature for each state. Based on the data from the National Solar Radiation Database (NSRDB) (NSRDB, 2024) and the season and year combinations in Table 2.3, we provide the range of ambient temperature for each state throughout the data collection period. We observe that the temperatures of the states can go below -20 degC, and they can be as high as 40 degC. Based on the mean values of the box plots, the average temperature is around 15 degC.

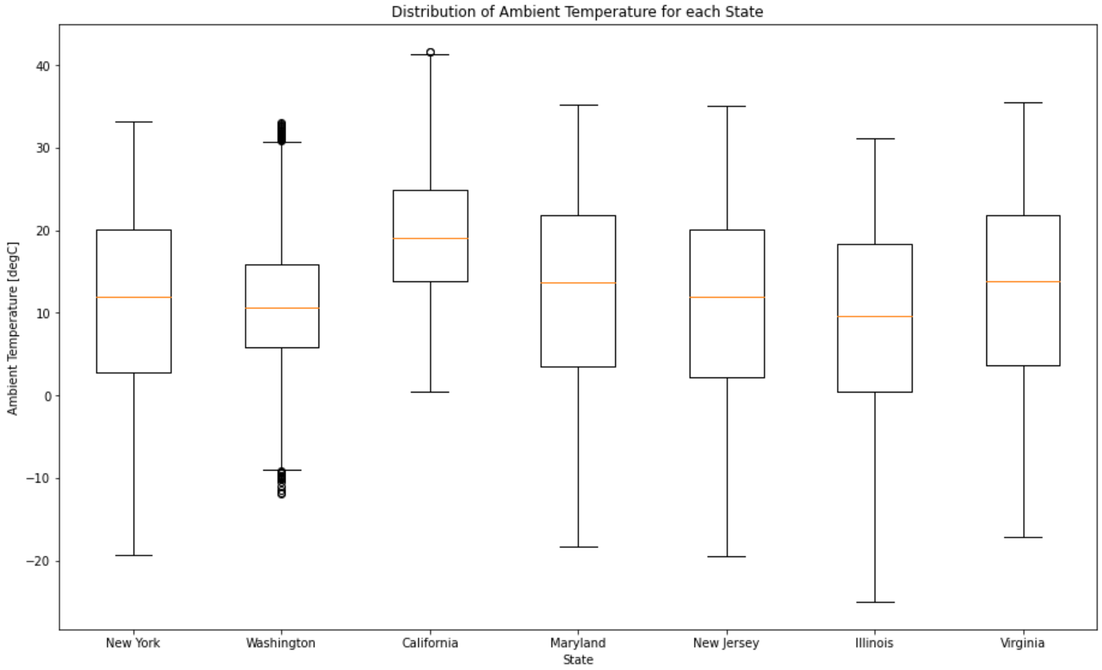


Figure 2.1 Distribution of Ambient Temperature for each State based on National Solar Radiation Database (NSRDB) data

2.2 Data Pre-processing and Feature Extraction

As we aim to make the energy consumption prediction at an aggregate level, we need to identify the trips of the trucks from the NREL Fleet DNA data. In this section, we first devise a trip identification algorithm. Then, we compute a set of features for the trips. Section 2.2.1, presents an initial trip identification algorithm. Section 2.2.2 discusses the motivation and proposes a modified trip identification algorithm. Finally, in Section 2.2.3, we perform feature extraction and compute a set of features for the trips.

2.2.1 Initial Trip Identification Algorithm

We identify the trips from the NREL Fleet DNA data by using Algorithm 1. Identifying a trip means identifying the beginning and end points of the trip. As we are working with data vectors, points correspond to indices of the vectors. Therefore, instead of point, the term *index* will be used hereafter. In Algorithm 1, we define trip beginning index as an index where the motor is on and speed is positive. Trip end index is defined as an index where the motor is not on (i.e., either the motor is off or the battery is being charged) and speed is zero for at least Ψ_1 seconds. There are two main reasons for introducing such a time lapse. First, we want our algorithm to be robust against erroneous measurements. For example, the GPS device for speed may malfunction, and record a value of zero although the speed is nonzero. We do not want the trip to end abruptly in that case. Second, we acknowledge that the speed may become zero temporarily due to natural occurrences such as stopping at a red light or encountering congestion. We do not want the trip to end abruptly in such cases either. We take the value of Ψ_1 as 300 seconds. Algorithm 1 has three vector inputs and one scalar input. The vector inputs are the vectors of battery management mode, speed, and timestamp. The scalar input is Ψ_1 which is defined as the minimum time lapse for trip completion. In line 2, we initialize the Boolean variable *searching_beginning_index* which indicates whether a beginning index or an end index is being searched. As stated in line 10, while searching a beginning index, we look for an index i where the motor is on and speed is positive. If such an index is found, we mark it as the beginning index of a trip, and set *searching_beginning_index* to False (lines 11-13). On the other hand, while searching an end index, we first find an index p such that the time lapse between p and the current index i is at least Ψ_1 seconds (line 18). Then, between lines 19 and 23, we check whether the motor is not on and speed is zero throughout the range from i to p. If this is the case, we mark i as the end index of the trip, and set *searching_beginning_index* to True to identify the next trip.

Once all trips are identified, we calculate the distance traveled, duration, and average speed for each trip using Equations (2.1)-(2.3) , respectively for $k = 1, \ldots, |\mathcal{K}_0|$

$$x_k \approx \sum_{i=b_k}^{e_k-1} \frac{v_i + v_{i+1}}{2} \cdot (t_{i+1} - t_i) \tag{2.1}$$

$$\tau_k = \left(t_{e_k} - t_{b_k}\right) \tag{2.2}$$

$$\bar{s}_k = \frac{x_k}{t_k} \tag{2.3}$$

where x_k, τ_k, and $\bar{s}_k$ are the distance travelled, duration, and average speed during trip k, respectively. $\mathcal{K}_0$ is the set of all trips such that $|\mathcal{K}_0| = 70,046$. For a particular trip k, b_k and e_k refer to the beginning and end indices of the trip, respectively. In equation (2.1), we estimate distance travelled using the trapezoidal rule, and v is in meters per second.

Algorithm 1 Identification of Trips

```
Inputs:   bmu, $v$, $t, \Psi_1$
$searching\_beginning\_index$ = True
$n = length($ bmu)
$trip\_tuples = [\,]$
$i = 1$                                                    ▷ current index
$b_k = 0$                                      ▷ beginning index of trip $k$
$e_k = 0$                                            ▷ end index of trip $k$
while $i \leq n - \Psi_1$ do
    if searching_beginning_index = True then
        if (bmu$_i = 2$)  and  ($v_i > 0$) then
            $b_k = i$
            $i = i + 1$
            searching_beginning_index = False
        else
            $i = i + 1$
        end if
    else
        $p :=$ index such that $t_p - t_i \geq \Psi_1$
        if (bmu$_{i:p} \neq 2$)  and  ($v_{i:p} = 0$) then
            $e_k = i$
            trip_tuples = trip_tuples + $\{(b_k, e_k)\}$
            $i = e_k + 1$
            searching_beginning_index = True
        else
            $i = i + 1$
        end if
    end if
end while
```

As we build the trip identification algorithm from scratch, we want to verify that our algorithm can divide the daily records into trips in a consistent way with the NREL's findings. For this reason, we use *average daily distance per truck* as a benchmark metric. The reference value for this metric is provided as 41.7 km in the Smith Newton Vehicle Performance Evaluation report (NREL, 2017b). At this point, it is important to emphasize a limitation of our study: Although the NREL Fleet DNA data consists of 200 trucks, we are provided a partial dataset comprising only 30 trucks. Thus, we need to customize the reference value to our dataset. We compute the customized reference value by using Algorithm 2. For each combination of operating day d and truck j, we calculate $\mathbf{X}_{dj}$ which is the distance travelled by truck j on day d. At the end of an operating day d, we calculate the average distance travelled by a truck on day d (x_d) by dividing the total distance travelled by the number of trucks performing trip on day d. At the end, the customized reference value is equal to the mean of $\mathbf{x}$. We obtain the customized reference value as 40.8 km. One detail regarding the NREL Fleet DNA data is that the speed vectors (v) may contain null values. If the speed is null at a particular index, we remove that index from both the speed and timestamp vectors while calculating $\mathbf{X}_{dj}$. If the speed vector is all null, we assume $\mathbf{X}_{dj}$ as 0.

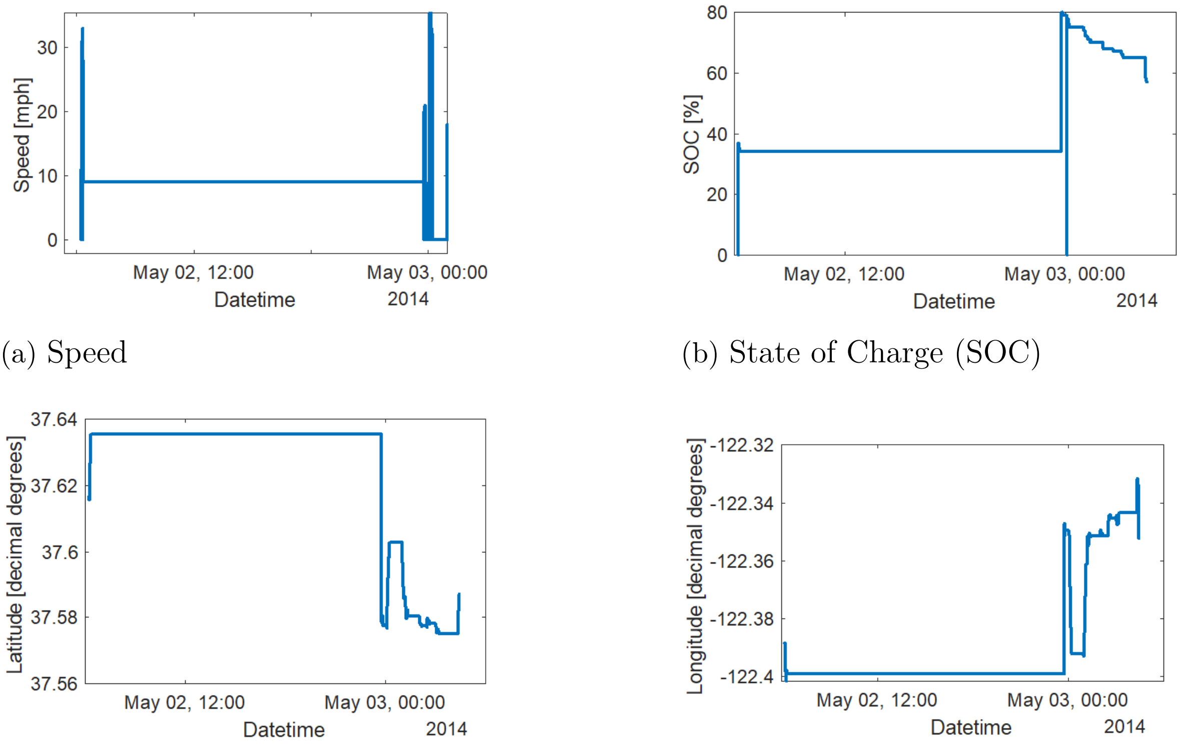
Algorithm 2 Customized Reference Value

```
$\mathbf{X}$ := Daily distance travelled matrix
$\mathbf{x}$ := Average daily distance vector
for each operating day $d$:
    for each truck $j$:
        Calculate $\mathbf{X}_{dj}$, the distance travelled on day $d$ by truck $j$
    $n_d$ := Number of trucks performing trip on day $d$
    $x_d = sum(\mathbf{X}_d)/n_d$
customized_reference_value = $mean(\mathbf{x})$
```

We calculate the average daily distance per truck resulting from the trips in $\mathcal{K}_0$ as 41.5 km. One may find this result counterintuitive as we partition the daily records into trips, and obtain a larger value than the customized reference value (which is calculated by using the daily records directly, i.e., without any partitioning). However, it should be noted that we compute the *average* daily distance per truck, not the *total*. Given that the absolute difference between the customized reference value (40.8 km) and the distance resulting from the trips (41.5 km) is less than 1 km, we conclude that Algorithm 1 can divide the daily records into trips in a consistent way with the NREL's findings.

2.2.2 Modified Trip Identification Algorithm

Although Algorithm 1 can partition the daily records into trips in a consistent way with the NREL's findings, when we analyze the obtained trips in detail, we spot some trips which have contradictory characteristics within themselves. For example, there are trips with portions during which the speed is constant at a positive value for a long time, yet the SOC is also constant. In addition, the coordinates of the vehicle (i.e. latitude and longitude measurements) do not change either. On average, these trips have a duration of 14.03 h and an average speed of 3.85 mph. Such combinations of long duration and low average speed are unlikely. These observations indicate that the trip has ended at the beginning of the portion, yet the algorithm failed to determine it. An example of a trip with contradictory characteristics is shown in Figure 2.2. The speed is constant at 9 mph for more than 17 hours, yet the SOC, latitude, and longitude are also constant. If this trip were to be performed, constant SOC would mean that such a long trip is achieved with almost no energy consumption! All considered, we claim that Algorithm 1 is prone to identifying trips with contradictory characteristics which might stem from measurement errors.



(a) Speed

(b) State of Charge (SOC)

(c) Latitude

(d) Longitude

Figure 2.2 Example of a trip with contradictory characteristics

Given the trips with contradictory characteristics, we make some modifications to

Algorithm 1. These modifications are explained below:

- We hypothesize that the trips with contradictory characteristics may result from erroneous speed measurements. It is probable that the speed was not constant for such a long time, but it appeared to be so because of a malfunction in the GPS device. As a remedy, we introduce a new parameter, Ψ_2, which is defined as the maximum amount of time speed is allowed to be constant within a trip. With this modification, the average percentage of time speed is constant becomes 16.28%. In addition, the percentage of trips with duration greater than 1 hour drops from 6.23% to 0.89%.

- As visualized in Figure 2.3, another issue is that the average speed during some trips exceeds the top speed of the truck. Although the top speed of the truck is specified as 50 mph (Table 2.1), we observe average speed values as high as 91 mph. When we investigate the speed vectors of these trips, we see that they contain values greater than the top speed. To alleviate this issue, we introduce a new parameter, Ψ_3, which is defined as the maximum amount of time speed is allowed to exceed the top speed within a trip. After this modification, the highest average speed during a trip is only 58 mph.

- In line with the items above, we modify the definition of trip end index as follows: An index starting from which either the motor is not on for at least Ψ_1 seconds, or speed has been constant for Ψ_2 seconds, or speed has exceeded the top speed for Ψ_3 seconds. With this new definition of the trip end index, we aim to design an algorithm which will not identify trips with contradictory characteristics.

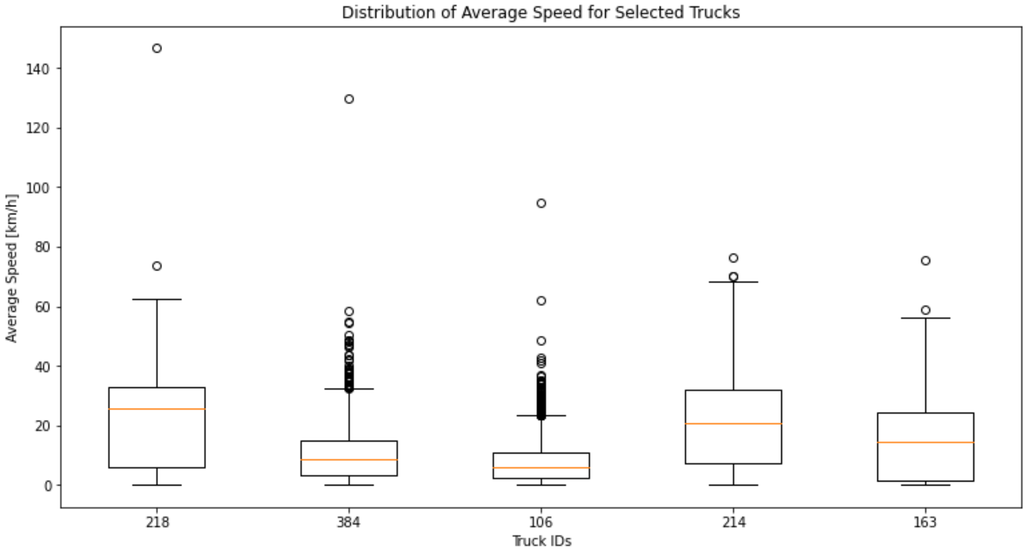


Figure 2.3 Distribution of Average Speed for Selected Vehicles

The modified algorithm is presented in Algorithm 3. We define trip beginning index in the same way as in Algorithm 1. However, we modify the definition of trip end index. We now have three cases for the trip end index. First, if the motor is not on for at least Ψ_1 seconds starting from the current index i, we check the last index speed was nonzero (u). If the time lapse between u and i is less than or equal to Ψ_4 seconds, we set the trip end index as i; otherwise, we set it as u. This is because it might be undesirable to have a trip in which speed is zero for a long time towards the end. We take the value of Ψ_4 as 30 seconds. Second, if speed has exceeded the top speed ($v_{\max}$) for Ψ_3 seconds starting from an index p', we set p' as the trip end index. Then, we look for the first index (u') starting from which speed is less than or equal to the top speed for at least 300 seconds. If such an index is found, we set it as the new value of i. Otherwise, we terminate the algorithm. Finally, if speed has been constant for Ψ_2 seconds starting from an index p'', we set p'' as the trip end index. Then, we look for the first index (u'') at which the speed changes. If such an index is found, we set it as the new value of i. Otherwise, we terminate the algorithm. When a trip end index is found, we add the trip tuple (b_k, e_k) to the list *trip_tuples*, and set the Boolean variable *searching_beginning_point* to True so that the next trip can be identified. If none of the three cases is satisfied, we increment the value of i, and continue searching for an end index. We take the values of Ψ_1, Ψ_2, and Ψ_3 as 300, 300, and 5 seconds, respectively. Let $\mathcal{K}_m$ be the set of all trips identified by Algorithm 3 where $|\mathcal{K}_m| = 93,262$.

Algorithm 3 Identification of Trips - Modified

```
Inputs: bmu, $v$, $t$
$v_{\max}$ := Top speed
$\Psi_1$ := minimum amount of time motor is required not to be on for trip completion
$\Psi_2$ := maximum amount of time speed is allowed to be constant within a trip
$\Psi_3$ := maximum amount of time speed is allowed to exceed the top speed within a trip
$\Psi_4$ := maximum amount of time speed is allowed to be zero at the end of a trip
$n = length($ bmu$)$; $trip\_tuples = [\ ]$; $searching\_beginning\_index$ = True
$i = 1$                                                        ▷ current index
$b_k = 0$                                                      ▷ beginning index of trip $k$
$e_k = 0$                                                      ▷ end index of trip $k$
while $i \leq n - max(\Psi_1, \Psi_2, \Psi_3, \Psi_4)$ do
    if searching_beginning_index = True then
        $i = i + 1$
        if $(\text{bmu}_i = 2)$ and $(v_i > 0)$ then
            $b_k = i$; searching_beginning_index = False
        end if
    else
        $p$ := index such that $t_p - t_i \geq \Psi_1$
        $p'$ := index such that $t_i - t_{p'} = \Psi_3$
        $p''$ := index such that $t_i - t_{p''} = \Psi_2$
        if $(\text{bmu}_{i:p} \neq 2)$ then                    ▷ Case 1: Motor is not on
            $u$ := last index before $i$ such that $v_u \neq 0$
            $e_k = i$ if $t_i - t_u \leq \Psi_4$ else $e_k = u$
            trip_tuples = trip_tuples + $\{(b_k, e_k)\}$; searching_beginning_index = True
            $i = i + 1$
        else if $v_{i-p':i} > v_{\max}$ then                   ▷ Case 2: Speed exceeds the top speed
            $e_k = p'$; $u'$ := first index after $i$ such that $v_{u':u'+300} \leq v_{\max}$
            if $\exists\ u'$ then
                $i = u'$
                trip_tuples = trip_tuples + $\{(b_k, e_k)\}$; searching_beginning_index = True
            else
                Stop
            end if
        else if $v_{i-p'':i} = v_{i-p''}$ then                 ▷ Case 3: Speed is constant
            $e_k = p''$; $u''$ := first index after $i$ such that $v_{u''} \neq v_i$
            if $\exists\ u''$ then
                $i = u''$
                trip_tuples = trip_tuples + $\{(b_k, e_k)\}$; searching_beginning_index = True
            else
                Stop
            end if
        else
            $i = i + 1$
        end if
    end if
end while
```

The average daily distance per truck resulting from $\mathcal{K}_m$ is 31.5 km. The difference between this value and the customized reference value (40.8 km) may seem significant, however, when we analyze $\mathcal{K}_m$ further, we observe that the trip characteristics have improved significantly. Only 0.89 % of the trips have duration greater than 1 hour, and the average percentage of time speed is constant within a trip is 16.28 %. Previously, we had trips whose average duration was 14.03 hours, and in those trips, speed used to be constant for large amounts of time.

2.2.3 Feature Extraction

For each trip in $\mathcal{K}_m$, we compute the features listed in Table 2.4.

Table 2.4 Trip Features

Feature	**Symbol**	**Unit**	**Equation**
Beginning hour	τ^B	-	-
End hour	τ^E	-	-
Distance	x_k	[km]	Equation (2.4)
Duration	τ_k	[h]	Equation (2.5)
Average speed	$\bar{v}_k$	[km/h]	Equation (2.6)
Maximum speed	v_k^*	[km/h]	-
Second maximum speed	v_k^{**}	[km/h]	-
Third maximum speed	v_k^{***}	[km/h]	-
Haversine distance	x_k^{H}	[km]	Equation (2.7)
Absolute difference for distance	x_k^{Diff}	[km]	Equation (2.8)
Sum of positive accelerations	a_k^+	$[m/s^2]$	Equation (2.10)
Average positive acceleration	$\bar{a}_k$	$[m/s^2]$	Equation (2.11)
Kinetic intensity	Φ_k	[1/km]	Equation (2.12)
Net change in SOC during the trip	$\Delta\mathrm{SOC}_k$	%	Equation (2.13)
Average ambient temperature	$\bar{T}_k$	[degC]	-
Net elevation change per distance	Δh_k	-	Equation (2.14)
Gross vehicle weight	w_k	[ton]	-
Energy consumption	E_k	[kWh]	Equation (2.15)
Unit energy consumption	ϵ_k	[kWh/km]	Equation (2.16)

A trip is primarily identified by its beginning and end indices, and beginning hour and end hour correspond to the timestamps at the beginning and end indices, respectively. Then, duration is the difference between the end hour and beginning hour. The most essential feature in Table 2.4 might be the unit energy consumption since we aim to construct an empirical model to predict the unit energy consumption of a medium-duty battery electric delivery truck. The unit energy consumption is defined as the energy consumption per distance travelled, and hence it is calculated as the ratio of energy consumption to distance as given in Equation (2.16). We ensure that the energy consumption is in line with the net change in SOC during the trip. Similarly, we crosscheck the distance values against another distance measure, namely Haversine distance. As stated in Equation (2.7), Haversine distance measures the distance travelled from a geometric point of view. For a particular trip, we expect the distance and Haversine distance to be close to each other, and we quantify the closeness by calculating the absolute difference between the two measures (Equation (2.8)). We observe from the literature that speed and acceleration related measures are important while predicting the energy consumption (Heide and Mohazzabi, 2013; Lee and Thomas, 2017; Fetene et al., 2017; Modi et al., 2020; Ahmed et al., 2022). Therefore, as speed related measures, we compute the average speed, maximum speed, second maximum speed, and third maximum speed during each trip. As acceleration related measures, we consider sum of positive accelerations and average positive acceleration. Another essential measure for energy consumption prediction is driving aggressiveness (i.e. aggressive versus gentle driving profile). As described in the work of O'Keefe et al. (2007), kinetic intensity is a measure of driving aggressiveness. It is the ratio of characteristic acceleration to aerodynamic speed, and it might be used to detect drive cycles which are good for hybridization. In general, the drive cycles with high kinetic intensity are considered suitable for hybridization. Based on this motivation, we compute kinetic intensity during each trip as shown in Equation (2.12). As stated in the work of Prohaska et al. (2016), another factor energy consumption is sensitive to is road profile. We attempt to capture the effect of road profile through the net elevation change per distance. As given in Equation (2.14), the net elevation change per distance is the net change in altitude per unit distance. Woody et al. (2022c) demonstrates that the impact of vehicle weight on energy consumption may be significant. Thus, we associate a gross vehicle weight with each trip based on the particular truck the trip is performed by. Last but not least, ambient temperature is a prominent factor in energy consumption prediction. Yuksel and Michalek (2015) show that the ambient temperature may affect the driving range of a BEV up to 40% on too cold winter or hot summer days. We compute the average ambient temperature during each trip by using real data from NSRDB (NSRDB, 2024). Although we have ambient temperature among the parameters in the NREL Fleet DNA data (Table 2.2), the temperature vector is all null for around 10% of the trips in $\mathcal{K}_m$. Therefore, we decide to use real data for ambient temperature.

Equations (2.4)-(2.16) show how each feature is calculated, and more details about the calculations are provided below for $k = 1, \dots, |\mathcal{K}_m|$.

$$x_k \approx \sum_{i=b_k}^{e_k-1} \frac{\mathrm{v}_{k,i} + \mathrm{v}_{k,i+1}}{2} \cdot (\mathrm{t}_{i+1} - \mathrm{t}_i) \tag{2.4}$$

$$\tau_k = \left(t_{e_k} - t_{b_k}\right) \tag{2.5}$$

$$\bar{s}_k = \frac{x_k}{t_k} \tag{2.6}$$

(2.7)

$$\zeta_k = \sum_{i=b_k}^{e_k-1} \sin^2\left(\frac{\mathrm{lat}_{k,i+1} - \mathrm{lat}_{k,i}}{2}\right) + \cos\left(\mathrm{lat}_{k,i}\right) \cdot \cos\left(\mathrm{lat}_{k,i+1}\right) \cdot \sin^2\left(\frac{\mathrm{long}_{k,i+1} - \mathrm{long}_{k,i}}{2}\right)$$

$$x_k^{\mathrm{H}} = R\left(2 \cdot \arctan\left(\sqrt{\zeta_k} / \sqrt{1 - \zeta_k}\right)\right)$$

$$x_k^{\mathrm{Diff}} = \left| x_k - x_k^{\mathrm{H}} \right| \tag{2.8}$$

$$\mathrm{a}_{k,i} = \frac{v_{k,i} - v_{k,i-1}}{t_i - t_{i-1}} \qquad \forall i = 1, \dots, e_k \tag{2.9}$$

$$a_k^+ = \sum_{i=b_k}^{e_k} \mathbb{1}_{\mathrm{a}_{k,i} > 0}\{\mathrm{a}_{k,i}\} \tag{2.10}$$

$$\bar{a}_k = \frac{a_k^+}{\mathbb{1}_{\mathrm{a}_{k,i} > 0}\{1\}} \tag{2.11}$$

$$\Phi_k = \frac{\sum_{i=b_k}^{e_k-1} \max\left(0, 0.5 \cdot \left(v_{k,i+1}^2 - v_{k,i}^2\right) + g \cdot \left(h_{k,i+1} - h_{k,i}\right)\right)}{\sum_{i=b_k}^{e_k-1} \frac{v_{k,i+1}^3 + v_{k,i+1}^2 v_{k,i} + v_{k,i+1} v_{k,i}^2 + v_{k,i}^3}{4} \cdot (t_{i+1} - t_i)} \tag{2.12}$$

$$\Delta\mathrm{SOC}_k = \mathrm{soc}_{e_k} - \mathrm{soc}_{b_k} \tag{2.13}$$

$$\Delta h_k = \frac{h_{e_k} - h_{b_k}}{x_k} \tag{2.14}$$

$$E_k \approx -\sum_{i=b_k}^{e_k-1} \frac{V_{k,i} I_{k,i} + V_{k,i+1} I_{k,i+1}}{2} \cdot (t_{i+1} - t_i) \tag{2.15}$$

$$\epsilon_k = \frac{E_k}{x_k} \tag{2.16}$$

Below are further details about the calculations:

- In equation (2.7), we compute the Haversine distance during a trip. R is the Earth's radius in kilometers. We take $R = 6,373$ km.

  At this point, it is important to mention that we have spotted some spikes in the latitude and longitude vectors (Figure 2.4). We believe that these spikes may result from temporary malfunctions in the GPS device. Since these spikes may affect the Haversine distance results, we perform smoothing on the latitude and longitude vectors before computing the Haversine distance. The smoothing algorithm is presented in Algorithm 4.

  In Algorithm 4, we first define a smoothing threshold, s. If the magnitude of the difference between consecutive entries of the input vector is greater than s, we apply smoothing.

  To smooth the latitude and longitude vectors, we choose the value of s as 0.015 decimal degrees. According to Wisconsin State Cartographer's Office (Wisconsin, 2022), a change of 0.01 decimal degrees in latitude corresponds to a distance of 1.1 km; and at 47.7 degrees latitude, a change of 0.01 decimal degrees in longitude corresponds to a distance of 0.75 km. Given these pieces of information, we are aware that 0.015 decimal degrees is a loose threshold. However, one should note that the distance estimations are sensitive to location. We refrain from selecting a tight threshold and smoothing the latitude/longitude vector too much.

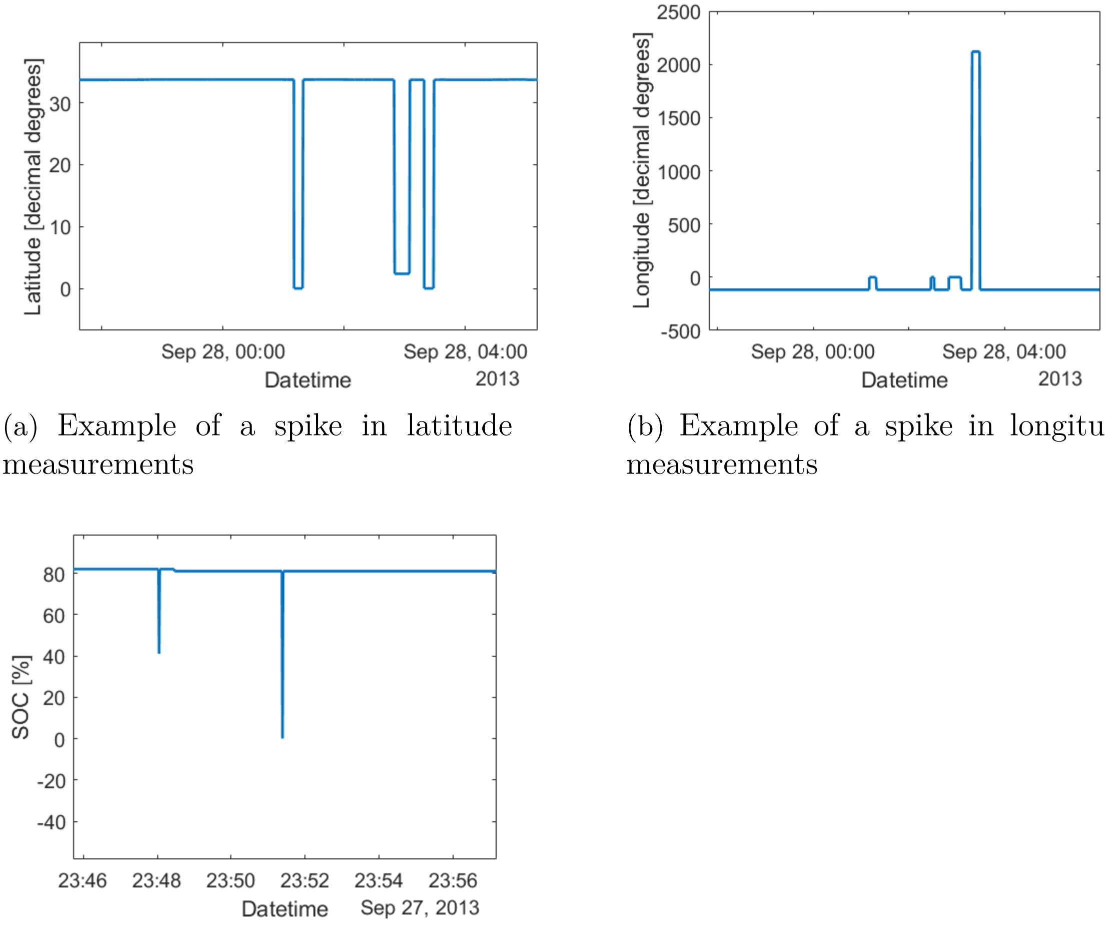
- In equation (2.9), acceleration vector of the trip is computed using the backward finite difference method. Initial acceleration at $i = 0$ is assumed to be zero.

- In equations (2.10) and (2.11), we calculate the sum of positive accelerations and average positive acceleration during a trip, respectively. $\mathbb{1}_{.}$ is the indicator function.

- In equation (2.12), we compute the kinetic intensity during a trip. g is the gravitational acceleration, and we take $g = 9.81\ m/s^2$.

- In equation (2.13), we calculate the net change in SOC during a trip. However, similar to the case for latitude and longitude, we observe spikes in the SOC vectors (Figure 2.4c). Therefore, using Algorithm 4, we apply smoothing on the SOC vector prior to the calculation. We take the value of s as 1%.

- In equation (2.15), we approximate the energy consumption of a trip using the trapezoidal rule. Minus sign is due to the sign convention of current: Positive current represents charging, and negative current represents discharging.



(a) Example of a spike in latitude measurements

(b) Example of a spike in longitude measurements

(c) Example of a spike in latitude measurements

Figure 2.4 Examples of spikes in latitude, longitude, and state of charge measurements

As a final touch to the set of trips, we apply post-processing on $\mathcal{K}_m$. There are two main reasons for post-processing: We do not perform any pre-processing at the beginning, and we still observe some undesired trip characteristics. As we do not perform any input checks, we encounter missing values for the speed-related features, energy consumption, and kinetic intensity. In addition, we notice that the speed-related features, energy consumption, kinetic intensity, and net elevation change per distance take values out of their bounds. Regarding the undesired trip characteristics, we see that some trips are very short (either by duration or distance) or have an unlikely duration and average speed combination. Besides, we observe that some trips have inconsistencies among their features. For example, both distance and Haversine distance measure the distance travelled during a trip, and hence

Algorithm 4 Smoothing

```
Input:  $\phi :=$ Vector to be smoothed
$s :=$ Smoothing threshold
$n = length(\phi)$
for $i = 1$ to $n - 1$ do
    if $i = 1$ then
        if $|\phi_2 - \phi_1| > s$ and $|\phi_3 - \phi_2| \leq s$ then
            $\phi_1 = \phi_2$
        end if
    else
        if $|\phi_{i+1} - \phi_i| > s$ then
            $\phi_{i+1} = \phi_i$
        end if
    end if
end for
```

we expect their values to be close to each other. However, we see that this is not the case for some trips.

Criteria for post-processing and details of the eliminations are shown in Table 2.5. We take the upper bound for the speed-related features as 50 mph since it is the top speed value specified in Table 2.1. To overcome the issue of very short trips, we eliminate a trip if its duration is less than 3 or its distance is lower than 1. To remove the trips with unlikely duration and average speed combinations, we eliminate all trips whose duration is greater than 1 h, but the average speed is below 5 km/h. Finally, to address inconsistencies among the features of a trip, we eliminate a trip for which the absolute difference for distance is greater than 10 km or both Haversine distance and change in SOC are zero. An important detail about Table 2.5 is that the second column shows the number of trips failing each criterion; it does not show the number of eliminated trips. In other words, the second column includes duplicates since a trip may fail multiple criteria. In total, $52,223$ trips are eliminated in post-processing.

Table 2.5 Post-Processing

Criterion	Number of Trips Failing the Criterion
$\tau_k < 3$ min	$34,897$
$\tau_k > 1$ h && $\bar{v}_k < 5$ km/h	165
$v_k^* =$ null || $v_k^* > 50$ mph	$8,358$
$v_k^{**} =$ null || $v_k^{**} > 50$ mph	$1,657$
$v_k^{***} =$ null || $v_k^{***} > 50$ mph	$3,649$
$\bar{v}_k > 50$ mph	12
$x_k < 1$ km	$40,235$
$x_k^{\text{Diff}} > 10$ km	$1,966$
$E_k =$ null || $E_k \leq 0$	654
$\Phi_k =$ null || $\Phi_k > 10.50$ 1/km	$37,041$
$x_k^H = 0$ && $\Delta\text{SOC}_k = 0$	$1,040$
$\|\Delta h\| > 0.4$	2

Let $\mathcal{K}_p$ be the set of remaining trips after post-processing where $|\mathcal{K}_p| = n_{\text{post}} = 41,037$. The average daily distance per truck resulting from $\mathcal{K}_p$ is 29.31 km. Distributions of daily number of trips, duration per trip, average ambient temperature per trip, and unit energy consumption per trip based on the trips in $\mathcal{K}_p$ are presented in Figures 2.6a, 2.6b, 2.6c, and 2.6d, respectively. A machine learning model to predict unit energy consumption based on $\mathcal{K}_p$ is developed in Chapter 5.

2.3 Descriptive Statistics of the Post-Processed Trips

In this section, we present some descriptive statistics based on the post-processed set of trips ($\mathcal{K}_p$). Figure 2.5 shows the monthly average energy consumption for each state. The black line labelled "ALL" shows the average energy consumption based on all states. We observe that the energy consumption is high in winter season, low in spring and fall, and then high again in winter. For all states (and the black line), the energy consumption is higher in winter compared to summer.

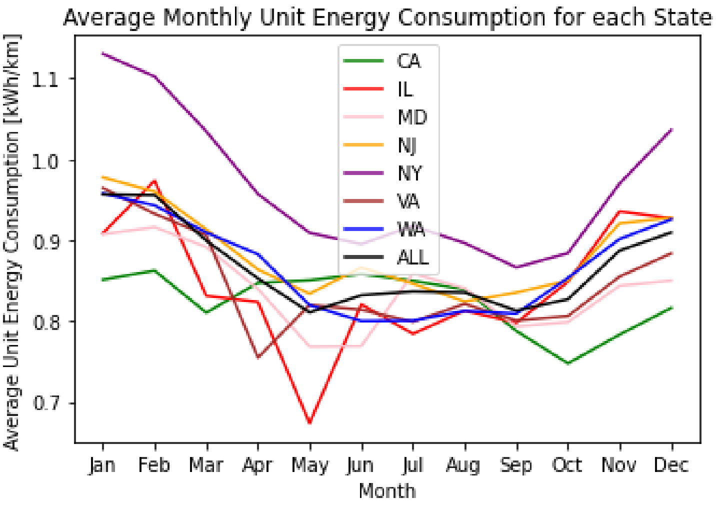
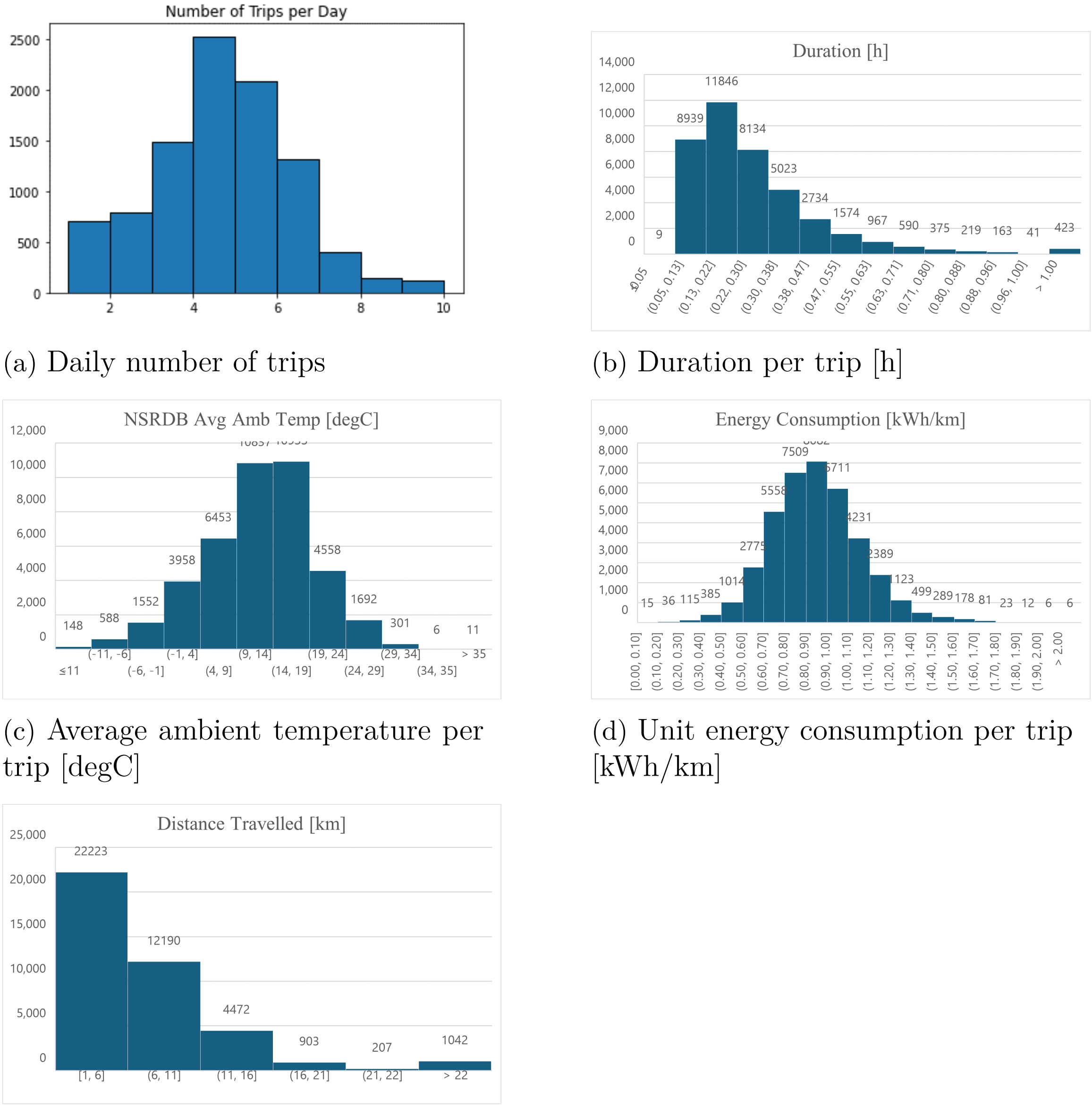


Figure 2.5 Average monthly energy consumption for each state. The black line labelled "ALL" shows the energy consumption based on all states.

Figure 2.6 shows some histograms for the post-processed set of trips. We present histograms for the daily number of trips, and the trip features duration, average ambient temperature, unit energy consumption, and distance travelled.



(a) Daily number of trips

(b) Duration per trip [h]

(c) Average ambient temperature per trip [degC]

(d) Unit energy consumption per trip [kWh/km]

(e) Distance Travelled per trip [km]

Figure 2.6 Distributions of daily number of trips, duration per trip, average ambient temperature per trip, unit energy consumption per trip, and distance travelled per trip

3. MACHINE LEARNING MODEL FOR UNIT ENERGY CONSUMPTION

Our purpose is to build a machine learning model to predict the unit energy consumption of a medium-duty battery electric delivery truck. To decide on the machine learning model, we first consider the tradeoff between the accuracy and explainability of a model. On one hand of the spectrum, we have linear regression models which are the most explainable, but also the least accurate in general. On the other hand, we have neural network models which can achieve high levels of accuracy, yet they usually have a low level of explainability (Qiu et al., 2023). We aim to build a model which is more accurate than the linear regression models, but also more explainable than the neural network models. Thus, we focus on tree-based algorithms. In their study on medium-duty and heavy-duty electrified passenger and delivery vehicles, Qiu et al. (2023) develop three tree-based algorithms (gradient boosted trees, random forest, and XGBoost) to predict the unit energy consumption of an electric vehicle. They observe that all three models yield very close R^2 values. Given this observation and the similarity of their study to our study (Table 1.1), we decide to build a Random Forest model although this might be a suboptimal decision. We develop a Random Forest model, and evaluate its performance relative to a Least Absolute Shrinkage and Selection Operator (LASSO) regression model. The LASSO model serves as our base model following the work of Qiu et al. (2023). In Section 2.2, we compute the features listed in Table 2.4 for each trip. As the predictors of Random Forest, we choose nine of these features which are listed in Table 3.1. In the literature, speed (Fetene et al., 2017; Modi et al., 2020; Ahmed et al., 2022), acceleration (Heide and Mohazzabi, 2013; Fetene et al., 2017; Ahmed et al., 2022), ambient temperature (Yuksel and Michalek, 2015; Woody et al., 2022c; Qiu et al., 2023), and weight (Weiss et al., 2020; Ahmed et al., 2022; Woody et al., 2022c) have been shown to be impactful factors while predicting energy consumption. Road profile (Prohaska et al., 2016) and driving profile (Karabasoglu and Michalek, 2013) are other factors which impact the energy consumption. Road profile is informative on the gradient of a road, i.e. whether a road is a level road (zero gradient) or a steep road. We account for the effect of road profile on energy consumption by including net elevation change per distance among the features. Driving profile is informative on whether a vehicle is driven under aggressive (e.g. sudden accelerations) or gentle conditions. We attempt to include the effect of driving profile through kinetic intensity.

As the predictors of LASSO, we use all the predictors in Table 3.1, and in addition, we populate some of the predictors. Based on our literature review, the relationship between unit energy consumption and some predictors may be polynomial. For example, Yuksel and Michalek (2015) demonstrates a polynomial relationship between unit energy consumption and average ambient temperature. Lee and Thomas (2017) presents a polynomial relationship between unit energy consumption and average speed. Thus, we decide to add polynomial terms for average speed, sum of positive accelerations, average positive acceleration, kinetic intensity, and average ambient temperature. For each predictor we populate, except for average ambient temperature, we add the square, natural logarithm, and reciprocal of the predictor to the model. For average ambient temperature, we consider a fifth order polynomial based on the work of Yuksel and Michalek (2015). As LASSO can force the regression coefficients of some predictors to be zero, it may help us identify the polynomial terms which are relevant for the energy consumption model. The list of predictors of LASSO is given in Table 3.2. X^{RF} and X^{LASSO} in Equations (3.1) and (3.2) are the predictor matrices for Random Forest and LASSO, respectively. Each column of X^{RF} corresponds to a predictor in Table 3.1 (in the same order), and each row corresponds to a trip in $\mathcal{K}_p$. X^{RF} is an n_{post} x 9 matrix as there are n_{post} many trips in $\mathcal{K}_p$ and nine predictors in the Random Forest model. Similarly, each column of X^{LASSO} corresponds to a predictor in Table 3.2, and each row corresponds to a trip in $\mathcal{K}_p$. Differently than X^{RF}, X^{LASSO} has 25 columns since there are 25 predictors in the LASSO model.

Table 3.1 Predictors of Random Forest

	Predictor	**Symbol**
(i)	Duration	τ_k
(ii)	Average speed	$\bar{v}_k$
(iii)	Maximum speed	v_k^*
(iv)	Sum of positive accelerations	a_k^+
(v)	Average positive acceleration	$\bar{a}_k$
(vi)	Kinetic intensity	Φ_k
(vii)	Average ambient temperature	$\bar{T}_k$
(viii)	Net elevation change per distance	Δh_k
(ix)	Gross vehicle weight	w_k

Table 3.2 Predictors of LASSO

	Predictor	**Self**	**Square**	**Log**	**Reciprocal**	**Cube**	**Power 4**	**Power 5**
(i)	τ_k	✓						
(ii)	$\bar{v}_k$	✓	✓	✓	✓			
(iii)	v_k^*	✓						
(iv)	a_k^+	✓	✓	✓	✓			
(v)	$\bar{a}_k$	✓	✓	✓	✓			
(vi)	Φ_k	✓	✓	✓	✓			
(vii)	$\bar{T}_k$	✓	✓			✓	✓	✓
(viii)	Δh_k	✓						
(ix)	w_k	✓						

$$\mathbf{X}^{\mathrm{RF}} = \left[(\mathbf{x}_1^{\mathrm{RF}})^\top \quad \dots \quad (\mathbf{x}_{n_{\mathrm{post}}}^{\mathrm{RF}})^\top\right]^\top \tag{3.1}$$

where the kth row is

$$\mathbf{x}_k^{\mathrm{RF}} = \left[\tau_k \quad \bar{V}_k \quad v_k^* \quad a_k^+ \quad \bar{a}_k \quad \Phi_k \quad \bar{T}_k \quad \Delta \mathrm{h}_k \quad w_k\right]$$

The feature matrix of LASSO can be written as a block matrix

$$\mathbf{X}^{\mathrm{LASSO}} = \left[\mathbf{X}^{\mathrm{RF}} \quad \mathbf{X}_{\bar{v}}^{\mathrm{LASSO}} \quad \mathbf{X}_{a^+}^{\mathrm{LASSO}} \quad \mathbf{X}_{\bar{a}}^{\mathrm{LASSO}} \quad \mathbf{X}_{\Phi}^{\mathrm{LASSO}} \quad \mathbf{X}_{T}^{\mathrm{LASSO}}\right] \tag{3.2}$$

where the matrices in the block are

$$\mathbf{X}_{\bar{v}}^{\mathrm{LASSO}} = \left[(\mathbf{x}_{\bar{v},1}^{\mathrm{LASSO}})^\top \quad \dots \quad (\mathbf{x}_{\bar{v},n_{\mathrm{post}}}^{\mathrm{LASSO}})^\top\right]^\top$$
$$\mathbf{X}_{a^+}^{\mathrm{LASSO}} = \left[(\mathbf{x}_{a^+,1}^{\mathrm{LASSO}})^\top \quad \dots \quad (\mathbf{x}_{a^+,n_{\mathrm{post}}}^{\mathrm{LASSO}})^\top\right]^\top$$
$$\mathbf{X}_{\bar{a}}^{\mathrm{LASSO}} = \left[(\mathbf{x}_{\bar{a},1}^{\mathrm{LASSO}})^\top \quad \dots \quad (\mathbf{x}_{\bar{a},n_{\mathrm{post}}}^{\mathrm{LASSO}})^\top\right]^\top$$
$$\mathbf{X}_{\Phi}^{\mathrm{LASSO}} = \left[(\mathbf{x}_{\Phi,1}^{\mathrm{LASSO}})^\top \quad \dots \quad (\mathbf{x}_{\Phi,n_{\mathrm{post}}}^{\mathrm{LASSO}})^\top\right]^\top$$
$$\mathbf{X}_{\bar{T}}^{\mathrm{LASSO}} = \left[(\mathbf{x}_{\bar{T},1}^{\mathrm{LASSO}})^\top \quad \dots \quad (\mathbf{x}_{\bar{T},n_{\mathrm{post}}}^{\mathrm{LASSO}})^\top\right]^\top$$

such that the rows of those matrices are

$$\mathbf{x}_{\bar{v},k}^{\mathrm{LASSO}} = \left[\bar{v}_k^2 \quad \log(\bar{v}_k) \quad \bar{v}_k^{-1}\right]$$
$$\mathbf{x}_{a^+,k}^{\mathrm{LASSO}} = \left[(a_k^+)^2 \quad \log(a_k^+) \quad (a_k^+)^{-1}\right]$$
$$\mathbf{x}_{\bar{a},k}^{\mathrm{LASSO}} = \left[\bar{a}_k^2 \quad \log(\bar{a}_k) \quad \bar{a}_k^{-1}\right]$$

$$\mathbf{x}_{\Phi,k}^{\text{LASSO}} = \begin{bmatrix}\Phi_k^2 & \log(\Phi_k) & \Phi_k^{-1}\end{bmatrix}$$
$$\mathbf{x}_{\bar{T},k}^{\text{LASSO}} = \begin{bmatrix}\bar{T}_k^2 & \bar{T}_k^3 & \bar{T}_k^4 & \bar{T}_k^5\end{bmatrix}.$$

We split $\mathbf{X}^{\text{RF}}$ and $\mathbf{X}^{\text{LASSO}}$ into train and test matrices using the default train-test split ratio of $80\% - 20\%$. The split is stratified with respect to the predictors in the Random Forest model (Table 3.1). In other words, the range of each predictor in the train and test datasets are similar. The minimum and maximum values of each predictor in the train and test datasets are reported in Table 3.3. The columns Interval{Train} and Interval{Test} show the range of a predictor in the train and test datasets, respectively.

Table 3.3 Range of each Predictor in the Train and Test Datasets

Predictor	**Interval{Train}**	**Interval{Test}**
Duration [h]	0.05 - 2.95	0.05 - 2.80
Average speed [km/h]	2.86 - 71.9	2.66 - 70.3
Maximum speed [km/h]	20.9 - 80.5	24.1 - 80.5
Average ambient temperature [degC]	-22.1 - 37.5	-22.5 - 37.6
Sum of positive accelerations [m/s^2]	5.36 - 1,584	6.26 - 1,312
Average positive acceleration [m/s^2]	0.447 - 2.53	0.447 - 1.84
Kinetic intensity [1/km]	0.112 - 10.5	0.0617 - 10.5
Net elevation change per distance [ ]	-0.309 - 0.0826	-0.159 - 0.0871
Gross vehicle weight [ton]	7.5 - 12.0	7.5 - 12.0

Let $K^{\text{Post, Train}}$ and $K^{\text{Post, Test}}$ be the set of trips in the train and test splits, respectively where $|K^{\text{Post, Train}}| = n_{\text{post, train}} = 32,829$ and $|K^{\text{Post, Test}}| = n_{\text{post, test}} = 8,208$. Then, for LASSO only, we standardize the train and test matrices. We ensure that the standardized train matrix is full-rank so that the ordinary least squares (OLS) solution exists. Rank of a matrix is defined as the number of independent columns, i.e., the number of independent predictors. Therefore, ensuring that the augmented matrix is full-rank also ensures that all predictors are independent of each other. Let $\mathbf{X}^{\text{LASSO, Train}}$ be the standardized and augmented train matrix for LASSO, $\mathbf{X}^{\text{LASSO, Test}}$ be the standardized test matrix for LASSO, and $\mathbf{X}^{\text{RF, Train}}$ and $\mathbf{X}^{\text{RF, Test}}$ be the train and test matrices for Random Forest, respectively. Also, let $\mathbf{y}^{\text{Train}}$ and $\mathbf{y}^{\text{Test}}$ in Equations (3.3) and (3.4) be the response vectors for train and test, respectively.

$$\mathbf{y}^{\text{Train}} = \begin{bmatrix}y_1^{\text{train}} & \dots & y_{n_{\text{post, train}}}^{\text{train}}\end{bmatrix}^\top \tag{3.3}$$

$$\mathbf{y}^{\text{Test}} = \begin{bmatrix}y_1^{\text{test}} & \dots & y_{n_{\text{post, test}}}^{\text{test}}\end{bmatrix}^\top. \tag{3.4}$$

We develop our machine learning models in Python. For LASSO, we use the *LASSOCV* package from the sklearn.linear_model library. The objective of the LASSO model is to minimize the error term given by

$$\frac{1}{n_{\text{post, test}}} \sum_{k=1}^{n_{\text{post, test}}} \left(y_k^{\text{Test}} - \beta_0 - \sum_{j=1}^{25} \beta_j \mathbf{X}_{kj}^{\text{LASSO, Test}} \right)^2 + \lambda \sum_{j=1}^{25} |\beta_j|, \tag{3.5}$$

where β is the vector of regression coefficients, β_0 is the intercept, and λ is a tuning parameter. To decide on the value of λ, we perform 10-fold cross validation, and pick the λ which yields the lowest mean squared error (MSE).

We construct the LASSO model with the hyperparameter values shown in Table 3.4. The hyperparameters which are not listed in the table are kept at their default values. In Table 3.4, *eps* defines the range of values which are cross validated for λ. In other words, it is the ratio of the minimum λ value to the maximum λ value. *max_iter* is the maximum number of iterations that can be performed to identify the best λ. *tol* defines the duality gap between the best value identified for λ and the optimal λ value. Finally, *cv* is the number of folds in cross validation.

Table 3.4 Hyperparameter Values for LASSO

Hyperparameter	**Value**
eps	1e-04
max_iter	1e+05
tol	1e-07
cv	10

After 10-fold cross validation, the best value for λ is identified to be $1.92e-05$. The intercept is 0.851. Such a low value for λ might seem counterintuitive, however, as visualized in Figure 3.1, the lowest MSE is yielded by $\lambda = 1.92e-05$. In Figure 3.1, the average MSEs resulting from the 10-folds are plotted against the cross validated λ values. The λ values are shown in logarithm base 10. The β coefficients for each predictor, in the descending order of the magnitudes of β_j's, is given in Table A.2.

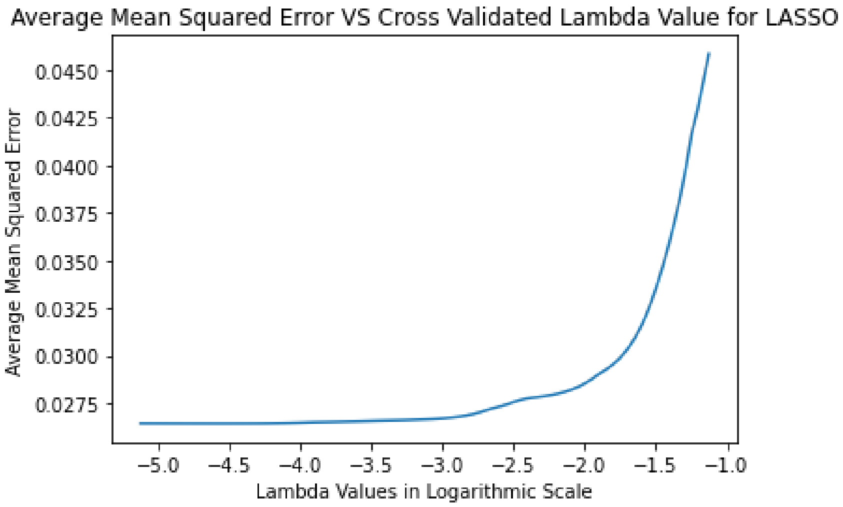


Figure 3.1 Average Mean Squared Error versus Cross Validated Lambda Value for LASSO

For Random Forest, we use the *RandomForestRegressor* package from the sklearn.ensemble library. We change the values of only three hyperparameters: *max_samples, max_features,* and *min_samples_split.* We keep the other hyperparameters at their default values. *max_samples* indicates the fraction of samples to be used while training each tree. *max_features* is the number of features to be considered while determining the best split of a node into its children. Finally, *min_samples_split* is the minimum number of samples necessary to split a node further. To decide on the values of these three hyperparameters, we perform hyperparameter tuning. For each hyperparameter, we consider a set of possible values. Then, we check all resulting combinations of the hyperparameters. We construct a Random Forest model with each combination, train the model with $\mathbf{X}^{\text{RF, Train}}$, test the model with $\mathbf{X}^{\text{RF, Test}}$, and record the R^2 value yielded at the test. At the end, we label the combination which yields the highest R^2 at the test as the best combination. The set of values considered for each hyperparameter is given in Table 3.5. The best hyperparameter combination is presented in Table 3.6.

Table 3.5 Set of Possible Hyperparameter Values for Random Forest

Hyperparameter	**Value**
max_samples	{0.2, 0.4, 0.6, 0.8, 1.0}
max_features	{0.2, 0.4, 0.6, 0.8, 1.0}
min_samples_split	{2, 4, 6}

The Gini importance values for each predictor, resulting from the best combination of hyperparameters, is given in Table 3.7. Gini importance quantifies the contribution of each feature to the Random Forest model. The Gini importance of a particular feature is calculated based on the reduction in Gini impurity achieved by splitting a node in a tree based on that feature. To calculate the Gini importance of a particular feature, all trees in the Random Forest are considered. At the end, the Gini importance of a particular feature is the overall reduction in Gini impurity achieved by splitting a node based on that feature. Features which provide greater reductions in Gini impurity are assigned higher importance values (GeeksForGeeks, 2024).

Table 3.6 The Best Hyperparameter Combination for Random Forest

Hyperparameter	**Value**
max_samples	1.0
max_features	0.6
min_samples_split	6

Table 3.7 Gini Importance Values for Random Forest

Predictor	**Gini Importance** [%]
Net elevation change per distance	33.3
Gross vehicle weight	10.4
Kinetic intensity	9.89
Sum of positive accelerations	9.16
Average ambient temperature	8.32
Average speed	8.22
Maximum speed	7.69
Duration	6.99
Average positive acceleration	6.07

For both LASSO and Random Forest, the summary of error statistics is presented in Table 3.8. For both models, we present the results of the best solutions (i.e. the LASSO model with $\lambda = 1.92e-05$ and the Random Forest model constructed with the hyperparameters specified in Table 3.6). To evaluate the performance of a model, we use four error statistics: mean squared error (MSE), mean absolute percentage error (MAPE), coefficient of determination (R^2), and adjusted R^2. The suffixes {Train} and {Test} refer to the results obtained in the train and test, respectively. Looking at the test results, we observe that Random Forest outperforms LASSO.

Table 3.8 Summary of Error Statistics for the Machine Learning Models

	LASSO	**Random Forest**
Intercept	0.851	-
MSE{Train}	0.0265	0.00328
MAPE{Train}	17.1%	5.63%
R^2{Train}	0.427	0.929
Adjusted R^2{Train}	0.427	0.929
MSE{Test}	0.0261	0.0140
MAPE{Test}	19.2%	11.6%
R^2{Test}	0.440	0.690
Adjusted R^2{Test}	0.438	0.690

The Random Forest model for unit energy consumption will be used in Chapter 4.

4. SIMULATION

We develop a simulation framework to predict the use phase emissions of medium-duty battery electric delivery trucks across the United States (US). There are three main reasons for building this framework: First, we aim to predict the emissions for the entire US; however the NREL Fleet DNA data comprises a limited number of counties. In other words, we do not know the trip details in all counties. Second, even if a county is present in the NREL Fleet DNA data, we do not have its trip records for the entire year. Finally, we perform simulation to quantify the effects of regional differences on the use phase emissions. The simulation framework is illustrated in Figure 4.1.

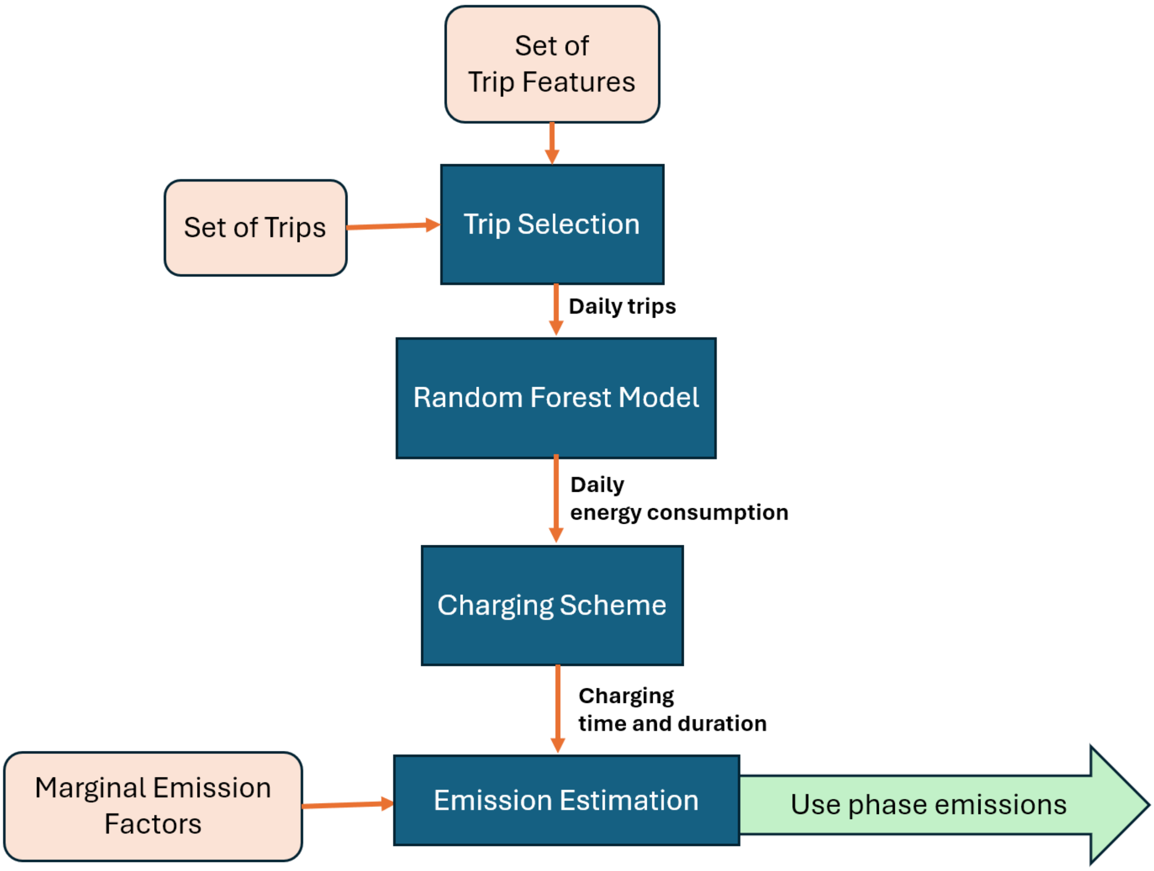


Figure 4.1 Illustration of the simulation framework

We feed the set of trips $\mathcal{K}_p$ as an input into the simulation framework. Then, based on a set of trip features which is determined by the user, we select a trip from $\mathcal{K}_p$. We calculate the energy consumption of the selected trip using the Random Forest model we developed in Chapter 3. Depending on the daily number of trips, we repeat these steps for all trips of a day. At the end of the day, we calculate the daily energy consumption, and apply a charging scheme. We combine the charging duration and time with the marginal emission factors (MEFs) from electricity generation to compute the use phase emissions. We repeat this procedure for one year to estimate the annual energy consumption and emissions. As both the energy consumption and emissions are random variables (i.e., they are sampled from mathematical distributions), we perform 100 Monte Carlo (MC) runs for our simulations to obtain more reliable results. Using our simulation framework, we perform two case studies. In the first case study, we quantify the effect of ambient temperature on the use phase emissions across the US, and in the second case study, we attempt to quantify the effect of driving profile.

4.1 Effect of Ambient Temperature on Use Phase Emissions

In the first case study, we aim to quantify the effect of ambient temperature on the use phase emissions. Details of trip generation, charging scheme, and computation of the use phase emissions are presented below.

1) Trip generation

We generate the trips of a particular county K_l^{Sim} as described in Algorithm 5. Let L and D be the set of counties and days, respectively where |L| = 3,052 and |D| = 365.

Algorithm 5 Trip Generation

```
$K_l^{\text{Sim}}$
for $d = 1$ to 365 do
    $n_d \sim \mathcal{F}_N$
    Schedule$_d$ := Schedule for the trips on day $d$
    flag_battery_capacity_failed = True
    while flag_battery_capacity_failed do                ▷ Rejection sampling
        for $k = 1$ to $n_d$ do
            trip$_k$ = generate_trip($S_k^{\text{F}}, S_k^{\text{V}}, S_k^{\text{h}}$)
        end for
        Calculate total daily energy consumption $\varepsilon^{Day}$ in kWh
        flag_battery_capacity_failed = False
        if $\varepsilon^{Day} > battery\_capacity$ then
            flag_battery_capacity_failed = True
        end if
    end while
    $K_{ld}^{\text{Sim}}$ := Set of trips on day $d$
    $K_l^{\text{Sim}} = K_l^{\text{Sim}} \cup K_{ld}^{\text{Sim}}$
end for
```

Details of Algorithm 5 are explained below:

- We generate trips for each day of the year. For a particular day, we first determine the number of trips (n_d) which comes from the discrete empirical distribution $\mathcal{F}_N$. $\mathcal{F}_N$ is the underlying empirical distribution of the histogram in Figure 2.6a.

- Once n_d is determined, we generate a schedule for the trips (Schedule$_d$) which shows the beginning hour, duration, and end hour of each trip. We generate the schedule based on a Gamma distribution. The scale parameter of the Gamma distribution is taken as 0.22 because 0.22 h is the average trip duration based on Figure 2.6b. The shape parameter of the distribution is adjusted such that the earliest beginning time for the first trip of the day is 8 am, the latest end time for the last trip of the day is 10 pm, and the average duration between consecutive trips is half an hour.

- We generate each trip of the day by using the subroutine *generate_trip($S_k^{\text{F}}, S_k^{\text{V}}, S_k^{\text{h}}$)*. The arguments S_k^{F}, S_k^{V}, and S_k^{h} are the set of features, feature values, and feature tolerances, respectively for generating the k^{th} trip. Details of *generate_trip(.)* are presented in Algorithm 6.

- At the end of the day, we check whether the total daily energy consumption in kWh ($\varepsilon_d^{\text{Day}}$) exceeds the battery capacity. If this is the case, we apply rejection sampling, and regenerate all trips until the battery capacity is not exceeded.

We take the value of battery capacity as 80 kWh in line with the specifications in Table 2.1.

Details of the subroutine *generate_trip(.)* are presented in Algorithm 6:

Algorithm 6 generate_trip

```
**Input:** $S^{\mathrm{F}}, S^{\mathrm{V}}, S^{\mathrm{h}}$ := Set of features, desired values, and tolerances
**Input:** $\mathbf{X}^{\mathrm{Sim}}$
$n = length(\mathbf{X}^{\mathrm{Sim}})$                                  ▷ Number of trips in $\mathbf{X}^{\mathrm{Sim}}$
$n_F = |S^{\mathrm{F}}|$                                              ▷ Number of features
$w_i = \prod_{j \in |S^F|} \exp\left\{ -\frac{1}{2S_j^V} \left( \mathbf{X}_i^{(sim),S_j^F} - S_j^V \right)^2 \right\}, \quad i = 1, \dots, n$
$p_i = \frac{w_i}{\sum_{j=1}^{n} w_j}, \quad i = 1, \dots, n$
$i^* \sim \text{Categorical}(p_1, \dots, p_n)$                        ▷ Random generation
trip $= \mathbf{X}_i^{\mathrm{Sim}}$
trip$_j$ $= S_j^{\mathrm{V}} \quad \forall j = 1, \dots, n_F$     ▷ Modify the selected trip to generate a new trip
```

Details of Algorithm 6 are explained below:

- S^{F} is a set of features, S^{V} is the set of desired values of the features, and S^{h} is the set of tolerances for the desired values. In this algorithm, we generate a trip by first selecting a trip from $\mathbf{X}^{\mathrm{Sim}}$, and then modifying the values of some of its features. We aim to select a trip such that the values of the features specified in S^{F} are as close as possible to those specified in S^{V}.
  - We utilize three features for trip generation: beginning hour, duration, and average ambient temperature, i.e., we take $S^{\mathrm{F}} = \{\tau^B, \tau, \bar{T}\}$.
  - Values of beginning hour and duration come from the schedule in Algorithm 5. We use the beginning hour as it is, but we further randomize duration as stated in Equation (4.1) where $\mathcal{U}$ denotes the uniform distribution, $\tau^{'}$ is the duration in the schedule, and 0.05 h (3 min) is the minimum duration of a trip in line with Table 2.5. For average ambient temperature, we use real data from NSRDB (NSRDB, 2024). As we perform our simulation for a typical year, we use the Typical Meteorological Year (TMY) temperature values.

$$\tau = \begin{cases} \mathcal{U}(0.05, \tau^{'}) & \text{if } \tau^{'} < 1 \\ \mathcal{U}(0.05, \tau^{'}/2) & \text{otherwise} \end{cases} \tag{4.1}$$

  - We define the tolerances for each feature as $S^{\mathrm{h}} = \{0.25 \text{ h}, 10 \text{ min}, 20 \text{ degC}\}$. Giving such a high tolerance for temperature has a similar effect to not including temperature in S^{F} at all. However, we want temperature to be one of the features since it is the only regional difference across the counties in our first simulation, and we believe that temperature may have an effect on other trip features although marginal. Therefore, we prefer to include average ambient temperature in S^{F}, but give it a high tolerance.

- $\mathbf{X}^{\mathrm{Sim}}$ is defined as $\mathbf{X}^{\mathrm{Sim}} = \begin{bmatrix} \tau^B & \mathbf{X}^{\mathrm{RF}} \end{bmatrix}$ where τ^B is the column vector of beginning hours for each trip, i.e.,

$$\tau^B = \begin{bmatrix} \tau_1^B & \dots & \tau_{n_{\mathrm{post}}}^B \end{bmatrix}^\top$$

- In Algorithm 6, we generate a trip by first selecting a trip from $\mathbf{X}^{\mathrm{Sim}}$, and then modifying some of its feature values. To select a trip whose feature values are as close as possible to those specified in S^{V}, we form Gaussian kernels around the trips. The dimension of the space the Gaussian kernels lie in is equal to the number of features in S^{F}. The Gaussian kernels are formed based on the feature values specified in S^{V} and the bandwidths specified in S^{h}. The closer the features of a trip are to the values in S^{V}, the higher the probability of that trip to be selected. In lines 5 and 6, we compute the probability of each trip being selected based on how close their feature values are to those in S^{V}. Then, in line 7, we select the i^{th} trip by generating a random number between 1 and n with respect to the selection probabilities of the trips.

- In this algorithm, we generate a new trip by modifying the values of some of the features of the selected trip. For all features in S^{F}, we replace the value of the feature in the trip with the value specified in S^{V}, and we keep the values of the remaining features unchanged. In other words, the new trip has exactly the same values for the features specified in S^{F}, and the other features are compensated from the selected trip. This way of trip generation may be a remedy for the following cases:

  - The user cannot provide the values of all the predictors in the energy consumption model (Table 3.1). The Random Forest model requires the values of all nine predictors to make a prediction, however, it might be difficult for a fleet owner to know the values of some predictors such as sum of positive accelerations, kinetic intensity, and net elevation change per distance. In this case, the fleet owner can provide as many predictors as they can, and the remaining predictors can be compensated from the selected trip.

- Even though the value of a predictor is available to the fleet owner, they might think that the value is erroneous, and may not want to use it. In this case, similar to the previous item, the value of that predictor can be substituted from the selected trip.
- There exist some physical relations between the predictors which must be respected in order to make a reliable prediction. Nevertheless, it might be difficult for a fleet owner to ensure that there are no measurement errors and hence the physical relations are satisfied among the feature values they provide. Imputing the values of some predictors from the selected trip increases the chance that the physical relations will be respected.

2) Charging scheme

As the charging scheme, we apply convenience-full charging. The charging begins when the last trip of the day ends, and the battery is charged up to 100% SOC. The charging duration is calculated by

$$\delta_{ld} = \frac{\varepsilon_{ld}^{\text{Day}}}{\eta r} \quad \forall l \in |L|,\ d \in D, \tag{4.2}$$

where δ_{ld} is the charging duration in hours, η is the charging efficiency, and r is the constant charging rate. We take $\eta = 85\%$ and $r = 15$ kW.

Then, as in the work of Yuksel et al. (2016), we distribute the charging duration into hourly bins using Equation (4.3):

(4.3)

$$\Delta_{ldh}^{\text{Charging}} = \sum_{n=0}^{1} \begin{cases} 1 & \text{if} \quad \tau_{ld}^{E} + 1 \leq t_n \leq \tau_{ld}^{E} + \delta_{ld} \\ 0 & \text{if} \quad \tau_{ld}^{E} \geq t_n \text{ or } \tau_{ld}^{E} + \delta_{ld} \leq t_n - 1 \\ \min(t_n, \tau_{ld}^{E} + \delta_{ld}) - \max(t_n - 1, \tau_{ld}^{E}) & \text{otherwise} \end{cases}$$

$$\forall l \in L,\ d \in D,\ h \in H$$

$$t_n = h + 24n,$$

where $\Delta_{ldh}^{\text{Charging}}$ is the charging duration falling into hourly bin h, and τ_{ld}^{E} is the end time of the last trip of the day. t_n is used to handle the cases where the charging continues on the following day.

3) Use phase emissions

We define the use phase emissions as the direct emissions from the electric grid as a result of recharging events. We compute the use phase emissions based on marginal emission factors (MEFs). Marginal emissions can be defined as the amount of GHG emissions released from the power plants which are utilized to meet the extra demand due to BEV charging. We only consider CO_2 emissions in this study. Therefore MEF shows the amount of CO_2 in kg that is released per each 1 MWh extra electricity generation. We utilize the seasonal hourly MEFs (CEDM, 2021) for the eight North American Electric Reliability Corporation (NERC) regions: Northeast Power Coordinating Council (NPCC), Florida Reliability Coordinating Council (FRCC), Texas Reliability Entity (TRE), Western Electricity Coordinating Council (WECC), SERC Reliability Corporation (SERC), ReliabilityFirst Corporation (RFC), Southwest Power Pool (SPP), and Midwest Reliability Organization (MRO). The MEFs include only the carbon dioxide (CO_2) emissions from the grid. Regional average CO_2 emissions in grams/km γ_l (averaged over all trips and days of the year) can be calculated by

$$\gamma_l = \frac{\sum\limits_d \sum\limits_h r \Delta_{ldh}^{\text{Charging}} \text{MEF}_{ldh}}{\sum\limits_d \sum\limits_{k \in K_{ld}^{\text{Sim}}} x_{ldk}} \qquad \forall l \in |L|, \tag{4.4}$$

where MEF_{ldh} is the regional time of day marginal emission factor for county l in kg-CO_2/MWh.

As the energy consumption and emissions are random variables (i.e. they are sampled from mathematical distributions), we perform 30 Monte Carlo (MC) runs for our simulations to obtain more reliable results.

4.1.1 Results

The average annual energy consumption and use phase emissions for each county are shown in Figures 4.2 and 4.3, respectively. We calculate the average annual energy consumption of a particular county $\varepsilon_l^{\text{Year}}$, in kWh/km, by

$$\varepsilon_l^{\text{Year}} = \frac{\sum\limits_{d \in D} \varepsilon_{ld}^{\text{Day}}}{\sum\limits_{d} \sum\limits_{k \in K_{ld}^{\text{Sim}}} x_{ldk}} \quad \forall l \in |L|, \tag{4.5}$$

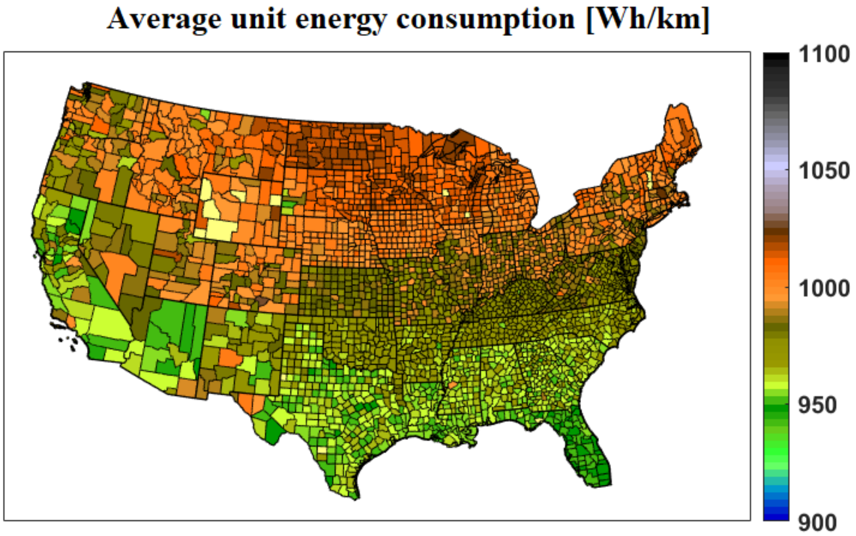


Figure 4.2 Average annual energy consumption for each county in the US. The energy consumption values are reported in Wh/km, and they are averages of 30 Monte Carlo runs

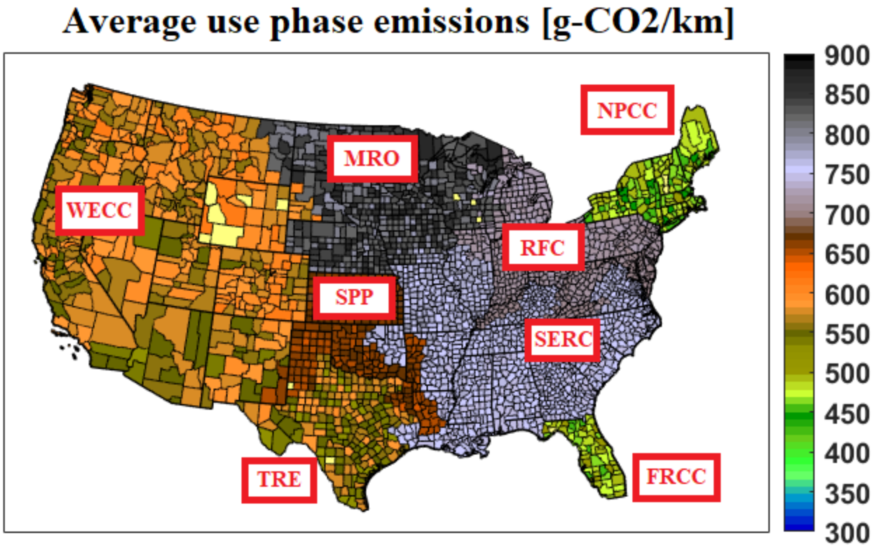


Figure 4.3 Average annual use phase emissions for each county in the US. The emissions values are reported in g-CO_2/km, and they are averages of 30 Monte Carlo runs

The average annual energy consumption per km (averaged over all trips over a full year and averaged over all MC runs) is shown for each county in Figure 4.2. We observe a 10% increase in average energy consumption in the Upper Midwest where a colder climate exists compared to Pacific Coast or Florida where milder temperatures are experienced throughout the year. This is consistent with previous literature which states BEVs consume more energy in cold days due to cabin heating and poorer battery temperature at lower temperatures (Barnitt et al., 2010; Yuksel and Michalek, 2015).

Similar to cold temperatures, hot temperatures can also increase energy consumption due to AC use in hot days. For example, Yuksel and Michalek (2015) reports that the electric driving range of a BEV can drop by 29% in Phoenix, Arizona compared to San Francisco, California due to temperatures that can reach to 41 degC in summer. However, according to our results, average energy consumption does not change between these two regions. To investigate this result, we looked into the relationship between unit energy consumption and average ambient temperature based on our Random Forest model. Figure 4.4 shows how unit energy consumption changes with respect to temperature when all other features are kept constant. We note that the energy consumption can increase up to 18% in cold temperatures compared to the case at 20 degC. However, after 20 degC, the energy consumption does not change. This shows that our model can capture the energy consumption increase in cold temperatures but not in hot temperatures. One reason for this might be the smaller number of trips at high temperatures compared to other temperatures (see Figure 2.6c). Another reason might be about the driver behaviour in hot temperatures. Looking at our trips, we note that the energy consumption values of some of the trips performed in hot temperatures are lower than anticipated. It is possible that the driver did not turn on AC during the trip. This might be reasonable if we consider that most of our trips are short. The average trip duration based on Figure 2.6b is around 15 minutes. The NREL Fleet DNA data does not provide any details about the driver behaviour; i.e. we do not know if AC was used in hot temperatures or not.

Based on the MEFs in Tables A.4 and A.5, respectively, NPCC and MRO regions have the least and most carbon intensive electric grids, respectively. Consequently, the lowest emissions are observed in northeast (NPCC) and southeast (FRCC) regions, and the highest emissions are attained in northern midwest (MRO) as shown in Figure 4.3. The high emissions in MRO are anticipated due to both high energy consumption and high emission factors. Although the energy consumption in MRO region was only 8% higher than NPCC region, we see that the emissions increase by 1.8 times. This shows that the most significant factor affecting the regional differences in emissions is the grid mix.

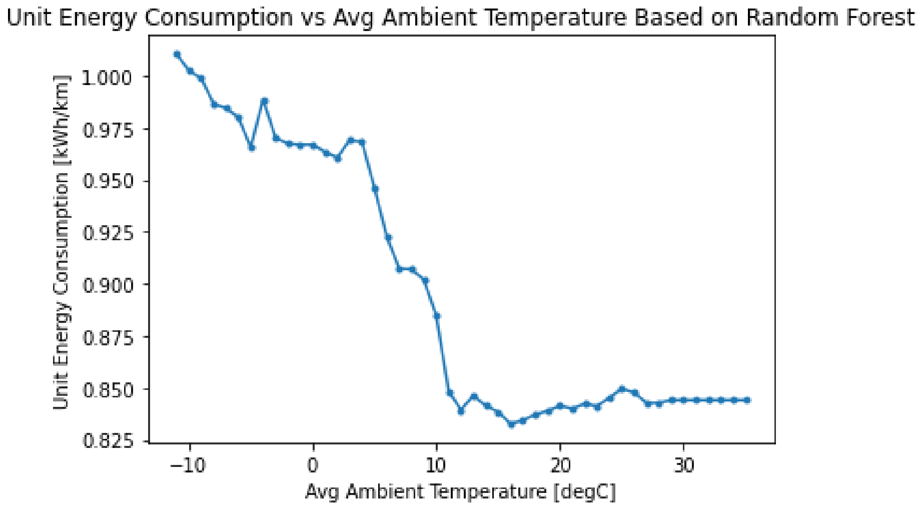


Figure 4.4 Unit Energy Consumption versus average ambient temperature based on the Random Forest model

4.2 Effect of Driving Profile on Use Phase Emissions

In this simulation, we attempt to quantify the effect of driving profile on the use phase emissions. In other words, we want to distinguish between a region with an aggressive driving profile and a region with a gentle profile. However, we were unable to come up with a feature which would allow us to classify the driving profiles of counties as aggressive or gentle. For this reason, we could not perform a simulation for the entire US. Instead, we create two hypothetical counties, County 1 and County 2, such that they represent the aggressive and gentle driving profiles, respectively. We compute the energy consumption and emissions of these two counties based on synthetic data. In addition, we use the MEFs of the NPCC and MRO regions for County 1 and County 2, respectively.

Similar to the first case study, we generate the trips of these two counties using Algorithm 5. However, we now consider four features for trip selection: average speed, sum of positive accelerations, kinetic intensity, and average ambient temperature. We distinguish between an aggressive and a gentle driving profile based on average speed, sum of positive accelerations, and kinetic intensity. Average ambient temperature is kept among the features for trip generation since we use real data for it from the NSRDB TMY database based on the day of the year. Since we use the temperature values of a typical year, we assume the same average ambient temperature for both counties. For County 1, we assume aggressive driving conditions. For County 2, we assume gentle driving conditions. In line with these assumptions, the average speed, sum of positive accelerations, and kinetic intensity values we consider for County 1 and County 2 are given in Table 4.1. The tolerances of the features are taken as 5 km/h, 10 m/s^2, 0.5 1/km, and 20 degC, respectively.

Table 4.1 Synthetic Data for Case Study 2

	County 1	**County 2**
Average Speed [km/h]	75	30
Sum of Positive Accelerations [m/s^2]	520	80
Kinetic Intensity [1/km]	4	1

Same as in the first case study, we calculate the annual average emissions and energy consumption (averaged over all trips and days of the year) using Equations (4.4) and (4.5), respectively. The results are shown in Table 4.2 below. The values in parantheses are the standard deviations.

Table 4.2 Average annual energy consumption and emissions for aggressive and gentle driving profiles

	County 1	**County 2**
Average annual energy consumption [Wh/km]	970 (5.35)	908 (3.74)
Average annual emissions [g-CO_2/km]	758 (1.63)	520 (3.09)

We observe a 7% increase in energy consumption in County 1 compared to County 2. Studies on passenger vehicles have shown that aggressive driving can increase energy consumption by $23 - 27\%$ (Karabasoglu and Michalek, 2013; Mohammadnazar et al., 2024). Our Random Forest model captures the increasing trend, however the amount of increase could have been larger. Detailed investigation of both the machine learning model and simulation framework is necessary to explore possible causes of the discrepancy.

5. CONCLUSION

In this study, we quantify the regional differences in the emission benefits of medium-duty delivery vehicle electrification. A vital part of the simulation framework is the Random Forest model which we develop to predict the unit energy consumption of a battery electric delivery truck at a trip level. In order to make predictions at the trip level, we devised an algorithm to identify trips from the source data. Our energy consumption model accounts for various factors such as ambient temperature, driving profile, and road profile. Using our Random Forest model, we develop a simulation framework which can predict the use phase emissions across the United States. Using our framework, we perform two case studies. In both case studies, we perform simulations for one year. In the first case study, we aim to quantify the effect of ambient temperature on the regional variation in use phase emissions. We simulate the annual energy consumption and annual use phase emissions in each county across the United States. We observe that our model predicts higher energy consumption values in regions with a colder climate. For example, we note a 10% increase in the average annual energy consumption in the Upper Midwest compared to a region with a milder climate such as Pacific Coast or Florida. This result is in line with the literature as the energy consumption is expected to increase at low temperatures due to the use of heater or decrease in the battery efficiency. However, our model cannot capture the increase in the energy consumption in regions with a hotter climate. The energy consumption is expected to increase at high temperatures as well due to the use of air conditioning. One reason for this unanticipated result might be that most of our trips were short, and it is possible that the driver did not turn on the AC during the trip. The source data does not provide any details regarding the driver behaviour during the trips. Regarding the use phase emissions, the lowest and highest emissions are observed in the NPCC and MRO regions, respectively point out the impact of electricity grid mix in regional emissions variations. In the second case study, we attempt to quantify the effect of driving profile on the use phase emissions. However, we were unable to come up with a feature which would help us decide whether a region has an aggressive or a gentle driving profile. Therefore, we perform our simulations with synthetic data. Our

Random Forest model still captures the increase in energy consumption under an aggressive driving profile, yet the amount of increase could have been larger if we could find a better way of distinguishing between aggressive and gentle driving.

As future work, the robustness of our results can be tested under different charging schemes or emission factors. Besides, to improve the accuracy of energy consumption prediction, a neural network model can be developed. Finally, to investigate the emission benefits potential of medium-duty delivery truck electrification further, one may attempt to differentiate between the driving profiles of regions or consider the impact of different factors on the use phase emissions such as road profiles.

BIBLIOGRAPHY

Ahmed, M., Mao, Z., Zheng, Y., Chen, T., and Chen, Z. (2022). Electric vehicle range estimation using regression techniques. *World Electric Vehicle Journal*, 13(6):105.

Al-Wreikat, Y., Serrano, C., and Sodré, J. R. (2021). Driving behaviour and trip condition effects on the energy consumption of an electric vehicle under real-world driving. *Applied Energy*, 297:117096.

Alvarez, A. D., Garcia, F. S., Naranjo, J. E., Anaya, J. J., and Jimenez, F. (2014). Modeling the driving behavior of electric vehicles using smartphones and neural networks. *IEEE Intelligent Transportation Systems Magazine*, 6(3):44–53.

Archsmith, J., Kendall, A., and Rapson, D. (2015). From cradle to junkyard: assessing the life cycle greenhouse gas benefits of electric vehicles. *Research in Transportation Economics*, 52:72–90.

Barnitt, R. A., Brooker, A. D., Ramroth, L., Rugh, J., and Smith, K. A. (2010). Analysis of off-board powered thermal preconditioning in electric drive vehicles.

Burnham, A., Lu, Z., Wang, M., and Elgowainy, A. (2021). Regional emissions analysis of light-duty battery electric vehicles. *Atmosphere*, 12(11):1482.

CEDM (2021). Electricity marginal factors estimates, climate and energy decision making center, carnegie mellon university. https://cedm.shinyapps.io/MarginalFactors/. Accessed: 2024-07-12.

Chen, Y. and Fan, Y. (2013). Transportation fuel portfolio design under evolving technology and regulation: a california case study. *Transportation Research Part D: Transport and Environment*, 24:76–82.

EVO (2024). Electric vehicle outlook. https://about.bnef.com/electric-vehicle-outlook/. Accessed: 2024-07-18.

Fernández, R. Á., Caraballo, S. C., and López, F. C. (2019). A probabilistic approach for determining the influence of urban traffic management policies on energy consumption and greenhouse gas emissions from a battery electric vehicle. *Journal of Cleaner Production*, 236:117604.

Fetene, G. M., Kaplan, S., Mabit, S. L., Jensen, A. F., and Prato, C. G. (2017). Harnessing big data for estimating the energy consumption and driving range of electric vehicles. *Transportation Research Part D: Transport and Environment*, 54:1–11.

Fiori, C., Ahn, K., and Rakha, H. A. (2016). Power-based electric vehicle energy consumption model: Model development and validation. *Applied Energy*, 168:257–268.

Fiori, C., Ahn, K., and Rakha, H. A. (2018). Microscopic series plug-in hybrid electric vehicle energy consumption model: Model development and validation. *Transportation Research Part D: Transport and Environment*, 63:175–185.

GeeksForGeeks (2024). Feature importance with random forests. https://www.geeksforgeeks.org/feature-importance-with-random-forests/. Accessed: 2024-08-05.

Genikomsakis, K. N. and Mitrentsis, G. (2017). A computationally efficient simulation model for estimating energy consumption of electric vehicles in the context of route planning applications. *Transportation Research Part D: Transport and Environment*, 50:98–118.

Giordano, A., Fischbeck, P., and Matthews, H. S. (2018). Environmental and economic comparison of diesel and battery electric delivery vans to inform city logistics fleet replacement strategies. *Transportation Research Part D: Transport and Environment*, 64:216–229.

Heide, C. H. and Mohazzabi, P. (2013). Fuel economy of a vehicle as a function of airspeed: the concept of parallel corridors. *International Journal of Energy and Environmental Engineering*, 4:1–7.

How, D. N., Hannan, M. A., Lipu, M. S. H., Sahari, K. S., Ker, P. J., and Muttaqi, K. M. (2020). State-of-charge estimation of li-ion battery in electric vehicles: A deep neural network approach. *IEEE Transactions on Industry Applications*, 56(5):5565–5574.

Hull, C., Giliomee, J., Collett, K. A., McCulloch, M. D., and Booysen, M. (2023). High fidelity estimates of paratransit energy consumption from per-second gps tracking data. *Transportation Research Part D: Transport and Environment*, 118:103695.

IEA (2024). Electrification of road transport goes beyond cars and the ambition is growing. https://www.iea.org/energy-system/transport/trucks-and-buses#tracking. Accessed: 2024-07-18.

Jiménez, D., Hernández, S., Fraile-Ardanuy, J., Serrano, J., Fernández, R., and Alvarez, F. (2018). Modelling the effect of driving events on electrical vehicle energy consumption using inertial sensors in smartphones. *Energies*, 11(2):412.

Karabasoglu, O. and Michalek, J. (2013). Influence of driving patterns on life cycle cost and emissions of hybrid and plug-in electric vehicle powertrains. *Energy policy*, 60:445–461.

Lee, D.-Y. and Thomas, V. M. (2017). Parametric modeling approach for economic and environmental life cycle assessment of medium-duty truck electrification. *Journal of Cleaner Production*, 142:3300–3321.

Li, Y., Zhang, L., Zheng, H., He, X., Peeta, S., Zheng, T., and Li, Y. (2015). Evaluating the energy consumption of electric vehicles based on car-following model under non-lane discipline. *Nonlinear Dynamics*, 82:629–641.

Maia, R., Silva, M., Araújo, R., and Nunes, U. (2015). Electrical vehicle modeling:

A fuzzy logic model for regenerative braking. *Expert systems with applications*, 42(22):8504–8519.

Marmiroli, B., Venditti, M., Dotelli, G., and Spessa, E. (2020). The transport of goods in the urban environment: A comparative life cycle assessment of electric, compressed natural gas and diesel light-duty vehicles. *Applied Energy*, 260:114236.

Modi, S., Bhattacharya, J., and Basak, P. (2020). Estimation of energy consumption of electric vehicles using deep convolutional neural network to reduce driver's range anxiety. *ISA transactions*, 98:454–470.

Mohammadnazar, A., Khattak, Z. H., and Khattak, A. J. (2024). Assessing driving behavior influence on fuel efficiency using machine-learning and drive-cycle simulations. *Transportation Research Part D: Transport and Environment*, 126:104025.

NASA (2023). July 2023 was the hottest month on record. https://earthobservatory.nasa.gov/images/151699/july-2023-was-the-hottest-month-on-record. Accessed: 2024-07-15.

NREL (2017a). National renewable energy laboratory fleet dna project data.

NREL (2017b). Smith newton vehicle performance evaluation-cumulative. https://www.nrel.gov/docs/fy15osti/61238.pdf. Accessed: 2024-07-08.

NSRDB (2024). Nsrdb: National solar radiation database. https://nsrdb.nrel.gov/data-viewer. Accessed: 2024-07-11.

Onat, N. C., Kucukvar, M., and Tatari, O. (2015). Conventional, hybrid, plug-in hybrid or electric vehicles? state-based comparative carbon and energy footprint analysis in the united states. *Applied Energy*, 150:36–49.

O'Connell, A., Pavlenko, N., Bieker, G., and Searle, S. (2023). A comparison of the life-cycle greenhouse gas emissions of european heavy-duty vehicles and fuels.

O'Keefe, M., Simpson, A., Kelly, K., and Pedersen, D. (2007). Duty cycle characterization and evaluation towards heavy hybrid vehicle applications. sae tech. https://www.nrel.gov/docs/gen/fy07/40929.pdf. Accessed: 2024-07-18.

Pan, Y., Fang, W., and Zhang, W. (2023). Development of an energy consumption prediction model for battery electric vehicles in real-world driving: a combined approach of short-trip segment division and deep learning. *Journal of Cleaner Production*, 400:136742.

Perugu, H., Collier, S., Tan, Y., Yoon, S., and Herner, J. (2023). Characterization of battery electric transit bus energy consumption by temporal and speed variation. *Energy*, 263:125914.

Prohaska, R., Simpson, M., Ragatz, A., Kelly, K., Smith, K., and Walkowicz, K. (2016). Field evaluation of medium-duty plug-in electric delivery trucks. https://www.osti.gov/biblio/1337010. Accessed: 2024-07-10.

Qiu, Y., Dobbelaere, C., and Song, S. (2023). Energy cost analysis and operational range prediction based on medium-and heavy-duty electric vehicle real-world deployments across the united states. *World Electric Vehicle Journal*, 14(12):330.

Sun, S., Zhang, J., Bi, J., and Wang, Y. (2019). A machine learning method for predicting driving range of battery electric vehicles. *Journal of Advanced Transportation*, 2019(1):4109148.

Tamayao, M.-A. M., Michalek, J. J., Hendrickson, C., and Azevedo, I. M. (2015). Regional variability and uncertainty of electric vehicle life cycle co2 emissions across the united states. *Environmental science & technology*, 49(14):8844–8855.

UN (2024). Causes and effects of climate change. https://www.un.org/en/climatechange/science/causes-effects-climate-change#:~:text=Fossil%20fuels%20%E2%80%93%20coal%2C%20oil%20and,they%20trap%20the%20sun's%20heat. Accessed: 2024-07-15.

Vepsäläinen, J., Otto, K., Lajunen, A., and Tammi, K. (2019). Computationally efficient model for energy demand prediction of electric city bus in varying operating conditions. *Energy*, 169:433–443.

Weiss, M., Cloos, K. C., and Helmers, E. (2020). Energy efficiency trade-offs in small to large electric vehicles. *Environmental Sciences Europe*, 32:1–17.

Wisconsin (2022). Wisconsin geospatial news, how big is a degree? https://www.sco.wisc.edu/2022/01/21/how-big-is-a-degree/. Accessed: 2024-07-11.

Woo, J., Choi, H., and Ahn, J. (2017). Well-to-wheel analysis of greenhouse gas emissions for electric vehicles based on electricity generation mix: A global perspective. *Transportation Research Part D: Transport and Environment*, 51:340–350.

Woody, M., Craig, M. T., Vaishnav, P. T., Lewis, G. M., and Keoleian, G. A. (2022a). Optimizing future cost and emissions of electric delivery vehicles. *Journal of Industrial Ecology*, 26(3):1108–1122.

Woody, M., Vaishnav, P., Craig, M. T., and Keoleian, G. A. (2022b). Life cycle greenhouse gas emissions of the usps next-generation delivery vehicle fleet. *Environmental Science & Technology*, 56(18):13391–13397.

Woody, M., Vaishnav, P., Craig, M. T., Lewis, G. M., and Keoleian, G. A. (2021). Charging strategies to minimize greenhouse gas emissions of electrified delivery vehicles. *Environmental Science & Technology*, 55(14):10108–10120.

Woody, M., Vaishnav, P., Keoleian, G. A., De Kleine, R., Kim, H. C., Anderson, J. E., and Wallington, T. J. (2022c). The role of pickup truck electrification in the decarbonization of light-duty vehicles. *Environmental Research Letters*, 17(3):034031.

Wu, D., Guo, F., Field III, F. R., De Kleine, R. D., Kim, H. C., Wallington, T. J., and Kirchain, R. E. (2019). Regional heterogeneity in the emissions benefits of electrified and lightweighted light-duty vehicles. *Environmental science & technology*, 53(18):10560–10570.

Yang, L., Hao, C., and Chai, Y. (2018). Life cycle assessment of commercial delivery trucks: Diesel, plug-in electric, and battery-swap electric. *Sustainability*, 10(12):4547.

Yang, S., Li, M., Lin, Y., and Tang, T. (2014). Electric vehicle's electricity consumption on a road with different slope. *Physica A: Statistical Mechanics and its Applications*, 402:41–48.

Yuksel, T. and Michalek, J. J. (2015). Effects of regional temperature on electric vehicle efficiency, range, and emissions in the united states. *Environmental science & technology*, 49(6):3974–3980.

Yuksel, T., Tamayao, M.-A. M., Hendrickson, C., Azevedo, I. M., and Michalek, J. J. (2016). Effect of regional grid mix, driving patterns and climate on the comparative carbon footprint of gasoline and plug-in electric vehicles in the united states. *Environmental Research Letters*, 11(4):044007.

Zhang, R. and Yao, E. (2015). Electric vehicles' energy consumption estimation with real driving condition data. *Transportation Research Part D: Transport and Environment*, 41:177–187.

Zivin, J. S. G., Kotchen, M. J., and Mansur, E. T. (2014). Spatial and temporal heterogeneity of marginal emissions: Implications for electric cars and other electricity-shifting policies. *Journal of Economic Behavior & Organization*, 107:248–268.

APPENDIX A

Table A.1 Whole Set of Parameters for the NREL Fleet DNA Data

Parameter	Description
'BMU_Mode_SYS'	Battery management mode
'Battery_Current_SYS'	Battery current
'Battery_Voltage_SYS'	Battery voltage
'CT_Air_Con_Current_RD'	An indication of if the AC is in use
'CT_Heater_Current_RD'	Cabin heater current
'GPS_Altitude'	Altitude
'GPS_Latitude'	Latitude
'GPS_Longitude'	Longitude
'GPS_Speed'	Speed
'Highest_Cell_Temperature_SBS1'	Highest battery cell temperature 1
'Highest_Cell_Temperature_SBS2'	Highest battery cell temperature 2
'Highest_Cell_Temperature_SBS3'	Highest battery cell temperature 3
'Highest_Cell_Temperature_SBS4'	Highest battery cell temperature 4
'Highest_Cell_Temperature_SBS5'	Highest battery cell temperature 5
'Highest_Cell_Temperature_SBS6'	Highest battery cell temperature 6
'Highest_Cell_Voltage_SBS1'	Highest battery cell voltage 1
'Highest_Cell_Voltage_SBS2'	Highest battery cell voltage 2
'Highest_Cell_Voltage_SBS3'	Highest battery cell voltage 3
'Highest_Cell_Voltage_SBS4'	Highest battery cell voltage 4
'Highest_Cell_Voltage_SBS5'	Highest battery cell voltage 5
'Highest_Cell_Voltage_SBS6'	Highest battery cell voltage 6
'Lowest_Cell_Temperature_SBS1'	Lowest battery cell temperature 1
'Lowest_Cell_Temperature_SBS2'	Lowest battery cell temperature 2
'Lowest_Cell_Temperature_SBS3'	Lowest battery cell temperature 3
'Lowest_Cell_Temperature_SBS4'	Lowest battery cell temperature 4
'Lowest_Cell_Temperature_SBS5'	Lowest battery cell temperature 5
'Lowest_Cell_Temperature_SBS6'	Lowest battery cell temperature 6
'Lowest_Cell_Voltage_SBS1'	Lowest battery cell voltage 1
'Lowest_Cell_Voltage_SBS2'	Lowest battery cell voltage 2
'Lowest_Cell_Voltage_SBS3'	Lowest battery cell voltage 3
'Lowest_Cell_Voltage_SBS4'	Lowest battery cell voltage 4
'Lowest_Cell_Voltage_SBS5'	Lowest battery cell voltage 5
'Lowest_Cell_Voltage_SBS6'	Lowest battery cell voltage 6
'RD_Ambient_Temp_degC'	Ambient temperature
'RD_Cab_Temp_degC'	Cabin temperature
'SOC_SYS'	State of charge
'Timestamp'	Timestamp
'VS_DCMD'	Accelerator pedal position
'ms_nmot'	Motor speed
'ms_ths1'	Motor temperature sensor 3
'ms_ths2'	Motor temperature sensor 4
'ms_ths3'	Motor temperature sensor 5
'ms_ths4'	Motor temperature sensor 6
'ms_ths5'	Motor temperature sensor 7
'ms_tmf1'	Motor temperature sensor 1
'ms_tmc1'	Motor temperature sensor 2
'vs_24vbat'	24V system voltage
'vs_bcmd'	Brake pedal position

Table A.2 Regression Coefficients for LASSO

Predictor	β	$\lvert\beta\rvert$
Average positive acceleration	0.12795	0.12795
Average speed	-0.12037	0.12037
Average speed - Squared	0.11789	0.11789
Gross vehicle weight	0.09813	0.09813
Average positive acceleration - Reciprocal	0.07067	0.07067
Duration	-0.06785	0.06785
Average ambient temperature	-0.06771	0.06771
Kinetic intensity - Log	0.06758	0.06758
Net elevation change per distance	0.06412	0.06412
Kinetic intensity	0.06392	0.06392
Average positive acceleration - Squared	-0.06031	0.06031
Average ambient temperature - Power 5	-0.05220	0.05220
Sum of positive accelerations - Log	0.05168	0.05168
Average ambient temperature - Cubed	0.05123	0.05123
Sum of positive accelerations - Squared	0.04427	0.04427
Sum of positive accelerations	-0.04385	0.04385
Kinetic intensity - Squared	-0.04079	0.04079
Sum of positive accelerations - Reciprocal	-0.04054	0.04054
Average ambient temperature - Power 4	0.03041	0.03041
Kinetic intensity - Reciprocal	0.02223	0.02223
Maximum speed	0.01678	0.01678
Average speed - Reciprocal	0.00329	0.00329
Average ambient temperature - Squared	-0.00020	0.00020
Average speed - Log	0.00000	0.00000
Average positive acceleration - Log	0.00000	0.00000

Table A.3 Marginal Emission Factors for FRCC [kg-CO_2/MWh]

Hour of Day	**Summer**	**Trans**	**Winter**
0	539	567	491
1	514	552	479
2	497	435	496
3	442	425	477
4	454	512	502
5	447	451	428
6	356	426	405
7	474	473	446
8	499	519	449
9	490	512	468
10	441	462	464
11	480	490	472
12	531	541	482
13	522	483	482
14	511	437	477
15	530	448	491
16	507	464	425
17	556	507	441
18	488	454	455
19	460	469	468
20	492	525	471
21	449	425	488
22	489	480	486
23	468	493	498

Table A.4 Marginal Emission Factors for NPCC [kg-CO_2/MWh]

Hour of Day	**Summer**	**Trans**	**Winter**
0	494	417	440
1	440	394	478
2	415	466	473
3	451	488	477
4	462	421	436
5	403	386	386
6	386	383	367
7	491	432	356
8	481	436	385
9	479	432	450
10	479	448	396
11	461	438	457
12	462	427	411
13	459	428	417
14	443	413	423
15	463	418	424
16	471	431	413
17	484	441	424
18	510	407	445
19	491	457	445
20	515	440	427
21	490	400	425
22	486	410	428
23	451	403	449

Table A.5 Marginal Emission Factors for MRO [kg-CO_2/MWh]

Hour of Day	**Summer**	**Trans**	**Winter**
0	749	709	766
1	832	818	846
2	844	815	871
3	837	790	830
4	843	787	822
5	854	732	766
6	847	755	722
7	809	838	749
8	836	744	841
9	844	761	794
10	741	769	795
11	709	735	786
12	695	729	744
13	775	703	775
14	761	775	783
15	794	768	810
16	771	794	774
17	777	783	764
18	751	757	813
19	735	723	781
20	739	756	788
21	763	745	818
22	756	823	823
23	767	824	860

Table A.6 Marginal Emission Factors for RFC [kg-CO_2/MWh]

Hour of Day	**Summer**	**Trans**	**Winter**
0	690	705	630
1	723	732	699
2	702	667	642
3	629	650	729
4	641	635	659
5	676	606	578
6	695	637	552
7	697	615	619
8	708	627	616
9	672	635	631
10	662	598	636
11	628	631	654
12	597	620	624
13	611	602	622
14	607	608	670
15	641	604	627
16	652	634	561
17	644	652	571
18	617	644	667
19	623	616	654
20	623	631	659
21	635	605	624
22	603	551	613
23	653	515	567

Table A.7 Marginal Emission Factors for SERC [kg-CO_2/MWh]

Hour of Day	**Summer**	**Trans**	**Winter**
0	599	738	559
1	669	657	594
2	664	584	609
3	599	607	603
4	592	594	624
5	645	635	619
6	689	609	595
7	623	633	596
8	674	630	639
9	634	639	584
10	616	600	584
11	620	615	591
12	595	617	582
13	636	618	591
14	621	642	574
15	656	619	570
16	662	647	590
17	632	608	625
18	631	621	581
19	629	651	637
20	626	640	588
21	620	627	628
22	641	592	619
23	625	653	569

Table A.8 Marginal Emission Factors for SPP [kg-CO_2/MWh]

Hour of Day	Summer	Trans	Winter
0	695	637	637
1	751	695	659
2	762	703	712
3	744	734	738
4	679	666	715
5	696	657	667
6	735	592	614
7	724	684	596
8	663	600	694
9	688	639	638
10	652	630	680
11	635	651	675
12	628	641	686
13	609	635	688
14	624	639	683
15	657	598	701
16	631	612	703
17	629	647	637
18	626	640	642
19	601	633	718
20	618	637	745
21	626	644	695
22	648	638	724
23	680	605	656

Table A.9 Marginal Emission Factors for TRE [kg-CO_2/MWh]

Hour of Day	**Summer**	**Trans**	**Winter**
0	649	588	606
1	702	640	630
2	664	676	648
3	659	751	658
4	656	677	655
5	659	625	577
6	593	542	530
7	684	692	569
8	632	544	572
9	648	592	577
10	639	581	595
11	582	531	565
12	524	543	520
13	512	537	541
14	514	525	576
15	506	552	588
16	513	573	605
17	515	549	581
18	504	564	555
19	527	527	628
20	514	571	641
21	485	549	594
22	510	547	556
23	617	550	553

Table A.10 Marginal Emission Factors for WECC [kg-CO_2/MWh]

Hour of Day	Summer	Trans	Winter
0	513	534	543
1	472	501	529
2	504	511	479
3	512	596	498
4	565	592	525
5	570	606	560
6	634	584	585
7	607	533	561
8	587	551	569
9	545	567	545
10	576	519	552
11	599	482	532
12	572	553	538
13	554	606	524
14	529	528	518
15	511	559	593
16	516	591	546
17	556	551	567
18	546	516	534
19	579	518	529
20	583	581	601
21	595	592	570
22	614	644	568
23	542	605	569