ABSTRACT

Title of Dissertation:	DEVELOPING A STATISTICAL VEHICLE DRIVER BEHAVIOR MODEL FOR ECO-ROUTING DEPLOYMENT
	Weiyi Zhou, Doctor of Philosophy, 2022
Dissertation Directed By:	Professor Lei Zhang, Department of Civil and Environmental Engineering

Predicting energy consumption accurately and reliably is critical for route optimization in eco-routing. State-of-the-practice methods for calculating energy consumption utilize second-by-second speed, acceleration, and power demand. Such models can achieve high accuracy but are not suitable for forecasting usages due to strict requirement of inputs and computing resources. Other methods used to predict energy consumption rely on average speed data to reduce data collection and computation efforts. However, they ignore the individuality of driving behavior, which is particularly important in near-term predictions of energy consumption, as shown in this paper. This study develops an input-output hidden Markov model (IOHMM) to cope with the influence of external environment and driving behaviors on individual driving features. The model is built and trained using passively collected geospatial location data. The approach furthermore improves the prediction of vehicle specific power (VSP) distribution, a critical parameter for energy predication, through predicted driving features. The model is tested in the Washington D.C. metropolitan area, and the performance is evaluated by comparing various indicators with the real-world values obtained from in-vehicle fuel recording devices. In general, the IOHMM behavior model demonstrates an overall cruising speed accuracy of 86.85% and acceleration rate accuracy

of 82.73%. The behavior-integrated energy prediction model outperforms the traditional approaches by increasing the energy prediction accuracy to 86.81%. Results obtained from this study corroborate the importance of behavioral richness, environmental dynamics, and computation efficiency.

DEVELOPING A STATISTICAL VEHICLE DRIVER BEHAVIOR MODEL FOR ECO-ROUTING DEPLOYMENT

by

Weiyi Zhou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Lei Zhang, Chair
Professor Yan Li
Professor Ali Haghani
Professor Paul Schonfeld
Professor Benjamin Kedem

Acknowledgements

This work would not have been possible without the great institution of the University of Maryland. Go Terps! It is truly an honor of a lifetime to meet so many caring, loving, and inspiring colleagues, staff, and professors at this magnificent place.

I am especially indebted to my advisor, Dr. Lei Zhang, at the Civil & Environmental Department at the University. Dr. Zhang's caring attitude, in-depth knowledge, effective teaching, and continuous encouragement, along with financial support through a research assistantship, have enabled me to reach my goal. I would be at a loss without Dr. Zhang's guidance.

I would also like to express my sincere appreciation to Professors Paul Schonfeld, Ali Haghani, Yan Li, and Benjamin Kedem for serving on my research thesis committee and offering me their valuable comments and encouragement on my research.

I want to thank all my teammates in Dr. Zhang's research group. Thank you for the ideas you have offered, the comments you have shared, and more than anything else, the friendship we have built.

Last, I would like to extend my heartfelt indebtedness to my parents. There are no words that can truly express my appreciation for their love. I love you!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	**Artificial Intelligence**
ARTEMIS	**Assessment and Reliability of Transport Emission Models and Inventory Systems**
ATIS	**Advanced Traffic Information System**
ATM	**Active Traffic Management**
BPTT	**Back Propagation Through Time**
CATT	**Center of Advanced Transportation Technology**
COPERT	**Computer Programmer to Calculate Emissions from Road Transport**
DNN	**Deep Neural Network**
DOE	**Department of Energy**
DOT	**Department of Transportation**
DQL	**Deep Q-Learning**
DTA	**Dynamic Traffic Assignment**
EPA	**Environmental Protection Agency**
EF	**Emission Factor**
EIA	**Energy Information Administration**
FHWA	**Federal Highway Administration**
GHG	**Greenhouse Gas**
GPS	**Global Position System**
HBEFA	**Handbook on Emissions Factors for Road Transport**
HMM	**Hidden Markov Model**
IAT	**Intake Air Temperature**
IOHMM	**Input Output-Hidden Markov Model**
ITMS	**Internet Traffic Monitoring System**
LSTM	**Long Short-Term Memory**
MAP	**Manifold Pressure**
MIT	**Massachusetts Institute of Technology**

MOVES	**Motor Vehicle Emission Simulator**
MTI	**Maryland Transportation Institute**
NN	**Neural Network**
OBU	**On-board Unit**
OD	**Origin-Destination**
RL	**Reinforcement Learning**
RITIS	**Regional Integrated Transportation Information System**
RTRL	**Real Time Recurrent Learning**
TEAD	**Transportation Energy Analytics Dashboard**
VSP	**Vehicle Specific Power**
VMT	**Vehicle Miles Traveled**
VOC	**Vehicle Operating Cost**
VOT	**Value of Time**
WMSE	**Weighted Mean Squared Error**
UMD	**University of Maryland**

Chapter 1 Introduction

1.1. Background

In an era of accelerated urbanization around the world, the ability to travel freely is more critical than ever before. Travel is no longer just for commuting to and from work. As a matter of fact, travel for other purposes such as entertainment, vacation, leisure activities, running errands, and shopping has long surpassed work-related trips. While trips through modes such as public transportation (e.g., bus, rail), transportation network companies (e.g., Uber, Lyft), bicycles, shared bikes, and scooters have increased, trips carried out through personally owned vehicles are still the predominant method of travel. This phenomenon results in continued increases on fuel demand, leading to increases in energy consumption, greenhouse gas (GHG) emissions, and air toxic emissions [2-4]. According to the Inventory of U.S. Greenhouse Gas Emissions and Sinks 1990-2017 (A national inventory that the U.S. Department of Energy prepares annually under the United Nations Framework Convention on Climate Change), transportation accounted for the largest portion (29%) of total U.S. GHG emissions in 2017 [5].

In the past few decades, researchers have developed a wide range of technologies and practices to mitigate vehicle travel congestion and improve air quality. On the public agency front, these efforts include the continued improvement of roadway infrastructure such as more efficient and interconnected roadway intersection controls, highly effective freeway ramp metering and variable speed limit managements, focused investment on new modes of transportation,

carpooling/vanpooling promotion, and HOV/HOT deployment to reduce single occupancy vehicles [5, 6]. On the private business and research front, trip planning and navigation apps developed and installed through smart phones and GPS devices are used by the public to enable efficiency for an individual traveler, achieving their goals of minimizing travel time or travel distance. Such apps analyze roadway conditions with near real-time inputs and make recommendations on the shortest route to travel from a driver's origin to their destination, which includes the least travel time paths, least travel distance paths, or toll-free ways.

While these apps primarily focus on travel distance, travel time, and toll, functional apps designed for fuel and emissions considerations by individual traveler are not widely used. Various research states the path with shortest travel time or shortest travel distance may not be the eco-route, especially under complicated traffic conditions. Regional-level research on energy and emission indicates the trip fuel efficiency can be increased up to 25% through alternative routing. The entire system can benefit from the eco-routing system with the fuel reduction ranges between 3.3% and 9.3% compared to traditional shortest-travel-time strategies [6].

As it stands now, the focus of both public entities and private businesses is congestion mitigation: reducing travel distance and travel time. Recently, several trip-based computation methods and algorithms on energy consumption have emerged, such as eco-routing and eco-driving. Eco-routing suggests the optimum path with minimum energy consumption, while eco-driving improves drivers' driving behavior [7-16]. Both methods alleviate concerns associated with energy consumption,

greenhouse gas emissions, and general air quality. Various research analyzes the benefits of eco-routing and suggests the overall reduction of energy consumption is considerable [17-23].

1.2. Vehicle Energy Consumption Modeling

Vehicle energy consumption estimation modeling has been researched with varying coverage and focus. Most macroscopic models take energy consumption as a function of traffic features, such as traffic speed, traffic flow density, and roadway types. Various bins are established with pre-defined energy consumption rates to reflect different driving modes. Some models also use driving features (e.g., average cruising speed) and vehicle features (e.g., vehicle type and vehicle age) to determine the energy consumption rates. The macroscopic models are easy to deploy but not fully consider the influence of the dynamic driving influences on energy consumption.

An energy consumption estimation model at microscale has gained significant attention, benefitting from the advancements in modern electronics, which enable data acquisition on a second-by-second scale. Detailed trip trajectory data with driving dynamic is recorded, which helps to estimate energy consumption more accurately. The two representative types of microscopic models are cycle-based models and modal-based models. Cycle-based models use driving features (e.g., number of stops, acceleration and deceleration ratios, speed variation, and braking time) as inputs to calculate energy consumption. A trip is separated into several cycles based on driving features, and energy consumption for each cycle is calculated. Modal-based models classify vehicle instantaneous operation modes into different bins, which are

determined by a second-by-second power demand called vehicle specific power (VSP) and cruising speed. VSP can be formulated as a function of speed, acceleration, vehicle mass, energy power, resistance factors, and rotation factors. Both models fully consider the influence of driving dynamics on energy consumption. Although models at microscale are highly accurate, low availability of input data and strict requirement of computing power prohibit their practical application for eco-routing. Specifically, the microscale models require detailed driving dynamics at a second-by-second level, which are usually hard to acquire at a roadway system or through area-wide levels. In addition, rates of energy consumption for various bins must be pre-calculated and stored in a database for any real-time application, considering the high computation demand.

Both macroscale and microscale models are not suitable for prediction purposes in an eco-routing system. To bridge the gaps between macroscale and microscale models, mesoscopic models based on VSP distribution have emerged. Recent research indicates that VSP on a link follows Gaussian distribution, whose mean and standard deviation can be estimated using average traffic speed and road type. The VSP-based models have simple structures and lower requirements for the inputs but consider the influence of driving dynamics.

1.3. Driving Behavior Modeling

Driving behavior covers a lot of parameters, including both macroscale phenomena and microscale actions. Macroscale travel behavior includes the number of trips a person makes in a day, modes the person uses, and purposes of the person's trips.

Microscale actions cover how a driver operates their vehicle. In the discussion here, travel behaviors refer to the microscale parameters related to how the driver completes a vehicle trip. Parameters of the driver's emotion, experience, pressure, reaction to emergency traffic, and other unquantifiable factors are referred to as driving behavior. The microscale driving behaviors discussed above are mainly influenced by traffic conditions (e.g., degree of congestion and weather) and roadway geometry features (e.g., both horizontal and vertical curvatures, lane width, median availability, and shoulder availability), as revealed by past research [19-21]. Meanwhile, both driving behavior and external environment influence driving features: for instance, average cruising speed, acceleration rate, and deceleration rate. Driving under low-traffic volumes on an uncongested, properly designed, and constructed roadway is more likely to a cruising experience with fewer sudden acceleration and deceleration actions. On the other hand, making a left turn without protected left-turn signal control may increase a driver's pressure, resulting in harsh acceleration and deceleration. Except the external environment, individuality of driving behaviors also leads to varied actual driving features and travel experiences. An example is that two drivers driving under the same external environment (same congestion level and same roadway) can present extremely different driving features, even with similar average speed. A new driver with poor driving experience may drive very cautiously with frequent deceleration under heavy traffic, while an experienced driver can operate the vehicle more smoothly.

Though the driving behavior is highly related to the individual, it can be improved through various approaches. For example, an eco-driving assistance system can recommend to the driver based on real-time operating conditions, improve driving

behaviors, result in changes of driving features, and ultimately lead to higher fuel efficiency. Additionally, more appropriate routes for lower energy consumption can be suggested to the driver based on individual driving behaviors. The deployment of such approaches shows that driving behavior changes could significantly increase fuel efficiency and reduce energy consumption and emission reduction.

1.4. Objective

The objective of this study is twofold. The first is to establish quantitative relationships among external environment, driving behaviors, and driving features. The second objective is to deploy these relationships to gain a preliminary understanding on their utility.

The study aims at developing an applicable mesoscopic model to predict energy consumption with high accuracy for an eco-routing system. The applicable model integrates the individual driving behavior model and VSP-based energy prediction model. The relationship between external environment, individual driving behavior, and driving features is studied, then quantitated with a well-defined statistical model. The driving behaviors and features predicted by the statistical model can be utilized to estimate more accurate VSP distribution for the VSP-based energy model.

External environment refers to traffic flow conditions, such as the degree of congestion and roadway features such as road type, speed limit, grade, curvature, and traffic control. Driving behavior refers to how drivers react to the external environment changes, which includes driving experience, reaction time, pressure, emotion, and other unquantifiable factors. Driving features are the result of external environment and

driving behavior. It includes information on cruising speed, acceleration rate, deceleration rate, continuous acceleration or deceleration, and speed variance.

The relationships are analyzed statistically through the generative approach of Input Output Hidden Markov Model (IOHMM), whose structure is consistent with the relationships demonstrated above. The model takes external environment variables as inputs, driving features as outputs, and driving behaviors as hidden states. The collected trip trajectory data is utilized to train the model to predict driving behaviors and driving features. The model outputs are then used to improve energy consumption prediction model for eco-routing system by amending the VSP distribution based on the driving features.

Compared with other related research, the present study makes contributions from the following aspects.

First, this study considers the driving behavior impact in energy consumption prediction. The study analyzes the relationship between external environment, driving behavior, and driving features using drivers' historical trip trajectory data. Then, the relationship is quantitated using a well-defined statistical approach to predict driving features, which are then utilized to achieve more accurate VSP distribution, thus, to improve energy consumption prediction.

Second, this study quantitates the relationship between external environment, driving behavior, and driving features in an innovative way. The route sequence is considered as a Hidden Markov Process with states transition, which is modeled

utilizing Input Output-Hidden Markov Model. The driving behaviors are included as hidden states of the Input Output-Hidden Markov Model to cope with the impact on driving features and energy consumptions.

The proposed behavior-integrated energy prediction model proves to be an appropriate approach for eco-routing system. The model does not require detailed driving dynamics and powerful computing recourses. Instead, the driving features can be predicted through the personalized IOHMM model, which is trained in advance using historical trip trajectories.

The study explores and successfully develops a comprehensive and innovative effective framework to integrate the proposed energy prediction model with traffic simulation models. The IOHMM approach offers the ability to generate different sequences of driving behaviors and driving features, which can be further integrated into traffic simulation models for system benefits analysis on energy consumption.

1.5. Paper Organization

The remainder of the dissertation is organized as follows. The state-of-the-practice energy modeling approaches and driving behaviors analysis are summarized in Chapter 2. Chapter 3 covers the proposed methodologies. The experimental study is demonstrated in Chapter 4. Chapter 5 discusses the extension of the proposed model in real-world applications, such as eco-routing and system energy monitoring. The final chapter ends the paper with conclusion and further research directions.

Chapter 2 Literature Review

2.1 Energy Modeling Approaches

2.1.1. Microscopic models

Microscopic models calculate the vehicle-level energy consumption and emissions by tracking second-by-second driving statistics that represent vehicles' real-time operational conditions. Among various microscopic models, cycle-based models are the simplest. Such models calculate emissions by recording driving characteristics (e.g., number of stops, vehicle miles traveled (VMT) or maximum acceleration) of the entire driving circle [24-26]. Cycle-based models are not considered as common strategies since the summary characteristics cannot fully represent all driving conditions. Regression models are another type of methods dealing with energy estimation. Generally, the regression models take second-by-second driving statistics as inputs and consider the energy consumption as a function of driving variables during an event (e.g., acceleration, deceleration, and idling) coupled with roadway geometry features (e.g., grade and curvature) [27-29]. Compared with cycle-based models, the regression models are of more details, and aim at calculating vehicular emissions and energy consumption at a second-by-second resolution. Another highly representative category at microscopic scale are modal-based emission models. Modal-based models calculate vehicle emissions based on engine-operating modes from detailed speed and acceleration [30, 31]. The Motor Vehicle Emission Simulator (MOVES), developed by the U.S. Environmental Protection Agency (EPA), is the most widely accepted modal-based models [32-36]. The latest version of MOVES was officially released in 2013,

fully replacing the previous version, namely MOBILE [37]. In MOVES, vehicle energy consumption and emissions are described as a combining effect of two factors – emissions sources and vehicle operating modes. As for emissions sources, various bins are categorized by vehicle characteristics such as vehicle type, fuel type, vehicle age, model, engine technologie, and average weight fraction. Operating modes refer to vehicle operating conditions, represented by a function of Vehicle Specific Power (VSP). VSP is a formalism used in the evaluation of vehicle emissions, which represents the combination of loads resulting from various vehicle manufacturing and operation parameters, including aerodynamic drag, acceleration, speed, rolling resistance, mass, and slope of speed [38-43]. MOVES model calculates the VSP at second level through Equation 1 [36]:

$$VSP = \frac{1}{M} \times (A * v + B * v^2 + C * v^3) + a * v + sin\theta * v \quad (1)$$

where v refers to the speed (m/s), a denotes acceleration (m/s^2) and θ is grade. A, B, and C stand for rolling term (metric ton), rotation term (metric ton/(m/s)) and drag term (metric ton/(m/s)^2), respectively. M refers to vehicle mass (metric ton).

As a practical model with detailed emissions factors, MOVES can conduct energy and emissions analysis on large-scale traffic networks with complex vehicle compositions. Although models of this type fully capture the influence of driving dynamics on energy and emissions, they require driving data at high resolution and detailed emissions rates for various combinations of source bins and operating mode bins. These data usually require a huge amount of memory for storage and powerful

computation resources for calculation, which are impracticable for real-time applications, especially for eco-routing system.

2.1.2. Macroscopic models

Macroscopic models, on the other hand, taking traffic flow properties as inputs to represent overall traffic conditions, estimate fleet emissions at regional or state level. Original studies of this type tried to estimate emissions using aggregates emissions factors (EF) for various vehicle types operating in different driving conditions [57, 58]. Traffic situations models and traffic variables models treat the traffic as a whole and use category-specific emissions factors to calculate emissions and fuel consumption [44, 47]. The difference between the two is that traffic situations models classify traffic conditions into pre-defined categories while traffic variables models formulate regional emissions as function of average traffic speed or traffic density. Research work carried out to analyze the relationships between driving dynamics and emissions/energy consumption, concluded that vehicle-specific emissions factors could be represented by a function of average cruising speed [47-49]. Because its ability to estimate emissions at both trip and regional levels, average speed models quickly came out ahead. Some well-known applications include COPERT (Computer programmer to calculate emissions from road transport) and HBEFA (Handbook on emissions factors for road transport [50, 51]. Research also suggested that levels of congestion and regional environmental impact should be considered in emissions calculation [52, 53]. To overcome the limitation of average speed model, Smit et al. (2008) replaced the average cruising speed with mean speed distribution on roadway links to better represent the impact of traffic conditions [54]. Results showed that all emissions

indexes presented noticeable increase when using mean speed distribution, which indicates the original average speed models likely underestimate the emissions. Tsanakas et al. (2017) reported similar conclusions by quantifying errors of average speed models and stated that using average speed could result in an underestimation of emissions [55, 56]. In Tsanakas' work, a post-processing based on quasi-dynamic approach was proposed in order to filling the gap of missing emissions by traffic dynamics.

2.1.3. Mesoscopic models

Obviously, both microscopic and macroscopic models have their drawbacks in various aspects. Microscopic models can produce accurate estimations but requires detailed travel dynamic data and consumes large amount of computation resources. Macroscopic models, on the other hand, is simpler and faster regarding computation but offers no consideration to the influence of travel dynamics. Over the last decade, more researchers contributed to improving emissions models regarding both their accuracy and efficiency. The vehicle specific power arises as a more representative variable that more broadly reflects the influence of travel dynamics. Jimenez (1999) first introduced the concept of VSP, which is the evaluation of vehicle emissions as the sum of the loads divided by the mass of the vehicle [59]. As mentioned earlier, some mobile source emission models (e.g., MOVES, PHEM) take VSP as the primary variable for predicting second-by-second emissions and show promising results.

However, these models are not suitable for real-time application. Instead of calculating for every second, some studies suggested the VSP distribution could be

formulated as a Gaussian distribution function of average cruising speed [60-63]. Similar findings were shown in the study of Li et al. (2016) by deriving the relationship between emissions factors and average cruising speeds based on an intermediate variable of VSP [64]. Moreover, Li also indicated that the relationship was obviously different on different types of roads and speed ranges. Yao et al. (2013) analyzed the impact of freeway grades and time-of-day factors on the characteristics of VSP using collected GPS data from different periods [65]. The results demonstrated that both variables significantly influence the variation of VSP distribution and should be considered in VSP-based emission models. Quaassdorff et al. (2016) integrated the microscopic traffic model VISSM with a VSP-based emissions model [66]. Different scenarios with various traffic conditions were simulated through the VISSIM model to provide vehicle-driving patterns as inputs for emissions model. While to some extent the integration of dynamic approaches covers both the spatial and temporal congestion pattern, the computation heavily relies on the availability of high-resolution data, both in terms of demand and calibration data.

2.2. Behavior Modeling Approaches

Several energy models introduced above were tested for eco-routing applications to predict route-based energy consumptions. However, all the applied models are in a simple structure with limited traffic parameters at aggregated level. Brundell and Ericsson (2005) analyzed the influence of various trip features resulted from different driving behaviors on energy consumption and concluded that frequent speed changes and continuous acceleration usually result in higher energy consumption [67]. Their study also analyzed the difference of these trip features between different

drivers. Results suggested that driver micro-level driving behaviors and traffic controls are significant contribution factors to trip features. Zheng et al. (2018) investigated the influence of driving behavior on vehicle emission through driver behavior questionnaires. The investigation found that drivers reported higher fuel consumptions are most likely exhibiting aggressive micro-level driving behavior. A study was conducted by Zhang et al. (2013) on the impact of various socio-demographic characteristics and driving behaviors on fuel efficiency. The study indicated that fuel efficiency is highly related to several factors, for instance, purpose, age and gender [68].

Various approaches have been developed to improve driver micro-level driving behaviors through training simulators or on-board assistant systems. Training simulators help drivers to get familiar with correct driving skills, while on-board assistant systems provide real-time suggestions on driving behaviors during a trip. By comparing a drivers' performance before and after the improvement programs, Guan et al. (2012) claimed the fuel efficiency could be obviously improved after maneuvering skills were fully improved [70]. Walnum et al. (2015) conducted a similar case study on heavy-duty truck drivers [69]. Truck drivers received real-time suggestions from an on-board assistance system and adjusted their driving behaviors according to the suggestions. A significant reduction of truck energy consumption and emissions were observed when drivers changed the pattern of using running idle, avoided frequent acceleration and deceleration, and decreased rolling without engine load. The study also suggested the results are specific to heavy-duty trucks, and the influence of behaviors varied between different vehicle classes, roadway types, and traffic conditions. Studies also showed the relationship between micro-level driving

behaviors and emissions and energy consumption using trajectory data collected from different drivers. It is highlighted that an experienced driver could travel in a highly saturated flow condition and kept fuel and consumptions and emissions low. Jimenez et al (2018) integrated the energy prediction model in a deep neural network to classify driving behavior into three types: turning, acceleration, and braking [71]. The significant improvement in classification indicates high correlation between energy consumption and the micro-level driving behavior. A lot of other research prove the individuality in driving behavior and the influence on driving features and fuel efficiency [72-98]. Some of the related research is summarized in Table 2-1.

Table 2-1: Summary of Previous Studies on Driving Behavior

Features	Related Studies	Findings
Road features: road type, speed limit, number of lanes, grade, intersections	• Van Der Horst, R., & De Ridder, S. (2007) • Zhao, Y. et al. (2018) • Boggio-Marzet, A. et al. (2021)	• local streets show 37% less vehicle's energy efficiency compared with highways. • Drivers' reactions to the traffic signal vary in gender, age, and vehicle type.
Traffic features: traffic speed, traffic volume	• Steg, L. et al. (2011) • Hu, L. et al. (2021)	• Driving behavior varies under highly demanding traffic environment • Older drivers drives more cautious on high-speed roads.
Vehicle features: type, age, fuel type	• Hu, L. et al. (2021) • Paleti, R. (2010) • Varella, R. A. et al. (2019)	• Driver with large SUV are easier to show aggressive behaviors • More Frequent acceleration and deceleration are observed in new vehicles
Driver features: age, emotion, press, driving years	• Magaña, V. C. et al. (2021) • Dogan, E., et al.(2011) • Hu, L. et al. (2021) • Paleti, R. et al. (2010)	• Eco-driving style is hard to be kept by the driver if the trip has stressful goal or time pressure. • Significant relationship can be observed between age, gender, and driving behaviors.
Other features: weather, humidity	• Colonna, P. et al. (2016) • Yan, X., & Wu, J. (2014) • Li, X et al. (2015)	• Different driving behaviors under severe weather increase the ratio of accidents.

2.3. Sequence Modeling

General energy prediction models usually consider a road segment as an independent frame. However, the continuous transmission and transformation of driving behaviors happen between different stages of a trip. Therefore, this study considers a route as a sequence of stages, where each stage refers to a road segment. The link- and route-based energy consumption can be estimated through sequence modeling approaches. Sequence data attracts many researchers in various fields, for instance, speech recognition, activity recognition, and sentiment analysis [99-106]. Various sequence modelling approaches has been developed and deployed to cope with sequence data, of which the two most popular ones are Recurrent Neural Network (RNN) and Hidden Markov Model (HMM). Ycart and Benetos (2017) applied the Long Short-Term Memory (LSTM) on modelling polyphonic music sequences [106]. The study analyzed the influence of various parameters on training process and prediction performance. Sequence labeling is critical for various sequence modeling problems. Jagannatha and Yu (2016) utilized RNN in sequence labeling to improve the exact phrase detection of various medical entities [107]. HMM approaches have similarities in basic principles and structures with RNN, but the two models have significant difference in latent variables [108, 109]. In RNN, a latent variable is deduced deterministically from the latent variables and historical observations in the previous stages. On contrary, HMM considers the sequence data as a Markov Chain that a latent variable is only determined by the latent variable in the last stage. Leveraging the depth video sequences of spatial temporal features, Kamal et al. developed a modified HMM for human body detection and recognition [109]. Similar analysis on human activity

recognition (HAR) is conducted by Lee and Cho (2011) and Segundo et al. (2016) using HMM approaches [111, 112]. Integrated the HMM with decision tree (DT), Reddy et al. (2010) developed a classification approach to separate various transportation modes, for instance, walking, driving, and biking [113]. The integrated approach proved to achieve an accuracy of 93.6%. Though HMM proves to be reliable approaches for sequence modelling, its performance is sometimes limited by the homogeneous assumption [114, 115]. Input Output-Hidden Markov Model (IOHMM), a variant of standard HMM, is proposed based on the non-homogeneous assumption that emission and state transition probabilities depend on inputs. Yin et al. (2018) developed an IOHMM to analyze travelers' daily travel pattern using collected call detail records (CDR) data [115]. The trained IOHMM is also capable to generate travel activity sequences for travel demand modelling. The work is extended by Lin et al. (2017) that integrated the IOHMM with LSTM to estimate location choice of the predicted activities [116]. Some recent studies innovatively integrated the RNN and HMM to enhance the performance in sequence modeling. An obvious example is the iterative re-alignment approach developed by Koller et al. (2017) [117]. Koller's team first trained end-to-end CNN-BLSTMs with considerable improvement in language and gesture recognition. The CNN-BLSTMs are then embedded into HMM to corrects labeling accuracy and improve the performance. Except the CNN-BLSTM, some other machine learning methods are integrated with HMM for better performance, for instance, HMM-DNN, HMM-GMM and HMM-RF.

2.4. Summary

This review underlines the difficulty and drawbacks of existing energy model, which can be summarized as follows: First, although vehicle emission is highly related to average cruising velocity, neither this single variable nor the mean speed distribution can adequately reflect a vehicle's operation under dynamic traffic conditions. Other parameters such as speed distribution, acceleration distribution, road configuration, traffic volume, congestion level and weather condition should be considered in emission and energy consumption calculations, especially for short-term prediction. Second, data source is another aspect preventing the development of a second by second in-situ link-based emission model. In-vehicle GPS data has great details and accuracy, but it is difficult to obtain a large size of data that offers complete driving behavior, leading to different trip patterns. Additionally, individual travel characteristics collected by microscopic simulation models have been found to be of questionable regarding its reliability. Random and volatile data on traffic congestions show poor fitness and low accuracy when used for forecasting through simulation models. Third, there is a relationship between average cruising speed and VSP distribution. However, the single average speed does not represent the dynamic traffic condition in real world. Studies has revealed that trips with the same average trip speed, the differences in speed-acceleration distribution are significant, resulting in both significantly different emissions and energy consumptions.

With these concerns, the purpose of this study is to explore the quantitative relationships among external environment, travel behavior, and trip features to predict

the link-level driving features based on real-time accessible traffic variables, thus, to predict short-term emissions and energy consumption at both link and system scales.

Compared with traditional VSP-based models using only average speed or mean speed distribution, the new model may not be the simplest and fast in prediction traffic emissions but offers more accurate and reliable data regarding emission and energy consumption.

The new model enables the understanding and explains the relationships among speed, and acceleration distributions associated with traffic activities and roadway geometric features. In addition, this new model study offers a robust short-term emission prediction model with high accuracy. With this new established modeling approach, both real-time emission estimation and energy consumption quantification and short-time emission and energy consumption prediction, covering both individual vehicle rerouting and large-scale network monitoring, can be carried out.

Chapter 3 Methodology

3.1. Data Introduction and Processing

3.1.1. Location Data Types

Mobile device location data (MDLD) with driving information is widely used in travel demand and travel pattern analysis. MDLD mainly includes three different types: GPS-enhanced travel survey data, call detailed record (CDR) data, and location-based service (LBS) data.

GPS-enhanced travel survey data is collected through on-board GPS loggers and further verified and amended by users' feedback through user feedback systems. CDR data is collected through the communication between smartphones and cell towers. As long as the density of cell towers is high, the collected CDR data could be in high reliability and accuracy. Similar with CDR data, LBS data is mainly collected when a mobile phone communicates with an app, and the communication is captured and recorded by various medias, for instance, Wi-Fi, Bluetooth, cell towers, or GPS loggers. Given the fact that a large portion of the population use mobile devices with various APPs, and the communication media sources have great coverage across the nation, the LBS data outperforms in spatial and temporal coverages, data quality, and user coverage over the other two data types.

3.1.2. Data Collection

Two sets of data gathered from UMD's incenTrip and i2D are utilized in this study. incenTrip (incentrip.org) is a multimodal trip-planning APP developed and

deployed by the National Transportation Center (NTC) at UMD. The APP is developed as a product for the 'Integrated, Personalized, Real-time Traveler Information and Incentive (iPretii)' project, funded by the U.S. Department of Energy (DOE). The app was officially launched in August 2019, supported by the Metropolitan Washington Council of Government (MWCOG), incentivizing travelers to choose eco-friendly travel options (e.g., transit, multimodal, biking, walking, and carpooling). The service area of incenTrip is illustrated in Figure 3-1.

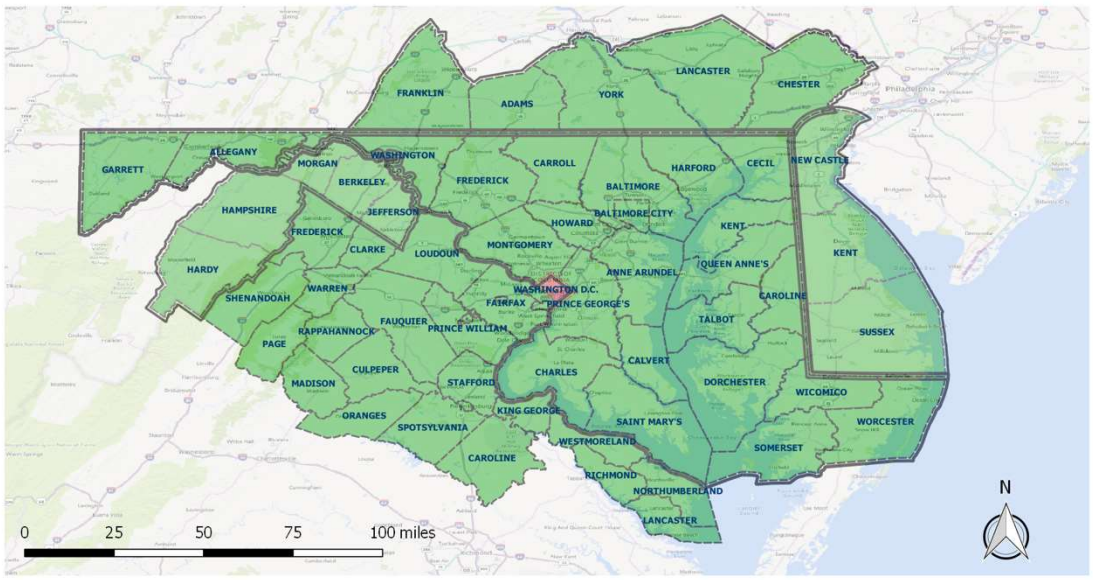

Figure 3-1: incenTrip Service Area

Users can plan trips through the APP, which records the executed trip trajectories with GPS points at 1HZ. Additionally, the APP also records unplanned trips using background logging system and trip identification algorithms in order to track and incentivize users' behaviors changes. GPS location data with driving statistics (i.e., timestamp, speed, and acceleration) are collected from Google Maps API with a pre-

defined sample rate (i.e., 1HZ) and stored through Amazon Web Services (AWS) for privacy and security concerns. When creating an incenTrip account, users also provide vehicle-related information, such as vehicle type, vehicle age, and fuel type. This study queries 5,269 driving trips from October 2019 to May 2021 to support the experimental studies. Detailed description of the collected data is summarized in Table 3-1.

"i2D" data was collected through an on-board logging system, which is consists of a unit, a communication system, and a cloud database. The unit collects multiple engine and travel dynamics statistics in real time at a resolution of 1 Hz through a GPS sensor along with a 3D accelerometer. The logged trips are uploaded to the online database for users queries and safety purpose. For each trip, the logging system collects trajectory (e.g., grade, longitude, and latitude), travel dynamics (e.g., speed and acceleration), fuel consumption statistics (e.g., cumulated fuel and fuel efficiency), and environmental factors (e.g., intake air temperature, humidity, and manifold pressure) in real time. Sixteen testers are involved in the trip collection process and 2,812 trips with the greatest level of detail were collected. A data summary is provided in Table 3-2. Vehicle information, such as age, type, mass, and power, is also stored in the cloud database.

Table 3-1: Summary of Trips Collected from incenTrip Application

Characteristics	Sedan	SUV	Van	Truck
Number of trips collected	*4326*	*777*	*114*	*52*
Number of drivers recorded	*192*	*45*	*8*	*9*
Average vehicle age (year)	*6.55*	*5.27*	*8*	*11*
Number of days recorded	*511*	*112*	*41*	*29*
Number of road segments covered	*52141*	*36114*	*6013*	*3114*
Total of vehicle miles travelled (mile)	*55996*	*13978*	*6294*	*423*
Total of vehicle hours travelled (hour)	*2210*	*582*	*43*	*30*
Percentage of VMT in peak period	*70%*	*68%*	*62%*	*54%*
Average trip length (min)	*30.65*	*45.01*	*22.66*	*35.01*
Average trip distance (mile)	*12.94*	*17.99*	*8.10*	*8.14*

Table 3-2: Summary of Trips Collected from i2D system

Characteristics	Sedan	SUV
Number of trips collected	*2434*	*378*
Number of drivers recorded	*14*	*2*
Average vehicle age (year)	*6.42*	*5.5*
Number of days recorded	*554*	*217*
Number of road segments covered	*34541*	*6324*
Total of vehicle miles travelled (mile)	*22344*	*7634*
Total of vehicle hours travelled (hour)	*1176*	*214*
Percentage of VMT in peak period	*63%*	*48%*
Average trip length (min)	*29*	*34*
Average trip distance (mile)	*9.18*	*11.84*

Taking advantage of detailed driving dynamics and trip trajectories, the data collected from the incenTrip APP and i2D system are used to analyze the relationship among external environment, driving behaviors, and driving features. To match the trip trajectories with roadway geometry information, a statewide network is extracted from HERE navigation network with the same coverage to the incenTrip service area, as demonstrated in Figure 3-2.

The network consists of 215,202 nodes and 366,464 links. These links are identified by their geometrics such as number of lanes, speed limits, facilities, intersections, and traffic controls. One point must be emphasized regarding the number of lanes. The lanes in the HERE navigation network includes normal travelling lanes, left-turn lanes, acceleration and deceleration lanes, and special-purpose lanes. In this study, only the normal travel lane and left-turn lanes are considered.

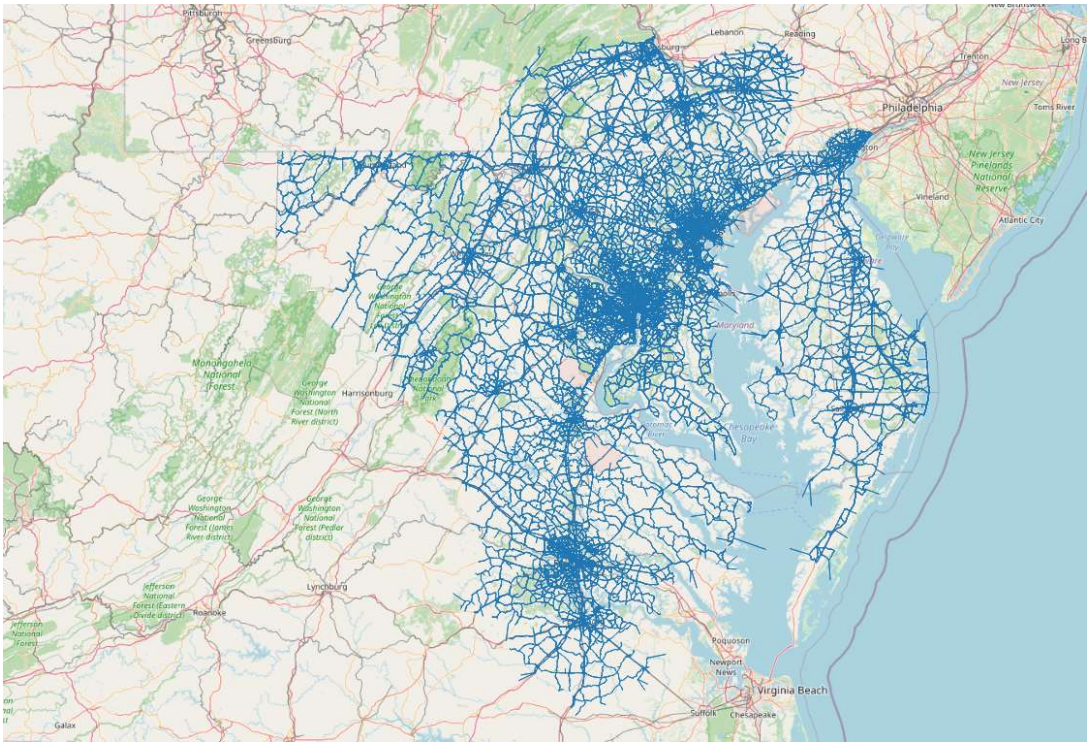

Figure 3-2: HERE Navigation Network with Various Roadway Geometry Features

3.1.3. Data Screening and Processing

The proposed model aims at predicting driving features and energy consumption at link level. To support model training, the trip trajectories must be converted into sequences of stages to achieve link-based driving features. The stage here refers to one or more continuous links with similar environmental features. The GPS points are first screened on the statewide network to gain roadway geometrics and traffic dynamics information. A reliable tool, called Rtree spatial-index function, is used for the screening process at network scale. Rtree is a ctypes Python wrapper of lib-spatial-index that provides several advanced spatial indexing features for spatial analysis, such as nearest neighbor search and clustered indexes [125]. Once a buffer

(e.g., 0.001 mile) is defined, the links intersected with the buffer around the GPS points are filtered. However, the statewide network developed in this study includes many local roads, where roadway segments are often short and near other roadways. To better match the GPS points with the state-level network, the study develops a shortest-path approach, which takes the link-based historical travel time as weight. The approach is coded based on the NetworkX function in Python, with a database recording historical traffic speed collected from RITIS. For some local roads with limited observations of historical traffic speed given their low traffic volume during most time period, the speed is assumed to be free-flow speed given the low volume conditions. The framework of the data processing is shown in Figure 3-3.

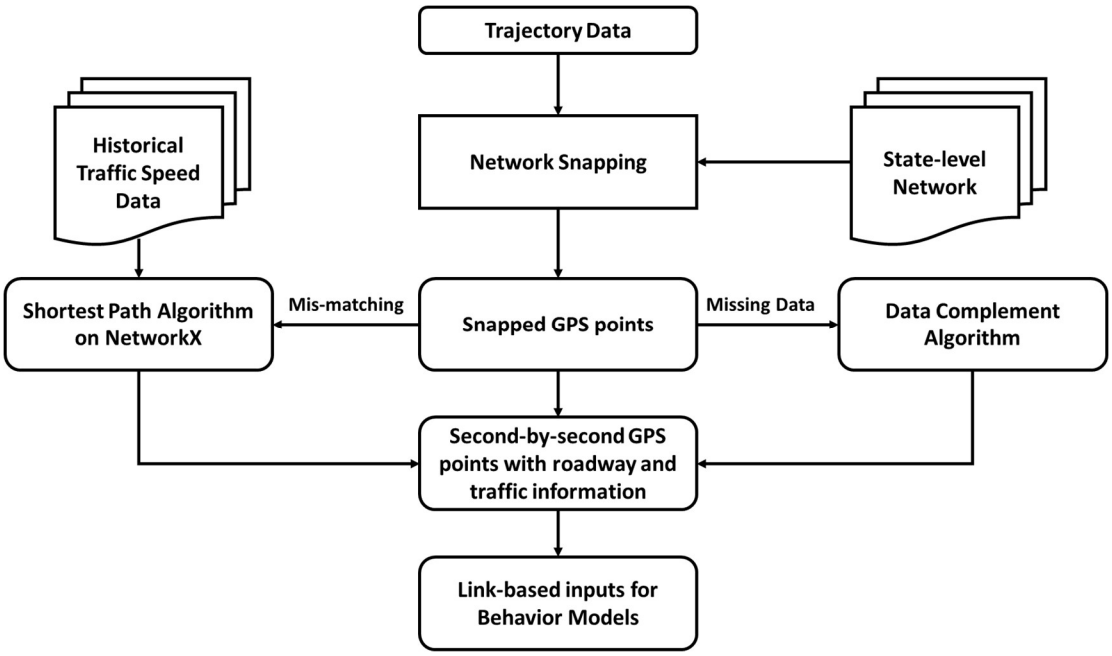

Figure 3-3: Framework of Data Screening and Processing

The screened GPS points are checked, and necessary reconstructions are conducted to insure appropriate matching. Special cases, as described below, are further processed as described.

1) Mismatching: When two or more roadway links are close to or on top of each other (e.g., a bidirectional link), GPS points may be assigned to the wrong links. This type of mismatching is common when the spatial-index tool is used, since only the vertical distances to the reference links are considered. To correct the mismatched links, the shortest-path function in the NetworkX package of Python is used to check if the link belongs to the feasible path of the trip and relocate the mismatched points if needed.

2) Missing data: In some cases, the observations on one link are completely missing due to a mobile phone signal lost or communication issues. For instance, there are four consecutive roadway segments named A, B, C, and D in sequential order. While segments A and D have the GPS data, segments B and C do not have observations. In this case, the shortest-path algorithm is utilized to decide how the missed data should be filled in. The algorithm takes the from-node of the upstream link and the to-node of the downstream as origin and destination, calculates the weighted cost of each feasible path, then returns the path with the minimum cost. When the correct middle segments are decided, the empty data cell is filled in with the assumption that the speed on the segment is uniformly distributed.

3) Outliers: Values are judged against the vehicle's technical parameters. For instance, points with speeds over 100 mph will be removed. For the removed GPS points, the driving features will be determined based on the assumption that the speed on the segment is uniformly distributed.

Suitability in this paper is defined as the number of links with GPS points over the total number of links of the path. Only the trips with a suitability of more than 90% will be used as training data. The processed dataset is summarized as Table 3-3.

Table 3-3: Summary of Trips after Data Processing

Features	Sources	Before Processing	After Processing
Number of Trips	incenTrip	5269	3897
	i2D	2812	2664
Number of Users	incenTrip	254	46
	i2D	16	16
Average Travel time	incenTrip	32.65	33.19
	i2D	29.67	29.78
Average Travel Distance	incenTrip	13.53	15.01
	i2D	8.19	8.19

3.1.4. Data Exploration

At the mesoscopic scale, using VSP-distribution function to calculate energy consumption at link level is a promising approach for eco-routing purposes. Less computation resources are needed, and input data does not need to be at a high temporal

resolution in VSP-based models. Existing studies suggest the mean and standard deviation of VSP distribution can be estimated using average cruising speed [41]. For prediction purpose, the average speed is not available, while posted speed limits and historical speed may be used instead. This study analyzes the relationship between VSP distribution and various driving features. The results indicate that using average cruising speed and acceleration rates can improve the accuracy of VSP distribution and energy consumption by 7% and 11%, respectively.

Four scenarios are compared, taking different inputs to estimate VSP distribution and energy consumption: posted speed limit, historical traffic speed, average cruising speed, and average cruising speed with acceleration rate as supplementary input. For convenience, the four groups are named as speed limit group, traffic speed group, cruising speed group, and cruising speed & acceleration rate group in further discussions. Only 'i2D' data is utilized as ground truth data at this stage since it includes information for both VSP distribution and energy consumption. In each scenario, the mean and standard deviation of VSP distribution is first calculated and compared with the observed statistics for 'i2D'. Utilizing the reference table (Table 3-4) between VSP and operation modes, and the reference table (Table 3-5) between operation modes and energy rates, the link-based energy consumption can be achieved [36].

Table 3-4: Reference Table between Operation Mode and VSP

Cruising Speed (mph)					
0~25		**25~50**		**50<**	
Mode	**VSP**	**Mode**	**VSP**	**MODE**	**VSP**
11	<0	21	<0	NA	NA
12	0~3	22	0~3	NA	NA
13	3~6	23	3~6	33	<6
14	6~9	24	6~9	35	6~12
15	9~12	25	9~12	NA	NA
16	12<	27	12~18	37	12~18
Other		28	18~24	38	18~24
0	Braking	29	24~30	39	24~30
1	Idling	30	30<	40	30<

Table 3-5: Reference Table Between Various Emission Factors and Operation Mode

Vehicle type	Age	Operating Mode	Total Energy (KJ/h)	CO2 (g/h)	NOX (g/h)
1	5	0	49108.1	3529.2	0.2297
1	5	1	45430.6	3264.9	0.0973
1	5	11	71438.5	5134.0	0.3407
1	5	12	98643.7	7089.2	0.5201
1	5	13	137092	9852.3	1.2175
1	5	14	173224	12449.0	2.1495
1	5	15	206566	14845.2	3.8096

The comparison among the four scenarios is summarized in Table 3-6 on distance-weighted error of VSP distribution mean, VSP distribution standard deviation, and trip-based energy consumption. As shown in the table, the speed-limit-group achieves the lowest accuracy, while the accuracy of the traffic-speed-group is slightly better. Significant improvement is observed using average cruising speed as input instead of speed limit or historical traffic speed. Adding the acceleration rate as supplementary, the performance is further improved for all criteria, especially for the standard deviation.

Table 3-6: Accuracy of Prediction of VSP Distribution and Energy Consumptions for Different Groups

Inputs	Mean Cruising Speed (%)	Standard Deviation (%)	Link Energy Consumption (%)	Trip Energy Consumption (%)
Speed Limit	70.1	61.7	68.8	74.6
Traffic Speed	79.1	72.2	76.1	78.3
Cruising Speed	88.6	78.2	82.6	86
Cruising Speed & Acceleration Rate	89.3	83.9	87.5	91.3

3.2. Model Selection

As discussed in section 3.1, the personalized driving features—for instance, average cruising speed and acceleration rate—could be predicted to amend VSP distribution. Given a feasible path from origin to destination, a model is needed to map a sequence of links to a sequence of link-based driving features. In general, the algorithms for sequence data modeling can be categorized as supervised learning, unsupervised learning, and semi supervised learning.

Through sufficient and reliable ground truth data, supervised learning is used to explore the relationship between inputs and desired outputs (supervisory signal). In other words, inputs are pre-labeled ground truth data, from which a relationship is established and the model accuracy can also be estimated. Supervised learning algorithms are modeled as classification or regression, and have been widely applied in text categorization, face detection, signature recognition, weather furcating, etc.

Unsupervised learning, contrary to supervised learning, takes inputs that are neither classified nor labeled. The machine needs to explore and learn the hidden patterns by itself, as there is a lack of knowledge on the underlying logic and the desired outputs. The strength of the unsupervised learning rests with its input flexibilities. In practice, unsupervised learning algorithms performs well in clustering and anomaly detection.

Falling between the supervised learning and unsupervised learning, Semi-supervised learning is a new approach. It can be flexibly applied when only limited amount of ground truth data is available. Typically, the training data for semi-supervised learning consists of a small amount of labeled data and a large amount of unlabeled data. The combination of labeled and unlabeled data could infer the hidden pattern and guide the learning process to perform well.

As stated in previous studies, various factors from external environment, such as road type, traffic condition, and traffic control, have an influence on driver behavior and driving features. In order to predict driving features, it is necessary to model driver's behavior and driving context jointly. Driver's driving behavior is an abstracted concept that is usually unobservable and hard to be quantified, though its influence on driving features is matter-of-course. The individuality of driving features under the same external environment is an embodiment of driving behavior. On the other hand, driving behavior varies stage by stage, while the current stage is mostly affected by the previous stage and the external environment at the current stage. Therefore, a model such as Hidden Markov Model (HMM) is better suited than discriminative models such

as Support Vector Machine (SVM) or Neural Network (NN) that do not consider these temporal aspects. Though some Recurrent Neural Networks (RNN), such as Long Short-Term Memory (LSTM) can also be used to do sequence classification and sequence prediction, the HMM model outperforms the RNN models considering the limited training data for personalized modeling. Moreover, the HMM is developed based on the assumption that the current state is only determined by the previous state, which is a sound hypothesis in the current research.

3.2.1. Hidden Markov Model

The Hidden Markov Model (HMM) is a statistical model treating a system as a sequence of possible events. In the system, the status of each event only depends on the previous event, namely the Markov Process (MP). Except the observational data, the underlying data of the Markov model is hidden or unknown. As a generative model, HMM is well known for its effectiveness in modeling the correlations between adjacent symbols, domains, and events [126]. It has been extensively developed and applied in action recognitions and digital communications. HMM consists of two parts (Figure 3-4). The first part is a set of observed states and hidden states. The second part is the probability model consisting of initialization, transition, and observed outputs. The hidden states in the model are the factors that influence the environment but are difficult to be captured or quantified (e.g., emotion, mental activities, propensity). Nevertheless, the impact of these factors can be represented by several observable variables, namely observed states.

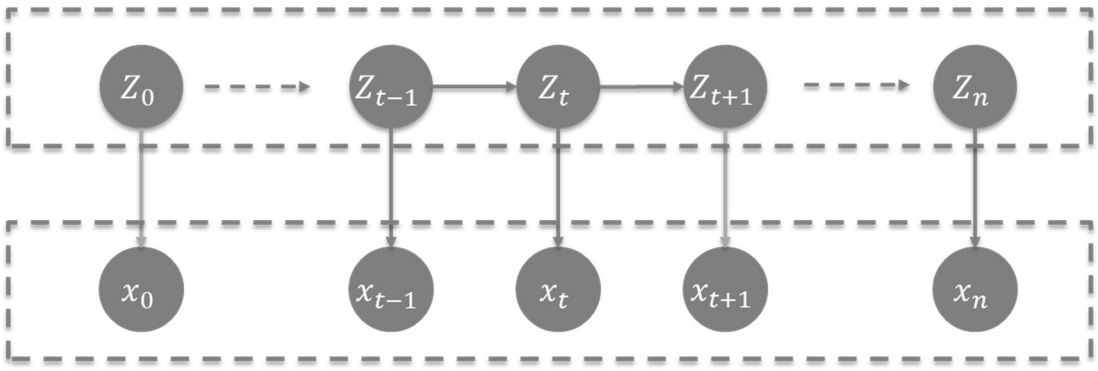

Figure 3-4: Structure of Standard Hidden Markov Model

A hidden Markov model is formulated as HMM: $\{\mathrm{N}, \mathrm{M}, \mathrm{A}, \mathrm{B}, \prod\}$ where:

1) N refers to the capacity of a hidden states set, which is pre-defined before training the model. The projection of hidden state can be either determinate or indeterminate depending on the learning method: supervised or unsupervised. Each individual hidden state can be denoted as s_i $(1 \leq i \leq N)$.

2) M stands for the length of the output set, where the output variables represent the impact of hidden states. It is worthwhile to mention that the output variables should only be observable after transitions occur (prediction process). Each individual output variable can be denoted as x_i $(1 \leq i \leq M)$.

3) The initial state probability distribution is a column vector with a dimension of $1 \times S$. In general, the initial probability can be determined by experience or historical records. While in some models, NN and RNN approaches can also be utilized to estimate the initial state probability. Cell π_i the value represents the initial probability given its initial hidden state is s_i, as represented by Equation 2:

$$\pi_i = P(q_1 = S_i) \quad 1 \leq i \leq N \tag{2}$$

4) A stands for the state transition probability matrix, where each cell indicates the probability that a hidden state transitioning to another state. Since the transition is modelled as Markov process, it assumes that the next state is only dependent on the current state. The probability for each state pair is formulated as:

$$\varphi_{i,j} = P(s_j | s_i), 1 \leq i \leq N, 1 \leq j \leq N \tag{3}$$

5) Except the initial probability matrix, state transition matrix, the HMM also includes an observation probability matrix, also known as the emission probability matrix ($N \times M$). Each cell in the observation probability matrix denotes the output variables distribution given a specific hidden state. It is expressed in the form of:

$$\theta_{i,k} = P(x_k | s_i), \quad 1 \leq i \leq N \; and \; 1 \leq k \leq M \tag{4}$$

3.2.2. Input Output-Hidden Markov Model

The standard HMM takes homogeneous transition and emission probability and assumes each hidden state is only decided by the previous hidden state. These assumptions, however, are too restrictive and one-sided in most cases. Some other factors may also affect the current hidden state. For instance, a driver driving aggressively on the immediate prior road segment is more likely to make a sharp deceleration if the current roadway segment has a traffic signal control or a reduced speed ahead. On the other hand, a less aggressive driver may act excessively cautious

on a high-traffic road, even though the traffic speed is still high. To make up the drawbacks of the standard HMM, the so-called Input Output Hidden Markov Model (IOHMM) is proposed as the selected architecture, which fully considers the influence of both the previous state and the system environment. As shown in Figure 3-5, the IOHMM shares similar structure and components as the standard HMM, except that it has an additional layer of observed contextual variables, namely the input layer. Different from the output layer, the observable variables in this input layer represent the system effects that should be observable before a state transition. In IOHMM, the inputs determine the current state through the previous state, then control the outputs through the current state. In other words, the distribution of both states and outputs are influenced by a set of input variables. Based on this modeling structure, the IOHMM could map the input sequence into output sequence through the similar process as RNN.

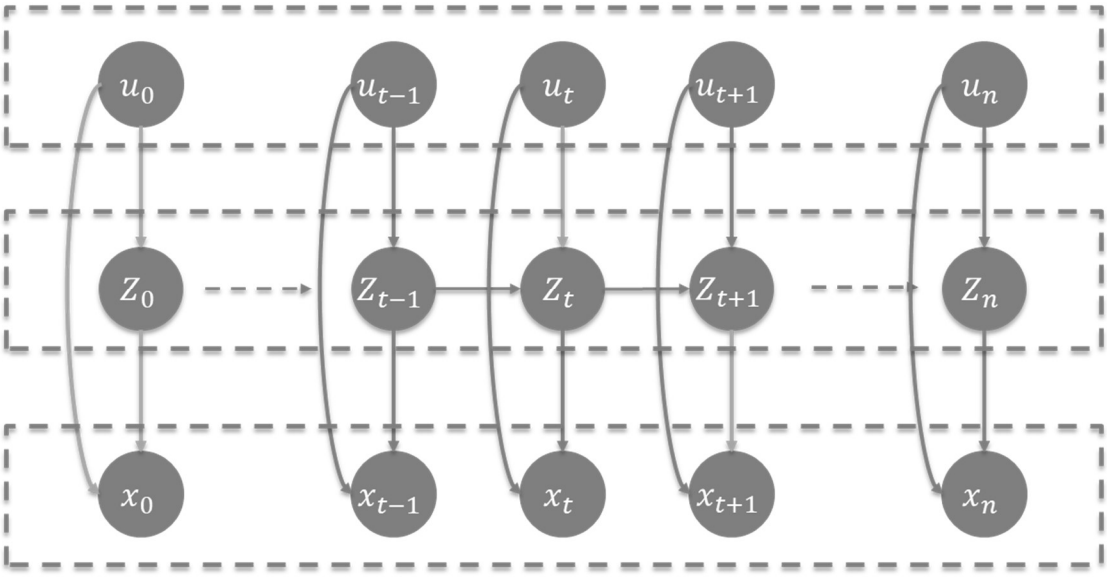

Figure 3-5: Structure of Input Output Hidden Markov Model

3.3. Model Generation

3.3.1. Feature Selection

The IOHMM learns the relationship between inputs, outputs, and hidden states, then maps the input sequence into output sequence. In this study, the observable features of external environments are treated as inputs, the driving features are considered as outputs of the model, and individual behaviors are considered as hidden states.

As demonstrated in the data exploration results in section 3.1.4, the average cruising speed and acceleration rate contribute to the formulation of VSP distribution. Meanwhile, the two variables are controlled by both external environment and individual behaviors and are only observable after state transition. Therefore, the average cruising speed and acceleration rate are involved as the output variables.

As stated in previous studies, various features from the external environment have impact on driving behaviors and driving features, such as roadway geometry features (i.e., number or lanes and speed limit) and traffic conditions (i.e., traffic speed and traffic volume). Though more features included in the input layer provide richer information, the model may not perform well in sequence prediction. A simpler structure with less inputs is preferred in sequence modeling. The most important reason is the overfitting issue caused by redundant inputs and complex models, which shows high performance in training datasets and poor performance in the testing ones. In general, the overfitting issue could be eliminated through appropriate feature selection and dimension reduction. Feature selection works by analyzing the relationship between features and selecting the most critical ones. Commonly applied feature

selection methods can be categorized as filter methods (e.g., linear discrimination analysis, Anova analysis, Pearson Correlation, and chi-square analysis) and wrapper methods (e.g., forward selection, backward selection, and recursive feature elimination). Dimension reductions alleviate the overfitting by generating new features with most critical impacts on the outputs. Both methods can drop low-impact features and reduce multi-collinearity of features. Some emerging approaches also prove reliable performance in reducing the effects of overfitting: 'early-stopping' approach that stops the training process before reaching optimum, and 'data-expansion' approach that generates a larger training dataset based on the original training dataset [120].

In this study, the authors conducted a Pearson correlation analysis on alternative features to select most critical and comprehensive variables for input layer. The average Pearson correlation coefficients among variables are summarized in Figure 3-6.

	Mean Cruising Speed	Acceleration Rate
Mean Cruising Speed		0.1765
Acceleration Rate	0.1765	
Number of Lanes	0.2135	0.1554
Speed Limit Ratio	0.6106	0.3100
Intersection	-0.2903	0.3955
Signal Control	-0.4220	0.4152
Peak Hour	-0.3772	-0.2099
Ramp	-0.0102	0.2114
Road Type	0.3867	-0.2178
Grade	-0.2312	-0.0999

Figure 3-6: Pearson Correlation Coefficients of Various External Variables and Driving Features

As indicated by the Pearson correlation analysis, the relationship between mean cruising speed and speed limit is the most significant, with the highest coefficient of 0.6106. The road type comes in second, achieving a coefficient of 0.3867. A clear negative correlation is observed between mean cruising speed and signal control, which indicates the segments with signal controls usually have lower speeds. The correlation score between mean cruising speed and intersection is -0.2903, and the value is -0.3772 for peak hour factor. Speed limit, intersection, and signal control present positive influence on acceleration rate, which indicates that drivers tend to frequently adjust

their speeds toward an intersection with signal controls. In addition, the road type has a negative influence on accelerate rate, which indicates that the speed is smoother on high-level roads. Five variables exerting substantive effects on mean speed and acceleration rate are included in the proposed input output-hidden Markov mode. These four variables are, road types, turning movements, signal controls, and speed limit. A detailed description and definition are as follows:

1) Road type: an integer value from 1 to 5 (1, a road with high volume and maximum-speed traffic; 2, a road with high volume and high-speed traffic; 3, a road with high-volume traffic; 4, a road with high-volume traffic and moderate-speed traffic between neighborhoods; and 5, a road whose volume and traffic flow are below the level of any other road types).

2) Peak hour indicator: a binary variable indicating if the segment is in peak period, which is defined as the period where the ratio between the historical traffic speed and posted speed limit is less than 0.85. It is worth noticing that the threshold used here is different from the traditional definition of peak period. The reason for this modification is the fact that traffic conditions on some segments are very different during different times of day.

3) Speed limit ratio: an integer number between 1 and 4 represents different ratio of historical speed over the posted speed limit (1, speed limit ratio less than or equal to 0.75; 2, speed limit ratio greater than 0.75 but less than or equal to 0.9; 3, speed limit ratio greater than 0.9 but less than or equal to 1.1; 4, speed limit ratio greater than 1.1).

4) Traffic control: an integer number between 1 and 3 represents if the segment has traffic control (1, with signal control; 2, left/right turn without signal control; and 3, others)

5) Speed Limit

Except for the variables discussed above, traffic features could also affect driving behaviors. Historical traffic statistics may be adopted as inputs in the training process of the model, but the deployment requires real-time traffic information. Therefore, a short-term traffic dynamic prediction model is needed for the present study to support a real-time eco-routing system.

Recurrent Neural Network (RNN) is a popular approach of learning sequential data through a generalization of feedforward NN with an internal memory. Different from a standard NN, the RNN learns from inputs through its internal memory, which helps deal with large data size. Traditional RNN performs well in modeling non-linear time series problems. Nevertheless, its performance is still limited due to time-consuming and poor performance in long-sequences modeling.

To overcome the limitations summarized above, a long short-term memory (LSTM) approach is deployed. LSTM is a special RNN approach that works better for long sequencesAs shown in Figure 3-7, LSTM makes a significant improvement on recurrent cells by adding forget gates, which allows removing or updating information to a cell state.

A standard LSTM model is composed of a cell and three gates: the input gate, the output gate, and the forget gate. The input gate decides the rate of the cell receiving a new value. The output gate controls the value to be used to compute the output variation of the LSTM unit. The forget gate controls the extent in which a value remains in the cell. In each iteration, the three gates control the status of flow of information in the cell: in, out, or remain.

In this study, the LSTM algorithm with the following structure is used for traffic dynamic prediction.

$$X = (x_1, x_2, x_{3,} \dots, x_n) \tag{5}$$

$$X = (x_1, x_2, x_{3,} \dots, x_n) \tag{6}$$

$$H = (h_1, h_2, h_{3,} \dots h_n) \tag{7}$$

where X stands for the input sequence (historical traffic dynamic), Y refers to the output sequence (predicted traffic dynamic), and H is the hidden state vector, also known as output vector of the LSTM unit. The training process of LSTM model starts by setting the cell state vector c_t and hidden state vector h_t as zero ($c_0 = 0, h_0 = 0$). The predicted average speed is calculated iteratively using historical data through the following equations [127-129]:

$$h_t = o_t \cdot \tanh(c_t) \tag{8}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t+1} + b_o) \tag{9}$$

$$f_t = \sigma_g(W_f x_t + U_f h_{t+1} + b_f) \tag{10}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t+1} + b_i) \tag{11}$$

$$c_t = f_t \cdot c_{t-1} + \ i_t \cdot \tanh\left(W_c x_t + U_c h_{t+1} + b_c\right) \tag{12}$$

where o_t refers to the output gate's activation vector, i_t represents the activation vectors of input and output gates, i_t stands for the forget gate's activation vector, and the cell state vector is denoted as c_t.

$\sigma_g(\cdot)$ is defined as the standard logistics sigmoid function with the formulation of:

$$\sigma(\cdot) = \frac{1}{1+ e^{-x}} \tag{13}$$

In addition to input and output layers, IOHMM model also includes a hidden layer to reflect the influence of unobservable hidden states. Hidden states are unobservable during training, but certainly influence the state transitions and outputs generation. Past research indicated that there have been large differences in judging the appropriate number of hidden states. Some studies stated that having more hidden states leads to a higher likelihood, which obviously increases the reliability of parameters. While small number of hidden states are preferred in practical applications considering the data quality and availability. Developing the IOHMM modeling structure also requires labeling the hidden states. If no ground truth data is available for labeling, only the number of hidden states is required, and the model will be formulated as unsupervised learning process. In this study, the hidden states are individual driving behavior. Though some research defines various rules to classify driving behaviors,

few of them are validated using ground truth data. Therefore, the authors first trained the model as unsupervised learning process without labeling the hidden states. Multiple rounds of training processes are repeated to test the performances under a different number of hidden states, and the results show that the optimum number of hidden states is five. Then, the personalized behavior models are developed with the optimum number of hidden states. The authors conducted deeper research on the characteristics of predicted hidden states and the relationship between input and output variables. The analysis suggested that the labeling rules with speed rate (defined as average cruising speed over average traffic speed), percentage of continuous acceleration duration, and percentage of continuous deceleration duration generate the most reliable hidden states. More detailed analysis and discussion will be presented in Chapter 5 in experimental results. The hidden states labeling rules are summarized below:

1) Idling:
    a. Percentage of continuous acceleration duration < 0.05
    b. Percentage of continuous deceleration duration < 0.05.
    c. Average cruising speed/average traffic speed < 0.2.
2) Very cautious:
    a. Average cruising speed/average traffic speed within (0.2 ~ 0.75).
3) Cautious:

a. Average cruising speed/average traffic speed within (0.75 m/s^2 ~ 0.9 m/s^2).

4) Normal:

a. Average cruising speed/average traffic speed within (0.9 m/s^2 ~ 1.1 m/s^2).

5) Aggressive:

a. Average cruising speed/average traffic speed > 1.1.

b. Percentage of continuous acceleration duration + percentage of continuous deceleration duration < 0.25.

6) Very aggressive:

a. Average cruising speed/average traffic speed > 1.1.

b. Percentage of continuous acceleration duration + percentage of continuous deceleration duration >= 0.25.

Note that the optimum number of hidden states is five based on multiple rounds of testing. The authors, however, found that many states with long duration of stationary are mis-classified into other states, which may influence the performance of the proposed model. Therefore, a separate hidden state for idling is added in the labeling rules of supervised learning process.

3.3.2. Parameter Estimation

With the structure of IOHMM, several parameters related to the hidden states cannot be directly determined based on the experimental data: initial probability parameters, transition probability parameters, and emission probability parameters. Various types of approaches have been developed and deployed for parameters estimation for all kinds of statistical problems, such as the Expectation-Maximization (EM) algorithms, NN algorithms, and rank regression. EM algorithms estimate the parameters by finding the maximum log-likelihood iteratively based on the unobserved latent variables in statistical models, while NN algorithms seek the optimum parameters through iterative optimization processes such as gradient decants or backpropagation. In this paper, the unknown parameters are learned through the EM algorithm as the hidden states of IOHMM model. EM algorithms include the Estimation step (E step) and Maximization step (M step). Based on the model structure and selected variables, the log-likelihood representing the summation of all routes through the hidden states can be formulated as follows.

$$L(x,u,\theta) = \sum_S [P(s_1|u_1;\theta_{in}) \cdot \prod_{t=2}^{T} P(s_t|s_{t-1},u_t;\theta_{tr}) \cdot \prod_{t=1}^{T} P(x_t|s_t,u_t;\theta_{em})] \quad (15)$$

Where s_t, u_t, x_t refer to the hidden state, input variables, and output variables at stage t, respectively. θ_{in}, θ_{tr}, θ_{em} stand for the initial probability, transition probability, and emission probability. T is the total number of stages of the sequence. $L(x,u,\theta)$ denotes the likelihood.

With the likelihood function formulated, the parameters are estimated and optimized through the EM algorithm.

E step: in the E step, the expected value of the likelihood is computed through Equation 15 based on the observed inputs and outputs at the current stage and the parameters estimated from the previous stage.

M step: the parameters are updated to maximize the expected log-likelihood formulated in Equation 16.

$$Q = \sum_{i=1}^{S} \gamma_{i,1}\, logP(s_1 = i|u_1; \theta_{in}) + \sum_{t=2}^{T}\sum_{i=1}^{S}\sum_{j=1}^{S} \xi_{i,j,t} logP(s_t = j|s_{t-1} = i, u_t; \theta_{tr}) + \sum_{t=1}^{T}\sum_{i=1}^{S} \gamma_{i,t} logP(x_t|s_t = i, u_t; \theta_{em}) \quad (16)$$

Where Q refers to the expected value of the log-likelihood. S is the set of state and i, and j represents a specific state from the set of states. T denotes the total stages of the sequence and t refers to one of the stages. s_t, u_t, x_t refer to the state, input variables, and output variables at stage t, respectively. θ_{in}, θ_{tr}, θ_{em} stand for the initial probability, transition probability, and emission probability. $\gamma_{i,t}$ stands for the posterior state probability for state i at stage t, while $\xi_{i,j,t}$ explains the posterior transition probability from state i to state j at stage t.

3.3.3. Model Specification

The unknown parameters: initial model, transition model, and emission model can be formulated based on the types of variables. A logistic regression model is defined as Equation 17 for the initial model based on the discrete variable in hidden state.

$$P(s_1 = i|u_1; \theta_{in}) = \frac{e^{\theta_{in}{}^{i} u_t}}{\sum_k^S e^{\theta_{in}{}^{k} u_t}} \tag{17}$$

Where s_1 and u_1 represent the initial state and given inputs, respectively. $\theta_{in}{}^{i}$ stands for the coefficients of initial state i in initial probability matrix. k denotes the state in state set S. u_t refers to the inputs at stage t.

For the discrete input variables, for instance, road type and traffic control could also be formulated as a logistic regression model:

$$P(s_t = j|s_{t-1} = i, u_t; \theta_{tr}) = \frac{e^{\theta_{tr}{}^{ij} u_t}}{\sum_k^S e^{\theta_{tr}{}^{ik} u_t}} \tag{18}$$

In the equation, s_t and u_t represent the state and inputs at stage t. θ^{ij} donates the transition probability to reach state j, given the current state i. S is the total number of states of the hidden state set and k refers to a single state in S.

For inputs defined as binary variables—for instance, peak hour indicator—a logistic regression in the following format also works:

$$P(s_t = 1|s_{t-1} = i, u_t; \theta_{tr}) = \frac{1}{1+e^{-\theta_{tr}{}^{i} \cdot u_t}} \tag{19}$$

The average cruising speed and acceleration rate defined as output variables are both continuous. A normal distribution could be assumed for emission model, as formulated in Equation 20.

$$P(x_t|s_t = i, u_t; \theta_{em}) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_t - {\theta_{em}}^i \cdot u_t)^2}{2{\sigma_i}^2}} \quad (20)$$

where σ_i and ${\theta_{em}}^i$ stand for the standard deviation and coefficients of the linear model for state i, respectively.

3.4. Modeling Framework

In summary, this chapter describes the methodology to analyze individual driving behavior and to develop statistical models to predict driving features, as shown in Figure 3-7.

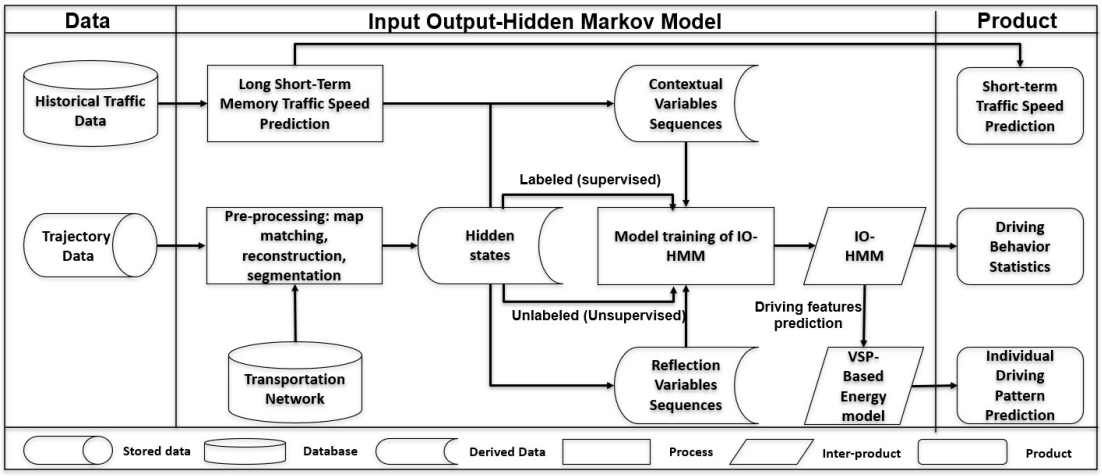

Figure 3-7: Methodology Framework

The chapter first describes the trip trajectory data collected from incenTrip and i2D, which is utilized to analyze individual driving behaviors. Then the data processing process is briefly introduced. To explore the relationship between external environment, individual driving behavior, and driving features, a statistical model is developed using Input-Output Hidden Markov model. The model considers a route as a sequence of stage, with each stage representing one or more links with the same characteristics. The model takes the driving behavior as hidden states and assumes the behavior at each stage is decided by the external environment at current stage and the behavior at previous stage. The driving features are involved as output variables, which are influenced by both the hidden states and external environment. The model is first trained by an unsupervised learning process without labeling the hidden states. The predicted hidden states from trained models are used to extract distinguishing characteristics to evaluate the labeling rules in previous studies. The model is then trained by supervised learning using labeled hidden states based on the optimum rules from previous studies. Several procedures of developing the model are then discussed: feature selection, parameter estimation, and model specification. Finally, the integration of IOHMM model and energy prediction model is presented.

Chapter 4 Experimental Results

4.1. Model Deployment

To evaluate the performance of the proposed driving behavior model with IOHMM structure, an experimental study is conducted using trip trajectory data collected from incenTrip APP and the i2D system. As indicated in Chapter 3, the data used in this study shows uneven distribution in travel distance, which may result in the imbalance of input sequences. The IOHMM could take the imbalanced inputs by setting the dimensions of transition and emission probability matrix as the length of the longest input vector. However, less information can be provided by the input vectors with short length, and the learning efficiency at the last stages will be significantly weakened. Therefore, the consecutive links with similar input variables are combined and the trips with a sequence length between 20 to 40 are used in the experimental study. This helps balancing the input data size and ensures learning efficiency. To ensure the data size is enough for training the personalized driving behavior model, drivers with 20 or more trips are selected. Additionally, to reduce the impact of familiarity of roadway on driving behavior, only drivers with five or more different paths are retained in the training dataset. A total of 2,254 trips from 75 incenTrip users and 2,011 trips from i2D testers meet these criteria and are used in the experimental study for personalized driving behavior modeling. The performance of the proposed model is evaluated on multiple levels through various measurements of effectiveness (MOE): (1) state transition and recognition, (2) driving features prediction, (3) energy consumption prediction at link- and trip- levels, (4) benchmark methods comparison, and (5) system

benefits analysis. It is worth stating that only the trips from i2D are using in evaluating energy consumption prediction since the trips from incenTrip does not include ground truth energy data.

4.2. State Transition and Recognition

As stated above, the IOHMM can be trained as both a supervised and unsupervised learning process depending on whether the hidden states are labeled. In the experiemental study, the personalized IOHMM models are first trained as unsupervised learning processes without labeling the hidden states. Multiple rounds of testing with different numbers of hidden states are conducted and the performance indicates the most approperiate number of hidden states is five. The personalized models are trained and the hidden states are predicted. The state transition probability matrix for four different drivers under same inputs are presented as examples through the heatmaps, as shown in Figure 4-1. The red cells represent lower transition probablity while the green cells stand for higher transition probability. As demonstrated in Figure 4-1, significant individual differences in state transition pattern can be observed among different drivers. Specifically speaking, the green cells of the first driver's pattern concentrate in the -45-degree axis of symmetry, which indicates that the driver tends to maintain the current state instead of shifting to other states. This phenomenon is commonly seen in experienced drivers who can operate the vehicle smoothly under various traffic conditions. On the contrary, the green cells of the second driver's state transition pattern concentrate in state D. In other words, the driver prefers to change from other states to state D when driving. Similar phenomemon can be observed in Figure 4-1 (c) for the third driver and 4-1 (d) for the forth driver. The green

cells of these two drivers' pattern concentrate more in state B and the drivers prefer to maintain state B during the trips. Additionally, a slight difference can be observed between the third and fourth driver. With the exception of state B with highest transition probability, the second highest state is state C and state A for the third and fourth driver, respectively. These observations prove that the proposed IOHMM models could capture the individually different driving behaviors. Moreover, these differences in state transition patterns among different drivers further prove the importance of considering individual driving behavior in energy prediction and building personalized driving behavior models.

State	A	B	C	D	E
A	0.0000	0.0000	0.9990	0.0010	0.0000
B	0.0000	0.6240	0.3530	0.0000	0.0000
C	0.0000	0.0000	0.6530	0.0000	0.3340
D	0.0000	0.0000	0.3180	0.6770	0.0003
E	0.0000	0.0320	0.0000	0.2350	0.7250

(a)

State	A	B	C	D	E
A	0.1605	0.0000	0.1170	0.7222	0.0000
B	0.0365	0.0372	0.0690	0.8867	0.0037
C	0.0432	0.0000	0.1480	0.8080	0.0000
D	0.0387	0.0000	0.0730	0.8880	0.0000
E	0.0386	0.038	0.037	0.8475	0.0381

(b)

State	A	B	C	D	E
A	0.2310	0.5950	0.1090	0.0640	0.0000
B	0.0592	0.7060	0.1780	0.0230	0.0330
C	0.0000	0.7940	0.2050	0.0000	0.0000
D	0.1058	0.5210	0.2653	0.0170	0.0000
E	0.0000	0.5305	0.2010	0.0000	0.2676

(c)

State	A	B	C	D	E
A	0.0600	0.8635	0.0640	0.0090	0.0020
B	0.3850	0.8112	0.0275	0.0000	0.0914
C	0.0158	0.8812	0.0000	0.0230	0.0790
D	0.0980	0.7445	0.0244	0.1306	0.0010
E	0.066	0.7334	0.0111	0.0108	0.1777

(d)

Figure 4-1: Heatmap of Transition Probability Matrix Between States for (a) Driver with 176 Trips, (b) Driver with 112 Trips, (c) Driver with 65 Trips, and (d) Driver with 39 Trips

The study then compares the predicted hidden states from unsupervised IOHMM with labeled states through various labeling rules. The comparison indicates that maximum matching occurs under the labeling rules based on average cruising speed, acceleration rate, and continuous acceleration rate. The IOHMM models are then trained as supervised learning process using the labeled hidden states. The three parameters are obtained, and the hidden states are predicted. The comparison between predicted hidden states and labeled hidden states for the testing trips (20% of total trips) is demonstrated through confusion matrix, as shown in Figure 4-2. Each cell in the confusion matrix stands for the number of observations, whose labeled state is A (row name) and the predicted state is B (column name). Additionally, the percentages of mismatching among different states are summarized as a heatmap in Figure 4-3, where red cells represent higher percentages of mismatching and green cells stand for lower

percentages of mismatching. The white cells in the -45-degree axis of symmetry of the figure are the percentage of properly matching among different states.

States		Predicted States					
		1	2	3	4	5	6
Labeled States	1	494	59	2	2	0	1
	2	122	725	60	23	6	4
	3	17	251	4145	504	175	119
	4	203	475	1263	12471	2103	374
	5	16	50	327	519	4499	489
	6	1	1	31	45	110	1326

Figure 4-2: Confusion Matrix between Labeled Hidden States and Predicted Hidden States

States		Predicted States					
		1	2	3	4	5	6
Labeled States	1	88.53047	10.57348	0.358423	0.358423	0	0.179211
	2	12.97872	77.12766	6.382979	2.446809	0.638298	0.425532
	3	0.326233	4.816734	79.54327	9.671848	3.358281	2.283631
	4	1.201966	2.812481	7.47824	73.84096	12.45189	2.214459
	5	0.271186	0.847458	5.542373	8.79661	76.25424	8.288136
	6	0.06605	0.06605	2.047556	2.972259	7.265522	87.58256

Figure 4-3: Mismatching Ratio Between States

In general, the proposed personalized IOHMM models achieve an accuracy of 76.32% in state recognition. The 'idling' state shows the highest accuracy of 88.53% due to the significant higher probability of the 'idling' state in the initial probability. The initial states of almost all the input sequences are labeled as the 'idling' states, which contributes to the higher initial probability. Except for the 'idling' state, the 'very cautious' state and 'very aggressive' state are the best represented in state matching, with the accuracy of 87.58% and 79.54%, respectively. The 'normal' state achieves the poorest accuracy of 73.84% with the most mismatching happening between the 'normal' state and the 'cautious' state. A high percentage of mismatching between the 'very cautious' state and the 'idling' state is observed. This confusion is natural since the 'very cautious' state is defined as the observations with higher deceleration rate and lower cruising speed than traffic, which is consistent with the driving features on the link that the vehicle starts from the 'idling' state. For instance, a vehicle approaching an intersection with a red signal is more likely to reduce its speed and stop for a considerable duration. A similar condition may also happen to an overly cautious driver approaching a ramp with a yield sign. The driver may perform 'stop and go' many times, especially when upcoming main-through traffic is heavy. A deeper analysis is conducted on the mismatched observations. Among the 59 mismatched observations, 51 are ramps or with signal controls—that's a ratio of 86%. However, the observations with signal control or ramps only contribute to 8.9% of the entire network and 17.2% of the entire testing data. Such disparities indicate the assumptions stated above: that the confusion rate between 'strong cautious' and 'normal' is reasonable.

Deeper analysis is conducted on the relationship between predicted states and various input variables. The portion under different states is summarized for the input variable of peak-hour indicator and traffic controls, as shown in Table 4-1 and Table 4-2, respectively. As shown in Table 4-1, the 'aggressive' state and 'very aggressive' state have observable higher percentages during off-peak hours, which are around 2.85% and 4.3% higher than peak hours. The 'Normal' state has a greater increase of 5.9%. This may result from the heavy traffic during peak hours that prevents drivers from operating vehicles aggressively. When traffic is free during off-peak hours, overspeed-driving and continuous acceleration are more commonly observable. On the contrary, the 'cautious' and 'strong cautious' states occur more in the peak hours, which also indicates more drivers prefer to take less aggressive actions during peak hours.

Distributions of the hidden states under different traffic control show perceptible differences, as indicated in Table 4-2. Among the three types of traffic controls, non-control has a significantly higher percentage of 'normal' state than signal control and turning control. However, the percentage of all other four states presents a perceivable higher percentage under signal control and turning control, especially the 'very cautious' state. Comparing the state distribution pattern under different peak-hour indicators and traffic control indicators gives various interesting findings. In general, more drivers prefer to be more cautious during peak hours with heavy traffic and are more aggressive during off-peak hours with little traffic. However, the tendency is different for traffic controls. Drivers show obvious polarization in the reactions to different traffic controls, which is presented by the significant increase in both 'aggressive' states and 'cautious' states. These findings indicate the influence of the

external environment on drivers' driving behavior and the importance of building personalized driving behavior models in predicting driving features.

Table 4-1: State Distribution in Peak Hours and Non-Peak Hours

State	Peak Hour	Off-peak Hour
1	6.71	3.86
2	21.64	17.34
3	48.19	42.23
4	18.5	26.58
5	4.96	9.98

Table 4-2: State Distribution on Roads with Signal Control, with Intersection, and without Controls

State	Signal Control	No Control	Turning Control
1	7.10	4.71	6.99
2	22.20	18.77	21.25
3	36.89	46.97	37.11
4	24.87	22.37	24.75
5	8.94	7.18	9.87

4.3. Output Variables

The proposed IOHMM has the ability to predict output sequences of driving features using input sequences of environmental variables. In this section, the accuracy

of outputs prediction under different input variables is summarized to evaluate the performance of the model. The predicted cruising speed and acceleration rates are compared with the observed values at link level through distance weighted error (DWE), as shown in Figure 4-4, 4-5, 4-6, and 4-7. As shown in Figure 4-4, the DWE of both cruising speed and acceleration rate have the tendency to increase with the decrease of roadway levels. Alternatively, it could simply mean that the prediction of both output variables on high-level roadways (e.g., freeway and highway) outperforms the low-level roadways, for instance, local roads. It could be speculated that this is due to the more complete and smoother observations on high-level roads. More complex traffic conditions may occur on low-level roads, for instance, traffic controls, pedestrians, work zones, and loading vehicles. For DWE under different road types, similar trends are observed, the high-speed roads present lower error for both speed rate and acceleration, and vice versa. However, the error for both output variables on the roads with a speed between 10 to 40 mph is hard to distinguish. It is difficult to explain such results with the context of current experimental studies. Additionally, significantly higher distance weight errors for both output variables are observed during peak hours. This appears to be a case of more frequent acceleration and deacceleration during peak hours, which results in a wider range of acceleration rates that are hard to predict. In the aspect of traffic controls, the roads with traffic controls obviously perform worse than the roads without control. Moreover, the roads with turning control present slightly lower error than signal-controlled roads.

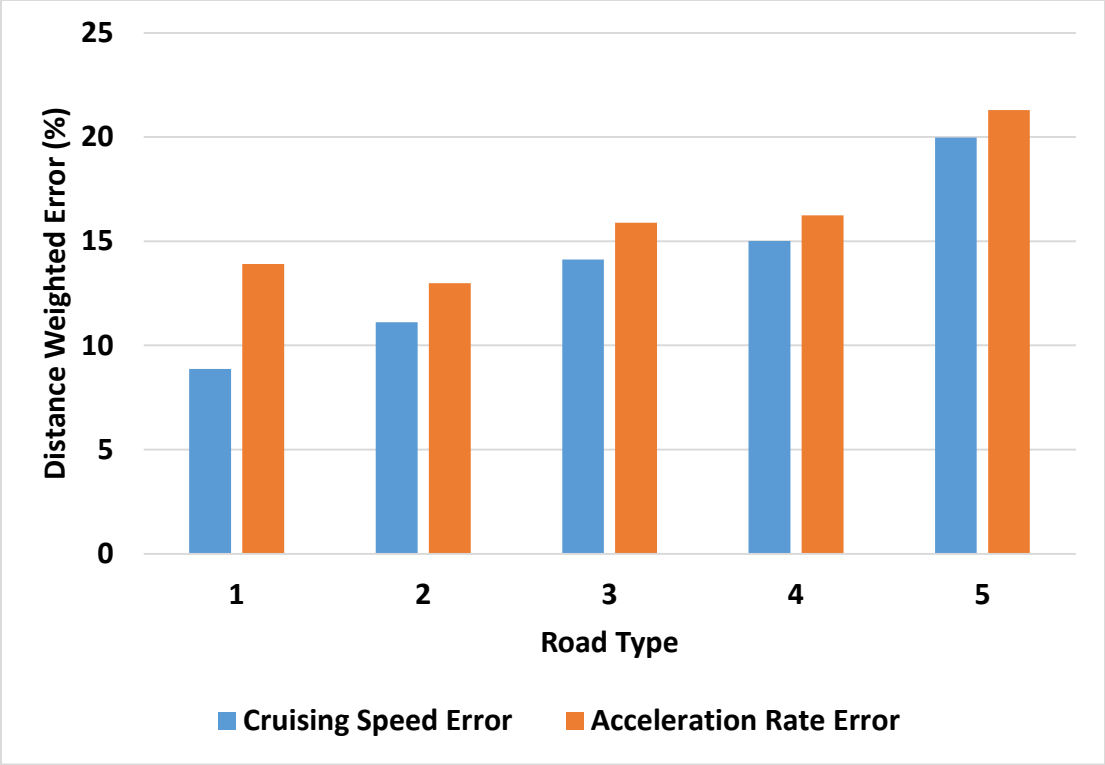

Figure 4-4: Errors of Output Variables for Different Road Types

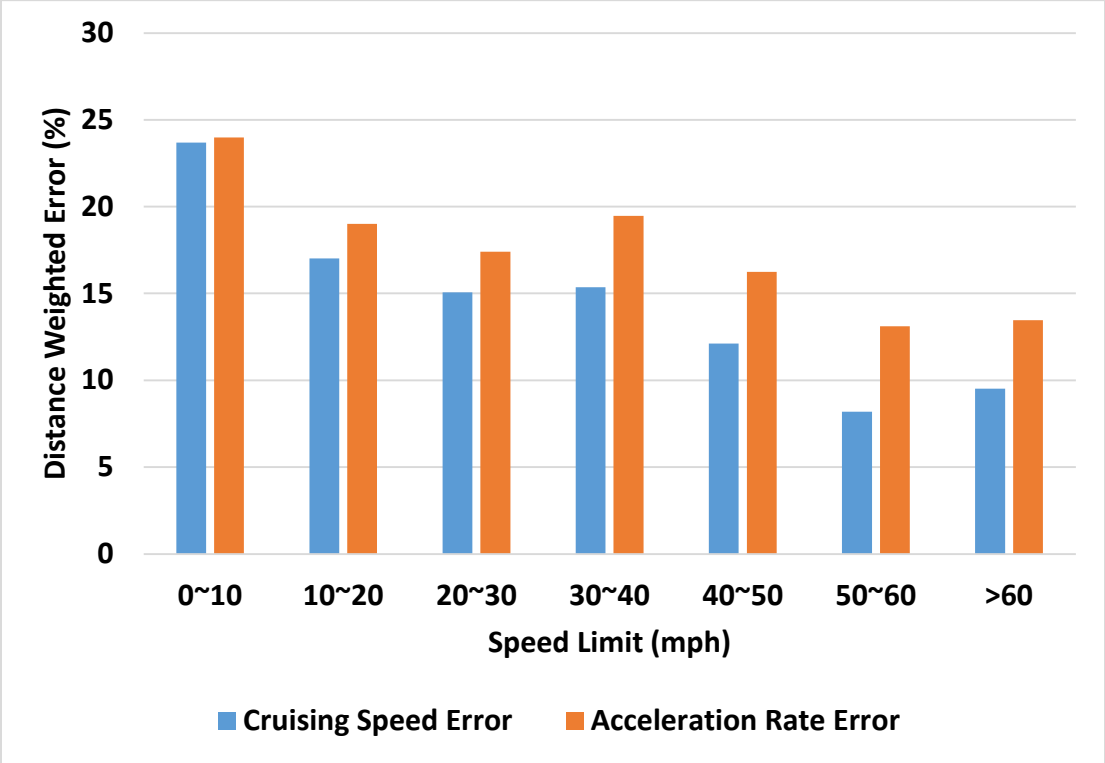

Figure 4-5: Errors of Output Variables for Different Speed Ranges

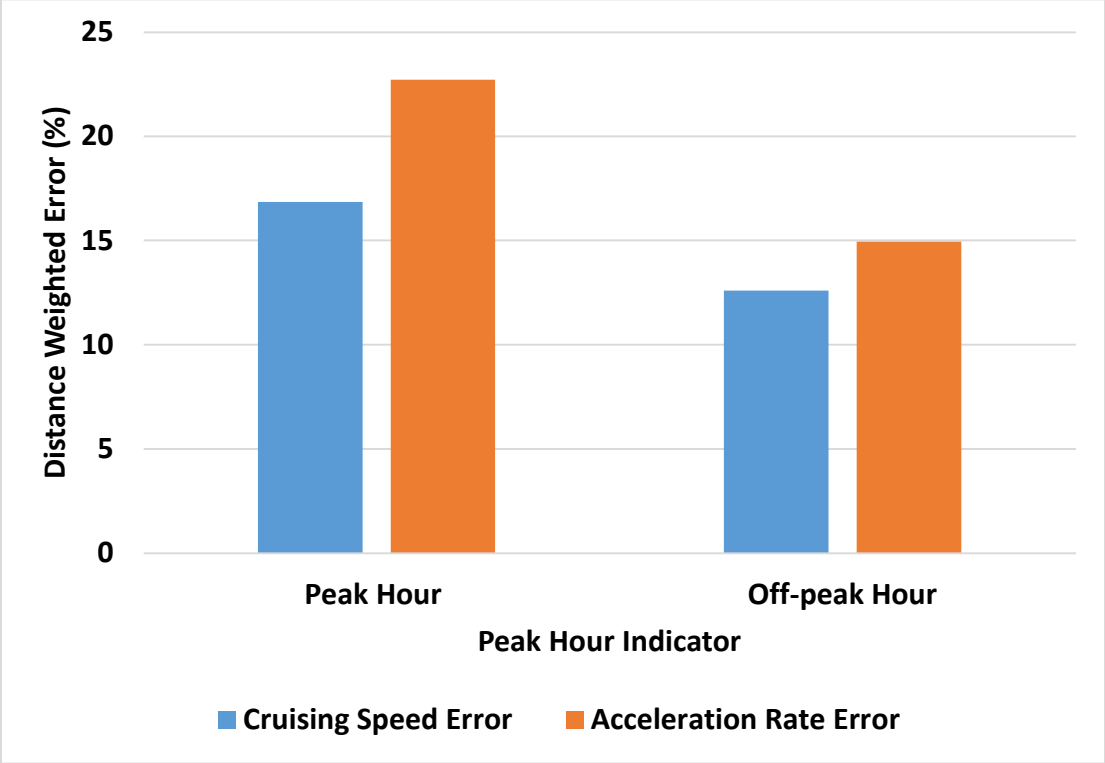

Figure 4-6: Errors of Output Variables for Peak Hours and Off-peak Hours

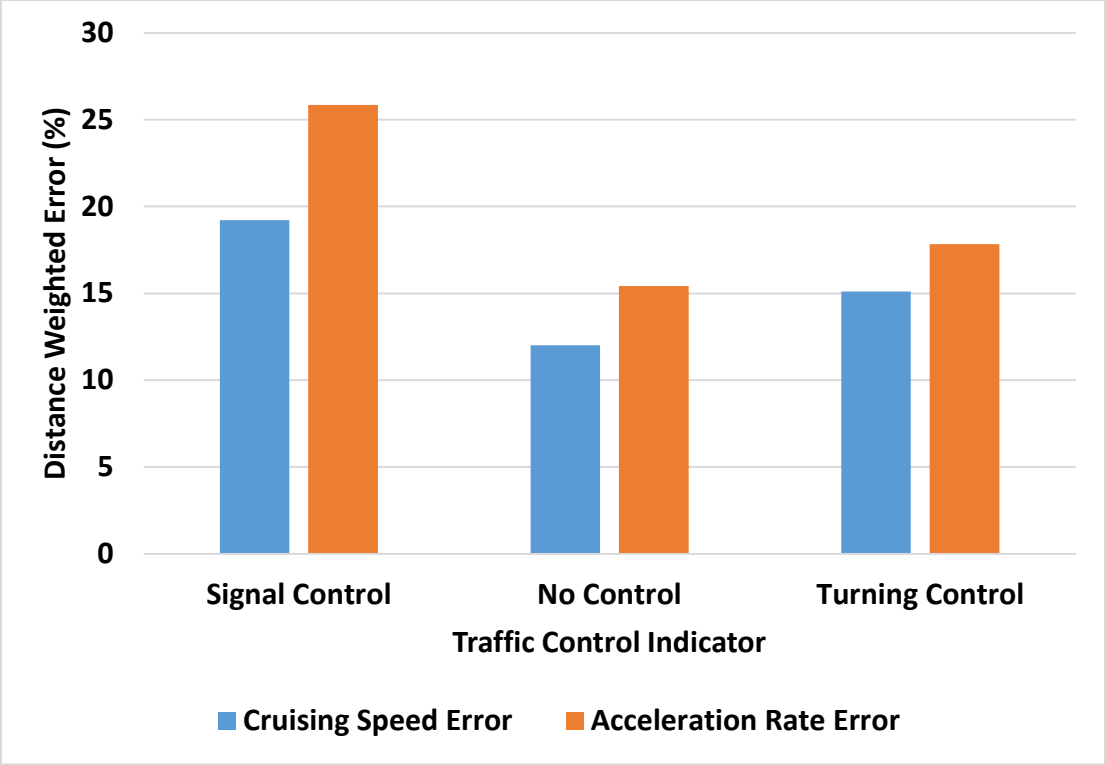

Figure 4-7: Errors of Output Variables for Different Traffic Controls

4.4. Energy Consumptions

Another critical ability of the proposed model is to accurately predict the VSP distribution, thus, enhancing the performance of energy consumption estimation. As introduced in Chapter 3.4, the average cruising speed and acceleration rates predicted by the behavior model are utilized to calculate the mean and standard deviation of VSP distribution through Equation 4 and Equation 5. Energy consumption can be estimated based on the integral of VSP distribution with energy factors, as illustrated in Equation 6.

Emission factors are derived from the simplified MOVES model developed by Frey and Liu in 2013, namely MOVES lite [121]. Similar to MOVES, MOVES lite

estimates the energy consumption based on the conversion table of VSP and operating models. The difference is that MOVES lite uses average energy and emission rates of various bins to reduce complexity [121-123]. As reported in Frey and Liu's work, the simplified MOVES model can represent more than 95% of the total traffic. Moreover, the method considers only limited vehicle types such as passenger car, passenger truck, light commercial truck, short haul truck, and long-haul truck. The relationship between VSP, operation modes, and energy factors can be referred to in Table 3-4 and Table 3-5.

To highlight the advantage of the proposed IOHMM, energy calculated based on VSP distribution with and without adjustment from the IOHMM are compared at link level, as illustrated in Figure 4-8. The X-axes in the two figures represent the observed energy consumptions, while the Y-axes stand for the energy estimation based on the old VSP-based model in Figure 4-8 (a) and the energy estimation based on the proposed behavior-integrated model in Figure 4-8 (b). Each dot in the figure refers to a link-based observation, with its color corresponding to its link-based average speed depicted in the accompanying vertical bar to the left of the chart.

Several phenomena shown in Figure 4-8 are worth noting. First, the density distribution is more concentrated to the cut-off line of y=x in Figure 4-8 (b), indicating that the integration of the IOHMM behavior model can achieve more accurate VSP distribution and energy consumption estimation. Second, the dots with dark shades in Figure 4-8 (a) are widely distributed in the upper half of the cut-off line, while the dots with undertint colors are mostly observed in the middle and lower half. This indicates

that the old VSP-based models are more likely to overestimate energy consumptions on low-speed roads and achieve undervalued results on high-speed roads, which are consistent with findings in other studies. Similar characteristics are observed in the results with IOHMM adjustments, as shown in Figure 4-8 (b). Though the darker shaded dots are still obvious in the upper half, the undertint dots are more evenly distributed near the cut-off line. The findings prove that the proposed IOHMM model outperforms the old VSL-based models in predicting energy consumption on high-level roads with higher travel speeds. For the low-level roads, the improvement of IOHMM is not as expected. The VSP distribution on lower-level roads, which does not fully follow normal distribution, may contribute to this. Moreover, some low-level roads are short with their lengths and the observations on the roads are limited, which also influences the estimation accuracy based on the normal distribution.

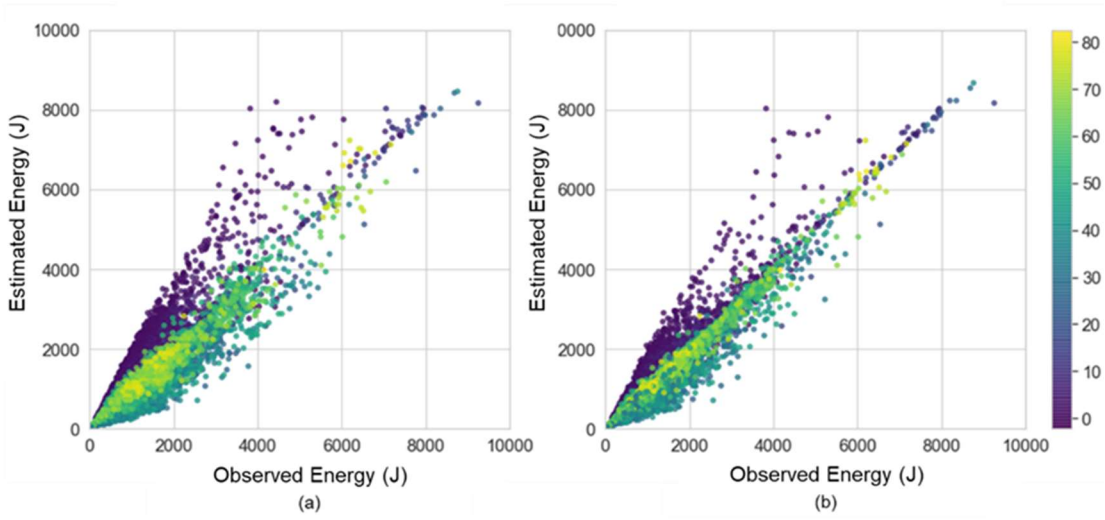

Figure 4-8: Comparison between Observed Energy and Estimated Energy Based on (a) Average Traffic Speed, and (b) Predicted Driving Features at Link Level

Similar comparisons are conducted at trip level, as shown in Figure 4-9. The X-axes in the two figures represent the observed trip energy consumptions, while Y-axes stand for the trip energy estimation based on the old VSP-based model in Figure 4-9 (a) and the trip energy estimation based on the proposed behavior-integrated model in Figure 4-9 (b). Each dot in the figure refers to a link-based observation, with its color corresponding to its trip-based average speed depicted in the accompanying vertical bar. The VSP-based model with IOHMM gives clearly better results than the traditional models without adjustments, and the distance weighted mean squared error significantly decreases from 20.20% to 13.19%.

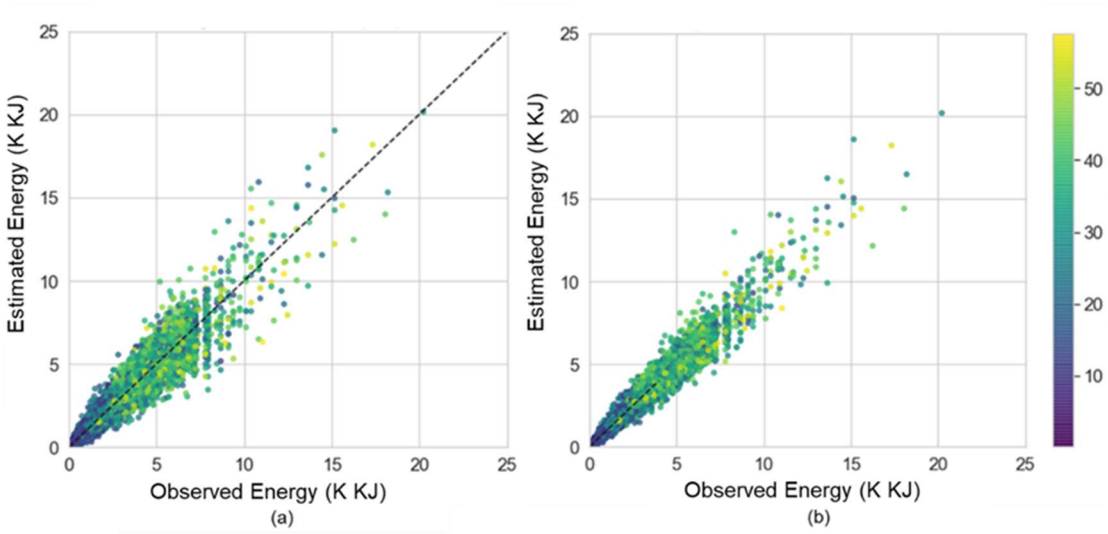

Figure 4-9: Comparison between Observed Energy and Estimated Energy Based on (a) Average Traffic Speed, and (b) Predicted Driving Features at Trip Level

4.5. Comparison Experiments

Here the study compares the results of the proposed IOHMM method with those of the traditional energy models including VSP-based model, aggregate model, modal-

based model, traffic-based model, and average-speed model. The modeling structure, inputs, and outputs of these traditional energy models have been described in Chapter 2. The absolute error and distance weighted error of energy consumption for trips are calculated and compared with the collected ground truth energy data from i2D, as summarized in Table 4-3. When comparing the results to those of previous studies, it must be pointed out that the model structures are directly referred to the instructions in the older papers without further calibrations or validations using existing data. Moreover, some hypotheses are made in the comparison due to the limitation of some input variables. For instance, some system-related factors of the modal-based model are hard to achieve, so the study takes the energy factors from the MOVES lite model instead of the MOVES model. Additionally, the traffic volume on some roads is missing as is the volume calculated from the fundamental relations of traffic flow.

Table 4-3: The Absolute Error and Distance Weighted Error of Various Energy Models

Model	Absolute Error (%)	Distance Weighted Error (%)
IOHMM Model	16.32	13.19
VSP-Based Model	21.65	19.27
Aggregate Modal	22.50	20.15
Modal Model	10.32	9.98
Traffic-Based Model	34.96	29.32
Average-Speed Model	29.20	24.60

Among the five traditional methods, the modal-based model leads to the best results, even if the strict requirements on inputs of detailed driving statistics at second level and consumed computing resources are negligible. A traditional VSP-based model gives the second-best performance, with 21.65% in absolute error and 19.27% in distance weighted error. The aggregate model, widely used in some simple eco-routing systems, is fractionally behind the VSP-based model with a slightly higher error. The traffic-based model that takes traffic speed and volume as inputs presents the highest error with 34.96% in absolute error and 29.32% in distance weighted error. however, the extent to which the accuracy of the traffic-based model can be improved using ground truth traffic volume data instead of estimated traffic volume is unknown. The average-speed model generates slight better results with 29.20% in absolute error and 24.60% in distance weighted error. These two models use energy factors at an aggregated level and do not consider the influence of external environment and individual driving features. Overall, the accuracy and ranks of these models are in accordance with the findings in previous studies. The results of the proposed IOHMM demonstrated in Table 4-3 match the state-of-the-art methods (i.e., modal-based model), with 6% more in absolute error and 3% more in distance weighted error. It must be pointed out that at this stage of understanding, the author believed the implementation of the modal-based model in an eco-routing system is just scarcely possible due to these strict requirements. The results of the proposed IOHMM model also go significantly beyond the traditional VSP-based model, showing that the consideration of the cruising speed and acceleration rates can obviously improve the accuracy of energy prediction. By comparing the results from other older methods, the

proposed IOHMM method takes simpler structure and less-detailed inputs but achieves high accuracy. Therefore, this study determines that the proposed method could be a more appropriate energy prediction model for an eco-routing system.

A further novel finding is that the distance-weighted errors for all methods are significantly smaller than the absolute errors, which suggests the absolute error on short-distance links and short-distance trips are higher than long-distance ones. This also indicates the energy prediction for short links and trips need more effort. A deeper comparison is conducted on various models by trips distance. The average absolute errors of different bins of trips distances are plotted in Figure 4-10. For all models, the short-distance trips achieve significant higher error than other distance bins. With the increase of trip distance, the average absolute error presents an obvious decreasing trend. This basic finding is consistent with the research showing that energy prediction is more difficult for short-distance trips. A popular explanation of this phenomenon is that the short-distance trips include more local roads, which are hard to predict. Moreover, the starting of the vehicle generates a constant amount of energy consumption, which is not considered in this study. The error caused by the dismiss of constant starting energy is more significant in short-distance trips with less energy consumption. Though the decreasing trend can be observed in all six models, the slope is more significant in the proposed model. The difference of average absolute error between the proposed IOHMM model and the modal-based model is about 12% for trips less than 1 mile, while this error is only 3% in the longest-distance bin. This phenomenon suggests the proposed IOHMM has more reliable prediction ability at long-distance links or long-distance trips. This also confirms the assumption that

individual behavior on higher-level roads with long distance and smooth transition is easier to capture with the proposed IOHMM model.

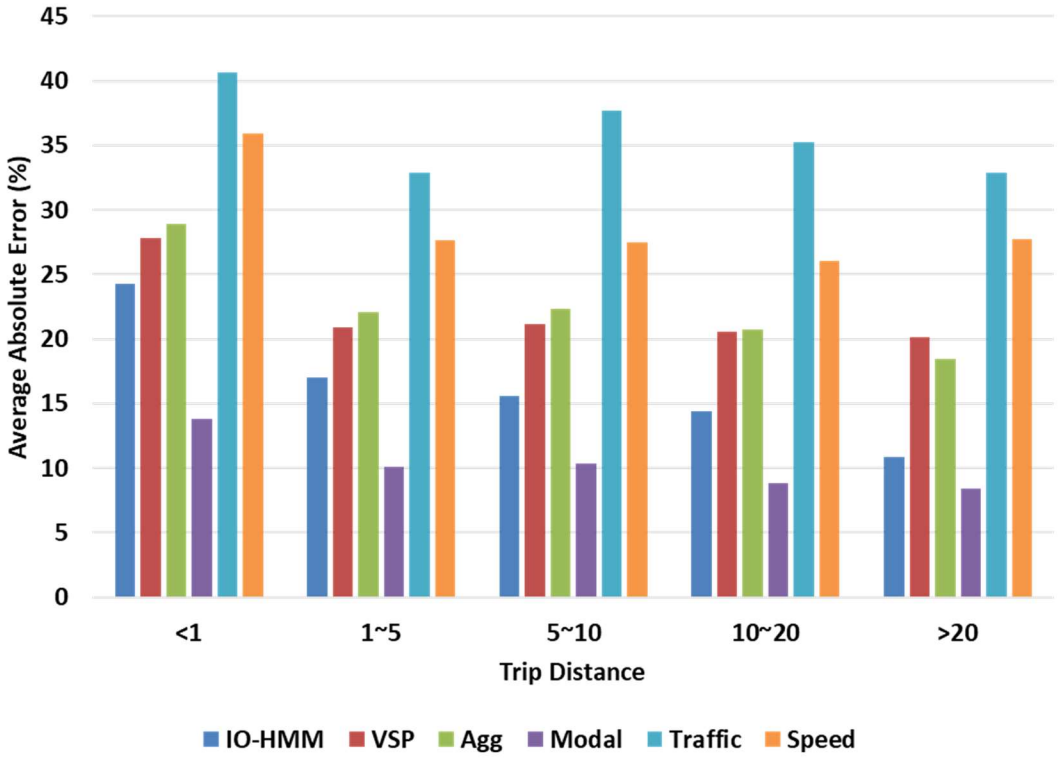

Figure 4-10: Average Absolute Errors for Various Energy Models in Various Bins of Trip Distances

Further statistical analyses are performed using the box plot to represent the minimum, maximum, median, 25th percentile, and 75th percentile, as demonstrated in Figure 4-11. The box plot reveals the absolute error of the modal-based model is more concentrated than other five models. With the exception of the modal-based model, which is not suitable as an energy prediction model, the proposed IOHMM achieves a statistically significant improvement in energy prediction accuracy by capturing individual driving behaviors. However, the model's outliers in the box plot are higher

than some other models, which indicates the proposed model may generate an unstable estimation under some situations. Further analysis of the outliers is worth studying.

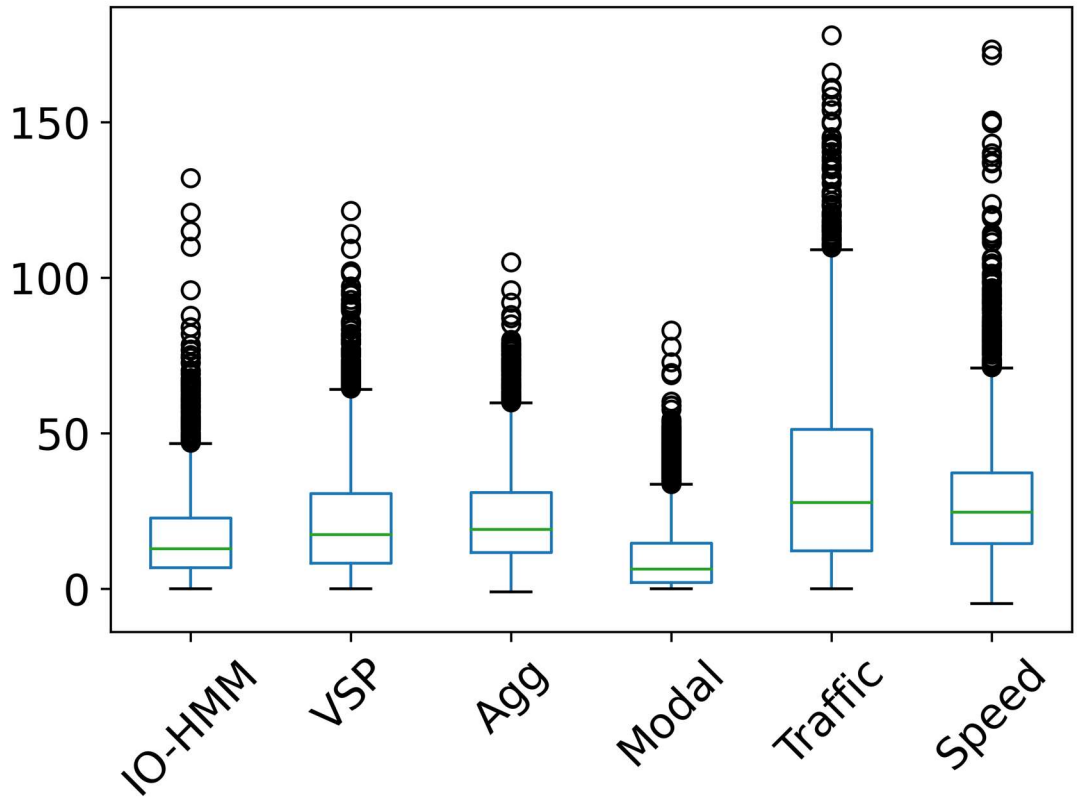

Figure 4-11: The Distribution of Absolute Errors for Various Energy Models

The study also compares the proposed IOHMM with other modeling approaches: Hidden Markov model (HMM), Neural Network (NN), Linear Regression (LR), and Long Short-term Memory (LSTM), as shown in Table 4-4. The proposed IOHMM performs well, giving clearly better results than other models. Long short-term memory models, a widely used model for sequence modeling, gives the second-best results with 20.16% in average absolute error and 17.30% in distance weighted error. When comparing the results from the proposed model to those of the LSTM

model, one thing must be noted. Though the proposed IOHMM model outperforms the LSTM model in general, the LSTM model achieves similar or even higher accuracy on the drivers with the large training dataset. NN and LSTM models result in 20.62% and 19.99% in distance weighted error, respectively. HMM achieves the poorest performance, with 26.15% in absolute error and 22.93% in distance weighted error. This demonstrates that the external environmental variables have significant impact on driving features and energy consumption, which should be fully considered in energy prediction.

Table 4-4: The Absolute Error and Distance Weighted Error of Various Modeling Approaches

Model	Absolute Error (%)	Distance Weighted Error (%)
IOHMM Model	16.32	13.19
HMM Model	26.15	22.93
Neural Network	23.41	19.99
Linear Regression	24.96	20.62
Long Short-Term	20.16	17.30

4.6. System Benefits Evaluation

To evaluate the performance of the proposed eco-routing method for real-world implementation and deployment, a statewide traffic simulation model is carried out to analyze system benefits. The traffic simulation model covers the Greater Washington metropolitan area, encompassing Maryland, Washington, D.C., and Northern Virginia. The model generation includes two steps: network generation and demand estimation.

The network is transformed from the HERE navigation network for the year 2019. Additional links are added to construct the signalized intersections, split the bi-directional links, and connect traffic analysis zones (TAZ). The reconstructed network includes 2,089 traffic analysis zones (TAZ), 215,202 nodes, and 366,464 links, as displayed in Figure 4-12. Leveraging the massive cell phone location data, more reliable traffic demand data could be used as seed inputs for the traffic simulation model. This proceeds in four stages: trip identification, mode imputation, trip chaining, and trip population [124].

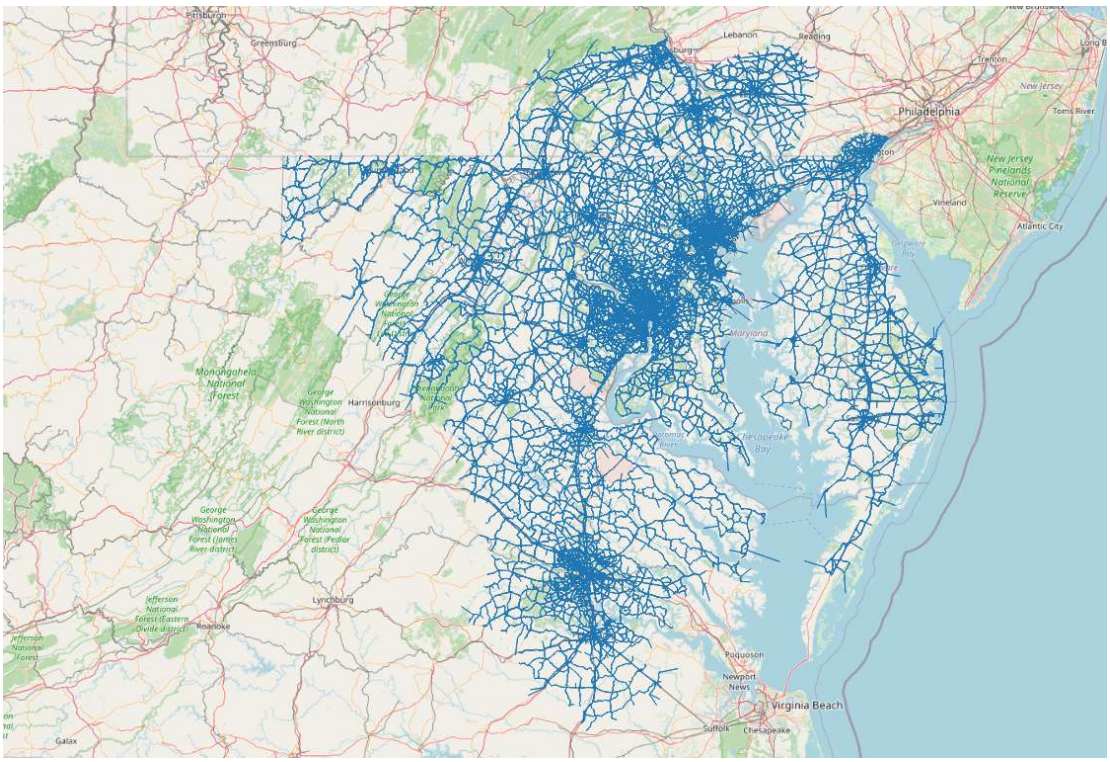

Figure 4-12: Simulation Model for Benefits Analysis

Over the course of the traffic demand estimation, calibration and validation play important roles. Several critical parameters—for instance traffic volume, traffic speed,

and other supply-side factors—must be calibrated to build a reliable simulation model. Link-based volumes are calibrated based on hourly traffic counts data collected from 179 field sensors in 2015, which reside in the Regional Integrated Transportation Information System (RITIS) and the State Highway Administration (SHA) Traffic Monitoring System (I-TMS) database. In addition to the Origin Destination Matrix Estimation (ODME) function in DTAlite, manual effort also contributes to the calibration and validation process to deal with complex roadway segments. The framework of calibration and validation process is summarized as Figure 4-13.

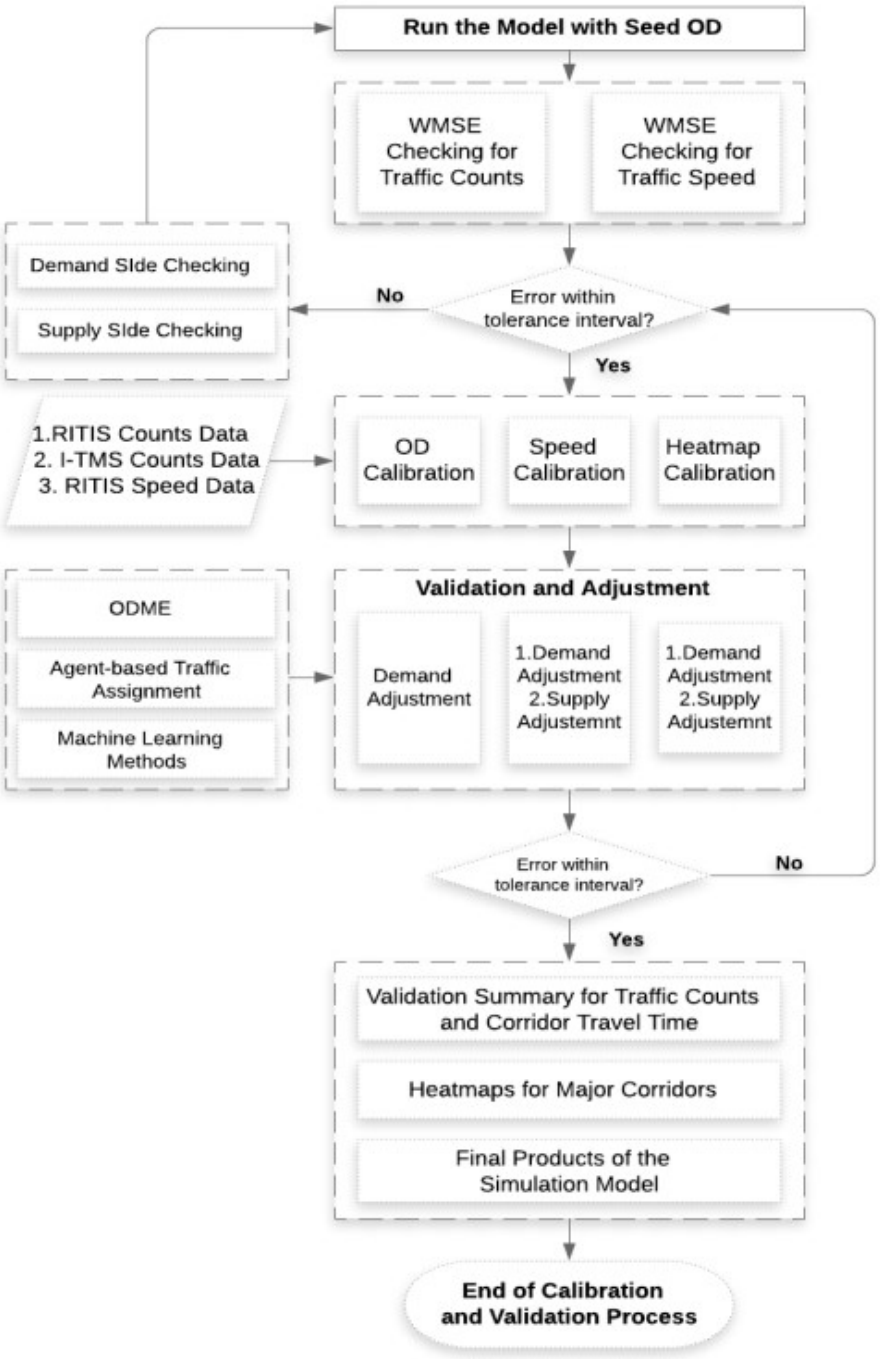

Figure 4-13: Framework of Calibration and Validation Process of Simulation Model

The proposed eco-routing algorithm is simulated with the calibrated travel demand model under different scenarios through a series of pre-defined levels of user penetration represented by the parameter-named ratio of audiences (i.e., a 5% ratio of audiences means 5% of travelers follow the guidance of the proposed eco-routing system). Though the personalized IOHMM models are trained for all drivers, the ratio of the model distribution in all populations is unknown. Therefore, the drivers are clustered into four groups through a simple K-means algorithm, and a separate IOHMM driving behavior model is trained using the trips from all users in each group. The trained modes are randomly assigned to the drivers with the same proportion as the training dataset (i.e., 46% for group 1, 24% for group 2, 19% for group 3, and 11% for group 4). For each driver, the top five routes with the least amount of travel time during the same departure time window are collected from the simulation model, and the energy consumption for the routes are estimated. Each route is assumed to start from the 'idling' state. Relevant contextual information, transition probabilities, and emission probabilities are then used to determine the states and output variables. The output variables for all stages are estimated, then the link based VSP distributions are obtained, thus the energy consumptions are calculated. The route with minimum fuel consumption is considered the eco-route for each driver, which will be updated as the driver's route in the next iteration. The simulation model is re-run using the updated routes, and the system performance is summarized.

Table 4-5: System Benefits under Various Percentage of Audience

Ratio	Energy Saving Per Vehicle (KJ)	Energy Saving Per vehicle without Eco-routing (KJ)	Energy Saving Per vehicle with Eco-routing (KJ)
5%	129.23	7.02	88.52
10%	131.15	10.11	149.75
15%	141.1	16.01	234.42
20%	147.17	25.41	335.46
25%	150.12	28.01	394.62

From the summary above, key findings emerge. First, the application of the proposed algorithm not only reduces the energy consumption of the app users, but also benefits the non-app users in the system. One possible reason accounts for this phenomenon. With the users changing their routes, the congestion on some high-occupied roads could be released. Therefore, the non-app users also achieve higher fuel efficiency, although their routes remain the same. Second, with the increase of the audience ratio, the system benefits increase. Extensive results carried out show that the increasing speed first speeds up then slows down. One possible explanation for this is that more users shifting to other routes may result in unexpected congestion on some roads. This is an important finding in understanding the optimal ratio of the audience.

Nevertheless, there are several limitations to the designed analysis. First, the traffic speed, one of the contextual input variables, is assumed to be constant under different scenarios. However, the case is not consistent with real world conditions. With users changing routes, the traffic conditions would change accordingly, which may

affect the prediction of driving behaviors and driving features. Second, though the routes of app users can be adjusted based on energy consumption, the routes of non-app users are assumed to be fixed. This also goes against the real-world condition that users usually seek the fastest routes under various traffic conditions. This may raise concerns about the difficulties of integrating an eco-routing system with dynamic traffic assignments, which can be addressed in further research.

4.7. Other Findings

In addition to the analysis stated above, the study reveals several other interesting findings. First, the study compares the routes with minimum energy consumption and routes with minimum travel time. Collected trips from i2D and the incenTrip app are first matched with simulated trips from the statewide simulation model based on trajectories and departure times. Then other feasible routes for the matched trips are filtered based on origin, destination, and departure time. Energy consumption of the other feasible routes is estimated through the trained personalized behavior-integrated energy prediction model. For each collected trip, the routes with minimum fuel cost, minimum user cost (defined as fuel cost plus value of time), and minimum system cost (defined as fuel cost plus societal cost) are achieved. The percentages of trips with minimum travel time in each group are plotted for different distances (defined as the OD distance under free flow travel time) bins, as presented in Figure 4-14.

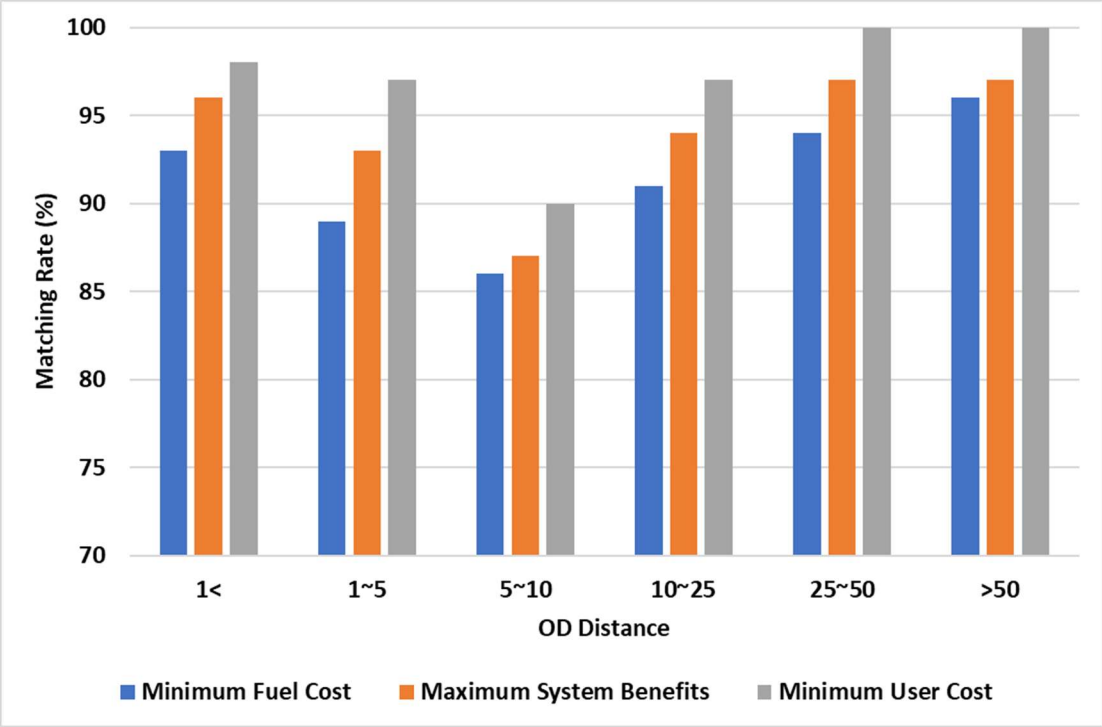

Figure 4-14: Matching Rate of Groups with Various Trips under Various OD Distance

In general, 90% of the routes with minimum energy consumption also have minimum travel time, and the percentages for maximum system benefits and minimum user cost are 93% and 95%, respectively. These indicate that the shortest route is not always the one with minimum energy consumption or minimum societal impact, which is consistent with the assumptions that are generally accepted these days. Moreover, the percentages in mid-distance bins (i.e., 5~10 miles and 10~25 miles) are obviously lower that the values in short-distance bins (1 mile< and 1~5 miles). The differences are more significant between mid-distance bins and long-distance bins (25~50 miles and >50 miles). One possible reason is that feasible routes with similar travel time are limited for a short-distance trip, so the route with the shortest travel time usually has the lowest fuel cost and user cost. The case is similar for a long-distance trip, in that the route is usually unique except under severe traffic congestion. On the other hand, a

mid-distance trip has more feasible routes with similar travel time, so the shortest-travel-time route is usually the most fuel-efficient option.

Additionally, the study undertakes the empirical analysis on the relationship between fuel efficiency and drivers' familiarity to the routes. Based on the number of trips per route, the routes are categorized into six groups. The fuel efficiency of each trip is calculated and the distance weighted fuel efficiency for all groups is summarized in Figure 4-15. As demonstrated in the figure, the distance weighted fuel efficiency increases with the increase of number of trips per route. This phenomenon indicates that a driver can operate the vehicle more smoothly on familiar roads and avoid less efficient driving behaviors. Driving on roads with frequent uphill and downhill segments would be a typical example. Though the traffic signs along the road may catch the driver's attention, the perception of the road ahead would still be very limited if the driver is unfamiliar with the road. The driver is more likely to take a sharp deacceleration when he encounters a curve or signal ahead after passing the uphill. The situation is quite the opposite for a driver who is familiar with the road. The driver could take advance deceleration based on the experiences, even when the curve or signal is not visible, to avoid sharp deceleration and increase fuel efficiency. Based on these assumptions, the familiarly could be treated as an input variable or a hidden state of the proposed behavior-integrated model in further deployments.

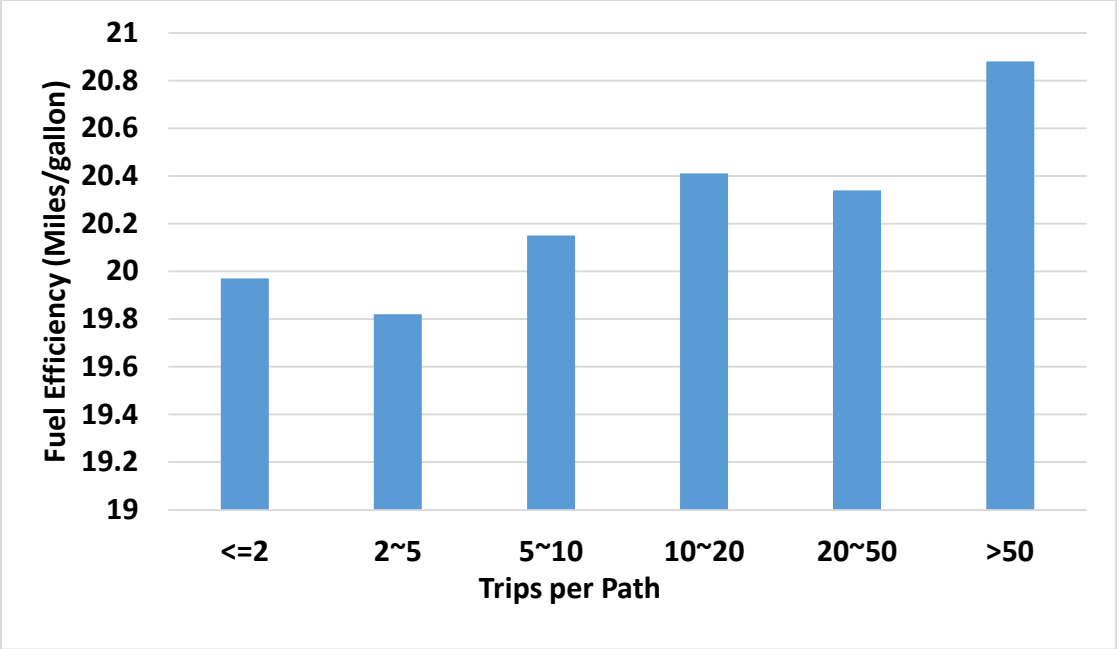

Figure 4-15: Average Fuel Efficiency of Groups with Various Number of Trips per Path

Chapter 5 Applications

This chapter presents the applications of the proposed behavior-integrated energy prediction approach on individual eco-routing and large-scale network monitoring and forecasting.

5.1. Eco-routing System

Supported by the Department of Energy (DOE), incenTrip, a comprehensive trip planning APP has been developed. incenTrip helps commuters in the Washington metropolitan region to find optimal commuting options. The application provides four distinct services to commuters: 1) provide cash rewards for a smart commute, 2) identify multimodal commute options, 3) avoid traffic delays, and 4) monitor the system impact [130]. However, the driving trips could be reduced, but could not be avoided in most cases. To reduce energy and emissions consumption of driving trips, the team is currently developing and deploying the proposed behavior-integrated energy prediction model in the incenTrip APP.

The incenTrip APP has thousands of active users, most of which are commuters in Washington metropolitan area. The APP records the trip trajectory data of the users to improve user experience. Trip trajectory data stored online includes rich travel information with high frequency (e.g., 1 HZ) such as speed, acceleration, and location, which can be matched with roadway geometry features through map matching. The rich travel information can then be used as inputs for training personalized driving behavior model, whose outputs contribute to achieving accurate energy estimation. As

demonstrated in Figure 5-1, the operation process of the eco-routing module consists of three modules: online database, behavior modeling, and eco-routing.

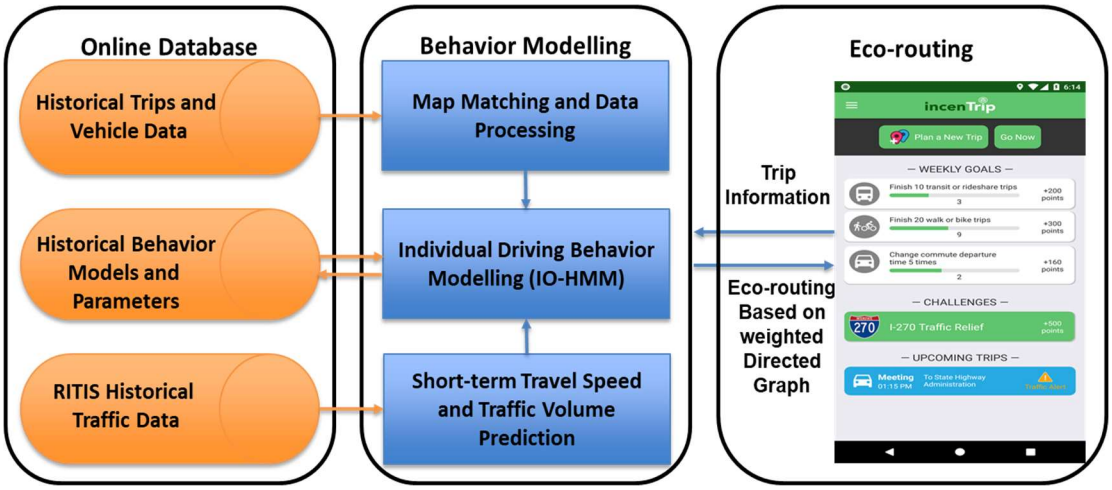

Figure 5-1: Framework of the Deployment of Eco-routing System on incenTrip Application

Online database stores historical trip trajectories, vehicle information, and user information. The historical traffic data is also stored in the online database, which contributes to the short-term traffic speed and volume estimation. Every time a user plans a trip, the front-end sends the trip information, for instance, departure time, origin, and destination to the behavior modeling module. Several feasible routes are then generated by the routing algorithm and the corresponding energy consumption is estimated using the pre-trained behavior-integrated energy prediction model stored in the online database. For the users with limited driving records, the general models for different behavior groups could be used instead. The routes with minimum fuel cost, user cost, or societal impacts are filtered and returned to the front-end as different travel options. User can select the most appropriate route fitting personal needs. Sometime the route with minimum energy consumptions may not be the optimum solution if the

travel time is too long. Therefore, we eco-routing function considers travel time, fuel consumption and system benefits. For each possible route, the total cost is denoted as:

$$C_{i,t} = T_{i,t} * c_{VOT} + F_{i,t} * c_F + \sum_{j=0}^{J} E_{i,t,j} * c_j \quad (21)$$

Where $C_{i,t}$ refers to the total cost of route i departing at time t, while $T_{i,t}$ and $F_{i,t}$ represent the total travel time (hours) and total fuel consumption (gallon). The unit cost of travel time and fuel consumption are indicated as c_{VOT} and c_F. c_{VOT} of this study is set as a \$25.13/hour which is the median hourly wage rate for all employees working in the Washington metropolitan area, as reported by the Bureau of Labor Statistics. c_F is a dynamic variable that depends on fuel type and real-time fuel price. The third part of the equation refers to the system impact of emission consumption, where $E_{i,t,j}$ stands for the consumption of emission item j, and c_j means the corresponding system impact. Cost for all emissions is considered as the societal cost generated by FY 2015-17 TDM program, as shown by Table 5-2. The eco-routing system aims at providing user the route with the largest benefits, so the objective function is formulated as:

$$Minimize\ C_{i,t}, \quad i \in I\ and\ t \in T \quad (22)$$

As an important input of hidden Markov behavior model, link-based average speed should be accurately predicted. As introduced previously, a long short-term memory model can be developed for average speed prediction, as described in section 3.3.1. One thing worth noticing is that the average speeds for all links in the network are predicted every 10 minutes. This is because running LSTM model takes some time.

Table 5-1: Societal Cost of Various Types of Emissions Generated by 2015-17 TDM Program

Societal Benefit	Benefit Unit	Cost Per Unit of Benefit
NOX	Tons removed	$1,612
VOC	Tons removed	$133
PM2.5	Tons removed	$15,107
PM2.5NOX	Tons removed	$1,612

After the user finishing the trip, the trip trajectory data is uploaded to the online database. Leveraging the latest sets of trip trajectories, the user's personalized behavior model is trained and updated.

5.2. System Energy Monitoring

Although the proposed behavior-integrated prediction model cannot be directly used to predict link-based energy consumption, its hidden states can help to improve the accuracy of links with high bias. The hidden set of hidden Markov model represents the unobservable factors that influence the output variables. Recall the estimation results using average-speed model, there are some segments with obvious differences between the predicted energy consumption and the observed value. Although some deep learning methods may help to decrease the difference, it is still necessary to understand the reasons and come up with general methods. The proposed system is composed of three parts: link-based average speed prediction, link-based volume prediction, and fleet type and age distribution prediction.

Similar with average speed prediction, link-based volume is predicted through a spatial-temporal traffic flow method using LSTM. LSTM can handle the long-term dependency in the traffic flow data and discover the latent feature representations hidden in the traffic flow, which yields better prediction performance. Historical traffic volume data from RITTIS is used to train the LSTM model.

In addition to traffic condition effect, energy consumption is also influenced by vehicle-related factors such as vehicle type, vehicle age, and fuel type. Some existing studies weight the traffic volume with fleet type distribution to improve accuracy. This study proposes tested a method to predict time-dependent link-based fleet distribution using trajectory data collected from mobile phone devices and vehicle registration data at zip code level. The rational for this testing is due to national or regional default fleet factors do not reflect local condition The trajectory data collected from mobile devices is used to get traffic demand, while vehicle registration data provides age and type distribution to weight traffic demand from each zone.

Chapter 6 Summary and Further Work

6.1. Summary of the Research

Theories of eco-routing and eco-driving have drawn increasing attention in both research and practice as the public becomes more concerned with environmental quality issues. However, the practical foundations of energy prediction models for eco-routing systems have not been well established. Microscopic models generate accurate results but require detailed inputs and considerable computing resources. Macroscopic models are simple to implement but fully dismiss driving dynamics. Some mesoscopic models are a trade-off between microscopic and macroscopic ones using VSP distribution, which can be estimated using various driving features. However, the existing VSP-based models estimate the VSP distribution using average traffic speed or speed limit, which achieves poor accuracy in energy estimation (73.22% at trip level and 64.37% at link level). Taking the instantaneous VSP under average traffic speed as the mean of VSP distribution assumes that the speed is normal distribution on a link and is linearly related with VSP, which is not true in most cases. Further analysis indicates that the mean and variance of VSP distribution are strongly related to other driving features, for instance, average cruising speed and acceleration rate.

External environmental factors prove to significantly influence driving features, affecting energy consumption and emissions. Additionally, various unobserved driver-related factors also influence the driving features—for instance, emotions, pressures, and sensitivity. These unobserved factors show obvious individual differences and are collectively referred to as driving behaviors. Understanding the relationships among

external roadway environment, driving behaviors, and driving features is critical for building a reliable eco-routing system.

Considering the interplay of these three groups of factors, a train-based IOHMM with similar structure is trained and deployed to estimate individual driving features using historical trip trajectory data. Compared with standard HMM, the IOHMM has an additional input layer, which consists of observed contextual variables such as historical traffic statistics and roadway geometry information. The hidden layer of IOHMM captures the influence of unobservable driving behaviors and the outputs are predicted driving features. Three types of parameters are involved in the proposed model: initial probability model, transition probability model, and emission model. The parameters are formulated based on the variables of different layers and are learned through Expectation-Maximization and Forward-Backward algorithms.

Next, the experimental study is given to explain the modeling process and the performance of the proposed algorithm. Personalized behavior models are developed based on drivers' historical trip trajectory data. Various measurements of effectiveness are conducted to evaluate the model performance: for example, state recognition, driving features estimation, and energy estimation at link and trip levels. State recognition is evaluated by comparing the estimated states with the labeled states. In general, an accuracy of 82% is achieved, which indicates that the proposed IOHMM model can capture the driving behavior with great success. Then, the driving features—including average cruising speed and acceleration rates—are compared with ground truth data at link level. The distance weighted errors are 13.15% and 17.27% for

average cruising speed and acceleration rates, respectively. Moreover, the results also prove that the proposed approach with IOHMM outperforms other methods in predicting energy consumption with an error of 12.80%. Additionally, a state-level simulation model is developed and tested to analyze system benefits. The system performance indicates that the widely implementation of the proposed approach not only benefits the users but also reduces the energy consumption and traffic congestion of the entire system.

In summary, this study presents a framework of modeling individual driving-behavior-generated driving features for effective short-term energy consumption and emissions prediction. In the proposed framework, the prediction process doesn't need the VSP at a second-by-second scale as it has been. Instead, the VSP distribution is predicted based on the personalized IOHMM training through drivers' historical trip trajectory data. The integration of IOHMM model overcomes the limitation of old energy prediction approaches in coping with driving behavior impact, resulting in more accurate VSP distribution and energy consumption prediction.

6.2. Further Work

6.2.1. Model Selection

In this study, the input output-hidden Markov model is utilized to map a sequence of input variables into output variables. Except for the statistical approaches, various machine learning algorithms—for instance, LSTM, deep neural network (DNN), and generic algorithm (GA)—can also be used to solve the prediction problems of sequence data. The selection of statistical methods and machine learning methods

are mainly depending on the size of input data and logistics assumptions. For limited size of inputs, the models with simple structure (e.g., hidden Markov model, input output-hidden Markov model, and Bayesian model) can produce better results than complex models. The most critical reason is that training with limited size of data with simpler models can avoid overfitting. When a lot of data is available, more complex models (DNN and LSTM) outperform the simpler ones since a more accurate relationship can be explored. This is also proved in the comparison experiments of this study. Though the proposed IOHMM achieves higher accuracy than the LSTM model in general, contradictory results are observed in some users with a large training dataset. The relationship between data size and model accuracy is worth further exploration. With the wider implementation of the application and communication with more data sources, richer personal trip records are available and more complex models could be utilized. Additionally, applying different models on drivers with different trip records is also an innovative approach in eco-routing applications.

6.2.2. Model Generation

Though the framework has been thoroughly designed, developed, and evaluated, several aspects still need further exploration for real-world implementation.

The selection of input and output variables of the proposed IOHMM may benefit from further analysis. Previous studies tend to determine these variables based on the researchers' own practical experiences. In this study, the Pearson correlation analysis is used to select the appropriate variables—high coefficients output variables. Similar analysis is conducted in deciding the input contextual variables; the candidate

variables are also selected this way. A more systematic evaluation of all relevant variables and a more comprehensive analysis of their sensitivity could provide more certainty on the appropriateness of selecting input and output variables.

Hidden states also worth further exploration. The number of hidden states in the input output-hidden Markov model is usually hard to determine. More hidden states can help the model handle a wide range of driving behaviors, but the performance is usually limited by the data size and data quality. A smaller size of hidden states can expedite the training process; a dynamic number of hidden states is a potential solution, instead of a constant value. The number of hidden states can be dynamically determined by the data size and data quality of the historical trips. Further quantitative research on the exact number of hidden states could provide more certainty in the process and potentially improve computation efficiency. The study first trains the IOHMM through unsupervised learning without labeling the hidden states. The results of unsupervised learning are used to select the most appropriate rule for hidden states labeling, which are later used for supervised learning. The experimental results of state recognition in this study highlight the capabilities of the proposed IOHMM in estimating the driving behavior of individuals. However, confusion can be observed in some states, which indicates the definitions of hidden states are not appropriate enough. The hidden states can be better defined through classification methods such as supper vector machine (SVM), random forest, decision tree, logistic regression, and K-nearest neighbor.

As shown by the experimental results, the proposed IOHMM achieves significantly low accuracy of energy prediction on local roads. One possible reason that

contributes to this phenomenon is that local roads usually have short link distances in the HERE navigation network. A limited number of observations indicates the VSP may follows other distributions (e.g., linear distribution) instead of Gaussian distribution. Merging several adjacent links with similar geometry features together could help to receive richer link-based GPS observations.

6.2.3. Model Implementation

Though the personalized input output-hidden Markov model can capture a driver's driving behavior, it is not applicable in some cases. For instance, the trip records of a newly registered driver are limited, so the data size is insufficient to support training a personalized model. For application purposes, more general driving behavior models are needed. Drivers could be categorized into different groups based on driving styles or behaviors through various categorized algorithms. In general, categorized algorithms include clustering algorithms (e.g., decision tree, Bayesian classifiers, logistic regression, and support vector machine) and classification algorithms (e.g., K-means algorithms and hierarchical clustering), based on the knowledge of the outputs. Clustering is framed by unsupervised learning, which takes a set of inputs without previously knowing the desired outputs. On the other hand, classification algorithms belong to supervised learning, which means that both the inputs and the outputs are pre-defined. This study suggests that K-means algorithm could be adopted to study the features of the collected data and classify the drivers. Pearson correlation coefficients between driving features and trip energy consumption are first calculated. Variables with significant correlations should be selected as inputs for the K-means model according to the standard guidance of Pearson correlation coefficient. Then, the trips

belonging to drivers in each group are combined as the training dataset for general behavior modeling, which can later be utilized to estimate energy consumption for drivers with limited driving records.

Though the proposed algorithm help users select eco-routes with minimum energy consumption, how to benefits the entire system worth further analysis. As discovered in Section 4.6, the increment of system benefits slows down after the ratio of audience reaches a specific level. Two possible reasons contribute to this inertia. First, with more users detour to eco-routes, some new bottlenecks arise when some local roads cannot afford the heavy influx of traffic. Second, the non-APP users are assumed not to change their routes, which is inconsistent with real-world situation. Here the study comes up with several ideas to extend the proposed algorithm from user optimum to system optimum.

1) Though real time simulation can help generate system optimal solutions, it is not acceptable for real world implementation due to time consuming. However, the primary group of incenTrip users are commuters with relatively stable travel patterns, for instance, work location, home location, and departure time. Therefore, these users' recommended routes could be predicted in advance by offline simulation models, which are simulated and calibrated using real-time and historical traffic statistics. Instead of setting minimum travel time as objective function, the simulation model aims at maximizing the system benefits. When simulating a user's trajectory, the driving features on a link are predicted using the pre-trained driving behavior model and the energy

consumption is estimated through the behavior-integrated energy prediction model. After predicting the routes with maximum system benefits, the routes are recommended to the users through the incenTrip APP in advance (e.g., 15 or 30 minutes prior to the normal departure time). To increase the efficiency and accuracy of the prediction, several offline models could be developed separately for different time periods.

2) For the non-commute trips, the process described in Section 5.1 could be used to predict eco-routes. Except the factors formulated in equation 21, the penalty factor representing the system impact from detouring (i.e., detouring from routes with least travel time) could also be considered. For instance, detouring from a heavy congested road will have a negative penalty, while positive penalty is applied if detouring to a road that near saturation. A simple penalty function can be formulated as:

$$P = \varepsilon(\frac{V}{C})^{\epsilon} \quad (22)$$

Where V stands for the existing traffic volume and C represents the capacity of the link. ε and ϵ are link-based penalty factors, which can be estimated in further analysis.

Autonomous vehicles, also known as self-driving vehicles, are vehicles that operate based on screening the external environment through multiple sensors and moving safely with little or no human effort. Recently, the autonomous driving technology has been promoted significantly by the rapid advances in computer vision and deep neural networks. More researchers suggest that the self-driving vehicle will

be an accepted trend for the future. Though the proposed model in this study focuses on drivers' behavior and driving features, it can also contribute to the autonomous driving vehicles. Similar with driver-driving vehicles, energy consumption of self-driving vehicles is also determined by various unobservable factors, for instance, vehicle performance and auto-driving algorithms. The vehicle performance and auto-driving algorithms can be involved in the hidden layer as hidden states to capture the influence on driving features. Additionally, most autonomous vehicles also have driver-driving mode for safety concerns. The changes between driver-driving status and autonomous driving status can influence driving features and should also be included as hidden states. Utilizing the historical trip trajectory data, the personalized IOHMM is trained and the integrated model is used to predict energy consumption for autonomous driving vehicles.

References

[1] Summary of Travel Trends – 2017 National Household Travel Survey. Federal Highway Administration, U.S. Department of Transportation, 2018. *https://nhts.ornl.gov/assets/2017_nhts_summary_travel_trends.pdf*

[2] Zeng, W., Miwa, T., & Morikawa, T. (2020). Eco-routing problem considering fuel consumption and probabilistic travel time budget. Transportation Research Part D: Transport and Environment, 78, 102219.

[3] Lave, L., MacLean, H., Hendrickson, C., & Lankey, R. (2000). Life-cycle analysis of alternative automobile fuel/propulsion technologies. Environmental Science & Technology, 34(17), 3598-3605.

[4] Samaras, C., & Meisterling, K. (2008). Life cycle assessment of greenhouse gas emissions from plug-in hybrid vehicles: implications for policy.

[5] Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990 - 2017. United States Environmental Protection Agency, 2019.

[6] Ahn, K., & Rakha, H. A. (2013). Network-wide impacts of eco-routing strategies: a large-scale case study. Transportation Research Part D: Transport and Environment, 25, 119-130.

[7] Boriboonsomsin, K., Barth, M. J., Zhu, W., & Vu, A. (2012). Eco-routing navigation system based on multisource historical and real-time traffic information. *IEEE Transactions on Intelligent Transportation Systems*, *13*(4), 1694-1704.

[8] Ahn, K., & Rakha, H. A. (2013). Network-wide impacts of eco-routing strategies: a large-scale case study. *Transportation Research Part D: Transport and Environment*, *25*, 119-130.

[9] Rakha, H. A., Ahn, K., & Moran, K. (2012). Integration framework for modeling eco-routing strategies: Logic and preliminary results. *International Journal of Transportation Science and Technology*, *1*(3), 259-274.

[10] Zeng, W., Miwa, T., & Morikawa, T. (2016). Prediction of vehicle CO2 emission and its application to eco-routing navigation. *Transportation Research Part C: Emerging Technologies*, *68*, 194-214.

[11] Yi, Z., & Bauer, P. H. (2018). Optimal stochastic eco-routing solutions for electric vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *19*(12), 3807-3817.

[12] Huang, X., & Peng, H. (2018). Eco-routing based on a data driven fuel consumption model. *arXiv preprint arXiv:1801.08602*.

[13] Sun, J., & Liu, H. X. (2015). Stochastic eco-routing in a signalized traffic network. *Transportation Research Procedia*, *7*, 110-128.

[14] Yi, Z., Smart, J., & Shirk, M. (2018). Energy impact evaluation for eco-routing and charging of autonomous electric vehicle fleet: Ambient temperature consideration. *Transportation Research Part C: Emerging Technologies*, *89*, 344-363.

[15] Bandeira, J. M., Fontes, T., Pereira, S. R., Fernandes, P., Khattak, A., & Coelho, M. C. (2014). Assessing the importance of vehicle type for the implementation of eco-routing systems. *Transportation Research Procedia*, *3*, 800-809.

[16] De Nunzio, G., Sciarretta, A., Gharbia, I. B., & Ojeda, L. L. (2018, November). A constrained eco-routing strategy for hybrid electric vehicles based on semi-analytical energy management. In *2018 21st international conference on intelligent transportation systems (itsc)* (pp. 355-361). IEEE.

[17] Boriboonsomsin, K., Barth, M. J., Zhu, W., & Vu, A. (2012). Eco-routing navigation system based on multisource historical and real-time traffic information. *IEEE Transactions on Intelligent Transportation Systems*, *13*(4), 1694-1704.

[18] Ahn, K., & Rakha, H. A. (2013). Network-wide impacts of eco-routing strategies: a large-scale case study. *Transportation Research Part D: Transport and Environment*, *25*, 119-130.

[19] Rakha, H. A., Ahn, K., & Moran, K. (2012). Integration framework for modeling eco-routing strategies: Logic and preliminary results. *International Journal of Transportation Science and Technology*, *1*(3), 259-274.

[20] Zeng, W., Miwa, T., & Morikawa, T. (2016). Prediction of vehicle CO2 emission and its application to eco-routing navigation. *Transportation Research Part C: Emerging Technologies*, *68*, 194-214.

[21] Yi, Z., & Bauer, P. H. (2018). Optimal stochastic eco-routing solutions for electric vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *19*(12), 3807-3817.

[22] Huang, X., & Peng, H. (2018). Eco-routing based on a data driven fuel consumption model. *arXiv preprint arXiv:1801.08602*.

[23] Sun, J., & Liu, H. X. (2015). Stochastic eco-routing in a signalized traffic network. *Transportation Research Procedia*, *7*, 110-128.

[24] Zhao, X., Ye, Y., Ma, J., Shi, P., & Chen, H. (2020). Construction of electric vehicle driving cycle for studying electric vehicle energy consumption and equivalent emissions. Environmental Science and Pollution Research, 27(30), 37395-37409.

[25] Chen, X., M. Hadi, and Y. Xiao . Development of Macroscopic Emission Estimation Model Based on Microscopic Operating Modes. Transportation Research Record: Journal of the Transportation Research Board, 2016. 2570(1):39-47.

[26] Lyons, T. J., Kenworthy, J. R., Austin, P. I., & Newman, P. W. G. (1986). The development of a driving cycle for fuel consumption and emissions evaluation. Transportation Research Part A: General, 20(6), 447-462.

[27] Fiori, C., Ahn, K., & 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.

[28] Ahn, K., Rakha, H., Trani, A., & Van Aerde, M. (2002). Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. Journal of transportation engineering, 128(2), 182-190.

[29] Liu, K., Yamamoto, T., & Morikawa, T. (2017). Impact of road gradient on energy consumption of electric vehicles. Transportation Research Part D: Transport and Environment, 54, 74-81.

[30] Liu, Y., Gao, J., Qin, D., Zhang, Y., & Lei, Z. (2018). Rule-corrected energy management strategy for hybrid electric vehicles based on operation-mode prediction. Journal of Cleaner Production, 188, 796-806.

[31] Xu, X., Aziz, H. A., & Guensler, R. (2019). A modal-based approach for estimating electric vehicle energy consumption in transportation networks. Transportation Research Part D: Transport and Environment, 75, 249-264.

[32] Zhao, Y., & Sadek, A. W. (2013). Computationally-efficient approaches to integrating the MOVES emissions model with traffic simulators. Procedia Computer Science, 19, 882-887.

[33] Koupal, J., Cumberworth, M., Michaels, H., Beardsley, M., & Brzezinski, D. (2003). Design and implementation of MOVES: EPA's new generation mobile source emission model. Ann Arbor, 1001(48), 105.

[34] Liu, J., Kockelman, K., & Nichols, A. (2017). Anticipating the emissions impacts of smoother driving by connected and autonomous vehicles, using the MOVES model. In transportation research board 96th annual meeting.

[35] Koupal, J., Beardsley, M., Brzezinski, D., Warila, J., & Faler, W. (2010). US EPA's MOVES2010 vehicle emission model: overview and considerations for international application. Ann Arbor, MI: US Environmental Protection Agency, Office of Transportation and Air Quality. http://www. epa. gov/oms/models/moves/MOVES2010a/paper137-tap2010. pdf.

[36] Zhou, X., Tanvir, S., Lei, H., Taylor, J., Liu, B., Rouphail, N. M., & Frey, H. C. (2015). Integrating a simplified emission estimation model and mesoscopic dynamic traffic simulator to efficiently evaluate emission impacts of traffic management strategies. Transportation Research Part D: Transport and Environment, 37, 123-136.

[37] Vallamsundar, S., & Lin, J. (2011). MOVES versus MOBILE: comparison of greenhouse gas and criterion pollutant emissions. Transportation research record, 2233(1), 27-35.

[38] Jimenez-Palacios, J. L. (1998). Understanding and quantifying motor vehicle emissions with vehicle specific power and TILDAS remote sensing. Massachusetts Institute of Technology.

[39] Jimenez, J. L., McClintock, P., McRae, G. J., Nelson, D. D., & Zahniser, M. S. (1999, April). Vehicle specific power: A useful parameter for remote sensing and

emission studies. In Ninth CRC On-Road Vehicle Emissions Workshop, San Diego, CA.

[40] Wang, H., & Fu, L. (2010). Developing a high-resolution vehicular emission inventory by integrating an emission model and a traffic model: Part 1—Modeling fuel consumption and emissions based on speed and vehicle-specific power. Journal of the Air & Waste Management Association, 60(12), 1463-1470.

[41] Song, G., Yu, L., & Tu, Z. (2012). Distribution characteristics of vehicle-specific power on urban restricted-access roadways. Journal of transportation engineering, 138(2), 202-209.

[42] Song, G., Zhou, X., & Yu, L. (2015). Delay correction model for estimating bus emissions at signalized intersections based on vehicle specific power distributions. Science of the Total Environment, 514, 108-118.

[43] Yao, Z., H. Wei, and H. Liu. Statistical Vehicle Specific Power Profiling for Urban Freeways. Procedia – Social and Behavioral Sciences, 2013. Volume 96: 2927-2938.

[44] De Haan, P., & Keller, M. (2004). Modelling fuel consumption and pollutant emissions based on real-world driving patterns: the HBEFA approach. International journal of environment and pollution, 22(3), 240-258.

[45] Kraschl-Hirschmann, K., Luz, R., & Fellendorf, M. (2016). Using trajectory data to estimate energy consumption for routing purposes. In Strategies for Sustainable Mobilities (pp. 159-174). Routledge.

[46] Lu, H., Song, G., Zhao, Q., Wang, J., He, W., & Yu, L. (2018). An investigation of the uncertainty of handbook of emission factors for road transport (HBEFA) for estimating greenhouse gas emissions: A case study in Beijing. Transportation Research Record, 2672(25), 79-88.

[47] Negrenti, E., Carrese, S., Beltran, B., Parenti, A., Giovannini, F., & Lapolla, V. (2007). Modelling vehicles kinematics and parking processes relevance on pollutant emissions in the city of Florence. Air Pollution XV., 1, 341-350.

[48] Bebkiewicz, K., Chłopek, Z., Lasocki, J., Szczepański, K., & Zimakowska-Laskowska, M. (2019). Inventory of pollutant emission from motor vehicles in Poland using the COPERT 5 software. Combustion Engines, 58.

[49] Zu, Y., Liu, C., Dai, R., Sharma, A., & Dong, J. (2018). Real-time energy-efficient traffic control via convex optimization. Transportation Research Part C: Emerging Technologies, 92, 119-136.

[50] Li, F., Zhuang, J., Cheng, X., Li, M., Wang, J., & Yan, Z. (2019). Investigation and prediction of heavy-duty diesel passenger bus emissions in Hainan using a COPERT model. Atmosphere, 10(3), 106.

[51] Hammarström, U., & Yahya, M. R. (2013). An analysis of the Swedish HGV fleet with driving resistance in focus: vehicle parameters as a basis for HBEFA emission factor estimation. Statens väg-och transportforskningsinstitut.

[52] Shankar, R., & Marco, J. (2013). Method for estimating the energy consumption of electric vehicles and plug-in hybrid electric vehicles under real-world driving conditions. IET intelligent transport systems, 7(1), 138-150.

[53] Greenwood, I. D., Dunn, R. C., & Raine, R. R. (2007). Estimating the effects of traffic congestion on fuel consumption and vehicle emissions based on acceleration noise. Journal of Transportation Engineering, 133(2), 96-104.

[54] Smit, R., Brown, A. L., & Chan, Y. C. (2008). Do air pollution emissions and fuel consumption models for roadways include the effects of congestion in the roadway traffic flow?. Environmental Modelling & Software, 23(10-11), 1262-1270.

[55] Tsanakas, N., Ekström, J., & Olstam, J. (2017). Reduction of errors when estimating emissions based on static traffic model outputs. Transportation research procedia, 22, 440-449.

[56] Tsanakas, N., Ekström, J., & Olstam, J. (2020). Estimating emissions from static traffic models: problems and solutions. Journal of Advanced Transportation, 2020.

[57] Skiba, U., Jones, S. K., Dragosits, U., Drewer, J., Fowler, D., Rees, R. M., ... & Manning, A. J. (2012). UK emissions of the greenhouse gas nitrous oxide. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1593), 1175-1185.

[58] Baggott, S. L., Lelland, A., Passant, N. P., & Watterson, J. (2002). Review of carbon emission factors in the UK Greenhouse gas inventory. change, 1990(8.73), 8-49.

[59] Jimenez, J. L., McClintock, P., McRae, G. J., Nelson, D. D., & Zahniser, M. S. (1999, April). Vehicle specific power: A useful parameter for remote sensing and emission studies. In Ninth CRC On-Road Vehicle Emissions Workshop, San Diego, CA.

[60] Song, G., Yu, L., & Tu, Z. (2012). Distribution characteristics of vehicle-specific power on urban restricted-access roadways. Journal of transportation engineering, 138(2), 202-209.

[61] Song, G., Zhou, X., & Yu, L. (2015). Delay correction model for estimating bus emissions at signalized intersections based on vehicle specific power distributions. Science of the Total Environment, 514, 108-118.

[62] Zhai, Z., Song, G., Lu, H., He, W., & Yu, L. (2016). A Validation of Temporal and Spatial Consistency of Facility-and Speed-Specific VSP Distribution for Emissions Estimation: A Case Study in Beijing (No. 16-3341).

[63] Lai, J., Yu, L., Song, G., Guo, P., & Chen, X. (2013). Development of city-specific driving cycles for transit buses based on VSP distributions: Case of Beijing. Journal of Transportation Engineering, 139(7), 749-757.

[64] Li, M and L. Yu. Development of Emission Factors for an Urban Road Network Based on Speed Distributions. Journal of Transportation Engineering, 2016. 142(9).

[65] Yao, Z., Wei, H., Liu, H., & Li, Z. (2013). Statistical vehicle specific power profiling for urban freeways. Procedia-Social and Behavioral Sciences, 96, 2927-2938.

[66] Quaassdorff, C., Borge, R., Pérez, J., Lumbreras, J., de la Paz, D., & de Andrés, J. M. (2016). Microscale traffic simulation and emission estimation in a heavily trafficked roundabout in Madrid (Spain). Science of the Total Environment, 566, 416-427.

[67] Brundell-Freij, K., & Ericsson, E. (2005). Influence of street characteristics, driver category and car performance on urban driving patterns. Transportation Research Part D: Transport and Environment, 10(3), 213-229.

[68] Zhang, H., Sun, J., & Tian, Y. (2020). The impact of socio-demographic characteristics and driving behaviors on fuel efficiency. Transportation Research Part D: Transport and Environment, 88, 102565.

[69] Walnum, H. J., & Simonsen, M. (2015). Does driving behavior matter? An analysis of fuel consumption data from heavy-duty trucks. Transportation research part D: transport and environment, 36, 107-120.

[70] Guan, T., & Frey, C. W. (2012, September). Fuel efficiency driver assistance system for manufacturer independent solutions. In 2012 15th International IEEE Conference on Intelligent Transportation Systems (pp. 212-217). IEEE.

[71] Jimenez, D., S. Hernandez, and J. Fraile-Ardanuy. Modelling the Effect of Driving Events on Electrical Vehicle Energy Consumption Using Inertial Sensors in Smartphones. Energies, 2018. 11(2):412.

[72] Meseguer, J. E., Calafate, C. T., Cano, J. C., & Manzoni, P. (2015, January). Assessing the impact of driving behavior on instantaneous fuel consumption. In 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC) (pp. 443-448). IEEE.

[73] Xu, Z., Wei, T., Easa, S., Zhao, X., & Qu, X. (2018). Modeling relationship between truck fuel consumption and driving behavior using data from internet of vehicles. Computer-Aided Civil and Infrastructure Engineering, 33(3), 209-219.

[74] Chen, C., Zhao, X., Yao, Y., Zhang, Y., Rong, J., & Liu, X. (2018). Driver's eco-driving behavior evaluation modeling based on driving events. Journal of Advanced Transportation, 2018.

[75] Magaña, V. C., Pañeda, X. G., Garcia, R., Paiva, S., & Pozueco, L. (2021). Beside and behind the wheel: factors that influence driving stress and driving behavior. Sustainability, 13(9), 4775.

[76] Paefgen, J., Kehr, F., Zhai, Y., & Michahelles, F. (2012, December). Driving behavior analysis with smartphones: insights from a controlled field study. In Proceedings of the 11th International Conference on mobile and ubiquitous multimedia (pp. 1-8).

[77] Van Der Horst, R., & De Ridder, S. (2007). Influence of roadside infrastructure on driving behavior: driving simulator study. Transportation Research Record, 2018(1), 36-44.

[78] Ranjitkar, P., Nakatsuji, T., Azuta, Y., & Gurusinghe, G. S. (2003). Stability analysis based on instantaneous driving behavior using car-following data. Transportation Research Record, 1852(1), 140-151.

[79] Dogan, E., Steg, L., & Delhomme, P. (2011). The influence of multiple goals on driving behavior: The case of safety, time saving, and fuel saving. Accident Analysis & Prevention, 43(5), 1635-1643.

[80] Hu, L., Bao, X., Lin, M., Yu, C., & Wang, F. (2021). Research on risky driving behavior evaluation model based on CIDAS real data. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 235(8), 2176-2187.

[81] Paleti, R., Eluru, N., & Bhat, C. R. (2010). Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. Accident Analysis & Prevention, 42(6), 1839-1854.

[82] Colonna, P., Intini, P., Berloco, N., & Ranieri, V. (2016). The influence of memory on driving behavior: How route familiarity is related to speed choice. An on-road study. Safety science, 82, 456-468.

[83] Yan, X., & Wu, J. (2014). Effectiveness of variable message signs on driving behavior based on a driving simulation experiment. Discrete dynamics in nature and society, 2014.

[84] Varella, R. A., Faria, M. V., Mendoza-Villafuerte, P., Baptista, P. C., Sousa, L., & Duarte, G. O. (2019). Assessing the influence of boundary conditions, driving behavior and data analysis methods on real driving CO2 and NOx emissions. Science of the total environment, 658, 879-894.

[85] Li, X., Yan, X., & Wong, S. C. (2015). Effects of fog, driver experience and gender on driving behavior on S-curved road segments. Accident Analysis & Prevention, 77, 91-104.

[86] Boggio-Marzet, A., Monzon, A., Rodriguez-Alloza, A. M., & Wang, Y. (2021). Combined influence of traffic conditions, driving behavior, and type of road on fuel consumption. Real driving data from Madrid Area. International Journal of Sustainable Transportation, 1-13.

[87] Zhao, Y., Yamamoto, T., & Morikawa, T. (2018). An analysis on older driver's driving behavior by GPS tracking data: Road selection, left/right turn, and driving speed. Journal of traffic and transportation engineering (English edition), 5(1), 56-65.

[88] Arafa, A., El-Setouhy, M., & Hirshon, J. M. (2019). Driving behavior and road traffic crashes among professional and nonprofessional drivers in South Egypt. International journal of injury control and safety promotion, 26(4), 372-378.

[89] Classen, S., Wang, Y., Winter, S. M., Velozo, C. A., Lanford, D. N., & Bédard, M. (2013). Concurrent criterion validity of the safe driving behavior measure: A predictor of on-road driving outcomes. American journal of occupational therapy, 67(1), 108-116.

[90] Faria, M. V., Duarte, G. O., Varella, R. A., Farias, T. L., & Baptista, P. C. (2019). How do road grade, road type and driving aggressiveness impact vehicle fuel consumption? Assessing potential fuel savings in Lisbon, Portugal. Transportation Research Part D: Transport and Environment, 72, 148-161.

[91] Li, Y. M. (2007). Road traffic casualties and risky driving behavior in Hualien County, 2001–2005. Tzu Chi Medical Journal, 19(3), 152-158.

[92] Kim, H., Yoon, D., Lee, S. J., Kim, W., & Park, C. H. (2018, January). A study on the cognitive workload characteristics according to the driving behavior in the urban road. In 2018 International Conference on Electronics, Information, and Communication (ICEIC) (pp. 1-4). IEEE.

[93] Lárusdóttir, E. B., & Ulfarsson, G. F. (2015). Effect of driving behavior and vehicle characteristics on energy consumption of road vehicles running on alternative energy sources. International Journal of Sustainable Transportation, 9(8), 592-601.

[94] Haque, F., & Abas, M. A. (2018). Review of Driving Behavior Towards Fuel Consumption and Road Safety. Jurnal Mekanikal.

[95] Park, J., Lim, J., Joo, S., & Lee, S. (2015). A Study on the Compensation of the Difference of Driving Behavior between the Driving Vehicle and Driving Simulator. International Journal of Highway Engineering, 17(2), 107-122.

[96] Park, J., Lim, J., Joo, S., & Lee, S. (2015). A Study on the Compensation of the Difference of Driving Behavior between the Driving Vehicle and Driving Simulator. International Journal of Highway Engineering, 17(2), 107-122.

[97] Hawirko, J. D., & Checkel, M. D. (2002). Real-time, on-road measurement of driving behavior, engine parameters and exhaust emissions (No. 2002-01-1714). SAE Technical Paper.

[98] Hui, L., Yong, L., Shibo, Z., Yanfei, S., Xiaohan, L., Jian, L., & Xia, L. (2011). RESEARCH BETWEEN AGGRESSIVE DRIVING BEHAVIOR AND TYPE A BEHAVIOR 2. In 3rd International Conference on Road Safety and SimulationPurdue UniversityTransportation Research Board.

[99] Tax, N., Teinemaa, I., & van Zelst, S. J. (2018). An interdisciplinary comparison of sequence modeling methods for next-element prediction. arXiv preprint arXiv:1811.00062.

[100] Chen, Z., & Droppo, J. (2018, April). Sequence modeling in unsupervised single-channel overlapped speech recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4809-4813). IEEE.

[101] Ren, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2019, May). Almost unsupervised text to speech and automatic speech recognition. In International Conference on Machine Learning (pp. 5410-5419). PMLR.

[102] Liu, X., & Zhou, M. (2011, August). Sentence-level sentiment analysis via sequence modeling. In International Conference on Applied Informatics and Communication (pp. 337-343). Springer, Berlin, Heidelberg.

[103] Robertson, N., & Reid, I. (2006). A general method for human activity recognition in video. Computer Vision and Image Understanding, 104(2-3), 232-248.

[104] Liu, H., Hartmann, Y., & Schultz, T. (2021, August). Motion Units: Generalized Sequence Modeling of Human Activities for Sensor-Based Activity Recognition. In 2021 29th European Signal Processing Conference (EUSIPCO) (pp. 1506-1510). IEEE.

[105] Zhang, L., Wu, X., & Luo, D. (2015, July). Human activity recognition with HMM-DNN model. In 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC) (pp. 192-197). IEEE.

[106] Ycart, A., & Benetos, E. (2017, October). A study on LSTM networks for polyphonic music sequence modelling. ISMIR.

[107] Jagannatha, A. N., & Yu, H. (2016, November). Structured prediction models for RNN based sequence labeling in clinical text. In Proceedings of the conference

on empirical methods in natural language processing. conference on empirical methods in natural language processing (Vol. 2016, p. 856). NIH Public Access.

[108] Salaün, A., Petetin, Y., & Desbouvries, F. (2019, December). Comparing the modeling powers of RNN and HMM. In 2019 18th IEEE International Conference on Machine Learning And Applications (ICMLA) (pp. 1496-1499). IEEE.

[109] Kamal, S., Jalal, A., & Kim, D. (2016). Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM. Journal of Electrical engineering and technology, 11(6), 1857-1862.

[110] Just, A., Bernier, O., & Marcel, S. (2004). HMM and IOHMM for the recognition of mono-and bi-manual 3D hand gestures (No. REP_WORK). IDIAP.

[111] Lee, Y. S., & Cho, S. B. (2011, May). Activity recognition using hierarchical hidden markov models on a smartphone with 3D accelerometer. In International conference on hybrid artificial intelligence systems (pp. 460-467). Springer, Berlin, Heidelberg.

[112] San-Segundo, R., Montero, J. M., Moreno-Pimentel, J., & Pardo, J. M. (2016). HMM adaptation for improving a human activity recognition system. Algorithms, 9(3), 60.

[113] Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks (TOSN), 6(2), 1-27.

[114] Bacciu, D., Micheli, A., & Sperduti, A. (2013). An input–output hidden Markov model for tree transductions. Neurocomputing, 112, 34-46.

[115] Yin, M., Sheehan, M., Feygin, S., Paiement, J. F., & Pozdnoukhov, A. (2017). A generative model of urban activities from cellular data. IEEE Transactions on Intelligent Transportation Systems, 19(6), 1682-1696.

[116] Lin, Z., Yin, M., Feygin, S., Sheehan, M., Paiement, J. F., & Pozdnoukhov, A. (2017). Deep generative models of urban mobility. IEEE Transactions on Intelligent Transportation Systems.

[117] Koller, O., Zargaran, S., & Ney, H. (2017). Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4297-4305).

[118] Cunningham, P. (2000). Overfitting and diversity in classification ensembles based on feature selection. Trinity College Dublin, Department of Computer Science.

[119] Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research, 3(Mar), 1371-1382.

[120] Ying, X. (2019, February). An overview of overfitting and its solutions. In Journal of Physics: Conference Series (Vol. 1168, No. 2, p. 022022). IOP Publishing.

[121] Liu, B. and H.C. Frey. Development and Evaluation of a Simplified Version of MOVES for Coupling with a Traffic Simulation Model. Proceedings of Air and Wastes Management Association's Annual Conference and Exhibition, 2012.

[122] Guensler, R. L., Liu, H., Xu, X., Xu, Y., & Rodgers, M. O. (2016). MOVES-Matrix: setup, implementation, and application (No. 16-6362).

[123] Liu, H., Guensler, R., Lu, H., Xu, Y., Xu, X., & Rodgers, M. O. (2019). MOVES-Matrix for high-performance on-road energy and running emission rate modeling applications. Journal of the Air & Waste Management Association, 69(12), 1415-1428.

[124] Pan, Y., Darzi, A., Kabiri, A., Zhao, G., Luo, W., Xiong, C., & Zhang, L. (2020). Quantifying human mobility behaviour changes during the COVID-19 outbreak in the United States. Scientific Reports, 10(1), 1-9.

[125] Rtree 0.9.7 documentation >> Rtree: Spatial indexing for Python. https://rtree.readthedocs.io/en/latest/

[126] 7 Types of Generative Models for Your Next Machine Learning Project. https://analyticsindiamag.com/7-types-of-generative-models-for-your-next-machine-learning-project/

[127] https://en.wikipedia.org/wiki/Long_short-term_memory

[128] Wang, J., Chen, R., & He, Z. (2019). Traffic speed prediction for urban transportation network: A path based deep learning approach. *Transportation Research Part C: Emerging Technologies*, *100*, 372-385.

[129] Roy, K. C., Hasan, S., Culotta, A., & Eluru, N. (2021). Predicting traffic demand during hurricane evacuation using real-time data from transportation systems and social media. *Transportation research part C: emerging technologies*, *131*, 103339.

[130] https://www.commuterconnections.org/incentrip-app/