VEHICULAR MOBILITY MODELING ON A LARGE SCALE: AN APPROACH TO COMBINE STATIONARY SENSING AND MOBILE SENSING

BY YU YANG

A thesis submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Computer Science

Written under the direction of

Desheng Zhang

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2017

ABSTRACT OF THE THESIS

Vehicular Mobility Modeling on A Large Scale: An Approach to Combine Stationary Sensing and Mobile Sensing

by Yu Yang

Thesis Director: Desheng Zhang

Real-time mobility is important for many real-world applications, e.g., transportation, urban planning given different level administrative jurisdiction. However, most of the existing work focuses at small scale with limited data samples (e.g. region or city level with samples over all the taxis). Recently, with upgrades of transportation infrastructures, we have new opportunities to capture real-time mobility at larger scale. With emerging of multiple sensors e.g., traffic cameras, toll systems, traffic loop sensors and GPS equipped vehicle fleets, we have unprecedented opportunities to capture real-time state-level mobility

In this dissertation, we analyze the challenges and opportunities for mobility modeling on a large scale and design a mobility prediction model called StateFlow to capture real-time intra and inter city vehicular mobility. In particular, StateFlow is based on (i) a stationary sensor network capturing aggregated mobility at the highway toll station level; (ii) a mobile sensor network capturing individual mobility at the local grid level. The key novelty of StateFlow is in its two-level structure where we investigate the correlation between highway station-level mobility and grid-level mobility for fine-grained mobility modeling. With multiple models built upon the two-level structure, we address

a key intellectual challenge of sensing heterogeneity in terms of spatiotemporal granularity. In station level, we use Bayesian Inference to predict the exit stations based on vehicle historical travel records including when and where they enter the highways and use K Nearest Neighbors to predict the travel time between two stations considering both real-time including real-time traffic condition and weather condition and historical information including personal driving habits. In grid leve, we build a random-based model to predict vehicle final destinations based on personlized features and crowd features. Based on these two level prediction, we can track individual vehicles from entering the highways to arriving the final destionations. More importantly, we implement StateFlow in Guangdong Province, China with (i) an electric toll collection system with tracking devices at 1439 highway entrances and exits in Guangdong, functioning as a stationary sensing part of StateFlow; (ii) a vehicle fleet system consisting of both commercial logistics and private vehicles in Guangdong with in total 114 thousand GPS-equipped vehicles, functioning as a mobile sensing part of StateFlow. We compared StateFlow with the two benchmark mobility models based on our data, and the experimental results show that StateFlow outperforms others in terms of accuracy.

Acknowledgements

This dissertation was only possible through the contribution, encouragement, advice, and support from a large number of people and their previous work. Thanks for their efforts of paving the way for my work. Especially, I wish to thank my advisor, Professor Desheng Zhang, and the members of my committee, Professor Jingjin Yu, Professor Zheng Zhang, Professor Richard Martin for their generosity in taking the time and energy to guide and review my work. And also, I wish to thank all the members at Department of Computer Science for their encouragements and helps throughout my study period. And I also wish to thank my family for supporting me to take my every step following my heart. They give me life and teach me how to live. This dissertation is not possible achieved without their supports

Dedication

To my father Haizhou Hu and my mother Daolan Yang.

Table of Contents

Chapter 1

Introduction

1.1 Thesis

The thesis of the dissertation state that:

By combining stationary sensing and mobile sensing, we can predict vehicular inter and intra city mobility at individual level in real time.

1.2 Background

Cities or regions within the same states typically share similar features and are well-connected due to their similar state-level administrative jurisdiction. For example, road networks between regions in the same state are typically more developed than regions across different states [34]; the economy between cities or regions within the same states is typically higher than cities across states. As a result, it is essential to understand in-state real-time mobility in terms of travel time or vehicle volumes for many real-world applications, e.g., transportation, region planning and in-state business development [7].

To date, almost all existing work on real-time mobility modeling focuses on mobility patterns at city scale instead of state scale, e.g., mobility modeling based on data from taxis, buses, smartcards, cellphones, and social networks [13] [7] [26] [27]. All these models are based on particular systems, e.g., transportation infrastructure, telecommunication, finance, or online social networks, which are typically at city level [1]. Thus, little work, if any, has been proposed to study real-time mobility between cities within the same states. We argue that city-level models cannot be directly applied to learn state-level mobility because city-level infrastructures cannot be scaled to the state level.

Recently, with updates of state infrastructures, e.g., traffic cameras, toll systems,

traffic loop sensors and GPS equipped vehicle fleets, we have unprecedented opportunities to capture real-time state-level mobility [32, 44]. We divide these state-level infrastructures into two categories from sensing perspectives with complementary features: (i) stationary sensing systems where we can track all vehicles passing fixed locations; (ii) mobile sensing systems where we can track a subset of vehicles at grid level with detailed GPS devices. In our setting, the stationary sensing systems can capture all vehicles traveling between cities in the same state using highways, when they pass highway entrances and exits. But it cannot capture any vehicles when not using highways or traveling on these highways. In contrast, the mobile sensing systems can track all participating vehicles in real time with onboard GPS devices regardless of routes they are taking. Thus, the stationary sensing can cover all vehicles but only at fixed locations in the station level; whereas the mobile sensing can cover all locations in grid level but only for limited vehicles. Therefore, our core idea is to design a hybrid sensing system utilizing complementary features of stationary and mobile sensing to address their individual limitation.

In this paper, we motivate, design, implement and evaluate a two-level mobility model called StateFlow based on a hybrid sensing system to capture real-time state-level mobility. The key novelty of StateFlow is in its two-level structure where we investigate the correlation between station-level mobility and grid-level mobility for fine-grained mobility modeling. With multiple models built upon two-level structure of the state flow, we address a key intellectual challenge of sensing heterogeneity in terms of spatiotemporal granularity. More importantly, we implement StateFlow in Guangdong Province, China with an electric toll collection system and a commercial logistics fleet. Even though some tracking systems using toll data or GPS devices have been proposed to model mobility within the city level [32] [44] [1], we believe the key difference between StateFlow and them is in state-level real-time mobility modeling with both toll data and GPS data, along with a real-world implementation and an independent application evaluation.

The rest of the paper is organized as follows. Section 2 presents the motivation for StateFlow. Section 3 provides the overview of StateFlow system. Section 4 gives

the physical layer design of StateFlow system which collects real-time sensing data for mobility modeling. Section 5 and section 6 elaborate on the design and the evaluation of the two-level mobility model proposed in StateFlow to predict the traffic flow. Section 7 presents a real-world application on top of the StateFlow mobility modeling, followed by the related work and discussion in Section 8 and Section 9. Finally, Section 10 concludes the paper.

1.3 Motivation

To motivate our design, we show how the two sensing components, i.e., stationary sensing and mobile sensing, complement each other to capture state-scale vehicular mobility. In particular, we utilize an electric toll collection system (ETC) and a mobile vehicle fleet in the Chinese province Guangdong (similar to a state in U.S.) as concrete implementations of the stationary and mobile sensing components in our StateFlow system. The details of these two systems and their data will be introduced in Section 3.

1.3.1 ETC for Stationary Sensing

In this subsection, we study the vehicular mobility captured by the ETC system at 1439 toll collection stations in the Guangdong highway network. Note that the ETC system captures all vehicles paying with cash or electric toll devices on the highway network. We differentiate each vehicle with a toll collection serial number, instead of plate IDs, to protect the privacy of the drivers, and more discussion about privacy in Section 9. Figure 1.1 gives the number of vehicles captured by the ETC system during June 2016. We found that the ETC system captures around 1 million vehicles every day on average. However, we also found that most of these vehicles travel regularly. In particular, we study the unknown vehicles in the ETC system, which are defined as the vehicles that appear in the ETC system without being captured historically. We found the number of the unknown vehicles decreases significantly with the accumulation of historical ETC data. This phenomenon motivates us to build a predictive model on the individual

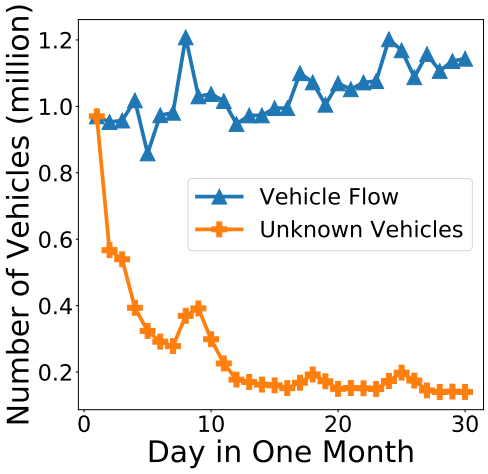

Figure 1.1: ETC Tracking

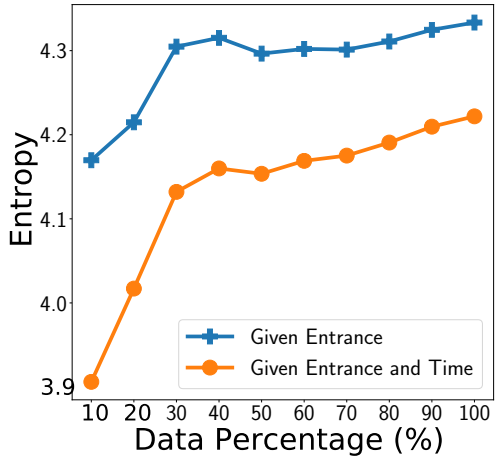

Figure 1.2: Travel Pattern

vehicle level to predict a vehicle's exit location and time based on its entrance locations and time in the highway system.

To further explore this motivation, we investigate the entropy distributions given different data scale, i.e., length of data history. Two entropies, the entropy of destination given entrance and the entropy of destination given entrance and entering time are computed. Figure 1.2 shows that entropies given both entrance and entering time or only given entrance are both slightly higher than 4. For instance, entropy given entrance and entering time is less than 4.3, meaning that there are around $2^{4.3}$ possible exits compared to total 1439 toll stations in the whole network. This finding also implies that we can use a sparse Bayesian model to efficiently model the highway traffic mobility because the average out-degree of a node in the Bayesian Network is no more than 4.3.

1.3.2 Fleets for Mobile Sensing

In this section, we validate the opportunities to utilize fleets traveling regularly within a state as a mobile sensing component to capture and predict the state-scale mobility. We divide Guangdong province with 119 regions, which are districts, counties, or cities based on the administrative level 6 defined by OpenStreetMap [28]. And in a lower granularity, we divide Guangdong into grids of size $5km * 5km$. In particular, we use a vehicular fleet with more than 14 thousand vehicles to validate the potential of using

them to model mobility at grid level within a state. In Figure 1.3, we show the region coverage of fleets. We found that with the increase in the number of vehicles used in this fleet, we can cover more regions, e.g., if 50% of vehicles are used, we can travel 85% of all 119 regions in Guangdong. For the grid level in Figure 1.4, we can cover more that 50% grids to track the mobility in the lower level. It indicates that this vehicle fleet has the potential to cover the mobility outside highways.

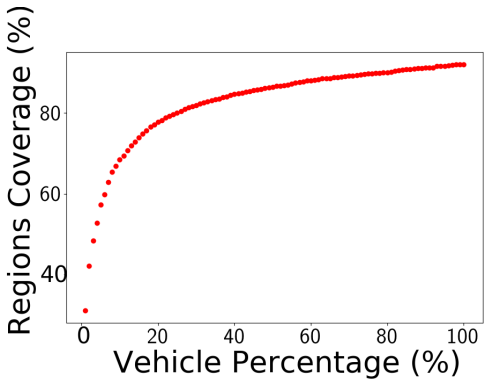

Figure 1.3: Region Coverage

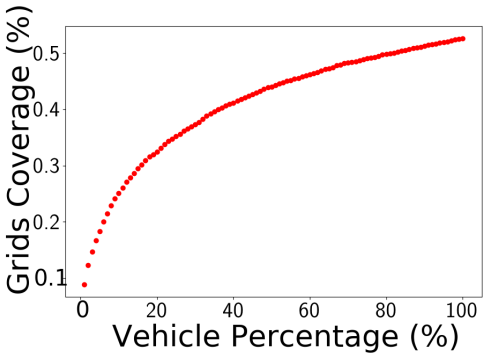

Figure 1.4: Grids Coverage

1.3.3 Summary

Based on the above two subsections, we found that (i) the stationary ETC system can be used to predict vehicle mobility at station level between different regions within a state with high accuracy; (ii) a mobile vehicle fleet can provide high coverage outside the highways with only 14 thousand vehicles. These findings motivate us to combine the two sensing components together from a hybrid perspective in our StateFlow system to predict and model the vehicle mobility at state scale, instead of utilizing them separately.

1.4 State of The Art

Our study relies on the data collected from transportation sensor network and models the mobility at state level. Therefore, mobility modeling by utilizing empirical data and vehicular crowdsourcing system are both related to our work.

1.4.1 Mobility Modeling

Mobility modeling through single data source: Historically, mobility modeling relies on either trajectory data, i.e., time-stamped positions, or statistics data collected from urban infrastructure, e.g., inductive loop, RFID-based toll stations. At city level, trajectories from different data sources have been utilized, e.g., taxi cab trajectory [7], bike-sharing system transactions [37], smart card transactions [39], cellular phone records [13, 17, 6, 35], and social network data [26, 27]. Traffic count and speed information collected from static sensors are also used to recover the traffic flows (real-time traffic demand between given origin-destination (O-D) pairs) [24]. Traffic flow can be extracted from wireless access points at different locations, by WiFi connection [19] or Bluetooth connections [22]. At national-wide scale, mobility modeling is also conducted using private and commercial vehicle trajectories [41]. The problem with single source data is either low penetration issue with fine-grained trajectory data or data sparsity and static nature with infrastructure sensors or coarse-grained O-D information.

Mobility modeling through data fusion: Due to the limitations of single data source, data fusion among different datasets gains significant attention recently: the fine-grained trajectory data to provide high resolution and static sensing data to cover large population. Transit trajectory and cellular data are fused together to model human mobility in metropolitan regions [42] [40]. Trajectories from sample traffic and total traffic count from sample road segments are used to recover the total traffic flow in the city [1, 23]. However, the mobility modeling using data fusion has been limited within city scale because these models are based on particular systems which are typically at the city level.

Summary: Our work combines the dataset from both a stationary network capturing aggregated mobility and a mobile network capturing individual mobility, and thus avoids the limitations of single data source modeling. Our model is built upon state scale, demanding different methodology from current data fusion method since the city-level infrastructures cannot be simply scaled to the state level.

1.4.2 Vehicular Applications

Due to the rich mobile sensors, various vehicular data-driven systems are proposed for urban sensing and recommendation purposes: vehicles are tracked in real-time [15, 43, 45, 30, 31] to infer map changes including road segment [33] and traffic light regulators (e.g. traffic light) [16], to monitor traffic speed [32, 44], volume [1] and pollution [36, 14, 3], to estimate parking status [25], to recommend time-efficient [9, 38, 12] and fuel-efficient [8] driving route, to predict passenger demand for taxi drivers [11] or recommend optimal pickup locations for passengers [10] [21], to detect the taxi anomaly [29], to estimate arrival time of bus [45], and taxi trip duration and fares [2].

Compared to the above systems, our paper presents a vehicular sensing system for a different purpose: we model human mobility pattern, i.e., sensing and predicting traffic flow at the state level using both static and mobile sensors.

1.5 Contributions

The key contributions of this thesis are as follows.

- To our knowledge, we conduct the first systematic investigation on real-time mobility at state scale. Our work is based on 7.8 million vehicles at the highway level at entrance and exit locations and 114 thousand vehicles at the grid level at GPS locations.

- We present a two-layer mobility model called StateFlow to capture vehicular mobility at state scale. In StateFlow, we utilize (i) a stationary sensor network capturing aggregated mobility at the highway entrance/exit station level, and (ii) a mobile sensor network capturing individual mobility at the grid level with GPS devices.

- At the station level, StateFlow utilizes a Bayesian model to infer exit locations and exit time of vehicles based on their entrance locations and time. At the grid level, StateFlow maps vehicles exiting from the highway to specific grids based on mobility patterns learned from the mobile sensing components tracking individual

vehicles with GPS devices. As a result, we use this two-level structure to address sensing heterogeneity in terms of spatiotemporal granularity.

- More importantly, we implement the StateFlow in Guangdong Province, China with (i) an electric toll collection system functioning as a stationary sensing part of StateFlow, which has tracking devices at 1439 highway entrances and exits in Guangdong and captures around 1 million vehicles per day based on toll records; (ii) a vehicle fleet system including 14-thousand commercial logistics vehicles and 100-thousand private vehicles functioning as a mobile sensing part of StateFlow, which has 14-thousand GPS-equipped vehicles in Guangdong.

- We evaluate StateFlow through a one-month dataset collected from both the electric toll collection system and the vehicle fleet system in Guangdong. Compared with two benchmark methods, the proposed approach provides an improvement in terms of traffic flow prediction accuracy.

Chapter 2

Three-Layer Architecture

In the StateFlow system, we consider a set of mobility sensors, including stationary sensors (e.g., ETC stations), and mobile sensors (e.g., onboard GPS devices collecting vehicle trajectories), as a hybrid sensor network to track state-level vehicle mobility in real time. By an integration of multiple sensors, StateFlow provides mobility dynamics for both individual vehicles and traffic flow as a whole under fine-grained spatiotemporal resolutions to support real-world services. Figure 2.1 shows the overview of a three-layer architecture.

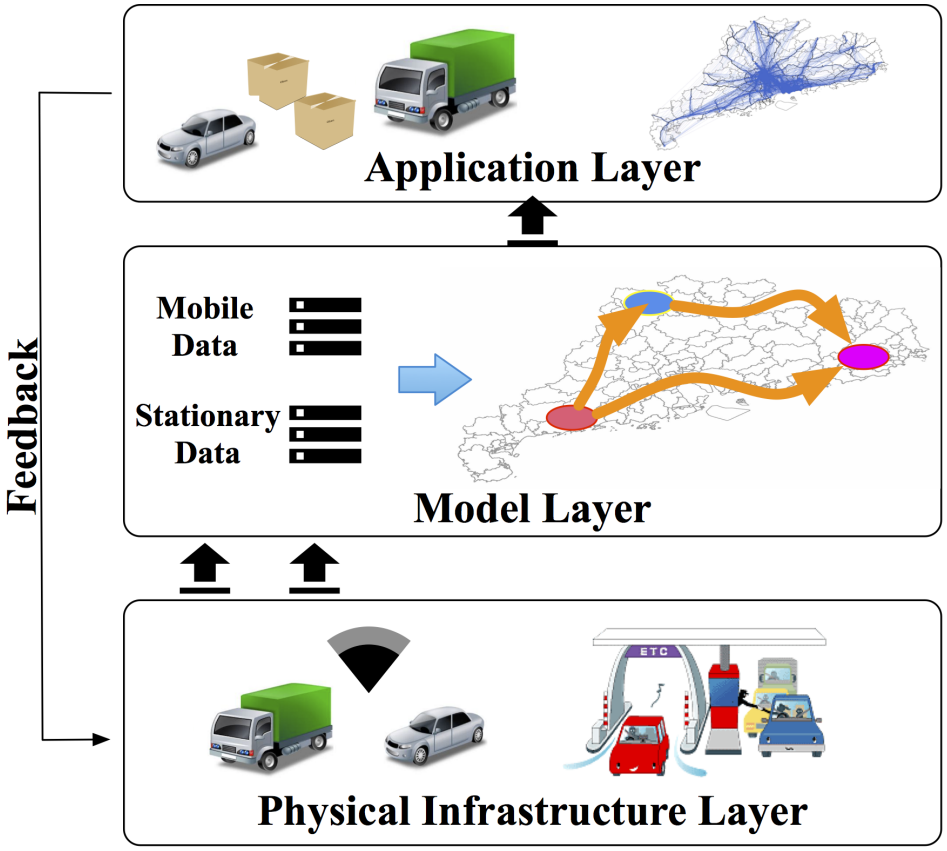

Figure 2.1: System Architecture

Infrastructure Layer: At the bottom, the physical infrastructure layer provides real-time data feeds through both stationary sensors and mobile sensors. The stationary sensors capture the traffic flow as fixed point, e.g. an ETC system based on RFID or cameras collecting traffic information of entering/leaving the highway system but they may not be sufficient to reveal the fine-grained mobility. The mobile sensors, e.g. a smartphone navigation system, on the other hand, collect the trajectory of a vehicle, but may not cover all the traffic in the road network.

Model Layer: The model layer takes the multi-source data from the physical layer and models the mobility dynamics by using a two-level model. Stationary and mobile sensing data are fused through the model and complement each other. As a result, StateFlow performs fine-grained mobility model in state-level for both the individual and the traffic flow perspective.

Application Layer: The output of the model layer is digested by the application layer to provide real-time application services. By using the prediction of individual driver's mobility, we could build applications such as routing services using predicted destination to shorten individual driver's travel time. The traffic flow prediction, on the other hand, helps the transportation authority to do predictive control, e.g. proactive planning for potential traffic congestion.

Based on the three-layer architecture, we use the Guangdong province, one of the most populous and wealthy provinces in China with the total area of 179,800 km^2, as a testbed to implement our StateFlow System. The details of the implementation generate the roadmap for the rest part of the paper: The physical layer details are presented in Section 3. The model layer design and implementation are detailed in Section 4. To close the loop, we also design and evaluate a route suggestion system in Section 6 based on vehicle's predicted destination in the application layer to show the benefit of StateFlow.

Chapter 3

Sensing Infrastructure Layer

StateFlow aims to fuse stationary sensing data and mobile sensing data in real time. In this section, we provide the details of these two sensing components based on two real-world systems, an ETC system in Guangdong as a concrete stationary sensing component and a vehicle fleet as a concrete mobile sensing component.

3.1 Stationary Sensing Component

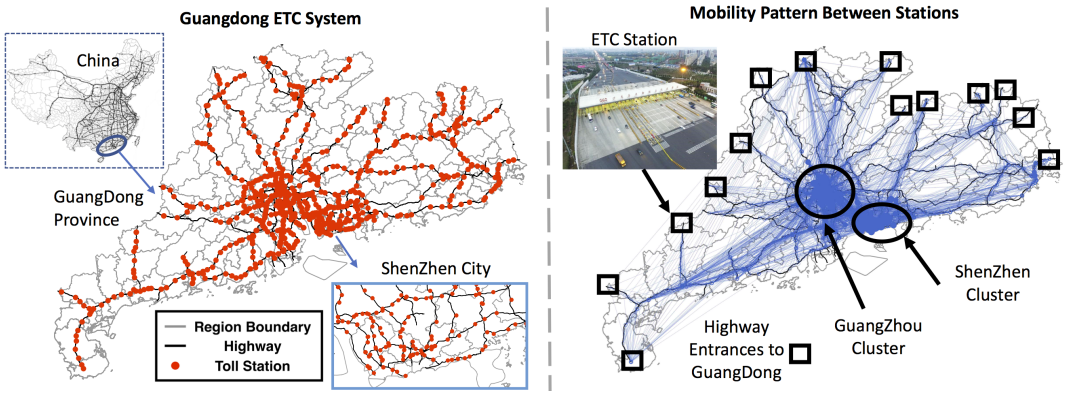

Figure 3.1: the ETC System in 119 regions partition and Mobility Patterns between ETC Stations

Figure 3.1 gives overview of the ETC system in Guangdong, a network of 121 highways and 1439 ETC toll stations. As shown in Figure 3.1, the ETC stations have very high density, even inside downtown areas of cities, e.g., in the Shenzhen city. We divide Guangdong into 119 regions based on the administrative levels 6 defined by OpenStreetMap [28] to study mobility at both the station level and grid level. Figure 3.1 also gives the mobility patterns at station levels. We found two big urban clusters in the provincial capital Guangzhou and the economic hub Shenzhen. Furthermore, we also found that there are strong mobility patterns between the highway entrances to

Guangdong and these two major clusters.

Realtime data from the ETC system contains all the enter/leave transactions from 1439 toll stations in Guangdong province, no matter whether a vehicle pays in cash and using an electric device. Each transaction includes the time, road id, toll station id, vehicle plate, vehicle type (e.g. Bus, Truck, Private vehicle), and so on.

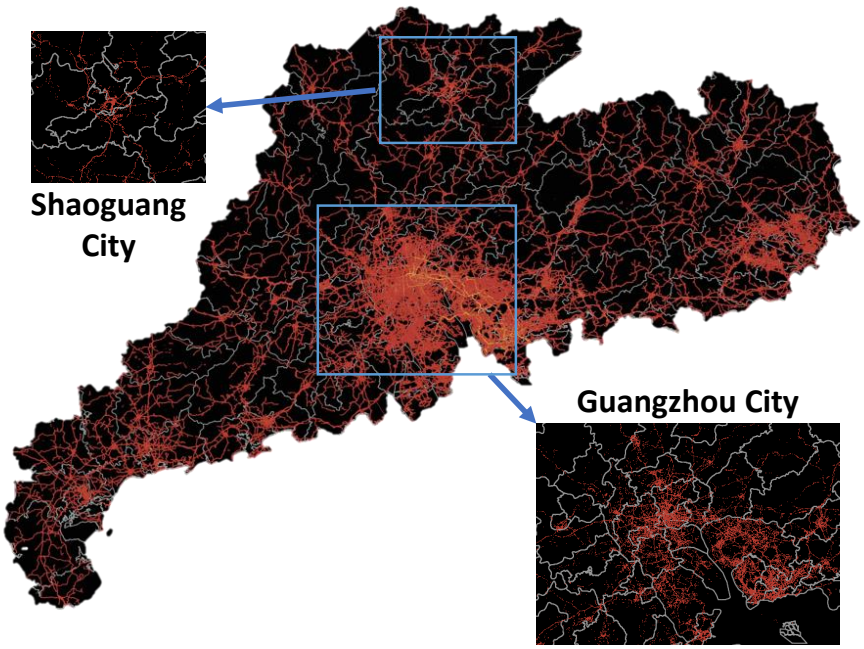

Figure 3.2: Fleet Visualization

3.2 Mobile Sensing Component

For the mobile sensing component, we acknowledge the differences in mobility patterns of commercial vehicles and private vehicles, and therefore we collect real-time trajectories from both types of vehicles: 14 thousand commercial vehicles and 100 thousand private vehicles. Figure 3.2 shows the aggregated traces of these vehicles.

- For the commercial vehicle network, we collaborate with a commercial logistics company with 45 thousand trucks, among which 14 thousand are operating in Guangdong. These vehicles upload their status including the GPS location and the travel speed to central management system every 15 seconds on average, which are redirected to our server in real time.

- For the private vehicle network, we collaborate with a navigation service provider, which serves 295 thousand vehicles on the national scale, among which 100 thousand vehicles are active in the Guangdong. We access a database of the real-time location of the vehicles using a navigation service, each of which uploads its real-time GPS location with a frequency of about 10 seconds to a cloud server through cellular networks.

3.3 Summary

One of the major strengths of StateFlow is the fusion of both the stationary sensing system and the mobile sensing system. With these sensors covered in the state, our physical infrastructure layer can achieve large-scale real-time vehicle mobility, which is unprecedented in terms of both quantity and quality.

Chapter 4

Mobility Modeling Layer: Design

In this section, we first introduce the key idea of our StateFlow system and then introduce our detailed design.

4.1 Key Idea

Given a regular vehicle without onboard GPS devices entering the highway system, we aim to predict its exiting highway station and final destination. The key idea of StateFlow is a two-level prediction structure. (i) When a vehicle is entering the highway, StateFlow utilizes the ETC system as a stationary sensing system to predict the exiting station and time of a vehicle based on its real-time data (e.g., including the entering station and time) and its historical data (e.g., historical entering/exiting stations and time), along with data from other vehicles; (ii) to utilize an urban fleet as a mobile sensing system to predict a vehicle's final destination when it is exiting the highway based on its real time data (e.g., including the entering station and time) and the historical data of sensing fleets. Figure 4.1 gives an example of StateFlow. We have three high-level regions (e.g., cities or large metro areas), and each of them has a toll station, i.e., three toll stations, A, B, and C. We divide all three regions into smaller grids to fine-grained modeling, e.g., we have six grids A1 to C2. Assuming a vehicle is traveling from A1 to C2. We introduce our two-level prediction as follows.

(i) Station-Level Prediction: When this vehicle starts from A1, StateFlow cannot capture this vehicle until it enters a toll station, e.g., A. When it enters A, StateFlow obtains its real-time data and historical travel patterns, and then StateFlow predicts which station it will exit, e.g., B or C, and when it will exit. This is the station-level mobility prediction of StateFlow for inter regional mobility on highways, which is based

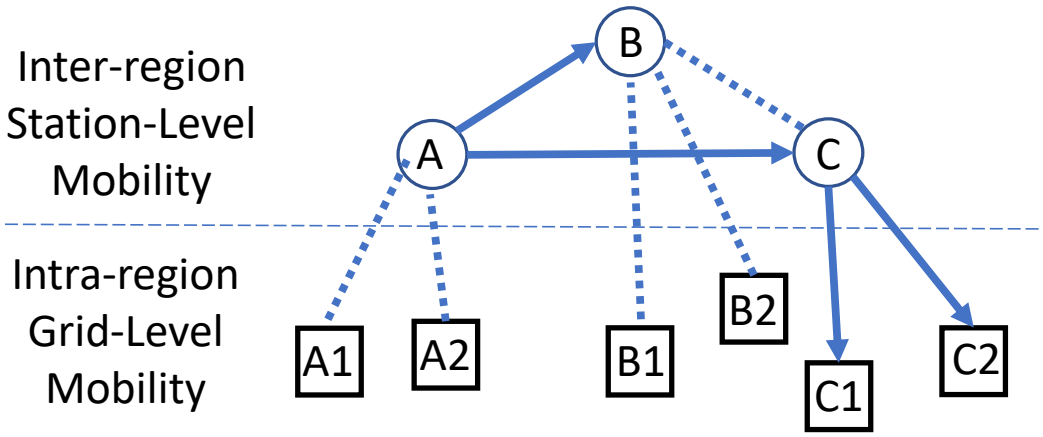

Figure 4.1: Key Idea for Two-Level Prediction

on the ETC system alone.

(ii) Grid-Level Prediction: When this vehicle is exiting at a station (e.g., C), StateFlow predicts its final destinations in a grid level (e.g., C1 and C2) based on which station it entered (e.g., A), how long it uses to travel from the entering and exiting stations, etc. With these real-time contextual information, if we have detailed historical GPS about this vehicle, we can easily predict its final destination. But since we aim to target more general vehicles without GPS devices, we utilize data from our mobile sensing fleet to predict the final destination of this vehicle. In particular, we utilize destinations of a few vehicles in our fleet that have similar features with this vehicle as its potential destinations on the grid level. This is based on both ETC and fleet systems.

Note that it is intuitive to just have a one level prediction where StateFlow predicts the final destination of this vehicle when it is entering the highway at a toll station, e.g., at Station A, we can predict it will go to B1, B2, C1 or C2. But when it is entering the highway network, we only have very limited contextual information about it, and a large search space for final destinations on the grid level. In contrast, when it is leaving the highway station at station C, we have more contextual information, e.g., travel time from A to C, and a smaller search space, e.g., C1 and C2. This justify why StateFlow has a two level prediction structure.

Based on this example, the key challenges we aim to address in StateFlow are as

given as follows. (i) For station-level prediction: how to predict exiting stations and exiting time? (ii) For grid-level prediction: how to predict the final destination on the grid level? As following two subsections, we introduce our detailed design to address these challenges.

4.2 Station-Level Prediction

In this subsection, we introduce how to predict the exiting station and exiting time on the highway level for inter-regional travels.

4.2.1 Exiting Station Prediction

To predict a vehicle's exiting station, we essentially need to assign an exiting probability to a station in the highway network. Specially, we want to estimate $p(\theta_d|\pi, \theta_s, t_0)$ the probability that a vehicle π entering the highway via an entering station θ_s during time t_0 has its exiting station as θ_d. By applying the Bayesian rule, we have

$$p(\theta_d|\pi, \theta_s, t_0) = \frac{p(\pi, \theta_s, t_0|\theta_d) \times p(\theta_d)}{p(\pi, \theta_s, t_0)}$$

where $p(\pi, \theta_s, t_0|\theta_d)$ is the probability that a vehicle that exited at the exiting station θ_d was entering from the station θ_s during t_0; $p(\theta_d)$ is the probability that any vehicle has the station θ_d as its exiting station; $p(\pi, \theta_s, t_0)$ is the probability that the vehicle π enters the highway via the station θ_s during t_0. Given large-scale historical data for all highway traffics, we obtain $p(\pi, \theta_s, t_0|\theta_d)$, $p(\theta_d)$ and $p(\pi, \theta_s, t_0)$ for each vehicle, each station at each time interval with a statistical method.

4.2.2 Exiting Time Prediction

To predict the exiting time, we need to infer the travel time given an entering station, an entering time period, and a predicted exiting station. However, the travel time prediction is challenging because it is influenced by a combination of real-world factors, e.g., driver's habits or skills, weather conditions, traffic conditions, and time of day. For instance, Figure 4.2 shows the travel time distribution between two stations with the

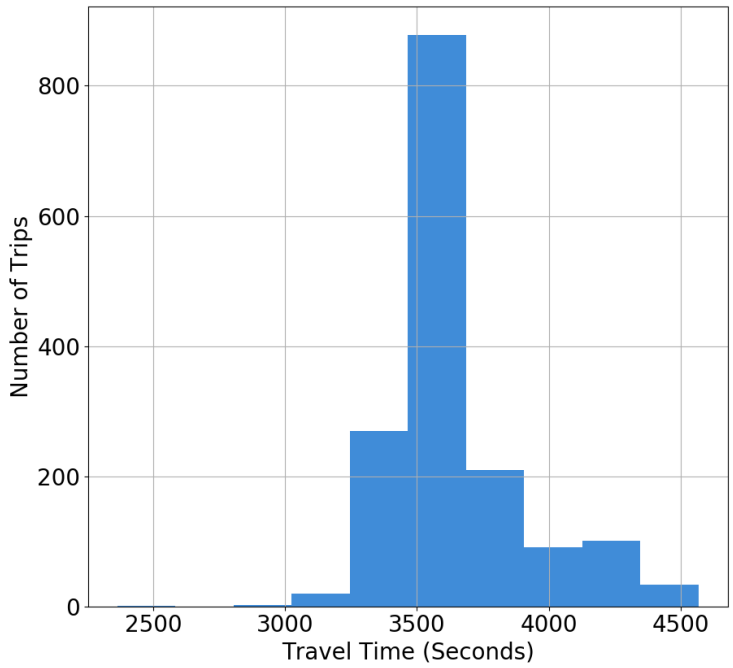

Figure 4.2: Travel Time between Two Stations

highest traffics. We can see that the travel time can vary more than 50% between the same stations given the same starting time.

In StateFlow, we solve the travel time prediction problem in a unsupervised manner. We assume drivers with similar driving habit would share similar travel time between the same locations under similar context like weather and time of day. Based on this assumption, we use the K-Nearest Neighbors method to find the top k most similar travel records then use the average travel time of these records as predicted travel time. To compute the similarity, we take four factors into consideration:

- **Driving Habit:** In the context of travel time, we consider that drivers with similar driving time between the same locations have similar driving habits. Since we do not have detailed information, we assume each vehicle has only one driver so the driving habit is identified by a specific vehicle. Then we use the historical average travel time as the specific feature. Considering the factor that highways are generally used as long distance travel, we use one minute as our time granularity.

- **Traffic Condition:** We use the average travel time between stations in the last time interval (e.g. last half hour) to represent the traffic condition.
- **Weather Condition:** Bad weathers like heavy rain or snow would dramatically influence driving speeds in general. We category weather into three cases: heavy rain/snow, small rain and none-rain and set their numerical values as 1, 0.6, and 0.3, respectively.
- **Time of Day:** Driving time is also important since people generally drive slower at night. To quantify the driving time, we present it as a Gaussian distribution which uses mid-night as mean. When the driving time comes to near the mid-night, the value is higher.

To uniform the similarity calculations, we normalize these features into the range [0, 1] and then use Euclidean distance to generate the overall similarity between records.

4.2.3 Putting them together

Since both destination prediction and travel time prediction is from the perceptive of individual vehicles, it is easy to put our model into practice with online updating. For exiting station predictions, we can update $p(\pi, \theta_s, t_0|\theta_d)$, $p(\theta_d)$ and $p(\pi, \theta_s, t_0)$ in real time as we receive updates from the ETC system. For travel time, the driving habits of a driver and traffic conditions can be updated when vehicles finish their travels from one station to another station. Real-time weather conditions can also be obtained from many online resources, e.g., National Centers for Environmental Information.

Based on our interactions with the ETC system operators, they are also interested in flow estimations at different stations for them to understand the status of their systems. But the technical challenges in the aggregating all these vehicles at individual levels to obtain flow estimation is the data volume. In particular, we need to predict the exiting station and time station time for every vehicle entering all stations. A high-performance cluster can handle the computation, but the ETC system operator can only budget a normal commodity server. To address this issues, we perform two optional approximations on StateFlow.

- **Limiting Exiting Station Candidates:** We only assign a vehicle to the exiting station with the highest predicted probability $p(\theta_d|\pi, \theta_s, t_0)$, i.e., the vehicle goes to that exiting station with a probability of 1. This approximation, though seems biased for each vehicle, achieves reasonable flow aggregation results. This may be because the biases from multiple vehicles will balance from each other.

- **Spatial and Temporal Pruning:** For the flow aggregation at an exiting station, we only consider the vehicles entering from 800 toll stations (about 50% of total toll stations based on their flow contributions towards the exiting station historically) within 2 hours. This approximation is based on the observation of spatiotemporal locality of the traffics in the highway system, i.e., the dominating traffics arriving at a station are from a limited number of stations within a relatively short time range. We justify this approximation by showing the spatial and temporal CDF in Figure 4.3 and Figure 4.4. We found that 800 toll stations account for more than 95% of the total trips, and the trips shorter than 2 hours, e.g., 120 mins, account for more than 90% of the total trips.

After applying these two approximation, with one CPU thread, we can compute the contribution of each vehicle in 100 milliseconds (including time for both exiting station prediction and travel time prediction) and we finish the flow aggregation at each station in 10 milliseconds. Given that maximum throughput is about 30 vehicles entering the highway per second, a commodity server is more than enough to provide real time service for the whole Guangdong province based on our current data volume.

4.3 Grid-Level Prediction

One of the key strengths of the StateFlow is that it not only captures the mobility between toll stations on highways but also captures the mobility after vehicles leave toll stations using our mobile sensing fleet data. This helps us to predict the final destination of vehicles without GPS in a lower spatial level, e.g., grid, which provides a finer-grained modeling for vehicles after they exit the highway ETC systems.

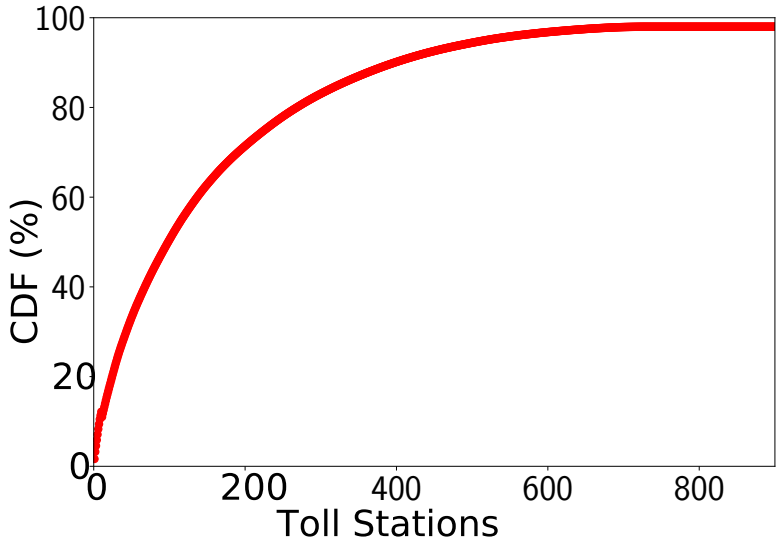

Figure 4.3: Spatial CDF

4.3.1 Grid Representation

For fine-gained mobility modeling inside a region, we use a grid representation to divide a region into grids, which are commonly used in map division [20]. Considering the whole area of the Guangdong province is 179,800 km^2, we use a $5km \times 5km$ grid. As a result, on average, each region for highway mobility modeling is divided into over 60 cells. Figure 4.5 gives an example of how the cell layout for a major city Shenzhen in Guangdong province which has 6 regions in the state level.

4.3.2 Trips Extraction

We extract trips from fleet GPS location data to infer potential destinations of regular vehicles. The location updates of a vehicle is represented as a sequence of (time, location) tuples. For our model, we focus on the source and destination of a vehicle, which are used to define a *Trip*. However, not all tuples are directly useful since a vehicle may be stopped in some time periods. Moreover, a sequence of tuples may belong to different trips because a vehicle may have multiple trips in a day. To extract trips from a trajectory, we use a time gap k to define if two locations belong to the same trip. Specially, we construct one trip record if and only if the time gap between two

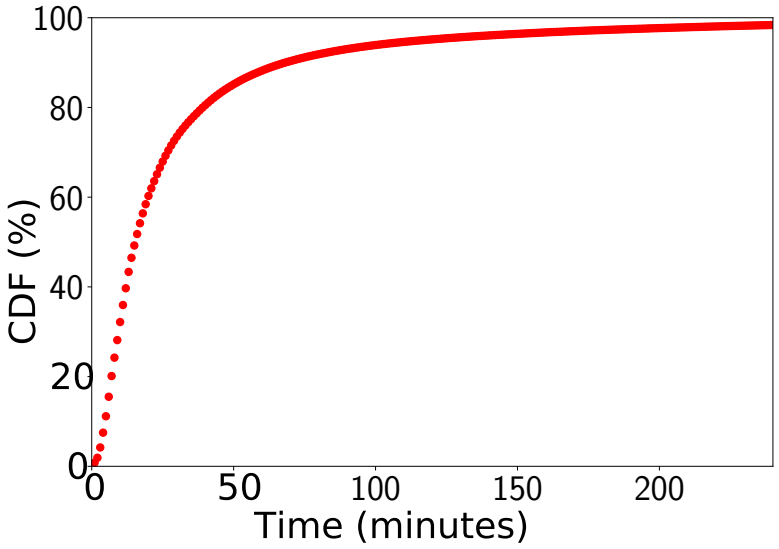

Figure 4.4: Temporal CDF

consecutive locations is no more than 30 minutes [5]. Once the trips are identified, the destination grid of the trips can be determined by matching the last GPS coordinates for a trip to the predefined $5km \times 5km$ grid.

4.3.3 Map Matching

As mentioned previously, our intra-region mobility model aims to predict the destination of the individual vehicles after they exiting from the highways. As a result, we need to match the trip trajectories extracted in the previous subsection to highways, and then only examine those on highways for mobility modeling. Fortunately, our highway matching problem is significantly easier than the map matching algorithm in dense urban road network [1]. This is because our GPS sampling frequency is very high, i.e., 10 to 15 seconds on average, due to the fact that we obtain GPS data for navigation services, and the highways in Guangdong are usually located far from other local roads. We therefore applied a simple projection distance based algorithm [20] to match trip trajectories on highways.

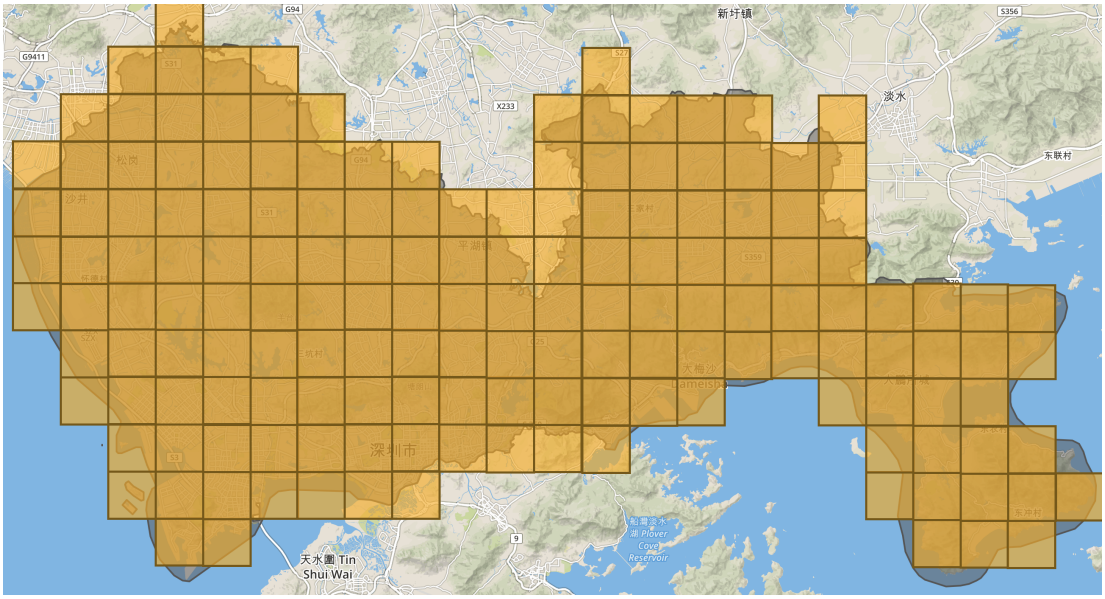

Figure 4.5: Grid Representation in China City Shenzhen

4.3.4 Feature Selection and Prediction

The intra-region mobility modeling (i.e., predicting the final destination) is different from highway-level modeling ((i.e., predicting the exiting station). This is because by the time we predict the final destination, we already have both real-time highway data and historical local data as features. But for exiting station predictions, we only have limited contextual information about entering station and time. As follows, we show how to select a few features to predict the final destination at grid levels for a particular vehicle without GPS devices based on our mobile sensing fleet with GPS devices.

- **Regular Vehicle Features**: Since we do not have GPS data for regular vehicles, we extract their features of trips from the ETC system. Where and when these regular vehicle vehicles getting into the highway system and getting off the highway system have a strong correlation with where they finally go. We choose the time-stamped entering station, time-stamped exiting station and travel time from the entering station to the exiting station as our highway features.

- **Mobile Sensing Fleet Features**: Since we have both GPS data and ETC data for all the vehicles in the mobile sensing fleet, we can extract more features from them compared to regular vehicles. But since our key objective is to utilize the

GPS data of mobile sensing fleet to examine possible final destinations at grid level for regular vehicles, we extract features similar to regular vehicles. We have where and when these fleet vehicles getting into the highway system and getting off the highway system. But the key difference is that for mobile sensing fleet, we extract their final destination on the grid level as an additional feature.

Based on these two sets of features, we train a Random Forest model [4] with both mobile sensing fleet data and regular vehicle data. As a result, when a regular vehicle exiting a toll station, we utilize its regular vehicle features to predict its potential final destination based on this model, along with mobile sensing fleet data.

Chapter 5

Mobility Modeling Layer: Evaluation

StateFlow conducts traffic flow estimation based on prediction of individual's highway exiting toll station and the final destination grid after vehicles getting off the highway. Therefore, in this chapter, we in general divide the evaluation to answer the following two groups of questions for the individual perspective and the flow perspective:

1. At the individual level, how accurately StateFlow predicts driver's exit toll station after getting on highways, and how accurately it predicts driver's final destination grid after getting off the highway.

2. At flow level, how accurately StateFlow estimates the flow arriving at an exit toll station on highways, and whether the estimated flow direction for the traffic getting off a highway exit station is correct.

5.1 Methodology and Baseline

Our evaluation of StateFlow is based on two sets of real world data in Guangdong, China: (i) 52 million toll transactions generated by 7.8 million vehicles at 1439 toll station during 30 days, and (ii) fleets trajectories in a week including 114 thousand vehicles. During the evaluation, we chronologically partition the dataset into the training set and the test set and present the number of metrics for both individual prediction and aggregated flow prediction.

For individual prediction, we predict the vehicle's exit toll station when the vehicle enters the highway and predict the vehicle's final destination (i.e., a grid) when the vehicle leaves from the highway. For individual prediction evaluation, we use precision as the metrics to represent the prediction accuracy. Precision is defined as the number

of trips whose final destinations (i.e., the toll station for highway prediction and the grid for intra-region prediction) is correctly predicted over the total number of trips. As for the comparison algorithms, we use the most frequent destination of a vehicle and a random selection in a vehicle's historical destinations as baselines.

For flow prediction evaluation, we set the estimation interval as one hour, i.e., StateFlow predicts the total traffic flow arriving at a highway exit or a grid within every hour in a day. For highway flow prediction, since we have the ground truth traffic volume collected in the toll transactions dataset, we compute the Root Mean Square Error (RMSE) to measure the difference between the estimated total flows and the actual total flows.

As for the intra-region flow prediction, since it is extremely challenging to track all the vehicles in reality at a state level to obtain the ground-truth of the mobility from highway exits to each grid, we conduct our prediction for the vehicles whose trajectories are tracked in our mobile sensing component. The comparison is based on the ground truth flow generated with tracked vehicles and predicted flow from the same group of vehicles. Given a large amount of vehicles tracked in our mobile sensing system, we envision it reflects the overall mobility trend of the total traffic.

5.2 Individuals Prediction

5.2.1 Individual Highway Exit Station Prediction

Our highway mobility model relies on the historical data because they provide the personalized mobility patterns of individual vehicles without GPS. The longer history we have, the more reliably we can understand a vehicle's mobility pattern. For this evaluation, we use the data from first 29 days for training and the 30th day for testing. To test the influence of different history lengths, we generate 29 training set setting, by accumulatively using only the first day as the training set, the first two days as the training set, until the first 29 days as the training set. The precisions of the prediction on the test set, i.e., the data on the 30th day, using different training set size are presented in Figure 5.1. As Figure 5.1 shows, StateFlow achieves significantly higher

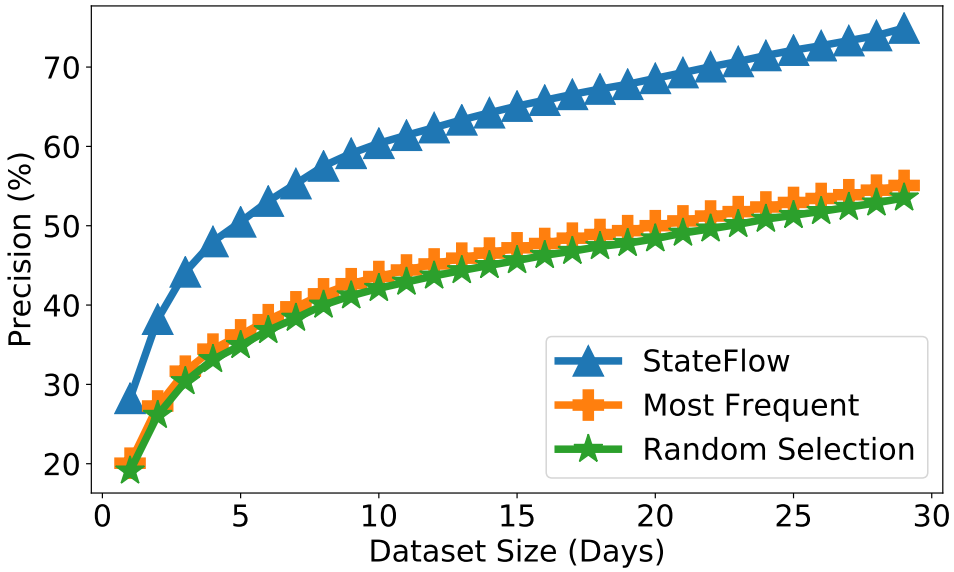

Figure 5.1: Destination Prediction

precision for the highway exit station prediction. In particular, StateFlow correctly predicts the exit highway station for about 75% of all the vehicles. With the increase of the data size, the precision increases rapidly. However, since there are certain amount of new vehicles without any historical data coming into the highway system every day, there is an upper bound for the increment of the precision. Based on our data, we found that the bound is close to around 80%.

5.2.2 Individual Highway Travel Time Prediction

Highway travel time prediction influences the expected arrival time at the exit station, given the time when the vehicle enters the highway via a toll station is known. We evaluate the highway travel time prediction using two case studies over a short-distance station pair and a long distance station pair. Vehicles need about 20 minutes to travel from a station to the other station for the short distance station pairs, and about 1.5 hours for the long-distance station pairs. By comparison, we use the mean value between two stations as the baseline and calculate the precision gain PG over the baseline using

the following formula.

$$PG_i = \frac{|pred_i - base_i|}{actual_i}$$

where PG_i is the Performance Gain for the ith trip; $pred_i$ and $base_i$ are the predicted and baseline travel times of the ith trip; $actural_i$ is the actual travel time of the ith trip.

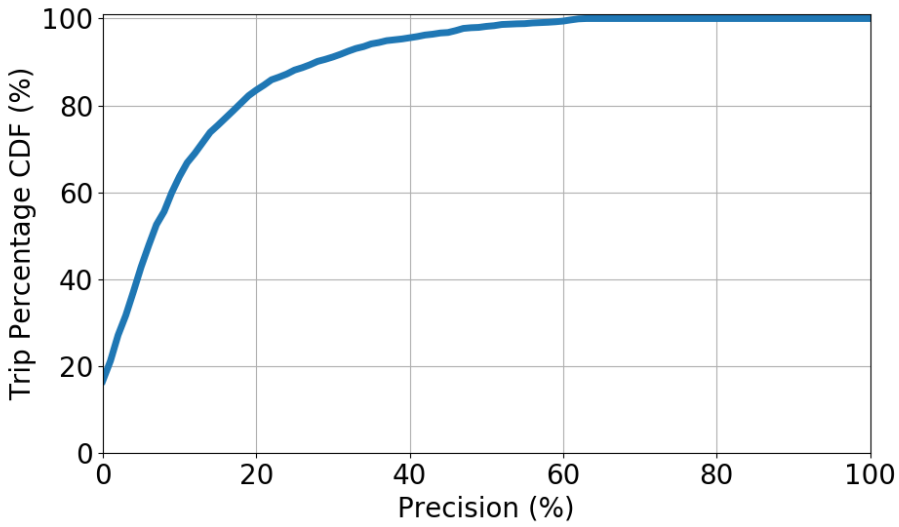

Figure 5.2: Short-Time Travel Prediction Improvement

In the short time travel, the average error of our model is around 21 seconds while the average error is around 185 seconds for the baseline. Figure 5.2 gives the detailed PG distribution. There are about 20% of trips with the precision less than the baseline and 80% of trips have better predicted results. It seems that 60% is the upper bound PG value for the precision based on the current data for short time travel.

In the long time travel, the average error for our model is around 101 seconds, while the average error for the baseline is around 312 seconds. Figure 5.3 gives the detailed PG distribution. There are still around 80% of trips that have better predictions than the baseline. It seems that the upper bound of the PG) value is around 20% based on current data for long time travel.

The above results show that for both long distance travels and short distance travels, StateFlow travel time prediction component outperforms the baseline.

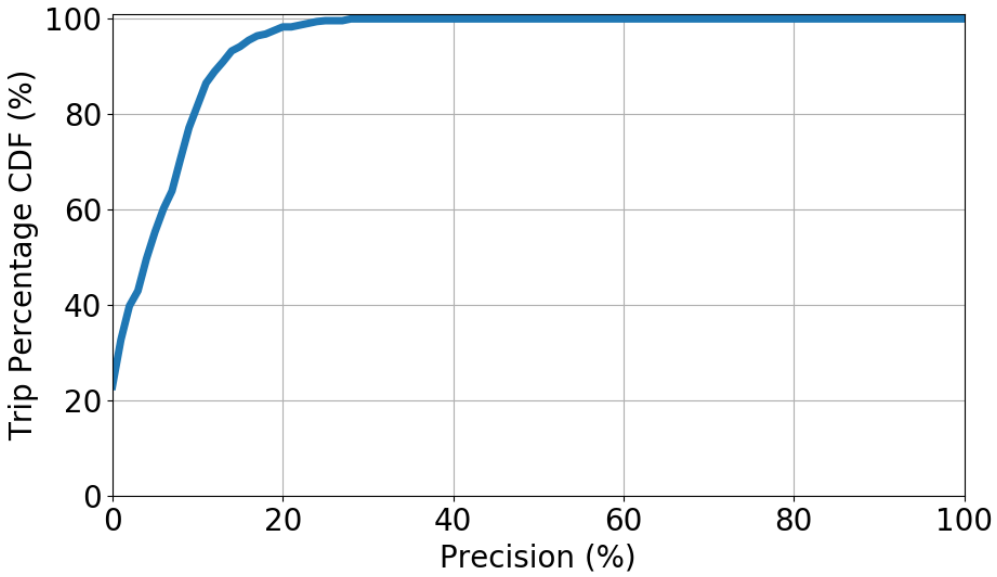

Figure 5.3: Long-Time Travel Prediction Improvement

5.2.3 Individual Local Destination Prediction

In this part, we test the individual destination prediction over two baselines which are similar to the chapter 5.2.1: drivers' most frequent destination and random selection of a destination from drivers' historical destinations. Considering that we only have limited data for grid-level modeling, we use cross-validation to test the performance. Figure 5.4 presents the precision in one-week data when using the data from different data as the test set, and the rest as the training set.

It is clear in Figure 5.4 that the StateFlow grid level mobility prediction components generally achieve better prediction accuracy than the baselines. In particular, StateFlow can correctly predict the grid level destinations (5km $\times$ 5km square) for 70% of the traffic when they exit the highway network.

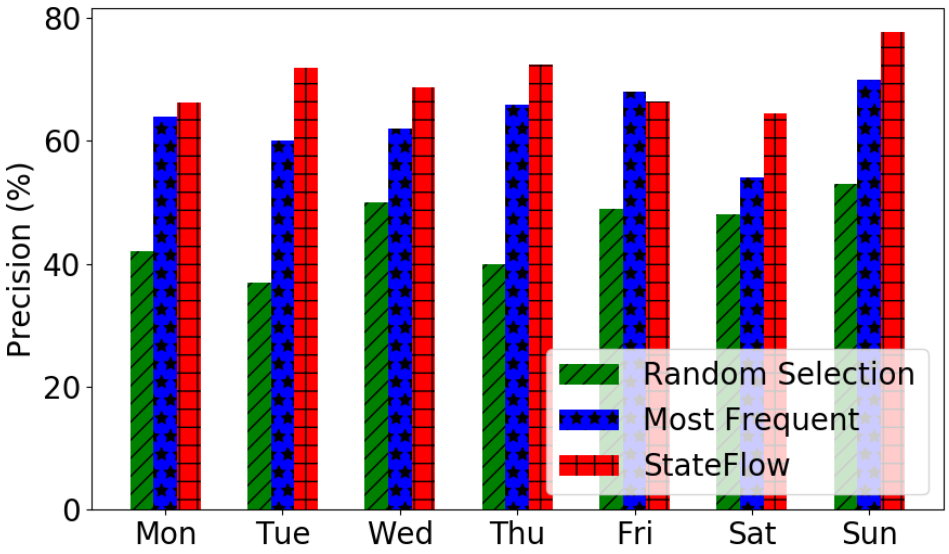

Figure 5.4: Sub-Region Individuals Precision

5.3 Flow Prediction

5.3.1 Highway Flow Prediction

StateFlow accurately models individual's mobility, in terms of exit toll stations and travel time, which lead to results in the arrival flow as a highway exit station by aggregation. The ETC system operators are interested in these results to understand their system. We use the data from the first 29 days as the training data and the data on the 30th day as the test set. We also calculate the traffic flow using vehicles' actual toll transactions at each station, and the average flow in the same interval in the first 29 days as the benchmarks.

Figure 5.5 shows the average prediction precision of arrival flow over all stations using StateFlow and the benchmark results. It shows that StateFlow has the estimation very close to the actual flow during the night time and early morning, i.e., 8 pm to 6 am, where the traffic is in general low compared to other time periods. However, during the day time when the traffic volume is high, StateFlow performs better than the average flow prediction.

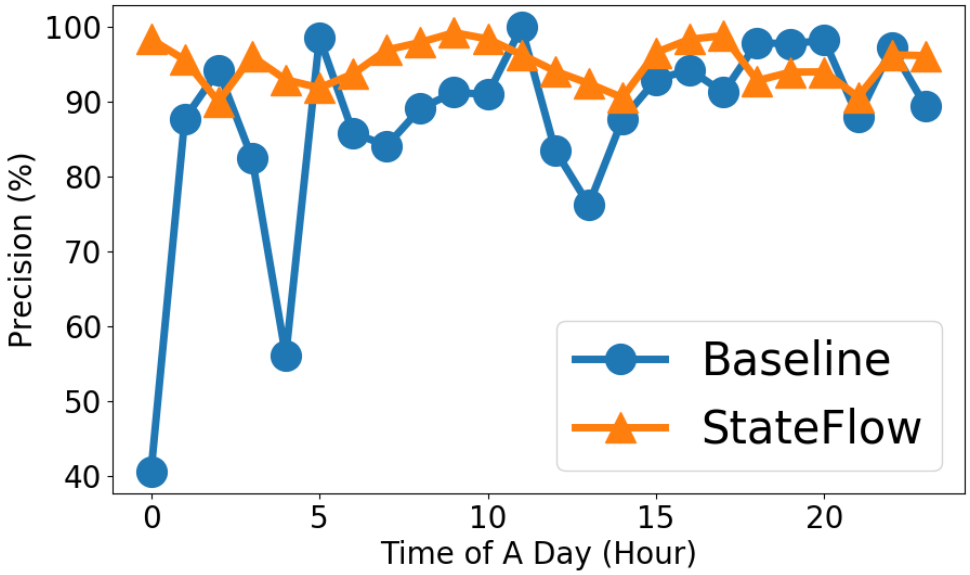

Figure 5.5: Exit-Flow Comparison

Numerically, we define RMSE over all intervals as

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=0}^{T}(Pred_t - Real_t)^2}, \tag{5.1}$$

where $Pred_t$ is the estimated arrival flow at time interval t; $Real_t$ is the ground truth arrival flow at time interval t. T is the total number of time intervals in a day. We set one hour as the time interval for arrival flow evaluation, so there is total 24 estimated value during a day. The RMSE of StateFlow arrival flow prediction is 67.4 compared to 147.7 of the average flow baseline algorithm.

5.3.2 Grid Flow Prediction

Since it is extremely hard in the real world to track the mobility of all the vehicles after they get off the highway, we study the grid level arrival flow in each grid based on the vehicles whose trajectories are tracked by our mobile sensing component. Figure 5.6 shows the grid percentage CDF of precision when we use the first 6 days in our trajectory dataset as the training set and the last day as the test set. We found from Figure 5.6 that 50% of grids has a precision no higher than 70%.

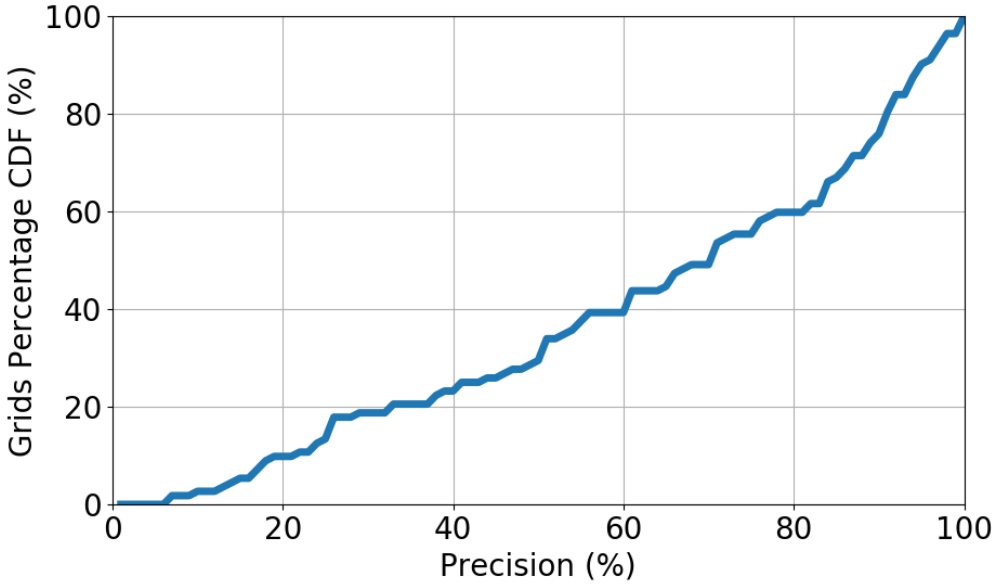

Figure 5.6: Grid Flow Precision Distribution

5.4 Evaluation Summary

Based on the experiments in various contexts and multiple levels, we make the following observations:

1. The exit station, travel time and exit time in highway are highly predictable given vehicle's historical mobility data and latest real-time traffic conditions. These findings are supported by Figure 5.1, Figure 5.2, and Figure 5.3.

2. The destination grid of the driver who uses the highway is also predictable, given the vehicle's highway and local mobility history and traffic conditions. This observation is supported by Figure 5.4.

3. The accurate prediction of individual driver's destination and travel time provides an aggregation of the arrival flow at highway stations (Figure 5.5) and grids in each region (Figure 5.6), which together make the state-level mobility flow predictable.

Chapter 6

Application Layer

With the availability of the mobility modeling and prediction system, many urban applications can be suggested and built with much fewer efforts. In this chapter, we demonstrated one case study that we built using our mobility modeling system: a personalized dynamic message sign (DMS) system for route suggestion at the entrance and the exit toll stations of the highway.

6.1 Background

Figure 6.1: Current en-route message sign

Figure 6.2: A typical toll entrance message sign

Nowadays, the DMSs in the highway system primarily target at the total traffic stream instead of individual drivers (as shown in Fig. 6.1), because they are shown to all drivers and do not have individual driver's travel information. DMSs at the highway toll station, though only visible for targeted drivers, now are used for displaying simple vehicle information or payment status (e.g., Fig. 6.2 show the vehicle plate number and payment information). In this case study, we propose to modify the toll station

DMS system to show route suggestions for the driver: at the entrance station, the DMS shows the route suggestion to the drivers to the predicted exit station (e.g., multiple ring routes between the same entrance and exit station pair in Figure 6.3 with different travel time due to real-time traffic); at the exit station, the DMS shows the suggested route to the driver's final destination at the grid level.

A successful DMS system for route suggestion relies on accurate travel time estimation and route suggestion algorithms but more importantly determined by the accuracy of drivers' exit station and final destination prediction. This is because vehicles traveling on the highway usually do not notify our system about their travel plans. Moreover, a driver usually only stops for a few seconds when entering the toll station. The limited display space and visibility time of DMS suggest that route for the only very limited number of destinations can be shown at the toll station DMS. We next elaborate on how the proposed route suggestion DMS system predicts driver's destination and calculate the route for the drivers.

6.2 Destination Prediction

When a vehicle arrives at the entrance toll station, the RFID reader gets access to its historical data. The current toll station ID, the current time, and the historical data are provided to the StateFlow to obtain the list of possible exit stations together with the associated probability. Based on the probability, we choose the exit toll stations with the highest probabilities as the predicted exit toll station.

When the vehicle reaches the exit toll station, we will predict the driver's final destination grids using our grid level mobility model. We choose the grids with the highest destination probabilities as the predicted final destination and translate the grid into road and town name using reverse geocoding, and then show that in the exit station display.

6.3 Route Computation

Once either the predicted exit stations or predicted final destinations are determined, the route suggestion DMS system needs to compute the highway route suggestion (from the highway entrance station to the predicted highway exit stations) and local route suggestions (from highway exit station to the predicted final destinations).

For highway route suggestion, since we have plenty of highway travel time observations extracted from toll transactions, we estimate the travel time between adjacent toll stations and then use Dijkstra's algorithm to find the fastest path. Specifically, we define a highway road segment as the highway section connecting two consecutive toll stations and use the travel time allocation and travel time aggregation algorithm presented in [18] to estimate average travel time of every segment during an estimation interval (e.g., 30 mins).

For local route suggestion, depending on whether the trajectory data is dense enough to estimate the traffic speed, we either apply the algorithm presented in [18] to estimate road segment level travel time during an estimation interval (e.g., 30 mins) or use speed limit in OpenStreetMap to build the suggested the route.

6.4 Application Evaluation

In this subsection, we evaluate the route suggestion DMS system in terms the efficiency of the route suggestion result. Clearly, if there is no route suggestion DMS system, a vehicle's travel time (TT) is captured by our existing dataset (i.e., highway toll transactions and local vehicle trajectory). This is defined as the baseline scenario. In contrast, for the situation when a route suggestion DMS system exists, we assume that a vehicle will follow the suggested route if the route suggestion for the vehicle's exit station (or the final destination grid) is shown in the DMS at the entrance station (or at the exit station), i.e., the destination prediction is correct. In this case, the driver's travel time can be estimated using other drivers travel time. If the destination prediction is inaccurate, the vehicle follows its actual driving route and therefore, the travel time would be its actual travel time. In this way, we synthesize the travel time

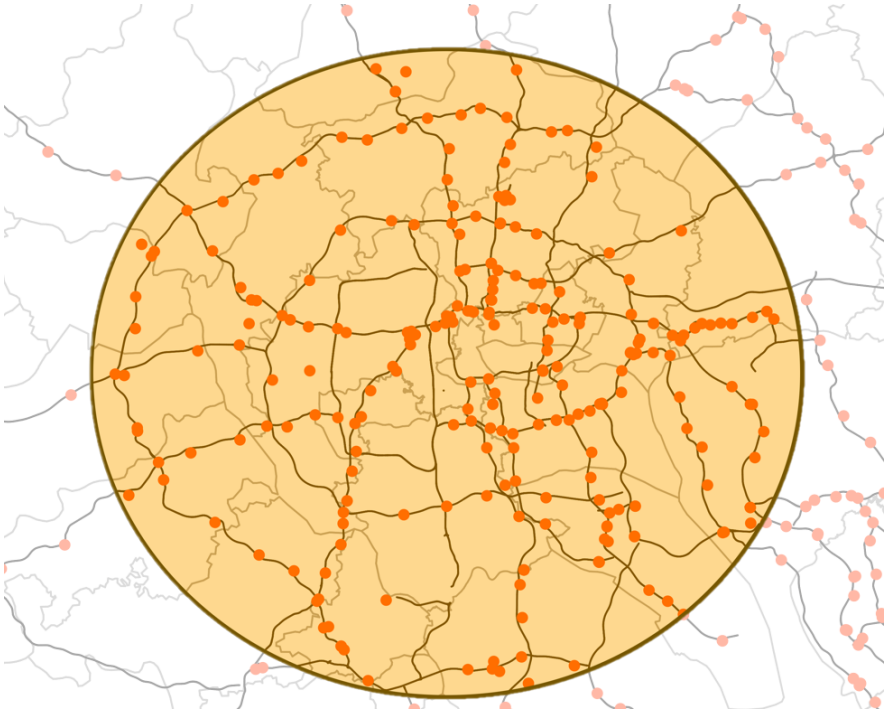

Figure 6.3: Guangzhou Highway Network

(ETT) when DMS works. As for the experiment scenario, we choose Guangzhou city which has the most complex highway road network in Guangdong province. Figure 6.3 shows the highway network. As shown by the yellow circle, we can see that there are multiple routes for people to choose between two locations with different real-time traffic speed. Our application is proposed to solve this problem.

We define relative travel time reduction (RTTR) as follows:

$$RTTR = (TT_{w/\,DMS} - TT_{w/o\,DMS})/TT_{w/o\,DMS},$$

to represent the ratio of reduced travel time if route suggestion DMS system is introduced in Guangdong Highway system.

Figure 6.4 shows the average RTTR for all the travels during different time intervals of a day. It can be easily found that when route suggestion DMS system is introduced, the average travel time is reduced. In particular, vehicle's average travel time can be reduced by 36% on average and by 60% at maximum during the rush hour, e.g., 10 am. We can see the trend of Figure 6.4 is similar to the overall flow shown in Figure 5.5. When the traffic is heavy, StateFlow can achieve better results because it knows the

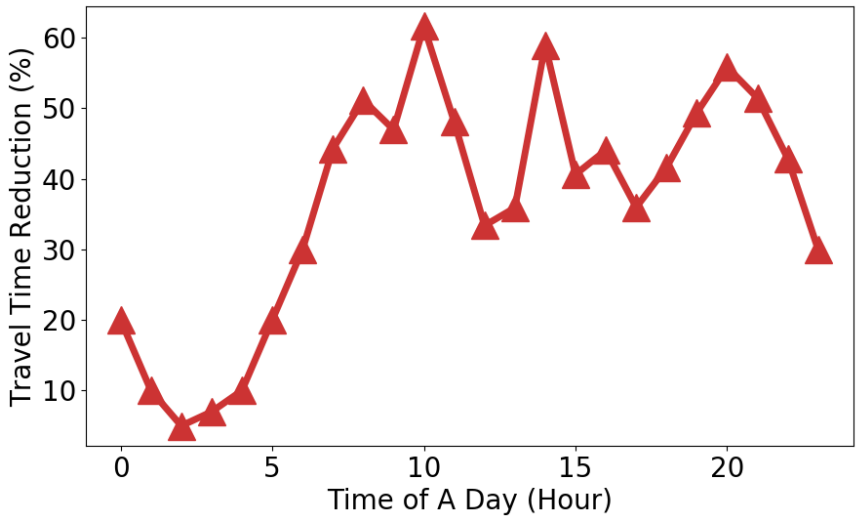

Figure 6.4: RTTR in A Day

real time traffic conditions in multiple road segments, which can be used to provide an optimal route. While the traffic is not heavy in the morning, drivers always choose the route with the shortest distance, which is the best route.

Chapter 7

Discussion

In this chapter, we provide some discussions as follows:

Scalability: StateFlow fuses stationary sensor data and mobile sensor data together to provide a fine-grained mobility modeling. In our context, we use transactions from ETC systems as stationary data and trajectories from fleets as mobile data as an example to conduct our experiments. However, in real life, there are other sensors that can be integrated into our model. For example, stationary sensors like road cameras, traffic loop sensors, cell towers, WiFi stations can be helpful to detect more mobility under other conditions even in the indoor environments. Mobile sensors like smartphones can be used to track mobility with a higher coverage. These sensors can be integrated into StateFlow as multiple levels and stages to perform mobility modeling.

Privacy: Privacy is a major concern in most of the systems related to location data. Travel records are all concerned with privacy. In StateFlow, all the vehicle plates are anonymized as globe IDs. The state flows are also presented as aggregated flows of individuals which can also benefit privacy protection. When the data is large enough, we can perform some sampling and differential privacy technologies to protect as much information as possible.

Data Incompleteness: Data incompleteness is also another concern that appears in many data related systems since it is impossible to obtain the complete data like human mobility and vehicles trajectories in real life. In our grid level mobility estimation, we only use the vehicles we tracked to study their final destination which is not on the same scale as the real number of vehicles getting off the highways. Scaling the mobility by estimating its distribution over different grids is a possible solution for this problem, but requires more investigation.

Chapter 8

Conclusion

In this dissertation, we have identified the challenges and opportunities in vehicular mobility modeling in real time. With the massive data from real life, we provide detailed analysis and visualization of the vehicular intra-city mobility and inter-city mobility. Based on the features of the data, we have presented a novel approach to address these challenges by combining stationary sensing and mobile sensing into a two level structure and demonstrated their feasibility through various experiments.

The main contribution of this dissertation is to provide the insights to deeper understanding of inter and intra city vehicular mobility and design a novel system called StateFlow to model the inter and intra mobility in state level. StateFlow is designed as a two-level model, which separates mobility into station level and grid level. Based on these two-level prediction, we can track individual vehicles from entering the highways to arriving the final destinations.

- **Stationary Sensing:** Stationary sensors such as electric toll system and cameras provide detailed and complete sensing in fixed locations with all time monitoring. With stationary sensing, we use Bayesian Inference to predict the exit stations and use K-Nearest Neighbors to predict the travel time between two stations in the station level.

- **Mobile Sensing:** Mobile sensors such as vehicles and smartphones provide better spatial coverages and personalized mobility information. With mobile sensing, we build a random-based model to predict vehicle final destinations in the grid level.

Combining stationary sensors and mobile sensors together compensates the shortages

of individual sensors which provides better spatiotemporal coverages. Two-level mobility tracking is also a flexible framework that can be extended to multiple levels with more sensors involvement in different levels. To further demonstration the practicability of the StateFlow, we apply it into a novel dynamic message sign system which is demonstrated to improve people's travel efficiency in real life.

Vita

Yu Yang

2011-15 B.E. in Software Engineering from Northeastern University, China.

References

[1] Aslam, J., Lim, S., Pan, X., and Rus, D. City-scale traffic estimation from a roving sensor network. In *Proceedings of 10th ACM Conference on Embedded Network Sensor Systems*, SenSys '12.

[2] Balan, R. K., Nguyen, K. X., and Jiang, L. Real-time trip information service for a large taxi fleet. In *Proceedings of the international conference on Mobile systems, applications, and services*, MobiSys '11.

[3] Cheng, Y., Li, X., Li, Z., Jiang, S., Li, Y., Jia, J., and Jiang, X. Aircloud: a cloud-based air-quality monitoring system for everyone. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems* (2014), ACM, pp. 251–265.

[4] De Poalo, T., and Howard, J. Predictive modeling in practice: A case study from sprint. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2014), KDD '14, ACM, pp. 1517–1517.

[5] Du, B., Liu, C., Zhou, W., Hou, Z., and Xiong, H. Catch me if you can: detecting pickpocket suspects from large-scale transit records. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), ACM, pp. 87–96.

[6] Fan, Z., Song, X., Shibasaki, R., and Adachi, R. Citymomentum: an online approach for crowd behavior prediction at a citywide level. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015), ACM, pp. 559–569.

[7] Ganti, R., Srivatsa, M., Ranganathan, A., and Han, J. Inferring human mobility patterns from taxicab location traces. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (New York, NY, USA, 2013), UbiComp '13, ACM, pp. 459–468.

[8] Ganti, R. K., Pham, N., Ahmadi, H., Nangia, S., and Abdelzaher, T. F. Greengps: A participatory sensing fuel-efficient maps application. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services* (New York, NY, USA, 2010), MobiSys '10, ACM, pp. 151–164.

[9] Gao, Y., Swaminathan, K., Cui, Z., and Su, L. Predictive traffic assignment: A new method and system for optimal balancing of road traffic. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on* (2015), IEEE, pp. 400–407.

[10] Ge, Y., Liu, C., Xiong, H., and Chen, J. A taxi business intelligence system. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11.

[11] Ge, Y., Xiong, H., Tuzhilin, A., Xiao, K., Gruteser, M., and Pazzani, M. An energy-efficient mobile recommender system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (2010), KDD '10.

[12] Gonzalez, H., Han, J., Li, X., Myslinska, M., and Sondag, J. P. Adaptive fastest path computation on a road network: a traffic mining approach. In *Proceedings of the 33rd international conference on Very large data bases* (2007), VLDB '07.

[13] Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. Understanding individual human mobility patterns. *Nature 453*, 7196 (2008), 779–782.

[14] Hasenfratz, D., Saukh, O., Sturzenegger, S., and Thiele, L. Participatory air pollution monitoring using smartphones. *Mobile Sensing* (2012), 1–5.

[15] Ho, B.-J., Martin, P. D., Swaminathan, P., and Srivastava, M. B. From pressure to path: Barometer-based vehicle tracking. In *BuildSys@SenSys* (2015).

[16] Hu, S., Su, L., Liu, H., Wang, H., and Abdelzaher, T. F. Smartroad: Smartphone-based crowd sensing for traffic regulator detection and identification. *ACM Transactions on Sensor Networks (TOSN) 11*, 4 (2015), 55.

[17] Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., and Willinger, W. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services* (2012), Acm, pp. 239–252.

[18] Liu, R., Liu, H., Kwak, D., Xiang, Y., Borcea, C., Nath, B., and Iftode, L. Balanced traffic routing: Design, implementation, and evaluation. *Ad Hoc Networks 37* (2016), 14–28.

[19] Liu, W., Liu, J., Jiang, H., Xu, B., Lin, H., Jiang, G., and Xing, J. Wilocator: Wifi-sensing based real-time bus tracking and arrival time prediction in urban environments. In *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)* (2016).

[20] Marchal, F., Hackney, J., and Axhausen, K. Efficient map matching of large global positioning system data sets: Tests on speed-monitoring experiment in zürich. *Transportation Research Record: Journal of the Transportation Research Board*, 1935 (2005), 93–100.

[21] Miao, F., Lin, S., Munir, S., Stankovic, J., Huang, H., Zhang, D., He, T., and Pappas., G. J. Taxi dispatch with real-time data in metropolitan areas. ACM ICCPS 2015.

[22] Michau, G., Nantes, A., Bhaskar, A., Chung, E., Abry, P., and Borgnat, P. Bluetooth data in an urban context: Retrieving vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems* (2017).

[23] Michau, G., Pustelnik, N., Borgnat, P., Abry, P., Nantes, A., Bhaskar, A., and Chung, E. A primal-dual algorithm for link dependent origin destination matrix estimation. *IEEE Transactions on Signal and Information Processing over Networks 3*, 1 (2017), 104–113.

[24] Moghadam, K. R., Nguyen, Q., Krishnamachari, B., and Demiryurek, U. Traffic matrix estimation from road sensor data: A case study. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2015), ACM, p. 65.

[25] Nawaz, S., Efstratiou, C., and Mascolo, C. Parksense: A smartphone based sensing system for on-street parking. In *Proceedings of the 19th annual international conference on Mobile computing & networking* (2013), ACM, pp. 75–86.

[26] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. A tale of many cities: universal patterns in human urban mobility. *PloS one 7*, 5 (2012), e37027.

[27] Noulas, A., Scellato, S., Lathia, N., and Mascolo, C. Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), 2012 IEEE 12th international conference on* (2012), IEEE, pp. 1038–1043.

[28] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017.

[29] Sen, R., and Balan, R. K. Challenges and opportunities in taxi fleet anomaly detection. SENSEMINE'13.

[30] Song, T., Capurso, N., Cheng, X., Yu, J., Chen, B., and Zhao, W. Enhancing gps with lane-level navigation to facilitate highway driving. *IEEE Transactions on Vehicular Technology* (2017).

[31] Thiagarajan, A., Biagioni, J., Gerlich, T., and Eriksson, J. Cooperative transit tracking using smart-phones. In *SenSys* (2010).

[32] Thiagarajan, A., Ravindranath, L., LaCurts, K., Madden, S., Balakrishnan, H., Toledo, S., and Eriksson, J. Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *SenSys* (2009).

[33] Wang, Y., Liu, X., Wei, H., Forman, G., Chen, C., and Zhu, Y. Crowdatlas: Self-updating maps for cloud and personal use. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services* (2013), ACM, pp. 27–40.

[34] Wikipedia contributors. Expressways of China. `https://en.wikipedia.org/wiki/Expressways_of_China`, 2017.

[35] Xu, F., Zhang, P., and Li, Y. Context-aware real-time population estimation for metropolis. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2016), ACM, pp. 1064–1075.

[36] Xu, X., Zhang, P., and Zhang, L. Gotcha: a mobile urban sensing system. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems* (2014), ACM, pp. 316–317.

[37] Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., and Moscibroda, T. Mobility modeling and prediction in bike-sharing systems. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* (2016), ACM, pp. 165–178.

[38] Yuan, J., Zheng, Y., Xie, X., and Sun, G. Driving with knowledge from the physical world. In *Proceedings of the international conference on Knowledge discovery and data mining*, KDD '11.

[39] Yuan, N. J., Wang, Y., Zhang, F., Xie, X., and Sun, G. Reconstructing individual mobility from smart card transactions: A space alignment approach. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on* (2013), IEEE, pp. 877–886.

[40] Zhang, D., Huang, J., Li, Y., Zhang, F., Xu, C., and He, T. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the 20th annual international conference on Mobile computing and networking* (2014), ACM, pp. 201–212.

[41] Zhang, D., Zhang, F., and He, T. Multicalib: national-scale traffic model calibration in real time with multi-source incomplete data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2016), ACM, p. 19.

[42] Zhang, D., Zhao, J., Zhang, F., and He, T. comobile: Real-time human mobility modeling at urban scale using multi-view learning. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2015), ACM, p. 40.

[43] Zhao, M., Ye, T., Gao, R., Ye, F., Wang, Y., and Luo, G. Vetrack: Real time vehicle tracking in uninstrumented indoor environments. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems* (2015), ACM, pp. 99–112.

[44] Zhou, P., Chen, Z., and Li, M. Smart traffic monitoring with participatory sensing. In *SenSys* (2013).

[45] Zhou, P., Zheng, Y., and Li, M. How long to wait?: predicting bus arrival time with mobile phone based participatory sensing. In *Proceedings of the 10th international conference on Mobile systems, applications, and services* (2012), ACM, pp. 379–392.