Empirical Study of Connected Vehicle Data and Its Application to Traffic Measurement

by

Jared Mitchell Porter

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Mechanical Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Kameshwar Poolla, Chair
Professor Roberto Horowitz
Professor Francesco Borrelli
Dr. Alex Kurzhanskiy

Fall 2022

Empirical Study of Connected Vehicle Data and Its Application to Traffic Measurement

Jared Mitchell Porter

Abstract

Empirical Study of Connected Vehicle Data and Its Application to Traffic Measurement

by

Jared Mitchell Porter

Doctor of Philosophy in Engineering - Mechanical Engineering

University of California, Berkeley

Professor Kameshwar Poolla, Chair

Connected Vehicle data is an emerging technology that offers to broadly measure traffic at a scope not accessible with traditional traffic measurement systems. Using GPS and vehicles onboard data, we can gain a broader perspective of traffic. The current adoption rate of connected vehicles is near 2-3 percent, but offers wide-area coverage. The resulting sparsity of data offers new challenges in extracting useful measures of the traffic state. This work focuses on an empirical study of available connected vehicle data collected from the San Francisco and Los Angeles areas. We explore how the penetration rate of connected vehicles in these areas influence the fidelity of extracted traffic flow measurements. Leveraging the accuracy of the GPS samples, we explore separating data into lane-level distinctions. We developed penetration rate agnostic estimates of queue length distributions at intersections. Utilizing the repeatability of timed intersections, we derive spatiotemporal diagrams of a major throughway and connect them to a macroscopic fundamental diagram. Finally, we explore new accident risk calculations based on extracted maneuver level information. Connected vehicle data offers new and exciting insights that will continue to improve with increased penetration rate.

I'd like to dedicate my thesis to Professor Pravin Varaiya. You were instrumental to the completion of this work and are dearly missed. I most enjoyed our economic discussions over coffee or a walk through Tilden Park.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **vii**

1 Introduction **1**

2 Description of Datasets **5**
2.1 Wejo Connected Vehicle Dataset 5
2.2 Supporting Loop Detector Datasets 8
2.3 Accident Database - TIMS 9

3 Penetration Rate and Estimation of Traffic Flow **12**
3.1 Comparing to San Francisco PEMS 13
3.2 Comparing to Sensys Detector Data 23
3.3 Linear Projection of Sample Counts 25

4 Connected Vehicle Lane Identification **30**
4.1 Fitting Lane Positions 30
4.2 Viterbi Search 33
4.3 Results 35

5 Queue Length Estimation **42**
5.1 Analytical Model 43
5.2 Non-Parametric Estimation 44
5.3 Parametric Estimation 47
5.4 Average Queue Length 47
5.5 Numerical Simulation 48
5.6 Empirical Verification - Comparison with Detector Data 50
5.7 Additional Discussion & Conclusion 52

6 Macroscopic Properties from Timed Intersections **56**

6.1 Problems with Sparsity 56
6.2 Aligning Trajectories 57
6.3 Clustering Time Intervals 59
6.4 Estimating Macroscopic Properties 61
6.5 Macroscopic Fundamental Diagrams 63
6.6 Discussion 64

7 Maneuvers and Risk **68**
7.1 Maneuver-Level Crash Risk 68
7.2 Maneuver-Level Data 69
7.3 Results 69
7.4 Discussion 70

8 Concluding Remarks **74**

Bibliography **75**

List of Figures

2.1 Distribution of sampling frequency taken from trips recorded in SF in the month of September 2021 . . . 7
2.2 Distribution of vehicle's model year in the connected vehicle trace dataset. . . . 7
2.3 Number of segments and total miles in SF by Segment Type . . . 8
2.4 Boxplot Distribution of Average Detection Volume in SF by Segment Type . . . 8
2.5 SF snapshot of a single 15 min period captured by the Aggregated Dataset. Speed is illustrated by the color of the segment while number of detected vehicles is illustrated by the thickness of the line. . . . 8
2.6 SF snapshot of active trips captured by the Trace Dateset. . . . 8
2.7 Placement of PEMS Stations on SF Highways . . . 10
2.8 Location of intersections outfitted with Sensys Detectors . . . 10
2.9 Placement of Sensys Sensors at an intersection on Hwy 107 . . . 10
2.10 Map of recorded accidents involving left turns at an intersection in SF between the years of 2011-2020. Larger dots indicate multiple accidents at that location. 11

3.1 Example Aggregated Wejo Segment on NB 101 and the corresponding PEMS stations. . . . 14
3.2 1 week sample of NB101 segment from PEMS and Wejo datasets . . . 14
3.3 1 Week Sample of Wejo's penetration rate estimated using PEMS flow data. . . . 15
3.4 Autocorrelation Plot of Weekday Penetration Rate for NB101 Segment . . . 16
3.5 Scatter plot of NB101 weekday penetration rates by time of day . . . 17
3.6 Scatter plot of NB101 weekend penetration rates by time of day . . . 17
3.7 Traffic Flow vs. Connected Vehicle Penetration Rate . . . 18
3.8 Quarter-hourly average penetration rates of SF Highways . . . 18
3.9 Locations, I-280 and Hwy 101 . . . 20
3.10 Penetration Rate Correlation Matrix for I-280 and Hwy 101 locations . . . 20
3.11 Flow Rate Correlation Matrix for I-280 and Hwy 101 locations . . . 20
3.12 Correlation Locations, Hwy 101 . . . 21
3.13 Penetration Rate Correlation Matrix for Hwy 101 locations . . . 21
3.14 Penetration Rate BoxPlots by Location . . . 22
3.15 Penetration rate distribution plots on opposing directions . . . 22
3.16 Week to week total penetration for select highways in SF . . . 23
3.17 Flow rate capture setup for Sensys Detectors and Wejo Connected Vehicle Data. 24

3.18 Weekday quarter-hourly average penetration rate for LA arterial. 24
3.19 Weekend quarter-hourly average penetration rate for LA arterial. 24
3.20 Mixed quarter-hourly penetration rate profiles of LA arterials. 25
3.21 Penetration Rate Correlation Map 26
3.22 LA Intersection Penetration Rate Correlation 26
3.23 SF Freeway flow estimation error for different estimates of penetration rate . . . 28
3.24 LA arterial flow estimation error for different estimates of penetration rate . . . 28
3.25 Optimal weighted moving average window to smooth flow estimation on SF Freeways 28
3.26 Optimal weighted moving average window to smooth flow estimation on LA arterials 28
3.27 Smoothed flow estimation error for SF. 28
3.28 Smoothed flow estimation error for LA. 28
3.29 True flow vs smoothed estimated flow. 29

4.1 Projection of connected vehicle GPS points to segment x-y plane. 31
4.2 Fitting Gaussian Mixture Model to a road segment of NB 101. 31
4.3 GMM means from connected vehicle data on NB101 32
4.4 Naive assignment resulting in excessive lane changes 32
4.5 Sequence of lane states and observations 33
4.6 Trellis diagram for lane assignment HMM 34
4.7 Viterbi search through the trellis diagram 35
4.8 Viterbi lane assignment of trajectory shown in figure 4.4. 36
4.9 Viterbi lane assignments of all points. 36
4.10 Lateral position distribution from naive lane assignment 37
4.11 Lateral position distribution from Viterbi lane assignment 37
4.12 Lateral position distribution from naive lane assignment 38
4.13 Lateral position distribution from Viterbi lane assignment 38
4.14 Left lane change outflow compared to the net right side flow of a downstream fork. 39
4.15 Left lane change outflow compared to average lane speed. 39
4.16 Distribution of lane change locations out of each lane. 40
4.17 Speed of each lane compared to PEMS g-factor speed estimation. 41

5.1 Queue formation behind an intersection 43
5.2 Resulting QP fit of a smooth monotone distribution to a taken empirical distribution 46
5.3 Numerical simulation for the distribution of queue - length X (in meters) vs $8\ (m) \times N$ (number of cars). 48
5.4 Numerical simulation for penetration rate $\alpha = 0.5\%$ 49
5.5 Numerical simulation for penetration rate $\alpha = 1.5\%$ 49
5.6 Numerical simulation for penetration rate $\alpha = 5\%$ 49
5.7 Numerical simulation - the estimation error for queue length percentiles F_X^{-1} . . 49
5.8 Illustrating detection of queue discharge from max gap data. 51
5.9 Distribution estimations from detector data with varying max gap parameters. . 51

5.10 Queue Estimations during morning on Lomita. 52
5.11 Estimated queue length distribution on lane 2 (NB) for intersections on CA-107 from connected vehicles with $\alpha \simeq 3\%$ vs. queue length estimation from in-ground vehicle detection sensors. 53
5.12 The difference between estimated queue length percentiles F_X^{-1} using connected vehicles vs detectors. 54
5.13 Auto-Correlation of consecutive Queue Lengths 55

6.1 San Francisco, Upper 19th Avenue (CA-1) Map 57
6.2 All captured northbound trips on upper 19th from 4-5pm, 9/5/2021 57
6.3 Aligned NB 19th Avenue trajectories for weekdays, 3:15pm - 7pm 58
6.4 Stopped samples of vehicles before moving again along with estimated discharge shockwaves. 58
6.5 Binary splitting of time intervals into groups with similar traffic patterns 60
6.6 Results of separating groups of similar traffic patterns. 60
6.7 Calculated spillback probability for each intersection and grouping. 61
6.8 Average flow of separated 15 minute intervals. 62
6.9 Average density of separated 15 minute intervals. 62
6.10 Average velocity of separated 15 minute intervals 62
6.11 Histogram of Hwy 1 penetration rates for Weekdays, 5-7pm 63
6.12 Comparing the standard deviation of sample density and flow vs. the standard deviation of the penetration rate 64
6.13 Macroscopic Diagram for northbound upper 19th Avenue with a cycletime of 90 seconds. 65
6.14 Comparison of Macroscopic Diagrams for northbound upper 19th. Comparing cycletimes of 80 and 90 seconds. 65
6.15 Map of lower portion of 19th Avenue. 66
6.16 Spacetime clusters for lower 19th Avenue 100 second cycletimes 66
6.17 MFD comparisons for northbound on lower 19th Avenue. Compares 90 second vs 100 second cycletimes. 67

7.1 Number of accidents before 2016 vs number of accidents after 2016 70
7.2 Comparison of total through crashes (left) and through crash risk (right) 71
7.3 Comparison of total left-turn crashes (left) and left-turn crash risk (right) . . . 72
7.4 Comparison of total right-turn crashes (left) and right-turn crash risk (right) . 73

List of Tables

2.1 Slice of the trace dataset 6
2.2 Slice of segment description table 7
2.3 Slice of segment level aggregation 7
2.4 Slice of PEMS Detector Data 9
2.5 Slice of Sensys Detector Data 11
2.6 Slice of SF accident data from TIMS Database. 11

7.1 Top 5 intersections for through risk 71
7.2 Top 5 intersections for left-turn risk 72
7.3 Top 5 intersections for right-turn risk 73

Acknowledgments

Thank you Kameshwar for advising me all these years and accepting me back after I stepped away for a bit.

Thank you Pravin and Alex for advising and holding weekly meetings and discussions.

Thank you to the members of my committee, Dr. Kurzhanskiy, Prof. Borrelli, Prof. Horowitz, and Prof. Poolla for helping review my manuscript.

Thank you to Christopher Flores from Sensys Networks, and Jo Birch from Wejo for providing the connected vehicle data.

Thank you Akhil and Hamid for working with me on this research and helping inspire confidence in myself.

Thank you Emma for handling the stress and setting aside plans and outings. Thank you for believing in me and keeping me rested and fed.

Thank you Mom and Dad for continuing to have confidence in me and pushing me forward.

Thank you Matt, Adam, Chase for always being supportive siblings.

Thank you all members of the BCCI lab group. Our shared experience and mutual encouragement were invaluable.

Thank you Yawo and all the ME Department supporting staff and your work to guide me through the PhD program.

Thank you to all the friends and family who have always supported and encouraged me throughout this process.

Chapter 1

Introduction

Traffic Congestion and Safety Impact

Transit is an integral part of society serving as the backbone of an economy moving people and resources around. As populations and economies grow and expand, the need to understand and optimize the road network is clear. A consultant report from Centre for Economics and Business Research [50] estimated the cost of road congestion. The study reports the road congestion cost for the US in 2013 as $124 billion through increases in fuel, emissions,and time and an indirect increase in the price of goods. Further, it projects a 50% increase in costs by 2030 resulting in a cumulative cost of $2.8 trillion over 17 years. It is evident that the road network fails to meet traffic demand.

A survey conducted by U.S. Travel Association measures the impact of congestion on travel demand [29]. In 2018, Americans avoided an estimated 47.5 million road trips due to traffic congestion. Moreover, infrastructure investment is popular with travelers. A majority of travelers share willingness to pay more for less traffic.

The other major concern found among travelers was improvement to safety. In 2010, motor vehicles crashes incurred an economic cost of $242 billion and resulted in 33 thousand fatalities [9]. Cities today are adopting the Vision Zero philosophy. Vision Zero promotes improving the traffic network with a systems approach with the ultimate goal of reducing traffic fatalities and severe injuries to zero [1].

Why do We measure Traffic/Applications

In order to improve safety and mitigate congestion data must be collected and analyzed to understand where and how traffic infrastructure can be targeted. Data gathered is used to create and calibrate models that in turn predict how individuals and traffic behave as a whole. There are many applications of these models.

In traffic control, highway ramp meter and intersection signals are optimized to maximize

flow or prevent jams [44] [15] [40]. Economic externalities of congestion can be modeled and reflected in dynamic pricing of tolls[57] [33]. Knowing the traffic state and demand can be used to optimally plan routes for individuals or globally. [58]

In planning, different road designs can be simulated to understand the impact of each design [34]. Data collected long-term reveals daily demand patterns, congestion bottlenecks, or accident prone areas. Modeling of demand can help to optimize the location of transportation services [62]. The more data available, the more accurate and useful our models are.

Traffic Models

Traffic flow is generally broken up into microscopic and macroscopic levels of aggregation.

Microscopic models focus on characterizing individual behavior, describing how vehicles accelerate,brake, or change lanes. A widely used microscopic model is the Intelligent Driver Model (IDM) [53]. IDM models the acceleration of a vehicle as a differential equation dependent on the spatial and velocity gap of the car in front. These models tune a variety of parameters that summarize individual drivers. Parameters of IDM include maximum acceleration, comfortable deceleration, minimum gap distance, safe headway time, and desired velocity.

Car following models can be tuned by measuring traffic and then in turn be used to create more accurate simulations. Typically traffic simulations will employ multiple agent types with a specifically tuned car following model creating a heterogeneous make-up of traffic. Simulations also employ probability distributions that model user choices or routing [34]. Measurement of demand patterns over time build accurate distributions for modeling agent choices.

Macroscopic models aim to model larger scale traffic metrics that arise from the aggregate flow of many agents. Most macroscopic models reference the fundamental diagram of section of road. The fundamental diagram describes for a point in the road the relation of traffic flow (q - number of cars per unit of time), with traffic density (k - the number of cars per unit distance). It can also be described as the relationship of traffic velocity v with traffic density where all 3 metrics are related via the equation: $q(k) = kv(k)$.

Lighthill-Whitham-Richards models or first order models combine the fundamental diagram, where flow and speed are dependent on density, with a continuity equation to get a wave equation for traffic density[31][54][19]. These models can be used to predict the development of traffic waves and how traffic flow changes over space and time.

Studying macroscopic models involves collecting observations of traffic flow, density, speed, and waves.

Traditional Traffic Data Sources

One of the difficulties of studying traffic is obtaining high quality data that holds of complete picture of the traffic state. A complete state would involve knowing the position, velocity, acceleration of every vehicle on the road as well other factors like pedestrian and traffic signal states. One of the most complete measurements of traffic state involves recording the road from a high vantage point and analyzing the video feed. These cameras can be permanent installations mounted on buildings, intersection lights, or freeway signs. They are mainly used in more temporary surveying work [7] [39].

Problems with cameras involve their difficult installation and maintenance as well as the large amounts of data and processing required. Video data consumes large amounts of storage and processing can be a labor intensive endeavour. New AI techniques have been developed to automate the analysis of video. Such techniques were used with the Pnuema dataset [7] and by Street Simplified which conducts video surveys of intersection performance. Still, AI requires tuning and manual verification.

Stationary detectors are inductive loops buried in ground that detect when a vehicle is over them. They can be used to collect flow, density, and speed information at a single point in the road. Detectors are expensive to install and maintain. Thus they are generally reserved for select locations of interest [8] [11].

To get samples of the broader network, traditionally, probe vehicles were outfitted to record speed and position. These probe vehicle could be a set of taxis or specifically outfitted test vehicles. Cell phone data offers an opportunity for a broad view into the traffic network. The near ubiquitous ownership of a smartphone means everyone is carrying a GPS on hand. Unfortunately the availability of cell phone data is very limited or expensive to obtain.

Connected Vehicle Data

Connected vehicle data is an emerging technology that collects data from vehicles transmitting onboard computer data. Connected vehicles (CVs) report GPS traces, speedometer readings, acceleration, heading, and more. CV technology is equipped in modern consumer vehicles and so provide a very general sample of traffic on the network. The level of coverage is ideal for traffic state estimation whose fidelity will only increase with higher adoption. The technology also scales well as more connected vehicles on the road only requires scaling for the higher data volume.

CV data is readily available as aggregators have created a business model of packaging this data and reselling it. Wejo Group Ltd. and Otonomo Technologies Ltd are two such companies with $85M and $58M market caps. Although currently the sample of total vehicles on the road is low, the data available now is shown to be valuable in measuring traffic.

Thesis Outline

Partnered with Sensys, Inc, they obtained samples of connected vehicle data from Wejo Ltd. to analyze. This thesis highlights the results from analyzing these CV samples.

In Chapter 2, we describe the sample dataset and all the supporting datasets we use.

In Chapter 3 we explore the penetration rate of CVs and the effectiveness of linear projection for flow estimation.

In Chapter 4, we develop an algorithm for assigning CV traces to lanes.

In Chapter 5, we propose an estimation for queue length that is agnostic to the penetration rate.

In Chapter 6, we explore recovering a spatiotemporal diagram from sparse trajectories and relate back to a measurement of the Macroscopic Fundamental Diagram.

Lastly in Chapter 7, we present the results from measuring maneuver level risk at intersections.

Chapter 2

Description of Datasets

This work analyzes data obtained from connected vehicles. To confer with CV results we also rely on some support detector datasets. We also explore a new way on analyzing traffic accident data combining cv data with an accident database. This chapter offers a description of all datasets used in this work.

2.1 Wejo Connected Vehicle Dataset

The focus of this work is centered around connected vehicle datasets from Wejo Ltd. provided by Sensys Networks Inc.

Wejo collects GPS trace measurements of connected vehicles along with data recorded by the vehicle's onboard computer. Wejo forms contracts with car manufacturers to obtain this data. Their sample sets mainly contain the data of recently manufactured vehicles by General Motors(GM).

At various price points Wejo differentiates data products based of what information they report and granularity of data. We obtained 3 different samples.

- September 2019 - November 2019: Vehicle Trace Dataset along the corridors of CA-107 and CA-1 in the southwest corner of Los Angeles County
- September 2021: Vehicle Trace Dataset covering San Francisco County.
- April 2021 - October 2021: 15-minute Aggregated Dataset covering San Francisco County, southwest Los Angeles County, Sacramento County, and the City of Arcadia.

Trace Dataset

Vehicle measurements are provided at a sampling time of 3 seconds. Each vehicle is equipped with "enhanced" GPS and by Wejo's claims provides latitude/longitude coordinates with 95% accuracy within a 3m radius. This should enable some degree of lane-level analysis we will explore in later sections. In addition to coordinates, the trace level data provides speed readings from the vehicles speedometer and bearing relative to north.

Every point is also assigned to a unique trip identification number. Trip ids are assigned to vehicles from key-on, ignition, to key-off. Table 2.1 shows a slice of the trace data. For the month of September 2021, there are 930,000 individual trips recorded in San Francisco resulting in 257 million points.

timestamp	trip_id	latitude	longitude	speed	heading
2021-09-14 00:00:01	464724	37.761527	-122.426034	33.6	356.7
2021-09-14 00:00:04	464724	37.761781	-122.426054	41.6	356.2
2021-09-14 00:00:07	464724	37.762105	-122.426088	43.84	356.6

Table 2.1: Slice of the trace dataset

From Figure 2.1 we see that the sampling time is fairly stable at every 3 seconds. The dropping of samples resulting in 6 and 9 second sampling times is not significant.

The September 2021 dataset also includes the vehicle make, model, and year. Data was mostly collected from Buick, Cadillac, Chevrolet, and GMC makes. Figure 2.2 shows the distribution of model years. Note that 25% of vehicle information was unknown. This figure shows that connected vehicle data is an emerging technology as most data is collected from newer vehicles and no data from any vehicle manufactured before 2015. As the technology is more widely adapted and older vehicles are replaced, the sample ratio of connected vehicles to total vehicles on the road is expected to increase.

15 Minute Aggregated Dataset

A cheaper offering of connected vehicle data creates 15 minute summaries of CV's traveling through a specific road segment. As shown in Table 2.2 the road is partitioned into a set of segments that generate the road network of interest. Segment traces and type identification are consistent with edges downloaded from OpenStreetMap.

The aggregated dataset provides 15 minute summaries of each segment with the total number of detected vehicles, average speed of samples, and the average stationary time for each vehicle. In addition to the summation of trace data, the aggregated set also provides additional count of samples where hard braking or hard acceleration are detected.

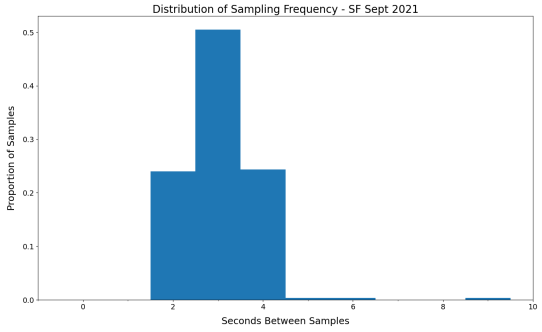


Figure 2.1: Distribution of sampling frequency taken from trips recorded in SF in the month of September 2021

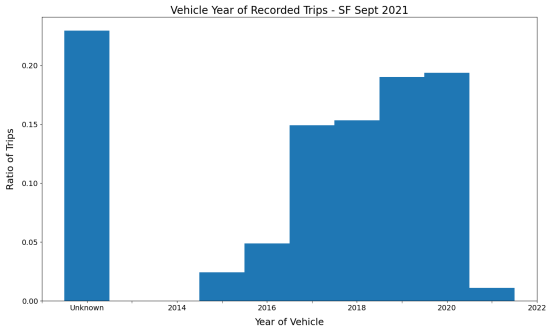


Figure 2.2: Distribution of vehicle's model year in the connected vehicle trace dataset.

seg_id	seg_road_name	seg_type	seg_heading	seg_len_miles	seg_trace
6812	divisadero street	Secondary	350.875671	0.063723	{geojson}
6813	jackson street	Residential	262.798843	0.091249	{geojson}
6814	fillmore street	Tertiary	168.618691	0.016214	{geojson}

Table 2.2: Slice of segment description table

seg_id	timestamp	car_count	avg_speed	avg_stat_time	hard_accel	hard_brake
1000	2021-04-04 10:00:00	35	55.4	0.0	0	0
1000	2021-04-04 10:15:00	25	55.8	0.0	0	0
1000	2021-04-04 10:30:00	35	52.6	0.0	0	0

Table 2.3: Slice of segment level aggregation

Figure 2.3 illustrates the break down of segment type by total number and miles in the city of SF. Figure 2.4 shows the breakdown of traffic flow by segment type. As expected, major through-ways make up the least number of segments but account for the majority of flow.

Snapshot Comparison

Figure 2.5 and 2.6 illustrate snapshots of the summarized aggregate dataset and the trace datasets. In figure 2.5 segments are highlighted based of the number of detected vehicles and average speed of the segment. Thicker segments illustrate higher flow, while the color shows the speed. The aggregate grants a nice large picture view of the city state and can be used to track trends in segments of long periods of time. However, the aggregate is derived from the trace set and so there is a clear loss of information that would otherwise prove

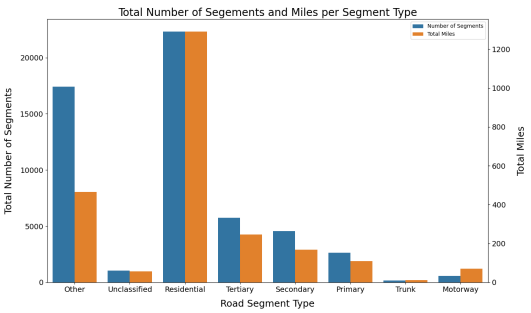


Figure 2.3: Number of segments and total miles in SF by Segment Type

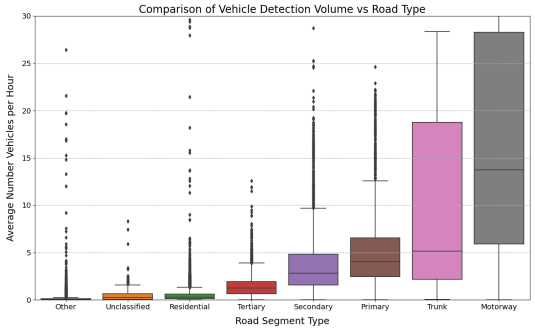


Figure 2.4: Boxplot Distribution of Average Detection Volume in SF by Segment Type

to be useful. Maneuvers, origin/destination, path choice, stop position, lane position, and acceleration are found in the trace set and lost in the aggregate. Figure 2.6 shows a pause in the animation of the trace set.

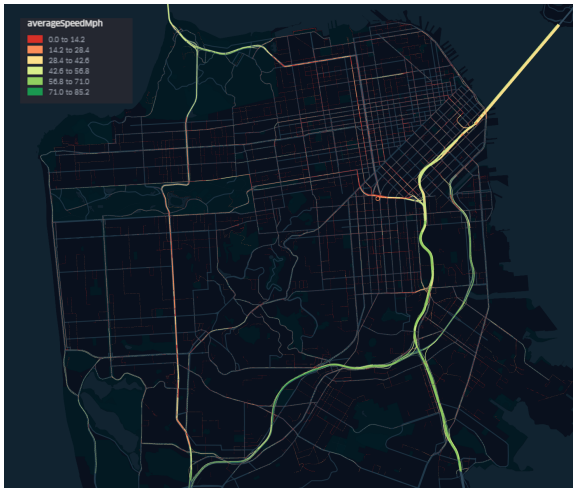


Figure 2.5: SF snapshot of a single 15 min period captured by the Aggregated Dataset. Speed is illustrated by the color of the segment while number of detected vehicles is illustrated by the thickness of the line.

Figure 2.6: SF snapshot of active trips captured by the Trace Dateset.

2.2 Supporting Loop Detector Datasets

Connected vehicles have a limited and variable sampling of the general driving population. To understand the impact of the penetration rate and the accuracy of connected vehicle measurements we compare CV results to available detector datasets. Vehicle Detector Sensors

have widespread and long term established use measuring traffic flows. In ground sensors detect whether or not a vehicle is over them or not. These provide accurate counts of vehicles crossing over them.

CalTrans PEMS Dataset

CalTrans currently collects and publishes detector data covering California's major freeways. The Freeway Performance Measurement System (PEMS) aggregates detector data and provides 5 minute reports. The primary direct measurements are flow, the number of vehicles in 5 minutes, and occupancy, the percentage of time detectors have a vehicle over them. From these measurements estimates of speed are produced using a g-factor estimation algorithm. The system also has algorithms to detect sensor health and report on the number of healthy sensors with the "percent observed" value. In the event of unhealthy detectors values are imputed using best estimates from healthy detectors and past values [8]. Individual lanes as well as flow weighted aggregates of all lanes are reported. Figure 2.7 shows the placement of PEMS detectors on highways running through SF.

station	timestamp	num samples	pct observed	flow	occupancy	speed
401409	09/01/2021 12:00:00	30	75	503.0	0.1332	57.6
401409	09/01/2021 12:05:00	30	75	527.0	0.1416	55.2
401409	09/01/2021 12:10:00	30	75	550.0	0.1346	56.7

Table 2.4: Slice of PEMS Detector Data

Sensys In-Ground Detectors

Sensys Networks, Inc. provided data from September 2019 - November 2019 to compare to the Los Angeles corridor CV trace data. Sensys has many intersections on Hwy 1 and Hwy 107 outfitted on the outflows of legs of intersections. Figure 2.8 shows the location of monitored intersections and Figure 2.9 illustrates the typical sensor placement for these intersections.

The Sensys data reported gave the up and down times of triggered detection sensors. We use this information to get total outflows out of intersection directions and the temporal spacing between vehicle detections.

2.3 Accident Database - TIMS

The last dataset that we utilized is accident data obtained by the Transportation Injury Mapping System (TIMS) [52]. The TIMS database collects records of traffic accidents reported by California Highway Patrol. Data entry takes some time to finalize so the most

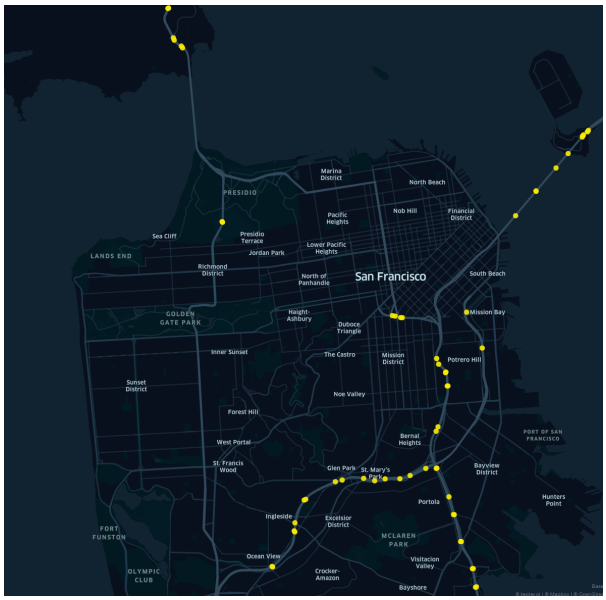


Figure 2.7: Placement of PEMS Stations on SF Highways

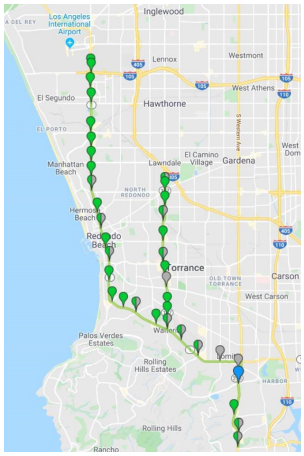


Figure 2.8: Location of intersections outfitted with Sensys Detectors

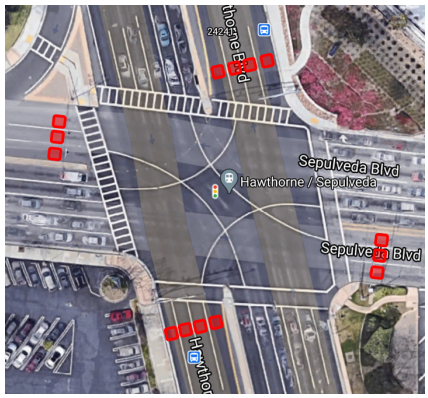


Figure 2.9: Placement of Sensys Sensors at an intersection on Hwy 107

recent years, 2021 and 2022 are not available. From the database we obtain accident data of SF from the years 2011 through 2020. TIMS provides many details on reports relating to severity of accident, injuries involved, the involvement of bicycles/pedestrians, road conditions, etc. We mainly utilize this to filter accidents that occur at intersections with specific "movements that precede the accident". Table 2.6 grabs a slice of the database giving the location, date, and maneuver code of accidents. Figure 2.10 plots all the recorded "left-turn"

start	end	dt	sensor	direction	apcc
2019-09-01 00:00:10.039	2019-09-01 00:00:10.054	15.0	1	SB	6703
2019-09-01 00:00:21.882	2019-09-01 00:00:22.390	508.0	12	SB	6703
2019-09-01 00:00:21.789	2019-09-01 00:00:22.304	515.0	13	SB	6703

Table 2.5: Slice of Sensys Detector Data

accidents at intersections in SF.

CASE ID	COLLISION DATE	COLLISION TIME	MOVE PRE ACC	INTERSECTION	POINT X	POINT Y
9219287	2020-11-20	831	E	Y	-122.389687	37.786221
9225894	2020-12-28	1355	E	Y	-122.447540	37.781040
9242419	2020-12-14	1445	E	Y	-122.431580	37.769321

Table 2.6: Slice of SF accident data from TIMS Database.

Figure 2.10: Map of recorded accidents involving left turns at an intersection in SF between the years of 2011-2020. Larger dots indicate multiple accidents at that location.

Chapter 3

Penetration Rate and Estimation of Traffic Flow

Measures of traffic volume are important in assessing the general health of the road network and making network design and control decisions. Traditional method of counting traffic using in ground detectors provides high resolution, accurate measurements of traffic flow. However, installation and maintenance of these sensors is laborious and expensive. Thus, the number of installed sensors and availability of data is sparse and limited to major throughways.

There is great interest in collecting a broader view of the traffic network. Traditionally probe vehicle data came from specially equipped vehicles that captured their data [54]. Other technologies have been used to provide wide-area data. Authors of [10] estimate traffic flow from movement of cell phones from cell tower to cell tower. In [55], authors try to evaluate the effectiveness of using mobile device GPS data. The recent availability of connected vehicle data offers an exciting window into the traffic state.

In order to understand how representative the CV dataset is of traffic, we need to understand what percent of traffic, or penetration rate, it measures.There has not been much work exploring the penetration rates of emerging connected vehicle datasets. In [45], authors examine the spatiotemporal penetration rate of a probe vehicle dataset over 2 highway sections in Washington DC over the course of a week. They model total traffic volume as a negative binomial distribution conditioned on the number of sample vehicles, time and location. [28] calculate the average penetration rate for Wejo connected vehicle data across detectors in Indiana, Ohio, and Pennsylvania. They analyze select Wednesdays and Saturdays across 7 months. The authors of [61] developed a method for an unbiased estimate of penetration rate using only the connected vehicle data at signalized intersections.

In this chapter, we utilize existing detector datasets, PEMS and Sensys, to evaluate of the

penetration rates of San Francisco freeways and some major arterials of Los Angeles. We find that the connected vehicle sample is not random day to day, as the same vehicles remain in our dataset. This creates a sampling bias but also generates a periodicity to the penetration rate signal. We find that these signals can be highly localized preventing wide-area use of the signal, however the periodicity can be used to improve flow estimations locally.

3.1 Comparing to San Francisco PEMS

We are interested in analyzing the penetration rate within our connected vehicle datasets and how it changes with time of day and location.

$$\alpha_i(t_m) = S_i(t_m)/N_i(t_m) * 100 \quad \text{for } m \in [0, ..., n]$$

Where $\alpha_i(t_m)$ is the measured penetration rate at location i over the time interval $[t_m, t_{m+1})$. $S_i(t)$ is the number of connected vehicles detected and $N_i(t_m)$ is the total number of vehicles measured at location i in the interval $[t_m, t_{m+1})$. $\Delta t = t_{m+1} - t_m$ is equal to 15 minutes.

We use PEMS detector data [11] pulled from the CalTrans Database to obtain a measure of the total number of vehicles on the road at available locations in San Francisco. The PEMS Dataset provides data for all 5 min intervals but any missing data or bad detection data gets filled with imputed values. To get the best estimate of the ground truth flow rates combined with an adequate number of samples, we look only at 15 minute intervals where the percentage of observed samples from PEMS is greater than or equal to 80%.

Using the aggregated dataset we can use observations over the course of 6 months. Sometimes aggregated segments cover more than one PEMS station.

In the case of multiple PEMS stations per Wejo segment we take the maximum observed flow rate across PEMS sensors at that location:

$$N_i(t_m) = \max_j \quad \mathbb{1}_{\{O_{i,j}(t_m) \geq 80\}} * N_{i,j}(t_m)$$

Where $O_{i,j}(t_m)$ and $N_{i,j}(t_m)$ are the percentage of observed samples and total vehicle count for station j at location i. Measurements are dropped when no stations meet the observation percentage threshold.

Figure 3.1 shows a segment on Northbound Highway 101 where connected vehicle data is aggregated over 15 minute intervals. Points on the segment display the approximated location of PEMS stations associated with this segment.

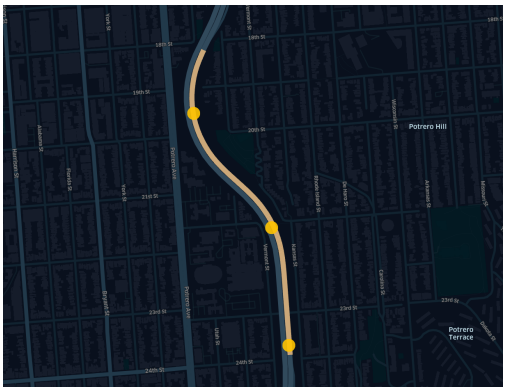


Figure 3.1: Example Aggregated Wejo Segment on NB 101 and the corresponding PEMS stations.

In figure 3.2 we see a sample week of measurements $N_i(t_m)$ and $\hat{N}_i(t_m)$ of the corresponding NB 101 segment. Figure 3.3 shows the calculated penetration rate $\alpha_i(t)$ for that week.

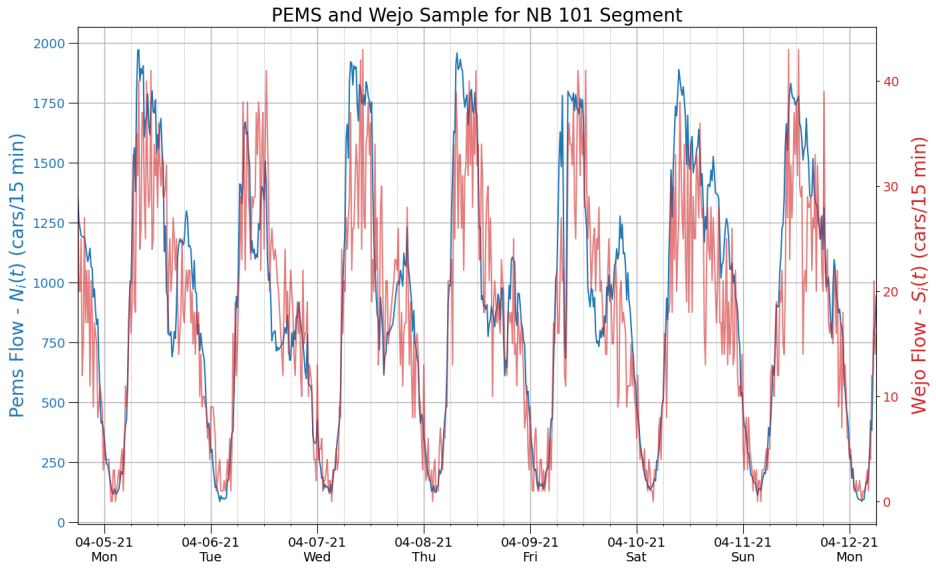


Figure 3.2: 1 week sample of NB101 segment from PEMS and Wejo datasets

Pen Rate Autocorrelation

From figure 3.3 we can see that there is a strong daily pattern for the penetration rate and that there is a difference between the weekday and weekend patterns. We can run an autocorrelation on the weekday portion of the penetration rate signal to highlight the time of

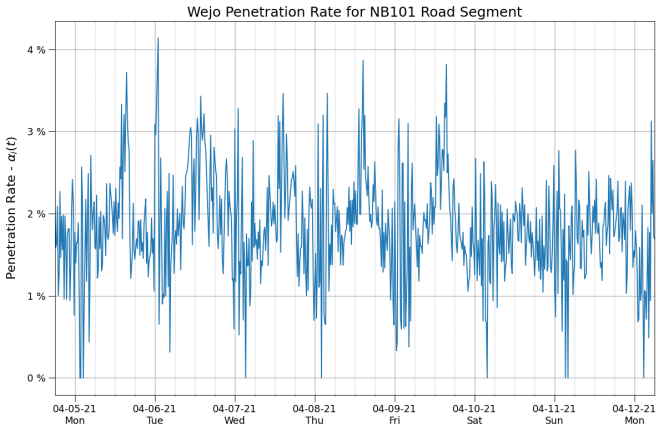


Figure 3.3: 1 Week Sample of Wejo's penetration rate estimated using PEMS flow data.

day correlation.

To do this we first modify the autocorrelation estimation to only look at the correlation across weekdays. Let $\mathbb{T}$ be all times that correspond to a weekday (Mon-Fri). We use an indicator function to average correlations where both the current and lagged interval are within a weekday.

$$\mathbb{1}_{\mathbb{T}}(m,k) = \begin{cases} 1, & \text{if } t_m, t_{m+k} \in \mathbb{T}, \\ 0, & \text{otherwise} \end{cases}$$
$$\mathbb{1}_{\mathbb{T}}(m) = \mathbb{1}_{\mathbb{T}}(m,m)$$

We calculate the weekday average and variance:

$$\mu_{\mathbb{T}} = \frac{\sum_{m=1}^{n} \mathbb{1}_{\mathbb{T}}(m)\alpha_i(t_m)}{\sum_{m=1}^{n} \mathbb{1}_{\mathbb{T}}(m)}$$

$$\sigma^2_{\mathbb{T}} = \frac{\sum_{m=1}^{n} \mathbb{1}_{\mathbb{T}}(m)(\alpha_i(t_m) - \mu_{\mathbb{T}})^2}{\sum_{m=1}^{n} \mathbb{1}_{\mathbb{T}}(m)}$$

We estimate the autocorrelation coefficient using the following formula and plot the results in figure 3.4:

$$\hat{R}(k) = \frac{1}{\sigma^2_{\mathbb{T}} \sum_{m=1}^{n-k} \mathbb{1}_{\mathbb{T}}(m,k)} \sum_{m=1}^{n-k} \mathbb{1}_{\mathbb{T}}(m,k)(\alpha_i(t_m) - \mu_{\mathbb{T}})(\alpha_i(t_{m+k}) - \mu_{\mathbb{T}})$$

Also plotted are the 95% confidence interval for a white noise process:

$$\mathrm{CI}(k) = \pm 1.96\Big(\sum_{m=1}^{n-k} \mathbb{1}_{\mathbb{T}}(m,k)\Big)^{-1/2}$$

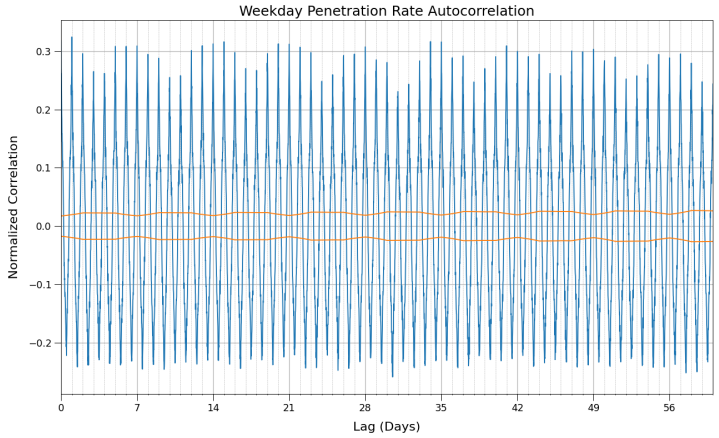


Figure 3.4: Autocorrelation Plot of Weekday Penetration Rate for NB101 Segment

From 3.4, we can see there is a strong periodicity to the penetration rate signal. This suggests a need to adjust any estimation from the connected vehicle data, to time of day.

Time of Day Profile

To examine the variability across days, we analyze the scatter of penetration rates that occur different times of day. We compare samples that occur at the same 15 minute interval in a day on either weekdays or weekends. Let τ_q specify the set of t_k's that occur at a certain time of day, for example, 8:00 - 8:15am weekdays. Define the quarter-hourly average for location i as such:

$$Q_\alpha(\tau_q, i) = \frac{1}{\sum_{t_k \in \tau_q} 1} \sum_{t_k \in \tau_q} \alpha_i(t_k)$$

From figures 3.5 and 3.6 we see the daily profile of a segment of NB101. They show the scatter of penetration rates for different times of day along with the average and standard deviation. We note that the late evening and early morning hours generate a wider spread and the average is a more chaotic. Then during more typical driving hours, the spread tightens up and the average takes a more concave regular shape. This is characteristic of other locations as well.

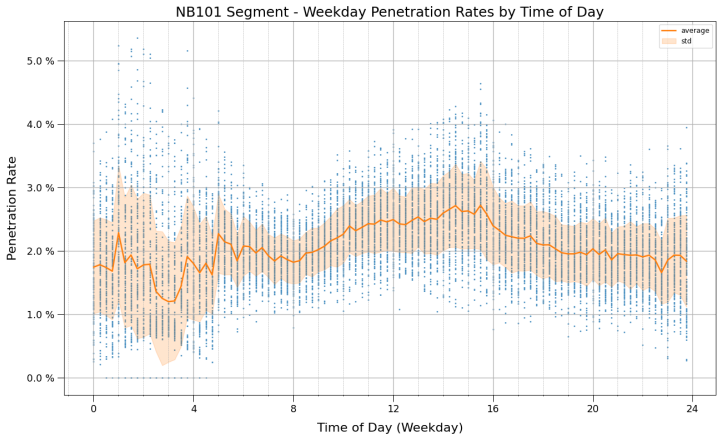


Figure 3.5: Scatter plot of NB101 weekday penetration rates by time of day

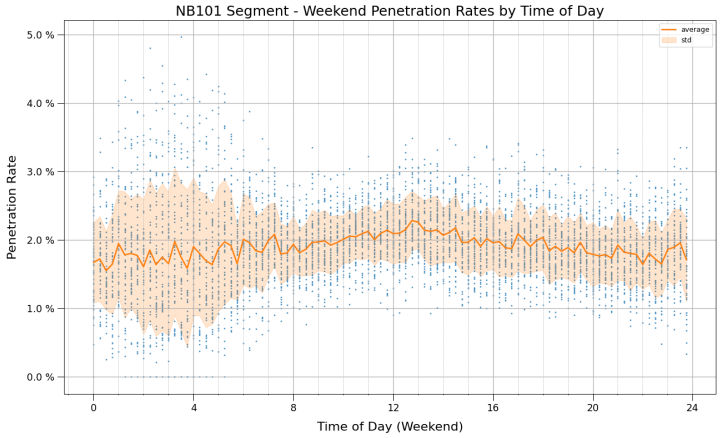


Figure 3.6: Scatter plot of NB101 weekend penetration rates by time of day

The more chaotic nature of early morning times can be attributed to high variance with low flow and the sampling bias, using the same set of connected vehicles everyday. Fig 3.7 plots the scatter points for pairs of observed flow $N_i(t_k)$ and penetration rate $\alpha_i(t_k)$ for all locations and times, i, t_k. The variance in penetration rate it clearly larger for measurements of low traffic volume. This is expected, at a flow of 100 vehicles per hour, a difference in 1 vehicle detection creates a 1% swing. Where as at 1000 vehicles per hour, a difference of 1 vehicle is only .1%.

Combine this with the sampling bias created by observing the same set of vehicles everyday. Connected vehicle measurements will capture the daily driving habits of the vehicles in the set. For example, say on average segment i witnesses the same 2 vehicles between 5-5:15am. Say car 1 does not travel on Wednesdays. That single car at a total flow of 100 vehicles

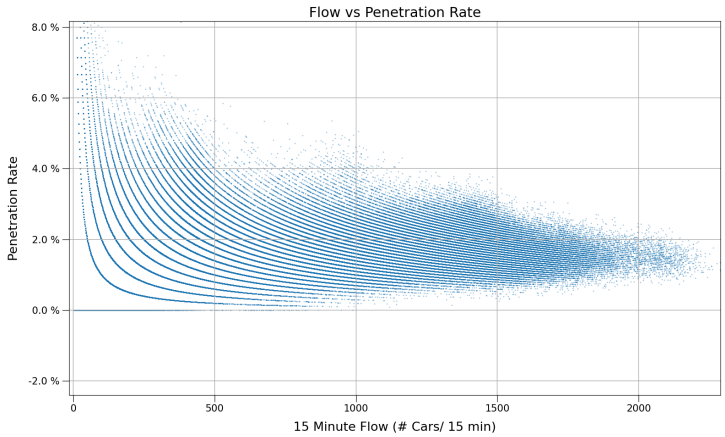


Figure 3.7: Traffic Flow vs. Connected Vehicle Penetration Rate

causes a 1% swing for Wednesdays on that segment/interval. In times of low flow the average penetration rate is very sensitive to day-to-day behavior of the sample set and creates a non-smooth time of day profile.

From figure 3.8 we see average penetration profiles on weekdays for various locations. Each location exhibits a chaotic rate for low flow hours to a more well defined trend for working hours. The plot also shows how locations exhibit similar behavior but produce there own individual penetration rate profiles. This highlights the need to calibrate for location. We'll next look at how well locations correlate with each other.

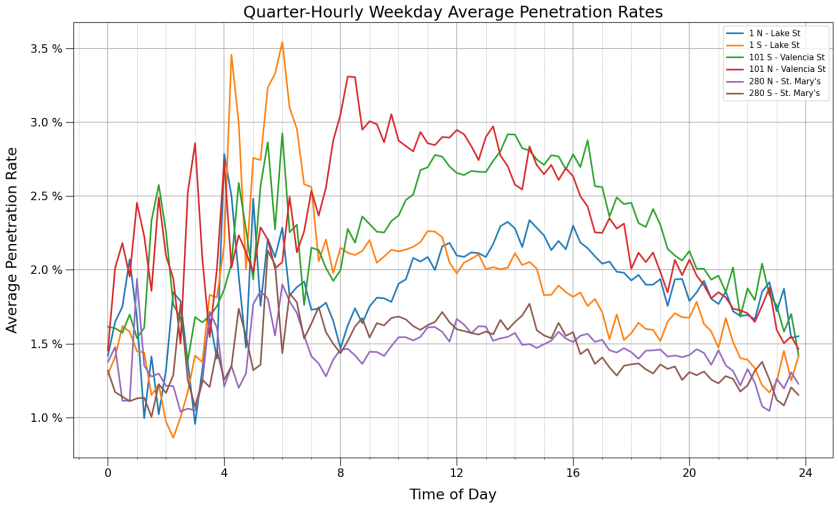


Figure 3.8: Quarter-hourly average penetration rates of SF Highways

Spatial Correlation

In ground detector measurements do not exist everywhere and so it is useful to see how penetration rate changes with locations. We would like to know how useful the using the flow and penetration rate from a near buy in-ground detector would be for a location where we only have connected vehicle data. To assess, we look at the correlation between locations that we have detector data for. We calculate the Pearson correlation coefficient between two locations:

$$\rho_{i,j} = \frac{\mathbb{E}\left[(\alpha_i - \mu_i)(\alpha_j - \mu_j)\right]}{\sigma_i \sigma_j}$$

where μ_i and σ_i are the mean and standard deviation of the penetration rate at location i, $\alpha_i(t_k)$

We show two routes to exemplify the patterns seen across highways. The first route shows correlation starting from Interstate 280 heading north onto 101 and then turning around and heading back in the south direction. The locations are shown on the map in Figure 3.9 where the color of the highlighted segment matches the labels in the correlation matrix in Figure 3.10.

This first route demonstrates a couple typical patterns seen. We see neighbors on the same highway following the same direction are correlated well as shown by the brighter 2x2 squares along the diagonal. It is expected that vehicles in the sample set mostly stay on the highway. Further, looking at movement across an interchange, there is a slight correlation shown by the 4x4 areas on the diagonals. It is not as strong as direct neighbors but expected as some percentage of traffic splits choosing one highway vs the other. Lastly, opposing flows do not demonstrate a significant correlation in penetration rate illustrated by large dark 4x4 off-diagonals.

By comparing the penetration rate correlation matrix in figure 3.10 to the correlation of flow rates shown in figure 3.11, the effect of sampling a portion of traffic becomes apparent. Total flow shows a strong correlation across all locations where that is clearly not true with penetration rate.

The second route demonstrates the effect of encounter a major interchange while staying on the same highway. Shown in 3.12, we analyze one location on Hwy 101 below (Paul Ave.) the I-280 interchange and two locations above it (Faith and Vermont). Here we see the that running through a major interchange drops the penetration rate correlation between neighbors. This highlights that the for highway traffic the only useful information from neighbors

Figure 3.9: Locations, I-280 and Hwy 101

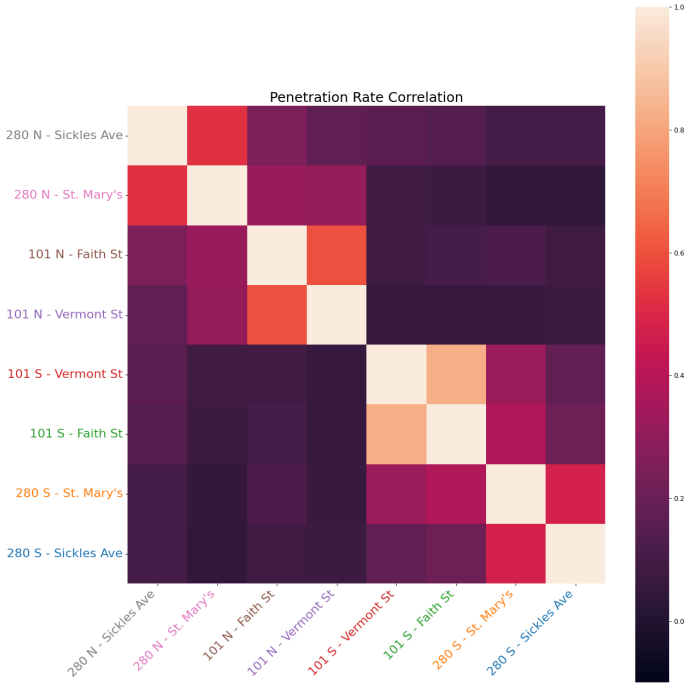


Figure 3.10: Penetration Rate Correlation Matrix for I-280 and Hwy 101 locations

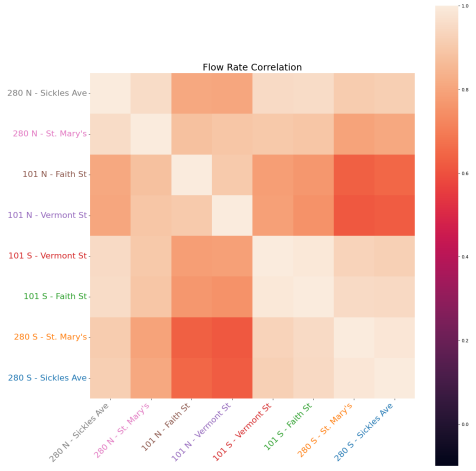


Figure 3.11: Flow Rate Correlation Matrix for I-280 and Hwy 101 locations

may be from those neighboring locations that are in the same direction between interchanges.

Figure 3.12: Correlation Locations, Hwy 101

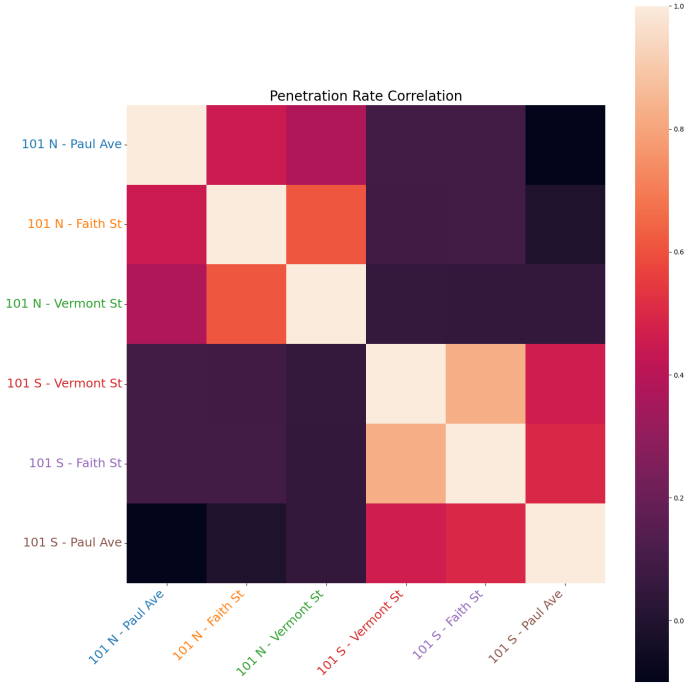


Figure 3.13: Penetration Rate Correlation Matrix for Hwy 101 locations

An interesting outcome between opposing directions, even though they don't share a significant correlation, they produce similar distributions of penetration rates. 3.14 shows the boxplots of all measured penetration rates for locations with more than 100 valid days of PEMS Data. The boxplots show the 4 quartiles of measured rates along with the means. Plotted below the box plot are corresponding average observed flow for each locations.

Some speculation as to what is happening. Opposing directions of traffic tend to experience different daily flow profiles but similar average flows. This could be do to commute round trips using one direction at some point in the day, more than likely, return to use the other direction.
From figure 3.7, the variance in penetration rate is related to the total flow. Thus, opposing directions with similar average flows should exhibit similar variance in penetration rate. The same commuter loop argument could be applied to the number of sampled vehicles observed in each direction and so opposing directions see similar averages of vehicle counts. This would give them similar penetration rate means. Variance and mean could be enough to characterize the distribution of penetration rates and so as seen in figure 3.15 opposing directions have closely matched distributions.

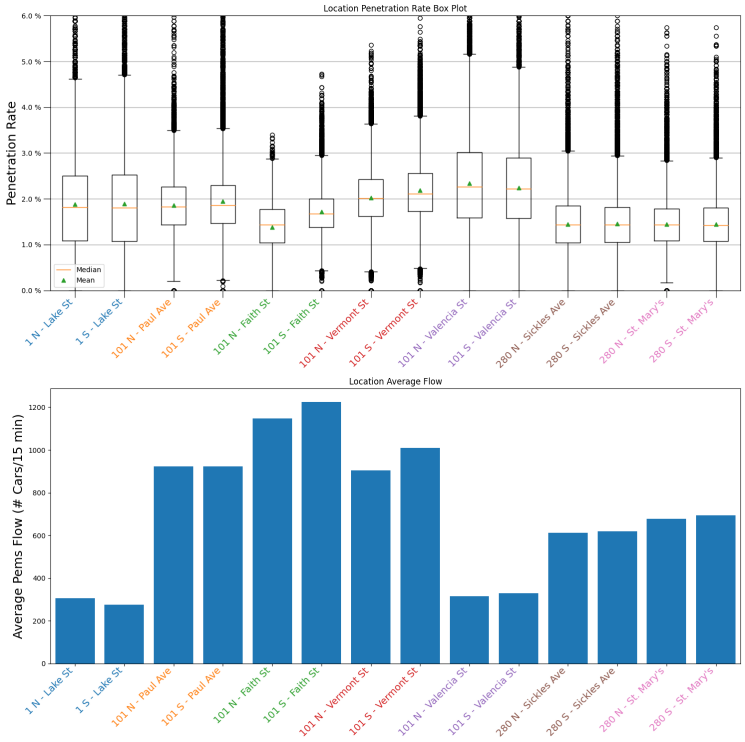


Figure 3.14: Penetration Rate BoxPlots by Location

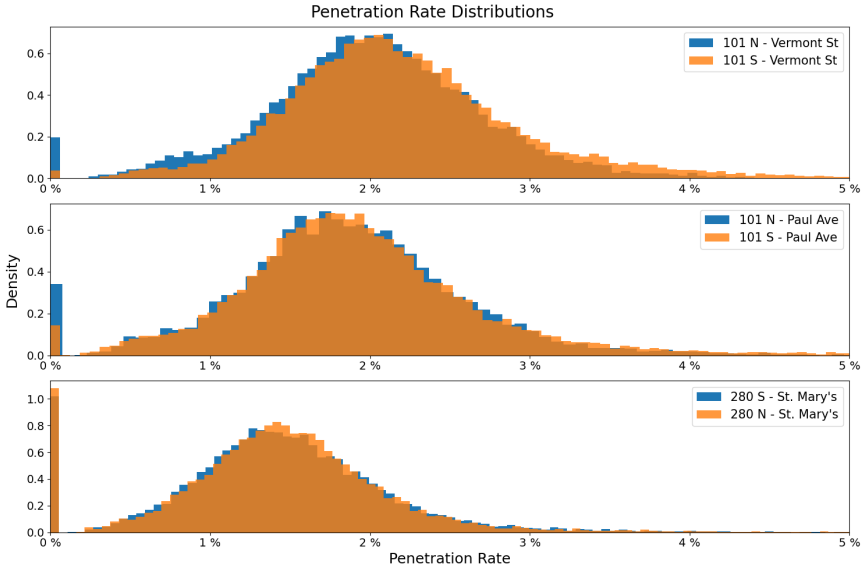


Figure 3.15: Penetration rate distribution plots on opposing directions

Weekly Total Penetration Trends

As more connected vehicles come online, it is expected that the penetration rates will increase over time. 6 months of data is a significant portion of time to investigate the trend.

Let τ_w specify the set of t_k's that occur in the week w. Then the weekly total penetration for location i is the total observed connected vehicles at location i divide by the total observed PEMS Flow:

$$W_\alpha(\tau_w, i) = \sum_{t_k \in \tau_w} \frac{\hat{N}_i(t_k)}{N_i(t_k)}$$

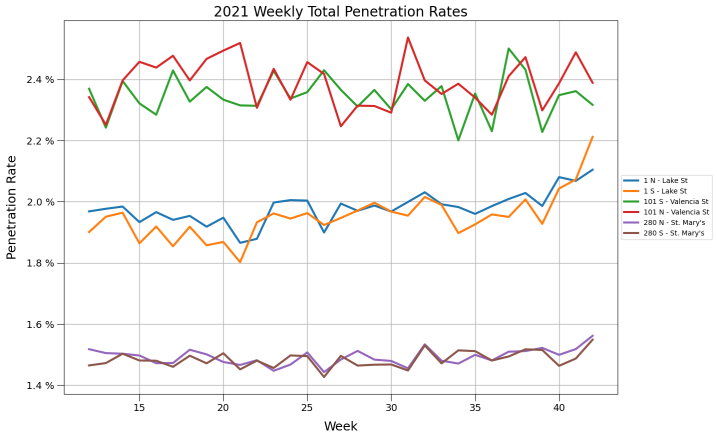


Figure 3.16: Week to week total penetration for select highways in SF

From figure 3.16 we there are a variety of week to week signals depending on location. At 101 Valencia area, pen rates can change dramatically from week to week with no general trend. In contrast, the 280 location maintains a relatively stable pen rate week to week with no significant trend. Last there is Hwy 1 location where pen rate stability is somewhere between the other two and it has a slight trend up.

3.2 Comparing to Sensys Detector Data

Caltrans PEMS data grants a window into how connected vehicle data is representative of the freeway network. Using the Sensys dataset we get a view into the arterial road network. Detectors are placed on major intersections on the Hwy 1 and Hwy 107 in southwest Los Angeles. Detector are on the outgoing lanes of each intersection. Figure 3.17 highlights the setup. The sum of vehicle counts in 15 minute intervals from in ground detectors across the red line are taken as the total flow on the network $N_i(t_k)$ Connected vehicle data is filtered

for the geofenced area after the intersection, shown as yellow scatter points. Unique vehicles seen in each 15 minute interval are our connected vehicle sample $S_i(t_k)$
We repeat the some of the same analysis as the PEMS Data.

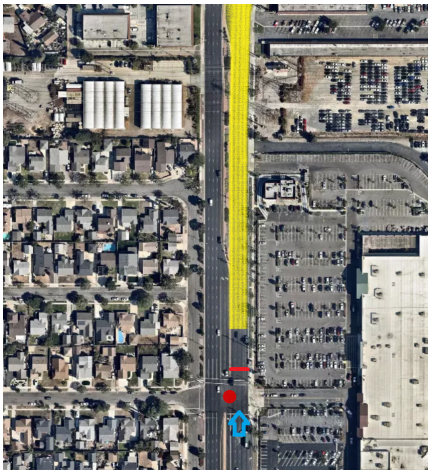


Figure 3.17: Flow rate capture setup for Sensys Detectors and Wejo Connected Vehicle Data.

Time of Day Profile

Figures 3.18 and 3.19 show the average quarter-hourly penetration rates for a sample intersection going northbound 107. Early morning hours are omitted because they spike wildly due to low flow and hide the detail of working hours. Note that the arterial street exhibits an average profile that is less smooth and has higher variance than the highway samples. This is mostly due to operating at lower flow rates. Also there is less of a concave trend moving through the day and the profile remains relatively flat during the busy hours of the day. This isn't just one isolated intersection.

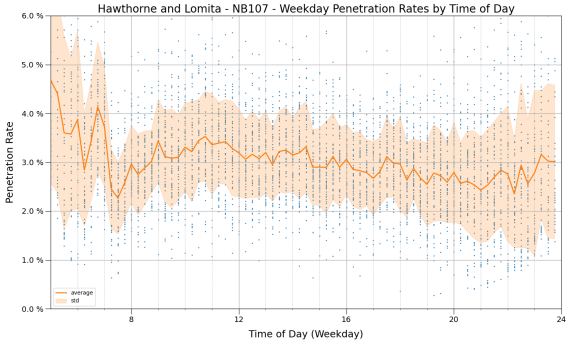


Figure 3.18: Weekday quarter-hourly average penetration rate for LA arterial.

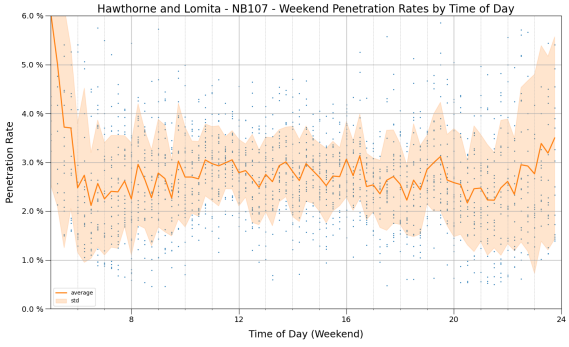


Figure 3.19: Weekend quarter-hourly average penetration rate for LA arterial.

Figure 3.20 shows the average profiles of multiple intersections on both Hwy 1 and Hwy 107. These samples are distantly spread out and they all exhibit similar profiles. This could suggest that a wide-area penetration rate is more applicable to the arterial streets. It is unclear why they exhibit a more flat profile as compared to the highway. Possibly, this is a consequence of the different composition of commercial vehicles and consumer vehicles on the different types of road network.

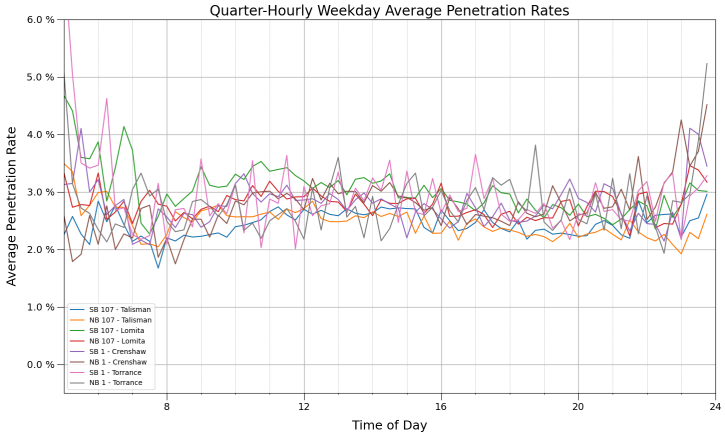


Figure 3.20: Mixed quarter-hourly penetration rate profiles of LA arterials.

Spatial Correlation

Looking at the correlation of penetration rates across Hwy 107, figure 3.22 shows that the drop-off in correlation on a major arterial is mostly related to distance. The largest break in correlation coefficients in the same direction is between the largest gap between intersections. Opposing directions again show no correlation.
Penetration rate correlation on these roads may not have any significance when the distribution of penetration rates is relatively the same across working hours and location. We can see the implication of this in the flow estimations made from connected vehicle data.

3.3 Linear Projection of Sample Counts

Ultimately, the goal is to be able to extract flow estimations from the connected vehicle dataset. The question becomes what characterization of the penetration rate do we need to get reliable information. We don't have ground truth everywhere, if we did we wouldn't need to use the CV data to measure it. However the ground truth is needed to understand what the penetration rate is to be able to project the connected vehicle sample to total flow.

The authors of [61] use estimates of queue position at signalized intersections to get estimates of the penetration rate solely from connected vehicle data. Other options include running

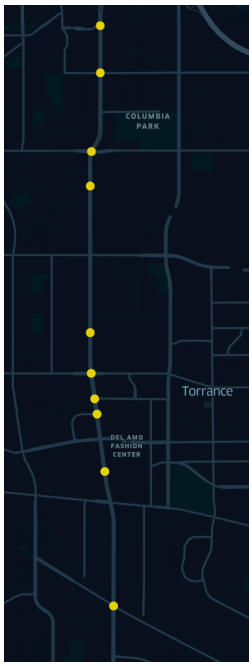


Figure 3.21: Penetration Rate Correlation Map

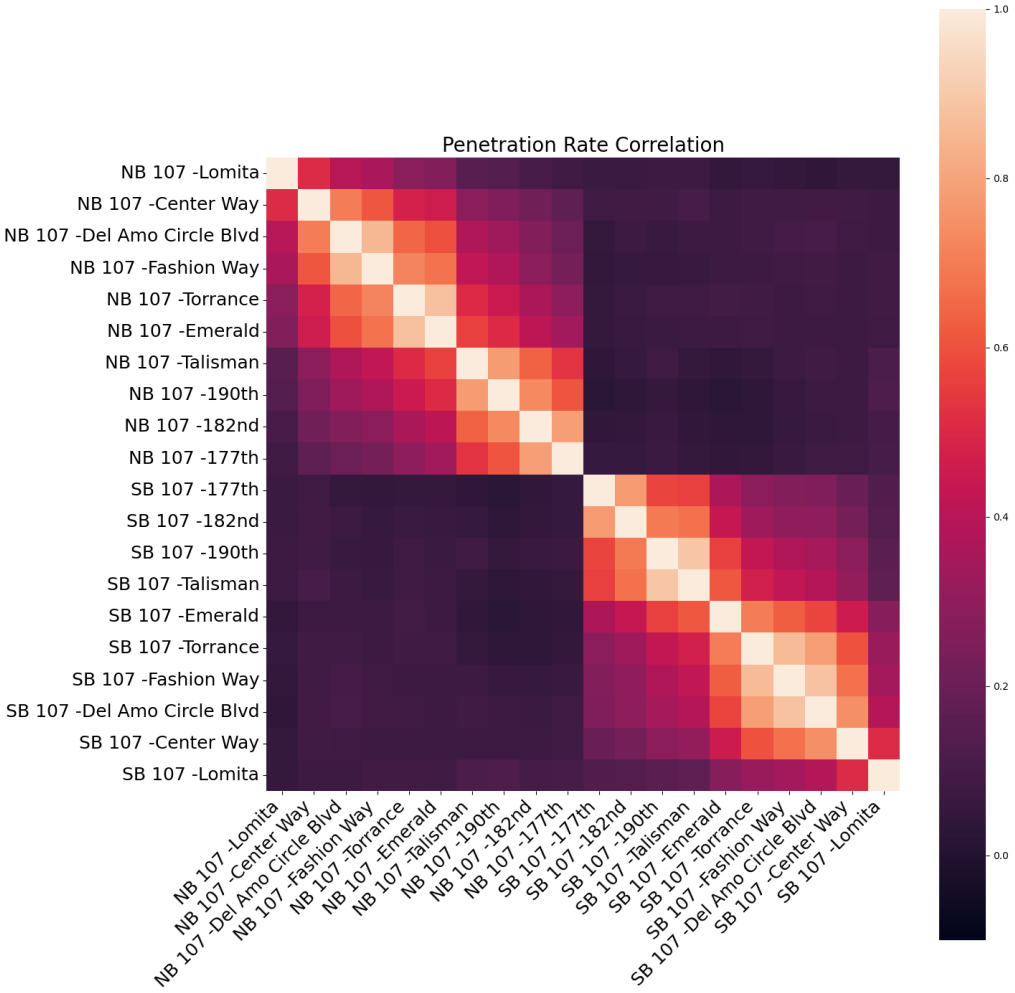


Figure 3.22: LA Intersection Penetration Rate Correlation

single day or week surveys. Here we explore given limited estimates of the penetration rate, how well can we estimate total flow from the connected vehicle data using linear projection.

For this analysis we will focus on estimation error generated for weekdays from 6am - 9pm.

Limited Penetration Rate Profiles

We estimate total flow of a location by linear projection, dividing the count of connected vehicles by an estimated penetration rate:

$$\hat{N}_i(t_k) = S_i(t_k)/\hat{\alpha}_i(t_k)$$

where we choose some estimation of the penetration rate for location i and time t_k, $\hat{\alpha}_i(t_k)$.

We compare five estimates of the penetration rate. Reminder that τ_q encapsulates a specific time of day, on weekends or weekdays. Say 8:15am on weekdays.

1 Day estimation of quarter-hourly penetration rate:

$$\hat{\alpha}_i(t_k) = \hat{Q}_d(\tau_q(t_k), i) = |\tau_q \cap \mathbb{T}_d|^{-1} \sum_{t_m \in \tau_q \cap \mathbb{T}_d} \alpha_i(t_m) \quad \text{where } \mathbb{T}_d = \{t_m | t_m \in \text{Survey Day}\}$$

1 Week estimation of quarter-hourly penetration rate:

$$\hat{\alpha}_i(t_k) = \hat{Q}_w(\tau_q(t_k), i) = |\tau_q \cap \mathbb{T}_w|^{-1} \sum_{t_m \in \tau_q \cap \mathbb{T}_w} \alpha_i(t_m) \quad \text{where } \mathbb{T}_w = \{t_m | t_m \in \text{Survey Week}\}$$

Full Dataset quater-hourly average penetration rate:

$$\hat{\alpha}_i(t_k) = \hat{Q}_\alpha(\tau_q(t_k), i)$$

Local Total Penetration Rate:

$$\hat{\alpha}_i(t_k) = \frac{\sum_{t_m} S_i(t_m)}{\sum_{t_m} N_i(t_m)}$$

Citywide Total Penetration Rate:

$$\hat{\alpha}_i(t_k) = \frac{\sum_i \sum_{t_m} S_i(t_m)}{\sum_i \sum_{t_m} N_i(t_m)}$$

We calculate the mean absolute percentage error on weekday working hours to compare results. Let $\mathbb{T}_{wkdy}$ be the set of all times from 6am-9pm Weekdays.

$$MAPE_i = \frac{\sum_{t_m \in \mathbb{T}_{wkdy}} |N_i(t_m) - \hat{N}_i(t_m)|/N_i(t_m)}{\sum_{t_m \in \mathbb{T}_{wkdy}} 1}$$

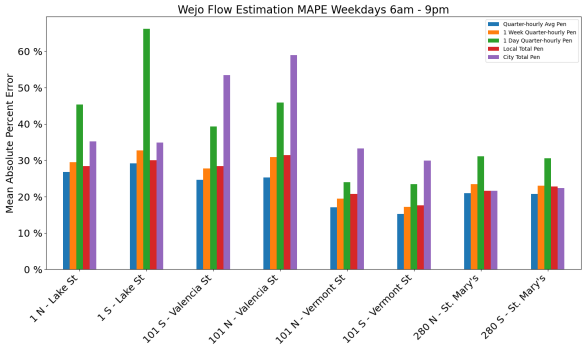


Figure 3.23: SF Freeway flow estimation error for different estimates of penetration rate

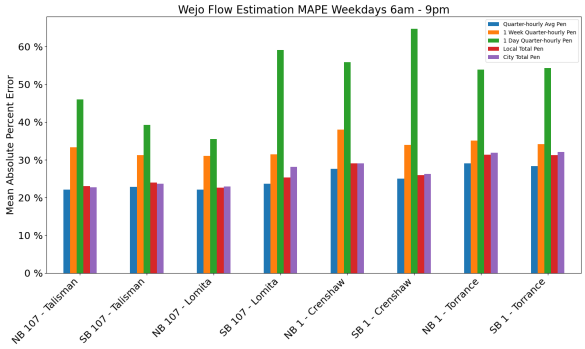


Figure 3.24: LA arterial flow estimation error for different estimates of penetration rate

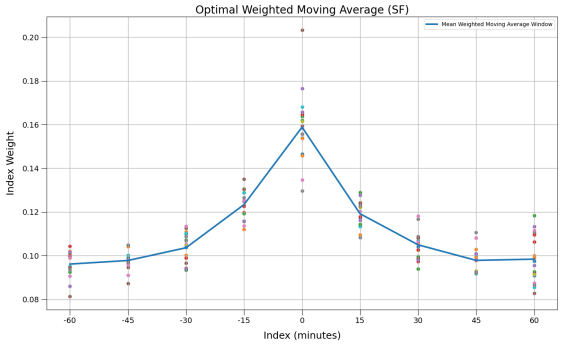


Figure 3.25: Optimal weighted moving average window to smooth flow estimation on SF Freeways

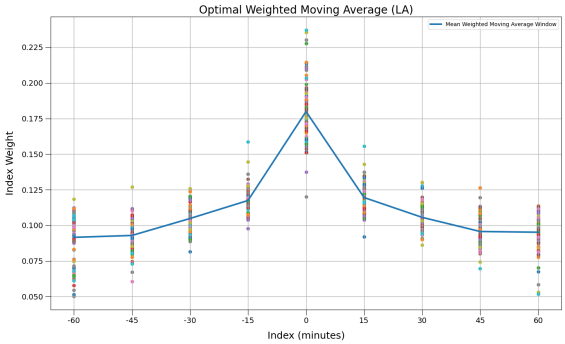


Figure 3.26: Optimal weighted moving average window to smooth flow estimation on LA arterials

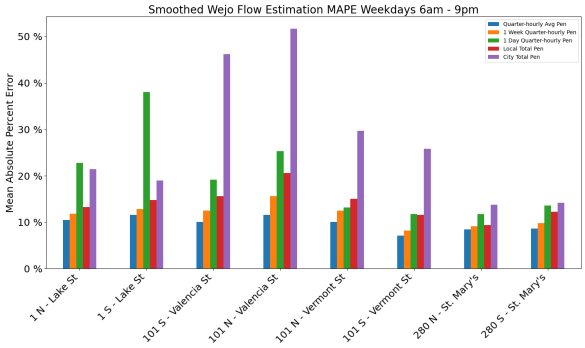


Figure 3.27: Smoothed flow estimation error for SF.

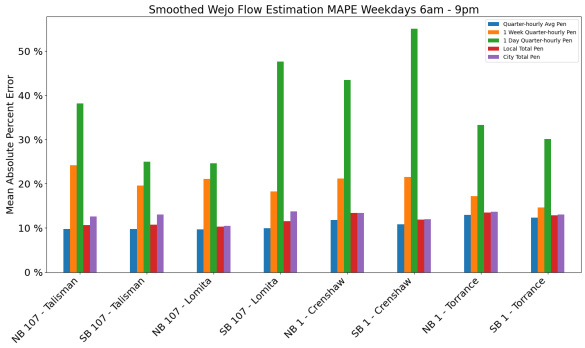


Figure 3.28: Smoothed flow estimation error for LA.

From figures 3.23 and 3.24 we see that the mean absolute percentage error is large for both SF and LA, even with a knowledge of the average penetration rate for each time of day.

However, connected vehicle data is quite noisy and can benefit greatly from some general smoothing. We applied a weighted moving average to the flow estimations and ran a linear program to calculate the optimal weights for each location. Figures 3.25 and 3.26 show the resulting moving average windows extending out 1 hour ahead and 1 hour behind. The optimal averaging window is close to an exponential window. Figures 3.27 and 3.28 show the smoothing significantly reduced the error percentage to below 10% with knowledge of the mean time of day penetration rate. Figure 3.29 shows the flow signal measured by detector data and the smooth projection of connected vehicle data.

It's clear that a one day survey is not enough information to get a reliable estimation of the penetration rate. However 1 week did very well predicting months of traffic flow from the connected vehicle sets. It is not clear why 1 week did much better on SF highways vs. LA arterials. It could be a factor of total flow with penetration rates having a narrower standard deviation on the SF highways. Interestingly the LA arterials all had similar pen rate distributions across location and time. Thus the flat average pen rate performed very well. It is possible that sampling of arterial roads is more homogeneous as in the LA case vs. motorways in the SF case. Distribution of commercial vehicles could be a factor. More work needs to be done on what factors affect the penetration rate signals seen across cities. As penetration rates increase estimation errors are expected to decrease, but understanding confounding factors will aid to reduce that error further.

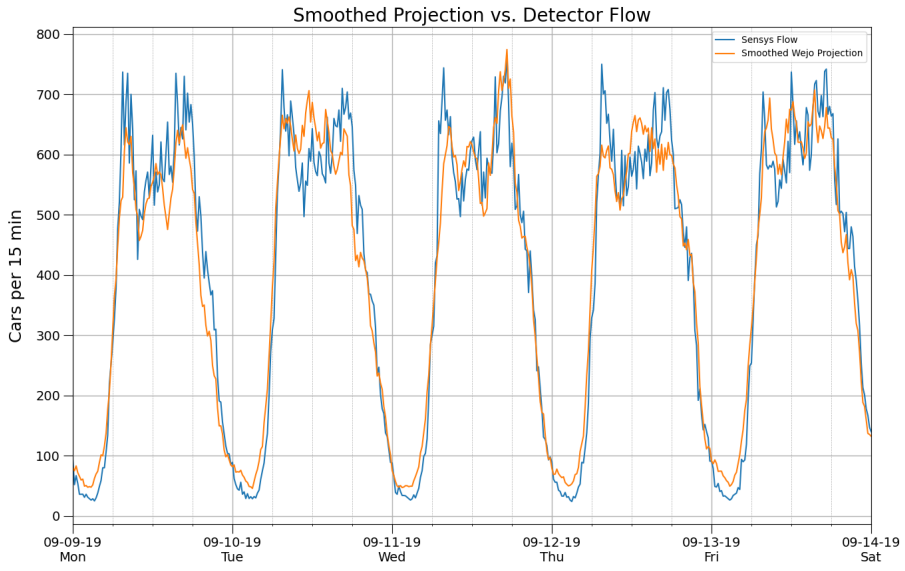


Figure 3.29: True flow vs smoothed estimated flow.

Chapter 4

Connected Vehicle Lane Identification

Connected vehicle data has potential to extract lane level information. If samples can be assigned to the correct lanes then we can extract valuable information on each lane: speed, flow, density, queue length, number of lane changes. Understanding where and when lane changes occur can help improve simulations of traffic flow when modeling agent decisions.

4.1 Fitting Lane Positions

It has been established that the distribution of vehicle positions across the road can be modeled as a Gaussian Mixture Model [13]. For a vehicle in a particular lane, aggregation of GPS inaccuracies and driver drift approximately create a Gaussian distribution of positions relative to the center of a lane. The combination of Gaussian distributions for each lane creates a mixture model.

For a particular segment we can fit a mixture model using connected vehicle data. Figure 4.1 demonstrates the filtering of connected vehicle data for a discretized segment of NB Hwy 101. With a limited segment length the road can be approximated as straight and the data projected onto an x-y plane. The y-axis here runs parallel to the flow of traffic and the x-axis perpendicular across the road. Taking the histogram of all GPS points projected onto the x-axis reveals the distribution of lane positions. From this we can fit a GMM model.

To approximate each lane as equal width we tie the variance for each lane's Gaussian and fix the number of lanes to a known number of lanes. We use the Expectation Maximization Algorithm to find a local maximum likelihood given the data:

$$\max_{\omega,\mu,\sigma} \prod_{j=1}^{N} p(x_j|\omega_i,\mu_i,\sigma) = \max_{\omega,\mu,\sigma} \prod_{j=1}^{N} \sum_{i=1}^{\ell} \omega_i \mathcal{N}(x_j|\mu_i,\sigma)$$

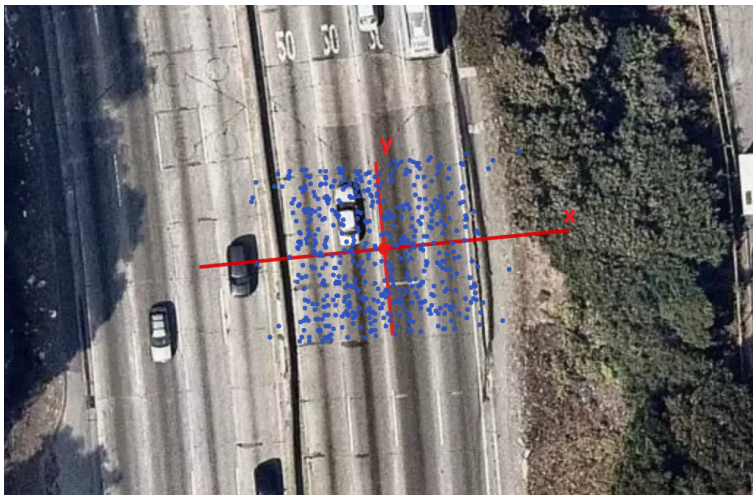


Figure 4.1: Projection of connected vehicle GPS points to segment x-y plane.

4.2 demonstrates the collection of connected vehicle data for a road segment and the fitted GMM distribution.

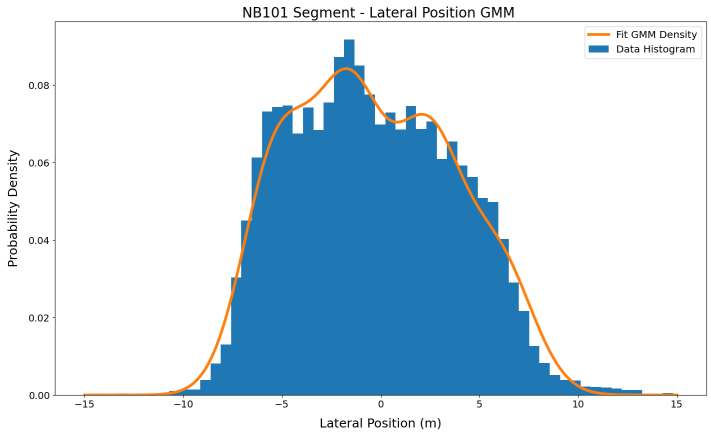


Figure 4.2: Fitting Gaussian Mixture Model to a road segment of NB 101.

For a corridor of interest we section the road into straight, discrete segments. For each segment we fit an individual GMM using the connected vehicle data. Each segment has a set of fit means approximating the center position of the lane and shared variance across the lanes. Figure 4.3 shows the resulting lane means of this process.

Thus, for any particular GPS trace point, we can bin its position to a localized segment, project onto the segments aligned axis, and predict which lane the sample belongs to by

Figure 4.3: GMM means from connected vehicle data on NB101

choosing the closest lane to the projected x-coordinate. Unfortunately, the GPS traces are not accurate enough and measurements from each lane bleed into each other. Therefore naively assigning based off proximity to each lane will result in many errors. In figure 4.4 we see the results of this assignment for one trip displaying an excess of lane changes.

Figure 4.4: Naive assignment resulting in excessive lane changes

4.2 Viterbi Search

To get a better approximation of lane assignment, we consider the entire trajectory of the vehicle. If a vehicle on a three lane road makes two significant movements to the left, we can make a reasonable guess that they started in the farthest right lane. The Viterbi algorithm has been used with noisy GPS traces to snap a full trip trajectory to the map network. (cite Map Snapping Resources) We use these ideas to snap noisy GPS measurements to lanes.

To estimate the total lane assignments for a whole trajectory we model the sequence of lanes as a hidden Markov model. For every sample point in a vehicle's trajectory, it's hidden state is the lane the vehicle is in, and the observed state is the measured projected x-axis position across the road. Modeling the HMM process and utilizing the Viterbi algorithm we can determine the maximum a posteriori probability estimate of the sequence of lanes. Let $s_k \in 0, 1, ..., L$ be the hidden sequence of assigned lanes for a road with up to L lanes. Let $o_k \in \mathbb{R}$ be the sequence of observed states as the projected horizontal distance from the center of the left most lane.

Consider figure 4.5 showing the sequence of 2 samples, k and $k+1$. Between the two samples the vehicle shifts lanes from 0 to 1, $s_k = 0$, $s_{k+1} = 1$. $\mu_{i,k}$ is the lane center for lane i relative to lane 0 ($\mu_{0,k} = 0$). The lane center is derived from the mean of the Gaussian Model fit at the location of observation k. σ_k is the shared variance of the GMM at k.

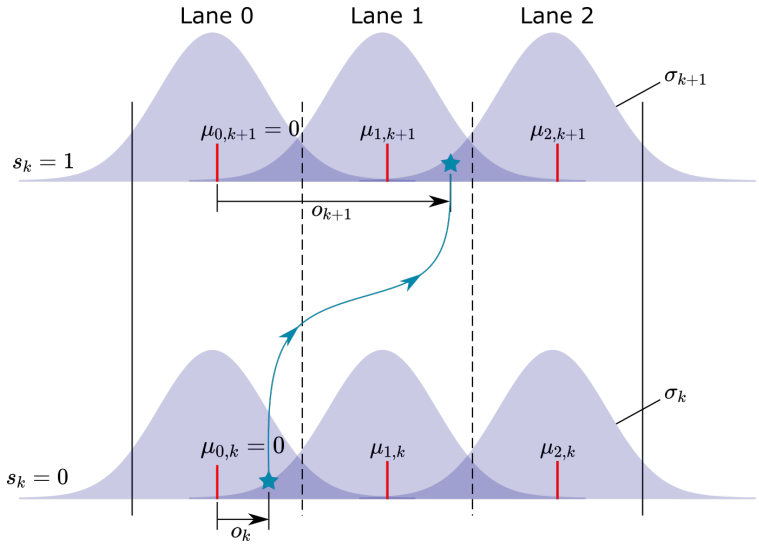


Figure 4.5: Sequence of lane states and observations

For the Hidden Markov Model we need to define the emission and transmission probabilities. The emission probability is the likelihood of an observation given we are in some state or $p(o_k|s_k = \ell)$. This probability will simply be the likelihood derived from our Gaussian

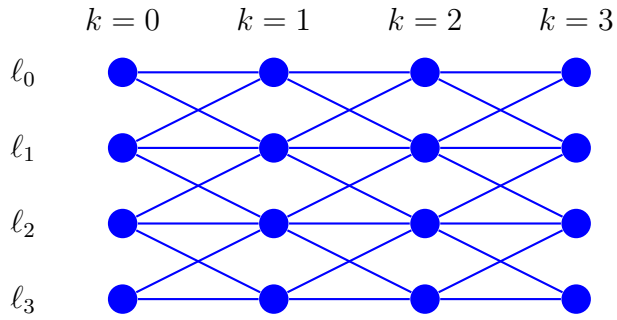


Figure 4.6: Trellis diagram for lane assignment HMM

mixture models.

$$p_e(o_k|s_k = \ell) = \mathcal{N}(o_k|\mu_{\ell,k}, \sigma_k) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left(\frac{1}{2}\left(\frac{o_k - \mu_{\ell,k}}{\sigma_k}\right)^2\right)$$

For the transmission probability we take inspiration from [38]. The authors use the Viterbi Search to snap GPS traces to the road map. Here they take the difference in distance of the mapped route to the greater circle distance between all GPS points. That difference is mapped to an exponential distribution. This acts as a filter for improbable paths.
Along that line of thinking we create a transmission probability that depends on the lateral distance traveled between samples. A lateral distance that is close in length between lanes should have a higher probability of transition. Where if the lateral distance traveled is close to zero then lane transition has a low probability. This will act to filter lane changes when a vehicle is consistently sampled near the border of two lanes.

The transmission probability from lane i to lane j from observations o_k and o_{k+1}:

$$p_t(j|i, o_k, o_{k+1}) \propto \frac{1}{\beta} \exp(-\beta\|(\mu_{j,k+1} - \mu_{i,k}) - (o_{k+1} - o_k)\|)$$

Consider the trellis diagram in figure 4.7. This represents all paths available when deciding the lane assignment (assuming only 1 lane change is possible between samples). The likelihood of any particular path is the product of all nodal emissions and edge transmissions through the path. The Viterbi Search Algorithm uses dynamic programming to calculate the path with the greatest likelihood. It does so by computing and retaining the optimal path to each node at every time step. At the last observation we pick the node with the greatest likelihood and backtrack through the path that led to that node.

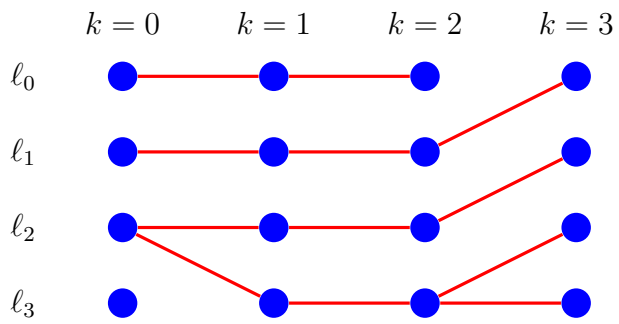


Figure 4.7: Viterbi search through the trellis diagram

Viterbi Algorithm:

$$
\begin{aligned}
V_0(j) &= p_e(o_0|s_0 = j) \\
V_{k+1}(j) &= \max_{i \in L} \quad V_k(i) p_e(o_{k+1}|s_{k+1} = j) p_t(j|i, o_k, o_{k+1}) \\
\bar{s}_{k+1}(j) &= arg\max_{i \in L} \quad V_k(i) p_e(o_{k+1}|s_{k+1} = j) p_t(j|i, o_k, o_{k+1})
\end{aligned}
$$

In our implementation, emission probabilities are determined by closest empirical Gaussian mixture model of lane positions, and transmission probabilities are tuned using the parameter β. With no real ground truth to compare and optimize the β parameter to, β is selected for what yields subjectively reasonable results. Setting β too high and the algorithm is too strict on transitions and lane changes are left undetected. Setting β too low and lateral movement becomes irrelevant and the Viterbi algorithm produces the same results as matching each point to the closest lane center. We found that $\beta = 1.5$ produced reasonable results between these two outcomes.

In figure 4.8 we see the reassignment of lanes from the Viterbi algorithm over the naive assignment seen in 4.4. Because the trajectory runs close to the lane boundary, many single noisy samples are interpreted as a change of lane. Here the additional transmission probability acts to smooth the lane change detections.

4.3 Results

The results of lane assignment via the Viterbi algorithm are displayed in Figure 4.9. To get a sense of improvement over assignment of the closest lane, we plot the distributions of lateral positions for each lane. We see from Figure 4.11 that assigning from Viterbi search results in lane distributions that more closely approach the fit Gaussian Mixture Models as lane distributions are allowed to overlap.

Figure 4.8: Viterbi lane assignment of trajectory shown in figure 4.4.

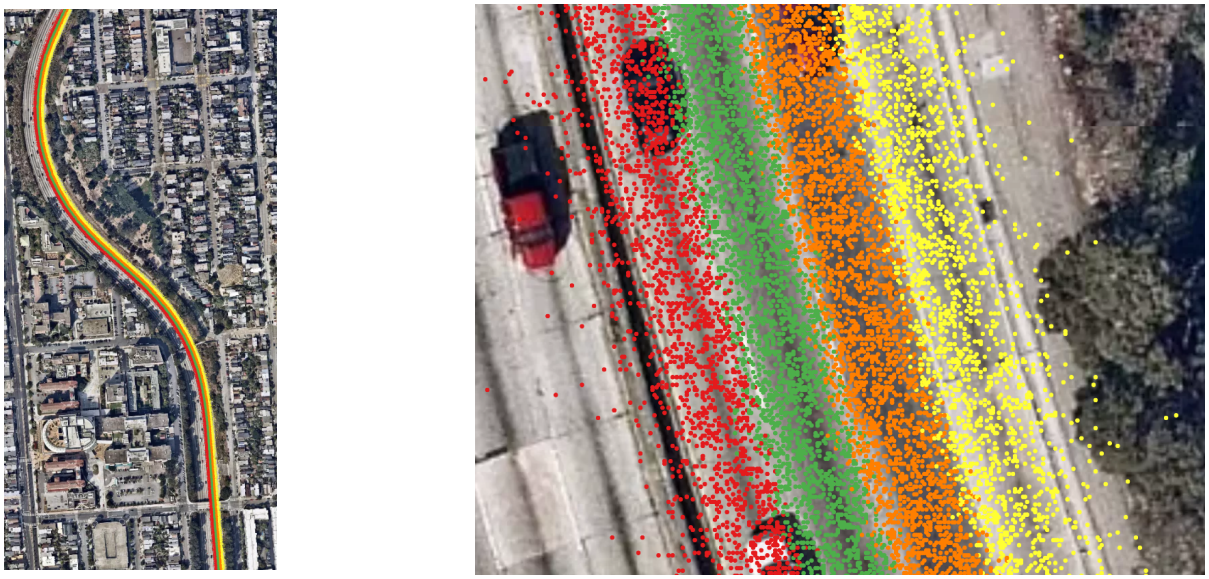


Figure 4.9: Viterbi lane assignments of all points.

To compare the different assignment methods measures of total lane changes we plot the average lane change rate for each lane by time of day on weekdays. We calculate the average change rate out, in and net for each lane. For example, Let $\delta_{i,o}$ be the number of lane changes out of lane i for time interval t_k. $S(t_k)$ is the total number of detected vehicles on the segment for time t_k. Then the average outflow rate for time of day tau is calculated as

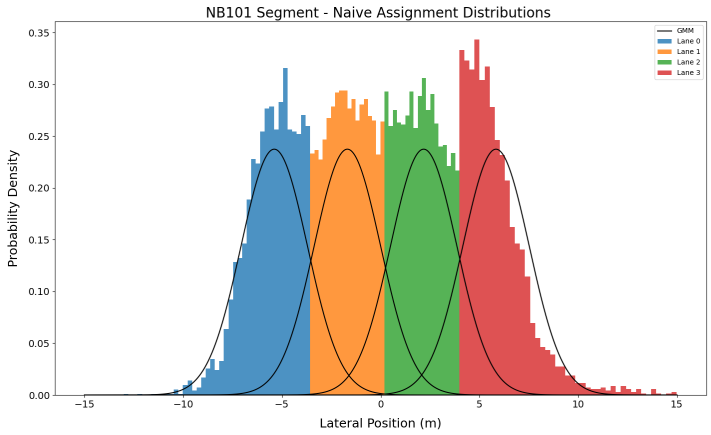


Figure 4.10: Lateral position distribution from naive lane assignment

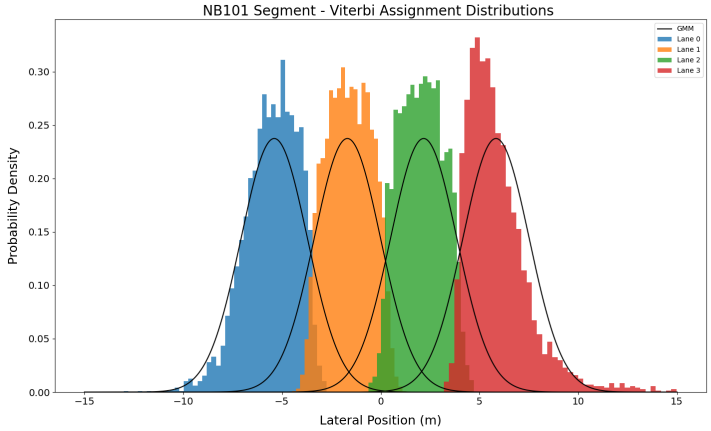


Figure 4.11: Lateral position distribution from Viterbi lane assignment

follows:

$$\alpha_{i,o}(\tau) = |\mathbb{T}_\tau|^{-1} \sum_{t_k \in \mathbb{T}_\tau} \frac{\delta_{i,o}(t_k)}{S(t_k)} \quad \text{where} \quad \mathbb{T}_\tau = \{t_k | t_k \text{ in weekday time of day } \tau\}$$

Similar average rates are calculated for the inflow of each lane and the net flow out of each lane. Net outflow average rate is a result of averaging the difference of outflow and inflow over each time of day. Figures 4.12 and 4.13 show the calculated rates for each the naive and Viterbi assignment methods. Notice that the Viterbi assignment significantly reduces the inflow and outflow rates of each lane, however the net outflow rates remain stable. This suggests that that the algorithm does well in filtering out jitter across the lane lines which would average out to zero net changes in and out of the lane.

We wanted to explore the lane change rate out of lane 0 a bit more. This is the farthest left

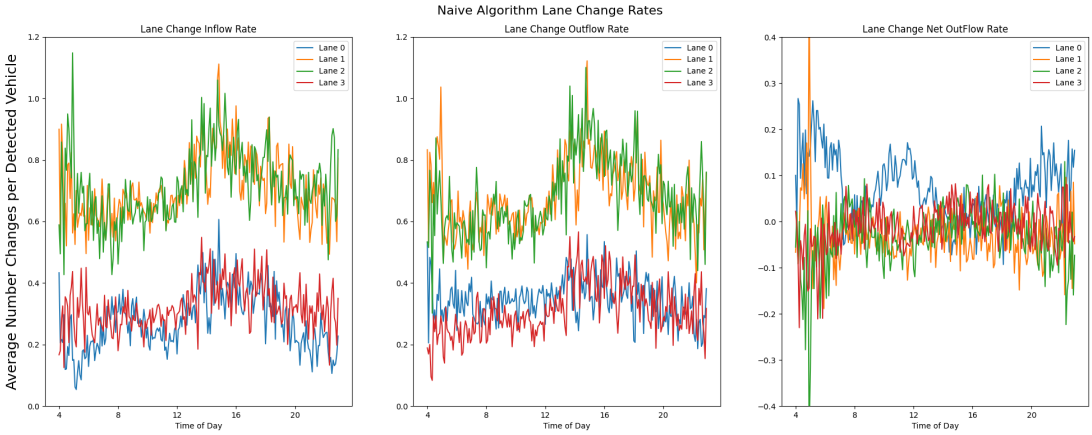


Figure 4.12: Lateral position distribution from naive lane assignment

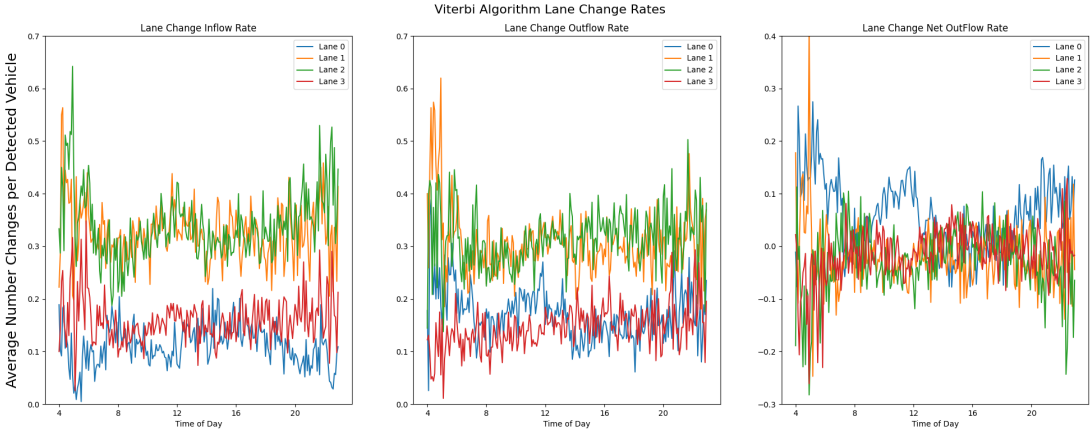


Figure 4.13: Lateral position distribution from Viterbi lane assignment

lane and this particular segment of NB101 occurs right before a major fork in the highway. This left lane becomes and exit only lane downstream. We found two signals that this lane change rate correlates well with. First we compare the average net vehicle detections taking the right fork. Let $S_R(t_k)$ be the number of connected vehicles taking the right fork for time t_k and $S_L(t_k)$ be the number taking the left. The average net flow on the right fork, $Q_R(\tau)$ is:

$$Q_R(\tau) = |\mathbb{T}_\tau|^{-1} \sum_{t_k \in \mathbb{T}_\tau} S_R(t_k) - S_L(t_k) \quad \text{where} \quad \mathbb{T}_\tau = \{t_k | t_k \text{ in weekday time of day } \tau\}$$

Figure 4.14 shows the plot of average net changes out of the left lane with the net right side fork flow.

The outflow rate also correlates well with the average speed in the lane. Figure 4.15 plots the

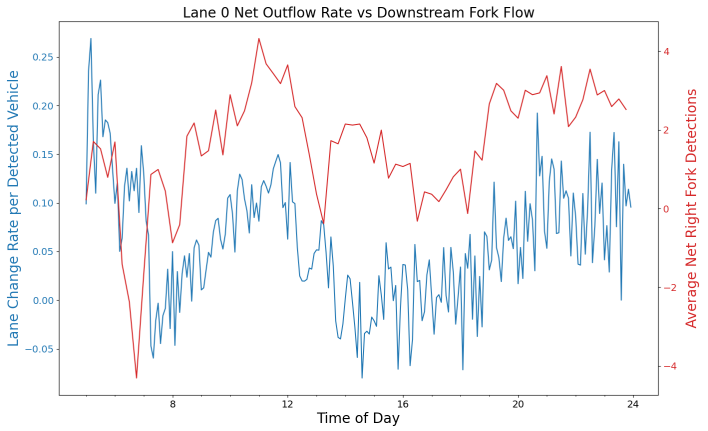


Figure 4.14: Left lane change outflow compared to the net right side flow of a downstream fork.

two signals for comparison. To speculate on what is happening, during times of congestion, flows on both sides of the fork are more even and speed difference between each lane is negligible. Therefore you get a more even mix of lane choice to destination pairs. During non congestion times the left lane is faster which creates a bias in preferred lane until drivers need to move over to go right.

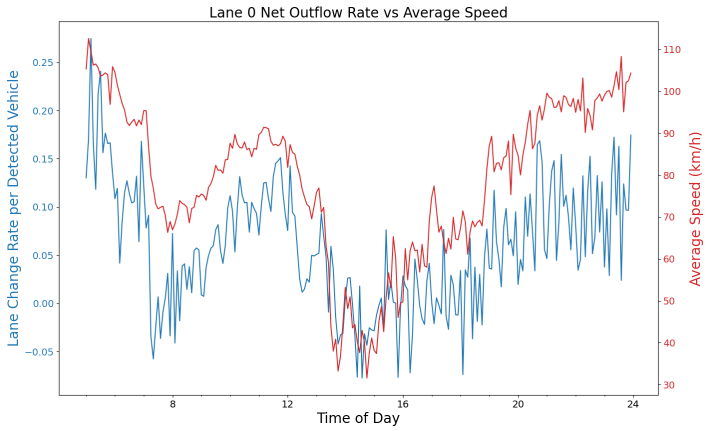


Figure 4.15: Left lane change outflow compared to average lane speed.

Plotting the distribution of where lane changes occur tells an interesting story about when drivers choose to change lanes. From Figure 4.16 we see a clear spike in lane changes out of lane 0 at the 700m mark. This is about 400m before the first sign indicating a fork in the highway. We see a slightly later spike for lane 1 and a little later spike for lane 2. Lane 3, the rightmost lane has the highest distribution around 0m and 900m. This may be the result of a on ramp right before this road segment and an off ramp right after.

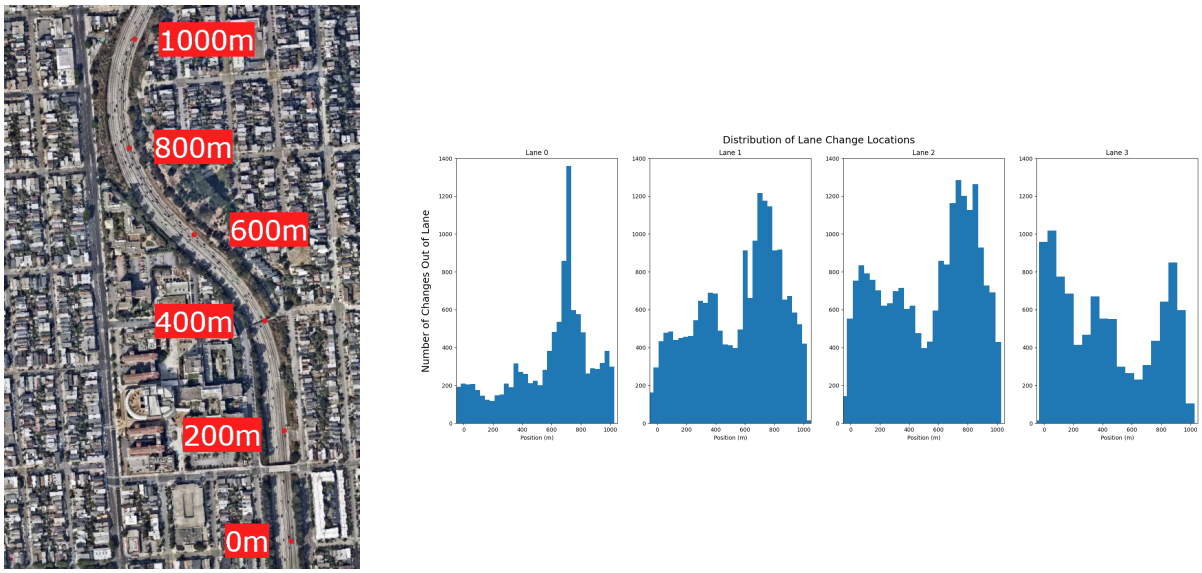


Figure 4.16: Distribution of lane change locations out of each lane.

Lastly, we look at the speed of each lane and compare it to the lane speed measurements of a PEMS station. We take the time of day average of the speedometer readings of points near the PEMS station location. Using the lane assignment we get the average speed curve for each lane at the PEMS station. We compare this to the time of day average of the PEMS g-speed estimation. Figure 4.17 shows the results. Note, the PEMS station was down for lane 1 and so the speed calculations are corrupted. For the other three lanes the results are interesting. The two signals overlap okay but there is clear error between the two. The expectation was the connected vehicle data would approximate the average speed well as sampling errors would average out over the month creating an unbiased estimate. Curiously, the overall trends are in agreement, both signals dip and spike at the same time. This is possibly a scaling issue linked to the g-factor estimation. The g-factor estimation does rely on the assumption of a known free flow speed that scales its estimation. More work is needed to explore the discrepancy between these two signals.

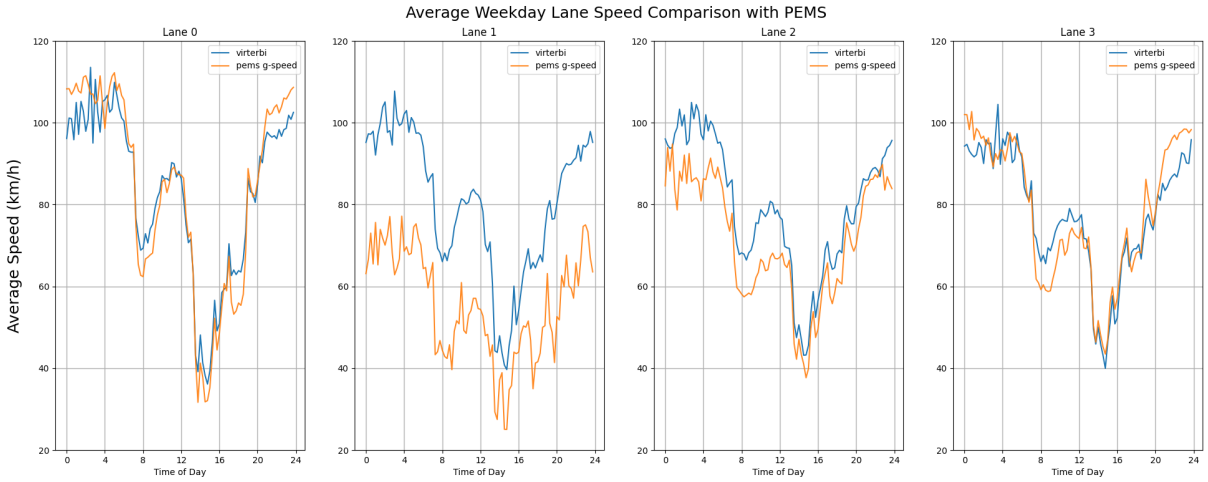


Figure 4.17: Speed of each lane compared to PEMS g-factor speed estimation.

Chapter 5

Queue Length Estimation

The distribution of queue length at signalized intersections provides crucial information for performance evaluation [35, 5], design and optimization [12, 36], and real-time operation [56] of traffic signals in road networks. Traditionally, estimation algorithms are developed based on information collected via manual surveys, in-ground fixed location sensors (*e.g.* loop detectors), or cameras. These algorithms are based on two types of models: (i) input-output models [42, 22, 64, 3] that attempt to estimate the queue by considering the cumulative arrivals and departures at an intersection; and (ii) shock-wave models [32, 44] that consider the dynamic process of formation and dissipation of queues at an intersection.

The main disadvantage of the traditional methods is their requirement for installation, operation, and maintenance of physical hardware at every intersection. As such, they have high capital costs, between \$30,000-\$60,000 per intersection [65]. Therefore, they cannot be implemented at a large scale to cover all intersections in a network. In the U.S., it is estimated that only about 3% of intersections are instrumented and monitored in real-time, and the majority of intersections receive signal re-timing updates only once every three to five years.

Recent advancement and deployment of connected vehicle technology has created new possibilities to address the high capital and operational costs of traditional measurements by physical sensors such as loop detectors. As a result, there now are several studies of algorithms for queue length estimation based on CV data. The authors in [14, 30, 63] develop algorithms based on shock-wave models in which they try to identify critical points for each CV, marking the time it joins and leaves the queue. The estimation methods in [6, 25] are based on the travel time for CVs through an intersection. In [18, 17, 16], the authors develop a stochastic framework to estimate the queue length at the end of each cycle based on the location of CVs. Alternative stochastic estimation methods are proposed in [27, 26, 47] based on the estimation of arrival/departure processes.

A practical limitation of existing queue length estimation algorithms based on CVs is that

they typically require a penetration rate $\sim$10% or greater to perform well. While reaching such targets is plausible in the future, our analysis from Chapter 3 of the current penetration rate of CVs in California, suggests $1.5\% - 3\%$ as the average penetration rate of CVs from data collected by Wejo, Inc. Additionally, many estimation algorithms [18, 17, 16, 51, 27, 26], require the knowledge of penetration rate at an intersection which is difficult to obtain without knowing the ground truth for the total numbers of vehicles in advance. This is especially challenging as our analysis shows that the average penetration rate can differ considerably even between two neighboring corridors, *e.g.* $\sim$10% relative difference between CA-107 and CA-1 corridors. Further the spatial correlation between penetration rates drops considerably after a few intersections.

This chapter consists of work presented at ITSC 2021[48]. We describe a novel approach to estimate queue length at signalized intersections and address the above limitations. The estimation algorithms do not require knowledge of the penetration rate and is applicable for penetration rates smaller than 1%. In contrast with existing methods that estimate the queue length on a cycle-by-cycle basis, we estimate the probability distribution and average value of queue lengths during selected time intervals, *e.g.* morning peaks during weekdays. The seasonal patterns of traffic at intersections are exploited to compensate for the low value of penetration rate of CVs. The algorithms are agnostic to the penetration rate and makes no additional assumption about the queue dynamics.

5.1 Analytical Model

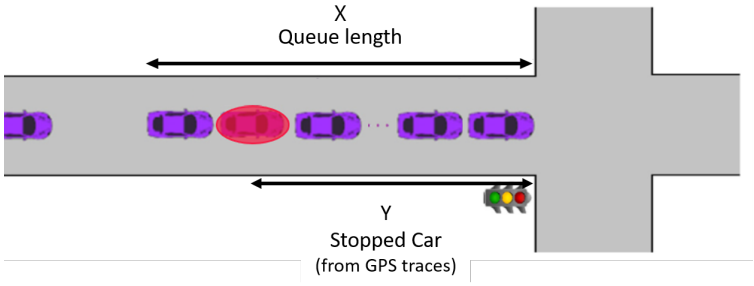


Figure 5.1: Queue formation behind an intersection

Consider a signalized intersection as in Fig. 5.1. For each leg of the intersection and a time interval T (*e.g.* morning peaks), we assume that the maximum queue length X (in meters) for each traffic cycle in each lane has a probability distribution f_X where

$$\mathbb{P}\{\underline{x} \leq X \leq \overline{x}\} = \int_{\underline{x}}^{\overline{x}} f_X(x)dx.$$

Our goal is to estimate f_X from observation of GPS traces. Let α denotes the penetration level of connected vehicles. That is, the GPS traces of each vehicle is observed with probability α; in our dataset $\alpha \simeq 3\%$. Let Y denote the distance/position from the intersection

of each stopped vehicle we observe from its GPS trace. Conditioned on the queue length X, Y is a random variable with approximate distribution

$$Y \sim \text{Uniform}[0, X] \text{ conditioned on } X.$$

So the probability distribution of Y (when we do not know X) is given by

$$f_Y(y) = \int_y^\infty \frac{1}{x} f_X(x) dx. \tag{5.1}$$

Consequently,

$$f_X(x) = -x \frac{df_Y(x)}{dx}. \tag{5.2}$$

Note, the simple collection of Connected Vehicle stop positions has a sampling bias. This is due to cycles with longer queues, on average, contain more probe vehicles. The bias becomes more significant with wider queue length distributions. The approach above takes into account the sampling bias explicitly by conditioning on X. Further, the distribution estimation of X is only reliant on the distribution of Y and does not require knowledge of the penetration rate α nor any timing information for the intersection.

This formulation does makes a few key assumptions. First, in constructing the empirical distribution for f_Y we assume that the position of observed stopped vehicles are independent. This assumption is reasonable when the penetration level is so low that we practically observe only a few samples for each queue length X. As we show through numerical simulations, even for high penetration rate the estimation approach proposed above gives satisfactory performance. Second, this formulation does not account for intersection spillover. If a queue overflows to the next intersection, we cannot distinguish stopped vehicle samples from the upstream queue corresponding to the downstream spillover.

5.2 Non-Parametric Estimation

Consider the empirical histogram for f_Y. Let $B = \{b_0, b_1, ..., b_K\}$, $b_0 = 0$, denote the bin edges and $Y = \{y_1, ..., y_K\}$ denote the associated values for the histogram. Assume that $b_{i+1} - b_i = \Delta b > 0$ for all i. We fit a curve $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_K\}$ to the empirical distribution $Y = \{y_1, ..., y_K\}$ by solving the following optimization problem.

Let $z_i := \frac{y_{i+1} - y_i}{\Delta b}$, $1 \leq i < K$, denote the slope of the histogram moving from bin i to $i+1$. For the K^{th} bin, define $z_K := \frac{0 - y_K}{\Delta b}$, *i.e.* set $y_{K+1} = 0$. Then we can write $y_i = -\Delta b \sum_{j=i}^{K} z_j$. Our model requires that $z_i \leq 0$. Therefore, we estimate a smooth curve $\hat{Y} = \{\hat{y}_1, \cdots, \hat{y}_K\}$, parameterized by its slope $\hat{Z} = \{\hat{z}_1, \cdots, \hat{z}_K\}$, by solving the following optimization problem:

$$\min \sum_{i=1}^{K} \|y_i - \hat{y}_i\|_2^2 + \beta \sum_{i=1}^{K-1} \|\hat{z}_{i+1} - \hat{z}_i\|_2 \tag{5.3}$$

subject to

$$\hat{z}_i \leq 0, \tag{5.4}$$

$$\hat{y}_i = -\Delta b \sum_{j=i}^{K} \hat{z}_j, \tag{5.5}$$

$$\Delta b \sum_{j=1}^{K} \hat{y}_i = 1 \tag{5.6}$$

Here we explicitly require the fitted curve to satisfy the monotonicity condition of our model by (5.4). Our fit of the empirical distribution, $\hat{Y}$ must also be a distribution and sum to 1 by (5.6)

The first term in the objective function denotes the estimation error, while the second term penalizes high variations in the slope of the fitted curve to ensure a smooth curve. As such, parameter β controls the trade-off between estimation error and smoothness of the curve. We note that $\{\hat{z}_1 \Delta b, 2\hat{z}_2 \Delta b, ..., K\hat{z}_K \Delta b\}$ gives us the estimate of f_X at bin edges $x = \{b_1, b_2, \cdots, b_K\}$ via (5.2). The optimization problem above can be written as a quadratic program (QP),

$$\begin{aligned} \min \ & \hat{z}^T P \hat{z} + q^T \hat{z} \\ \text{subject to} \\ & G\hat{z} \leq h \\ & C\hat{z} = 1, \end{aligned}$$

where P, q, G, h, C are defined as follows:

$$
\begin{aligned}
P &= A^T A + \beta D^T D, \\
G &= I_{N\times N}, \\
h &= [0, 0, \cdots, 0]^T, \\
q &= -2A^T \left[y_1, y_2, \cdots, y_N\right]^T, \\
A &= -\Delta b \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 \\ 0 & 1 & \cdots & 1 & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 1 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}, \\
D &= \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}, \\
C &= [1, 1, \cdots, 1]\, A
\end{aligned}
$$

Figure 5.2 demonstrates the resulting estimated $\hat{Y}$ density from enforcing smoothness and monotonicity on an empirical distribution of measured queue lengths.

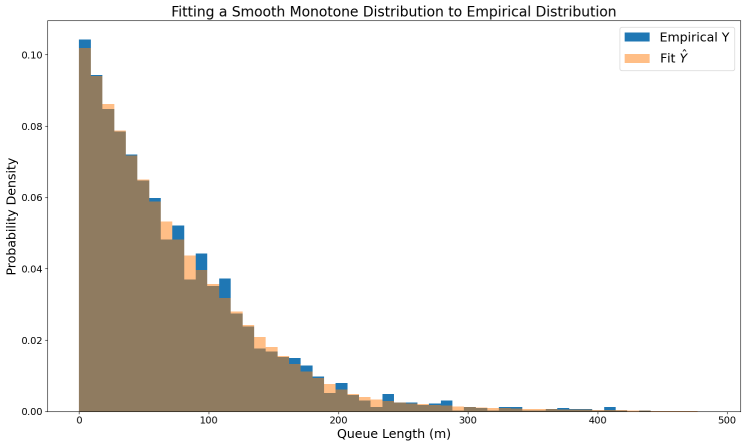


Figure 5.2: Resulting QP fit of a smooth monotone distribution to a taken empirical distribution

5.3 Parametric Estimation

Below, we consider an alternative approach by making an assumption about the family of the distribution for the queue length. We assume that vehicle arrivals to the intersection follows a non-homogeneous Poisson Process [66] . Let $\lambda(t)$ denote the arrival rate to the intersection at time t. Assuming that the duration of red phase is T_{red}, the maximum number of cars N in the queue at the end of red phase is a Poisson random variable with parameter $\lambda_Q(t) := \int_t^{t+T_{\text{red}}} \lambda(\tau)d\tau$ [60]. We assume that the traffic pattern is similar during the morning peak. As such $\lambda_Q(t)$ is similar for all cycles during 7-11AM; we denote the average value of it by λ_Q.

We note that the realized queue length for each cycle would be still different even if we assume that $\lambda_Q(t)$ is the same during 7-11AM; they correspond to different realizations of Poisson(λ_Q).

Assuming that queue length is Poisson(λ_Q), the average queue length is λ_Q. In Section 5.4, we present the estimation algorithm for average queue length. Therefore, a model-based estimation for probability distribution of queue length is given by Poisson($\hat{\lambda}_Q$), where $\hat{\lambda}_Q$ denotes the estimated average queue length.

5.4 Average Queue Length

One can estimate the expected queue length $\mathbb{E}\{X\}$ using the estimated distribution for f_X. However, we show below that $\mathbb{E}\{X\}$ can be estimated directly from data, without first estimating f_X; as such, it may have the advantage of avoiding the approximation errors arising in solving the QP.

For the expected queue length $\mathbb{E}\{X\}$ we have,

$$\begin{aligned}\mathbb{E}\{X\} &= \int_0^\infty x f_X(x)dx = \int_0^\infty -x^2 \frac{f_Y(x)}{dx}dx \\ &= -x^2 f_Y(x)\big|_0^\infty + 2\int_0^\infty x f_Y(x)dx = 2\mathbb{E}\{Y\}. \end{aligned} \tag{5.7}$$

We note that the confidence bounds for $\mathbb{E}\{X\}$ is twice the confidence bounds for $\mathbb{E}\{Y\}$. Therefore, 95% confidence interval for the average queue length can by computed as $2\overline{Y} \pm \frac{1.96}{\#\text{observations}} s_Y$ where $\overline{Y}$ and s_Y denote the empirical mean and standard deviation for Y.

5.5 Numerical Simulation

We numerically evaluate the ability of our proposed non-parametric algorithm to recover the true distribution of queue length $f_X(x)$. The simulation setup is as follows.

We consider queue lengths during morning peak hours 7-11 AM on working days for one month at an intersection. Let $\lambda(t)$ denote the average arrival rate to the intersection at time t. Assuming that the duration of red phase is T_{red}, the maximum number of cars N in the queue at the end of red phase is a Poisson random variable with parameter $\lambda_Q(t) := \int_t^{t+T_{\text{red}}} \lambda(\tau)d\tau$. We assume that the traffic pattern is similar during the morning peaks. As such $\lambda_Q(t)$ is similar for all cycles during 7-11AM; we denote its average value by λ_Q. We note that the realized queue length for each cycle would be still different even if we assume that $\lambda_Q(t)$ is the same during 7-11AM; they correspond to different realizations of Poisson(λ_Q). For our simulation, we set $\lambda_Q = 15$ and assume that $N \sim$ Poisson(15) for each cycle. Moreover, each traffic cycle lasts 120 (s).

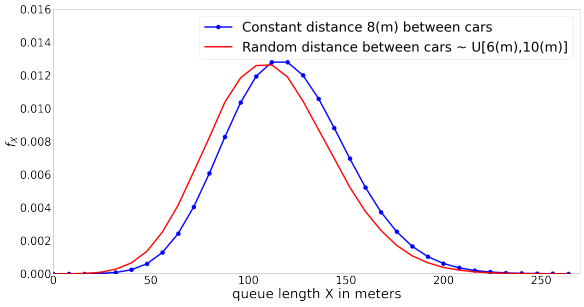


Figure 5.3: Numerical simulation for the distribution of queue - length X (in meters) vs $8\ (m) \times N$ (number of cars).

Now consider a queue of vehicles at the end of a red phase. The distance between every two consecutive vehicles in the queue d_i, $1 \leq i \leq N$ is random. We assume that this distance is on average 8 (m) (including one vehicle length) and set $d_i \sim U[8 - 2\ (m), 8 + 2\ (m)]$; that is, each distance d_i may deviate from 8 (m) by up to 2 (m). Accordingly, given a realization for the number of cars in the queue N, we simulate the length of the queue in meters by assuming that the distance between successive cars is $U[8 - 2\ (m), 8 + 2\ (m)]$. Figure 5.3 shows the empirical distribution for queue length X (in meters). We would like to point out that the randomness in car distances d_i, makes the distribution $f_X(x)$ slightly different from $f_N([x/8])$. Most importantly, $f_X(x)$ (queue length in meters) is a continuous distribution while $f_N(n)$, where $n = [x/8]$ (number of queued cars), is a discrete distribution. Additionally, f_X is slightly left skewed compared to f_N.

For the numerical simulation, we consider three penetration levels for connected vehicles $\alpha \in \{0.5\%, 1.5\%, 5\%\}$. Figures 5.4-5.6 depict the estimated probability distribution for queue length (in meters) $\hat{f}_X$, and compares them with the true distributions f_X and f_N. Moreover,

the estimated average queue length for each case $\{0.5\% : 111.4\,(m), 1.5\% : 117.4\,(m), 5\% : 122.0\,(m)\}$, along with the confidence bounds, are compared against the true sample average of queue lengths at $120.5\,(m)$. As it can be seen, the estimation accuracy is satisfactory even for very low penetration $\alpha = 0.5\%$. Additionally, the estimation accuracy for both the average queue length and the probability distribution improves as the penetration rate α increases from 0.5% to 1.5%, and 5%.

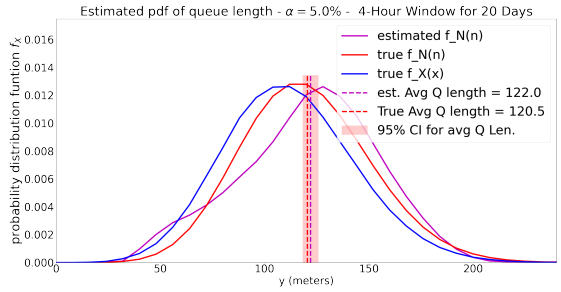


Figure 5.4: Numerical simulation for penetration rate $\alpha = 0.5\%$

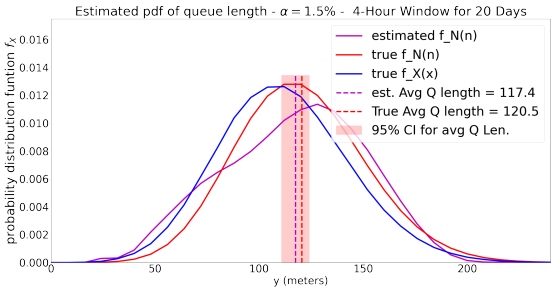


Figure 5.5: Numerical simulation for penetration rate $\alpha = 1.5\%$

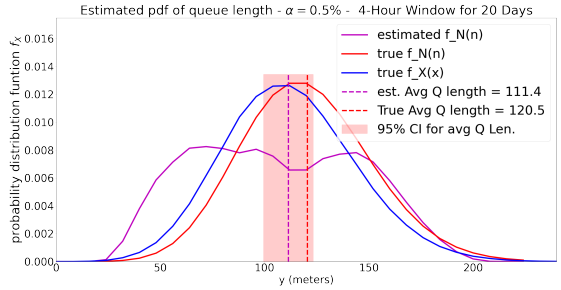


Figure 5.6: Numerical simulation for penetration rate $\alpha = 5\%$

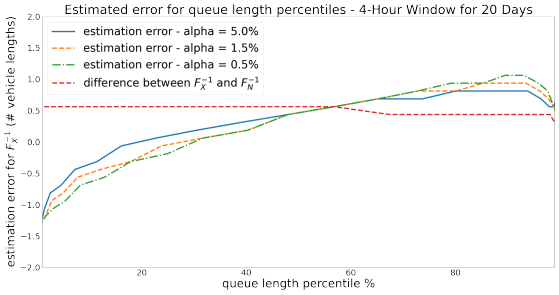


Figure 5.7: Numerical simulation - the estimation error for queue length percentiles F_X^{-1}

A particular value that is critical in both performance evaluation and the design of timing plan at intersections is the the queue length at specific quantiles (*e.g.* 95%). Figure 5.7 shows the estimation error vs. quantiles. We note that the estimation error for quantiles above 60% is always less than two times the average distance between queued vehicles. Moreover, this error is comparable to the distance between the two distribution F_x (queue length in meters) and F_N (queue length in number of vehicles $\times$ average vehicle length) which describe the true queue length distribution in two different ways

5.6 Empirical Verification - Comparison with Detector Data

We demonstrate the implementation of our approach for intersections on CA-107 corridor using real data. We present the results for three signalized intersections with different characteristics: (i) Lomita Blvd & CA-107 is a major intersection with one left-turn lane, four through lanes, and no dedicated right-turn lane; (ii) Torrance Blvd & CA-107 is a major intersection with two left-turn lanes, four through lanes, and one right-turn lane; (iii) Talisman St & CA-107 is an intersection located in front of an outlet, with one left-turn lane, four through lanes, and one-right turn lane leading to the outlet. To empirically verify the result of our approach, we estimate the queue length using vehicle detection sensors located at the downstream of the intersection links as well and compare the results.

We adopt the estimation algorithm proposed in [32] based on shock-wave theory. The key idea used in the algorithm is that the discharge rate for vehicles in the queue, which is observed via sensors at the downstream of the intersection, is higher than the normal free flow of vehicles when there is no queue. As such, one can determine the time gap between consecutive vehicle detections at the downstream, and identify the time instance where the queue is cleared, and thereby estimate the number of cars in the queue.

To determine the end of the queue, we identify the first time (skipping the first vehicle in the green phase) the detection gap exceeds a given a threshold *max gap*.We choose the value of max gap based the nominal time gap between consecutive vehicles leaving the queue.

We estimate the nominal gap between detections of two vehicles leaving the queue to be approximately $3(s)$. This is based on our estimate of $14.5-20(km/h)$ for how fast the shock-wave of vehicles clearing the queue propagates backward into the intersection upstream. See our note in [49] for the detailed calculation of $14.5(km/h)$ using drone footage. Later chapter computes upwards of $20\,(km/h)$ from connected vehicle traces. Assuming that the average distance (including a vehicle length) between two queued vehicles is $8\ (m)$, vehicles leave the queue with $\sim 2\,(s)$ time gaps. Additionally, for each two consecutive vehicles, the rear vehicle has to travel the additional distance of $8\,(m)$ to reach the location of the detectors, which takes $\sim 1\,(s)$ assuming an average speed of $30\,(km/h)$. The setting of max gap and average distance can change the estimated queue distribution. Increasing max gap will skew the distribution towards the right and increasing distance per vehicle stretches it out. We found that the parameters of $9(m)$ average distance at a max gap of $6(s)$ most closely aligned with the results from connected vehicle estimation. These are not unreasonable numbers. Figure 5.9 illustrates how different max gap numbers changed the estimated queue length from the detectors.

Aside from the sensitivity to these parameters, this estimation from detectors has a few more drawbacks. Identifying the end of the queue is sensitive to drivers' delays and distractions. That is, if a driver starts moving with some delay after the front vehicle leaves the queue, the

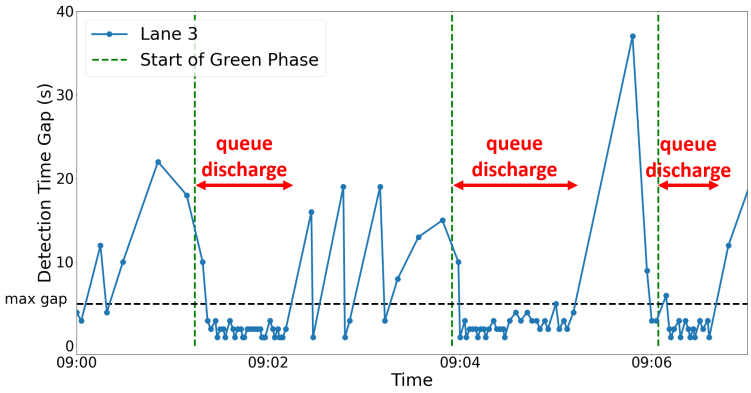


Figure 5.8: Illustrating detection of queue discharge from max gap data.

recorded time gap between the two vehicles can exceed the max gap threshold. As a result, we may underestimate the queue length for that cycle. Further, the traffic signals on CA-107 are actuated, and thus, the phase lengths vary slightly from cycle to cycle. Unfortunately, the realized phase lengths and cycle timing are not recorded and cannot incorporated in our estimation algorithm based on detections. As a result, we develop an algorithm to estimate the timing of the green phase based on the vehicle detection data on all four links of the intersections.

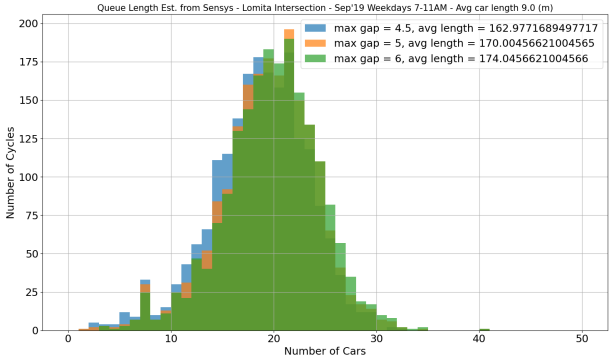


Figure 5.9: Distribution estimations from detector data with varying max gap parameters.

Figure 5.10 depicts an example of the estimated queue lengths for lane 2 during the morning peak 7-11AM on September 12, 2019, using the detection data.

Results & Comparison

We implement the proposed estimation algorithm based on connected vehicles for three intersections on CA-107 and compare the results with those based on detector data.

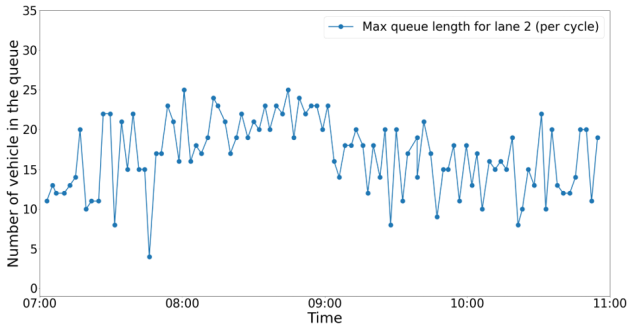


Figure 5.10: Queue Estimations during morning on Lomita.

We set $\beta = 150$ and utilize the dataset during morning-peak (7-11AM) of business days in September, 2019, consisting of approximately 2000 traffic cycles. We make the comparison only for lane 2 of NB direction at each intersection; we note that vehicles queuing in lane 1 (left-most lane) and lane 4 (right-most lane) may end up turning at the intersection, and thus do not end up being detected by the in-ground detection sensors at the outflows of the intersections.

Figure 5.11 depicts the estimated probability distribution as well as the average queue length using our non-parametric and parametric estimation algorithms. We note that the non-parametric estimation and average queue length are very close to the results using the in-ground detection sensors. Notably, the difference in the average queue length is less than or equal to an average distance between queued vehicles at $9 : /(m)$. Granted 9 was chosen to give a good match between the distribution estimates it is a reasonable value. Note the Poisson parametric estimation did not perform as the detector distributions are clearly not Poisson.

Figure 5.12 shows the difference between estimated percentiles F_X^{-1} using connected vehicles vs. detection sensors. As can be seen, the difference in queue length percentiles for 90%-98% is less than or equal to twice the average distance between two queued vehicle for the non-parametric estimation. Hence the empirical results suggest that the estimation method we propose in this paper can generate information about queue lengths that is accurate enough for both performance monitoring and the design of time-of-day traffic cycle plans.

5.7 Additional Discussion & Conclusion

The simulation and empirical results demonstrate the effectiveness of the estimation algorithm we propose based on trace data from CV with very low penetration levels. A major advantage of estimation from trace data is the full coverage it provides at all intersections without any hardware equipment and traffic interruption for their installation and maintenance. Additionally, it has negligible marginal costs since the trace data is being collected by

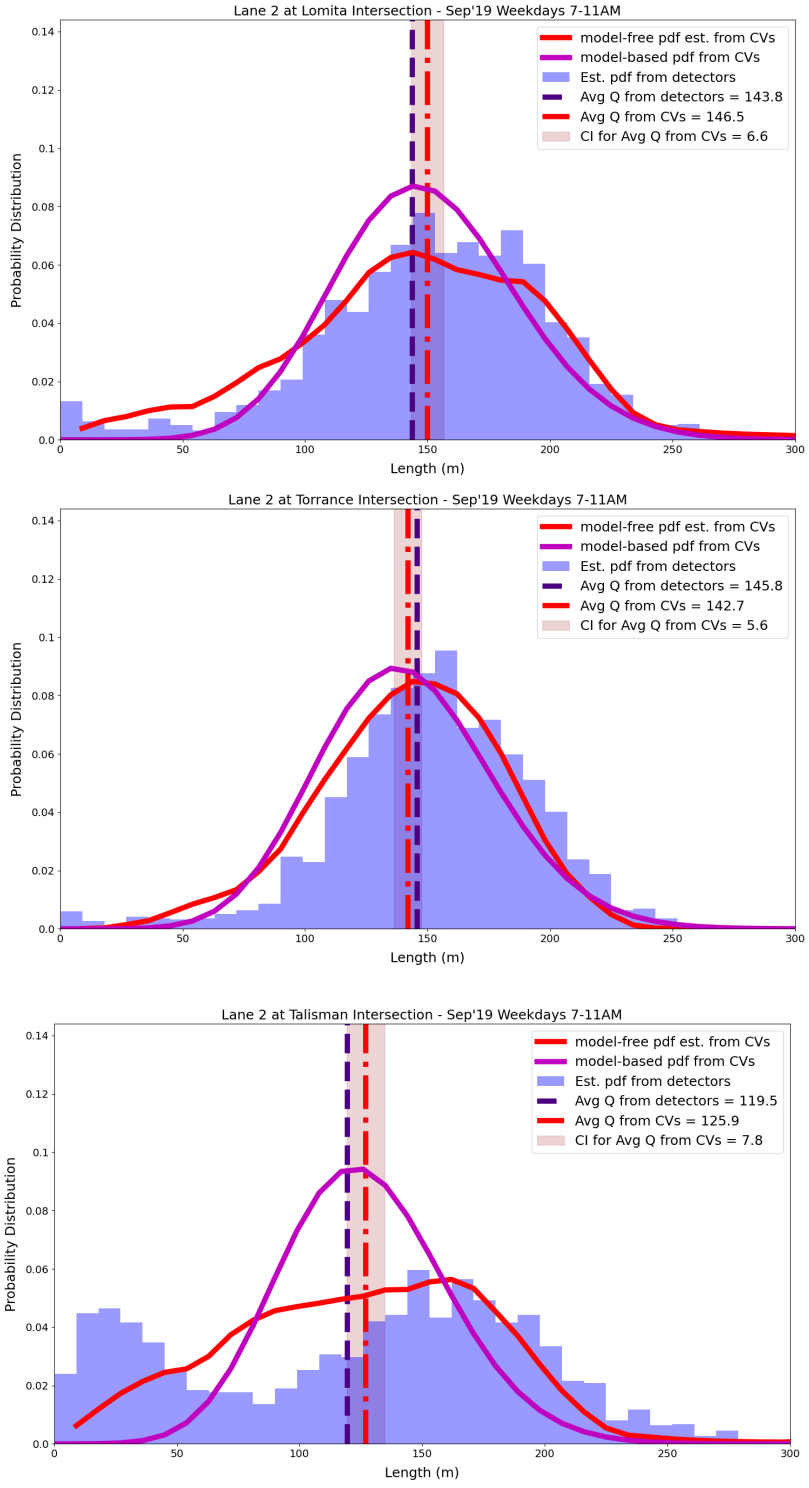


Figure 5.11: Estimated queue length distribution on lane 2 (NB) for intersections on CA-107 from connected vehicles with $\alpha \simeq 3\%$ vs. queue length estimation from in-ground vehicle detection sensors.

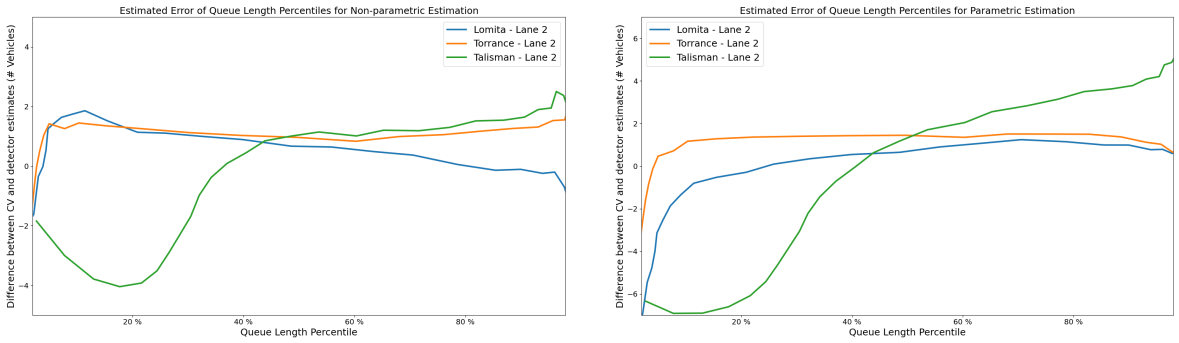


Figure 5.12: The difference between estimated queue length percentiles F_X^{-1} using connected vehicles vs detectors.

various OEMs. As such, our algorithm can already be utilized to monitor the performance of intersections across the road network, and to estimate the queue lengths necessary for the design and optimization of time-of-day traffic cycle plans for each intersection.

The main disadvantage of this approach is its limited ability to provide accurate real-time queue length measurements for each cycle individually. However, the existing algorithms for cycle-by-cycle estimation of queue length rely heavily on the assumption that they observe at least one CV in each cycle. At a penetration rate of $\sim 3\%$ with an average queue length of 15 cars, we do not observe any CV in 60% of cycles. A recent work by [47] suggests to address such an issue by utilizing information from historical trends. As such, our estimation algorithm can provide information to estimate such recent historical trends. Specifically, one can use our algorithm to estimate the distribution and/or average queue length during a narrower recent window of time (*e.g.* past few hours) in order to be used in such a hybrid scheme as suggested in [47].

Moreover, we argue that during peak hours when the traffic flow tends to be more predictable, the performance of a well-optimized time-of-day traffic cycle plan is as good as an adaptive traffic cycle control. The authors in [15] found that in fact that the time-of-day traffic cycle plan performs better than the adaptive traffic cycle control for a test site in Anaheim, CA, which is located only 35-40 km away from the intersections on CA-107. While we do not aim to provide a similar detailed comparison as in [15] here, we investigate the correlation among consecutive queue lengths as a proxy for the potential value of adaptive traffic cycle control. We note that currently, all intersection on CA-107 employs a coordinated actuated time-of-day cycle plan. Figure **??** shows the auto-correlation among consecutive queue lengths (estimated from detection sensors) at CA-107 & Lomita Blvd intersections. As it can be seen, the realized queue length at the end of each cycle includes very limited additional information about the queue lengths and traffic during the next few cycles, beyond the information already contained in the historical distribution. This suggests that the use of

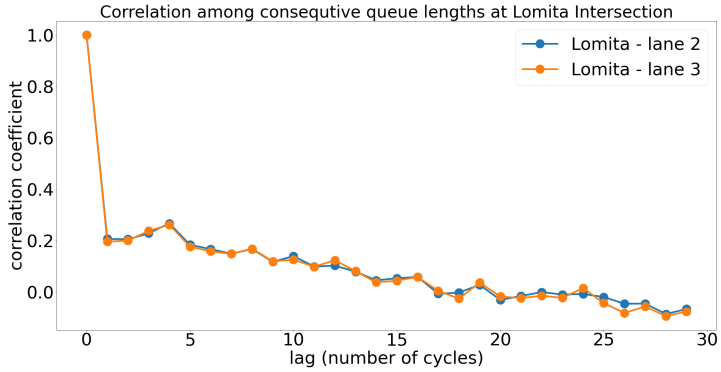


Figure 5.13: Auto-Correlation of consecutive Queue Lengths

adaptive traffic cycle has potentially very limited positive impact on the performance of the traffic signals during peak hours with normal traffic patterns that do not deviate significantly from the historical distribution.

Chapter 6

Macroscopic Properties from Timed Intersections

The work of Daganzo and Geroliminis in [20] [4] [23] [24]. proposed and analyzed the existence of a macroscopic fundamental diagram (MFD). An strong relationship between the average flow, average density, and average speed arises from an aggregated network. This relationship is an inherent to the network and time invariant given the network does not change and signal timing remains constant. From the first proposal a large body of research was conducted finding conditions for an MFD's existence, searching for them in traffic, analyzing what factors affect the MFD, and applying MFDs to traffic control.

Connected vehicle data has a large potential for studying macroscopic fundamental diagrams. In [37] measures for network flow and density from probe vehicle data are proposed with analytical variances. Multiple penetration rates are tested on simulated data and compare observed MFDs. From simulations they conclude that a penetration rate of 20% is need for an accurate measurement. At a 2% penetration rate, we explore what is available with our current level of data.

6.1 Problems with Sparsity

With a low penetration rate, measures of traffic have high variances as seen in Chapter 3. Plotting flow and density values measured from this sparse set, smears the macroscopic fundamental diagram and makes it difficult to understand the state of the network. Figure 6.2 plots all the trajectories for one hour going northbound on 19th avenue. On its own, we can see a transition from congestion in the upper half at 4pm to free flow after 4:10pm, but the overall state is not very clear.

We chose 19th Avenue as a corridor of study because it has favorable properties for the existence of a Macroscopic Fundamental Diagram. The road is straight with a constant 3

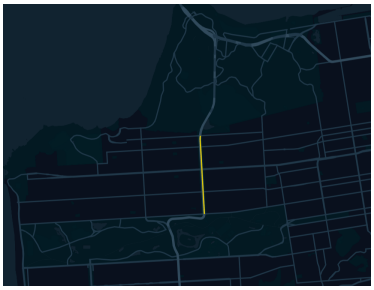


Figure 6.1: San Francisco, Upper 19th Avenue (CA-1) Map

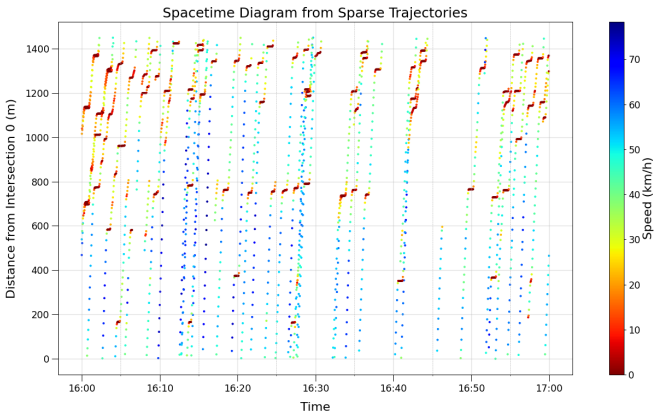


Figure 6.2: All captured northbound trips on upper 19th from 4-5pm, 9/5/2021

lanes. For long periods of time the signals appear to have set timing cycles all synchronized to each other (Note: There may be some adaptive timing of the signals but it is evident that there is at least a consistent average cycle time where signals re-synchronize). Distance between the signalized intersections is consistent at about 200m and left turns are banned for smoother flow. The road is also a major through-way for CA-1 and thus experiences high flow and consistent cycles of congestion. Further their exists a PEMS detector just north of the last intersection giving us a highly correlated measurement of the penetration rate.

6.2 Aligning Trajectories

To get a clearer spacetime diagram of traffic states we align the trajectories to their time within a cycle. With a cycle length of c seconds and a sample's time in seconds from midnight t_k, the time within the cycle τ_k for sample k is calculated simply.

$$\tau_k = t_k \bmod c$$

For this upper leg of 19th Avenue there exists two weekday cycle settings that experience congestion. From 9:45am - 3:00pm the cycle is set to 80 seconds. From 3:15pm - 7pm the cycle is set to 90 seconds. (We skip the intermediary period where the cycles switch settings) Figure 6.3 shows the result of plotting all the aligned northbound trajectories from weekdays, 3:15pm- 7pm, September 2021. The same 90 second spacetime diagram is copied and translated by 90 seconds for clarity.

From Figure 6.3 the backwards propagation of queue discharge is clearly seen. To get a measurement of the discharge shockwave velocity we plot the samples of vehicles right before they are seen moving again. Figure 6.4 plots these samples along with linear estimations of

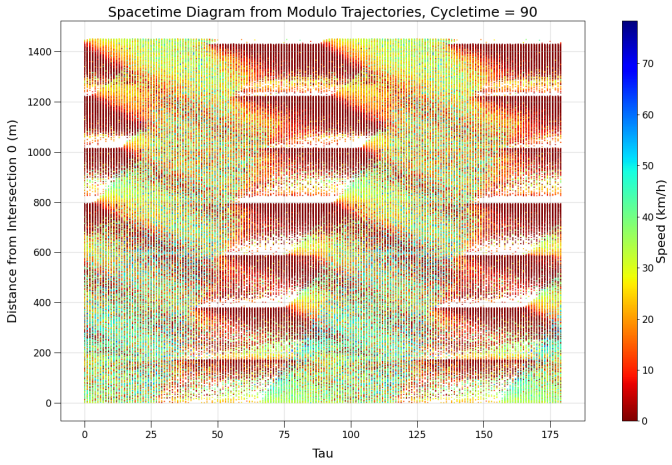


Figure 6.3: Aligned NB 19th Avenue trajectories for weekdays, 3:15pm - 7pm

the shockwaves. We see that there is a bit of spread where exactly the queue begins moving. This can be due to the variable delays of drivers or possible adaptive timing lengthening or shortening the stop cycle. We estimate the discharge rate by fitting to the outside envelope. Estimates range from -19 to -22 km/hr with the average estimate at -20 km/hr.

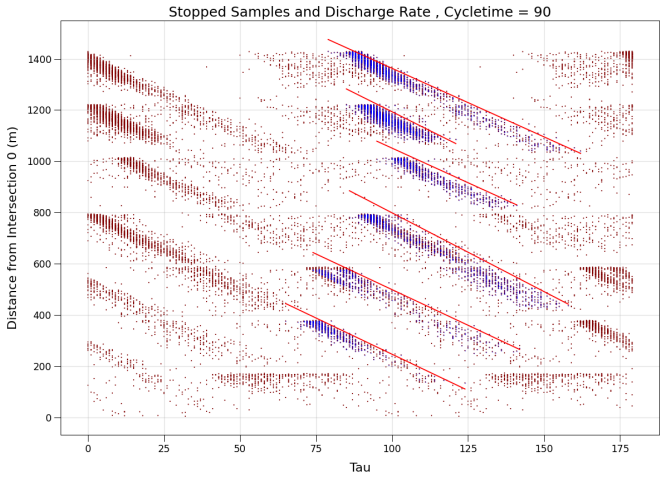


Figure 6.4: Stopped samples of vehicles before moving again along with estimated discharge shockwaves.

6.3 Clustering Time Intervals

We want to be cluster time intervals that exhibit the similar congestion patterns. If traffic patterns are fairly repeatable throughout the month then aligned trajectories within the same pattern should produce a clearer picture. Using this idea we separate time intervals based off how trajectories within the interval compare to the average spacetime diagram.

Let $G = \{(d_k, t_k, \tau_k, x_k, v_k)\}$ be the set of samples in a grouping we want to split. d_k is the date of sample k, d_k is the date, t_k is the timestamp in seconds from midnight, τ_k is the time within a cycle as defined above, x_t is the position in meters along the road, and v_k is the velocity. We start with G_0 as the set of all samples falling within the cycle time of study. G_0 is the set of all weekday samples between the times of 5:15pm and 9:00pm with a cycletime of 90s.

First we discretize the spacetime for the length of upper 19th and time in a cycle. Define $i_k = \tau_k/\tau_{div}$ and $j_k = x_k/x_{div}$. i and j are the discretized time and position within the spacetime diagram. We use $\tau_{div} = 1$ and $x_{div} = 5$. Let K_{ij} be the set of all samples that fall within pixel (i,j). Then the average velocity and standard deviation of each pixel is defined as:

$$\bar{V}(i,j) = \frac{\sum_{k \in K_{ij}} v_k}{|K_{ij}|}$$

$$\Sigma(i,j) = \left(\frac{\sum_{k \in K_{ij}} (v_k - \bar{V}(i,j))^2}{|K_{ij}|} \right)^{-1/2}$$

For each sample in a group we assign a score:

$$s_k = (v_k - \bar{V}(i_k, j_k)) \frac{\Sigma(i_k, j_k)}{1 + \bar{V}(i_k, j_k)}$$

Scores are based on how much they differ from the mean speed of samples in the same pixel. We multiply by the standard deviation of the pixel to prioritize splitting where there is a high variance in speed. We also scale by 1 plus the mean speed of the pixel to give higher weight locations where vehicles have low velocity. This is because there is higher variance from vehicle to vehicle in there acceleration and desired speed. However when they are forced to queue that variation between actors decreases.

A time interval is scored by the total sum of sample scores within its interval. If an interval's score is positive it is split right, negative split left. This process is repeated as illustrated in Figure 6.5 separating out intervals into different groupings.

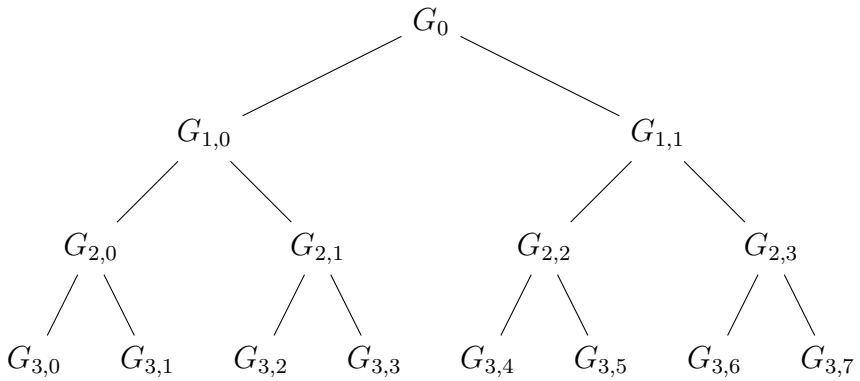


Figure 6.5: Binary splitting of time intervals into groups with similar traffic patterns

Figure 6.6 shows the results of separating out 15 minute intervals. We could continue splitting up clusters of into separate groups but there are a limited number of intervals for 1 month of measurement. We wanted enough samples in each group for density and flow measures.

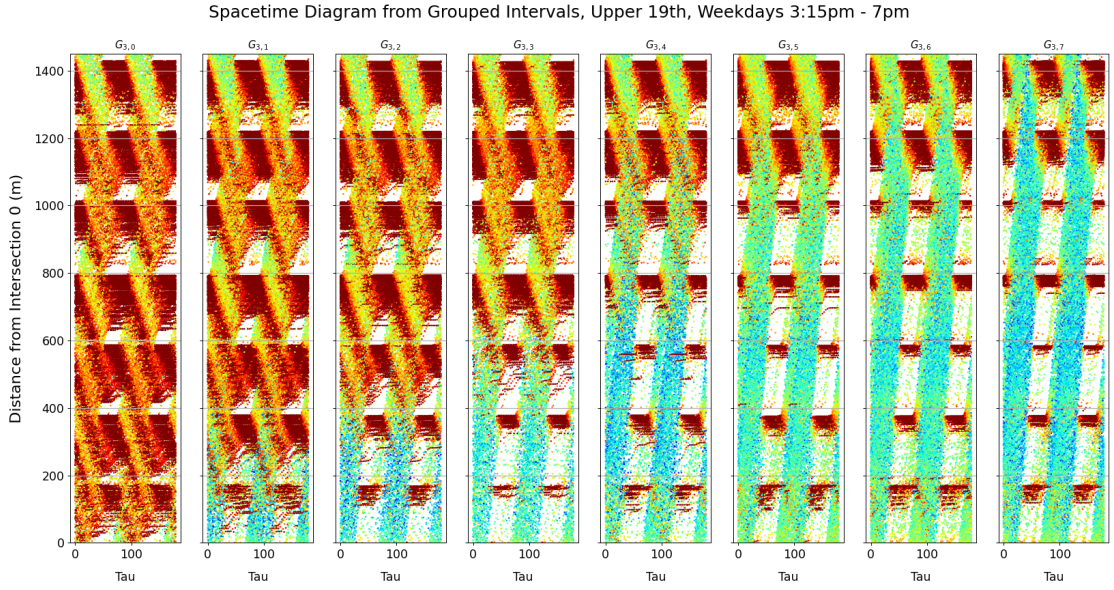


Figure 6.6: Results of separating groups of similar traffic patterns.

By examining sections of the space time diagram where spillback for intersections occur, we can get a rough estimation of the probability of spillback from each intersection for each grouping. We look at the area 100m back from the start of the preceding intersection and divide the number of vehicles that are observed to stop by the total number of vehicles that pass. Figure 6.7 plots the calculated probabilities starting with the most north intersection.

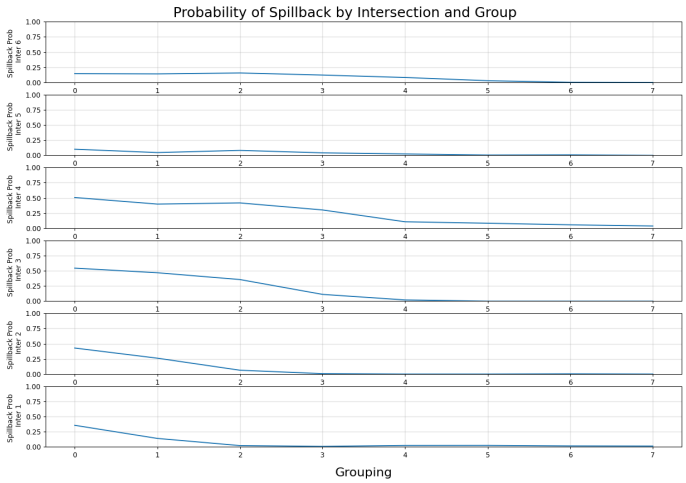


Figure 6.7: Calculated spillback probability for each intersection and grouping.

6.4 Estimating Macroscopic Properties

The generalized definitions for network flow - q, and network density - k are defined in [21].

$$q = \frac{\sum_i d_i}{LT}$$
$$k = \frac{\sum_i t_i}{LT}$$

Here, T is the total time of measurement, L is the total distance of the network. d_i is the the distance traveled by trip i in the network for the time of measurement. t_i is the total time trip i spends in the network. This also produces a generalize velocity of the network, which is the total vehicle distance covered divided by the total vehicle time.

$$v = \frac{q}{k} = \frac{\sum_i d_i}{\sum_i t_i}$$

To estimate traffic density and traffic flow from connected vehicle data we use the approximations proposed by [37]. They use the total sums from their probe vehicle sets and divide by the penetration rate α.

$$\hat{q} = \frac{\sum_{j \in \mathcal{D}} d_j}{\alpha LT}$$
$$\hat{k} = \frac{\sum_{j \in \mathcal{D}} t_j}{\alpha LT}$$

Just north of the last intersection there is a PEMS sensor where we can obtain penetration rates in close proximity to the corridor of interest. We use the calculated time of day penetration rates for the month. The macroscopic flow, density, and velocity values for upper 19th Avenue were calculated for each 15 minute interval from 5:15pm - 7pm, Weekdays Figures 6.8, 6.9, and 6.10 illustrate the spread of values for each of the groups.

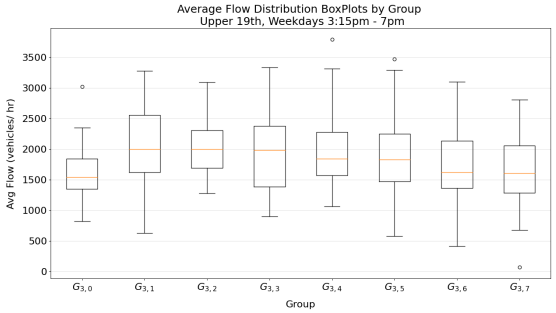


Figure 6.8: Average flow of separated 15 minute intervals.

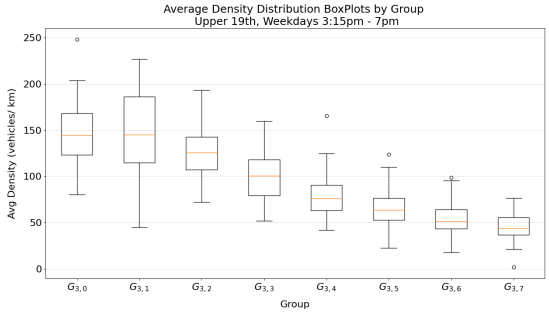


Figure 6.9: Average density of separated 15 minute intervals.

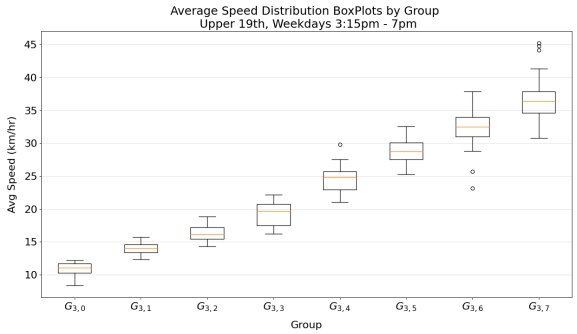


Figure 6.10: Average velocity of separated 15 minute intervals

We'd like to note that a similar cluster separation can be performed by simply bisecting by the observed average speed. However this creates disjoint sets of average speed where the scoring method described has overlapping average speeds among 2 or more groups. This is seen in 6.10 where the distributions clearly overlap. The scoring method may produce better representations of the different states. Overlapping is expected with the high variance of flow and density values.

To understand the variance in measurements assume, for each group, the density measurement is a constant $\bar{k}_g$. Then the sample sample set will see a density

$$\alpha\bar{k}_g = \hat{\bar{k}}_g = \frac{\sum_{j\in\mathcal{D}} t_j}{LT}$$

Note this is the measured density of the connected vehicle sample set (i.e. no penetration rate scaling)

If the total network density is constant for each group $\bar{k}_g$, and $\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha)$ then we would expect the density of sampled vehicles to be

$$\hat{\bar{k}}_g \sim \mathcal{N}(\bar{k}_g \mu_\alpha, \bar{k}_g^2 \sigma_\alpha^2)$$

Figure 6.11 shows the histogram of penetration rates for September 2021, Weekdays, 5-7pm.

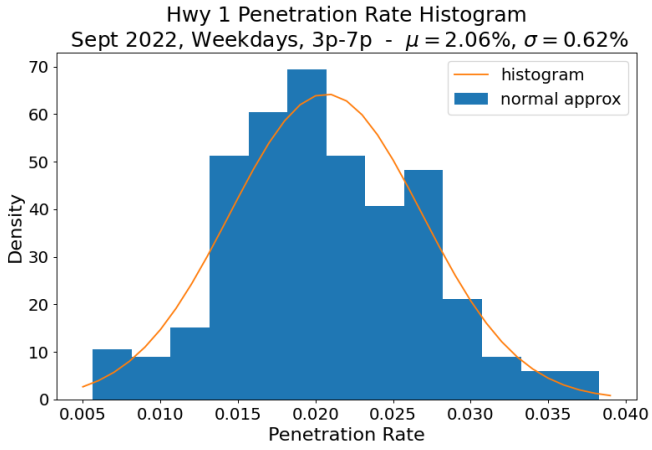


Figure 6.11: Histogram of Hwy 1 penetration rates for Weekdays, 5-7pm

We don't have enough samples from each group to compare probability densities, however we can compare the empirical variance vs the expected variance given $\bar{k}_g$ is the average of the approximated network densities $\hat{k}_g$. Similar comparisons can also be done with the flow measurements. Figure 6.11 shows the histogram of penetration rates for September 2021, Weekdays, 5-7pm. From the estimated normal distribution of the penetration rate we compute the expected variance. Figure 6.12 compares the expected standard deviation to the standard deviation for flow and density measurements seen in the sample set.

From the similar standard deviations measured from the CV set and predicted from the penetration rate, we can argue that the spread of flow and density values in each group is mainly a result of the variability in the penetration rate.

6.5 Macroscopic Fundamental Diagrams

By plotting the macroscopic flow versus density a concave fundamental diagram should be evident for this corridor. Figure 6.13 shows the resulting scatter plot from plotting all 15

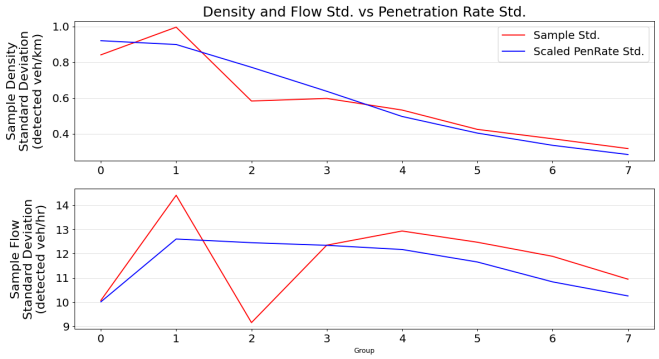


Figure 6.12: Comparing the standard deviation of sample density and flow vs. the standard deviation of the penetration rate

minute intervals from weekdays 5:15p-7p.

We see that the variance in the measurements smears out the scatter and makes for a hardly discernible macroscopic fundamental diagram. To deal with some of the variability we apply a windowed moving average to the flow and density signals. We use the same window discussed in Chapter 3 on penetration rates. This consolidates the scatter considerably and the MFD becomes much more clear.

Plotting each group's average shows the trend.

Similar analysis is performed for the morning/afternoon cycles set at 80 seconds. Plotting the two MFD's allows us to compare the effects of cycle time on the MFD. Figure **??** shows the comparison of the two cycletimes. The plot reveals a slight reduction in road capacity from the reduced cycle length.

The lower portion of the 19th avenue (mapped on Figure 6.15 also exhibits similar properties as the upper portion that create good conditions for a MFD. This collection of intersections have a 90 second cycletime from 6:00am- 10:45am and a 100 second cycletime from 11:00am - 8:00pm on weekdays. Figure 6.16 illustrates the results of grouping the 100 second cycles.

Figure 6.17 plots the resulting MFDs for the two cycletimes. The same reduction in capacity seen in the upper portion is visible here in the reduction from 100 second to 90 second cycles.

6.6 Discussion

Even at low penetration rate, there is a wealth of information on the MFD and state evolution of the traffic network. From the large scatter we can filter out a measurable change in capacity resulting from the change in signal timing. Further work would explore the MFD

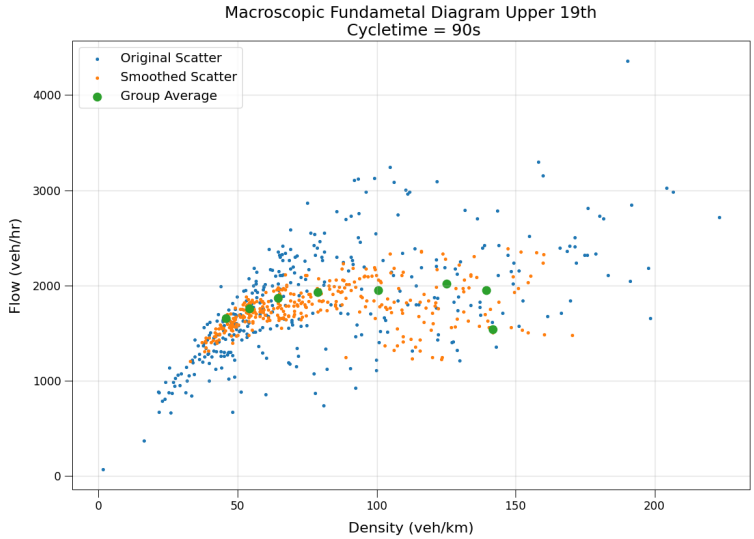


Figure 6.13: Macroscopic Diagram for northbound upper 19th Avenue with a cycletime of 90 seconds.

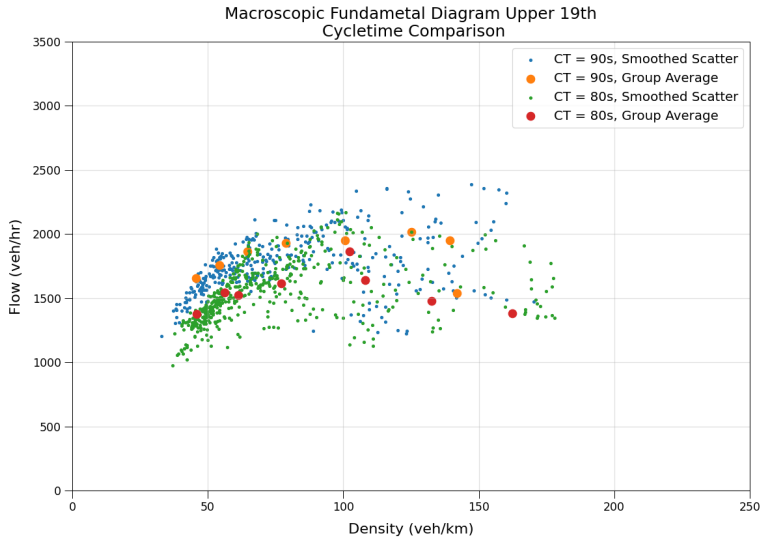


Figure 6.14: Comparison of Macroscopic Diagrams for northbound upper 19th. Comparing cycletimes of 80 and 90 seconds.

of the surrounding streets flowing into 19th and how their states are coupled.

Figure 6.15: Map of lower portion of 19th Avenue.

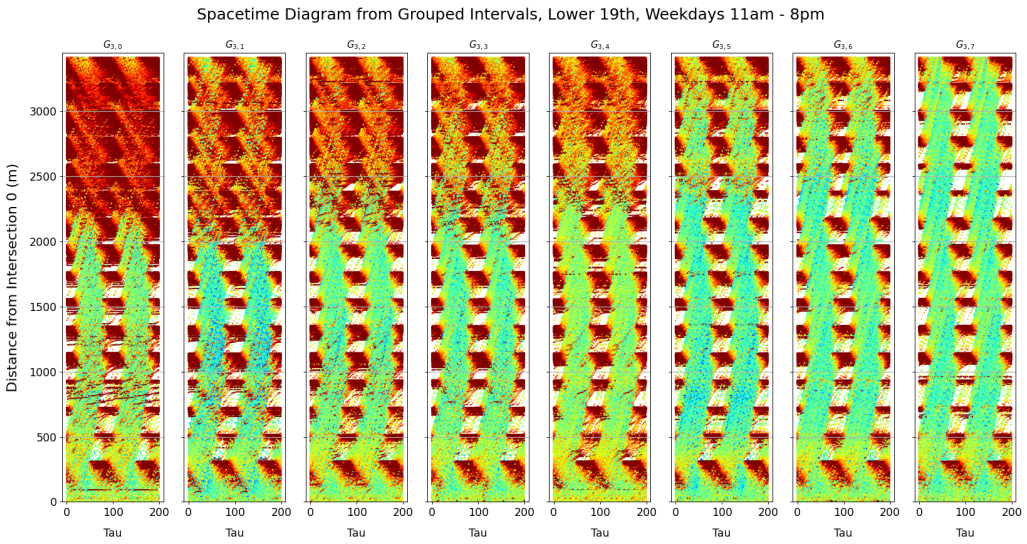


Figure 6.16: Spacetime clusters for lower 19th Avenue 100 second cycletimes

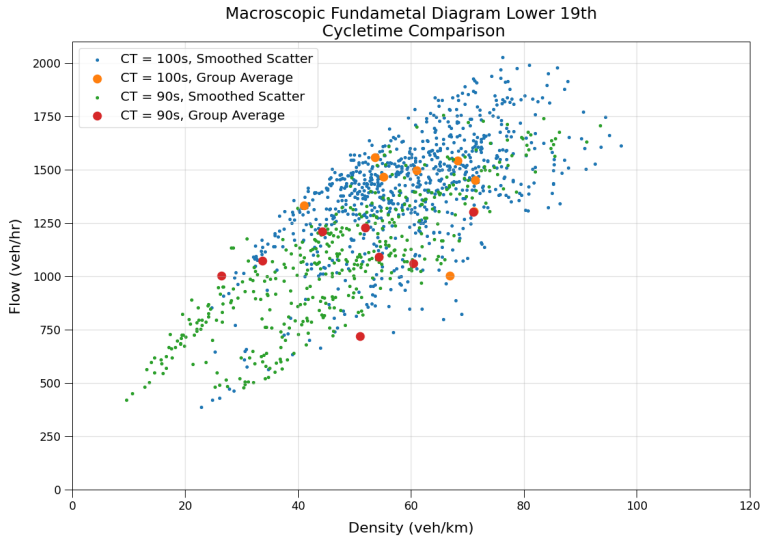


Figure 6.17: MFD comparisons for northbound on lower 19th Avenue. Compares 90 second vs 100 second cycletimes.

Chapter 7

Maneuvers and Risk

At 36,000 driving related deaths in 2018, improvements to traffic safety are a major area of research and investment[2]. Large investments have been made towards the development of autonomous vehicles (AVs) with hopes to remove human drivers from the equation and reduce the number of traffic deaths to zero. However, driving is an inherently risky task, where the risks involved are not unique to human drivers [59].

A common metric for evaluating the safety performance of AVs is the total miles driven. Here the goal for AVs is to out scale humans on the number of miles driven without incident. This metric is shown to be infeasible in achieving a statistically significant measure of reliability for AVs [41]. Also, it does not address many of the nuanced risks associated with driving. For example, 7% of crashes involve left turns but left turns make up a much smaller percentage of total driving time [46]. Left turns are a much higher risk maneuver.

In [43] a metric for evaluating maneuver-level crash risk on the roads. This metric can be used to serve as a baseline for judging AVs driving. it also provides the means for calculating more risk averse pathing. Further, by comparing maneuver level risk over an area, we can highlight locations of elevated risk to further understand the factors that increase risk.

The work in originally evaluated maneuver-level risk using the September 2019 trace dataset for LA corridors. Here we expand on that work to view the wider SF area.

7.1 Maneuver-Level Crash Risk

We use the model introduced in [43]. The total number crashes involving a maneuver at an intersection, C_m^i, is modeled as a binomial random variable parameterized by the number of executions of that maneuver, N_m^i and the inherent probability of an accident involving that maneuver, p_m^i.

$$C_m^i \sim \text{Bin}(N_m^i, p_m^i)$$

Thus the maximum likelihood estimation of p_m^i is the simply,

$$\hat{p}_m^i = \frac{C_m^i}{N_m^i}$$

7.2 Maneuver-Level Data

For these estimates we need data on the number of crashes involving that maneuver and the number of maneuvers taken.

To get the number of accidents per intersection we utilize the Transportation Injury Mapping System (TIMS) [52]. From the database we can extract data from the years 2011-2020 and filter out intersection accidents involving the maneuver of interest.

Accidents are relatively infrequent for each intersection. The maximum number of crashes for 1 intersection is only 63 over the 10 years of data. This means we have to consider data over a large time frame to get significant values for the number of crashes. This also means that the crash data and maneuver data from our cv samples are from different years.

To address this we bin the number of accidents in SF by traffic analysis zone (TAZ). We then plot the number of accidents before 2016 vs after 2016 for each TAZ.

Figure 7.1 shows that there is a strong correlation of accidents to location. The number of accidents per TAZ is consistent across the two 5 year periods.

From the connected vehicle trace data we can extract maneuvers taken by each vehicle for each intersection they cross. A map snapping algorithm was used to snap GPS traces to a network path and extract the maneuvers. Total counts of maneuvers are scaled by a flat penetration rate of 2% for the entire city. This rough estimate is okay as we could only reasonably expect evaluations close in order of magnitude. Total counts for the total survey time of 32 days are evaluated for an average hourly rate and extrapolated to expected totals over the course of 10 years.

7.3 Results

For risk estimates we apply some light filtering of the data to reduce outliers.

First, we only calculate maneuver level risk for intersections that experience 2 or more crashes involving that maneuver. We can reason that there is a baseline risk across the city and so some intersections around the baseline will have an accident occur. However because of the infrequency that single sample may skew the risk estimate for the intersection.

Second, we only evaluate intersections where flow rates for specific maneuvers are high. This is done because of the low penetration rate of the dataset. We have to consider locations

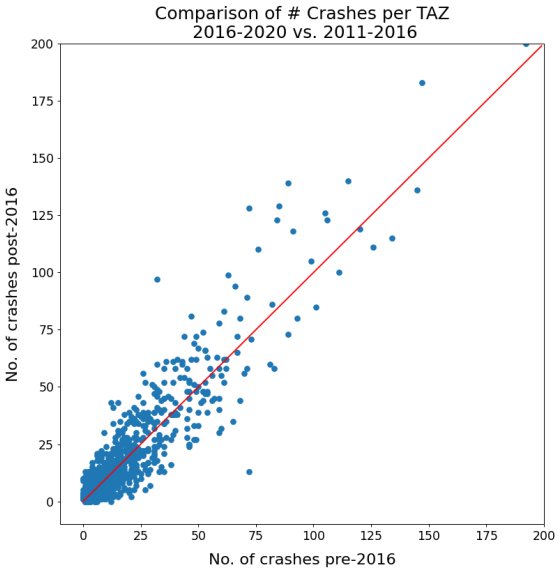


Figure 7.1: Number of accidents before 2016 vs number of accidents after 2016

of high flow rate to consider the sample set representative. We cut off intersections at less than 65 maneuvers per hour.

Tables 7.1, 7.2, 7.3, show the evaluations of the top 5 intersections with the highest risk for each maneuver. Figures 7.2,7.3, and 7.4 map the results of these risk calculations. The left side map of the figure compares the total number of maneuver type crashes across intersections. The larger dot indicates more crashes relative to each other. The right side map compares the calculated risk for that maneuver across intersections. Large dots indicate higher relative risk.

7.4 Discussion

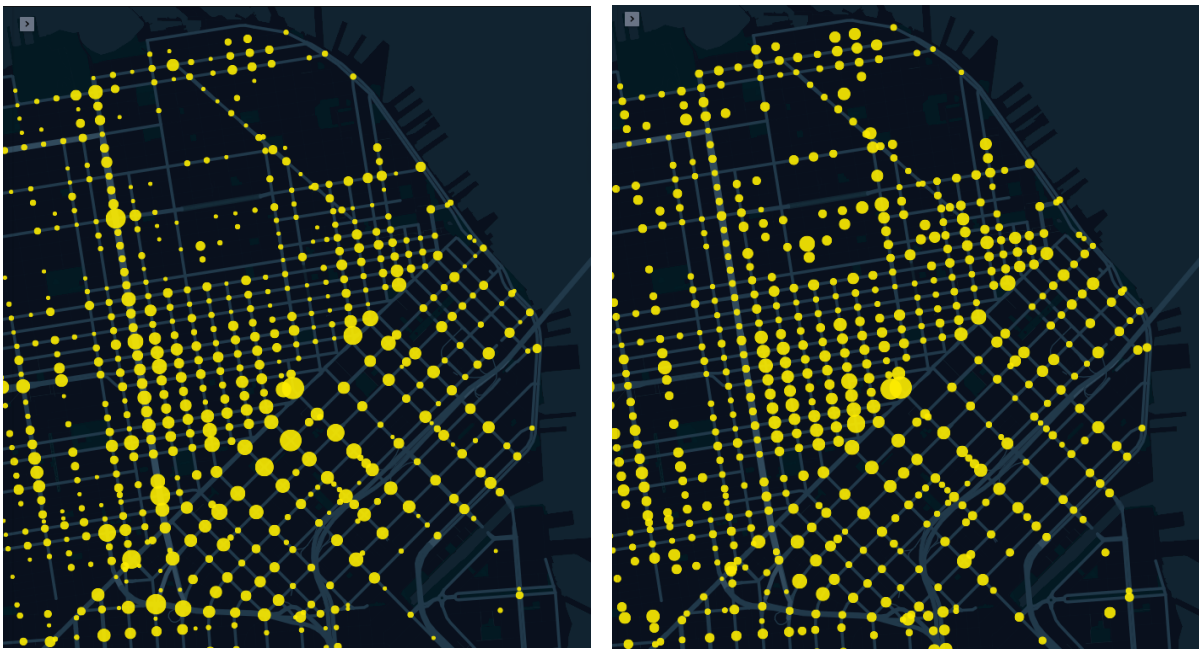
With the probabilities calculated the total risk of a particular path can be estimated by the sum total of risks for each maneuver taken on the path. This could be used to evaluate the risk of paths already traveled by AVs. We can also optimize travel between origin and destination to choose the path of least risk. Further we can study areas in the city that are the most difficult to travel too and areas that are most difficult to travel from. With more than a month of data, we could get more reliable data for intersections with low flow and more representative sample for maneuver rates across the city. Particularly, intersections with low flow for a maneuver but high number of crashes would reveal a high crash risk.

Through Risk

OSM ID	Thru Crash Probability	Lower 95% CI	Upper 95% CI	Thru Crashes	Thru Rate (num/hr)
65313131	2.2E-06	1.6E-06	3.0E-06	44	225.7
65326744	1.9E-06	1.2E-06	2.9E-06	22	132.0
65334878	1.6E-06	9.5E-07	2.6E-06	16	114.2
65342049	1.5E-06	9.5E-07	2.3E-06	20	151.4
65308264	1.4E-06	6.9E-07	2.5E-06	10	83.7

Table 7.1: Top 5 intersections for through risk

Through Risk



Total Through Crashes Through Risk

Figure 7.2: Comparison of total through crashes (left) and through crash risk (right)

Left Risk

OSM ID	LT Crash Probability	Lower 95% CI	Upper 95% CI	LT Crashes	LT Rate (num/hr)
65317399	2.4E-06	1.4E-06	4.0E-06	14	67.3
65296837	2.3E-06	1.4E-06	3.8E-06	15	73.3
65306810	2.3E-06	1.3E-06	3.9E-06	14	68.8
65317572	2.0E-06	1.1E-06	3.4E-06	13	75.4
65317045	1.9E-06	1.1E-06	3.3E-06	13	77.3

Table 7.2: Top 5 intersections for left-turn risk

Left Risk

Total LT Crashes LT Risk

Figure 7.3: Comparison of total left-turn crashes (left) and left-turn crash risk (right)

Right Risk

OSM ID	RT Crash Probability	Lower 95% CI	Upper 95% CI	RT Crashes	RT Rate (num/hr)
65363166	1.1E-06	5.4E-07	2.1E-06	9	93.4
65310823	1.0E-06	4.5E-07	2.1E-06	7	77.9
65303696	8.0E-07	3.2E-07	1.7E-06	6	85.7
65352330	6.9E-07	2.5E-07	1.6E-06	5	82.4
295083424	6.9E-07	2.8E-07	1.5E-06	6	99.8

Table 7.3: Top 5 intersections for right-turn risk

Right Turn Risk

Total RT Crashes RT Risk

Figure 7.4: Comparison of total right-turn crashes (left) and right-turn crash risk (right)

Chapter 8

Concluding Remarks

There is large potential for connected vehicle data to improve measurement and understanding of the traffic network. This work exemplifies that even at low penetration rates there is a wealth of information readily available. The level of detail is set to improve as more vehicles on the road adapt the technology. This work only scratches the surface of potential applications. It will be exciting to see traffic design improvements and the implementation of real-time network estimation and control based off this technology.

Bibliography

[1] Vision Zero Network 2022. URL: https://visionzeronetwork.org/about/vision-zero-network/ (visited on 12/03/2022).

[2] *2018 Fatal Motor Vehicle Crashes: Overview*. Tech. rep. DOT HS 812 826. Washington, D.C., United States: National Center for Statistics and Analysis, Aug. 2019.

[3] Zahra Amini et al. "Queue-length estimation using real-time traffic data". In: *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. 2016.

[4] "An analytical approximation for the macroscopic fundamental diagram of urban traffic". In: *Transportation Research Part B: Methodological* 42.9 (2008), pp. 771–781. ISSN: 0191-2615. DOI: https://doi.org/10.1016/j.trb.2008.06.008. URL: https://www.sciencedirect.com/science/article/pii/S0191261508000799.

[5] Kevin N Balke, Hassan A Charara, and Ricky Parker. *Development of a traffic signal performance measurement system (TSPMS)*. Tech. rep. Texas Transportation Institute, Texas A & M University System College, 2005.

[6] Xuegang Jeff Ban, Peng Hao, and Zhanbo Sun. "Real time queue length estimation for signalized intersections using travel times from mobile sensors". In: *Transportation Research Part C: Emerging Technologies* (2011).

[7] Emmanouil Barmpounakis and Nikolas Geroliminis. "On the new era of urban traffic monitoring with massive drone data: The pNEUMA large-scale field experiment". In: *Transportation Research Part C: Emerging Technologies* 111 (2020), pp. 50–71. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.11.023. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19310320.

[8] Peter J. Bickel et al. "Measuring Traffic". In: *Statistical Science* 22.4 (Nov. 2007). ISSN: 0883-4237. DOI: 10.1214/07-sts238. URL: http://dx.doi.org/10.1214/07-STS238.

[9] Lawrence Blincoe et al. *The Economic and Societal Impact Of Motor Vehicle Crashes, 2010 (Revised)*. Tech. rep. DOT HS 812 013. Washington, D.C., United States: U.S. Department of Transportation, National Highway Traffic Safety Administration, May 2015.

[10] Noelia Caceres et al. "Traffic Flow Estimation Models Using Cellular Phone Data". In: *IEEE Transactions on Intelligent Transportation Systems* 13 (2012). ISSN: 1524-9050.

[11] *California Department of Transportation Performance Measurement System (PEMS)*. CalTrans. URL: https://pems.dot.ca.gov (visited on 03/02/2022).

[12] Tang-Hsien Chang and Jen-Ting Lin. "Optimal signal timing for an oversaturated intersection". In: *Transportation Research Part B: Methodological* (2000).

[13] Yihua Chen and John Krumm. "Probabilistic Modeling of Traffic Lanes from GPS Traces". In: *ACM GIS* (2010).

[14] Yang Cheng et al. "An exploratory shockwave approach to estimating queue length using probe trajectories". In: *Journal of intelligent transportation systems* (2012).

[15] Irene Chia et al. "Evaluation of Actuated, Coordinated, and Adaptive Signal Control Systems: A Case Study". In: *Journal of Transportation Engineering, Part A: Systems* 143.9 (2017), p. 05017007.

[16] Gurcan Comert. "Queue length estimation from probe vehicles at isolated intersections: Estimators for primary parameters". In: *European Journal of Operational Research* (2016).

[17] Gurcan Comert. "Simple analytical models for estimating the queue lengths from probe vehicles at traffic signals". In: *Transportation Research Part B: Methodological* (2013).

[18] Gurcan Comert and Mecit Cetin. "Analytical evaluation of the error in queue length estimation at traffic signals from probe vehicle data". In: *IEEE Transactions on Intelligent Transportation Systems* (2011).

[19] Carlos F. Daganzo. "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory". In: *Transportation Research Part B: Methodological* 28 (1994).

[20] Carlos F. Daganzo. "Urban gridlock: Macroscopic modeling and mitigation approaches". In: *Transportation Research Part B: Methodological* 41.1 (2007), pp. 49–62. ISSN: 0191-2615. DOI: https://doi.org/10.1016/j.trb.2006.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0191261506000282.

[21] L. C. Edie. "Discussion of Traffic Stream Measurements and Definitions". In: *2nd Symposium on the Theory of Traffic FLow* (1965).

[22] Nikolaos Geroliminis and Alexander Skabardonis. "Prediction of arrival profiles and queue lengths along signalized arterials by using a Markov decision process". In: *Transportation Research Record* (2005).

[23] Nikolas Geroliminis and Carlos F. Daganzo. "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings". In: *Transportation Research Part B: Methodological* 42.9 (2008), pp. 759–770. ISSN: 0191-2615. DOI: https://doi.org/10.1016/j.trb.2008.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0191261508000180.

[24] Nikolas Geroliminis and Jie Sun. "Properties of a well-defined macroscopic fundamental diagram for urban traffic". In: *Transportation Research Part B: Methodological* 45.3 (2011), pp. 605–617. ISSN: 0191-2615. DOI: https://doi.org/10.1016/j.trb.2010.11.004. URL: https://www.sciencedirect.com/science/article/pii/S0191261510001372.

[25] Peng Hao and Xuegang Ban. "Long queue estimation for signalized intersections using mobile data". In: *Transportation Research Part B: Methodological* (2015).

[26] Peng Hao et al. "Cycle-by-cycle intersection queue length distribution estimation using sample travel times". In: *Transportation research part B: methodological* 68 (2014), pp. 185–204.

[27] Peng Hao et al. "Vehicle index estimation for signalized intersections using sample travel times". In: *Procedia-Social and Behavioral Sciences* (2013).

[28] M. Hunter et al. "Estimation of Connected Vehicle Penetration on US Roads in Indiana, Ohio, and Pennsylvania". In: *Journal of Transportation Tecnologies* 11.4 (2021), pp. 597–610. DOI: 10.4236/jtts.2021.114037.

[29] *Infrastructure/Road Congestion Economic Impact Study and Survey*. Tech. rep. Washington, D.C., United States: U.S. Travel Association, May 2019.

[30] Fuliang Li et al. "Real-time queue length estimation for signalized intersections using vehicle trajectory data". In: *Transportation Research Record* (2017).

[31] Michael James Lighthill and Gerald Beresford Whitham. "On kinematic waves II. A theory of traffic flow on long crowded roads". In: *Proceedings of the Royal Society A* 229 (1178 May 1955). DOI: https://doi.org/10.1098/rspa.1955.0089.

[32] Henry X Liu et al. "Real-time queue length estimation for congested signalized intersections". In: *Transportation research part C: emerging technologies* (2009).

[33] Claudio Lombari, Luis Picado-Santos, and Anuradha M. Annaswamy. "Model-Based Dynamic Toll Pricing: An Overview". In: *Applied Sciences* (2021). DOI: https://doi.org/10.3390/app11114778.

[34] Pablo Alvarez Lopez et al. "Microscopic Traffic Simulation using SUMO". In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 2018, pp. 2575–2582. DOI: 10.1109/ITSC.2018.8569938.

[35] Highway Capacity Manual. "HCM2010". In: *Transportation Research Board, National Research Council, Washington, DC* (2010), p. 1207.

[36] Pitu B Mirchandani and Ning Zou. "Queuing models for analysis of traffic adaptive signal control". In: *IEEE Transactions on Intelligent Transportation Systems* (2007).

[37] Andrew S. Nagle and Vikash V. Gayah. "Accuracy of Networkwide Traffic States Estimated from Mobile Probe Data". In: *Transportation Research Record: Journal of the Transportation Research Board* (2014). DOI: 10.3141/2421-01.

[38] Paul Newson and John Krumm. "Hidden Markov Map Matching Through Noise and Sparseness". In: *ACM GIS* (2009).

[39] *Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data*. Tech. rep. U.S Department of Transportation Federal Highway Administration, 2016. DOI: http://doi.org/10.21949/1504477.

[40] Markos Papageorgiou and Apostolos Kotsialos. "Freeway Ramp Metering: An Overview". In: *Transactions on Intelligent Transportation Systems* 3 (2002).

[41] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. "On a Formal Model of Safe and Scalable Self-driving Cars". In: *CoRR* abs/1708.06374 (2017). arXiv: 1708.06374. URL: http://arxiv.org/abs/1708.06374.

[42] Anuj Sharma, Darcy M Bullock, and James A Bonneson. "Input–output and hybrid techniques for real-time prediction of delay and maximum queue length at signalized intersections". In: *Transportation Research Record* (2007).

[43] Akhil Shetty et al. "Risk Assessment of Autonomous Vehicles across Diverse Driving Contexts". In: *International Intelligent Transportation Systems Conference* (2021). DOI: 10.1109/ITSC48978.2021.9564744.

[44] Alexander Skabardonis and Nikolas Geroliminis. "Real-time monitoring and control on signalized arterials". In: *Journal of Intelligent Transportation Systems* 12.2 (2008), pp. 64–74.

[45] Jack Snowdon et al. "Spatiotemporal Traffic Volume Estimation Model Based on GPS Samples". In: *Proceedings of the Fifth International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data*. GeoRich'18. Houston, TX, USA: Association for Computing Machinery, 2018, pp. 1–6. ISBN: 9781450358323. DOI: 10.1145/3210272.3210273. URL: https://doi.org/10.1145/3210272.3210273.

[46] *Statistics of Light-Vehicle Pre-Crash Scenarios Based on 2011-2015 National Crash Data*. Tech. rep. DOT HS 812 745. Washington, D.C., United States: National Tansportation Systems Center, Aug. 2019. URL: https://rosap.ntl.bts.gov/view/dot/41932.

[47] Chaopeng Tan et al. "Cycle-Based Queue Length Estimation for Signalized Intersections Using Sparse Vehicle Trajectory Data". In: *IEEE Transactions on Intelligent Transportation Systems* (2021).

[48] Hamidreza Tavafoghi et al. "Queue Length Estimation from Connected Vehicles with Low and Unknown Penetration Level". In: *International Intelligent Transportation Systems Conference* (2021). DOI: 10.1109/ITSC48978.2021.9564477.

[49] Hamidreza Tavafoghi et al. "Report for UAS4T Competition". In: *23rd IEEE International Conference on Intelligent Transportation* (2020). URL: www.hamidtavaf.github.io/UAS4T.pdf.

[50] *The future economic and envrionmental costs of gridlock in 2030*. Tech. rep. London, United Kingdom: CEBR (Center for Economics and Buisness Research), July 2014.

[51] Kamonthep Tiaprasert et al. "Queue length estimation using connected vehicle technology for adaptive signal control". In: *IEEE Transactions on Intelligent Transportation Systems* (2015).

[52] *Transportation Injury Mapping System (TIMS)*. Safe Transportation Research and Education Center. URL: https://tims.berkeley.edu (visited on 03/02/2022).

[53] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. "Congested traffic states in empirical observations and microscopic simulations". In: *Physical Review E* 62.2 (Aug. 2000), pp. 1805–1824. DOI: 10.1103/physreve.62.1805. URL: https://doi.org/10.1103%2Fphysreve.62.1805.

[54] Martin Treiber and Arne Kesting. *Traffic Flow Dynamics*. Berlin: Springer, 2013. ISBN: 978-3-642-32459-8. DOI: 10.1007/978-3-642-32460-4.

[55] *Using Mobile Device Samples to Estimate Traffic Volumes*. Tech. rep. 395 John ireland Boulevard, MS 330, St. Paul, Minnesota 55155: Minnesota Department of Transportation, Dec. 2017.

[56] Pravin Varaiya. "Max pressure control of a network of signalized intersections". In: *Transportation Research Part C: Emerging Technologies* (2013).

[57] W. Vickrey. "Congestion in midtown Manhattan in relation to marginal cost pricing". In: *Economics of Transportation* (2020). DOI: https://doi.org/10.1016/j.ecotra.2019.100152.

[58] Miao Wang et al. "Real-Time Path Planning Based on Hybrid-VANET_Enhanced Transportation System". In: *Transaction on Vehicular Technology* (2015). DOI: 10.1109/TVT.2014.2335201.

[59] "What humanlike errors do autonomous vehicles need to avoid to maximize safety?" In: *Journal of Safety Research* 75 (2020), pp. 310–318. ISSN: 0022-4375. DOI: https://doi.org/10.1016/j.jsr.2020.10.005. URL: https://www.sciencedirect.com/science/article/pii/S0022437520301262.

[60] Ward Whitt. "Time-varying queues". In: *Queueing models and service management* 1.2 (2018).

[61] Wai Wong et al. "On the estimation of connected vehicle penetration rate based on single-source connected vehicle data". In: *Transportation Research Part B: Methodological* 126 (2019), pp. 169–191. ISSN: 0191-2615. DOI: https://doi.org/10.1016/j.trb.2019.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0191261518309834.

[62] Feng Xie and David Levinson. "Modeling the Growth of Transportation Networks: A Comprehensive Review". In: *Network and Spatial Economics* (2007). DOI: 10.1007/s11067-007-9037-4.

[63] Juyuan Yin, Jian Sun, and Keshuang Tang. "A Kalman filter-based queue length estimation method with low-penetration mobile sensor data at signalized intersections". In: *Transportation Research Record* (2018).

[64] Xianyuan Zhan, Ruimin Li, and Satish V Ukkusuri. "Lane-based real-time queue length estimation using license plate recognition data". In: *Transportation Research Part C: Emerging Technologies* (2015).

[65] Yi Zhao and Zong Tian. "An overview of the usage of adaptive signal control system in the United States of America". In: *Applied Mechanics and Materials*. 2012.

[66] Jianfeng Zheng and Henry X Liu. "Estimating traffic volumes for signalized intersections using connected vehicle data". In: *Transportation Research Part C: Emerging Technologies* (2017).