# Implementing a Community-Based Mobility Lab: improving Traffic, Protecting Data Privacy

July 2024

A Research Report from the Pacific Southwest Region University Transportation Center

Tyler Reeb, Ph.D., College of Professional and Continuing Education, CSULB

Gwen Shaffer, Ph.D., College of Liberal Arts, California State University - Long Beach (CSULB)

Hossein Jula, Ph.D., College of Engineering, CSULB

Pacific Southwest Region UTC

University Transportation Center

CALIFORNIA STATE UNIVERSITY LONG BEACH 1949

METRANS Transportation Consortium

# TECHNICAL REPORT DOCUMENTATION PAGE

**1. Report No.**
PSR-22-02 TO 063

**2. Government Accession No.**
N/A

**3. Recipient's Catalog No.**
N/A

**4. Title and Subtitle**
Implementing a Community-Based Mobility Lab: improving Traffic, Protecting Data Privacy

**5. Report Date**
7/31/24

**6. Performing Organization Code**
N/A

**7. Author(s)**
Tyler Reeb, 0000-0002-8056-9939
Gwen Shaffer, 0000-0001-6425-960X
Hossein Jula, 0000-0002-6380-1839

**8. Performing Organization Report No.**
PSR-22-02 TO 063

**9. Performing Organization Name and Address**
METRANS Transportation Center
University of Southern California
University Park Campus, RGL 216
Los Angeles, CA 90089-0626

**10. Work Unit No.**
N/A

**11. Contract or Grant No.**
USDOT Grant 69A3551747109
[65A0674TO063]

**12. Sponsoring Agency Name and Address**
U.S. Department of Transportation
Office of the Assistant Secretary for Research and Technology
1200 New Jersey Avenue, SE, Washington, DC 20590

**13. Type of Report and Period Covered**
Final report (8/1/2023 – 7/31/2023)

**14. Sponsoring Agency Code**
USDOT OST-R

**15. Supplementary Notes**
https://metrans.org/research/implementing-a-community-based-mobility-lab-improving-traffic-protecting-data-privacy

**16. Abstract**
This study pilots a scalable methodology to conduct community-based mobility research to address critical safety, environmental, and traffic congestion issues while also addressing the data privacy issues such as transparency and accountability associated with the extensive data collections necessary for predictive traffic modeling.

The study provides a blueprint for a three-phase scalable methodological approach to community-based mobility studies including the collection of video footage, the data processing for the predictive traffic modeling technology and the approach to addressing related data privacy concerns. A key focus of this study regards the establishment of an efficient and practical way of balancing between (1) gathering the extensive data required to model the intersection of study as it relates to addressing critical safety, environmental, and traffic congestion issues with (2) a corresponding need to respect residents' concerns about smart technologies and data privacy. Using data from the video recordings, the research team used an artificial neural network to develop a predictive traffic model. To address data privacy concerns for residents living near the intersection of study during the footage collection, researchers conducted data walks – a 1.5-mile guided walk through the study area during which researchers prompted study participants to reflect on various smart technologies, including traffic cameras and the recording sessions in place.

The study focuses on a singular node in the West Long Beach transportation network, but its findings can be used as a resource for community-based mobility research addressing critical safety, environmental, and traffic congestion issues at other intersections, corridors, and networks of intersections and corridors not only across California but across the country.

**17. Key Words**
Artificial intelligence, digital privacy, urban mobility

**18. Distribution Statement**
No restrictions.

| 19. Security Classif. (of this report) | 20. Security Classif. (of this page) | 21. No. of Pages | 22. Price |
|---|---|---|---|
| Unclassified | Unclassified | 56 | N/A |

# Contents

# List of Figures

# List of Tables

# Appendix

# List of Acronyms and Abbreviations

| | |
|---|---|
| Pacific Coast Highway | PCH |
| Levels of Service | LOS |
| Intelligent Transportation Systems | ITS |
| Location-based Services | LBS |
| Connected Vehicle Data | CVD |
| Heavy Goods Vehicle | HGV |
| Vehicle Hours of Delay | VHD |
| Turning Movement Count | TMC |
| Department of Transportation | DOT |
| Artificial Neural Network | ANN |

# About the Pacific Southwest Region University Transportation Center

The Pacific Southwest Region University Transportation Center (UTC) is the Region 9 University Transportation Center funded under the US Department of Transportation's University Transportation Centers Program. Established in 2016, the Pacific Southwest Region UTC (PSR) is led by the University of Southern California and includes seven partners: Long Beach State University; University of California, Davis; University of California, Irvine; University of California, Los Angeles; University of Hawaii; Northern Arizona University; Pima Community College.

The Pacific Southwest Region UTC conducts an integrated, multidisciplinary program of research, education and technology transfer aimed at *improving the mobility of people and goods throughout the region.* Our program is organized around four themes: 1) technology to address transportation problems and improve mobility; 2) improving mobility for vulnerable populations; 3) Improving resilience and protecting the environment; and 4) managing mobility in high growth areas.

# U.S. Department of Transportation (USDOT) Disclaimer



# California Department of Transportation (CALTRANS) Disclaimer



# Disclosure

Principal Investigator Tyler Reeb, Co-Principal Investigators Gwen Shaffer, Hossein Jula and others, conducted this research titled, "Implementing a Community-Based Mobility Lab: improving Traffic, Protecting Data Privacy" at California State University, Long Beach. The research took place from August 1, 2023 to July 31, 2024 and was funded by a Pacific Southwest Region University Transportation Center (PSR) grant in the amount of $104,400.00. The research was conducted as part of the PSR research program.

Pacific Southwest Region UTC
University Transportation Center

## Acknowledgements

The authors acknowledge the time, expertise, and thoughtfulness that our partners at Caltrans, City of Long Beach, StreetLight, Iteris, and Jet Propulsion Laboratory contributed to the development of this project. The Center for International Trade and Transportation (CITT) would like to thank CITT Project Manager Lena Wild and CITT Research Associate Devin Martinez-Flores for their range of contributions to this project. The Center for International Trade and Transportation would like to thank Research Assistants Hung Jui Chang, Kiana Nguyen, and Sai Nikhila Chittari Vyacharouthu for their contributions.

# Abstract

This study pilots a scalable methodology to conduct community-based mobility research to address critical safety, environmental, and traffic congestion issues while also addressing the data privacy issues such as transparency and accountability associated with the extensive data collections necessary for predictive traffic modeling.

The study provides a blueprint for a three-phase scalable methodological approach to community-based mobility studies including the collection of video footage, the data processing for the predictive traffic modeling technology and the approach to addressing related data privacy concerns. A key focus of this study regards the establishment of an efficient and practical way of balancing between (1) gathering the extensive data required to model the intersection of study as it relates to addressing critical safety, environmental, and traffic congestion issues with (2) a corresponding need to respect residents' concerns about smart technologies and data privacy. Using data from the video recordings, the research team used an artificial neural network to develop a predictive traffic model. To address data privacy concerns for residents living near the intersection of study during the footage collection, researchers conducted data walks – a 1.5-mile guided walk through the study area during which researchers prompted study participants to reflect on various smart technologies, including traffic cameras and the recording sessions in place.

The study focuses on a singular node in the West Long Beach transportation network, but its findings can be used as a resource for community-based mobility research addressing critical safety, environmental, and traffic congestion issues at other intersections, corridors, and networks of intersections and corridors not only across California but across the country.

# Implementing a Community-Based Mobility Lab: improving Traffic, Protecting Data Privacy

## Executive Summary

In this study, the Center for International Trade and Transportation (CITT) research team piloted a scalable methodology to conduct community-based mobility research to address critical safety, environmental, and traffic congestion issues while also addressing the data privacy issues associated with the extensive data collections necessary for predictive traffic modeling. Based upon the preliminary findings from a previous Pacific Southwest Region UTC (PSR) study entitled "Using Artificial Intelligence to Improve Traffic Flows, with Consideration of Data Privacy," this study provides a blueprint for a three-phase methodological approach to community-based mobility studies including the collection of video footage, the data processing for the predictive traffic modeling technology and the approach to addressing related data-privacy concerns.

To develop a scalable methodology for community-based research that called for 1) recording footage of the intersection of study, 2) extracting mobility data from that recorded footage into predictive traffic modeling technology, and 3) accounting for related data privacy concerns, the research team implemented three phases to complete the study.

The first phase called for the collection of 25-30 hours of footage from the chosen intersection of study at Pacific Coast Highway (PCH) and Santa Fe Avenue in West Long Beach. The research team identified several possible ways to collect the necessary footage, including a formal request to Long Beach City Hall requesting permission to access closed-circuit television (CCTV) footage from traffic cameras, but ended up utilizing Do-It-Yourself (DIY) approach to obtain the necessary footage. This approach required the use of two GoPro cameras on mounted tripods to record the footage manually. The research team worked with a data platform called StreetLight, and later on with the private-sector firm, Iteris, which granted free access to its real-time data platform, Iteris ClearGuide, to gain a baseline understanding of traffic flow at the intersection of study.

To process the data captured in the footage, researchers utilized DataFromSky, an Intelligent Transportation Systems company, to convert the raw data, originally in MP4 files, into a format which an Artificial Neural Network (ANN) would then use to train the model simulator. The utilization of a data processing company allowed for a quick and insightful analysis of the collected footage and would prove indispensable in efforts to scale up community-based transportation research.

To develop a predictive traffic simulator for the chosen transportation node, researchers analyzed and evaluated multiple modeling methods. To train the ANN, researchers used data from the Caltrans Performance Measurement System (PeMS). Once the recorded footage was processed by DataFromSky, researchers then used the extracted data to train the predictive traffic simulator.

To address data privacy concerns for residents living near the intersection of study during the footage collection, researchers conducted a 1.5-mile data walk around the neighborhood prompting study participants to reflect on various smart technologies, including traffic cameras and the recording sessions in place.

The findings of this study serve as a blueprint for future community-based mobility research at other intersections, corridors, and networks of intersections and corridors not only across California but across the country. The research team used the findings of this study to make policy recommendations such as deploying high-altitude balloons or using low-earth and/or geosynchronous orbiting satellites to capture traffic footage for future community-based mobility studies.

# Introduction

For this project, the Center for International Trade and Transportation (CITT) research team piloted a scalable methodology to conduct community-based mobility research to address critical safety, environmental, and traffic congestion issues while also addressing the data privacy issues associated with collecting the necessary data. This study is a continuation of a previous 2022 Pacific Southwest Region UTC (PSR) study entitled "Using Artificial Intelligence to Improve Traffic Flows, with Consideration of Data Privacy."[i] That study sought to develop an artificial neural network (ANN), "a machine learning program, or model, that makes decisions in a manner similar to the human brain"[ii] to model an intersection. The development of a predictive traffic model requires footage from the intersection, which raises data privacy concerns for anyone who passes through the intersection being recorded. The initial study faced a range of data gathering challenges related to the COVID-19 global pandemic, which limited the subsequent development of the predictive traffic model. This project builds on the preliminary findings from that report and, using the same intersection of study (Pacific Coast Highway [PCH] and Santa Fe Avenue), incorporates a methodology to ensure the gathering of 25-30 hours of footage necessary to develop the ANN and related predictive traffic model.

CITT researchers chose the intersection of PCH and Santa Fe Avenue for the following reasons:

- It was identified as a failing intersection in the Long Beach 2040 General Plan;
- The intersection is also a designated truck route on both PCH and Santa Fe Avenue (See Appendix Section 1); and
- The proximity of the intersection to the San Pedro Port Complex, a high school, a public park, a public library, a police substation, and a surrounding residential neighborhood.

Figure 1 depicts the intersection of study with a Level of Service (LOS) of E indicating poor operation in both the morning and evening. The poor LOS designation is due to a mix of heavy truck, passenger vehicle, and pedestrian traffic in all four directions as observed by researchers during multiple visits. South of the intersection, there is an industrial cluster catering to automobiles and heavy goods vehicles (HGVs) but north of the intersection is a community enclosed by PCH (State Route 1), State Route 103, which fall under Caltrans jurisdiction, and Interstate Freeways 710 and 405, which fall under Federal jurisdiction. The study focuses on just one node in a network; and what is presented in this study can be used as a resource for community-based mobility research addressing critical safety, environmental, and traffic congestion issues at other intersections, corridors, and networks of intersections and corridors. Historically, community stakeholders have had to rely on local governments to view their own data. This report suggests ways that community-based research efforts—paired with modern digital-recording, processing, analytical technologies—make it possible for residents in a community to document their own unique mobility challenges.

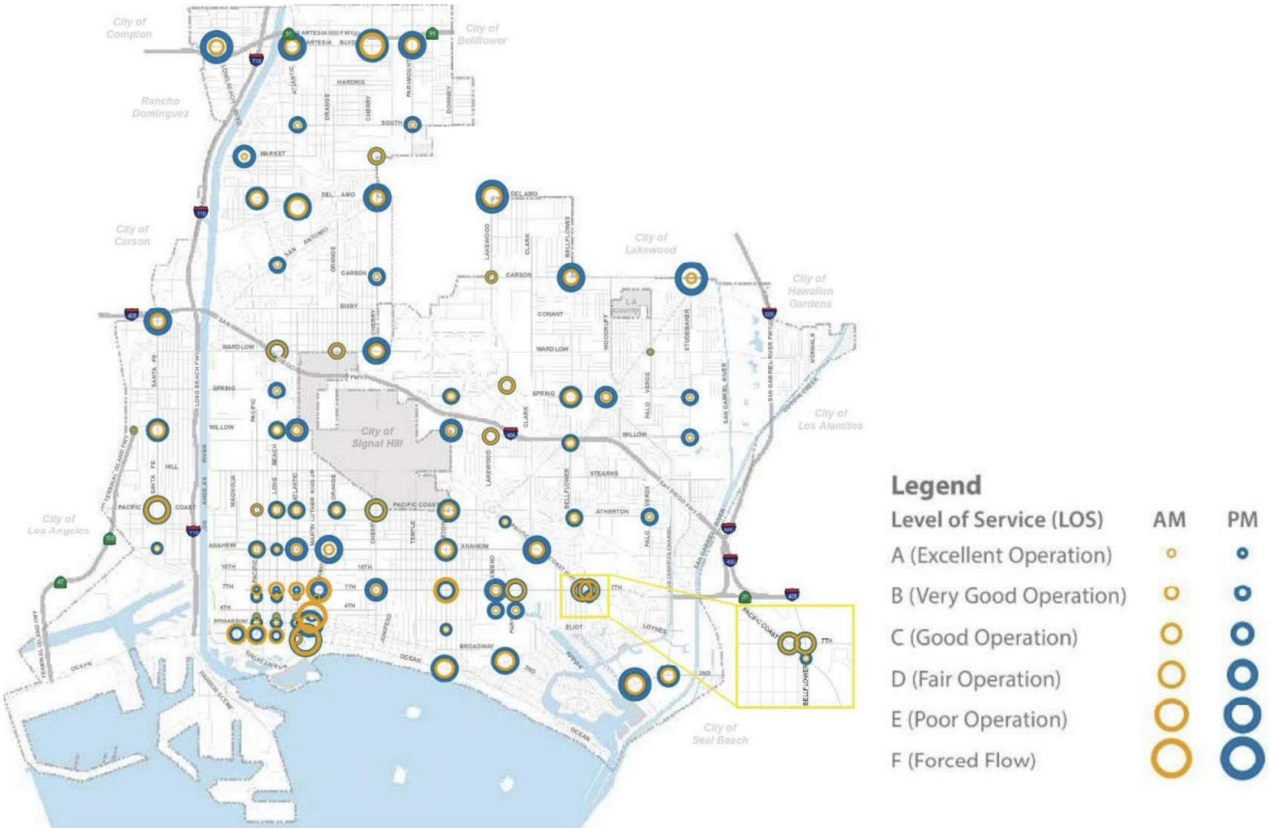



**Figure 1. Screenshot of Long Beach 2040 General Plan**

# Methodology

To develop a scalable methodology for community-based research that called for 1) recording footage of the intersection of study, 2) extracting mobility data from that recorded footage into predictive traffic modeling technology, and 3) accounting for related data privacy concerns, the research team implemented three phases to complete the study.

The first phase called for the collection of 25-30 hours of footage from the chosen intersection at PCH and Santa Fe Avenue. Researchers identified several possible ways to collect the necessary footage, starting with a formal request to Long Beach City Hall requesting permission to access closed-circuit television (CCTV) footage from traffic cameras located at the intersection. Appendix Section 2 documents the final email (from a months-long correspondence) from the City of Long Beach's Technology and Innovation Department informing the research team it would not be possible to access the CCTV feed within the timeframe of the project. The research team also explored accessing live stream footage from traffic cameras operated by Caltrans (see Figure 2 for various locations), but none of these traffic cameras were located at the intersection of study.

In anticipation that researchers would be unable to access CCTV footage from traffic cameras, researchers designed a Do-It-Yourself (DIY) approach to obtain the necessary footage. This approach included researchers using several GoPro cameras on mounted tripods to record the footage manually. This approach proved successful in meeting data gather deliverables for the study.

An unexpected resource not originally included in the scope of work for this project was the inclusion and utilization of a data platform called StreetLight, which Caltans made available to the researcher team by granting access to their vendor license. During the course of the study, researchers also worked with the private-sector firm, Iteris, which granted free access to its real-time data platform, Iteris ClearGuide.

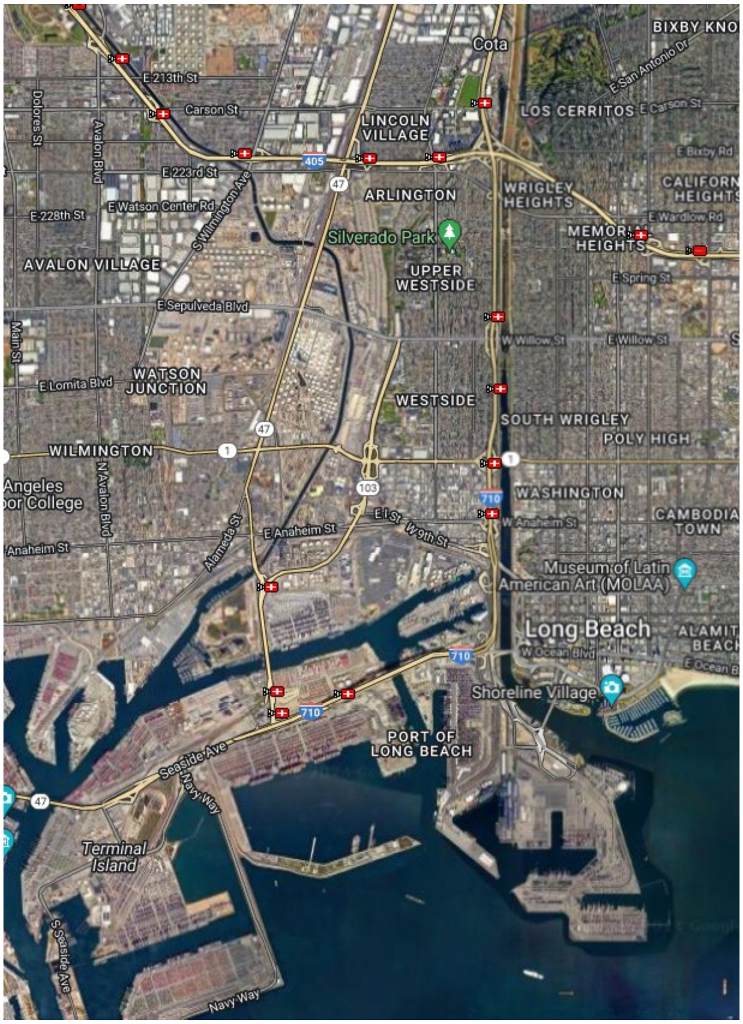



**Figure** *2*. **Caltrans CCTV map**

To address data privacy concerns of those crossing the intersection and nearby residents researchers conducted a 1.5-mile data walk around the neighborhood prompting study participants to reflect on various smart technologies, including traffic cameras and the recording sessions in place. These reflections were collected through a survey and focused discussions addressing residents' thoughts and potential concerns with smart technology and with further/other iterations of similar community-based research. Researchers then used the resulting data to make policy recommendations for future community-based mobility studies.

To process the data captured in the footage, researchers utilized an ITS company called DataFromSky to convert the raw data, originally in MP4 files, into a format which an Artificial Neural Network (ANN) would then use to train the model simulator. The utilization of a data processing company allowed for a quick and insightful analysis of the collected footage and would prove indispensable in efforts to scale up community-based transportation research.

To develop a predictive traffic simulator for the chosen transportation node, researchers analyzed and evaluated multiple modeling methods. These evaluations considered parametric models such as the Autoregressive Integrated and Moving Average (ARIMA) and the Seasonal Autoregressive Moving Average (SARIMA) as well as non-parametric models such as Feed-Forward Neural Networks (FNN) and other variations of Deep Neural networks (DNN). To train the Artificial Neural Network (ANN), researchers used data from the Caltrans Performance Measurement System (PeMS). Once the recorded footage was processed by DataFromSky, researchers then used the extracted data to train the predictive traffic simulator.

# Data Collection Findings

The community-based mobility lab model developed and implemented in this report was designed with the intention that it could be scaled and expanded upon in subsequent iterations. This study focuses on one intersection, or one node in a larger transportation network, to test the effectiveness and scalability of the methodology. The first phase in the data collection process was to find the most efficient and practical way to gather the necessary data needed to model the intersection. Determining an efficient and practical way to gather intersection data was important from a scalability standpoint because the research team wanted to implement a methodology that could easily be modified and repeated in other urban contexts.

Prior to recording footage at the intersection of study, the research team used two different traffic data platforms to gain a baseline understanding of mobility and traffic dynamics at the intersection. Traffic data platforms are a great way to begin the research process because such applications can reveal the dynamics of the intersection based upon historical data. CITT researchers used StreetLight and Iteris ClearGuide to first gain an understanding of how the intersection of study functioned. Researchers then used GoPro cameras mounted on tripods to capture footage of the intersection to confirm baseline findings from StreetLight and Iteris ClearGuide. Although traffic platforms provide valuable information about the dynamics of the intersection, they are ultimately not a substitute to actual video data. During the recording sessions researchers observed various issues with the intersection that were not present in either data platform.

## Pre-Collection

In a previous study[iii] researchers created a simulation of the intersection using 10-minute videos of the intersection that researchers captured. This process was repeated on a second day to help diversify the data collected from the video. Ultimately, researchers concluded that data from various times of day and days of the week was needed to better capture the dynamics of the intersection.

To make the simulation in this study as accurate as possible, researchers gathered 25 hours of footage at the intersection during both peak traffic hours and non-peak hours as well as various days of the week. The purpose in doing so was to provide a bigger sample size so that the simulator could more accurately reflect how the intersection behaves at various times of day throughout the week.

CITT researchers initially attempted to collect the first phase of the research by accessing the sources already in operation by the City of Long Beach. Researchers submitted several requests to the city to access the CCTV camera footage already present at the intersection since utilizing the present infrastructure would be the most scalable solution. This was a necessary first step, because, in theory, any community could make the same request. Researchers reached out to a contact in the Technology and Innovation Department of the City of Long Beach. That contact recommended finding out who owns the cameras at that intersection, although they suspected it was either Caltrans or Long Beach Department of Public Works. The research team learned that Caltrans has its cameras set up for public livestreaming but there were none posted on the intersection of interest. Reaching out to the Long Beah Public Works line, a City of Long Beach representative recommended creating a public records request to determine whether the traffic cameras at the intersection fell under the jurisdiction of local law enforcement, the City of Long Beach, or Caltrans.

Filing the Public Records request resulted in the researchers being bounced back and forth among the City of Long Beach's departments of Technology and Innovation, Police, and Public Works to determine who oversees the cameras at the intersection. Finally, an Infrastructure Systems Officer informed the researchers that the CCTV system was only available to internal staff at Long Beach Public Works and the intersection was supposed to be maintained by the California Department of Transportation. If researchers wished to gain access to the cameras, they would have to send a request to the California Department of Transportation or enter a partnership with the City of Long Beach that would require approval from the Director of Public works, the Long Beach City Council, the City Attorney, and the City Manager. The approval process for such agreements would have taken far longer than the timetable of the study, which compelled researchers to enact the DIY method.

The Streetlight platform granted researchers access to tools that made it possible to understand the traffic dynamics of the intersection prior to the footage gathering period. Researchers initially reached out to the Iteris in hopes of gaining access to their video-radar hybrid detection sensors and their existing relation to Caltrans. Unfortunately, Iteris was unable to provide researchers with any video footage of the intersection but were able to set up a trial account for their ClearGuide platform. Similarly to StreetLight, The Iteris platform allows researchers to view roadway data which alongside the StreetLight data can create a clear image of how the intersection typically operates. Unable to find any alternatives to the data gathering process through industry solutions, researchers captured necessary footage using GoPro cameras and multiple tripods.

## Traffic Findings

### StreetLight

To better inform the data collection process, researchers used StreetLight and Iteris Clear Guide to extract as much data from the intersection as possible. StreetLight InSight creates analyses based on location-based services (LBS), connected vehicle (CVD), and aggregated GPS data. The analytical tools used for this report were limited to LBS data from 2016 to 2022 and CVD from 2022 to 2023. The StreetLight platform has the capacity to differentiate between trucks, automobiles, and pedestrians, but the specific analytic tools used for this report did not have the capacity to do so. This limitation means that analyses performed by the platform takes all vehicles on the road into consideration rather than focusing exclusively on heavy goods vehicles

(HGVs).

The initial analysis from the StreetLight paltform was a corridor study analysis on a segment of PCH that begins and ends a few streets away from the intrsection of interest. The analysis provided the following metrics, roadway volume, average speed, 85th speed percentile, vehicle hours of delay, travel time index and congested operational groups. Of the provided metrics researchers were most interested in roadway volume, vehicle hours of delay and the travel time index. A separate corridor analyses using the same parameters was created for Santa Fe Avenue in order to compare the metrics of each intersection with each other.

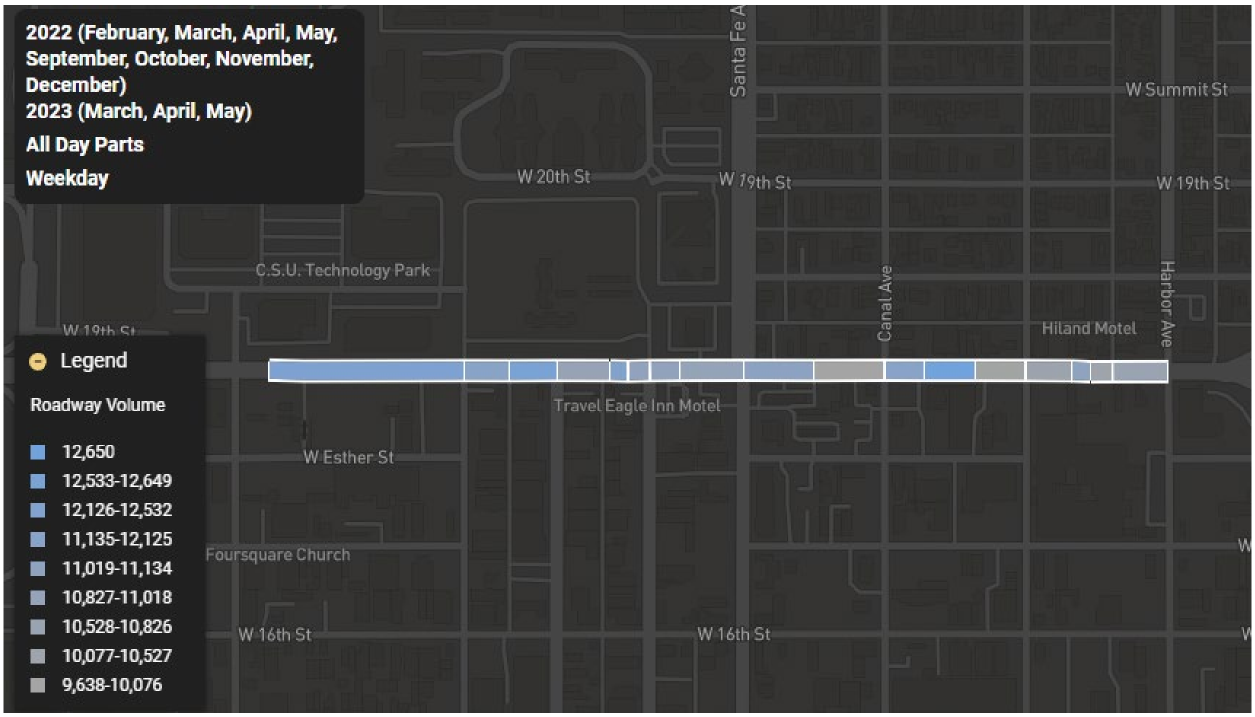



**Figure 3. StreetLight Roadway Volume data Pacific Coast Highway**

As seen in Figure 3, the roadway volume on PCH peaks roughly at 12,000. In comparison, Figure 4 shows that before the intersection the roadway volume is only about 7,000 and after the intersection it drops to only about 5,000. The sharp drop in volume suggests that a significant portion of the traffic is moving southbound and is doing so with the purpose of getting onto PCH.

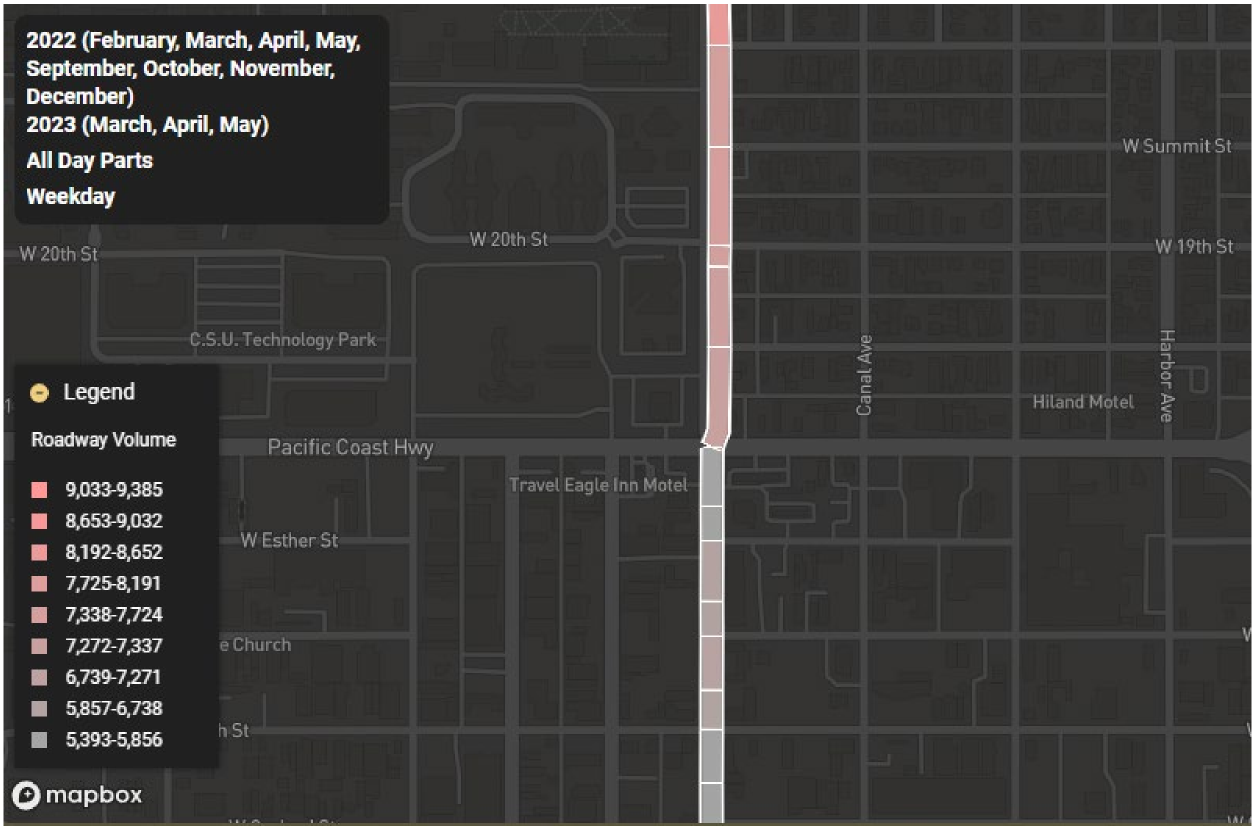



**Figure 4 StreetLight Roadway Volume data Santa Fe Avenue**

The next metric researchers looked at was the travel time index which describes how congested an area is compared to normal traffic. An index of 1 indicates normal traffic while an index above 1 indicates that traffic is more congested. By exporting the data into an excel file, researchers were able to organize the travel time index from highest to lowest which suggested that 2-5pm is when there is the most congestion on PCH.

*Table 1. StreetLight Travel Time Index*

| Day Part | Travel Time Index | Avg Segment Speed (mph) |
|---|---|---|
| 20: 4pm (4pm-5pm) | 2.59 | 16 |
| 20: 4pm (4pm-5pm) | 2.45 | 18 |
| 19: 3pm (3pm-4pm) | 2.39 | 17 |
| 18: 2pm (2pm-3pm) | 2.33 | 18 |
| 20: 4pm (4pm-5pm) | 2.31 | 17 |
| 20: 4pm (4pm-5pm) | 2.31 | 19 |
| 18: 2pm (2pm-3pm) | 2.25 | 17 |
| 19: 3pm (3pm-4pm) | 2.2 | 18 |
| 18: 2pm (2pm-3pm) | 2.17 | 17 |
| 19: 3pm (3pm-4pm) | 2.16 | 20 |
| 20: 4pm (4pm-5pm) | 2.14 | 17 |
| 19: 3pm (3pm-4pm) | 2.08 | 18 |
| 18: 2pm (2pm-3pm) | 2.05 | 21 |
| 18: 2pm (2pm-3pm) | 2.04 | 16 |
| 06: 4am (4am-5am) | 2.02 | 10 |
| 20: 4pm (4pm-5pm) | 2 | 21 |
| 22: 5pm (5pm-6pm) | 1.99 | 21 |
| 20: 4pm (4pm-5pm) | 1.99 | 21 |
| 05: 3am (3am-4am) | 1.97 | 11 |

Figure 5 shows the vehicle hours of delay (VHD) by each of its segments. Researchers noticed that Santa Fe Avenue had the highest levels before the intersection. Combined with the fact that a significant portion of vehicles seem to turn onto PCH it seemed like the majority of delays are caused by vehicles trying to turn onto PCH.

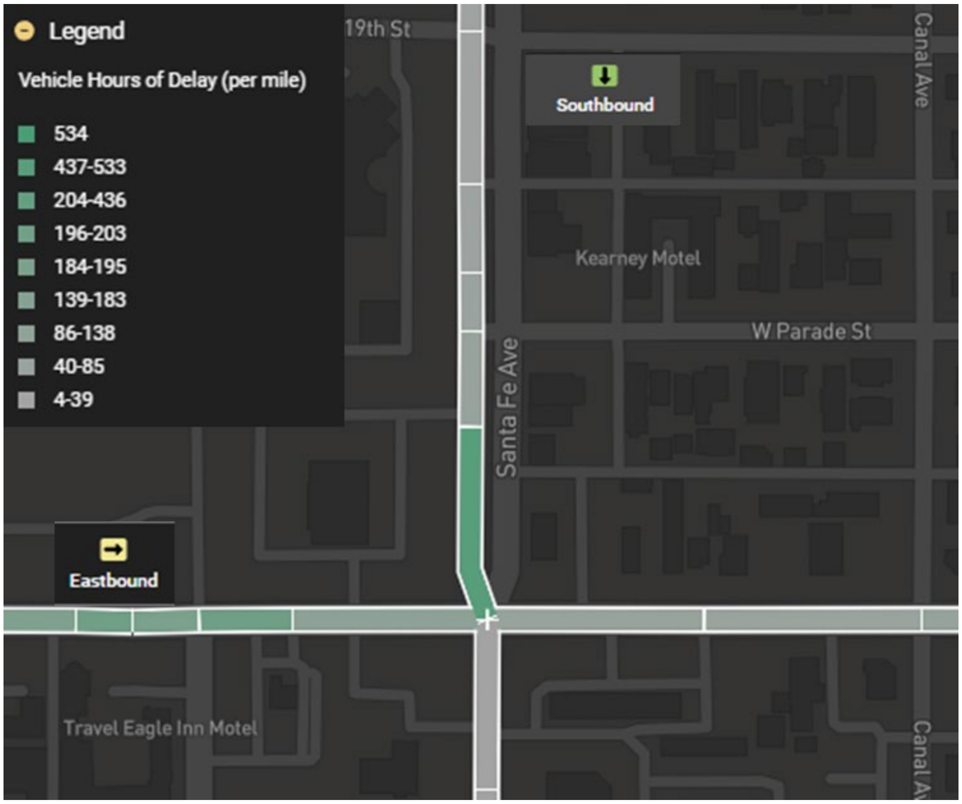



**Figure 5. StreetLight reported Vehicle Hours of Delay**

By creating a turning movement count (TMC) analysis researchers were able to get a closer look at the roadway volume present in the corridor study analysis. Streetlight and many other traffic data platforms use segments to organize the raw data into something useful, but the segment system has its limitations. At intersection vehicles have three choices on where to go, left, right, or straight. The segment system is unable to separate the three lanes and instead combines the three into one segment. When a segment shows high vehicle hours of delay before the intersection it could mean vehicles are struggling to turn onto the new street or the segment straight ahead is so full they cannot enter the intersection in fear of being unable to clear it when the light changes. In this case there is a drop in roadway volume in the following segment, so it is much more likely that the delay is being caused by vehicles turning onto PCH. Figure 6 shows that most of the traffic on PCH continues straight through the intersection but over half of the southbound vehicles on Santa Fe Avenue are turning either left or right.


**Figure 6 StreetLight Turning Movement Count**

Based on all the analysis performed by StreetLight researchers concluded that most of the delays that the intersection experiences are due to left and right turns and the most congested hours of the day fall in between 2-5pm. Further examination of the vehicle hours of delay table seemed to indicate that morning peak traffic occurs from 7-9am.

## Iteris ClearGuide

Iteris Clear Guide uses data from the companies' ClearData engine which combines speed data captured from connected vehicles, mobile applications, Bluetooth, and roadside sensors, with incident data captured from public sources agencies, and media partners. Iteris ClearGuide provides real-time data which can be compared to historical averages and can identify what time of day is the most congested

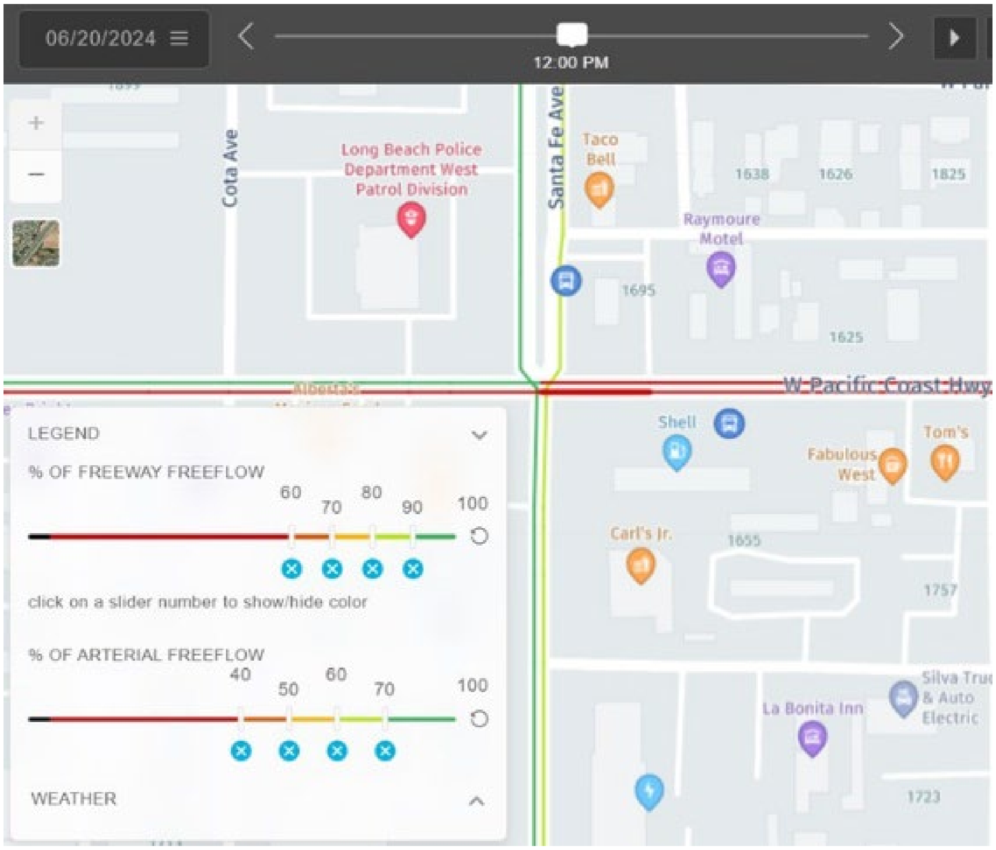



**Figure 7. Iteris free flow percentages**

The Iteris platform unfortunately does not provide traffic volume or TMC analysis which means the platforms cannot be directly compared in that analysis. Figure 8 shows the time travel indices generated by the platform which indicate that traffic volume begins to rise at 8am and creates the most congestion around 2pm. The platform is also capable of generating VHD analysis, but it was seemingly unavailable for the segments making up the intersection. The flat portions of the graph indicate that there is no congestion but that can be explained by the fact that they occur on the weekend and there was likely reduced congestion on June 19, 2024 due to the National Holiday Juneteenth.

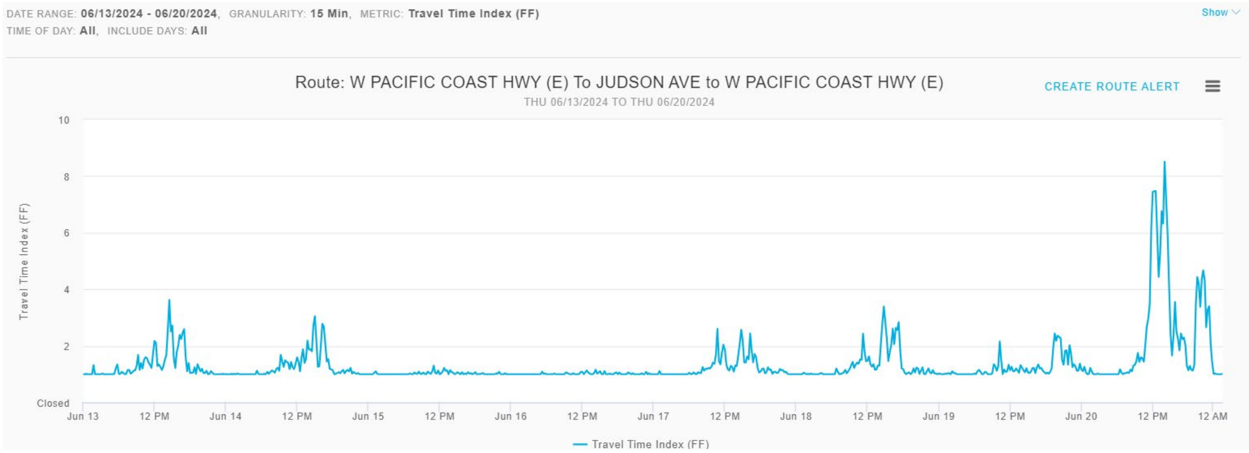



**Figure** *8*. **Travel time index eastbound traffic**

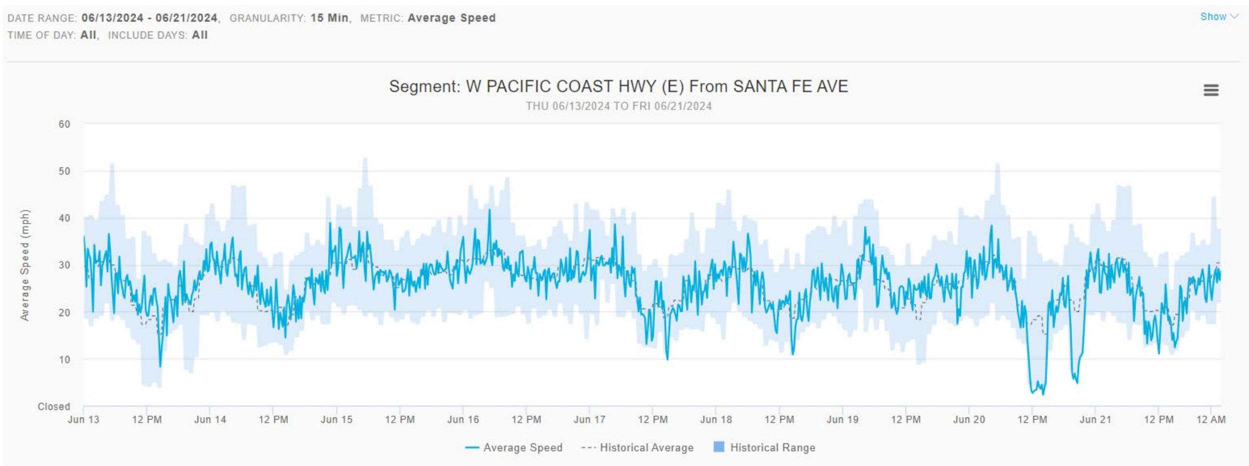



**Figure 9 Average Speed Over Time Eastbound**

The Iteris platform can also display historical average speed data which should have an inverse relationship with the travel time index. Figure 9 shows that the average speed is typically the lowest around 2-3pm which falls in line with the travel time index being highest around that time. Historical data also shows that this is usually how this stretch of road typically operates. June 20, 2024 shows that the average speed was well below the average, but it is not outside the historical range.

## Filming

The real-time data that Iteris provides combined with the analysis provided by Streetlight solidifies that peak traffic occurs from 1-3pm but researchers also wanted to explore the possibility of a peak morning traffic as indicated by StreetLight.

Researchers visited the intersection at 7am to survey the intersection and find an ideal spot to place the cameras. The data from StreetLight suggested that left turns were a major source of delays in the intersection which is why researchers wanted a location that would allow them to closely monitor left turns and have the main intersection visible.

**Figure *10*. Northeast camera placement**

Figure 10 shows the tripod set at a height of six feet and placed on the northeast corner of the intersection since it allowed researchers to see how the intersection functions while ensuring all left turns are captured on camera. Shortly after the filming began researchers observed thatthe intersection had additional issues not conveyed by either data platform.


**Figure 11. Screenshot of Long Beach Police vehicle struggling to make left turn in response to emergency. Research team recordings captured similar left-turn challenges with related Long Beach Fire Department firetrucks and ambulances.**

Researchers noted that emergency vehicles struggled to safely pass through the intersection (see Figure 11). The nearby police station on Santa Fe Avenue sent out multiple vehicles to respond to an emergency call, requiring a left turn on PCH. All three emergency-response vehicles were delayed due to the nature of the intersection. The designated left turn lane on Santa Fe Avenue was full forcing the emergency-response vehicles into the next lane meant for traffic traveling in the opposite direction. From that position, emergency-response vehicles have poor visibility of the intersection and the other vehicles coming from the right cannot visibly identify where the emergency vehicles are. The situation alone creates an unsafe environment, but because of the high number of trucks traveling along PCH, the possibility of a fatal collision is increased since those heavy trucks need more time to come to a complete stop. In subsequent recording sessions, researchers observed this behavior repeat itself on several occasions. During periods of lighter traffic, the problem is minimized but the possibility of a collision still exists.

The peak morning traffic that the data from Streetlight suggested was not present while researchers observed the intersection. One possible reason for this may be that the data from StreetLight was too outdated since it largely pulled its data set from the years 2020 to 2022. Since researchers could not find evidence of a morning spike in traffic volume, they decided to revolve the filming schedule around the 1-3pm peak. The researchers shared their findings with the engineering team which recommended adding a secondary camera on the opposite end of the intersection so they can capture the queue length of every side of the intersection at any given time. They also suggested raising the height of the camera since HGVs could obstruct the view of the intersection making them unable to accurately model the intersection.

After sharing the initial footage with the engineering team, they made several suggestions to help them process the footage (see Figure 12). In the following recording sessions researchers implemented the recommendations by the engineering team and set up an additional camera on a 12ft tripod and set the height of the main camera to 25ft. The engineering team recommended a second camera since one camera is not able to capture the queue length from all 4 directions of traffic. The change in height was recommended to place the camera at an angle where it is able to see over the HGV so the vehicles in the adjacent lanes are still visible while the HGV crosses the intersection. Researchers observed higher levels of traffic during the first few hours of the recording session which subsequently tapered off after 3pm.

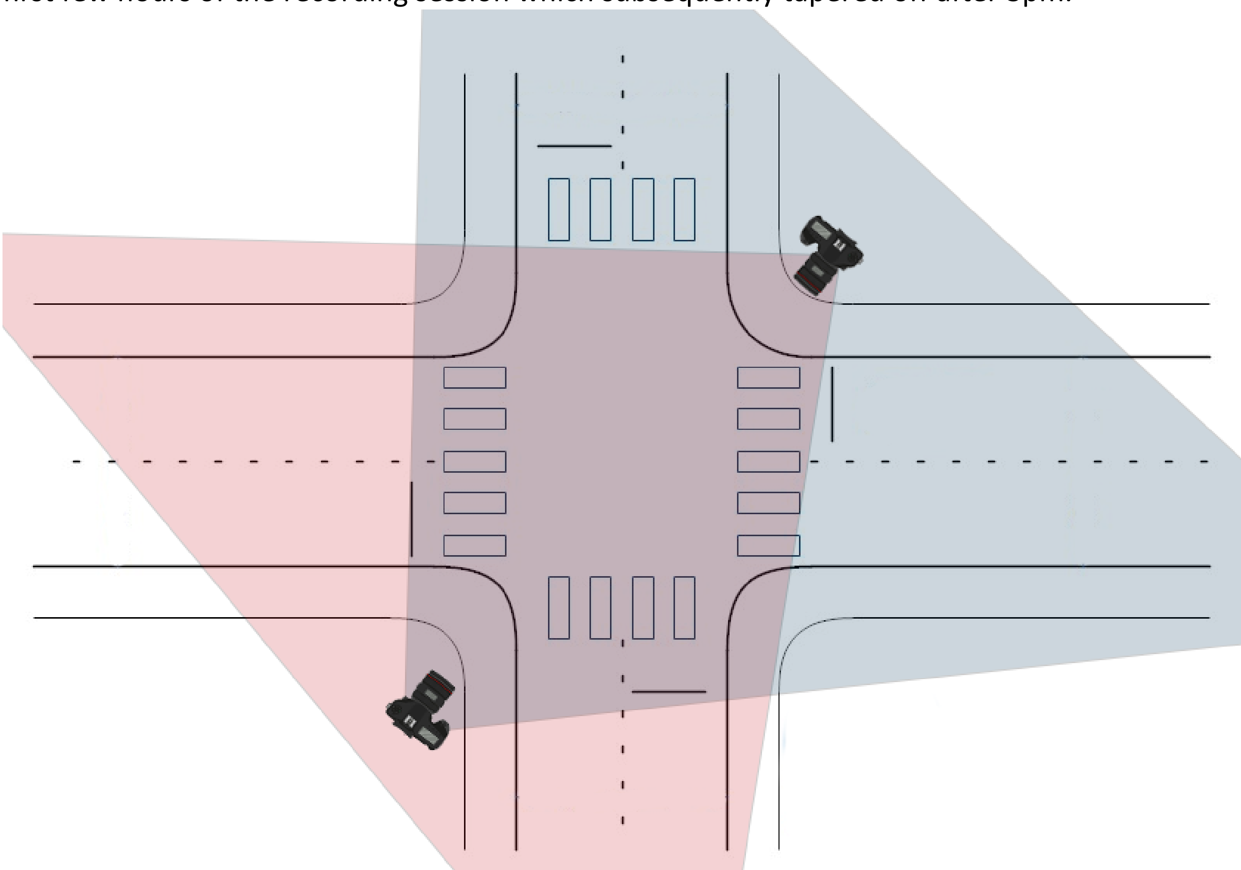


**Figure 12. Camera placement diagram**

During this new recording time researchers observed more issues that were not present in StreetLight or Iteris.

During peak traffic periods the east segment of PCH can quickly fill up or lose capacity from a variety of diverse sources. Along the right-most lane on PCH, there are three points that reduce the road's service flow rate. The first point is a bus stop used by Long Beach and Torrance Transit. When busses stop at their designated location they move as close to the sidewalk as they can to allow passengers to board and deboard the vehicle but often they are still partially blocking the lane they stopped in which forces drivers to wait or move into the left lane to get around them. Automobiles can easily navigate the blockage, but trucks carrying large loads cannot unless the left lane is completely free. Immediately following the bus stop is a gas station designed for trucks carrying containers as it has wider and longer spaces compared to the regular station next to it. Although these pumps are designed for longer vehicles, trucks still need to physically pull into the gas station which temporarily creates delays for the vehicles behind them. Furthermore, when the gas station lane becomes too full, the road can become partially obstructed like when the bus stops to board passengers. The last point of congestion is the entrance ramp onto the I-710 freeway. The southern entrance to the 710 freeway leads to the San Pedro port complex, but when the freeway is congested, the entrance ramp can become full which creates a blockage on PCH. When the I-710 entrance is full, it prevents trucks from moving forward, which when combined with the other factors creates the observed situation where the intersection of PCH and Santa Fe Avenue becomes obstructed. The blocked intersection issue cannot be properly conveyed via the data analytic platforms because of how rare the issue is when considering every intersection in a city. In this intersection, blockages are a common problem during heavy periods of congestion, but the issue is rare at other intersections because their traffic is typically composed of passenger vehicles rather than having a significant portion of the traffic being from HGVs. Most intersections provide higher levels of service, and an obstructed intersection is more akin to an anomaly like a car crash, so it is not something that these programs are particularly focused on.

**Figure 13. Blocked Intersection at Pacific Coast Highway and Santa Fe Avenue**

Researchers observed that Figure 3 from Streetlight seemed to be accurate since it shows that most of the traffic passing through the intersection comes from PCH. Roughly 12,000 vehicles travel eastbound through the intersection with a similar number of vehicles traveling in the opposite direction within any given weekday. As predicted by the program, researchers observed that most of the traffic moving through the intersection came from PCH while there was much less traffic coming from Santa Fe Avenue. A sizable portion of the northbound vehicular traffic moving through Santa Fe Avenue turned left or right onto PCH; but most of the southbound traffic continued straight through the intersection.

Based on Figure 5 portraying the vehicle hours of delay of each section, researchers anticipated that the main cause of congestion was due to vehicles attempting to turn onto PCH from Santa Fe Avenue. Just before the intersection, the southbound traffic has the highest level of congestion at 232 vehicle hours of delay (VHD) at the section just before the intersection. Researchers observed that there was some level of delay experienced caused by the volume vehicles trying to take a left turn, but it did not dramatically reduce the level of service of the intersection. During the recording sessions researchers observed two possible reasons why there could be a significant level of  delays present on Santa Fe Avenue, the first being inconsistency of the protected left turn signal. The protected left turn signal would only turn on occasionally while the rest of the time only the regular green light would turn on forcing the vehicles on Santa Fe Avenue to yield to oncoming traffic. Whenever there was a protected left turn signal, the queue of vehicles in the left turn lane could be cleared out but when they needed to yield to the northbound traffic only about two vehicles could pass at a time which contributed to the VHD reported in that segment of the road in StreetLight. The second possibility is that vehicles attempting to turn onto PCH physically have no room to do so. While recording the intersection, researchers observed that during peak traffic hours the intersection would become partially blocked due to heavy goods vehicles being unable to fully clear the intersection when the east segment of PCH was at maximum capacity. When the intersection is partially blocked, vehicles attempting to turn onto PCH cannot safely start their turn because if they did so they would also be stuck in the middle of the intersection, so even if they did not have to yield to oncoming traffic, vehicles could easily miss the turning cycle entirely and must wait for another opportunity to attempt a turn. Observing this behavior, researchers revisited the VHD data for PCH.

According to StreetLight, most of the congestion on PCH occurs within the segments before the intersection. This behavior implies that the vehicles at the light accelerate slower than average, which creates congestion through the compounding wait times. Once the first vehicle sees the greenlight and starts to accelerate to the speed limit, the vehicle behind must do the same as the first vehicle but instead of waiting for the greenlight, they are waiting for the vehicle in front of them. This behavior applies to every vehicle in the queue which results in each vehicle taking longer to move forward thus creating congestion. The reported congestion on PCH was not as large as Santa Fe Avenue as it only had a peak of 196 vehicle hours of delay for eastbound traffic and 151 vehicle hours of delay for westbound vehicles. Although this behavior was observed at the intersection, it did not seem to be a major source of congestion. A possible reason why StreetLight reports congestion in the affected segments is the extra time vehicles spend at a standstill when the segment after the intersection is full.

Figure 6, the Turning Movement Count (TMC)[1] shows that a significant portion of the traffic from Santa Fe Avenue was turning onto PCH which was thought to be the source of the congestion. Streetlight data showed that of the total northbound traffic 61.26% continued through the intersection, 28.03% turned left, and 10.71% turned right. Of the total southbound traffic, 46.27% continued through the intersection, 23.47% turned right, and 30.26% turned left. In comparison, the East and Westbound traffic have about 80% of vehicles continue through the intersection and the remaining 20% split between left and right turns.

Traffic analysis platforms such as Streetlight and Iteris provide resonably accurate information on the overall trends of the intersection, but it is unable to provide granular information, specifically the difference between vehicle type, that would be required to implement

[1] Researchers define Turning Movements Count (TMC) as the count of movement of vehicles at an intersection going left, right, or continuing straight, from a particular gate. A gate is an imaginary zone placed across a road at a specific location to capture the movement of vehicles, i.e., these gates act as counters that record when a vehicle crosses them. This data is a significant factor to determine signal-timing, design intersections, and corridor studies.

Intelligent Transportation Systems (ITS) technologies on a day-to-day basis. Streetlight functionality provides helpful tools to better understand the mechanics of an intersection before committing to other forms of analysis that provide more specific data. The Iteris Platform is seemingly more geared to understanding how any given segment is preforming on that day rather than providing a clear idea of how the intersection normally behaves but supplemented with the data from StreetLight they provide a clear image of what the problems of the intersection are. The free flow percentage map from Iteris shows in real-time the most critical segments of the intersection but without the context of StreetLight and a strong understanding of how the intersection functions in real life the data is difficult to comprehend. StreetLight or any other data analysis should not be used as the sole source of information because of the limitations in identifying and differentiating between truck types and other transportation modes. Other limitations of the program include: some dates and zones are unavailable, and mixed data sets—leading to confusion in identifying vehicle type and, in some cases, high-speed cyclists and pedestrians occasionally being counted as vehicles. Each platform has their own way of gathering and aggregating data which could lead to some biases so by comparing reports from multiple platforms researchers can find what the most definitive problems are. In the end data analysis platforms are no substitution for obtaining real world footage but they make the process of gathering that footage go much more smoothly.

## Developing a Predictive Traffic Model

Traffic prediction is a crucial aspect of an efficient and sustainable urban mobility system. By accurately predicting traffic flows and delays, authorities can reduce congestion, identify potential traffic jams and hazardous conditions, and enhance public safety and wellness. However, predicting traffic - in large cities - is a daunting task due to its complex, volatile, and nonlinear nature. Artificial Neural Networks offers a potential solution due to their capabilities to model nonlinear functions and systems. This section reviews current traffic prediction models, describes various types of neural networks, including parametric models such as the Autoregressive Integrated and Moving Average (ARIMA), and the Seasonal Autoregressive Integrated Moving Average (SARIMA), as well as non-parametric models such as Deep Neural Networks (DNN). This section further focuses on neural networks' effectiveness – compared to traditional traffic prediction methods – in predicting traffic characteristics.

## Developing Artificial Neural Networks and Predictive Traffic Models

The research in the traffic prediction field can be dated back to the 1970's, when Levin and Tsao applied Box-Jenkins time-series analyses to predict freeway traffic flow using ARIMA (Auto-Regressive Integrated Moving Average)[iv]. The method proved to be statistically significant, however, its traffic forecast error was also significant during irregular traffic fluctuations. Due to its *parametric nature*, its forecast depended on the data distribution rather than the data alone. Despite this drawback, ARIMA is used till date as a benchmark to estimate the efficiency of traffic prediction model. To tackle the obstacle of a parametric prediction method, the focus shifted towards *non-parametric* methods such as Neural Networks, Kalman Filters, Support Vector Machines, etc., that do not make assumption based on the data distribution. These nonparametric methods are best fitted to those skewed and sparse data obtained from volatile traffic data.

David Alexander Tedjopurnomo et al. performed a study in April 2022, focusing on one such non-parametric method - Deep Neural Networks (DNNs). They reviewed state-of-the-art DNNs based on model architecture and the data used for traffic prediction[v]. These DNNs have three types of neural networks forming their foundation – (1) Feed-Forward Neural Networks (FNN), (2) Recurrent Neural Networks (RNN), and (3) Convolutional Neural Networks (CNN).

Each of these three DNNs specializes in capturing one aspect of the traffic. CNNs capture spatial features such as the flow of traffic through different road networks. RNNs focus on temporal features such as the dependency of past traffic on future traffic. FNNs are commonly known to aggregate the outputs of different sub networks such as CNNs and RNNs, while adding external features such as weather, seasons, holidays, etc. Tedjopurnomo et al. also focus on the qualitative and quantitative importance of the data used to train these models. In the traffic prediction problem, the quality of data is determined by the number of features available that describe the traffic. Among all the DNNs evaluated, the authors found that the most popular data, fit for training a DNN, is provided by Caltrans PEMS[vi]. This conclusion was reached due to the features available in the data. These features include the date, timestamp, traffic flow per lane, aggregated traffic flow, occupancy of the vehicle, speed of the vehicle, class of the vehicle, etc., which make the data source rich. Rich data sources make it more useful for training the DNN, as it carries more detailed information and customizing features.

The data granularity or the interval between each captured data sample is also an important consideration in how well the model identifies the traffic patterns in the data. Data granularity has become an important hyperparameters that dictates the quality of data. Having low granularity between 2-5 minutes can lead to sparsity (*i.e.*, many zeros in data that will not significantly impact a calculation). Higher granularity between 45-60 minutes, on the other hand, might lead to major information loss in the traffic trends. Hence, a 15-minute granularity is deemed best to set as the hyperparameter that is used to capture a data sample[vii]. Lastly, the number of data samples fed to the model also corroborates with the richness of the data. The study[viii] also finds that having data samples from just one year might not be enough to capture the *seasonal trends* that are caused due to holidays, festivals, weather changes, etc. These trends are best captured using data samples from at least 3 years.

Most data available in resources across the globe are historical, time-series data. To manage large volumes of historical data and understanding the traffic characteristics that evolve with time, the best suited neural network for this task is the Recurrent Neural Networks (RNNs) using Long-Short Term Memory (LSTM). With the ability to retain long-term and short-term memory, RNNs decide what memory to retain or discard, and provide precedence over traditional neural networks or ARIMA[ix] [x]

# Parametric Models

This section provides a detailed description of the parametric models used in traffic prediction models. The research team will use this approach later to develop our parametric linear models.

## ARIMA

The ARIMA model is a statistical tool for forecasting time series data. The ARIMA model combined Autoregressive (AR) Integrated (I) and Moving Average (MA). In an autoregressive model, the value of the variable at any given time point is predicted based on its past values. This type of model assumes that there is a linear relationship between the variable and its past values, and it captures the temporal dependencies within the data. Mathematically, an autoregressive model of order *p*, often denoted as *AR*(*p*), can be expressed as:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$$

Where $X_t$ is the value of the variable to be predicted at time *t*, *c* is a constant term, $\phi_i$ the $i$th model parameter representing the relationship between the variable $X$ at time t and its past values. and $\varepsilon_t$ is the error term at time $t$ .

A Moving Average model is based on the concept of averaging past observations to forecast future values. Unlike autoregressive models where past values of the variable itself are used for prediction, in Moving Average models, past forecast errors (residuals) are utilized. The Moving Average model of order *q*, often denoted as *MA(q)*, can be represented mathematically as:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Where here $\mu$ is the mean of the time series and $\theta_i$ are the parameters of the model representing the weights assigned to the past values at time $t - i$ .

In the ARIMA (Autoregressive Integrated Moving Average) model, the “Integrated” part refers to the differentiating operation applied to the time series data to make it stationary. This differentiating, denoted by parameter *d*, can be represented by

$$X_t' = (1 - B)^d X_t = (1 - B)(1 - B)^{d-1} X_t$$

Where $X_t'$ is the differentiated series after applying differentiating of order *d* on the original non-stationary time series $X_t$, and *B* is the backshift operator, representing the lag operator (*i.e.*, $BX_t = X_{t-1}$).

## SARIMA

Seasonal Autoregressive Integrated Moving Average (SARIMA) is an extension of the ARIMA model that accounts for seasonality in time series data. This model could potentially be more applicable to the traffic prediction problem due to the seasonal patterns and trends in traffic. In SARIMA model, we have

$$X_t' = \left(1 - \phi_1 B - \cdots - \phi_p B^p\right)(1 - B)^d \times \left(1 - \theta_1 B^s - \cdots - \theta_q B^{qs}\right) X_t$$

$$= \left(1 - \phi_1 B - \cdots - \phi_p B^p\right)(1 - B)^d \times \left(1 - \theta_1 B^s - \cdots - \theta_q B^{qs}\right)(1 - B)(1 - B^s)^{D-1} X_t$$

Where $X_t$ is the original non-stationary time series, $X_t'$ the differentiated series after applying non-seasonal differentiating of order d and seasonal differentiating of order *D*, *B* is the backshift operator, *s* is the seasonal period, $\phi_1, \dots, \phi_p$ the autoregressive parameters for the nonseasonal component, and $\theta_1, \dots, \theta_q$ the autoregressive parameters for the seasonal component.

# Non-parametric Models

This section details the fundamentals of neural networks used in traffic prediction, including feed-forward neural networks (FNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs).

## Feed-Forward Neural Network

A feed-forward neural network (FNN) is a type of deep neural network wherein connections between the nodes do not form a cycle. This is the simplest type of artificial neural network. In this network, the information moves in only one direction—forward—from the input nodes, through the hidden nodes, and to the output nodes. There are no cycles or loops in the network. The three basic layers in an FNN are - the input layer, hidden layer, and output layer. Figure 14 illustrates a typical FNN with one hidden layer.

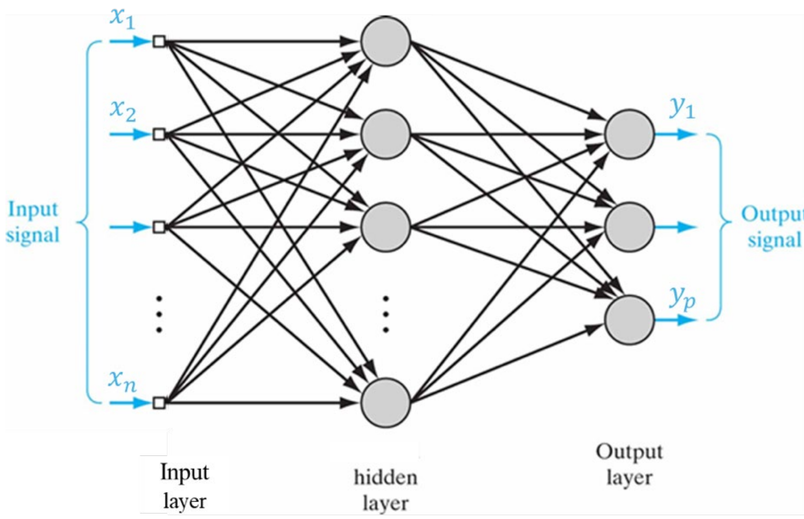



**Figure 14. A typical Forward Neural Network**

1) Input Layer - The input layer receives the input vector $x$ , which is a feature vector from the dataset. In the traffic prediction task, the input layer receives historical traffic data as feature vectors, including attributes such as traffic volume, weather conditions, time of day, and road network topology. Each component of the input vector, representing a specific feature of the traffic data, is passed to the hidden layer.
2) Hidden Layer - The nodes in this layer apply a weighted linear transformation followed by a non-linear activation function which allows the network to learn the complex patterns and relationships in traffic data. For a single hidden layer with H nodes, the output $h$ of the hidden layer can be represented as:
$$h = \lambda(W_1 x + b_1)$$
Here, $W_1$ is the weight matrix connecting the input layer to the hidden layer, $b_1$ the bias vector for the hidden layer and $\lambda$ the activation function.
3) Output Layer - The output layer receives the output from the hidden layer, again applies a linear transformation followed by a different activation function and produces the final output y, that produces the traffic prediction based on the learned features from the hidden layers. For a network aiming to predict K outputs, the output layer can be described as:
$$y = \lambda(W_2 h + b_2)$$
Here, $W_2$ is the weight matrix connecting the hidden layer to the output layer and $b_2$ the bias vector for the output layer.

The complexity of this type of neural network depends on the number of hidden layers used and the type of activation function. The choice of activation function lambda in the layers depends on the nature of the task. Activation functions are crucial as they introduce non-linear properties to the system, which allows the model to learn more complex patterns in the data. Some of the commonly used activation functions are -

- Sigmoid Activation - The sigmoid function is an activation function traditionally used in neural networks, particularly for binary classification at the output layer. It maps the input to an output value between 0 and 1, making it especially useful for models where we need to predict the probability as an output since the probability of anything exists only between 0 and 1. It is mathematically represented as:

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

- Hyperbolic Tangent Function (Tanh) - The tanh function is like the sigmoid but outputs values from -1 to 1. This nature of tanh makes its outputs zerocentered, thereby aiding the backpropagation process in neural networks, as it helps in data normalization and leads to faster convergence. Its mathematical representation is given as:

$$\sigma(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

- Rectified Linear Unit (ReLU) - ReLU function activates a node only if the input is above zero and is zero otherwise. It is computationally efficient and allows the model to converge very quickly. However, ReLU can lead to the dying ReLU problem, where neurons permanently die during training since they stop outputting anything other than zero. It is represented as:

$$ReLU(x) = \max\{0, x\}$$

- Leaky ReLU - It is a modified version of the ReLU function which allows a small negative value when the input is less than zero. This small slope ensures that leaky ReLU never dies, i.e., it can recover during training and always allows a small gradient when the unit is not active. Mathematically, it is represented as:

$$Leaky\ ReLU(x) = \max\{0.01x, x\}$$

- Exponential Linear Unit (ELU) - ELU combines the advantages of ReLU and leaky ReLU, offering a smooth and non-linear curve. Unlike ReLU, ELU can produce negative outputs, which allows it to push mean unit activations closer to zero like tanh and thus helps in speeding up learning. It is mathematically represented as:

$$ELU(x) = x, \qquad if\ x \geq 0$$

$$ELU(x) = \alpha(e^{x} - 1)\,, \quad if\ x < 0$$

where $\alpha$ is a hyperparameter.

- Softmax Activation - The softmax function is a generalization of the sigmoid function that squashes a Kdimensional vector of arbitrary real values to a Kdimensional vector of real values in the range (0, 1) that add up to 1. It is used in multi-class classification problems where it can output a probability distribution across various classes.

$$Soft\max(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

where, $z_i$ is the input to the softmax function for the *i*th class.

Any neural network is typically trained by adjusting the weights and biases to minimize a loss function, which measures the difference between the network output $y$ and the target output $\tilde{y}$. Common choices of loss functions include Mean Squared Error (MSE) for regression tasks and CrossEntropyLoss for classification tasks. Since traffic prediction is a classified as a regression task, its loss function is calculated as:

$$J = \frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y_i})^2$$

The training involves using an optimization technique - such as backpropagation with an algorithm like stochastic gradient descent (SGD) - to iteratively adjust $W_1$, $b_1$, $W_2$, and $b_2$ to minimize the loss function. SGD works by randomly picking a single data point or a small batch of data points from the dataset at each step, which is where it gets the "stochastic" part of its name. This random sampling reduces the computational burden, making the algorithm much faster and more scalable to large datasets compared to batch gradient descent, where the gradient is computed over the entire dataset. The update rule is applied iteratively, and over time, it is expected to converge to the minimum of the loss function, although the path taken towards the minimum may be noisy compared to that of batch gradient descent. This noise, however, can sometimes help the algorithm escape local minima, potentially leading to better solutions for complex models. It is mathematically represented by:

$$\theta_{t+1} = \theta_t - \eta\nabla_\theta J(\theta_t, x_i, y_i)$$

where, $\theta_t$ represents the parameter (weights $W$ or biases $b$ ) at iteration t, $\eta$ is the learning rate that determines the step size during the minimization process,$\nabla_\theta J$ is the gradient of the loss function $J$ with respect to the parameters $\theta$, and $x_i$, $y_i$ are the input and target output for the *i*th randomly selected sample from the dataset.

## Convolutional Neural Network

Convolutional Neural Networks (CNNs) are designed to learn spatial hierarchies of features automatically and adaptively from raw input data. It is effective in tasks like traffic flow prediction, congestion detection, and anomaly detection in traffic surveillance systems. Unlike an FNN, CNNs have the following layers:

1) Convolutional Layers - The core building blocks of CNNs are convolutional layers. These layers apply convolution operations to the input data using learnable filters or kernels. The filters convolve over the input data, extracting various features from distinct parts of the data. Each filter detects specific patterns or features in the input, helping in capturing local patterns and spatial relationships in the data. Mathematically, the output feature map $Y_k$ of a convolutional layer *k* can be computed as:

$$Y_k = f\left(\sum_i (X_i W_{k,i}) + b_k\right)$$

where $X_i$ is the input feature map representing the traffic data, $W_{k,i}$ the filter/kernel, $b_k$ the bias, and $f$ the activation function.

2) Pooling Layers - Pooling layers reduce the spatial dimensions of the input feature maps while retaining the most vital features. They achieve this by aggregating information from neighboring features or regions of the input, thus extracting relevant spatial patterns from traffic data. Common pooling operations include max pooling (selecting the maximum value within a window) and average pooling (calculating the average value within a window). Pooling layers help in making the representations more invariant to small translations and distortions in the input data, leading to improved generalization. Mathematically, the output of a max pooling layer can be computed as

$$Y_{j,k} = \max\{X_{i:i+m\,,\,j:j+n}\}$$

where $X$ is the input feature map, $m$ and $n$ the dimensions of the pooling window, $i{:}\,i + m\,,\ j{:}\,j + n$ is the region of the input to which the operation is applied, and $Y_{j,k}$ the output feature map after pooling.

3) Fully Connected Layers - Fully connected layers aggregate information from the spatially and temporally extracted features to make predictions about the input data. These layers are typically used at the end of the CNN architecture to map the high-level features learned by the convolutional layers to the output classes or labels. Fully connected layers perform a series of weighted linear transformations followed by non-linear activation functions. It is represented mathematically as:

$$Y = f(WX + b)$$

where, $Y$ is the final output, $X$ the input from the previous layer, $W$ the weight matrix, $b$ the bias vector, and $f$ the activation function.

A CNN is formed by layering several convolutional, pooling layers, and ending it with a fully connected layer. The training process of this model is like that of the FNN described above.

## Recurrent Neural Network

A Recurrent Neural Network (RNN) is a type of neural network suited for sequence prediction problems. This is because RNNs can maintain a hidden state that acts like memory. They are used extensively in time series analysis and other domains where temporal dynamics are crucial such as traffic prediction.

The hidden state in an RNN captures information about the sequence processed so far, acting as a memory. At each time step $t$, the new hidden state $h_t$ is updated based on the previous hidden state $h_{t-1}$ and the current input $x_t$. It is mathematically represented as:

$$h_t = f(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

where, $W_{hh}$ is the weight matrix for connections between the previous hidden state and the current hidden state, $W_{xh}$ the weight matrix for connections between the input and the hidden state, $b_h$ the bias vector for the hidden state, and f is an activation function.

The output at each step $y_t$ is calculated from the hidden state $h_t$, using an activation function $g$ . The output is computed as:

$$y_t = g\left(W_{hy}h_t + b_y\right)$$

where, $W_{hy}$ is the weight matrix for connections between the hidden state and the output, and $b_y$ the bias vector for the output.

The training of an RNN is like the FNN described above. Due to the recurrent nature, a method known as Backpropagation Through Time (BPTT) is used. BPTT involves unravelling the RNN through time and then applying backpropagation. This can lead to challenges such as exploding and vanishing gradients while calculating the gradients for the layers with long-time dependencies using optimization techniques such as SGD. This problem of vanishing or exploding gradients is tackled using Long-Short Term Memory (LSTM) networks.

### Long-Short Term Memory using RNN

Long Short-Term Memory networks (LSTMs) are a special kind of Recurrent Neural Network (RNN) capable of learning long-term dependencies in time-series data. LSTMs aim to address the “vanishing gradient problem” that affects standard RNNs. This problem occurs during backpropagation through time (BPTT) where gradients, which are values used to update a network’s weights, can become exceedingly small, effectively preventing the network from learning long-term dependencies. To overcome this, LSTMs incorporate several “gates” that regulate the flow of information. These gates decide what information should be kept or discarded at each time step of data processing, enabling the network to preserve important long-term information and discard non-relevant details. It is ideal for tasks like traffic prediction where past information is crucial for predicting future traffic conditions.

A single unit in an LSTM network consists of a cell with three gates - input gate, forget gate and output gate. The important components of the LSTM unit are:

1) Cell State - The key component of LSTM and acting like the memory of the network is the cell state, $C_t$. It runs through the entire chain of LSTM units, with only minor linear interactions. It's designed to make it easy for information to flow along with the network without being modified unnecessarily.

2) Forget Gate - This gate, $f_t$ decides what information is irrelevant and can be discarded from the cell state. It considers the previous hidden state $h_{t-1}$ and the current input $x_t$, and outputs a number between 0 and 1 for each number in the cell state $C_{t-1}$. Assigning 1 represents - "Completely retain this information", and 0 represents - "Completely discard this information." Mathematically, this gate computes this information, using the sigmoid function $\sigma$, as:

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right)$$

where, $f_t$ is Forget gate's activation vector at time t, determining which parts of the cell state to retain, $W_{xf}$ the weight matrix for the input $x_t$ affecting the forget gate, $W_{hf}$ the weight matrix for the previous hidden state $h_{t-1}$ affecting the forget gate and bf is the bias for the forget gate.

3) Input Gate - The input gate decides which values will be updated. It makes the decision using the sigmoid activation function and is computed as:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

where, $i_t$ is the input gate's activation vector at time $t$ , the weight matrices $W_{xi}$ and $W_{hi}$ weight matrices affecting the input gate and $b_i$ is the bias vector for input gate.

4) Candidate Layer - The candidate layer, $\widetilde{C_t}$ creates a vector of new candidate values that could be added to the state. It is computed using the tanh activation function as:

$$\widetilde{C_t} = \tanh(W_{xC}x_t + W_{hC}h_{t-1} + b_C)$$

where, $\widetilde{C_t}$ represent the Candidate values for the cell state, which are updates to the cell state vector, $W_{xC}$ and $W_{hC}$ the weight matrix affecting the candidate cell state, and the bias vector for the candidate cell state.

5) Cell State Update - The old cell state, $C_{t-1}$ is updated to the new cell state, $C_t$. It is updated to carry relevant information throughout the processing of the sequence. The dropped information discarded by the forget gate, and the new addition of information is updated into the new cell state as:

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t$$

where, $\odot$ represents the element-wise multiplication of the vectors/matrices.

6) Output Gate - The output gate, $o_t$ controls the output of the cell state into the next hidden state. This output is a filtered version of the cell state and computed using the sigmoid function as:

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$

where, $o_t$ is the output gate's activation vector at time t, $W_{xo}$ and $W_{ho}$ the weight matrices affecting the output gate, and $b_o$ is the bias vector for the output gate.

7) Hidden State Update - The old hidden state $h_{t-1}$ is updated to the new hidden state $h_t$ as part of the final output and for the next step's calculations, i.e., it is going to be used to calculate the gates and states represented by $t + 1$ . The state update is computed as:

$$h_t = o_t \odot \tanh(C_t)$$

In summary, LSTM uses the forget gate to evaluate both the previous hidden state and current input, and decides what information is irrelevant and should be removed from the cell state. Simultaneously, both the input gate and a set of candidate values decide what current information should be stored in the cell state. Based on these decisions, the cell state is then updated to reflect both the forgotten data and the new data added. The output gate determines which part of the cell state should be output as the hidden state, which can be used for predictions or passed to the next time step. This architecture allows LSTMs to effectively capture long-term dependencies and make them suitable tools for time series prediction. Figure 15 illustrates the structure of a single LSTM RNN cell.

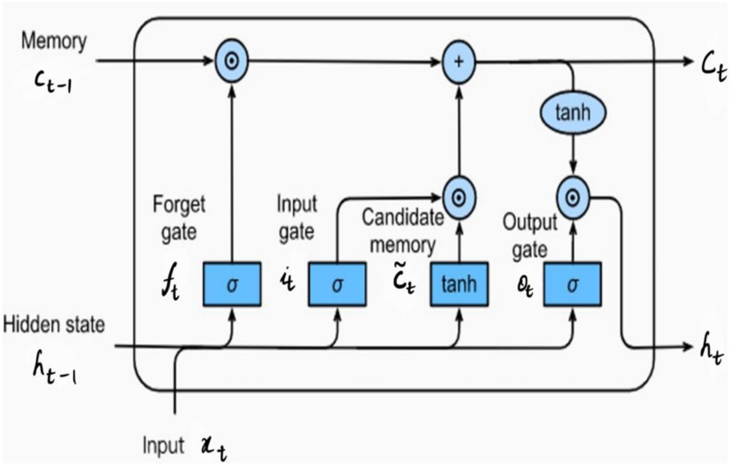



**Figure 15. The structure of a single Long Short-Term Memory network (LSTM) Recursive Neural Network (RNN)**

## Evaluation

The efficiency of the built traffic prediction models depends on parameters such as the data used to train the model, the hyperparameters used to tune and fit the model. To evaluate the efficiency of the prediction models, researchers compare the actual values of the data samples with the samples generated by the traffic prediction model.

Suppose $n$ is the total number of data samples, $y_i$ the actual traffic, and $\hat{y}_i$ the traffic predicted by the model, then the common measures used to test a model's efficiency in predicting traffic flow are:

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

- Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Mean Square Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Pacific Southwest Region UTC
University Transportation Center

## Simulation Results – Parametric Linear Models

Linear models, such as polynomial Autoregressive Integrated Moving Average (ARIMA), are of the most popular traffic prediction models that is set as a benchmark for comparison against any emerging models. To test and evaluate our models, we used StreetLight data from April 1 to May 31, 2022. Figure 16 illustrates a typical week's of traffic data.

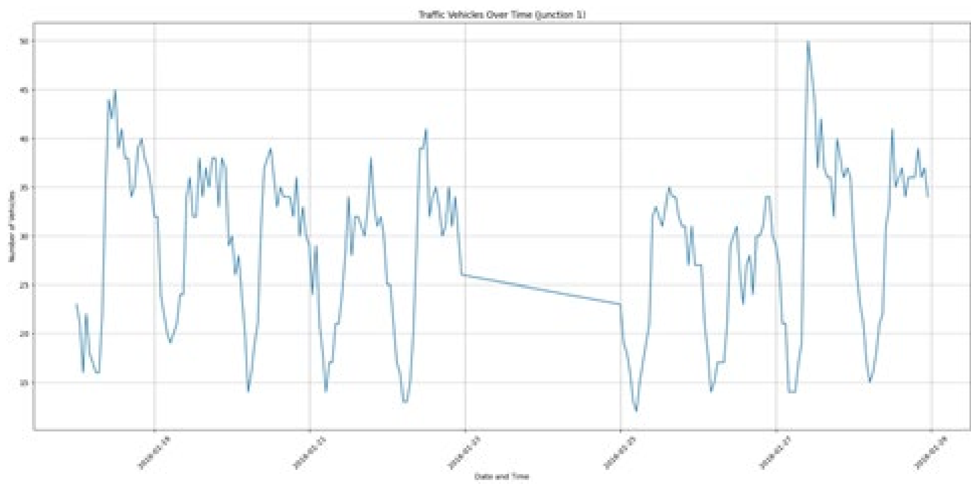



**Figure 16. Typical weekly StreetLight traffic data for the PCH and Santa Fe intersection. Data was not collected during weekends**

## Polynomial Models

A linear polynomial model is used to predict traffic during the day. This model can be described as a subclass of the linear models.

$$\hat{f}(x) = \theta_1 + \theta_2 x + \cdots + \theta_p x^p$$

Where, $\theta_i$ is the *i*th coefficient of the model, $x$ the time of the day, and $\hat{f}(x)$ is the predicted traffic at time $x$ . Researchers used the least squares (LS) method to find the optimal values of $\theta_i$ for $i = 1, \cdots, p$ during training phase. Figures 17 and 18 show the result of polynomial fitting to data for a fifth order polynomial, i.e., $p = 5$ .

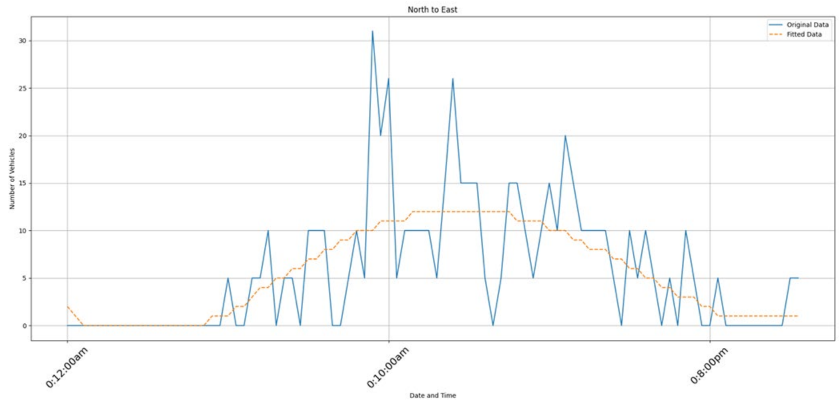



**Figure 17. Results of a polynomic traffic predictor; North-East bound when p=5**

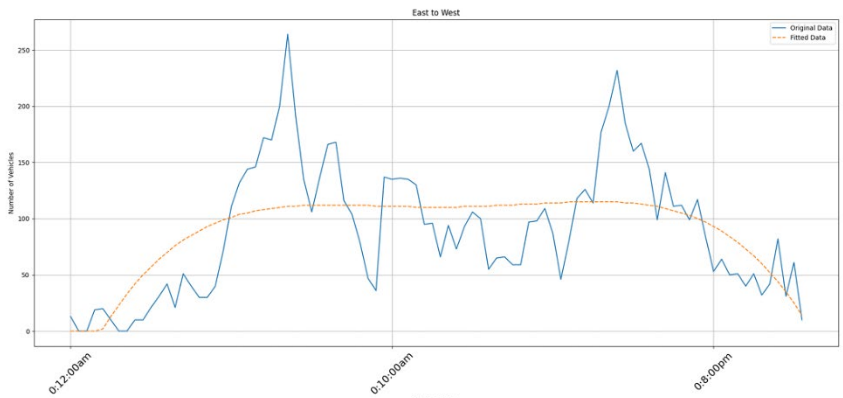



**Figure 18. Results of a polynomic traffic predictor; East-West bound when p=5**

As seen from figures 22 and 23, the polynomial models perform poorly when it comes to predicting traffic. One remedy for that is to increase the order of the polynomial model. Figure 24 shows the results for $p = 100$ .

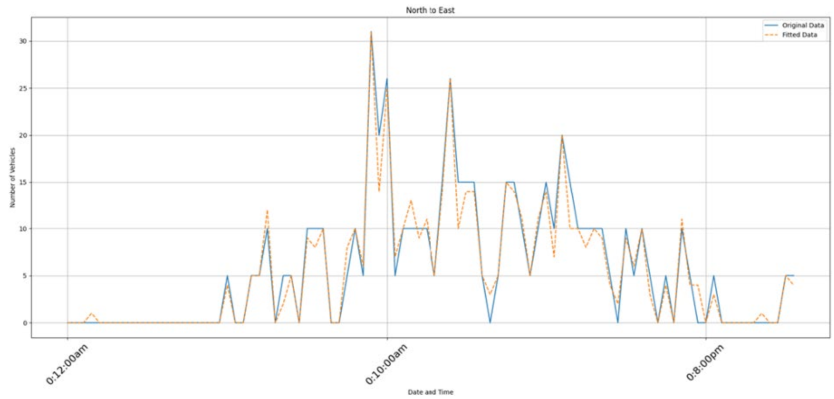



**Figure 19. Results of a polynomic traffic predictor; North-East bound when p=100**

Comparing Figure 19 to Figure 17 for the North-East bound, one can easily note that the performance of the polynomial predictor model has improved substantially. Despite this, very high order polynomials can lead to overfitting. That is, the model may perform well on training data but poorly on unseen data.

## Seasonal Models

Researchers note that traffic shows similar repetitive patterns day after day, hence a *seasonal model* could potentially lead to better performance over a polynomial model. The Seasonal Model

$$\hat{f}(x) = \widehat{f_{poly}}(x) + \widehat{f_{seas}}(x)$$

accounts for regular patterns, $\widehat{f_{poly}}(x)$, and periodic fluctuations, $\widehat{f_{seas}}(x)$, in traffic data. By incorporating both polynomial and seasonal components, this model can better handle variations that occur at regular intervals, such as daily or weekly traffic patterns. Figure 20 illustrates the effectiveness of a Seasonal model in capturing these recurring trends for the polynomial of order $p = 1$ (*i.e.*, a line).

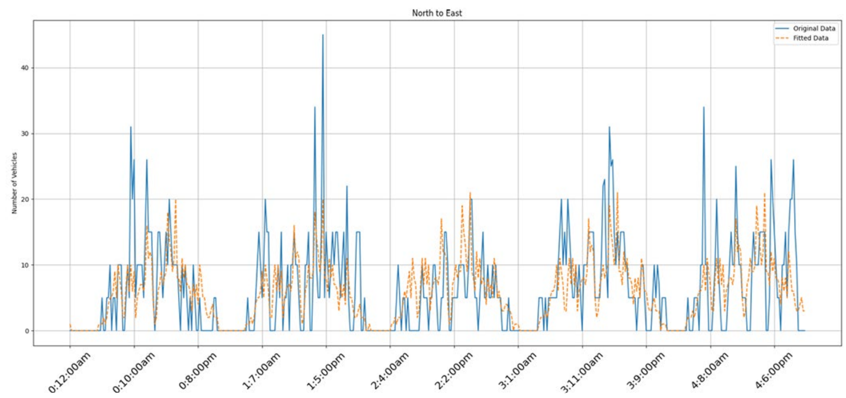



**Figure 20. Results of a seasonal traffic predictor for 5 consecutive days; North-East bound**

## Auto-Regressive (AR) Model

As discussed before, the Auto-regressive (AR) model predicts future traffic values based on past observations. The AR model considers a certain number of previous data points (known as lags) to make future predictions. The AR model can mathematically be described by

$$\hat{y}_{t+1} = \theta_1 y_t + \theta_2 y_{t-1} + \cdots + \theta_p y_{(t-P+1)}$$

Where here $t$ is the time, $\hat{y}_{t+1}$ the predicted traffic at time $t+1$ , $y_t$ the actual traffic at time $t$, and $P$ is the number of lags. Figures 21 and 22 show the performance of two AR models when the number of lags are $P = 8$ and $P = 24$ , respectively.

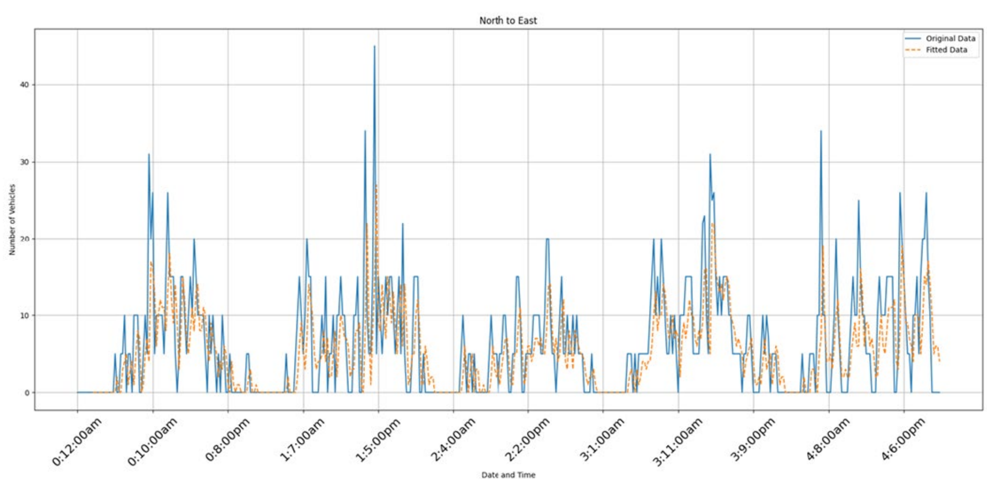



**Figure 21 Results of an AR; traffic for P=8; North-East bound**

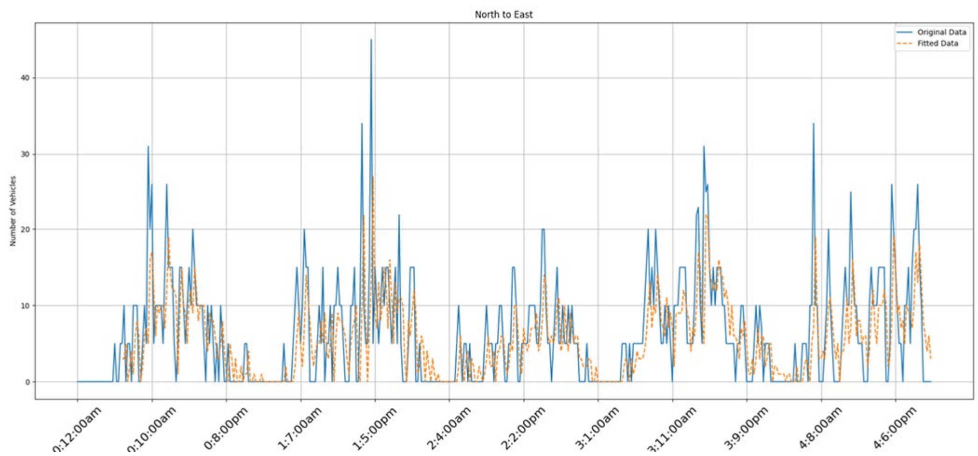



**Figure 22 Results of an AR; traffic for P=24; North-East bound**

As seen the performance of the AR models improves when choosing a higher number of lags, though it can also lead to overfitting if too many lags are used.

## Simulation Results – Neural Networks

An LSTM RNN traffic prediction model is developed and trained for up to 100 epochs with a batch size of 32, using 80% of data for training and the remaining 20% for model validation. The model is evaluated on the test set, and the loss and metrics are noted. Figure 23 illustrates the comparison between true and predicted values, providing insights into the model's performance.

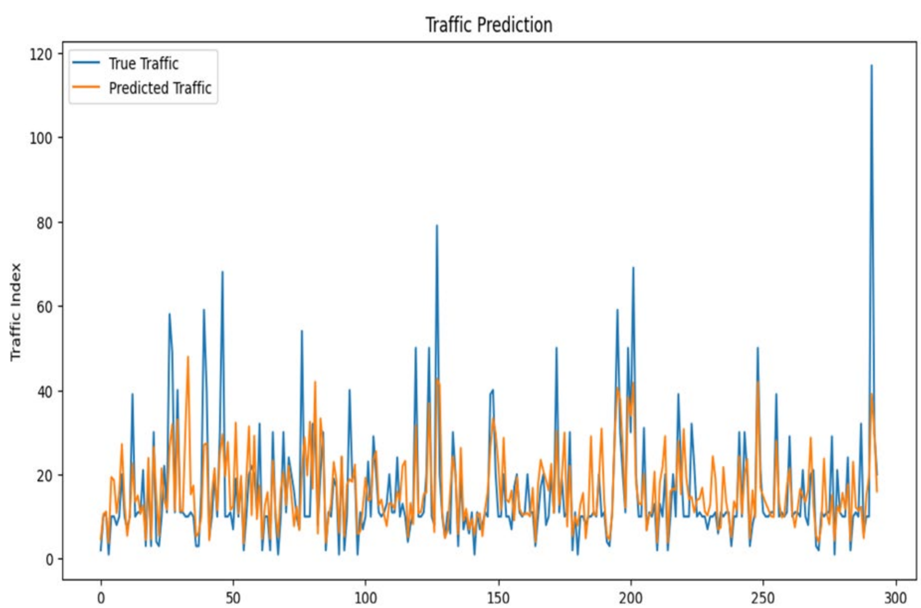



**Figure 23 Results of an LSTM RNN traffic predictor**

We note that the Mean Absolute Error (MAE) achieved for this simulation was 6.35 which is quite promising.

# Data Processing and Analysis

To extract data from the collected traffic footage, researchers used an application called TrafficSurveyViewer by ITS company, DataFromSky. The company provides a platform that allows for the extraction and analyses of video clips obtained from filming the traffic in the intersection. Using that software, researchers compared the data extracted from traffic footage with the data provided by StreetLight. Our initial analysis indicates that the data from the two different sources (i.e., video footage and StreetLight) relatively follow each other. Having a reliable resource for footage and data analysis allows for community-based research to progress smoothly and, as this pilot study has shown, proves to be an essential tool in an effort to produce a guidebook for scalable analysis of a range of different traffic nodes.

DataFromSky's TrafficSurveyViewer is designed for in-depth traffic video analysis, excelling in object detection and tracking within video footage, enabling the visualization of object trajectories and classification into up to 19 categories, such as vehicles, bicycles, and pedestrians. It captures critical traffic metrics including speed, acceleration, deceleration, headway, and gap time, which are fundamental for evaluating traffic flow and safety. The application supports detailed Turning Movement Counts (TMC) at intersections, offering insights into vehicle turning patterns and traffic volumes. Its Origin-Destination (O-D) analysis functionality maps vehicle movements between various segments and intersections. The safety analysis features computing metrics like time to collision, post-encroachment time, and heavy braking events, thereby identifying potential safety concerns. TrafficSurveyViewer allows users to configure gates and traffic regions to count vehicles and analyze specific areas of interest, making it an ideal choice for extracting raw data from the captured traffic footage.

## Results

To extract TMC data from the traffic footage, gates similar to those placed in the StreetLight analysis are placed on the idle video frame of the footage as shown in Figure 24.

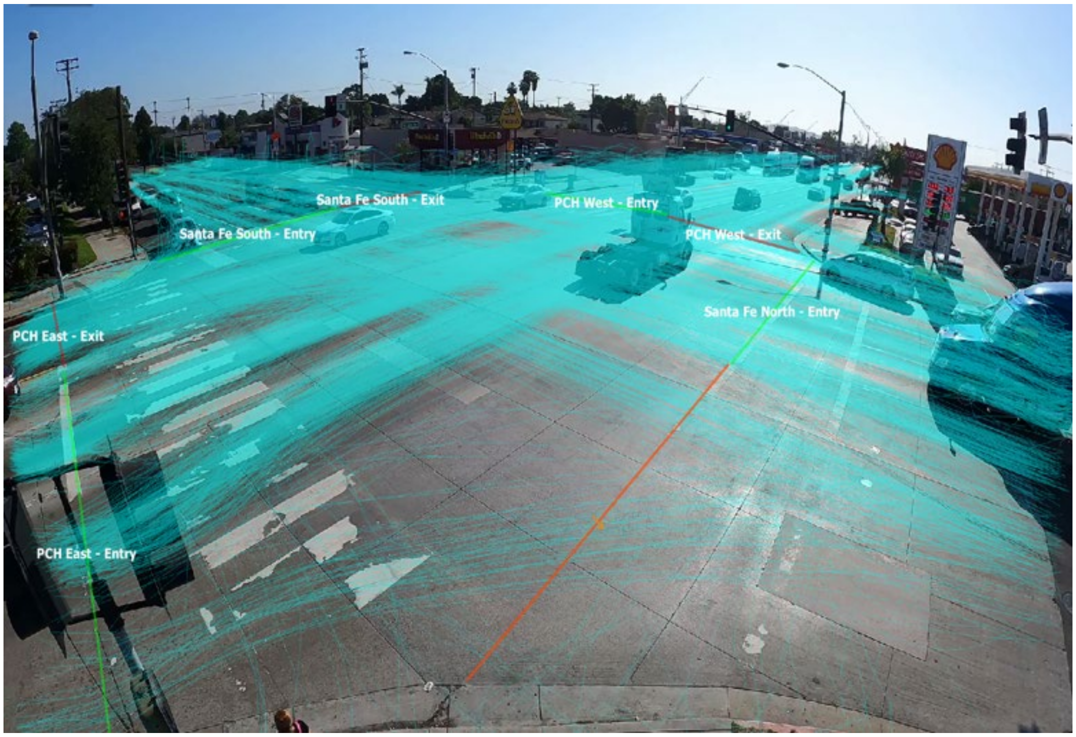



**Figure 24. Screenshot of DataFromSky**

Six days of traffic video footage was captured at the Santa Fe and PCH intersection. The recorded dates and times are as follows:

- Monday, April 22: From 7:00 AM to 9:00 PM
- Thursday, June 13: From 11:00 AM to 4:00 PM
- Friday, June 14: From 11:30 AM to 5:00 PM
- Saturday, June 15: From 11:00 AM to 5:00 PM
- Thursday, June 20: From 12:00 PM to 6:00 PM
- Friday, June 21: From 8:30 AM to 3:00 PM

Researchers selected the traffic footage captured on June 20, 2024, for further analysis. This date was chosen because it represents the longest continuous recording among all the days. The selected data covers traffic between 12:00 PM and 5:00 PM, totaling 4 hours, 14 minutes, and 14 seconds.

Researchers used this dataset to study truck activity in the area, emphasizing the role of trucks in the traffic congestion observed at the intersection. As previously mentioned, we've observed a high degree of similarity between the traffic data obtained from StreetLight and that from DataFromSky. Consequently, the findings presented in Table 2, based on DataFromSky, can be extended to the significantly larger dataset available from StreetLight.

Table 2 Overall Statistics of Vehicles Traversing the Intersection of Pacific Coast Highway and Santa Fe Avenue

| Vehicle Classification | Vehicle Count |
| --- | --- |
| All Vehicles | 11635 |
| Car | 9419 |
| Truck | 1446 |
| Pedestrian | 307 |
| Van | 215 |
| Motorcycle/ Bicycle | 134 |
| Bus | 114 |

From Table 2 we observed that trucks are the second highest number of vehicles travelling through the intersection. They make up 12.4% of the total traffic during the captured period, with a ratio of 3:25 between trucks and the total number of vehicles. This can be hypothesized that three in every twenty-five vehicles are trucks. Although this figure does not seem significant, researchers can take the inference from the analysis of truck traffic trends of PCH and Santa Fe Avenue compared to similar intersection (such as North Bellflower Boulevard and East Spring Street in Long Beach), to draw conclusion that the observed number of trucks is much higher than most regular intersections.

In addition to our previously discussed research, we've developed a traffic micro-simulator to get deeper insights into the behavior of traffic at the intersection. During this process, we evaluated two popular platforms for building our simulator: VisSim and SUMO (Simulation of Urban Mobility). Ultimately, we chose SUMO due to its simplicity and shorter learning curve.

Using the data collected from StreetLight and DataFromSky, our goal was to create a realistic and representative simulator. Although it's still in its final stages of development, this simulator looks very promising. Once fully functional, it will allow us to explore and evaluate a wide range of traffic scenarios.

# Data Privacy

## Methodology

*Route Design, Volunteer Recruitment and Participation*

Our research team designed a 1.5-mile data walk route that intentionally took volunteers through or past a variety of spaces, including a commercial corridor; a residential neighborhood; a 12-acre city park; and a public high school with enrollment of more than 2,000 students. Researchers also facilitated focus group discussions before and after each walk. These qualitative methods aimed to answer these research questions:

> **RQ1**: How do civic technologies in the public realm impact residents' attitudes and levels of trust in local government?
>
> **RQ2**: Would increased transparency and accountability associated with the use of smart technologies foster trust in local government?

Researchers undertook broad recruitment efforts starting about two weeks prior to the data walks. A recruitment flyer described the research and offered participants a $25 gift card as an incentive for a 2-hour commitment. The recruitment message also noted that volunteers must be at least 18 years old; work, live or attend school in Long Beach; and have access to a smart phone or tablet. Researchers disseminated the recruitment flyer to the following people/organizations.

- ☐ **All nine City Council members:** Mary Zendejas (District 1), Cindy Allen (District 2), Kristina Duggan (District 3), Daryl Supernaw (District 4), Megan Kerr (District 5), Suely Saro (District 6), Roberto Uranga (District 7), Al Austin (District 8), and Joni Ricks-Oddie (District 9)
- ☐ **City Departments:**

  City of Long Beach Technology and Innovation Department

- ☐ **Community-based and academic organizations:**

  ChildLane, Goldstar Manor, YMCA of Greater Long Beach, Gray Panthers, Downtown Long Beach Alliance, CITT newsletter and networks

- ☐ **Neighborhood associations**

  Belmont Heights Neighborhood Association, Bluff Heights Neighborhood Association, Bluff Park Neighborhood Association, Carroll Park Association, Deforest Park Neighborhood Association, Nehyam Neighborhood Association

- ☐ **Social media posts**:

  LinkedIn, Instagram

- ☐ **Emailed dozens of individual contacts**

Despite these far-ranging recruitment efforts, about a dozen people registered through a Microsoft Forms survey that enabled them to preference the date and time of walks. (Researchers attribute the low engagement rate, primarily, to the fact that the data walks coincided with the final days of school for Long Beach public schools—making it difficult for teachers and parents to attend a daytime activity). The form also collected basic demographic information, including age, ethnicity, gender, education level, annual income and zip code.

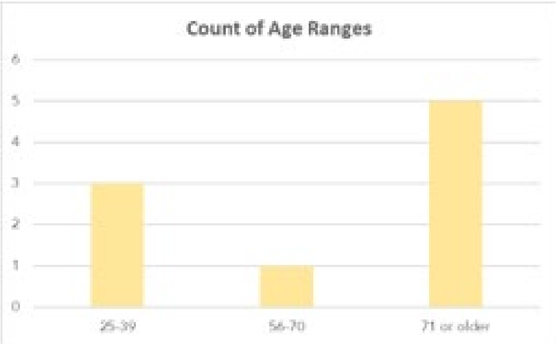



**Figure 25. Survey participant age range count**

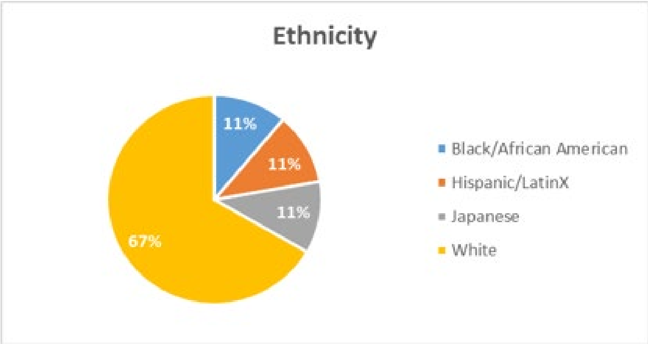



**Figure 26. Survey participant ethnicity**

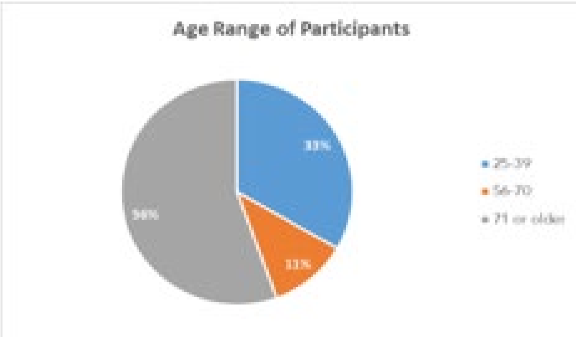



**Figure 27. Survey participant age ranges**

Several days prior to the start of each walk, researchers sent a confirmation email to registered volunteers. The message instructed them to access a custom version of an Esri mobile app, which they would need to submit responses to prompts during the walks. Our research team chose the Esri app, specifically, because it is cloud-based and does not collect or store personally identifiable information.

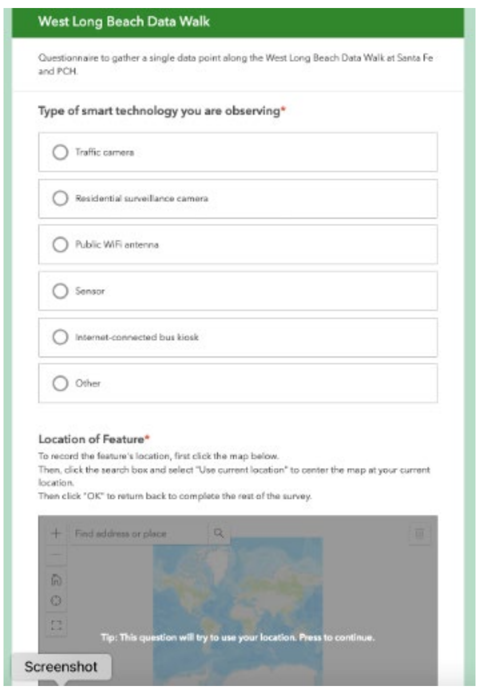



**Figure 28. Screenshot of Data walk App**

Participants met at the Bret Harte Neighborhood Library in West Long Beach. After everyone reviewed and signed an informed consent document, researchers explained the purpose of the study. Researchers then engaged in a semi-structured focus group discussion—lasting between

15 and 20 minutes—meant to sensitize volunteers to the general topic of data privacy. The conversations were semi-structured but centered around these guiding questions:

1. What technologies have already encountered today, just moving through your typical routine, that collect personally identifiable information?

2. Do you think about your data privacy and, if so, in what contexts?

3. Do you believe there is too much technology installed around the city?

4. What do you expect to observe during the walk?

Following the conversation, researchers started the walk.

On the data walk route, participants encountered more than a dozen smart technologies. They stopped to respond to prompts when interacting with these devices or platforms:

Stop 1: Traffic camera at Santa Fe Avenue  Avenue and Willow Street

Stop 2: Live video recording for CITT-developed traffic sensor at intersection Santa Fe Avenue Avenue and Willow Street

Stop 3: Private surveillance camera mounted in doorway of Hong Kong Express restaurant

Stop 4: Long Beach Transit apps at bus stop at intersection of Hill Street and Santa Fe Avenue Avenue

Stop 5: City-deployed security camera in Admiral Kidd Park

Stop 6: Bike Share docking station at entrance to Admiral Kidd Park

Stop 7: Car charger in Admiral Kidd Park parking lot

Stop 8: Public WiFi router at Admiral Kidd Park

Stop 9: Ring video doorbell at 2290 Cota Ave

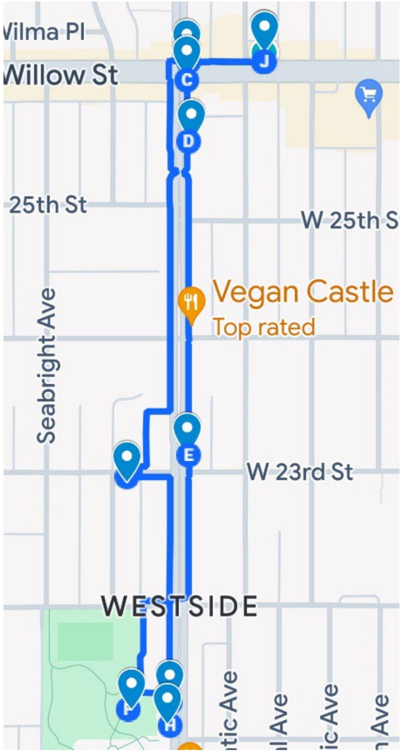



**Figure 29. Map of Data Walk Route**

Following each data walk, study participants returned to the Harte Library for a "debriefing" focus group discussion. While these conversations were also semi-structured, guiding questions included:

1. Did anything you observed during the walk surprise you?
2. Do you have any unique concerns about the traffic camera and future traffic sensor? Does anything about these technologies stand out for you?
3. Is it government's obligation to inform residents about data collection by civic technologies in the public realm, or are residents responsible for their own awareness of these smart technologies?
4. Do the benefits of data collection, such as traffic management, outweigh potential privacy violations?

For the analysis, researchers relied on a grounded theoretical approach in which themes emerged from the text of the pre- and post-walk interview transcripts and responses submitted to the mobile app. Specifically, researchers linked discourses of data privacy violations, trust, comfort levels and transparency with a critical perspective on smart city technologies. They then systematically coded the qualitative data (consisting of transcripts from the focus group discussions and comments typed into the Esri app), in accordance with the "constant comparative" method developed by Glaser and Strauss (1999)[xi]. Researchers read each transcript multiple times to form a systematic analysis through Nissenbaum's (2004)[xii] contextual integrity framework and theories associated with information economics, surveillance studies and cultures of trust literature. This process related data to ideas, then ideas to other ideas. Researchers used the qualitative data analysis platform Nvivo 14 to code the transcripts according to thematic relevance. The Nvivo 14 interface provided a streamlined structure, helping me draw links among data sources (comments made during focus group discussions and typed into the Esri app). Next, researchers compared these nodes against the conceptual arguments supporting this study and our research questions. They then analyzed and revisited these themes, using techniques that are cyclical and iterative.

## Findings

A reoccurring theme of focus group discussions was the perception that devices and platforms deployed by the City of Long Beach and other government entities are simply part of a ubiquitous surveillant culture. As one study participant in his early 30s commented, "I think if it's people who are younger, we grew up in this sort of environment. So I feel like it's at least partially common sense to probably assume that there are cameras out there." Another focus group participant commented that she is resigned to the reality that city-deployed devices and platforms will collect information about her. "I just accept what's happening because there's not really much we could do to change anything. So it's just like, okay, be aware...and take certain precautions...It's just like, eh, it's happening anyways." A response typed into the app during an encounter with electric vehicle chargers noted that the technology does collect significant amounts of personal information, "but we enter our credit card info so often."

During both focus group discussions and in comments submitted through the Esri app, study participants alluded to knowingly trading personal privacy in exchange for the benefits of paying utility bills online, reporting potholes through the Go Long Beach app, and checking the bus schedule on the Moovit app, among other routine tasks. Notably, a majority of study volunteers reported feeling ambivalence about these trade-offs. For example, a respondent said she doesn't love the idea of being recorded by surveillance cameras as she moves through her day but, at the same time, recognizes the benefits. "Seeing the security cameras in a weird way sometimes makes me feel safer than if they weren't there." Another study volunteer shared an anecdote about being grateful for traffic cameras in the wake of a recent car accident. "Once you get into this whole insurance situation, cameras can be helpful in proving that you're innocent, instead of relying on eyewitness statements." One focus group participant commented that she thinks about "the pros and cons" of surveillance technology, while another said, "If technology could sync the lights when I'm driving down the big streets, I'd be real happy with that." Finally, one participant noted that transparency is more important for certain technologies. "For example, the car chargers and the bike share, they get credit cards and personal information," she pointed out. An app user expressed data privacy concerns associated with using public Wi-Fi at the park, but that he/she would still use it "because of the convenience."

The prevailing sentiment emerging from each of the three data walks was that government entities have a responsibility to be transparent about data it collects, the purpose of collecting data, and storage practices. One study volunteer said that, because residents put their trust in public institutions, "we don't know the nuances of what questions to ask or where to research," so government should provide general information about data collection. "And then, if you want further information, the onus is on you to give that," she added. One person said she would appreciate more than a simple notice or privacy policy: "It would be really nice to have a live person to call and talk to." Another study participant commented, "When it comes to third-party vendors, you might want to know who your city is contracting with and if they had any issues in the past." During an encounter with traffic cameras, one participant expressed ambivalence with the technology. "It does concern me. If the data is used to make streets safer, then it is good. If data is sold to third parties for financial gain, then it is not good."

Of note, during all three data walks, study participants consistently referenced a November 2023 cyberattack against the City of Long Beach. Many commented that the incident, which forced the City to go offline for two weeks and to acknowledge stolen data, shook their trust in the City's ability to keep PII safe. "We need to know about security of our data. Lack of transparency is one of the reasons why trust in government is eroding. So it's imperative the city be honest about what they're doing."

## Data Privacy Policy Recommendations

While the findings are limited due to the small sample size, researchers can make several tentative policy recommendations based on data collected during data walks and focus group discussions:

1. Public engagement and education about a technology's intended purpose, functions and benefits are vital to instilling trust. When policymakers neglect to articulate the positive aspects of even benign smart city projects, members of the public are left to speculate—and to draw conclusions that sometimes clash with reality and, perhaps, unnecessarily foster distrust. Conversely, when civic technologies infringe upon residents' privacy, local government should be held accountable.
2. The City of Long Beach and other government entities that deploy surveillance technologies should develop data privacy policies and make these documents widely available. The policies should include actionable provisions. For example, government entities might create risk-based procedures and minimum thresholds for evaluating the privacy impacts of "smart" technology systems. Another actionable policy provision could mandate that technology systems posing a medium or high risk to data privacy go through an established privacy impact assessment before being adopted/deployed.

3. The City of Long Beach should continue to support deployment of its Digital Rights Platform, which that uses physical signage and an online portal to provide residents with a clear understanding of how local government applies predictive and diagnostic analytics to personal data. Further, City of Long Beach officials should work cross-jurisdictionally to advance adoption of the Digital Rights Platform standard in neighboring and peer municipalities throughout California.

# Conclusion

The purpose of this study was to develop a set of guidelines that would allow community groups across the country to conduct mobility studies of nodes in their respective transportation networks. The approaches employed in this study prove that such community-based mobility labs are scalable not just across the state of California, but across the United States. While Long Beach is a very forward-looking city with its incorporation of the Technology and Innovation Division, the acquisition of traffic footage presented a primary challenge to this project. Our research found that Caltrans traffic cameras were unsuitable because none were located at the intersection of study. Programs like StreetLight and Iteris were also inadequate as they did not provide live footage. StreetLight data was sometimes unavailable or “normalized,” and our attempts to understand this process were unsuccessful. Additionally, StreetLight did not provide truck traffic percentages, complicating the development of our NN predictor.

Prior to the start of this study, researchers determined that the use of drones for footage collection was not feasible, as their usage would have required a substantial financial commitment for insurance reasons and would not be a suitable approach community groups attempting to conduct similar studies. Researchers’ decision to approach and conduct the traffic recordings themselves proved successful, as they gained invaluable insights from being present on-site, the approach does not require any permits and thus proves the nation-wide scalability of community-based research.

## Data Privacy Results

Results from the data walks and related discussions showed that a majority of participants were ambivalent about the presence and usage of smart technologies employed by the city as long as it would be held accountable for any negative consequence thereof. It was further found that a lack of public engagement and outreach on behalf of the city and its officials regarding the use of smart technologies unintentionally increased public apprehension, as the lack of communication led to increased speculation about any negative consequences of smart technologies and left study participants in the dark about any positive aspects or outcomes. Study participants revealed that the establishment of the Long Beach Digital Rights Platform as well as the development and widespread communication of actionable privacy policies and associated risk evaluations could significantly decrease public apprehension about the use of smart technology by the City of Long Beach and other institutional actors.

## Policy Recommendations

The use of traffic cameras owned by Caltrans proved impossible for this study because there were no cameras located at the intersection of study. Thus, while these cameras potentially could be used in other places, their distribution imposes substantial limits and determines the area of study as opposed to the researchers recording traffic footage themselves. In order for community groups to efficiently identify an ideal approach for their study, identifying and mapping traffic cameras owned by institutional entities, such as Caltrans in the State of California, can greatly aid the development of studies such as this. However, there are many jurisdictional hurdles preventing easy access to such cameras.

During the course of this study, researchers identified two possible approaches for collecting traffic footage in future studies. The first option would be to deploy high-altitude balloons equipped with recording equipment over a desired area of study to obtain the necessary footage. The second option would be to access to low-earth or geosynchronous orbiting satellites to capture the footage. Initial outreach to NASA's Jet Propulsion Laboratory and a private satellite company called Maxar indicates that such a collaboration could indeed be possible pending further research. Continued and coordinated outreach efforts increase the scalability of community-based research and minimize bureaucratic hurdles for localized research efforts.

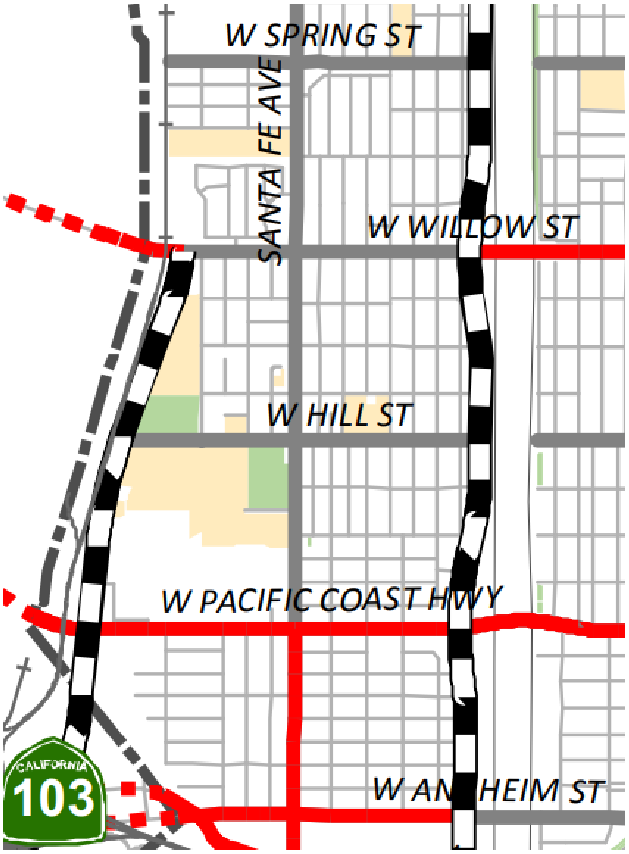
This study produced a somewhat unexpected finding, which is the need for safety protocols. During the study, two incidents highlighted the need for safety protocols. While Dr. Gwen Shaffer was facilitating a focus group discussion at the Bret Harte Neighborhood Library, a hit-and-run driver slammed into her parked car. According to witnesses, the driver crossed the median and drove on the wrong side of the street.

The second incident occurred when during the collection of traffic footage, research associate Devin Martinez-Flores watched as a commuter vehicle nearly crash into a tripod at the intersection of Pacific Coast Highway and Santa Fe Avenue. While nobody was injured in either incident, similar projects in the future would need to develop safety protocols to protect facilitating researchers and study participants.

This project overall met its goal to provide an example of community-based research in transportation networks and facilitate guidelines for the development and implementation of similar future projects. Considering the policy recommendations made, the potential of community-based research is abundant not only in here in California, but across the United States.

# Appendix

**1. Truck Routes in West Long Beach**




*Appendix Section 1 Truck Routes West Long Beach*

## 2. City of Long Beach Technology and Innovation Request Denial

Good afternoon, Devin. I apologize for all the circling around. I do see the issue and why your PRA was denied. The traffic CCTV system is accessible to our internal staff within Public Works. Systems such as these are not available for the public or, in this case, for such research requests. Please note that the Pacific Coast Highway corridor belongs to the California Department of Transportation and any request will need to be made directly to that agency. Something to consider is a partnership with the City of Long Beach(Public Works) regarding the research you are conducting. This will likely need approval from the Director of Public Works, the City of Long Beach Council, the City Attorney, and the City Manager.

Thank you,

**Hugo Gil**
*Infrastructure Systems Officer- Telecommunications Division*

**Technology & Innovation | Infrastructure Services Bureau**
411 W. Ocean Blvd., 7th Floor | Long Beach, CA 90802
Office: 562.570.4811

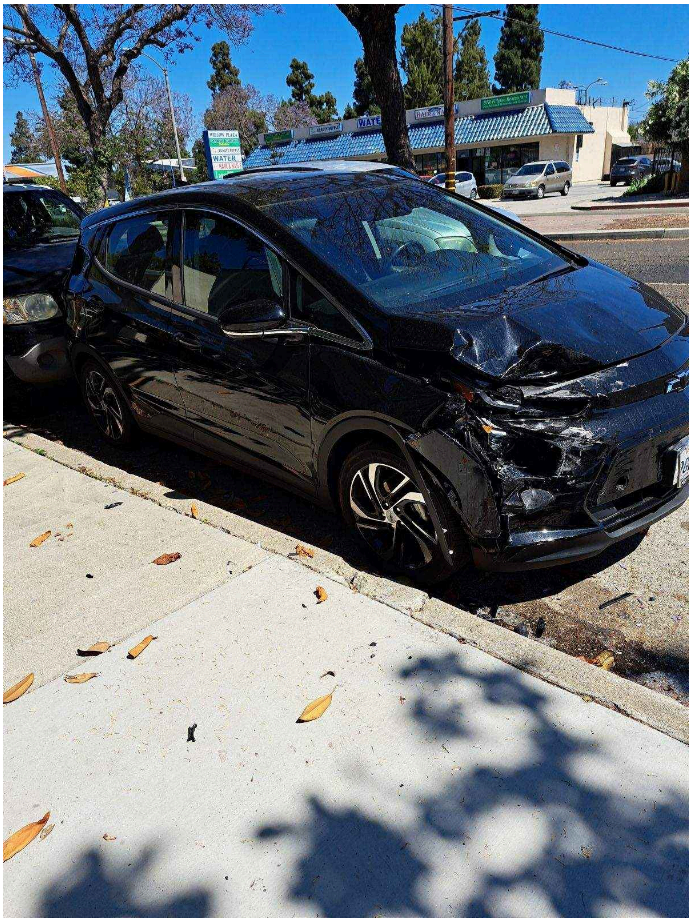
TECHNOLOGY
& INNOVATION
LONGBEACH

*Appendix Section 2 City of Long Beach Technology and Innovation Request Denial*

## 3. Camera Placement on the Intersection of Pacific Coast Highway

*Appendix Section 3 Camera Placement on the Intersection of Pacific Coast Highway*

**4. Picture of Dr. Gwen Shaffer's Car after a Hit and Run at Harte Neighborhood Library**



*Appendix Section 4 Picture of Dr. Gwen Shaffer's Car after a Hit and Run at Harte Neighborhood Library*

## 5. Footage of a failed cycle at Pacific Coast Highway and Santa Fe Avenue


*Appendix Section 5 Footage of a failed cycle: https://youtu.be/ivsqocnzW38I*

# References

[i] Shaffer, Gwen, Hossein Jula, Anastasios Chassiakos, and Tyler Reeb. "Using Artificial Intelligence to Improve Traffic Flows, with Consideration of Data Privacy," September 2022.

[ii] IBM. "What Is a Neural Network? | IBM," October 6, 2021. https://www.ibm.com/topics/neural-networks.

[iii] Shaffer, Gwen, Hossein Jula, Anastasios Chassiakos, and Tyler Reeb. "Using Artificial Intelligence to Improve Traffic Flows, with Consideration of Data Privacy," September 2022. https://www.metrans.org/assets/research/psr%2020-13%20final%20report_shaffer.pdf.

[iv] Levin and Y.-D. Tsao, "On forecasting freeway occupancies and volumes (abridgment)," Transp. Res. Rec., no. 773, pp. 47–49, 1980.

[v] Tedjopurnomo, David Alexander, Zhifeng Bao, Baihua Zheng, Farhana Murtaza Choudhury, and Alex Kai Qin. "A survey on modern deep neural network for traffic prediction: Trends, methods and challenges." IEEE Transactions on Knowledge and Data Engineering 34, no. 4 (2020): 1544-1561.

[vi] PeMS, California Department of Transportation s Performance Measurement System (PeMS). url: https://pems.dot.ca.gov/.

[vii] Tedjopurnomo, David Alexander, Zhifeng Bao, Baihua Zheng, Farhana Murtaza Choudhury, and Alex Kai Qin. "A survey on modern deep neural network for traffic prediction: Trends, methods and challenges." IEEE Transactions on Knowledge and Data Engineering 34, no. 4 (2020): 1544-1561.

[viii] Ibid.

[ix] Zhao, Zheng, Weihai Chen, Xingming Wu, Peter CY Chen, and Jing meng Liu. "LSTM network: a deep learning approach for short-term traffic forecast." IET intelligent transport systems 11, no. 2 (2017): 68 - 75.

[x] Kang, Danqing, Yisheng Lv, and Yuan-yuan Chen. "Short-term traffic flow prediction with LSTM recurrent neural network." In 2017 IEEE 20th international conference on intelligent transportation systems (ITSC), pp. 1-6. IEEE, 2017.

[xi] Glaser, A. & Strauss, A. (1999). *The Discovery of grounded theory: Strategies for qualitative research* (2nd Ed.). New York, NY: Transaction Publishers.

[xii] Nissenbaum, Helen. "Privacy as Contextual Integrity." *Washington Law Review* 79, no. 1 (2004): 119-157. https://crypto.stanford.edu/portia/papers/RevnissenbaumDTP31.pdf.