EXPLORING CROWDSOURCED BIG DATA TO ESTIMATE BORDER CROSSING TIMES

by

Xiao Li, Ehsan Jalilifar, Michael Martin, Bahar Dadashova, David Salgado, and Swapnil Samant

Project performed by

Center for International Intelligent Transportation Research

185921-00011
Exploring Crowdsourced Big Data to Estimate Border Crossing Times – Phase I

Estimating Border Wait Times Using Streamed Vehicle Data – Phase II

December 2022

Report prepared by
Center for International Intelligent Transportation Research
Texas A&M Transportation Institute
4050 Rio Bravo, Suite 151
El Paso, Texas 79902

TEXAS A&M TRANSPORTATION INSTITUTE
The Texas A&M University System
College Station, Texas 77843-3135

TABLE OF CONTENTS

Page

List of Figures ..... **vi**

List of Tables ..... **vii**

List of Acronyms ..... **viii**

Disclaimer and Acknowledgments ..... **ix**

Executive Summary **1**

Chapter 1: Introduction **2**
Problem Statement and Background 2
Objectives 2
Impacts and Implications 3
Outline of Report 3

Chapter 2: Literature Review **4**
Trends and Effects of Crossing times at U.S. Border Crossings 4
Existing Techniques for Border Crossing time Monitoring 5
Emerging Data Sources for Border Crossing time Estimation 6
Potential Solution for Border Crossing time Estimation 8

Chapter 3: Data Collection and Processing **10**
Study Site—Paso del Norte International Bridge 10
Border Crossing Information System 11
Connected Vehicle Data (Wejo) 12
Border Crossing time Calculation Based on Bluetooth Observations 13
Border Crossing time Calculation Based on Wejo CV Data 14

Chapter 4: Assessments of CV-Based Border Crossing times **17**
Temporal Coverage Assessment 17
Similarity Assessment 21

Chapter 5: Border Crossing Time Modeling Using CV Data **23**
Basic and Transformed OLS Models based on CV-Time Only 23
Additional CV-Generated Variables 26
Random Forest model 27
Gradient Boosting Model 28
Modeling Results Comparison 29

Chapter 6: Streaming data processing **31**

Chapter 7: Conclusions and Limitations **33**
Conclusions 33
Limitations 33

References **35**

LIST OF FIGURES

Page

Figure 1. Paso del Norte POE Aerial View. 10
Figure 2. Installation of Bluetooth Readers along the PDN POE.. 11
Figure 3. Illustration of Wejo CV Data at the PDN POE. 13
Figure 4. Illustration of Border Crossing Time Calculation. 15
Figure 5. Wejo Data Extraction Buffers around Each Bluetooth Reader at the PDN POE.......... 16
Figure 6. Comparsion between Bluetooth-Time and CV-Time in September 2022. 17
Figure 7. Monthly Average of CV-covered Slots Percentages for 2020 to 2022. 18
Figure 8. Temporal Coverage Assessment Results of the CV-Time in September 2022............ 20
Figure 9. Temporal Coverage Assessment Results of the CV-Time in October 2022. 20
Figure 10. Chart Correlation Plot between the Bluetooth-Time and the CV-Time in the Training Dataset in September and October 2022. 24
Figure 11. Chart Correlation Plot between the Square Root Transformed Bluetooth-Time and the CV-Time in the Training Dataset in September and October 2022. 25
Figure 12. Random Forest Model Feature Importance. 28
Figure 13. streaming data analysis process. 31
Figure 14. Streaming Data Ananlysis Prototype Flowchart. 32

LIST OF TABLES

Page

Table 1. Similarity Assessment Result of the CV-Times in September 2022. 22
Table 2. Similarity Assessment Result of the CV-Times in October 2022. 22
Table 3. Simple OLS Result Based on the CV-Time Only. 24
Table 4. Transformed OLS Result Based on the Square Root Transformed CV-Time Only.. 26
Table 5. Description of Additional CV-Generated Variables... 27
Table 6. Performance of Random Forest Models with All and Selected Variables. 28
Table 7. Performance of the Gradient Boosting Models with All and Selected Variables. 29
Table 8. Modeling Results Comparison based on the test dataset.. 29

LIST OF ACRONYMS

Abbreviation	Definition
AIC	Akaike's information criterion
BCIS	Border Crossing Information System
BLE	Bluetooth low-energy broadcasting
Bluetooth-Time	Bluetooth-based border crossing times
Bluetooth_Trans	Square root transformed Bluetooth-Time
CBP	Customs and Border Protection
CIITR	Center for International Intelligent Transportation Research
CV	Connected vehicle
CV-Time	CV-based border crossing times
CV_Trans	Square root transformed CV-Time
FHWA	Federal Highway Administration
GPS	Global positioning system
ITS	Intelligent transportation systems
MAC	Media access control
MAPE	Mean absolute percentage error
OLS	Ordinary least squares model
PDN	Paso del Norte
PDN-BT	Bluetooth readers installed at the PDN POE
POE	Port of entry
RFID	Radio frequency identification
RMSE	Root mean square error
TBBW	Toronto Buffalo Border Crossing time
TTI	Texas A&M Transportation Institute
TxDOT	Texas Department of Transportation
V2C	Vehicle-to-cloud communication
V2I	Vehicle-to-infrastructure communication
V2V	Vehicle-to-vehicle communication

DISCLAIMER AND ACKNOWLEDGMENTS

This research was performed by the Center for International Intelligent Transportation Research (CIITR), a part of the Texas A&M Transportation Institute (TTI). The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein.

The research team would like to thank CIITR for funding this research and Wejo for providing the testing data samples. TTI led the research. Any errors, inaccuracies, or omissions are the responsibility of the authors. The content of this research does not necessarily reflect the official views of the agencies mentioned.

EXECUTIVE SUMMARY

This project provides an in-depth assessment of market-available connected vehicle (CV) data—Wejo CV data in border crossing time estimation. This project evaluates both historical and streaming Wejo CV data collected at the Paso del Norte (PDN) port of entry (POE) in El Paso. The research team first developed a set of big data analytic tools to process Wejo datasets and generated CV-based border crossing times (CV-Time). Then, the research team evaluated the temporal coverage of CV-Time at the PDN POE and its correlation with the existing Bluetooth-generated border crossing times (Bluetooth-Time). The results demonstrate that the CV-Time is strongly correlated with the Bluetooth-Time with a correlation rate of approximately 0.8. The best fitted ordinary least squares model based on the combination of CV-Time and additional CV-generated variables can produce a temporally transferable model to estimate Bluetooth-Time with an R2 of 0.88, root mean square error (RMSE) of about 15 min and mean absolute percentage error (MAPE) of 52%. Random forest and gradient boosting models were also investigated to decrease the MAPE. The research findings show that the gradient boosting has the best performance among the introduced models with an RMSE of 15.5 min and MAPE of 26%. However, this evaluation was conducted based on the 2022 Wejo data with a relatively low penetration rate, resulting in around 30 percent of the testing hours lacking Wejo samples. Given these findings, the Wejo CV data can be a potential and promising data source for monitoring border crossing times, especially as sample penetration rates improve. The Wejo CV data can be used to supplement the existing Border Crossing Information System (BCIS) in El Paso and could be treated as an alternative solution when the BCIS is down for maintenance.

CHAPTER 1:
INTRODUCTION

PROBLEM STATEMENT AND BACKGROUND

Congestion at land ports of entry (POEs) has been socially recognized as a long-standing transportation issue for both commercial interests and the traveling public. El Paso, as one of the largest border cities in Texas, has a bi-national transportation network with about 20 million vehicle miles traveled per day. Driven by the exponential trade growth between the United States and Mexico, especially with the United States–Mexico–Canada Trade Agreement entered into force, the transportation efficiency at border crossings in El Paso is facing greater challenges. Effectively measuring border crossing times is of considerable interest and importance to various stakeholders. However, current measurement solutions (e.g., radio frequency identification [RFID]) usually require a large installed base of sensors (e.g., transponders and readers), which are costly to implement and maintain. These fixed sensors have spatial coverage limitations that preclude measuring long queues exceeding their coverage. Effectively and cost-efficiently monitor passing vehicles' crossing times at POEs remains a challenging problem in border transport management.

As the Internet of Things technology has developed over time, connected vehicles (CVs) are rapidly becoming the new paradigm of road transport, which has been widely believed to positively influence transportation safety, efficiency, and sustainability. CV data are collected from vehicles, directly reflecting the dynamics of traffic mobility. For example, Wejo, as a leading CV data start-up, provides high sampling and multi-dimensional vehicle movements and driving event (e.g., hard braking, hard acceleration, and speeding) data. These emerging CV data offer researchers unprecedented, detailed traffic data that are proving to be tremendously helpful in gaining new insights into traffic congestion.

OBJECTIVES

In this project, the research team explored the use of an emerging crowdsourced data source, Wejo CV data, to estimate crossing times for passenger vehicles at border crossings in El Paso. To the best of our knowledge, this project is among the first to comprehensively evaluate the effectiveness of the market-available CV data in congestion studies at border crossings. Through this project, we aimed to:

- Develop a set of big data analytic tools to process Wejo CV datasets.
- Evaluate the temporal coverage of Wejo CV data (both historical and streaming) at the Paso del Norte (PDN) POE in El Paso and its correlation with the existing Bluetooth-based border crossing times (Bluetooth-Time).
- Develop effective models to estimate the Bluetooth-Time based on CV-generated variables.

IMPACTS AND IMPLICATIONS

This project marks the first attempt to conduct an in-depth assessment of the market-available CV data in border crossing time monitoring. The research team partnered with Center for International Intelligent Transportation Research (CIITR) border study experts (David Salgado and Swapnil Samant) to develop new solutions to improve the estimations of border crossing times by leveraging the state-of-the-art CV data product and cloud computing and storage system. This data-driven practice can help transportation and planning agencies, such as CIITR, the Texas Department of Transportation (TxDOT), and the El Paso Metropolitan Planning Organization, better assess traffic conditions at a much finer spatiotemporal scale and effectively recognize highly congested roadway segments. The proposed big data analytic framework can be used as a prototype and applied and transferred to other border-crossing regions. In the future, once this framework is fully tested and validated, it could be potentially used as an infrastructure-less source of travel time data to enhance the current Border Crossing Information System (BCIS).

OUTLINE OF REPORT

The remaining chapters of this report include the following:

- **Chapter 2—Literature Review:** This chapter summarizes the reviewed literature and focuses on the existing techniques and emerging data sources for border crossing time estimation.
- **Chapter 3—Data Collection and Processing:** This chapter presents the study area and the different types of data (Bluetooth observations and CV data) processed and used in the analysis.
- **Chapter 4—Assessments of CV-Based Border Crossing Times:** This chapter presents the evaluation of the CV-based border crossing times. We calculated the hourly average border crossing times based on the Bluetooth data (Bluetooth-Time) and the Wejo CV-based border crossing times (CV-Time). Then, we conducted two assessments to evaluate the temporal coverage of CV-Time and its correlation with Bluetooth-Time.
- **Chapter 5—Border Crossing Time Modeling Using CV Data:** This chapter introduces the modeling efforts to improve border crossing time estimation based on CV data. Three linear regression models are developed based on the CV-generated variables.
- **Chapter 6—Streaming Data Processing:** This chapter summarizes the procedure implemented in this study to capture and analyze the steaming CV data and update the results in near real-time.
- **Chapter 7—Conclusions and Limitations:** This chapter summarizes the study and conclusions drawn from the study and covers the project limitations.

CHAPTER 2: LITERATURE REVIEW

The inspection of vehicles and passengers at border crossings is of great importance for ensuring homeland security; the inspection also incurs various spillover effects on traffic, such as increasing border crossing time and the queue length of vehicles. The border crossing time is commonly defined as "the time it takes, in minutes, for a vehicle to reach the primary inspection booth of the Customs and Border Protection (CBP) after arriving at the end of the queue" (Rajbhandari, Villa, Macias, et al. 2012). Providing an accurate and practical crossing time estimation is of great importance to advance the understanding of urban flow dynamics and support intelligent transportation systems (ITS). Effectively monitoring crossing times and trends can not only benefit government agencies for POE management and planning purposes but can also support a variety of needs (e.g., when and where to cross the border) of border commuters, travelers, and private-sector business (San Diego Association of Governments 2017). This chapter reviews the border crossing time trends and effects at the U.S. border and summarizes the existing border crossing time measuring techniques, emerging data sources, and potential solutions for border crossing time estimation.

TRENDS AND EFFECTS OF CROSSING TIMES AT U.S. BORDER CROSSINGS

The United States is bordered by Mexico in the south and Canada in the north, forming some of the longest international borders in the world—2,000 and 5,500 miles long, respectively. There are 110 border crossings at the U.S.–Canada border and 44 border crossings at the U.S.–Mexico border (Villa et al. 2017). The majority of the U.S. trade with its largest trading partners, Canada and Mexico, is conducted via land transport through its southern or northern borders. As reported by Villa et al. (2017), in 2015, more than 28 million private vehicles and 5.8 million commercial vehicles entered the United States from Canada, and 74 million private vehicles and 5.5 million commercial vehicles entered the United States from Mexico. From 2009 to 2015, the annual volumes of private and commercial vehicles entering the United States continued to increase with an annual growth rate of 1.1 percent and 2.4 percent, respectively. The increase in the crossing border vehicle volume results in high crossing times at land POEs.

Delays at border crossings are costly to bordering nations. Border delays have a negative impact on the ability of economies to attract investors, and the cost of goods to consumers is increased to cover losses. The tourism industry is also affected negatively, and there are environmental impacts of congestion on border communities that cannot be overlooked (Khan 2010). Estimated by the Ontario Chamber of Commerce, the U.S.–Canada border delays cost approximately $4.13 billion per year for the United States, which is about 40 percent of the total cost (Khan 2010). Roberts, Rose, et al. (2014) examined the influence of border crossing times on the U.S. economy. The authors' results suggested that crossing time at POEs has significant negative impacts on the economy. The value of reduced crossing time by adding one officer equals roughly $1.2 million savings for U.S. residents and $0.2 million savings for Mexico residents at each U.S.–Mexico land POE. The authors also estimated that the value of reduced crossing time at some specific ports by adding an officer is even much greater. For example, based on the authors' estimation, the value of time saved due to adding one officer at the San

Ysidro crossing is roughly $25 million (Roberts, Mcgonegal, et al. 2014). Meanwhile, lower crossing time will induce more cross-border trips.

EXISTING TECHNIQUES FOR BORDER CROSSING TIME MONITORING

For a long time, the only source of POE crossing time was provided by U.S. Customs and Border Protection (CBP), which is estimated by using a visual inspection method and random survey of truckers waiting at the bridge. However, these estimates are unreliable according to shippers and carriers (Rajbhandari and Villa 2012). Traditional crossing time monitoring is typically conducted through three main solutions (Villa et al. 2017):

- **Unaided visual observation:** The CBP officer records where the formed queue ends in relation to predetermined markers. Inspectors use their experience to estimate queue density and crossing times.
- **Camera snapshots:** Some civilian agencies have installed traffic cameras along the POEs. Camera snapshots are publicly available. CBP officers can use snapshots to estimate queue and crossing time.
- **Driver surveys:** Conducting surveys is the most commonly used method among crossing time measurement techniques. The officer working at the primary inspection asks the drivers to estimate how long they have been waiting in the queue.

As part of the effort to deploy ITS at international border crossings, a lot of new techniques were deployed to measure and monitor border-crossing times. Basically, three categories of approaches are used to measure crossing times at border crossings: queue length measurement, fixed point vehicle reidentification, and dynamic vehicle tracking (San Diego Association of Governments 2017; Sabean and Jones 2008; Rajbhandari, Villa, Macias, et al. 2012). Queue length can be estimated by using technology to measure the arrival and departure rates of vehicles and a calibrated model to estimate queue end. Fixed-point vehicle reidentification enables the calculation of time spent by a vehicle at a fixed point and between two fixed points. From these readings, crossing times can be found but only for archival purposes since such data are of little use for a real-time information system. Likewise, dynamic vehicle tracking based on the use of a wireless signal emitted by a device placed in the vehicle can generate data on time spent at a given spot or between various locations. However, the resulting information has archival value only (Khan 2010; Rajbhandari, Villa, Tate, et al. 2012; Ramezani and Geroliminis 2015).

- **Queue length measurement:** The queue length measurement approach uses humans or technology to measure the departure and arrival rates of vehicles and estimate the number of vehicles in the queue. This estimate is usually based on a measure of the length of the queue and an approximate average of the density of vehicles within it. The mainstream solutions include inductive loop detectors, ranging radar detectors, and video image processing.
- **Fixed-point vehicle reidentification:** This method uses technology to identify individual vehicles at a fixed point upstream of the queue, and then again at the primary inspection booth or at some point beyond the inspection facilities. The time difference between the two timestamps provides the travel time between the two points. The crossing time

attributed only to the queue can be calculated by subtracting out the average time required to travel that distance when there is no queue (i.e., under optimal conditions). The mainstream solutions include RFID and license plate recognition.
- **Dynamic vehicle tracking:** This approach uses some form of wireless signal to determine the location of a vehicle at multiple times along its route. The archived data can then be analyzed to determine how far the vehicle traveled between time intervals on the approach to the border. The mainstream solutions include cell phone tracking and global positioning system (GPS) tracking.

Comparing the number of arrivals and departures at land border crossings is one of the most straightforward solutions to quantify the congested queuing. Roberts, Mcgonegal, et al. 2014 applied this method to three land border crossings at El Paso, Texas. The length of the queue in terms of the number of vehicles is calculated by multiplying the crossing time in that hour by the number of departures per minute. RFID is one of the most commonly used technologies at POEs. Step-by-step guidance for implanting RFID to measure border crossing times is provided by Rajbhandari, Villa, Macias, et al. (2012) and Rajbhandari, Villa, Tate, et al. (2012).

RFID consists of three subsystems:

- **Field subsystem:** The field subsystem contains the transponder detection stations including the RFID readers, antennas, communication equipment, and power supply. The station identifies transponders carried by commercial vehicles and transmits the data to the central subsystem via wireless communication.
- **Central subsystem:** The central subsystem receives identification of transponders from field stations and performs all data processing to reidentify, filter, and archive the crossing and crossing time data.
- **User subsystem.** The user subsystem interacts with the central subsystem to provide access to users' archived data as well as real-time data via the internet (Rajbhandari and Villa 2012).

A similar system was deployed by the Niagara International Transportation Technology Coalition, which uses Bluetooth identification technology to provide more accurate delay estimates to motorists; the information is now updated every five minutes (Lin et al. 2014). McCord et al. (2016) developed and implemented an approach to document truck activity times associated with crossing POEs by using technologies that are already in use by truck fleets. The authors created onboard GPS-enabled data units containing positioning, navigation, and timing systems to track truck movements. These data were transmitted and spatially filtered using geofences to analyze truck crossing times at border crossings.

EMERGING DATA SOURCES FOR BORDER CROSSING TIME ESTIMATION

The basic ingredient for the new wave of smart cities that has emerged during the last decade are massive data sets concerning human mobility, fostered by the widespread distribution of sensors, such as GPS devices in many modes of transport, smart phones, and traffic fixed sensors (Ramezani and Geroliminis 2015).

For example, GPS-equipped floating cars can collect and transmit mobility data periodically (usually between every 15 second to 1 min), including longitude, latitude, speed, vehicle headings, and timestamps. Massive amounts of mobility data have been accumulated which could be exploited to characterize typical traffic conditions. Within the last decade, a broad range of transportation research has been conducted based on floating car mobility data. Research related to GPS-equipped vehicle mobility data is a hot topic in the field of ITS (An et al. 2016). Meanwhile, GPS-enabled devices (e.g., smart phones) or wireless magnetic sensors also can be used to identify vehicle movements providing a great potential for probe vehicles in ITS applications (Ramezani and Geroliminis 2015). A study—Travel Time Estimation Using Cell Phones (TTECP) for Highways and Roadways conducted by Florida Department of Transportation has verified that mobile phones and other mobile devices and their respective location data as viable sources for travel time; and the reliability, accuracy, and resolution of this data, continues to improve as smartphone manufacturers refine or adopt more capable components (The San Diego Association of Governments 2017).

Meanwhile, various wireless technologies (e.g., Bluetooth, Wi-Fi) have been intensively utilized in border crossing time estimation. Bluetooth wireless technology is a short-range communications technology originally intended to replace the cables connecting portable and/or fixed communications devices while maintaining high levels of security. Bluetooth technology is included commonly on devices such as smartphones, hands-free kits in vehicles, personal computers, wireless headsets, and other devices. The key features of Bluetooth technology are robustness, low power, and low cost. Bluetooth is a mature technology that has been in use for about 20 years (U.S. Customs and Border Protection 2016; The San Diego Association of Governments 2017). Like Bluetooth technology, Wi-Fi is another short-range communications technology, which has been utilized in border crossing time systems. Wi-Fi was initially designed to provide communications among devices while maintaining high levels of privacy. Wi-Fi technology is most often included commonly on modern devices such as smartphones, hands-free kits in cars, personal computers, other media streaming devices (The San Diego Association of Governments 2017). The research work conducted by U.S. Federal Highway Administration (FHWA) also recommend that the Bluetooth solution and the GPS/smartphone technology—showed the most promise as viable near-term approaches to automated crossing time measurement (Sabean, Lussier, and Pattan 2011).

CBP is moving forward with developing a hybrid data-driven solution (i.e., no hardware deployment required) to estimate border crossing times (U.S. Customs and Border Protection 2016). This hybrid data solution uses travel data generated from the public sector and CBP's vehicle throughput data to provide automated land border crossing time measurement. CBP believes this hybrid solution is developing to be the most viable and cost-effective solution to date because no port infrastructure or hardware is required.

Crowdsourced traffic data are also a promising option in transportation monitoring and management. Generally, crowdsourcing leverages the combined intelligence, knowledge data, or experience of a group of people (or their devices) to answer a question, solve a problem, or manage a process. With the maturation and advancement of mobile sensing techniques and social media applications, smartphones have become a promising platform for collecting and sharing traffic information. For travel time data collection, crowdsourced methods are the most used private-sector mechanism today. Mobile devices carried by drivers or installed in their vehicles can provide detailed information about their location, speed, and possibly additional information to a public or private entity, and that information is used to generate traffic/travel time information (San Diego Association of Governments 2017). For example, Waze, as a leading crowdsourcing mobile app, has millions of registered active users sharing their real-time traffic and road information, which can be potentially used to monitor traffic condition (Li et al. 2020). Meanwhile, different mobile apps, such as Toronto Buffalo Border Wait Time (TBBW), Border Wait Times, and Best Time to Cross Border, were also developed to collect border wait times from their users (Lin et al. 2015). Many third-party companies provide high-quality traffic data by aggregating crowdsourced traffic information, including but not limited to:

- INRIX™ (http://www.inrix.com/).
- HERE™ (http://here.com/).
- Cellint™ (http://www.cellint.com/).
- Telenav™ (http://www.telenav.com/).
- TrafficCast™ (http://trafficcast.com/).
- TomTom™ (http://www.tomtom.com/).
- Cuebiq™ (https://www.cuebiq.com/).
- Waze™ (https://www.waze.com).

Driving data collected from CVs are another emerging data source. CVs include short-range radio communications technologies for vehicle-to-vehicle communications (V2V), where vehicles on the roadway communicate with one another, and vehicle-to-infrastructure communications (V2I), where vehicles on the roadway communicate with roadside technologies and devices. The Texas A&M Transportation Institute (TTI) conducted a SWOT analysis to compare different emerging data sources in crossing time estimation; the result indicated that CV technology is the one that provides the highest value for the enhanced border crossing time measurement system (Villa et al. 2017). Although CV-related technologies are still in early stages of development, they have been maturing, prototyped, and tested for a couple of decades and will be more promising (San Diego Association of Governments 2017).

POTENTIAL SOLUTION FOR BORDER CROSSING TIME ESTIMATION

Vehicle mobility data are a valued data source for evaluating traffic congestion. An et al. (2016) used GPS-equipped vehicle mobility data to measure urban recurrent congestion evolution patterns. The authors divided the study area into an equal-size grid and summarized the characteristics of trajectories within each grid in terms of the number of trajectories and mean velocity of all trajectories. The authors calculated the Mahalanobis distance to measure the variations of trajectory characteristics for a specific grid over a series of timestamps to capture urban recurrent congestion patterns (one month, 16,000 taxi vehicles, and a sampling rate of 30 seconds). Iyer et al. (2018) proposed a deep learning-based congestion prediction model based on bus mobility trace data. The study calculated the speed of buses on each segment based on their GPS traces data. Based on the historical trends of speed at a given segment and its neighboring segments, a long- and short-term memory recurrent neural network model was built to predict travel speed in each segment.

Some estimation models were built based on real-time traffic volume and POE inspection time data. For example, Lin et al. (2014) developed two classes of multi-server models to predict border-crossing delay with vehicle arrival times and inspection times directly collected from a POE—Peace Bridge. Ramezani and Geroliminis (2015) proposed a method based on the position and instantaneous speed of probe vehicles to estimate the queue profile at intersections. Probe vehicles provide samples of their individual traffic state and aim to leverage the collective information of temporally and spatially dispersed probe data. The authors first used the velocity attribute of all probe data to classify them into two groups, including moving vehicles and stopped vehicles. Next, a clustering method was used to cluster the stopped vehicles into different groups representing cycles. Then, the authors associated the moving vehicles to cycles. Lastly, the authors calculated the discharging line and estimated the queue formation for each cycle to determine the front and the back of queue. The method was tested on the Next Generation Simulation dataset and simulated data. Khan (2010) proposed a cost-effective method for estimating the extent of queuing and delay at POEs based on microsimulation using VISSIM. A hybrid of sensor data was simulated to represent the accumulated number of arrivals and departures, which was then fed to an artificial neural network model to predict the queues and delays at POEs. Results showed that microsimulation and the artificial neural network modeling worked well for simulating border crossing time.

Signal strength detection is also a potential solution to monitor traffic flow. For example, Miyagawa and Ogawa (2017) proposed a new system by using Bluetooth low-energy (BLE) broadcasting transmitted from a mobile application. In this system, the smartphone was used as a broadcaster using BLE, and two receivers were installed to record time and the received signal strength indicator of BLE signals for estimating crossing time. Crowdsourced driving data collected from mobile devices (e.g., smartphones) are also a valued data source in border crossing time estimation. For example, Lin et al. (2015) created a mobile app, TBBW. This app can not only show the current waiting time lagged from a POE (updated every hour) but can also collect user-reported waiting time by app users in real time. By combining the official data and crowdsourced data, TBBW can provide a better estimation of border-crossing delays at a finer temporal scale.

With the advancement of big data analysis techniques, more studies are conducted to explore how to use big data analytics to cope with the emerging big datasets to provide a better estimation of crossing times. For example, Sankaranarayanan et al. (2016) used big data visualization approaches to investigate the patterns of airport crossing times, which can be potentially applied to visualize and extract the pattern of big mobility data crossing POEs.

CHAPTER 3:
DATA COLLECTION AND PROCESSING

This chapter describes the study site and research data analyzed in this study. We chose the PDN International Bridge as our testing site. We collected both the existing border-crossing information from the BCIS through Bluetooth and the CV data from Wejo for passenger vehicles. This study evaluated CV data collected from 3 consecutive years: from 2020 to 2022. In total 8 months of data, including November 2020, December 2020, November 2021, December 2021, September 2022 and October 2022, were evaluated in this study.

STUDY SITE—PASO DEL NORTE INTERNATIONAL BRIDGE

The PDN International Bridge is among the busiest POEs in the United States. Figure 1 illustrates the aerial view of the PDN POE, which is located at 1000 S. El Paso Street on the U.S. side and on Juárez Avenue on the Mexican side. This POE operates 24 hours a day, seven days a week, and is dedicated to pedestrians (northbound and southbound) and passenger vehicles (northbound), with more than 10 million people traveling from Mexico to the United States each year at this location. Since the Wejo CV data examined in this project are specifically collected from the passenger vehicles, the PDN POE is a perfect study site to evaluate the effectiveness of Wejo CV data in border travel time estimation.

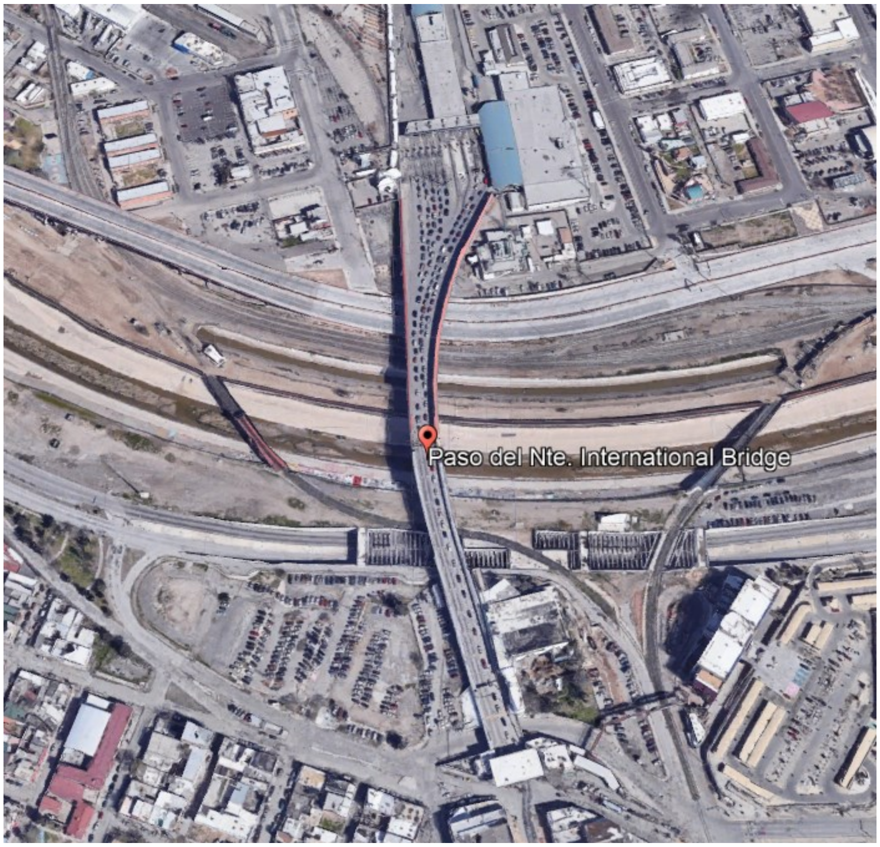


Figure 1. Paso del Norte POE Aerial View.

BORDER CROSSING INFORMATION SYSTEM

The current BCIS in El Paso was established by TTI with funds from TxDOT and FHWA. The BCIS was designed to measure travel times of commercial and passenger vehicles crossing the Texas–Mexico border. The BCIS uses RFID readers at commercial POEs and Bluetooth/Wi-Fi sensors at POEs serving passenger vehicles. The Bluetooth/Wi-Fi sensors will regularly capture mobile devices' signals (Bluetooth or Wi-Fi media access control [MAC] address) carried by the vehicles, drivers, or passengers at multiple locations along the path for each border-crossing trip. The captured MAC address will be transmitted to the BCIS server along with a captured timestamp and sensors' IDs. By matching the Bluetooth MAC address at different locations (readers), the BCIS can calculate crossing time based on the captured timestamps difference.

As illustrated in Figure 2, five Bluetooth readers are installed at the key locations along the PDN POE named PDN-BT 1–5:

- **PDN-BT1:** the starting point to observe the queue, near the intersection of Juárez Avenue and 16 de Septiembre Street.
- **PDN-BT2:** the intermediate point to observe the queue, near the intersection of Juárez Avenue and Ignacio Mejia Street.
- **PDN-BT3:** the exit of the Mexican toll booth.
- **PDN-BT4:** the international bridge.
- **PDN-BT5:** the U.S. CBP inspection facility.

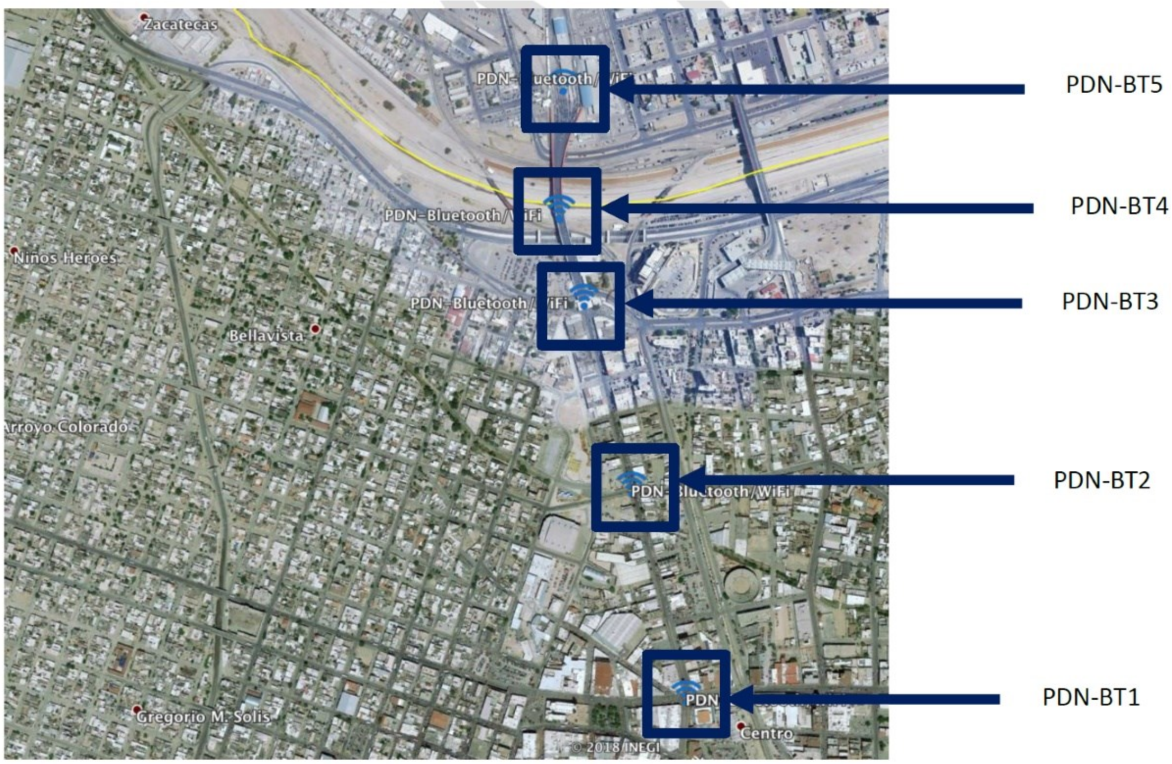


Figure 2. Installation of Bluetooth Readers along the PDN POE.

CONNECTED VEHICLE DATA (WEJO)

CVs are rapidly becoming the new paradigm of road transport, which has been widely believed to positively influence transportation safety, efficiency, and sustainability. CVs represent the unification of various connectivity technologies, enabling vehicles to communicate with other vehicles (V2V), transportation infrastructures (V2I), and the cloud (V2C) for achieving the goal of self-driving (Hoseinzadeh et al. 2020, Talebpour and Mahmassani 2016). Although most commercially available vehicles are still far from completely automating the driving task, most of them already could monitor the driving environment and vehicle movements through vehicular sensors. Many world-leading auto manufacturers, like Toyota™, General Motors™, BMW™, and Tesla™ among others, have ramped up the production of CVs, which could access and transmit vehicular sensors' data to the cloud (Miles 2019). Many automotive data companies have also emerged to facilitate the use of CV data. Like Wejo, Otonomo, Smartcar, Vinili, and CarAlgo, these data companies bridge the data providers—auto manufacturers—with data users by ingesting, aggregating, and normalizing the raw CV data and delivering the enriched and organized datasets to end users (Miles 2019).

The CV data are collected from vehicles, directly reflecting the dynamics of traffic mobility. For example, Wejo, as a leading CV data start-up, provides high sampling rates and multi-dimensional vehicle movements and driving event (e.g., hard braking, hard acceleration, and speeding) data. This data platform has currently partnered with multiple world-leading auto manufacturers and collected data from millions of vehicles with a sampling rate of three seconds per waypoint. Each waypoint describes the timestamp, location, and movement-related information (e.g., speed and heading) of a vehicle's trajectory. Wejo claims that its CV data products could access over 90 different vehicular sensors and cover 95 percent of road networks in the United States, with about 12 billion data points collected every day at its best temporal resolution of every three seconds. Our preliminary studies in Texas also demonstrated that Wejo data have good spatiotemporal coverage in both urban and rural regions of Texas. CV data show great superiority in data quality, volume, consistency, and richness compared to traditional mobility data sources, making them a promising data source for monitoring urban mobility dynamics.

In this project, we used the CV movement data provided by Wejo to estimate the travel time at the PDN POE. Each record in the movement data details the vehicle's status in a trip (e.g., speed, heading, and location) with a specific timestamp. The CV data were reprocessed by Wejo and delivered to the Azure Cloud Storage account. The data were organized in the Apache Parquet format. We used an online big data analytics platform, Azure Databricks™, to process the big CV dataset. Azure Databricks supports the latest version of Apache Spark™, allowing its users to seamlessly integrate with any open-source libraries and quickly establish a fully managed Apache Spark environment. We primarily used Apache Sedona™ to load, partition, process, and spatially analyze the big CV data and used other open-source libraries (e.g., Datashader) to visualize the large dataset.

The original Wejo CV data were pre-partitioned by geohash—a hierarchical geocode system dividing space into buckets of grid shape. In this study, we subset the Wejo data by using the geohashes, which covered the entire PDN POE. Figure 3 illustrates the geohash-partitioned CV data collected in October 2022 for the PDN POE, which were visualized through the Datashader.

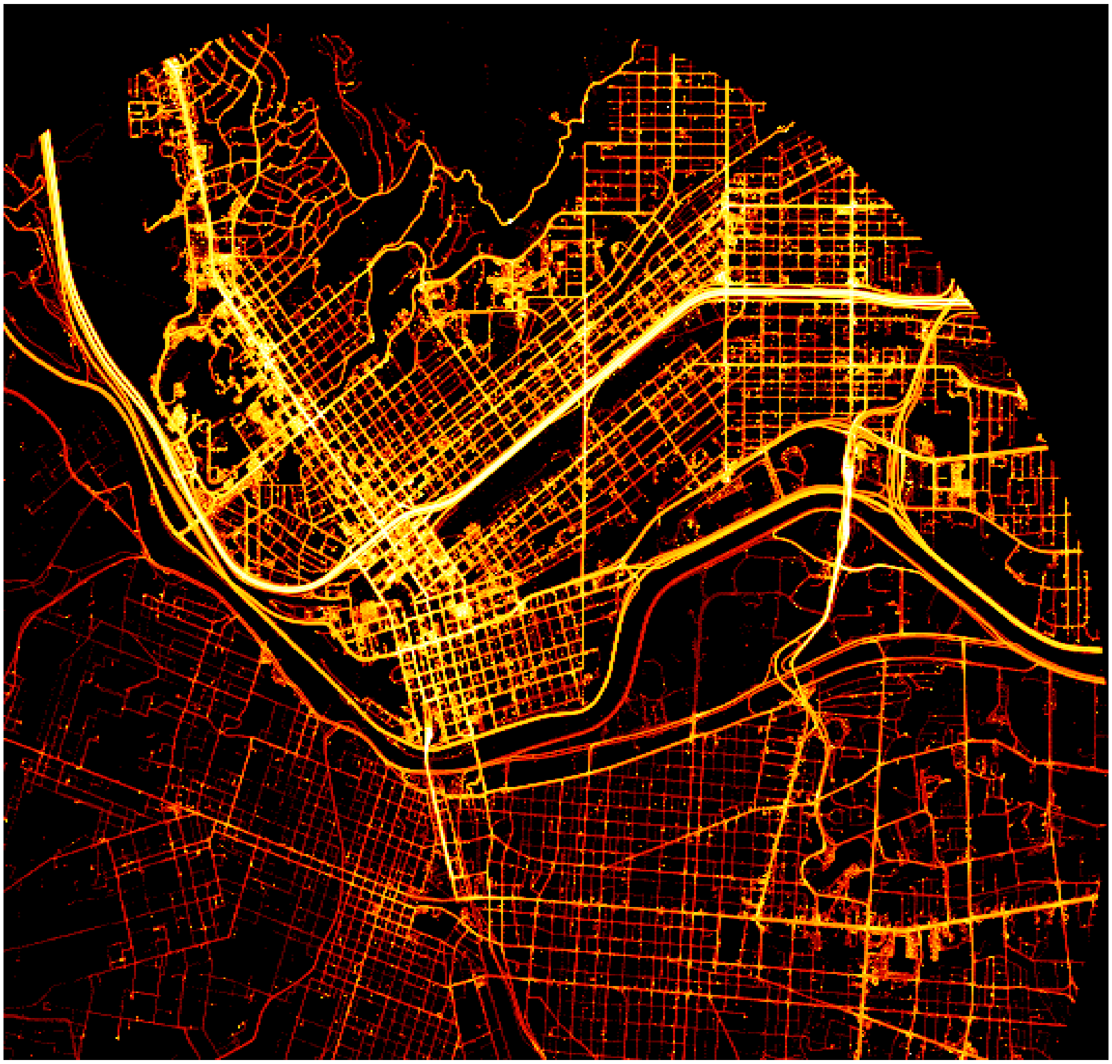


Figure 3. Illustration of Wejo CV Data at the PDN POE.

BORDER CROSSING TIME CALCULATION BASED ON BLUETOOTH OBSERVATIONS

The collected Bluetooth observations from the BCIS can be used to measure the passenger vehicle crossing time. Crossing time is defined as the time it takes, in minutes, for a vehicle to reach the U.S. CBP's primary inspection booth after arriving at the end of the queue (Sharma et al. 2018, Texas A&M Transportation Institute, 2021). This queue length is variable and depends on traffic volumes and processing times at each of the inspection facilities throughout the border-crossing process. The variable nature of the queue makes the installation location of the sensors critical. Since the queue of vehicles at the PDN POE usually can reach as far as PDN-BT1, as illustrated in Figure 1, this study defines the border crossing time as the time a vehicle takes, in minutes, to travel from PDN-BT1 to PDN-BT5.

As mentioned previously, five Bluetooth readers are installed at key locations along the PDN POE. For a single trip, the vehicle can be identified at multiple locations. By matching the captured Bluetooth MAC addresses at different pairs of Bluetooth readers, we could obtain the vehicle's travel times crossing different monitored segments, as illustrated in Figure 4. In this study, we aimed to calculate the hourly average border crossing times at the PDN POE. To achieve this aim, we first needed to figure out how to sum the hourly average travel times on different monitored segments to obtain the aggregated border crossing time at one-hour intervals from PDN-BT1 to PDN-BT5.

As illustrated in Figure 4, there were eight different combinations of segments (calculation options) we could choose to form a non-overlapping corridor reaching from PDN-BT1 to PDN-BT5. The longer segment usually requires more time to travel through it, leading to fewer vehicles' travel times captured on this segment. In this study, we assumed:

1. The shorter segment could capture more trips.
2. The calculation option composed of more segments could capture more trips and lead to a more reliable hourly average travel time.
3. The closer segments to the PDN-BT5 are more important than the farther segments in border crossing time calculation, which are more necessary to be kept because short segments ensuring more trips can be captured on them.

Based on these three assumptions, we ranked the priority of these calculation options. For each one-hour interval, we started with Option 1 by summing up the hourly average travel times on these four short segments. If any of the segments in Option 1 did not capture enough trips, we calculated the border crossing time with Option 2, and so on. If all these calculation options had data missing on their segments, we marked that one-hour interval as free flow with the border crossing time as zero.

BORDER CROSSING TIME CALCULATION BASED ON WEJO CV DATA

To evaluate the effectiveness of Wejo CV data in border crossing time monitoring, we used a similar procedure to the Bluetooth data processing to calculate the hourly average border crossing time from Wejo CV data.

Since the typical detection range of Bluetooth readers is around 100 meters (Honeywell 2020), we first created 100-meter buffers around each Bluetooth reader location to extract the Wejo CV samples within each buffer. Figure 5 shows the created Wejo extraction buffers around each Bluetooth reader for the PDN POE. The CV data were collected from each trip with a high-frequency sampling rate—up to three seconds per sample. Therefore, there could be multiple waypoints captured from the same trip within each buffer. To eliminate the data redundancy, we grouped the CV waypoints within each buffer based on their Journey ID—the unique identifier for every single trip. Then we kept the earliest waypoint for each trip within the first four buffers (PDN-BT1 through PDN-BT4) and kept the latest waypoint for each trip within the last buffer (PDN-BT5). After obtaining the unique trip IDs captured in each buffer, we applied the same procedure as illustrated in Figure 4 to calculate the hourly average travel times based on the CV data. The obtained results were evaluated and compared with the Bluetooth-based border travel times, which are introduced in Chapter 4.

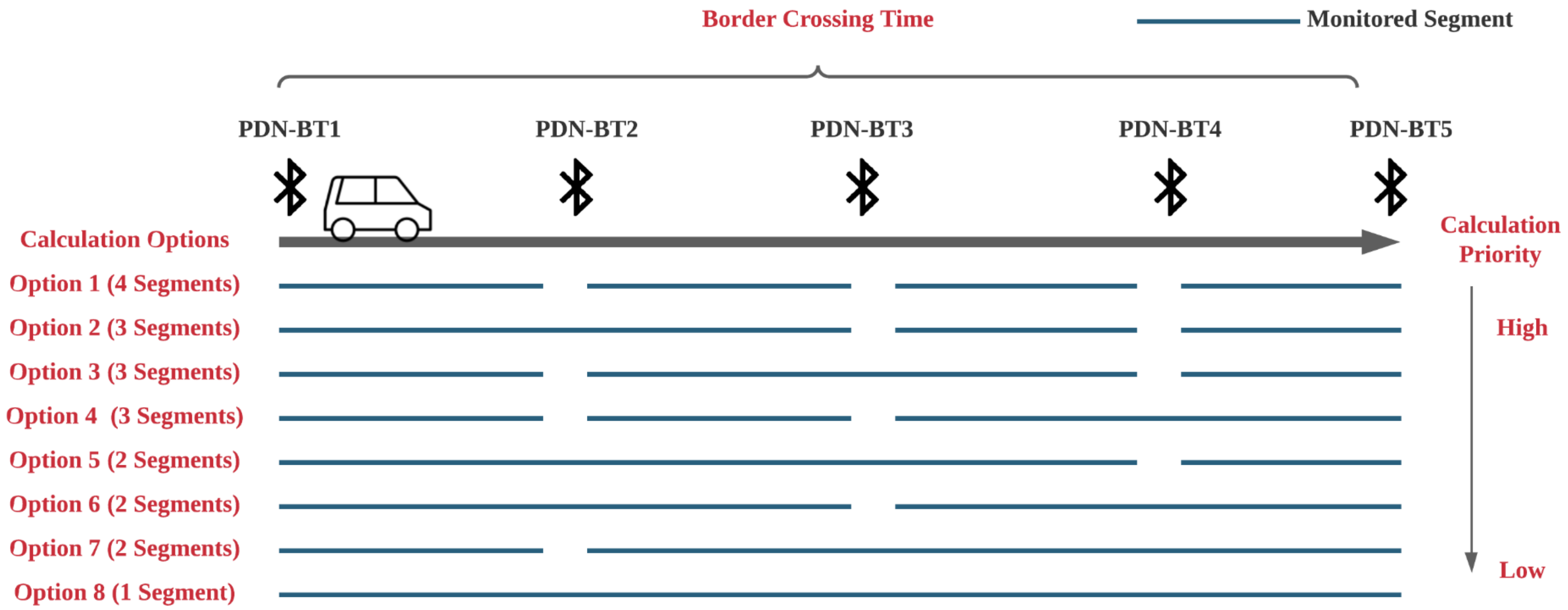


Figure 4. Illustration of Border Crossing Time Calculation.

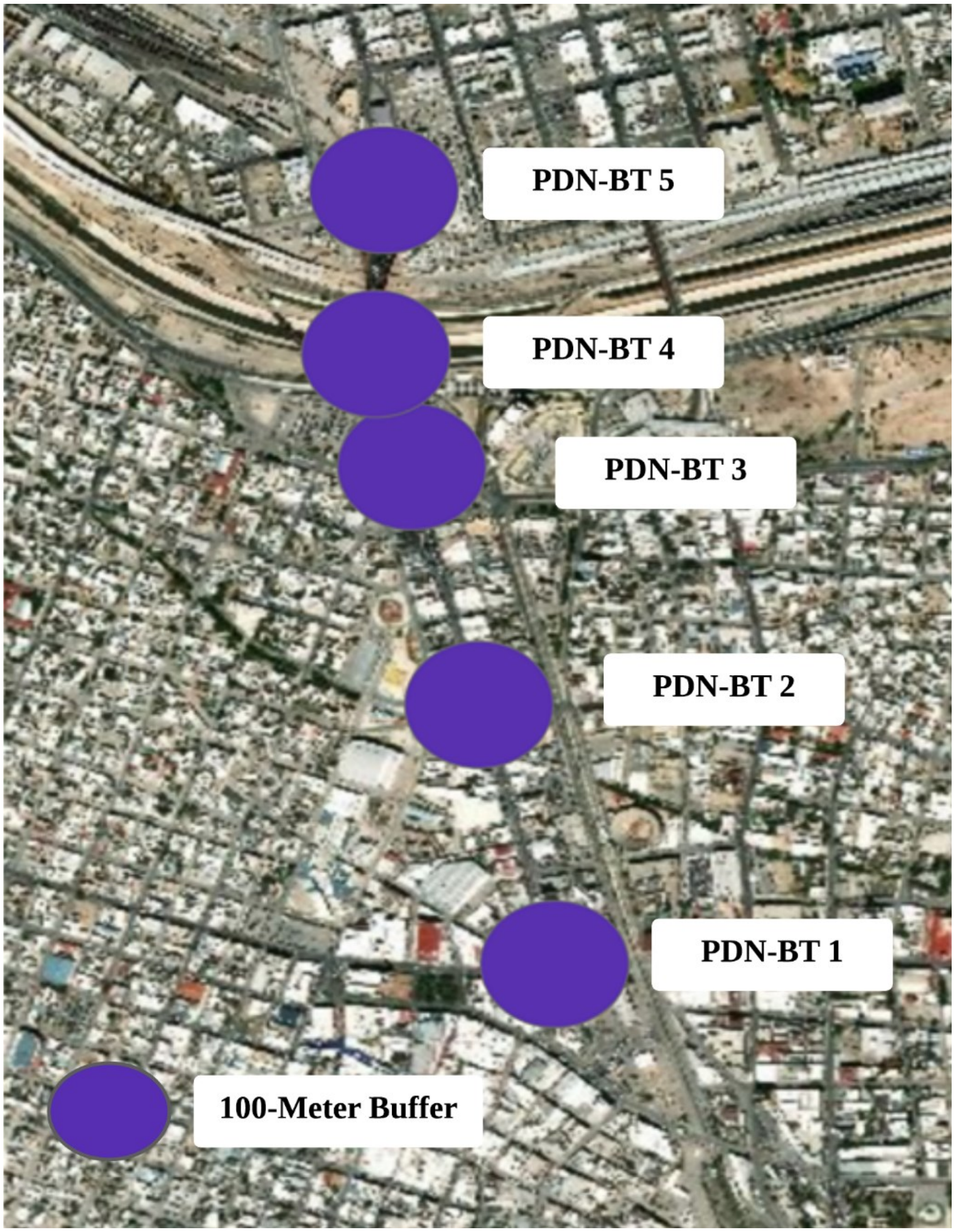


Figure 5. Wejo Data Extraction Buffers around Each Bluetooth Reader at the PDN POE.

CHAPTER 4:
ASSESSMENTS OF CV-BASED BORDER CROSSING TIMES

In this study, we calculated the hourly average border crossing times using CV data for each month and compared them with the Bluetooth-Time. For example, **Error! Reference source not found.** illustrate the calculated Bluetooth-Time and CV-Time for September 2022, which indicates that the Bluetooth-Time (red line) and CV-Time (blue line) appear to have similar results for the estimated border crossing times. To better assess the effectiveness of CV data, we further conducted two assessments:

- **Temporal coverage assessment:** This assessment aimed to investigate the percentage of one-hour slots in which we could obtain enough CV samples to estimate border crossing times.
- **Similarity assessment:** This assessment aimed to assess the statistical similarity between the Bluetooth-Time and CV-Time. We used three measures—the correlation coefficient, the root mean square error (RMSE), and the mean absolute percentage error (MAPE) to quantify the similarity between the Bluetooth-Time and CV-Time.

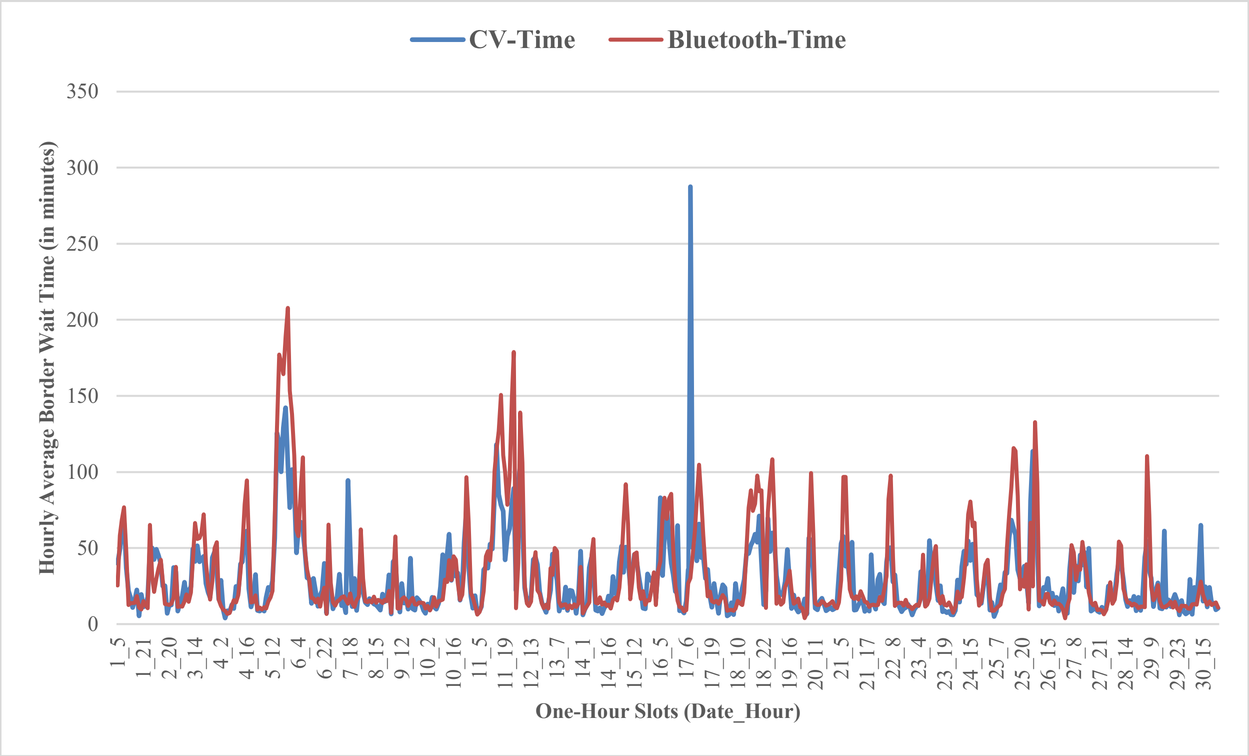


Figure 6. Comparsion between Bluetooth-Time and CV-Time in September 2022.

TEMPORAL COVERAGE ASSESSMENT

To evaluate the practical value of CV data in real-world border-crossing monitoring, we first need to figure out whether we could obtain sufficient CV data samples from the CV data provider, Wejo, to generate continuous monitoring results at the POE. In this study, we generated hourly averaged border travel times based on the CV data. We sampled the CV data from 3 consecutive years: 2020 to 2022. In total 8 months of data including November 2020, December 2020, November 2021, December 2021, September 2022 and October 2022 were available for this study. As introduced in Chapter 3, eight calculation options could be used to generate border crossing times. For each one-hour slot, if the border crossing time was generated through the calculation procedure introduced in Chapter 3, we marked this slot as a CV-covered slot; otherwise, it was marked as a CV-uncovered slot. We conducted the temporal coverage assessment for all eight months to calculate the percentage of the uncovered slots for each month.

Figure 7 illustrates the mean monthly coverage rate for each year. It clearly shows that although the average coverage rates of the data in 2020 and 2021 are low, an exponential growth was observed in 2022 in terms of the data coverage, implying the growing potential of CV data in real-world applications.

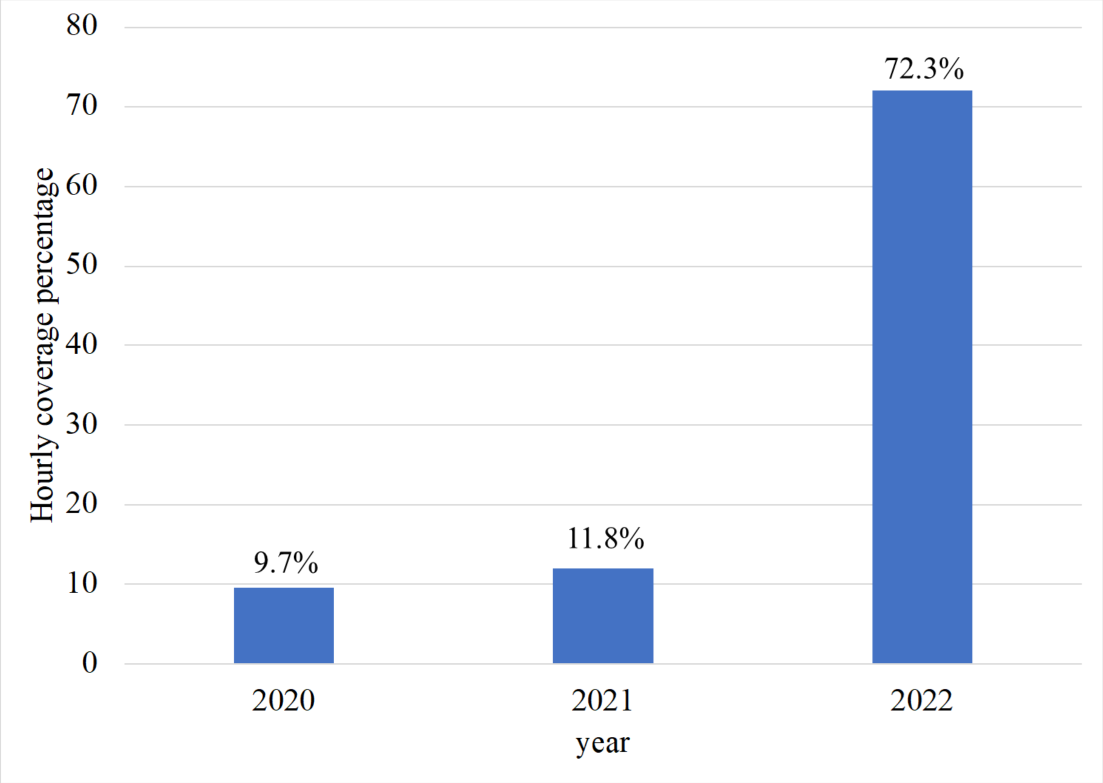


Figure 7. Monthly Average of CV-covered Slots Percentages for 2020 to 2022.

In this study, only the data for 2022 was further investigated due to its high coverage percentage. Figure 8 illustrates the results of the temporal coverage assessment for the September and October datasets 2022. The orange color is assigned to the cells with sufficient CV observations, and the white color in these figures indicates no sufficient CV observations to support any one of the eight calculation options to generate an estimation result for that one-hour slot, which is regarded as a CV-uncovered slot.

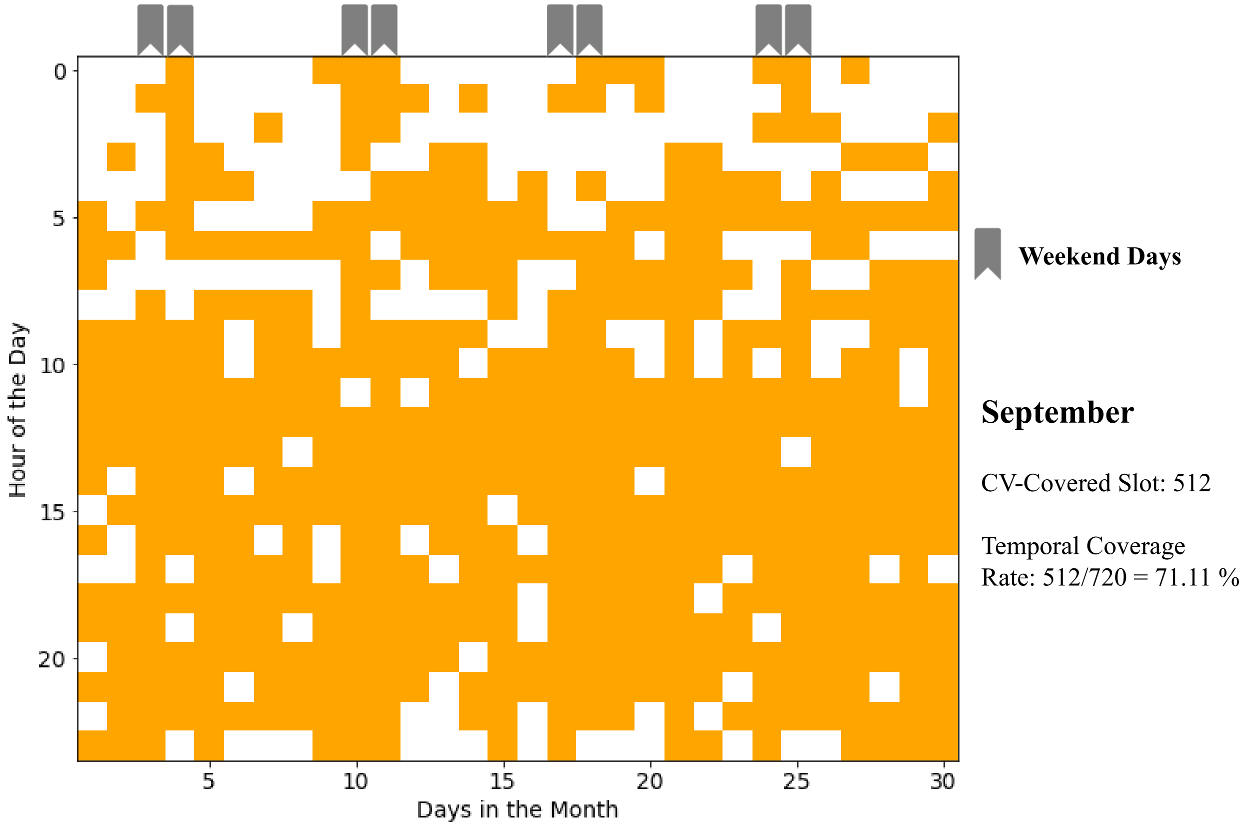
As illustrated in Figure 8, 512 of 720 one-hour slots (71.11 percent) in September 2022 were marked as CV-covered slots, which means we could generate border crossing times based on the

CV trips captured by Wejo in these slots. The October dataset showed a slightly broader temporal coverage, with 547 of 744 slots (73.52 percent) having sufficient CV trips to estimate border crossing time (see Figure 9).

Given the penetration rate of Wejo CVs in 2022, this evaluation showed roughly 70 to 75 percent of the one-hour slots with enough Wejo samples to estimate border travel times. The results also indicated that the Wejo CV data usually achieved a higher temporal coverage rate on weekend days (September had 79.69 percent, and October had 79.59 percent), up to 10 percent higher than the monthly average temporal coverage rates. We can also visually identify that the midnight and early morning hours (midnight–9 a.m.) have more CV-uncovered slots, which could be caused by the low volume of border-crossing passenger vehicles at these hours.

The reasons limited the temporal coverage of Wejo in the PDN POE could be twofold:

- Given the data usage agreements between the data provider (Wejo) and its partnered original equipment manufacturers (OEMs), only one OEM currently agreed to share their data for this pilot study. The actual existing CV data is much more than the amount we tested but not available to use due to the license issue.
- Access to the Mexico-registered vehicles' data was limited in the 2022 Wejo dataset. All the border-crossing trips used in this study were collected from U.S.-registered passenger vehicles.

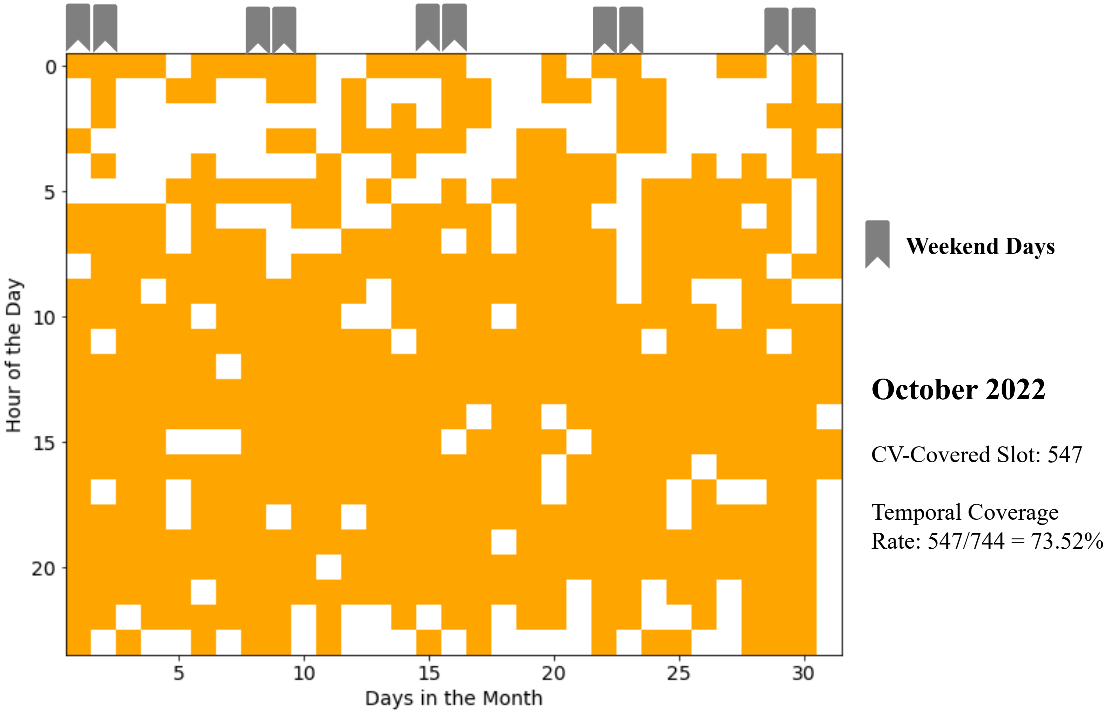


September

CV-Covered Slot: 512

Temporal Coverage Rate: 512/720 = 71.11 %

Figure 8. Temporal Coverage Assessment Results of the CV-Time in September 2022.



October 2022

CV-Covered Slot: 547

Temporal Coverage Rate: 547/744 = 73.52%

Figure 9. Temporal Coverage Assessment Results of the CV-Time in October 2022.

SIMILARITY ASSESSMENT

Besides assessing the temporal coverage of CV data at the border crossing, we also need to assess how accurate the CV-Time can represent the hourly averaged border crossing times. Since the Bluetooth observations collected from the BCIS were the only available data source to monitor border-crossing trips at the PDN POE, we evaluated the similarity between the Bluetooth-Time and CV-Time in September and October 2022 to assess the practical value of using CV data in border crossing time monitoring.

In this study, we first selected the one-hour slots with both Bluetooth-Time and CV-Time. Since the Bluetooth data resolution is 15 minutes, the hourly average for the Bluetooth-Time calculated as the average of 4 reading for each quarter in each hour. Then, we calculated the correlation coefficient, the RMSE, and the MAPE between the Bluetooth-Time and CV-Time to quantify their similarities and differences each month. The correlation coefficient is one of the most used measures to assess the strength of the relationship between two groups' values. The RMSE and MAPE are used to evaluate the difference of two groups' values, which are commonly used to assess the quality of predictions. The following equations illustrate how to calculate these measures based on two data groups with the same sample size:

$$Correlation\ Coefficient = \frac{\sum_{t=1}^{n}(A_t - \bar{A})(F_t - \bar{F})}{\sqrt{\sum_{t=1}^{n}(A_t - \bar{A})^2 \sum_{t=1}^{n}(F_t - \bar{F})^2}}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Where A and F represent two sample groups. Each of them contains n samples; A_t and F_t represent the $t - th$ sample in each group; and $\bar{A}$ and $\bar{F}$ represent the mean values of these two groups. Table 1 and

Table 2 summarize the similarity assessment results of the CV-Time for September and October 2022 datasets, respectively. Table 1 shows that when applying the base value of the calculation threshold (which means that as long as there is one trip captured on the segment, we could generate hourly travel times from it), the correlation between the Bluetooth-Time and CV-Time is 0.73, the RMSE is around 22.37 minutes, and the MAPE is around 43 percent in September 2022.

Table 1. Similarity Assessment Result of the CV-Times in September 2022.

Measures	***Values***
Bluetooth-covered one-hour slots	720
CV-covered one-hour slots	512
Correlation	0.73
RMSE (in minutes)	22.37
MAPE	43.58%

Table 2. Similarity Assessment Result of the CV-Times in October 2022.

Measures	***Values***
Bluetooth-covered one-hour slots	744
CV-covered one-hour slots	547
Correlation	0.81
RMSE (in minutes)	21.16
MAPE	39.77%

Table 2 shows the similarity assessment result of the October CV-Times. The CV-Time and the Bluetooth-Time in October show a slightly higher correlation (0.81 versus 0.73), a lower RMSE (21.16 versus 22.37), and a lower MAPE (39.77 percent versus 43.58 percent) when applying the base threshold.

The similarity assessment results indicated that the CV data are a valuable data source that is capable of producing similar estimations for border crossing times as the BCIS-generated results. We achieved a high correlation between the CV-Time and Bluetooth-Time in both September and October 2022 with a value of 0.73 and 0.81.

CHAPTER 5: BORDER CROSSING TIME MODELING USING CV DATA

Chapter 4 demonstrates that the CV data are a promising data source. The CV-Time is highly correlated with the Bluetooth-Time, with a correlation value of approximately 0.8. As previously discussed, the Bluetooth-based BCIS is costly in terms of installation and maintenance. Therefore, there is a practical necessity to explore further whether it is possible to accurately model the Bluetooth-Time based on the CV data.

This chapter introduces the modeling efforts to improve border crossing time estimation based on the CV data. We used the hourly averaged Bluetooth-Time as the response variable, and the CV-Time and other features as predicting variables. In this study, we used 70% of the data in each month as the training dataset to calibrate the model and tested the model on the rest 30% of the dataset. A total of 717 one-hour slots in the September and October 2022 dataset, containing both Bluetooth-Time and CV-Time, were used to train the model. A total of 336 slots in the dataset were used for testing and evaluating the model. This process could further evaluate the effectiveness of the proposed model. More importantly, it could demonstrate the capability of the pre-calibrated model for estimating future border crossing times.

In this study, we first built different ordinary least squares (OLS) linear regression models to estimate the Bluetooth-Time based on the CV data. Three measures, including R squared (R^2), RMSE, and MAPE were used to assess the performance of the models. R^2 is one of the most used statistics in model performance evaluation. R^2 represents the proportion of the variations in the response variable that is predicted from the predicting variable(s), which can be calculated through the following equation:

$$R^2 = 1 - \frac{\mathit{The\ sum\ of\ squares\ of\ residuals}}{\text{The total sum of squares (proportional to the variance of the data)}}$$

BASIC AND TRANSFORMED OLS MODELS BASED ON CV-TIME ONLY

In this study, we first built a basic OLS linear regression model (Simple OLS) using the CV-Time alone as the predicting variable and the Bluetooth-Time as the response variable. **Error! Reference source not found.** is a chart correlation plot of the CV-Time and the Bluetooth-Time in the training dataset (September and October 2022 dataset). The chart correlation plot visualizes the distribution of each variable on the diagonal, the bivariate scatter plots with a fitted line on the bottom of the diagonal, and the value of the correlation plus the significance level as stars on the top of the diagonal. Each significance level is associated with a p-value (0.001, 0.01, and 0.05) and a symbol (***, **, and *).

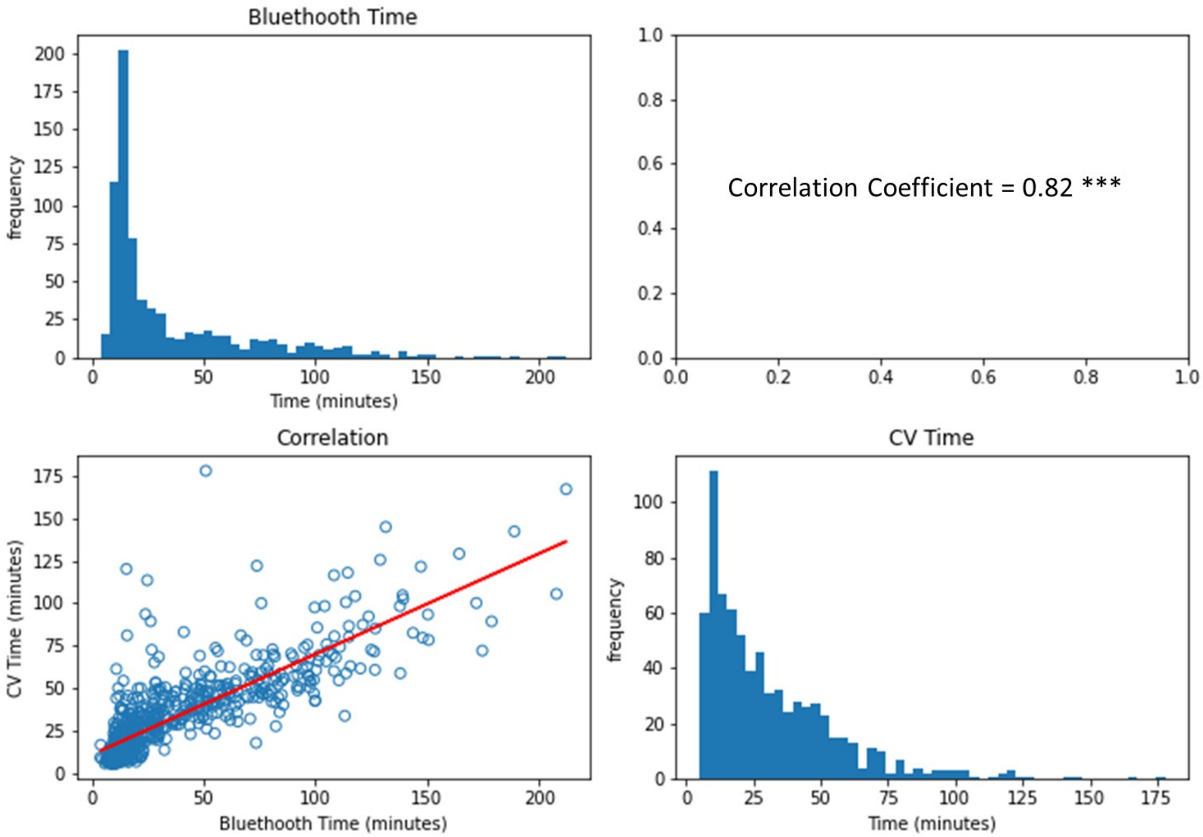


Figure 10. Chart Correlation Plot between the Bluetooth-Time and the CV-Time in the Training Dataset in September and October 2022.

The Bluetooth-Time and the CV-Time are significantly correlated with a high correlation coefficient (0.82) and follow a linear relationship, as illustrated in Figure 10. Therefore, we built an OLS model to represent the linear relationship between the Bluetooth-Time and the CV-Time (Simple OLS). The modeling result and the fitted functional form of Model 1 are detailed in Table 3. The result indicates that the CV-Time is a significant predictor for estimating the Bluetooth-Time. With the CV-Time as the sole predicting factor, 67 percent of the variance in the Bluetooth-Time can be explained by the fitted linear model.

Table 3. Simple OLS Result Based on the CV-Time Only.

Model Parameters	**Estimate (STD)**
Intercept	-0.44 (1.18)
CV-Time	1.13 (0.03) ***
R^2	0.67
Fitted functional form	$Bluetooth - Time = -0.44 + 1.13 * CV - Time$

The training dataset contains 717 observations; the testing dataset contains 336 observations.
STD stands for the standard errors, which are included in the parenthesis.
*** represents significant 99 percent level based on p-values.

Typically, all the variables in a linear regression model need to be normally distributed. However, the predicting and response variables in Model 1 are obviously right skewed, as illustrated in **Error! Reference source not found.**. Therefore, we performed a square root transformation on both variables to normalize them. The chart correlation plot of the transformed Bluetooth-Time (Bluetooth_Trans) and transformed CV-Time (CV_Trans) is shown in Figure 11. By performing the square root transformation, both variables in the training dataset can better follow the normal distribution, and the correlation coefficient increased from 0.82 to 0.84.

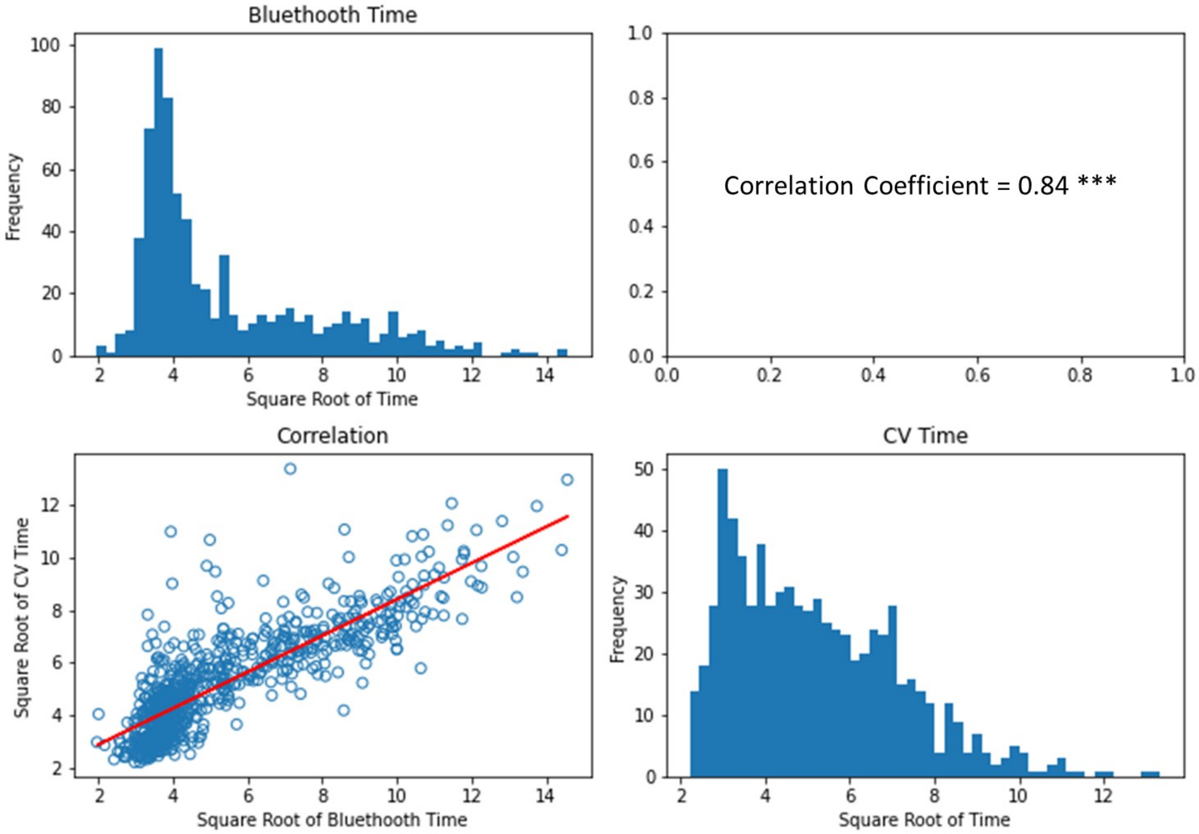


Figure 11. Chart Correlation Plot between the Square Root Transformed Bluetooth-Time and the CV-Time in the Training Dataset in September and October 2022.

Based on the transformed Bluetooth-Time and CV-Time, we built an additional OLS model (Transformed OLS). The modeling result and the fitted functional form of Model 2 are detailed in Table 4. The result indicates that the CV-Time is a significant predictor for estimating the Bluetooth-Time. By performing the square root transportation, Model 2 achieved a higher R^2 compared to Model 1 (0.70 versus 0.67), indicating that more variance in the Bluetooth-Time can be explained by the fitted Model 2 than Model 1.

Table 4. Transformed OLS Result Based on the Square Root Transformed CV-Time Only.

Model Parameters	**Estimate (STD)**
Intercept	0.06 (0.14)
CV_Trans	1.02 (0.03) ***
R^2	0.70
Fitted functional form	$Bluetooth - Time\ =\ \left(0.06\ +\ 1.02\ *\left\|\sqrt{CV - Time}\right\|\ \right)^2$

The response variable in this model is the square root transformed Bluetooth-Time.
The training dataset contains 717 observations; the testing dataset contains 336 observations.
STD stands for the standard errors, which are included in the parenthesis.
*** represents significant 99 percent level based on *p*-values.

ADDITIONAL CV-GENERATED VARIABLES

Besides the CV-Time, we also explored whether there are any other variables that could be generated from the Wejo CV data to improve the model performance in the border crossing time estimation. We believe the traffic congestion could directly influence the traffic volume and the vehicles' driving speeds. Therefore, the hourly aggregated Wejo CV volumes and driving speeds at different locations (Bluetooth readers) could be potential variables to be considered in the modeling effort. Additionally, the traffic volume would be different over the weekend or the rush hours which could potentially be effective as an input parameter.

In this study, we created 22 additional CV-generated variables to capture the traffic volume, and the driving speed dynamics at different locations along the PDN POE aggregated on an hourly basis. As introduced above, five Bluetooth readers are installed along with the PDN POE, and we created five 100-meter buffers around each reader to extract CV data. Therefore, we created four additional CV-generated variables to depict the hourly aggregated CV traffic volume and their driving speeds at five different locations when driving through the PDN POE. Since the traffic volume and crossing times are significantly different during rush hours (7am to 9am or 4pm to 7pm) and on weekends. Therefore, we added two additional temporal variables to distinguish whether a one-hour slot is during rush hours or on the weekend. (22 variables = 4 speed relevant variables × 5 locations [CV extraction buffers] + 2 temporal variables). These new variables are explained in Table 5.

Table 5. Description of Additional CV-Generated Variables.

Feature Name	***Feature Description***
avg_spd_btX	Hourly average CV driving speed at the Bluetooth reader X
85%_spd_btX	85th percentile of the hourly average CV driving speed at the Bluetooth reader X
max_spd_btX	Maximum hourly CV driving speed at the Bluetooth reader X
med_spd_btX	Median of the CV driving speed at the Bluetooth reader X
isWeekend	Boolean variable indicates whether it is a weekend day (1) or not (0)
isRushHour	Boolean variable indicates whether it is a rush hour (1) or not (0)

X = 1, 2, . . ., 5 in this study. For example, avg_spd_bt1 represents the hourly average speed within the 100-meter buffer around PDN-BT1.

RANDOM FOREST MODEL

Random Forest (RF) is one of the commonly used machine learning models capable of handling classification and regression problems. The RF model combines the result of series of decision trees (a forest) to make the model more robust. The RF method reduces risk of overfitting, provides flexibility, and it is easy to determine feature importance. As a result, a RF model for the months of September and October 2022 was generated assuming 70 percent of the data is dedicated to the training dataset and 30 percent is allocated to the test dataset. The model was trained by assuming 10 trees as estimator and squared error as the optimizing criterion. To investigate the contribution of each variable in the resulting model and select the features that have the majority of the contribution to minimize the model complexity, a feature selection process was run with the results shown in **Error! Reference source not found.**. The result shows that the topmost contributing features comprises of CV-Time, average speed at BT3, 85% speed at BT3, 85% speed at BT4 and average speed at BT2.

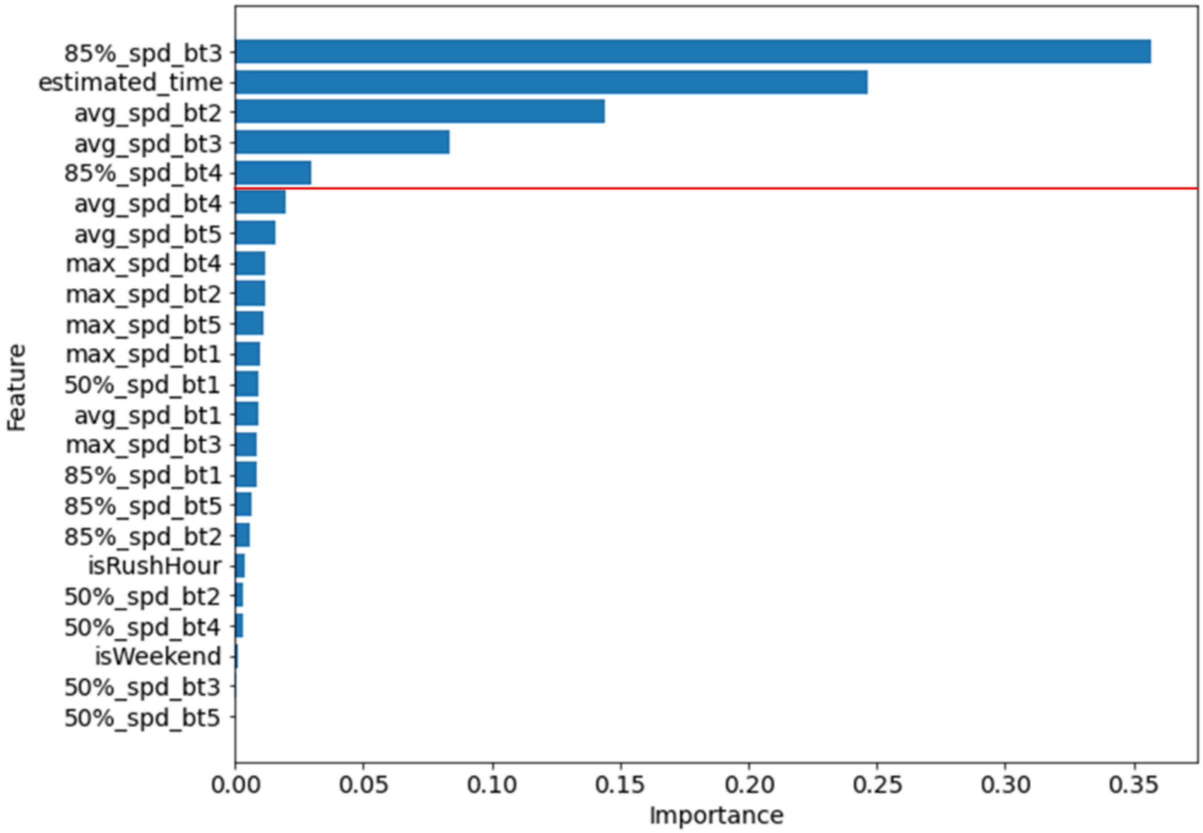


Figure 12. Random Forest Model Feature Importance.

Based on the 5 most contributing features in the wait time estimation, we rebuilt a simplified version of the random forest model. The performance for both the original model with all variables and the simplified model with selected variables is presented in **Error! Reference source not found.**.

Table 6. Performance of Random Forest Models with All and Selected Variables.

Dataset	All variables		Selected variables	
	RMSE	MAPE	RMSE	MAPE
Training dataset	7.31 minutes	15%	6.52 minutes	12%
Testing dataset	15.50 minutes	26%	14.19 minutes	29%

GRADIENT BOOSTING MODEL

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. Gradient boosting creates a small ensemble from the data and generate a model using that data. Then, based on the performance (gradient of the error) it creates the next model with sampling a new ensemble. We are using the XGBoost library to implement the gradient boosting algorithm, which is one the well-known libraries to implement this algorithm. The data used for this model is the same as the one used in the random forest model to compare the performance. In order to simply the model and make the modeling process more feasible, a new model was built which only includes the five most contributing features (**Error! Reference source not found.**). **Error! Reference source not found.** shows the RMSE and MAPE performance measures for both gradient boosting models using all 23 introduced variables and selected top 5 variables.

Table 7. Performance of the Gradient Boosting Models with All and Selected Variables.

Dataset	All variables		Selected variables	
	RMSE	MAPE	RMSE	MAPE
Training dataset	5.27 minutes	11%	7.31 minutes	15%
Testing dataset	14.13 minutes	25%	15.50 minutes	26%

MODELING RESULTS COMPARISON

In this study, we compared the performance of these three models based on two assessment measures: RMSE and MAPE. We also included the directly calculated CV-Time without modeling in the comparison to evaluate whether the modeling efforts could improve the estimation accuracy. The comparison results are detailed in Table 8.

Table 8. Modeling Results Comparison based on the test dataset.

Models	Without Modeling	Simple OLS	Transformed OLS	RF	GB
RMSE	21.67	18.90	19.41	16.04	**15.50**
MAPE	42%	45%	42%	28%	**25%**

“Without Modeling” indicates that we directly calculated the RMSE and MAPE between the Bluetooth-Time and CV-Time in the testing dataset.

Without modeling, the RMSE between the directly calculated CV-Times and the Bluetooth-Times in the testing dataset is 21.67, and the MAPE is 42 percent. By fitting a simple linear regression model with the CV-Time as the only predicting variable (Simple OLS), the estimate accuracy can be improved with the RMSE decreased to 20.60. The results also indicate that performing the square root transformation on the Bluetooth-Time and the CV-Time could significantly improve the model’s performance and robustness, resulting in a better estimation result. The R^2 increased from 0.66 to 0.70.

The OLS, random forest (RF) and gradient boosting (GB) algorithm results based on their significant features are summarized in **Error! Reference source not found.**. The result shows that the GB has the best performance among the introduced models considering both RMSE and MAPE factors. The OLS is significantly less accurate compared to the RF and GB if we only consider the MAPE performance measure. Consequently, it is suggested to utilize either RF or GB rather than OLS. The GB model has slightly better performance compared to RF. It is worthy to note that both RF and GB models have smaller number of features (only 5 features) compared to the OLS model while they performed better than the OLS.

CHAPTER 6:
STREAMING DATA PROCESSING

As we mentioned in chapter 5, the 2022 data we received is a streaming data product, which means the data was generated, aggregated, and delivered to us in near real time. The streaming data analysis is different in nature from the historical data although both of them represent the same amount of information. In order to analyze the streaming data, it is required to design an automated system that captures the data, analyzes the data, and updates the status (result) of the system. Figure 13 summarize the steps to achieve this goal. This system copies the data to a compatible storage service. Then, it runs the analysis and updates the dashboards waiting for the result.

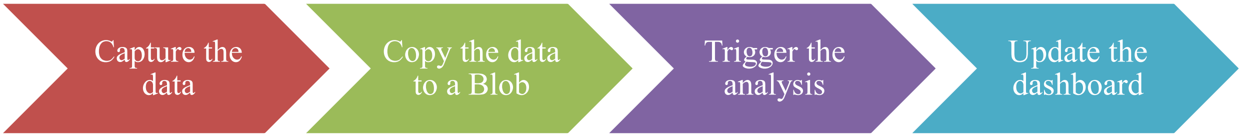


Figure 13. streaming data analysis process.

The main difference between the streaming data analysis and the historical data analysis is that it is necessary to define a trigger for streaming analysis to automatically execute the analysis as soon as the data is available. One of the most successful and saleable solutions to define a trigger and accomplish this task is Azure Data Factory (ADF). ADF is provided by Microsoft Azure and is able to automate the large scale analyzes defined in different Azure services. Among these services we used Azure Blob storage and Azure Databricks to store and analyze the received streaming data.

In this project, the raw data was being delivered to a S3 bucket by Wejo support team on hourly basis. This data was moved to a blob storage on Microsoft Azure service to be accessible by Azure Databricks. Although this step is expensive and time consuming, it is necessary for the prototype implementation. A better solution would be achieved by asking the data provider directly deliver the data to a Blob Storage on Azure which can be easily integrated with the rest of the analysis and visualization tools. The ADF has the ability to copy the data from AWS bucket to an Azure Blob storage securely and conveniently. By adding a short script to the ADF setting makes it possible to trigger the copy event whenever a new file is added to the S3 Bucket in AWS.

Figure 14 represents the flowchart of the designed prototype. The data provider (Wejo) delivers the data on a S3 bucket in AWS hourly. The designed trigger in ADF detects the new data and starts copying the new data to an Azure blob storage. Whenever all the new files covering one hour are successfully copied to the destination Blob in Azure, a flag will be raised that the data is ready. Then, the ADF initiates the analysis by Databricks. The program in Databricks that estimated the border crossing will be same as the historical one with only one difference. The streaming data analysis scope is much smaller (it only deals with one-hour data). Assuming the final figure showing the history of the border crossing over the past few hours and the current one will be shown on a publicly available website, the server can pull the data generated with Azure Databricks. The result is stored in the Azure Blob and again ADF is in charge to notify the server that the data is ready to be pulled.

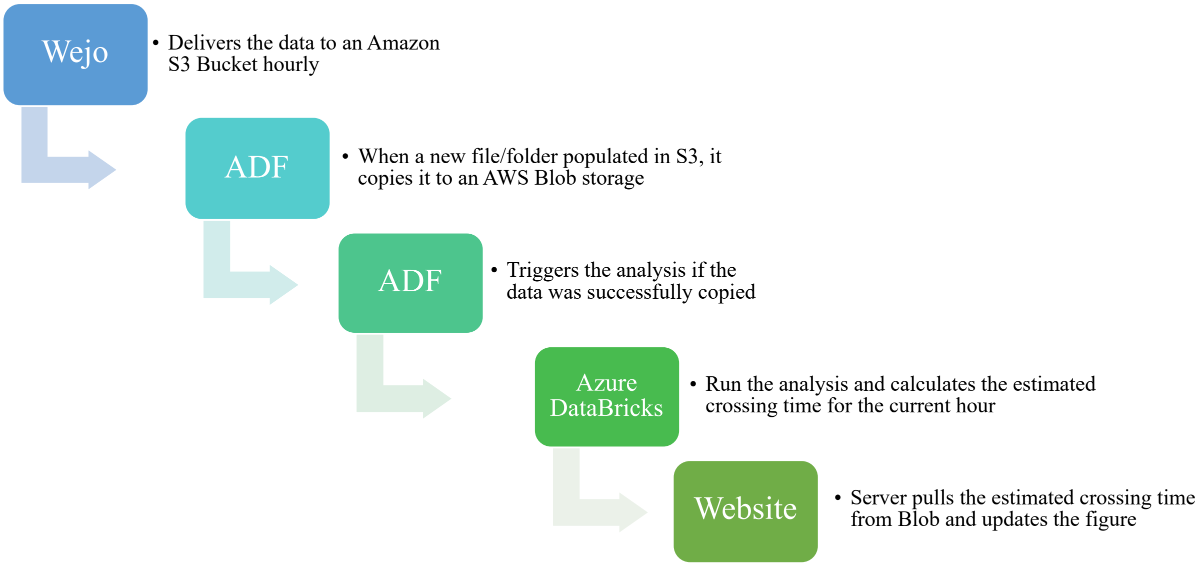

Wejo
• Delivers the data to an Amazon S3 Bucket hourly
ADF
• When a new file/folder populated in S3, it copies it to an AWS Blob storage
ADF
• Triggers the analysis if the data was successfully copied
Azure DataBricks
• Run the analysis and calculates the estimated crossing time for the current hour
Website
• Server pulls the estimated crossing time from Blob and updates the figure

Figure 14. Streaming Data Ananlysis Prototype Flowchart.

CHAPTER 7: CONCLUSIONS AND LIMITATIONS

CONCLUSIONS

POEs are gateways for international trades, which play an important role in the U.S. economy. Effectively and accurately monitoring border crossing times is of great importance in border transport management, which benefits various stakeholders. This project is among the first to comprehensively evaluate the effectiveness of the market-available CV data provided by Wejo in congestion studies at border crossings.

In this study, we developed a set of big data analytic tools to process Wejo CV datasets. Meanwhile, we evaluated the temporal coverage of Wejo CV data at the PDN POE and its correlation with the existing Bluetooth-Time. This study developed different models to estimate the Bluetooth-Time based on CV-generated variables. The main findings are summarized as follows:

- Bluetooth observations and Wejo CV data can generate highly similar results for border crossing times with a correlation efficient around 0.8.
- The best fitted model based on the combination of CV-Time and additional CV-generated variables can produce an effective model to estimate Bluetooth-Time with an RMSE of 15.5 mins and a MAPE of 25 percent.
- A significant increase was observed in 2022 CV data, implying the growing potential and practical value of CV data in border crossing management. However, we also need to acknowledge that the penetration rate of the 2022 Wejo dataset collected from a single OEM was still insufficient for our study site (the PDN POE), resulting in around 30 percent of the testing hours lacking Wejo samples.

Given these findings, the Wejo CV data are a potential and promising data source for monitoring border crossing times. The data can be used to supplement the existing BCIS at El Paso and could be treated as an alternative solution when the BCIS is down for maintenance. Although the streaming CV data for border regions is only available from one of Wejo's partnered OEMs due to the license issue, the highly accurate modelling results indicate the possibility of replacing the existing Bluetooth-based system with the CV data if Wejo could make all OEMs' data available and continuously provide sufficient CV samples at all hours every day.

LIMITATIONS

This current project could be improved from two perspectives: improve the experimental settings and extend the evaluation period. Due to data licensing limitations, we only examined several months of CV data at one POE—the PDN POE. Therefore, the performance of Wejo CV data for other POEs could be different. This evaluation was primarily conducted based on the 2022 Wejo data, majorly collected only from one OEM. Therefore, around 30 percent of the testing hours lack sufficient Wejo samples to generate valid estimates for border crossing times.

For further determining how to incorporate the CV data into the current border-monitoring system, there is a practical necessity to evaluate more CV data to see if a broader temporal coverage at different POEs over a more extended study period can be achieved.

REFERENCES

An, Shi, Haiqiang Yang, Jian Wang, Na Cui, and Jianxun Cui. 2016. "Mining Urban Recurrent Congestion Evolution Patterns from GPS-Equipped Vehicle Mobility Data." *Information Sciences* 373: 515–526. doi:10.1016/j.ins.2016.06.033.

Bruce, Peter, and Andrew Bruce. 2017. *Practical Statistics for Data Scientists*.

Honeywell. 2020. "What Is the Distance/Range a Bluetooth Scanner Can Be Used from the Base?" https://support.honeywellaidc.com/s/article/What-is-the-distance-range-a-Bluetooth-scanner-can-be-used-from-the-base.

Hoseinzadeh, Nima, Ramin Arvin, Asad J. Khattak, and Lee D. Han. 2020. "Integrating Safety and Mobility for Pathfinding Using Big Data Generated by Connected Vehicles." *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*. doi:10.1080/15472450.2019.1699077.

Iyer, Shiva R., Kate Boxer, and Lakshminarayanan Subramanian. 2018. "Urban Traffic Congestion Mapping Using Bus Mobility Data." *CEUR Workshop Proceedings* 2227: 7–13.

James, Gareth, Daniela Witten, and Trevor Hastie. 2019. *Introduction to Statistical Learning with Applications in R*. *Synthesis Lectures on Mathematics and Statistics*. Vol. 11.

Khan, Ata M. 2010. "Prediction and Display of Delay at Road Border Crossings." *The Open Transportation Journal* 4: 9–22.

Li, Xiao, Bahar Dadashova, Siyu Yu, and Zhe Zhang. 2020. "Rethinking Highway Safety Analysis by Leveraging Crowdsourced Waze Data." *Sustainability (Switzerland)* 12 (23). doi:10.3390/su122310127.

Lin, L., Q. Wang, A. Sadek, and G. Kott. 2015. "An Android Smartphone Application for Collecting, Sharing, and Predicting Border Crossing Wait Time." *Proceedings of the Transportation Research Board Annual Meeting (TRB'14)*, February 2019.

Lin, Lei, Qian Wang, and Adel W. Sadek. 2014. "Border Crossing Delay Prediction Using Transient Multi-server Queueing Models." *Transportation Research Part A: Policy and Practice* 64: 65–91. doi:10.1016/j.tra.2014.03.013.

McCord, Mark, Colin Brooks, and David Banach. 2016. *Truck Activity and Wait Times at International Border Crossings*. https://www.purdue.edu/discoverypark/nextrans/assets/pdfs/new/120OSU2.1 Truck Activity and Wait Times McCord and Brooks_NEXTRANS final report.pdf.

Miles, Stephanie. 2019. "6 Automotive Data Services Platforms." https://streetfightmag.com/2019/08/23/6-automotive-data-services-platforms/#.YMlWW_lKguU.

Miyagawa, Yujin, and Masakatsu Ogawa. 2017. "Immediate Cooperative Line Wait Time Estimation System Using BLE on Smartphone." *Journal of Signal Processing* 21 (4): 129–132. doi:10.2299/jsp.21.129.

Rajbhandari, Rajat, and Juan Villa. 2012. “Radio Frequency Identification System to Measure Crossing and Wait Times of Commercial Vehicles at U.S. Border Crossings.” *Transportation Research Record* 2285: 126–134. doi:10.3141/2285-15.

Rajbhandari, Rajat, Juan Villa, Roberto Macias, and William Tate. 2012. *Measuring Border Delay and Crossing Times at the U.S .–Mexico Border—Part II Guidebook for Analysis and Dissemination of Border Crossing Time and Wait Time Data Final Report*. https://rosap.ntl.bts.gov/view/dot/26032.

Rajbhandari, Rajat, Juan Villa, W. Tate, S. Samant, L. Ruback, and D. Kang. 2012. *Measuring Border Delay and Crossing Times at the US–Mexico Border Final Report*. https://ops.fhwa.dot.gov/publications/fhwahop12049/index.htm.

Ramezani, Mohsen, and Nikolas Geroliminis. 2015. “Queue Profile Estimation in Congested Urban Networks with Probe Data.” *Computer-Aided Civil and Infrastructure Engineering* 30 (6): 414–432. doi:10.1111/mice.12095.

Roberts, Bryan, Adam Rose, Nathaniel Heatwole, Dan Wei, Misak Avetisyan, Oswin Chan, and Isaac Maya. 2014. “The Impact on the US Economy of Changes in Wait Times at Ports of Entry.” *Transport Policy* 35: 162–175. doi:10.1016/j.tranpol.2014.05.010.

Roberts, Bryan, Steve Mcgonegal, Fynnwin Prager, Dan Wei, Adam Rose, Charles Baschnagel, Timothy Beggs, and Omeed Baghelai. 2014. *Analysis of Primary Inspection Wait Time at U.S. Ports of Entry*. http://raycomgroup.images.worldnow.com/library/08f5a31f-5773-4d9b-8926-7d85b06142cd.pdf.

Sabean, Jonathan, and Crystal Jones. 2008. *Inventory of Current Programs for Measuring Wait Times at Land Border Crossings*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.511.694&rep=rep1&type=pdf.

Sabean, Jonathan, Lynne Lussier, and Jim Pattan. 2011. *Effort to Test, Evaluate and Deploy Technologies to Automate the Measurement of Real-Time Border Wait Times at United States-Canada Land Border Crossings*. https://ops.fhwa.dot.gov/publications/fhwahop11025/bwt_techbrf.htm.

San Diego Association of Governments. 2017. *Border Wait Time Technologies and Information Systems White Paper*. https://www.sandag.org/uploads/publicationid/publicationid_4469_23227.pdf.

Sankaranarayanan, Hari Bhaskar, Gaurav Agarwal, and Viral Rathod. 2016. “An Exploratory Data Analysis of Airport Wait Times Using Big Data Visualisation Techniques.” *2016 International Conference on Computation System and Information Technology for Sustainable Solutions, CSITSS 2016*, 324–329. IEEE. doi:10.1109/CSITSS.2016.7779379.

Sharma, Sushant, Dong Hun Kang, Abhisek Mudgal, Jose Rivera, Montes De Oca, Swapnil Samant, and Gabriel A Valdez. 2018. *Developing Adaptive Border Crossing Mobility Measures and Short-Term Travel Time Prediction Model Using Multiple Data Sets*. https://static.tti.tamu.edu/tti.tamu.edu/documents/185917-00015.pdf.

Talebpour, Alireza, and Hani S. Mahmassani. 2016. “Influence of Connected and Autonomous Vehicles on Traffic Flow Stability and Throughput.” *Transportation Research Part C: Emerging Technologies* 71. doi:10.1016/j.trc.2016.07.007.

Texas A&M Transportation Institute. 2021. "How RFID Based Wait and Crossing Time System Works." https://bcis.tti.tamu.edu/Commercial/en-US/help-and-glossary.aspx.

U.S. Customs and Border Protection. 2016. *Automated Wait Time and Trade Facilitation Performance Measures*. https://www.dhs.gov/sites/default/files/publications/Customs and Border Protection - Automated Wait Time and Trade Facilitation Performance Measures_0.pdf.

Villa, Juan Carlos, Rajat Rajbhandari, and Melissa Tooley. 2017. *Developing a Concept of Operations for an Innovative System for Measuring Wait Times at Land Ports of Entry Final Report.* https://www.uh.edu/bti/research/Developing-a-Concept-of-Operations-for-an-Innovative-System-for-Measuring-Wait-Times-at-Land-Ports-of-Entry/Villa DHS Final Report 032817.pdf.