



Technical Report 121

Transportation Data Discovery Environment: Application for Connected Corridors

Research Supervisor: Natalia Ruiz Juri
Center for Transportation Research

September 2018

Data-Supported Transportation Operations & Planning Center (D-STOP)

A Tier 1 USDOT University Transportation Center at The University of Texas at Austin



**CENTER FOR
TRANSPORTATION
RESEARCH**



**Wireless Networking &
Communications Group**

D-STOP is a collaborative initiative by researchers at the Center for Transportation Research and the Wireless Networking and Communications Group at The University of Texas at Austin.

Technical Report Documentation Page

1. Report No. D-STOP/2018/121		2. Government Accession No.		3. Recipient's Catalog No.	
4. Title and Subtitle Transportation Data Discovery Environment: Application for Connected Corridors				5. Report Date September 2018	
				6. Performing Organization Code	
7. Author(s) Natalia Ruiz Juri, Nadia Florez, Ken Perrine, Heidi Ross				8. Performing Organization Report No. Report 121	
9. Performing Organization Name and Address Data-Supported Transportation Operations & Planning Center (D-STOP) The University of Texas at Austin 3925 W. Braker Lane, 4th Floor Austin, Texas 78759				10. Work Unit No. (TRAIS)	
				11. Contract or Grant No. DTRT13-G-UTC58	
12. Sponsoring Agency Name and Address Data-Supported Transportation Operations & Planning Center (D-STOP) The University of Texas at Austin 3925 W. Braker Lane, 4th Floor Austin, Texas 78759				13. Type of Report and Period Covered	
				14. Sponsoring Agency Code	
15. Supplementary Notes Supported by a grant from the U.S. Department of Transportation, University Transportation Centers Program. Project Title: Transportation Data Discovery Environment					
16. Abstract D-STOP project 121, Transportation Data Discovery Environment, began in January 2015 and encompassed a number of data-based activities before it ended in August 2018. Through this project, the research team leveraged the computing resources and expertise at UT Austin to develop a “data discovery environment” for transportation data to aid decision-making. The team began by creating that environment, termed the Data Rodeo: a web-based data clearinghouse that corrals transportation-related big data from regional partners in Central Texas. This dynamic ecosystem is available to the broader transportation community for usage and applications. From the Data Rodeo initiative grew additional sub-projects. The final activity was the creation of a work on connected corridors for the City of Austin. This work focused on identifying and implementing the workflows required to effectively evaluate the performance of arterial corridors based on data from multiple sources. In this context, D-STOP researchers developed and implemented workflows and a supporting system architecture to ingest, process, and catalog traffic data collected manually and through sensors. The workflows developed for this task, along with the corresponding decisions related to data storage and schemas, set the basis for developing an efficient framework to support the use of the Data Rodeo, which proved to be a valuable tool to motivate further discussion that may identify a variety of relevant use cases.					
17. Key Words Data Rodeo, connected corridor, sensors, mobility			18. Distribution Statement No restrictions. This document is available to the public through NTIS (http://www.ntis.gov): National Technical Information Service 5285 Port Royal Road Springfield, Virginia 22161		
19. Security Classif.(of this report) Unclassified		20. Security Classif.(of this page) Unclassified		21. No. of Pages 48	22. Price

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgements

The authors recognize that support for this research was provided by a grant from the U.S. Department of Transportation, University Transportation Centers.

Preface

D-STOP project 121, *Transportation Data Discovery Environment*, began in January 2015 and encompassed a number of data-based activities before it ended in August 2018. Through this project, the research team leveraged the computing resources and expertise at UT Austin to develop a “data discovery environment” for transportation data to aid decision-making. The team began by creating that environment, termed the Data Rodeo: a web-based data clearinghouse that corrals transportation-related big data from regional partners in Central Texas (the local MPO, the City of Austin, and the local transit agency). This dynamic ecosystem is available to the broader transportation community for usage and applications. Initially, the project focused on creating more meaning from existing data sources, and grew to include more novel data sources and methods. The Data Rodeo is web-based; part of the research included not only building the tool but developing appropriate protocols for access and governance.

From the Data Rodeo initiative grew additional sub-projects. The final activity was the creation of a work on connected corridors for the City of Austin. This work focused on identifying and implementing the workflows required to effectively evaluate the performance of arterial corridors based on data from multiple sources. In this context, D-STOP researchers developed and implemented workflows and a supporting system architecture to ingest, process, and catalog traffic data collected manually and through sensors.

The workflows developed for this task, along with the corresponding decisions related to data storage and schemas, set the basis for developing an efficient framework to support the use of the Data Rodeo, which proved to be a valuable tool to motivate further discussion that may identify a variety of relevant use cases.

Contents

Chapter 1. Executive Summary	1
Chapter 2. Introduction	3
Chapter 3. Data Description	4
3.1 Bluetooth Sensors	5
3.2 Wavetronix Sensors	6
3.3 GRIDSMA RT Cameras	7
3.4 Automated Vehicle Location.....	7
3.5 Automated Passenger Counts.....	8
3.6 INRIX Waypoint Data	8
3.7 Connected Vehicle Data	11
3.8 Custom Wi-Fi sensors	12
Chapter 4. Data Workflows	13
4.1 Existing Workflows and System Architecture	13
4.2 Future Workflows.....	20
Chapter 5. Data Processing and Analysis	21
5.1 Bluetooth	21
5.2 Wavetronix	24
5.3 GRIDSMA RT Cameras	28
5.4 Automated Vehicle Location Data	28
5.5 Automated Passenger Counts.....	29
5.6 INRIX Waypoint Data	30
Chapter 6. Bond Corridor Performance Analysis Tool	33
Chapter 7. Additional Considerations and Future Direction	38
7.1 Data Sharing and Collaboration with the Corridor Project Office	38
7.2 Coordination of Corridor Data Analysis Efforts across Stakeholders	39
7.3 Stakeholder Meetings	39
7.4 Future Efforts	39
Appendix 1: Source Code List	40

Chapter 1. Executive Summary

Task 2017-10, Connected Corridor, was focused on identifying and implementing the workflows required to effectively evaluate the performance of arterial corridors based on data from multiple sources. In this context, CTR developed and implemented workflows and a supporting system architecture to ingest, process, and catalog traffic data collected manually and through sensors. The original project scope included the analysis of connected vehicle data, but delays in the deployment of the corresponding technology prevented CTR from performing such analyses.

In order to inform the design of effective, flexible, and scalable data workflows, CTR considered the characteristics and intended use of multiple data sources and conducted preliminary work on data analysis and visualization. The workflows and architecture developed for this project were leveraged (in a separately funded project) to develop a web-based application for multi-modal corridor performance analysis. While the Bond Corridor Performance Analysis Tool (BCPAT) was designed with the goal of studying the evolution of performance on corridors to be improved under the Mobility Bond, the framework is available to the City of Austin and may be extended to the analysis of additional corridors.

The outcomes of this work include documented and implemented workflows to access and archive sensor data (Wavetronix, Bluetooth, Gridsmart), transit data (GTFS, AVL, APC), and manually collected traffic volumes/turning movement counts at selected locations. For some data sources, replicable data aggregation, cleaning, and visualization methods were also explored. The availability of considered data sources along bond corridors can be explored through a web application. The web tool interface includes a home page with a data summary table, a map that summarizes corridor data availability, and separate tabs for the interactive analysis of corridor travel time and traffic volume/speed at selected locations, among others.

In the context of Task 2017-10, CTR also supported the analysis of a large third-party dataset that contains individual vehicle trajectories collected through GPS over three months. The analysis explored the use of probe data to support corridor performance assessment at locations not covered by other data sources. Results suggested some limitations in the applicability of the analyzed data, which led to postponing the integration of this source to the analysis framework. Additional exploratory work to be conducted at a later stage includes the use of inexpensive, custom-developed Wi-Fi sensors to collect data similar to that produced by Bluetooth sensors. CTR developed scripts to extract and archive the data from prototype sensors and extended existing visualization tools for its analysis. While results are promising, Austin Transportation Department would like to further evaluate the effort required to deploy and maintain additional sensors.

The workflows developed for this task, along with the corresponding decisions related to data storage and schemas, set the basis for developing an efficient framework to support the use of archived traffic data collected by the City of Austin (CoA). The web application has been a valuable tool to demonstrate potential data uses, and to motivate

further discussion that may identify a variety of relevant use cases. The next stage of this project will focus on deploying the proposed architecture in a cloud-based system identified by the CoA, and implementing any refinements and extensions required to incorporate new data sources and address performance issues. The refinement and extension of the web application will continue to be funded separately by the Bond Corridor Program Office, and used to inform system design by identifying data use cases and related challenges.

Chapter 2. Introduction

The Austin Transportation Department (ATD) owns a variety of sensors that generate continuous data streams for system monitoring purposes. ATD also has access to speed data based on probe vehicles provided by a third-party vendor and archives of manual traffic volume and turning movement counts conducted periodically. Further, in the near future the CoA may operate its first “connected corridor,” where vehicle-to-infrastructure communications technologies could support advanced traffic management and safety strategies.

Real-time system monitoring data is already providing value to ATD, and further benefits are expected if archived data can be efficiently accessed and analyzed. The goal of this study is to identify and test the critical components of a framework that can support the archiving of, and systematic access to, traffic data collected by ATD. The framework is intended to support the development of tools and methods that may inform not only traffic operations, but also strategic decision making, planning, and eventually data sharing with a broader audience.

As an example of a valuable use case of the framework described above, CTR leveraged the corresponding data workflows and system architecture to support the development of a web-based tool, the Bond Corridor Performance Analysis Tool (BCPAT), for the analysis of corridor performance over time. While the BCPAT, separately funded by the Bond Corridor Program Office, is intended to analyze data along the corridors to be improved under the Mobility Bond, it may be applied to the study of any desired corridor. The development of the BCPAT provided valuable insights for the design of the methods implemented in this study.

The following sections describe each of the data types considered in this project, including [Bluetooth sensors](#), [Wavetronix sensors](#), [GRIDSMART cameras](#), [Automated Vehicle Location \(AVL\)](#), [Automated Passenger Counts \(APC\)](#), [INRIX Waypoint data](#), and manually collected traffic volumes and turning movement counts (Traffic Studies). We also present corresponding data access, processing, and archiving workflows, and discuss the outcomes of preliminary data analyses and visualizations when appropriate. While connected vehicle data and probe-based vehicle trajectories are not expected to be incorporated into the analysis framework in early stages, this reports includes preliminary analyses and considerations about both data types. Similarly, documentation of an exploratory analysis of the use of custom-made Wi-Fi sensors to complement Bluetooth sensor data is included, although the corresponding data workflows may not be integrated to the framework in the short term. A separate section provides a description of the capabilities and use of the web-based tool, while the final sections discusses a potential system architecture based on the developed methods and workflows. Appendix 1 lists the scripts/code developed to support this project’s work.

Chapter 3. Data Description

Integrated and proactive management of multimodal transportation corridors presents a compelling opportunity to improve the mobility of people and goods.¹ Performance data, such as average travel time, throughput (transit and passenger cars), transit speeds, and system use are critical in the evaluation of the impact of corridor management strategies and technologies. Developing meaningful metrics may require the use of multiple data sources, the use of which will greatly benefit from the availability of an efficient framework to access, and in some cases process/analyze, performance-related data.

To support the development of such framework, CTR worked with six data sources/types: Bluetooth sensors, Wavetronix sensors, GRIDSMART cameras, transit automated vehicle location (AVL) data, automated passenger counts (APC), and vendor vehicle trajectories from a vendor (INRIX Waypoint data). Data samples were used to understand data characteristics and, in some cases, develop data processing, analysis, and visualization methods. For some data types researchers also developed and implemented methods to systematically access and archive data continuously collected by sensors, along with techniques to process and ingest the data into a database to support replicable analyses and visualizations. The latter are accomplished through a web-based application developed with support from the Bond Corridor Program Office. Figure 1 displays the locations and availability of some of the sensors considered in this study on the corridors affected by the Mobility Bond.

The current framework uses computing resources hosted by UT Austin. Both workflows and the system architecture will be further developed in new 2018 tasks, seeking for a scalable deployment of the framework on systems/resources controlled by the City of Austin. Some of the challenges found while developing such application, including performance, will inform the refinement and final implementation.

Descriptions for each considered data type and sensor are provided below. Some of the challenges in data cleaning and analysis are also described, and we discuss how each data type is relevant in the context of connected corridors. The following section further discusses data processing and workflows.

¹ Robert Sheehan, “Integrated Corridor Management meets Connected Traveler,” presentation to ITS World Congress, Oct. 2015. https://www.its.dot.gov/presentations/world_congress2015/wc2015_CVMOD.pdf, last accessed Oct. 8, 2018.

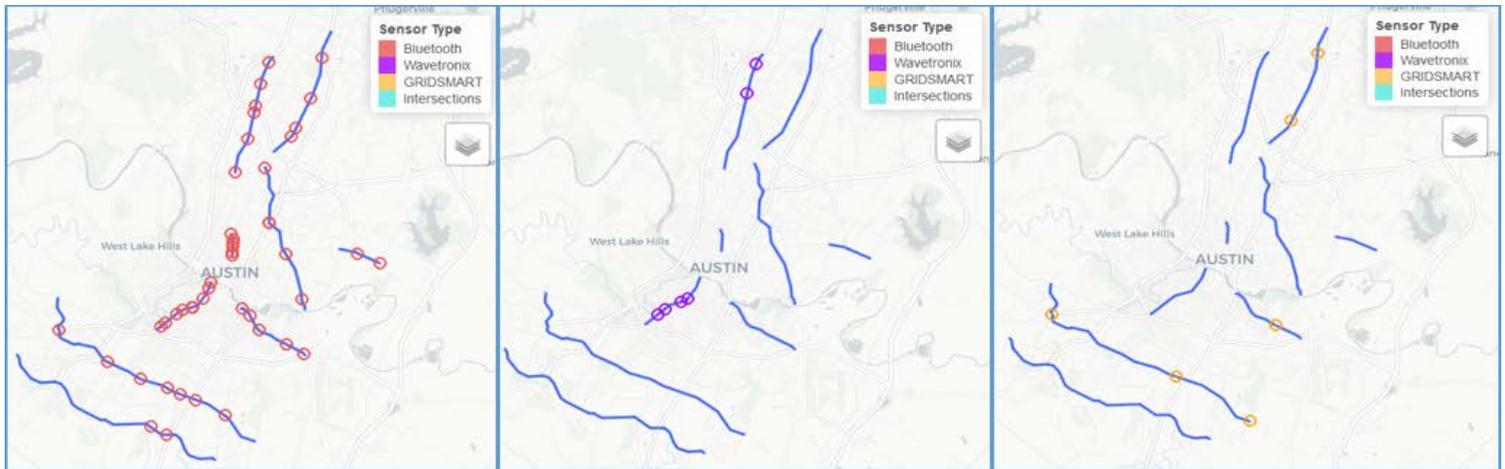


Figure 1. The location of all sensors, divided by sensor type, along the City of Austin Transportation Bond Corridors. From left: Bluetooth sensor locations; Wavetronix sensor locations; and, GRIDSMART sensor locations

3.1 Bluetooth Sensors

The City of Austin maintains 128 Bluetooth sensors across the city limits that log the Machine Access Control (MAC) address and timestamp of detected Bluetooth devices. By placing at least two Bluetooth sensors along a road segment, travel time can be calculated using the time difference associated with each MAC identifier (MAC addresses are valid for a twenty-four-hour period).

To calculate travel time along a corridor, two methods of travel-time analysis are utilized: end-to-end and segment-by-segment. The former filters MAC addresses and corresponding timestamps to those only matched with sensors at the two ends of a corridor, computing travel time directly using those observations. Segment-by-segment analysis leverages all Bluetooth sensors along a corridor and sums the travel time of each segment to measure travel time along the length of the corridor. Because a single MAC address rarely traverses a corridor end-to-end, average travel time of corridor segments are combined to obtain corridor travel time.

Segment travel time can be aggregated by considering the same time period, called instantaneous travel time, or in a time-dependent fashion that accounts for traveler experience, called experienced travel time. The methods currently implemented compute instantaneous corridor travel time (experienced corridor travel time will be studied in future work). Figure 2 illustrates the visualization of corridor travel times by time of day and sub-segment, and per travel direction.

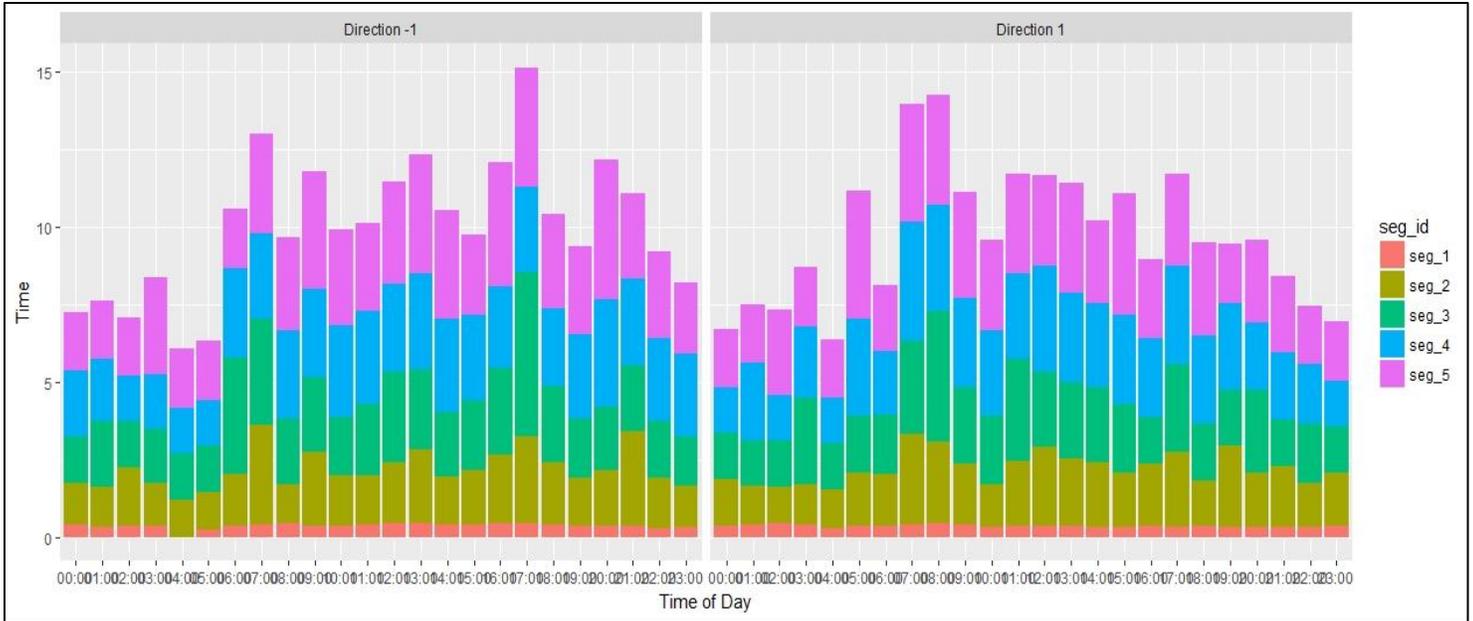


Figure 2. Average travel time distribution along Burnet Road

Bluetooth sensor data limitations include: only sampling a subset of travelers; inability to track individual traveler behavior across days; and, analyses may capture varying types of traveler behavior including turning movements and stops, as well as modes. Because not all travelers enable Bluetooth on mobile devices, the dataset encompasses only a subset of all traffic, estimated to be 2.02 to 8.13 percent of vehicles on the road.² Moreover, the unique identifiers of enabled Bluetooth devices change every twenty-four hours due to security concerns, making the differentiation between types of travelers (e.g., those traveling for work or leisure) and characterization of daily travel patterns difficult. Lastly, vehicle turning movements and stops may bias travel time upwards. For example, a vehicle stopped at an intersection waiting to turn left will record a higher travel time than a vehicle proceeding through the intersection.

3.2 Wavetronix Sensors

The Wavetronix radar is a nonintrusive vehicle detector that provides information on traffic volume, occupancy, and speed. Currently, the City of Austin has nineteen Wavetronix sensors at intersections across the city. The data collected between June 2017, when the sensors were first deployed, and the present is accessible via the City’s Open Data Portal. In June 2017, only three sensors were online; by August 2017, a total of twelve sensors were in use.

The output dataset, consisting of fifteen-minute interval counts and speeds-per-lane, can be aggregated to give details of traffic volume, occupancy, and speed per travel direction.

² Sharifi et al., “Analysis of Vehicle Detection Rate for Bluetooth Traffic Sensors: A Case Study in Maryland and Delaware,” 18th World Congress on Intelligent Transport Systems (Oct. 2011).

This data may provide valuable insights for planning and evaluation by allowing comparisons of system use and performance across time periods (e.g., before, during, and after construction) and/or time of day.

One of the challenges when analyzing data from Wavetronix sensors is that it is highly localized in nature (point speeds and volumes). Explaining detected trends may require the analysis of several sensors, or obtaining additional data from separate sources. Additionally, missing data may limit data availability for some locations/time of day.

3.3 GRIDSMART Cameras

GRIDSMART cameras use automated video data analysis to identify vehicles crossing through user-defined regions in an intersection. The resulting data may be processed to generate volumes per approach, turning movement counts, speeds, and delays, among others. The sensor system gathers data through of a fisheye camera mounted high above the respective location. Austin has more than thirty working GRIDSMART devices in operation, and many more are currently being installed throughout the city. CTR has conducted preliminary analysis and visualizations of traffic volumes and turning movement counts, per approach and time of day, at a single location. Future work will explore data aggregation and cleaning approaches, and consider additional data types provided by GRIDSMART cameras.

One challenge in using GRIDSMART data is knowing how accurate the vehicle counts are, as some of its images significantly distorted or low-resolution. To validate GRIDSMART accuracy, it would be necessary to manually count vehicles over a given time period and compare the results to the GRIDSMART data. CTR will include further data validation work in 2018 tasks. While GRIDSMART is geared toward measuring traffic, it is not suitable for measuring pedestrian or bicycle traffic.

3.4 Automated Vehicle Location

The Capital Metro Automatic Vehicle Location (AVL) system automatically determines and transmits the geographical location of transit vehicles, providing information on location and spot speed. The data, accessible at the State of Texas Open Data Portal, consists of location measures every two minutes for each Capital Metro bus. The information provided includes the trip ID, vehicle, and route identification numbers, among other variables that allow the real-time data to be fused with other transit datasets to provide meaningful analyses.

AVL information can provide insights into transit travel time, average speed, and delays, among other details. However, the main limitation of the data is that, due to the two-minute collection interval, location data can vary greatly and there's no additional information within this time window. Therefore, it is difficult to estimate delay and speed accurately at the route level.

3.5 Automated Passenger Counts

The Capital Metro Automatic Passenger Counter (APC) ridership dataset provides information—collected via door-mounted sensors—about transit vehicle location and corresponding passenger boarding and alighting at the stop level. Capital Metro shares the dataset twice per year through the State of Texas Open Data Portal. Its forty-seven columns of data include details about route identification number, dwell time, vehicle identification number, maximum load, timestamp, latitude and longitude, and other details.

APC data can support the estimation of transit use and reliability. For instance, measures like dwell time and occupancy can be compared across time and corridors. Currently, the main limitation of the data is the lack of a data dictionary with descriptions of all available variables, and the limited sharing.

3.6 INRIX Waypoint Data

The INRIX agreement with the City of Austin includes access to two months of raw waypoint data per year. Waypoint data consists of individual GPS points describing each trip performed by vehicles that participate in the INRIX data collection program. Figure 3 illustrates the trajectory of one such trip.

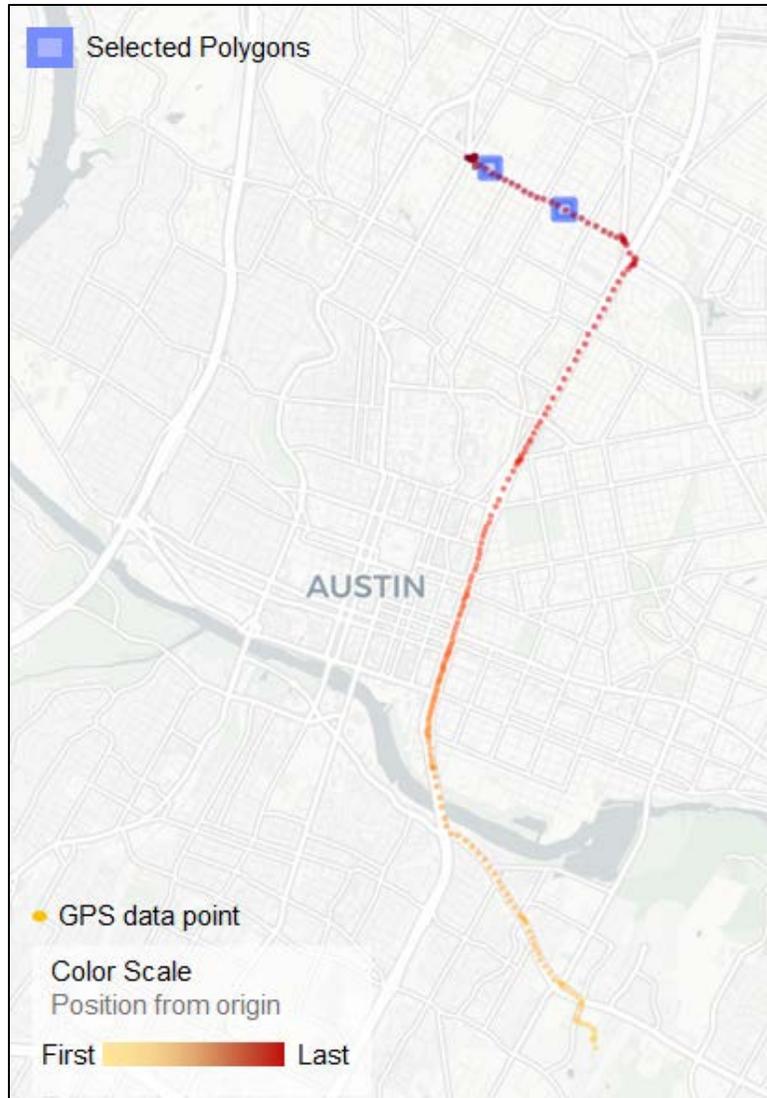


Figure 3: Waypoints for a trip that traverses two user-defined polygons

For each data point in Figure 3, INRIX provides a timestamp (date and time at which the data point was collected), geographic location, and speed. Table 1 provides a list of all the attributes associated to each data point.

Table 1: GPS point attributes

Attribute	Description
<i>ID</i>	GPS point identifier
<i>trip_id</i>	Trip identifier
<i>waypoint_seq</i>	Order of GPS point within the trip
<i>capture_time</i>	Date and time at which the GPS point is reported
<i>lat</i>	Latitude of GPS point
<i>lon</i>	Longitude of GPS point
<i>segment_id</i>	INRIX segment ID when applicable
<i>zone_name</i>	INRIX zone name (e.g. Williamson, Travis)
<i>device_id</i>	Device ID
<i>raw_speed</i>	Speed of vehicle when the GPS point is collected (kph)
<i>raw_speed_metric</i>	Unit used for speeds

For this research, a total of three months of data were available for analysis: March 2017, October 2017, and March 2018. Table 2 provides general statistics about the dataset.

Table 2: Waypoint data characteristics

Metric	All	3/2017	10/2017	3/2018
<i>Number of waypoints</i>	506 million			
<i>Distinct trips</i>	6.8 million	2.5 mill	2.3 mill	2.1 mill
<i>Distinct devices</i>	2.7 million	853 K	964 K	895 K
<i>Trips per device per day</i>	1.9	2.1	1.9	1.8
<i>Average trip length*</i>	18 mi	16.7 mi	18.7 mi	19.1 mi
<i>Average trip duration</i>	36 min	32 min	38 min	40 min
<i>Average points per trip</i>	62	56	59	72
<i>Average point frequency</i>	1'29"	1'26"	1'32"	1'29"
<i>Trip duration <1 minute</i>	3.8%	4.9%	2.0%	3.2%
<i>Trip length >50 miles</i>	32%	28%	33%	34%

* Computed for trips shorter than 100 miles, which comprise 78%, 83%, and 68% of all trips in 3/2017, 10/2017 and 2/2018, respectively

A significant portion of the recorded trips were longer than fifty miles and somewhat indirect (Figure 4), suggesting that they correspond to commercial vehicle routes rather than to personal vehicles. The data mapped in Figure 4 may also represent several consecutive trips taken by a single individual that have not been split appropriately. None of the attributes in the available data contain information on the type of vehicle for which trips are reported.

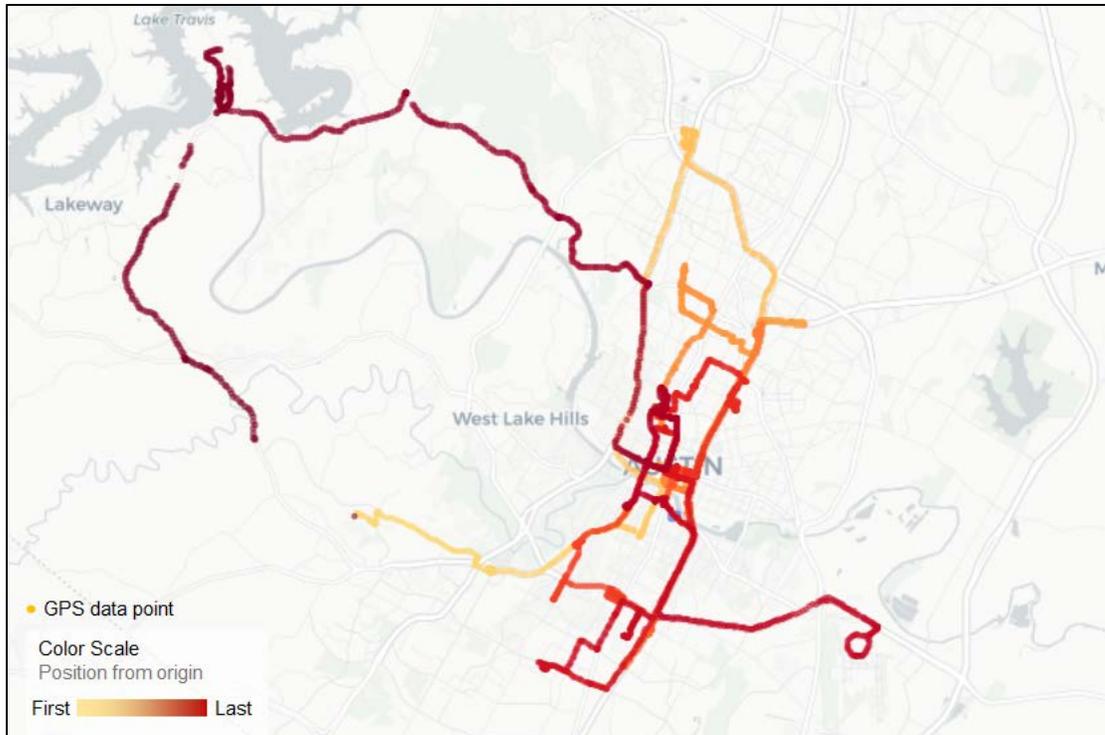


Figure 4: Example of long trip in waypoint dataset

One of the goals of analyzing waypoint data was to understand origin-destination travel times, but that effort requires further understanding of how trips are defined by INRIX and the ability to distinguish between commercial vehicles and passenger cars.

Another potential use of INRIX data was the analysis of speed/travel times at locations where INRIX does not currently report such data. A [case study of Manor Road](#), described later in this report, suggests that travel time analysis may not be feasible at all locations.

3.7 Connected Vehicle Data

The City of Austin is working with Cisco on a pilot study of vehicle-to-infrastructure communications technology on the Riverside Corridor. As part the study, Cisco will deploy roadside units that are capable of receiving messages from connected vehicles, and equip selected vehicles from the City of Austin fleet with on-board connectivity technology. Cisco will also make available to CTR some of its connected data platforms for data access and analysis.

In support of the pilot, researchers from CTR attended meetings with Cisco and the City of Austin—and also met separately with Cisco representatives—to gain better understanding of the capabilities of available data platforms. CTR currently has access to Cisco’s [Kinetic for Cities](#) application, a sensor data aggregation and analysis platform. Cisco has also offered to provide access to the Kinetic Edge and Fog Module (EFM), which may allow for more efficient data pre-processing and aggregation for storage purposes. The EFM can also enable more advanced real-time applications.

The planned pilot deployment of connected vehicle technologies was not finalized during the research period for Task 2017-10 so no data has been collected or analyzed by CTR. The time budgeted for the analysis of connected vehicle data analysis was redirected to support the deployment of new low-cost Wi-Fi sensors by the City of Austin as described in the following section. Discussions held in regard to Task 2017-10 will become the foundation for the work conducted in 2018 tasks if data becomes available.

3.8 Custom Wi-Fi sensors

In previous sections we described the analysis of data collected from Bluetooth sensors to understand travel times along corridors. Similar analyses may be possible using custom-built sensors to detect the unique addresses of portable devices through Wi-Fi. In Task 2017-10 (and other tasks for CoA), CTR explored the use of Raspberry Pi—a low-cost, consumer grade Internet-of-Things (IoT) hardware platform that runs free software. Raspberry Pi is appealing due to its affordability. Where proprietary Bluetooth detectors may cost on the order of thousands of dollars, a complete Raspberry Pi detector costs less than \$100. This low cost comes with caveats: the hardware must be monitored, understood, and reported and has a possible higher failure rate and increased system downtime than what would be expected from more sophisticated, hardened products.

The Wi-Fi detection effort is ultimately intended to characterize travel time between sensor locations by sampling the travel time of individual Wi-Fi enabled devices. Devices are identified by their MAC address, which is read by the custom-designed sensors when the device is within an appropriate range. The goal of this project was to establish a replicable workflow to remotely access sensors deployed on the field and extract/archive collected data. CTR also developed a method to process data from Raspberry Pi sensors in such way that it can be visualized using similar tools as those used for Bluetooth data. The testing of sensor connectivity involved two sensors deployed in South Austin, and one that runs at the CTR office at West Pickle Research Center.

In the final workflow, each unique Wi-Fi address detected on the field-deployed Raspberry Pi platforms is logged on the Wireshark Network Analyzer software that operates on the device. At the end of each day, the logs are collected and transmitted to CTR using appropriate scripts. Further analysis can then be executed to identify where common addresses are paired, travel times are computed, and erroneous pairings are filtered out. The goal is to reduce noise and identify outliers with the intent of producing reliable, consistent travel-time results—measurements that are similar to those produced by the Bluetooth detector scheme.

Chapter 4. Data Workflows

We have developed workflows to retrieve each data type from the City of Austin or other open sources, create a copy on our servers, and push the data into a database for further analysis, visualization, and integration with other data sources. The Bond Corridor Performance Analysis Tool (BCPAT) exemplifies the value of the proposed architecture by providing single-point access to data from multiple sources along selected corridors.

4.1 Existing Workflows and System Architecture

Within Task 2017-10, a preliminary back-end architecture was developed to convey data from the City of Austin to CTR for archiving and further analysis. The emphasis in developing the workflow and system architecture was to become functionally operational within a minimal amount of time while creating opportunities for learning about how to create a more robust, long-term design.

The current ingesters of each data source used for UT research and analysis purposes depend upon files that originate directly from City of Austin equipment. This initially includes Bluetooth data from the file sharing schema set up by the vendor software on the City's network, extractions from Socrata for the Wavetronix data (retrieved from Socrata due to the existing City of Austin open data scheme), and GRIDSMART counts files retrieved directly from GRIDSMART devices. Due to the different methodologies used by different vendors for providing data, the process for retrieving files for each type of data source varies significantly despite basic similarities. In addition to sensor data, we have developed workflows to systematically ingest and update data posted by Capital Metro (Automated Vehicle Location Data (AVL), Automated Passenger Counts (APC) and Generalized Transit Feed Specification (GTFS) data, and to catalog and provide access to manually collected traffic and turning movement counts available along selected corridors.

The "Current" section of Figure 5 shows the flow of data currently in operation for data generated by the City of Austin, and the preliminary system architecture. After gathering files, files are pushed to a CTR archive via file transfer protocols (e.g., secure shell or "SSH," and secure copy or "SCP"). As the cloud-based intermediate database at the City of Austin (hosted by Amazon Web Service's ETL service) becomes more fully adopted, the use of database queries via web-based protocols (e.g., secure hypertext transfer protocol or "HTTPS") and application programming interfaces (APIs) is expected to increase, as shown in the "Future" section.

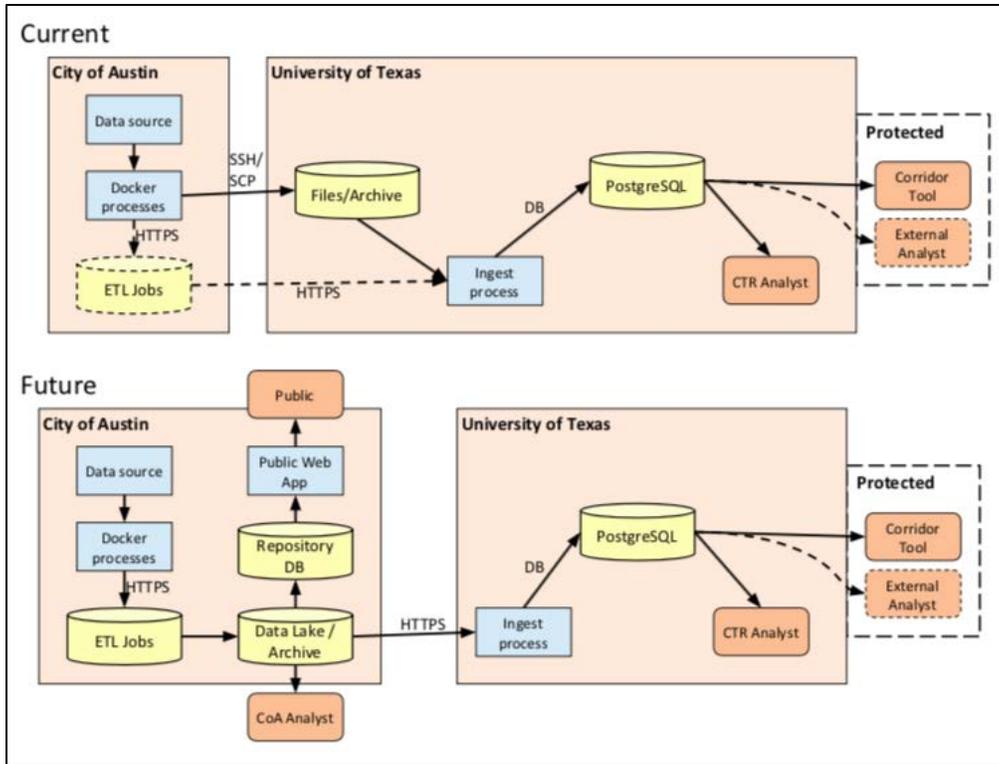


Figure 5: Current and future data flow architectures

Although each data type has different formats and sources, the workflows to make the data available for the tool and other applications are similar: the data is first accessed from its source, pushed in its raw form to CTR servers, and ingested into a database also hosted by CTR where schemas are in place to store and access the data. In order to support the analysis of data along specific corridors, we propose to use a layered database schema that links each data type and its corresponding metadata to corridor intersections and the corridors themselves, as shown in Figure 6.

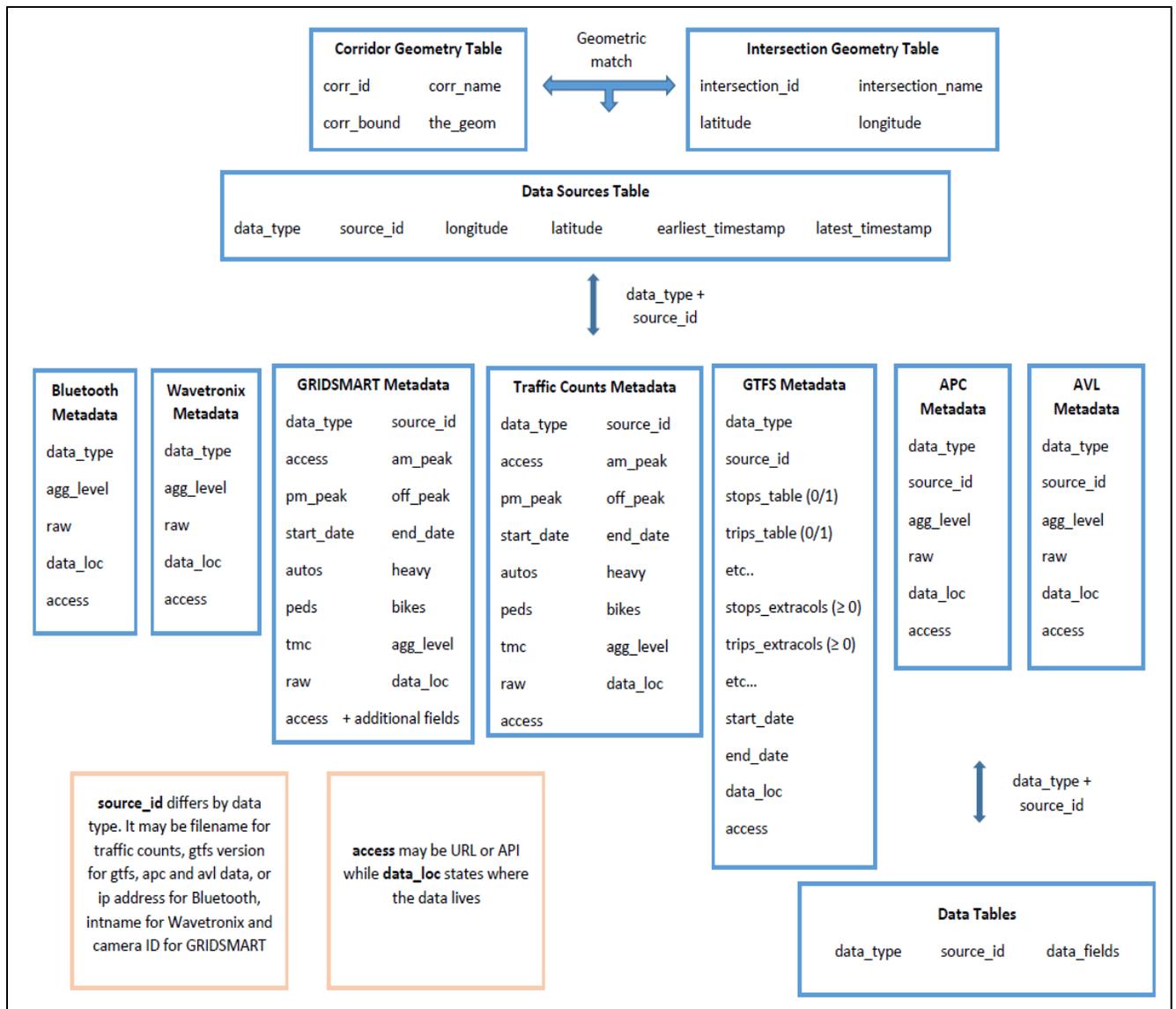


Figure 6. Data Model for City of Austin data types

Corridors and intersections are linked to data sources through the “Data Sources” table, which contains metadata on the location of Bluetooth, Wavetronix, and GRIDSMArt sensors across the city. The table aims to summarize what type of information is available for each intersection or corridor. Next, each data type has a table of metadata associated with it that contains more detail on the specific data type, including the date range for which the data applies, the data aggregation level, and the table name where the data is stored. The final layer is the data itself, which is queryable based on its metadata.

For example, the data model is first lets the application know that Burnet Road has a Wavetronix sensor at Burnet and Palm Way, and that there is a GRIDSMArt sensor at Lamar and Payton Gin Road. Next, the data model is set up to let the application know what type of information is collected by the Wavetronix sensor or GRIDSMArt device, including the range of dates for which data is available and aggregation level of the data. Finally, the metadata layer provides the application with the table name where the real

data is located so to retrieve the data based on metadata filters. Future work on the data model includes refining a versioning scheme as well as incorporating transit data into the model.

The following sections describe the workflows in place to ingest each of the considered data types into CTR's systems and database.

4.1.1 Bluetooth Sensors

Bluetooth sensor data are stored in files within the City of Austin network; a nightly process copies the files to CTR and CTR accesses the data through a file share system that requires VPN login. From there, an ingester process loads the data into a PostgreSQL database where a table schema is used for ease of querying.

4.1.2 Wavetronix Sensors

A script queries the Socrata service nightly to retrieve the available Wavetronix data, which is stored in CSV files on CTR servers. Files are named according to the date of retrieval. An ingester process then loads the tabular data into the analysis database.

4.1.3 GRIDSMART Cameras

Each GRIDSMART device contains records that are accessible through an API. CTR developed scripts that run on City of Austin systems and interact with GRIDSMART devices to extract historic counts data. Data is then further processed for its ingestion into the CTR database. Because of the large number of individual vehicle records, counts are currently aggregated in the database ingestion process to fifteen-minute increments and categorized according to passenger car or truck. This is similar to how manual counts are often recorded.

GRIDSMART devices archive a variety of information aside from turning movement counts, including speeds, zone definitions (regions within the fisheye camera field of view that correspond with specific turning movements), real-time activity, traffic light status, and approach type. Future work will support the archival and processing of additional data sources.

4.1.4 Automated Vehicle Location Data

Capital Metro streams AVL data in real-time and CTR accesses the data stream and stores each new post in a temporary local file. Data files are later moved to an archive folder from which they may be ingested into the analysis database.

4.1.5 Automated Passenger Counts

APC data is published sporadically throughout the year by Capital Metro. Data is accessed through the Austin City Data Portal, downloaded through Socrata, and manually ingested into the database on CTR servers. Scripts are in development for its ingestion in the analysis database.

Figure 7 illustrates total daily boardings and alightings along the Bond Corridors for the months of June 2016 and June 2017. The boarding and alighting values are not normalized by the number of stops along the corridor, so the comparison across corridors is difficult because all corridors do not have the same number of stops. However, evaluation within a corridor is likely to be useful. When comparing June 2016 to June 2017, all corridors showcase increased volume from 2016 to 2017.

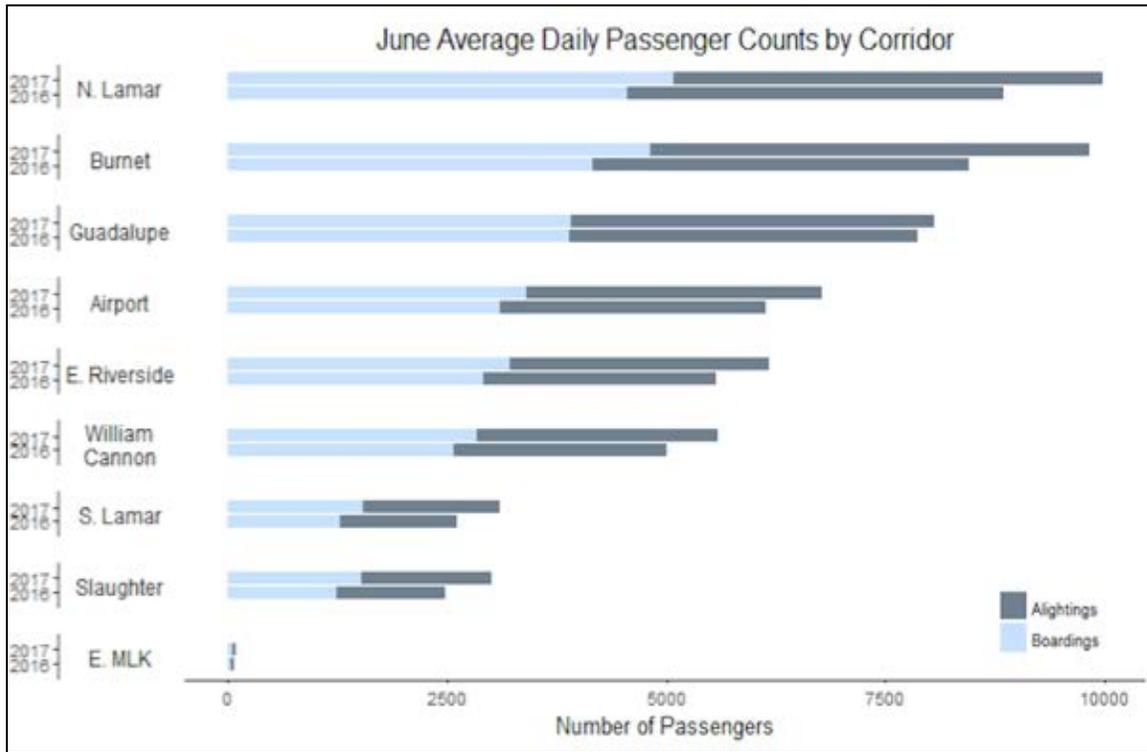


Figure 7. Average daily passenger counts per corridor

Similarly, Figure 8 denotes the average daily dwell time by corridor for the months of June 2016 and June 2017. Again, comparing one corridor from one year to the next is likely to be more useful than comparing corridor to corridor. Average dwell time tends to be greater in June 2016 than in June 2017 for most corridors except for E. Martin Luther King Boulevard where in 2017 there was a sharp increase from the year before. Further exploration on the causes of the increase is necessary to determine the validity of this value.

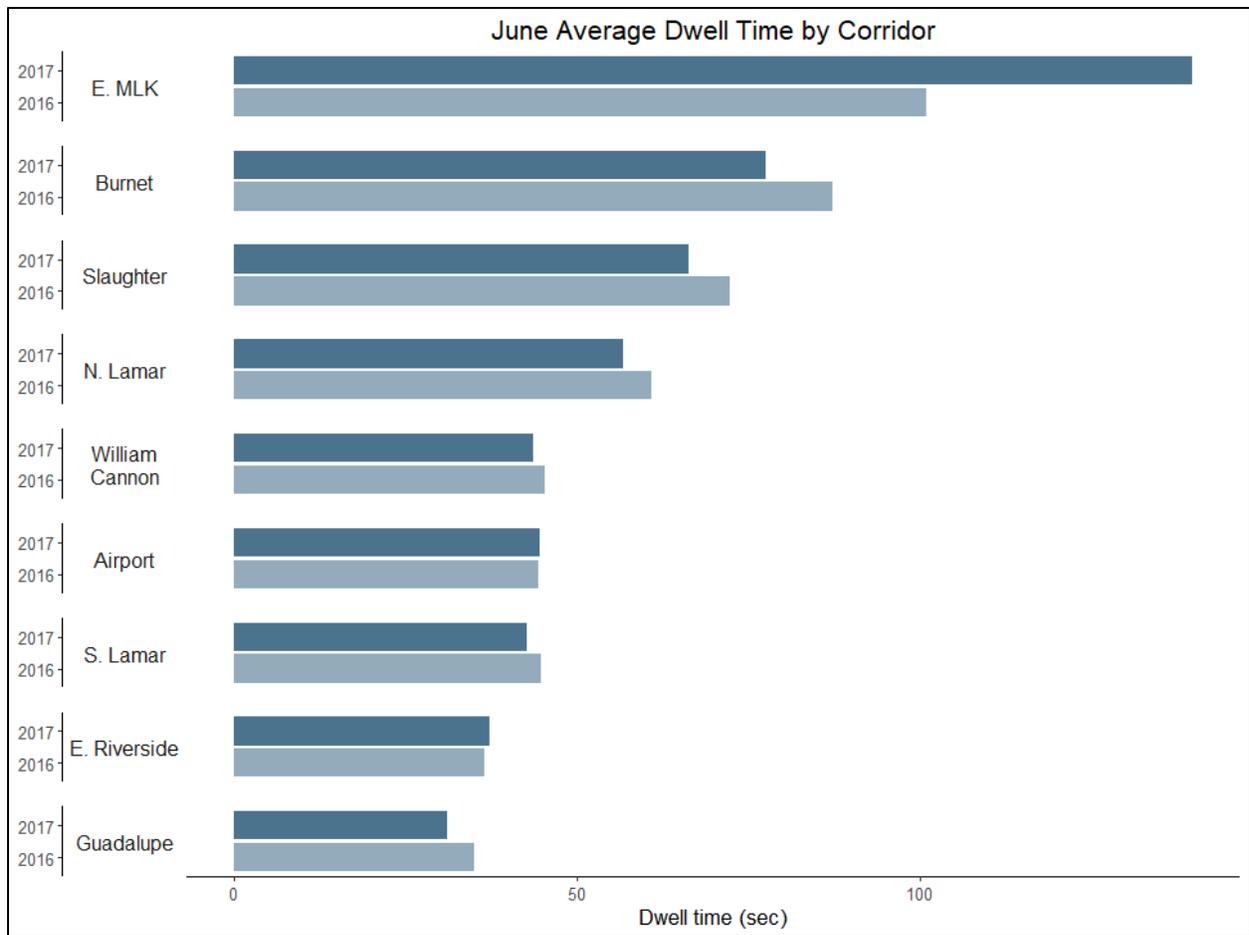


Figure 8. Average dwell time per corridor (June 2016)

4.1.6 INRIX Waypoint Data

Each month of waypoint data available to the City of Austin was provided in multiple files by INRIX through a dedicated website. The City of Austin developed scripts to ingest the data into a Postgres database hosted in Amazon Web Services. CTR replicated the database on its servers and built a web-based application to support the analysis of travel times between selected locations (Figures 9 and 10).

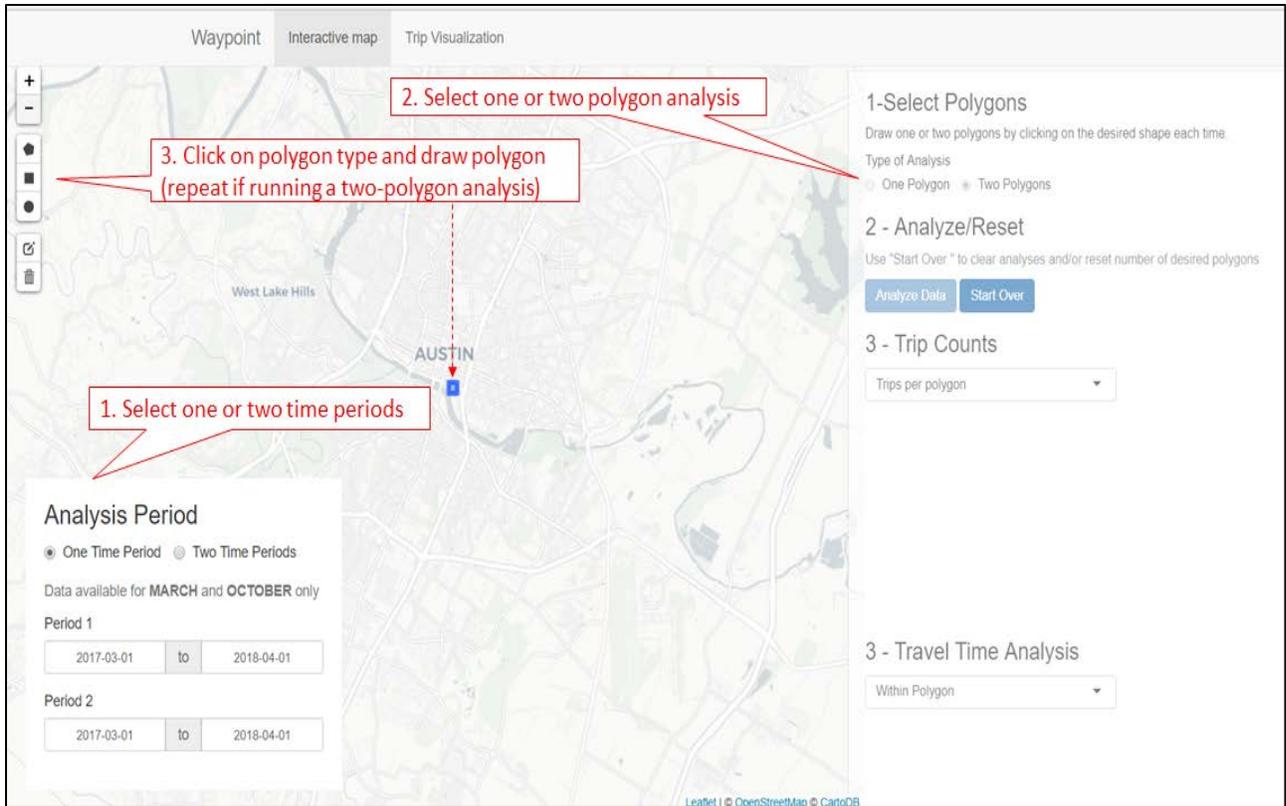


Figure 9. Waypoint analysis web application: analysis setup.

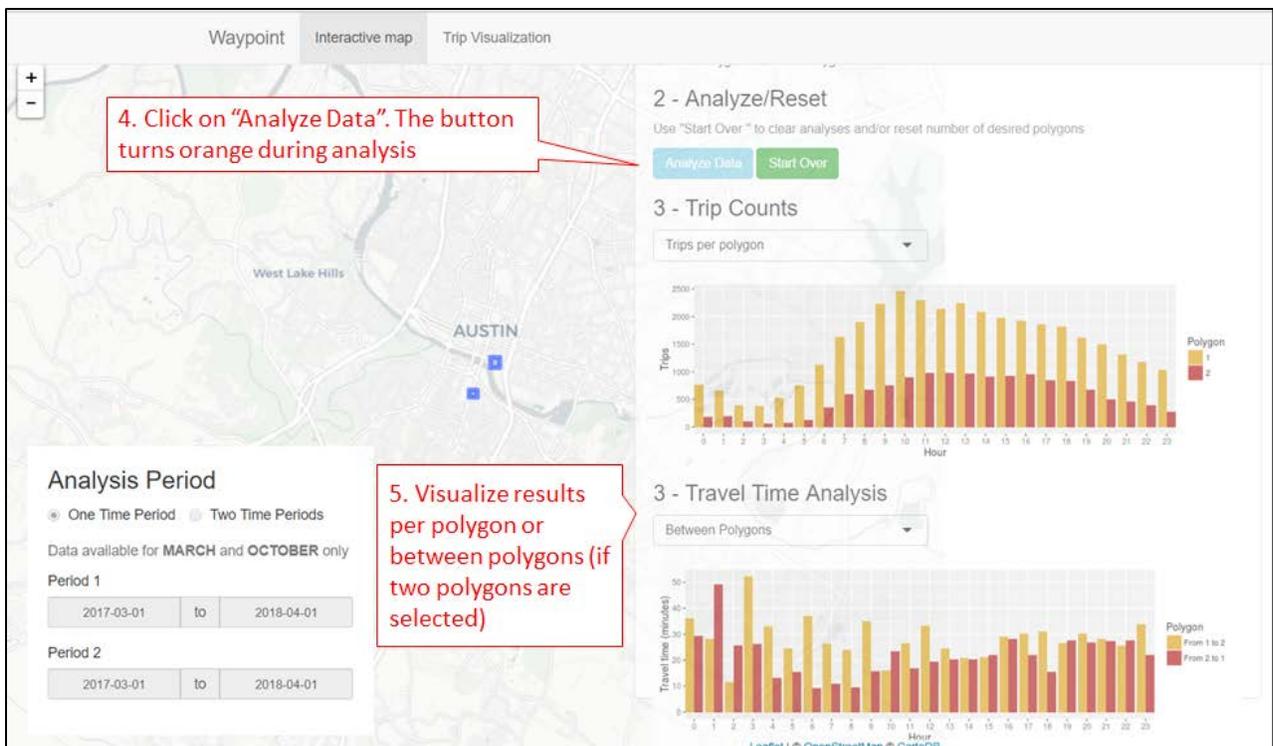


Figure 10. Waypoint analysis web application: travel time and trip counts by time of day.

The web application enables users to draw one or two polygons on a map and retrieve all trips within the polygon boundaries. When two polygons are selected, the tool computes average travel times, in both directions, between the polygons (as shown in Figure 10). It also enables users to visualize all relevant trips (Figure 4). The time period for which travel times are analyzed is a user input, and the application allows for comparing two time periods (Figure 9).

4.2 Future Workflows

In the future architecture, shown in the "Future" section of Figure 5, the use of files and archiving at CTR for City of Austin data will be phased out in favor of City of Austin hosting its own archives. The City of Austin's intermediate database (hosted by Amazon Web Service's ETL service) will be fully utilized, while data sources from other sources (e.g., Capital Metro AVL) may be archived on CTR resources.

While it remains to be seen if the data leveraged for analyses in applications like the Bond Corridor Performance Analysis Tool (BCPAT) will appear as a data lake—a large data storage and processing engine—at the City of Austin, or if CTR will interact with the ETL database, the analysis path remains the same. The benefit in CTR maintaining its own analysis database is to allow for research and experimentation to occur with maximum versatility. Meanwhile, because of the desire for data to be permanently hosted in the care of City of Austin, public access will also be most likely facilitated through City of Austin systems.

Chapter 5. Data Processing and Analysis

Data in its raw form often needs processing in order to extract valuable metrics and draw insights. For example, a single unique Bluetooth identifier with a date and time stamp is insufficient to derive corridor travel time. Two time stamps with the same unique identifier along a corridor are necessary to assess travel time. Data processing often involves a data quality assessment/cleaning step, which may be specific to the intended use of data. As an example, the treatment of missing data points is important when considering volume estimates from Wavetronix data. When Bluetooth data is used for vehicular corridor-level travel time estimation it is critical to identify and eliminate outliers, which may correspond to non-motorized transportation modes or vehicles not traveling on the corridor and could bias average travel time estimates.

The following sections further discuss currently implemented methods for data processing, cleaning, and analysis of considered data sources. Future work will further assess data quality and extend processing and analysis methodologies.

5.1 Bluetooth

Bluetooth sensor data consists of MAC addresses and the corresponding timestamps at which they are observed at any available Bluetooth sensor. Such sensors are identified by their unique IP addresses. Travel times between two sensors are estimated by finding matching MAC addresses across sensors and comparing corresponding time stamps. For the purpose of estimating travel times on corridors, consecutive sensors are first identified and sequenced. A data cleaning process is used to ensure that the considered travel times correspond to vehicles traveling on the corridor, and not to pedestrians or to vehicles that make a stop within the corridor or follow a different path between sensors. The following sections describe these processes, along with the final travel time estimation method. Finally a case study illustrates the application of the proposed methodology.

5.1.1 Sensor Sequencing

In order to facilitate the estimation of corridor travel times using Bluetooth sensor data it is necessary to map sensors to corridors and assign a sequence number that defines the order of the sensors in each travel direction. The distance between sensors is also computed and used later for data cleaning purposes.

To determine the sensor sequence along the corridor and thus directionality, the distance between “the beginning” of the corridor—as established by the corridor’s spatial geometry representation—and each sensor location is computed and sorted in ascending order. Directionality plays an important role when computing the average travel time using the “four-point” algorithm, outlined below.

5.1.2 Data Cleaning

Two layers of “cleaning” are implemented when estimating average travel times on a corridor between selected sensors. The first one is a simple threshold that eliminates

values considered too high under any likely traffic condition, and may reflect a stop within the corridor, the use of a different path, or the presence of a non-recurrent event, such as an incident. The second pass considers the median travel time by time of day at a pre-specified time step (e.g., fifteen minutes), and eliminates data points that are “too far” from the median using the median absolute deviation (MAD) technique. Due to its resiliency against outliers, MAD is a more robust measure of variability than the use of standard deviation (which measures the distance from the average value). It ensures data integrity, consistency, and thus the reliability of calculated average travel time along the corridor segments.

Ideally, average travel time for a corridor would be estimated by using the two-point algorithm for the first and last road segments, and the four-point method for the middle segments. However, the four-point method is substituted for the two-point algorithm when data is limited. For those periods without any data (whether by two-point method or four-point method) for specific road segment, default average travel time (the distance of the road segment divided by the posted speed limit) will be used to fill in the missing data. Once travel times are computed by segment, the average corridor travel time is factored by the summation of segment travel times.

5.1.3 Travel Time Estimation

The approach implemented by CTR to estimate end-to-end corridor travel time involves the following assumptions:

- **Spatial discretization.** For each corridor, CTR defines sub-segments between consecutive sensors. This approach allows for the computation of end-to-end travel times in long corridors where very few devices travel the entire length of the facility. Data cleaning and aggregation is performed at the sub-segment level using a pre-defined fifteen-minute time-step.
- **Instantaneous spatial aggregation.** Given the sub-segment approach described above, the total corridor travel time at any selected time-step is computed by adding the travel time of all sub-segments at such time-step. The approach may lead to estimation errors in long corridors, since vehicles starting their trip during one time-step may reach the last segments of the corridors at a later time-step, for which traffic conditions may be different. This is a limitation of many travel time estimation tools, and may be addressed by future work.
- **Through-traffic only.** Bluetooth sensors capture the presence of devices at an intersection, but cannot identify whether the device is entering/exiting the corridor by turning right or left, or passing through. Devices that are performing a turning movement when captured by either of the sensors used to compute their travel time may register longer travel times than those moving through the corridor at both sensor locations. Given that the traditional approach to measuring travel times involves using probe-vehicles that run on the corridor from end to end, CTR’s method includes an additional layer of data filtering (performed before the data cleaning) seeking to capture the travel time of

through traffic only. If we let I denote the position (sequence) of a sensor on a corridor, for each pair of consecutive sensors $(i, i+1)$, the filtering approach identifies “through” traffic by considering only devices that are also seen in $i-1$ and $i+2$ (if these are part of the corridor). This approach (four-point method) reduces the noise and bias introduced by turning movements, but also reduces the availability of data. If data availability is a concern, the additional filtering is not implemented.

The travel time estimation process begins with the selection of all MAC addresses captured by sensors along the desired corridor during the desired time period. Once the data is appropriately filtered and cleaned as described above, average travel time values per fifteen-minute intervals are computed at the sub-segment level, and added up to produce a single corridor-level travel time estimate.

5.1.4 Burnet Road Case Study

We selected several corridors to use in testing the model, including Burnet Road, South Lamar, and North Lamar. The Burnet Road results will be used here to show the performance of the model. Six Bluetooth sensors along the Burnet Road define five road segments. The travel time of the first and last road segments was obtained by using the two-point method; the average travel time of the middle three segments was obtained by using the four-point method. Results from weekdays were used to plot the corridor travel time distribution of the two directions.

Figure 11 shows travel time in minutes along the corridor by time of day. The different colors indicate the average travel time of its corresponding segment. The total height of each bar denotes the average travel time for a particular time of day.

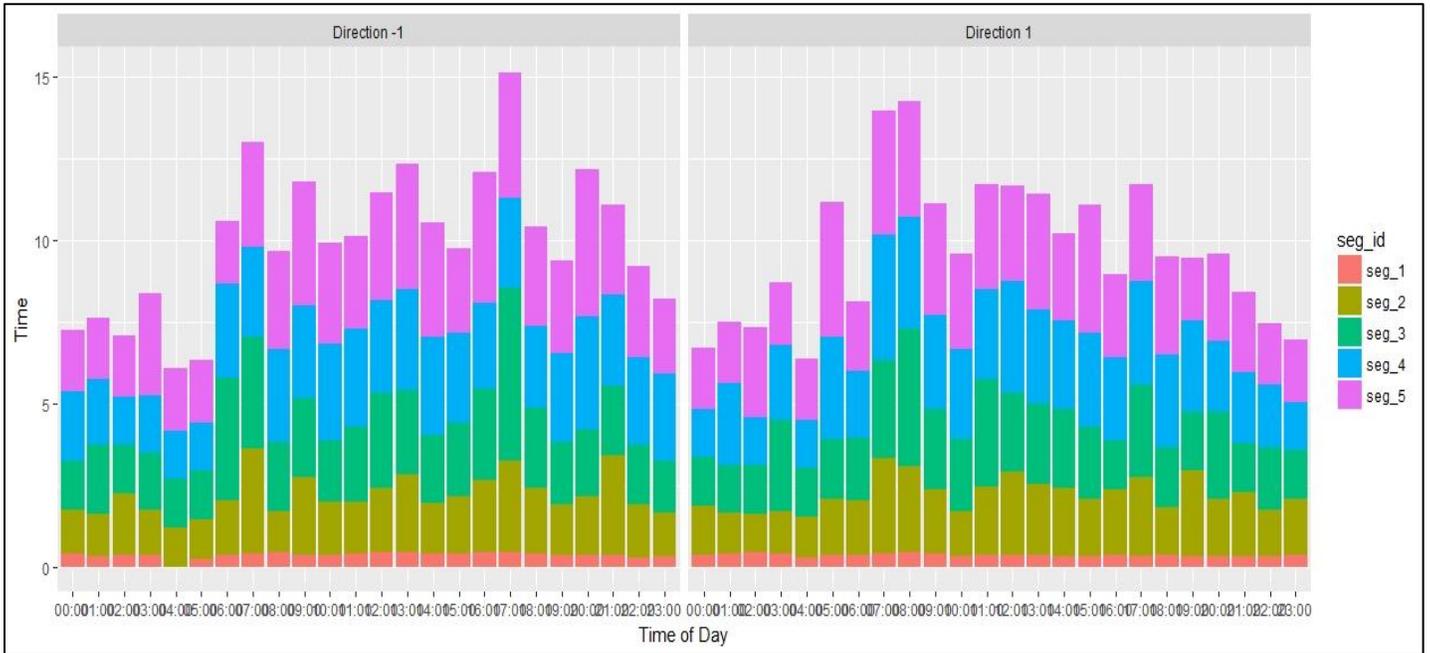


Figure 11. Average travel time distribution of Burnet Road

Note: “Direction -1” in Figure 11, showing that the travel time at 05:00 for segment 5 is zero minutes, is a clear error. The error stems from only one record with a travel time of zero minutes used in the calculation. Thus, the MAD method used to rule out the outliers was insufficient for this situation; a comprehensive filter method is required to achieve better results. For example, a minimum travel time could be assigned to each segment based on its length and any values below the minimum threshold would be filtered.

By leveraging the traffic data collected by the Bluetooth sensors, the details of average travel times along a corridor and the distribution of travel time by time of day may be estimated. The travel time information allows the researchers and city planners to understand the corridor performance and provides the opportunity to improve the traveler experience. For future work, additional sources of average travel time will be compared to test the accuracy of the model, outlier detection will be fine-tuned, and more data from the Bluetooth sensors will be used to test the performance of the model.

5.2 Wavetronix

CTR implements a running-sum approach to compute hourly volumes and average speeds every fifteen minutes using Wavetronix data. At any given time, the computed hourly volumes represent the total number of vehicles observed in the one-hour window that starts half hour before the selected time, and ends half hour after. The same time window is used to estimate average speed values. Computing the running sum and average of the data allows the hourly values to smooth out short-term fluctuations.

While the computation of the metrics described above is fairly straightforward, data cleaning considerations are important and will be the focus of further work, as described below. Finally a case study illustrates the implementation of the techniques described in this section.

5.2.1 Data Cleaning

Wavetronix volume and speed datasets may have periods of missing data or erroneous readings. Missing data by period may vary per sensor and cover both short- and long-term intervals. Short intervals may include a few hours of missing counts while longer time periods may be up to several days without information. Figure 12 provides an estimation of the percentage of correct fifteen-minute data counts per day at three different intersections: Lamar and Shoal Creek, Lamar and Manchaca, and Loop 360 and Lakewood. There are periods in which none of the three sensors are functioning, but there are also days for which the number of observations are greater than what was expected for a fifteen-minute interval rate within a day. This occurs, for example, at the 13:15 and 13:17 readings.

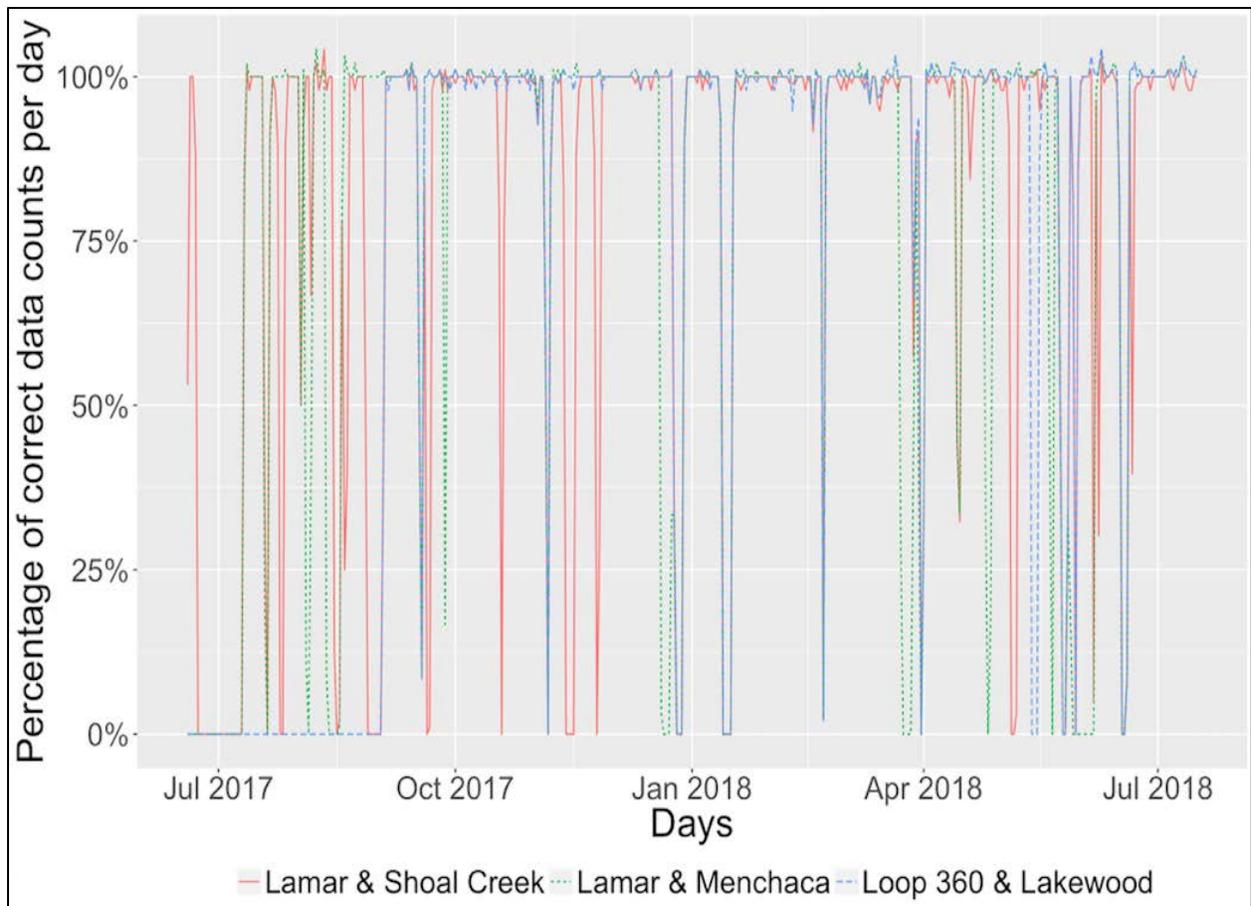


Figure 12. Example of data availability by location (NB direction only)

The data also shows extreme speed values, which are likely erroneous and may bias the analysis. For example, during the early morning hours, some sensors register high-speed values, upwards of 100 MPH in road segments with 35 MPH speed limits. Figure 13 illustrates an example of the average speed data during a week; we can observe an extreme measure during Thursday morning (approximately 4 AM).

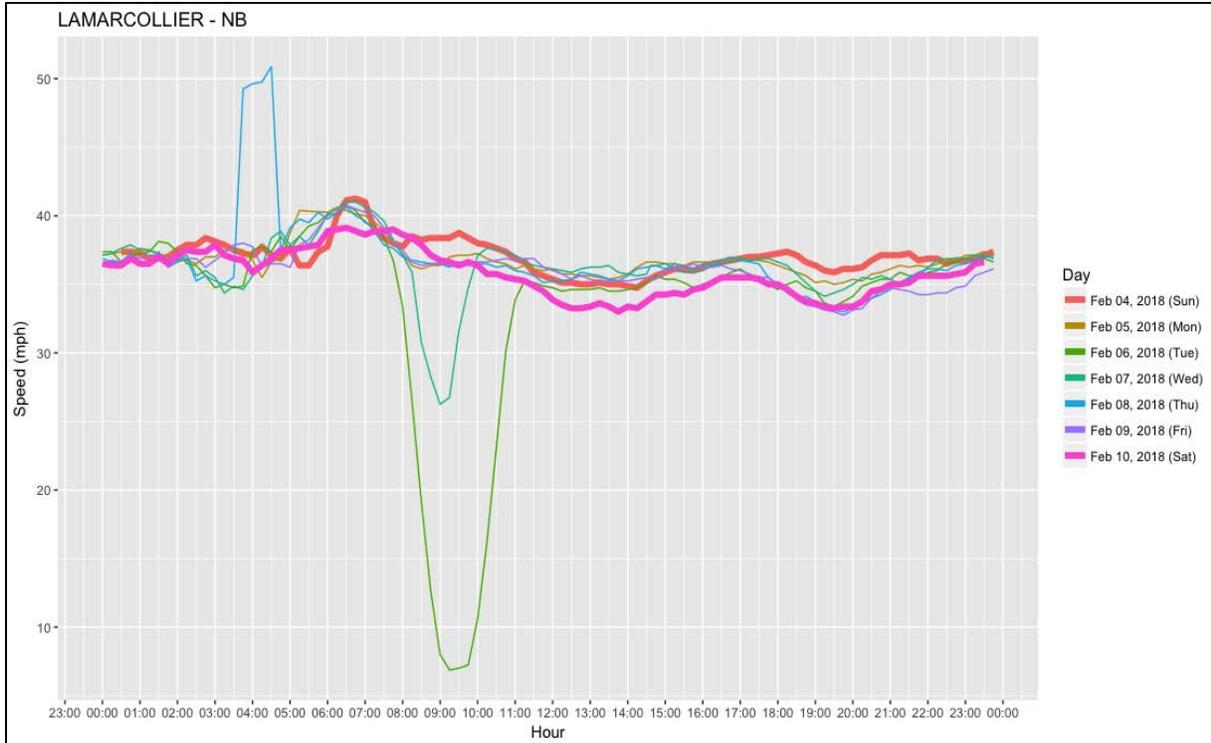


Figure 13. Example of hourly average speed per lane

Cleaning procedures for short-term missing data consists of using a Kalman Filter;³ for longer periods, other techniques are being assessed, such as interpolating speed and volume values based on historical data. Speed values greater than 60 MPH in an intersection are replaced by the hourly average. Lastly, any extra readings are averaged to the nearest fifteen-minute interval, so that a reading at 13:17 would be included and averaged with a reading at 13:15.

5.2.2 Lamar Collier Case Study

Figure 14 shows an example of hourly volume (top) and speed comparison (bottom) for the intersection of Lamar Collier Northbound for two periods. Period 1 corresponds to June 18 through June 23, 2017 while Period 2 corresponds to March 12 through March 16, 2018. Only weekdays are considered in the analysis, and the South by Southwest (SXSW) festival occurred during Period 2.

In Figure 14, the vehicles-per-hour volume and miles-per-hour speed are plotted by hour. The red trend line corresponds to Period 1 and the blue line to the second period, which when combined provide a direct comparison of the two periods.

³ Kalman filtering or linear quadratic estimation (LQE) is an algorithm that uses a series of measurements observed over time, containing statistical noise and other inaccuracies, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone, by estimating a joint probability distribution over the variables for each timeframe.

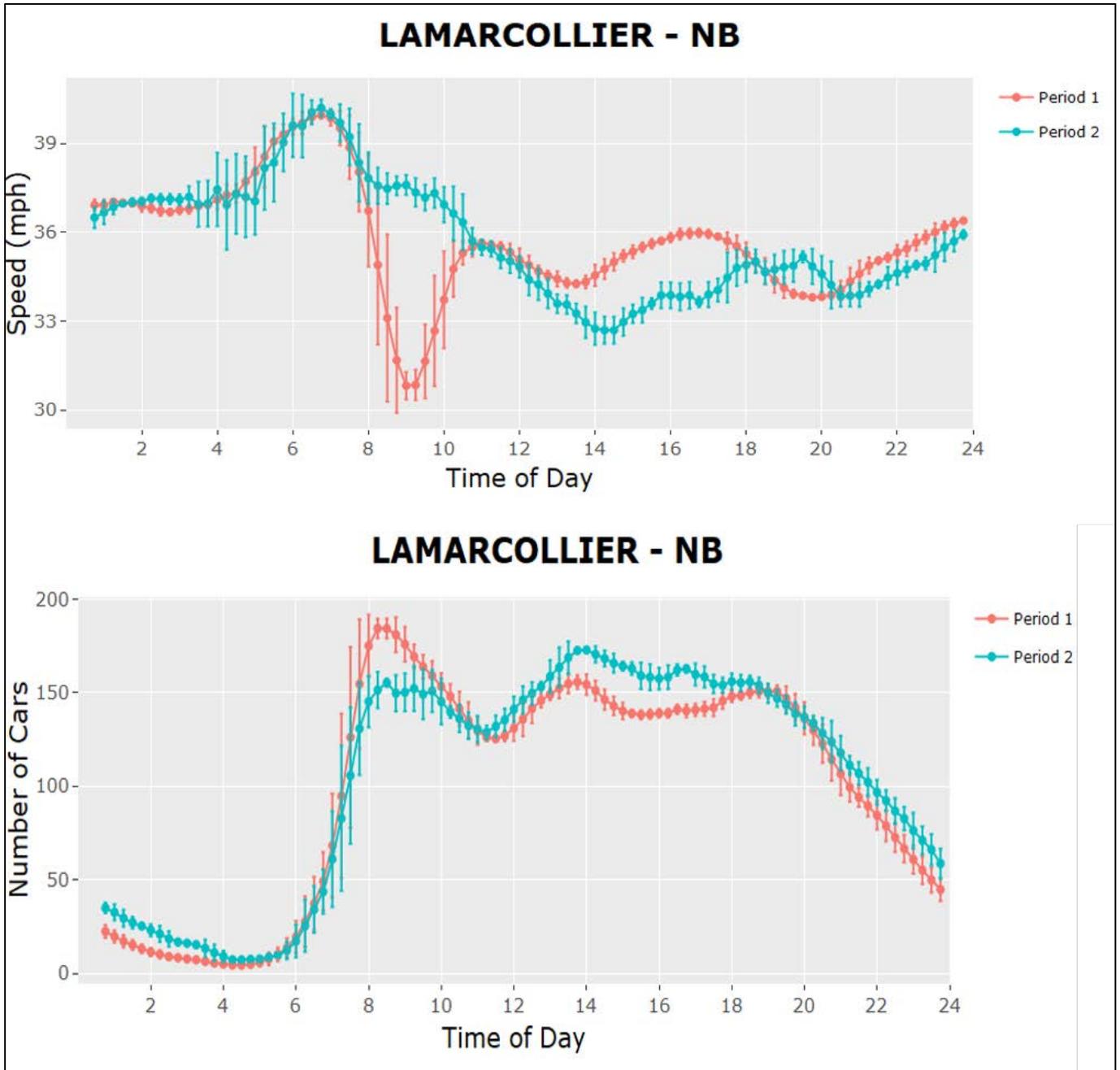


Figure 14. Wavetronix results hourly volume (top) and average speed (bottom)

In terms of average volume, the two periods follow a similar trend. However, while average speed for Period 1 follows the expected weekday trend—speed drops in the AM peak and PM peak hours—Period 2, the week of SXSW, does not illustrate the same troughs. Data from Wavetronix sensors provides detailed information on the volume and speed of passing vehicles, which may be leveraged to grasp corridor performance and enhance traveler experience.

5.3 GRIDSMART Cameras

Each GRIDSMART camera logs the speed, approach type, turning movement type, and vehicle classification of each vehicle detected. This data is aggregated to fifteen-minute intervals, providing the average speed and total volume for that time period. The aim of future GRIDSMART analyses is to mirror manual counts that provide critical traffic volume information on approach type, turning movement, and vehicle classification, as well as average speed. The GRIDSMART volume data will be included in the traffic volume tab in the BCPAT and be downloadable as a CSV file. Average speed data will be included in the speed tab and complement the Wavetronix speed data to create a more robust metric.

5.4 Automated Vehicle Location Data

AVL data records vehicle location (longitude and latitude), spot speed, and corresponding timestamp. Data is provided approximately every two minutes. In the context of corridor performance, the analysis of the data can provide insights on the average speed of traffic vehicles on the corridor. The reported speed is not necessarily representative of average values on the corridor because at any time the bus may be accelerating or decelerating. Similarly, the lack of regularity in the locations at which time stamps are provided makes travel time comparisons less meaningful.

CTR developed a method to estimate average corridor speeds based on available data by considering, for each trip, the first and last GPS point reported on the corridor, and estimating the corresponding travel time. The distance between such points along the corridor is used to derive an average speed value. This method of approximating transit speed includes delays at intersections and dwell time. Currently, methods to approximate delays are being evaluated by combining other transit datasets such as General Transit Feed Specification (GTFS) data and APC, thus providing a more accurate transit travel speed estimation. Further work will also consider potential errors/bias introduced by averaging the speed values of vehicles that traverse different portion of the corridors.

Figure 15 illustrates a proposed method on the corridors for which sufficient information was available in the analyzed time period: Burnet Road (corridor 2), N. Lamar Boulevard (corridor 3), Riverside Drive (corridor 4), S. Lamar Boulevard (corridor 6), Guadalupe Street (corridor 7), and Slaughter Lane (corridor 9). As shown in the figure, the approximate average levels of speed in MPH vary across corridors, but the speed trend throughout the day generally follows the expected pattern of dips in the morning and evening rush hours. Although the transit speed pattern follows expected trends, the actual values of speed include dwell time and stops at intersections, which creates a systematic downward bias in the values. Supplementing AVL data with APC data—specifically dwell time—could potentially lessen the bias.

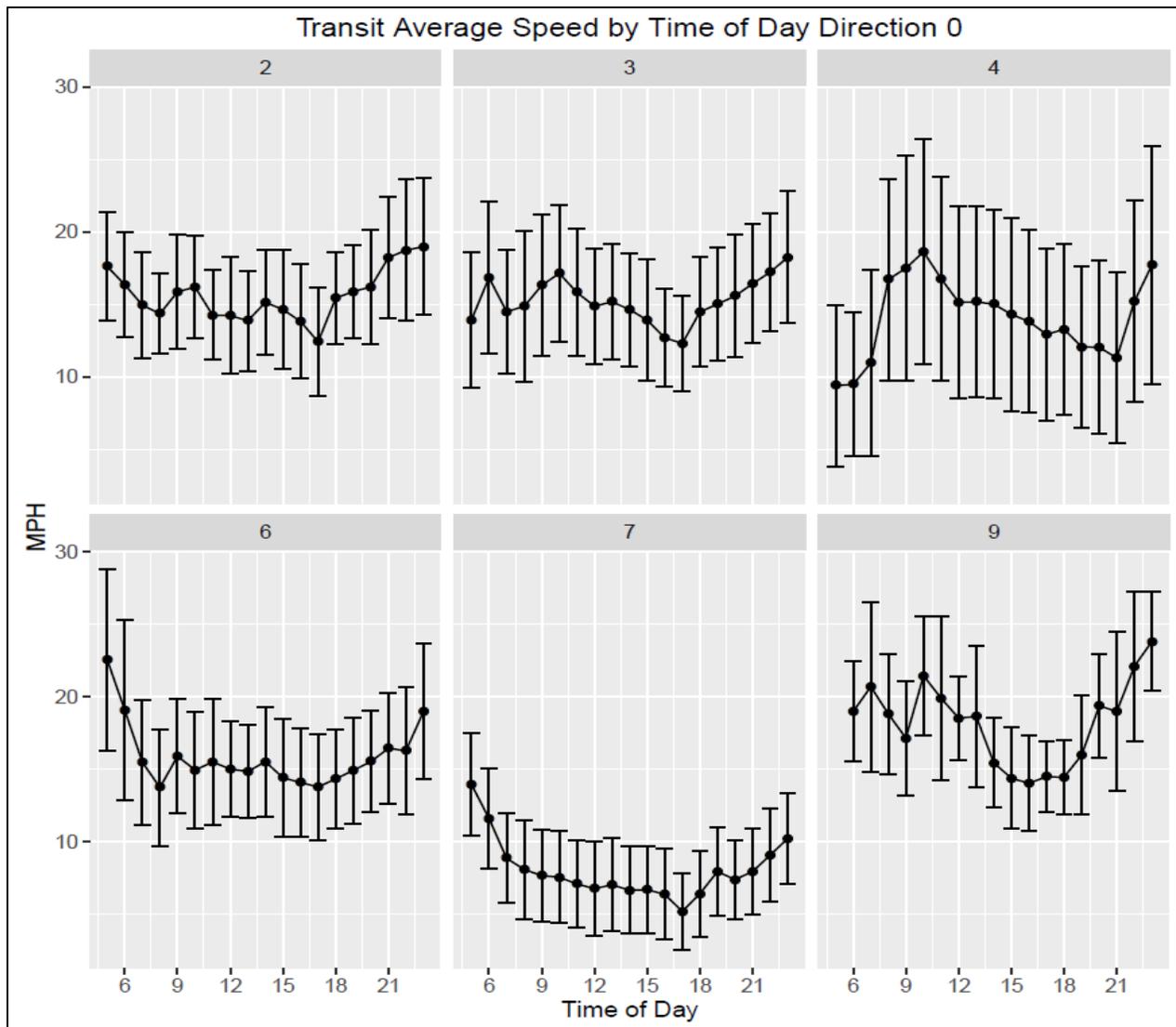


Figure 15. Transit speed by time of day on six corridors

Next steps when considering the APC and AVL data sets include boarding and alighting normalization by number of corridors stops; outlier detection methodology for both APC and AVL datasets; and, combining APC and AVL information to gather more accurate measures of transit speed. The estimate of the occupancy percentage per corridor using APC data is still under review.

5.5 Automated Passenger Counts

The APC dataset obtained through the City of Austin online data portal includes transit information on vehicle location, number of boardings, number of alightings, dwell time, timestamp, and trip identifier for all bus vehicles and stops. For this analysis, the APC information was filtered to include only those routes corresponding to the corridors under evaluation.

CTR has analyzed a data sample to understand data characteristics and design an efficient analysis approach. Figure 16 presents the distribution of boarding, alightings and dwell times, which may be used in the future to detect and eliminate outliers.

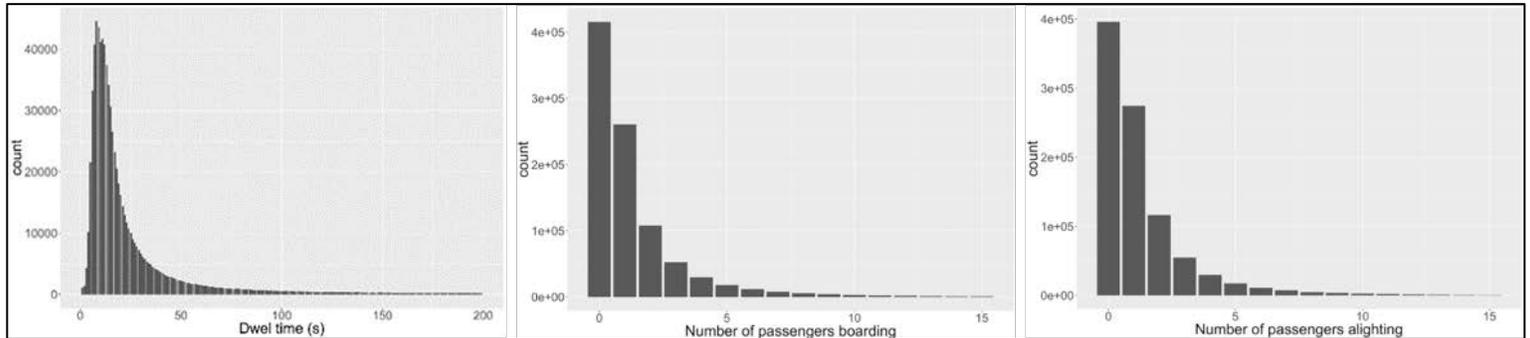


Figure 16. Histograms for (a) dwell time, (b) boarding count, and (c) alighting count

The information provided in the APC data allows for a valuable route-level performance assessment. For instance, the dwell time provides information on the delays at every stop. Also, the boarding and alighting information and the vehicle capacity enable estimation of the bus occupancy percentage, which is a variable used for estimating transit level of service. In order to obtain occupancy, it is necessary to process the information from each individual bus trip so that we can approximate the number of passengers on board. This process is not available yet but it will be included in the work for the next fiscal year.

5.6 INRIX Waypoint Data

INRIX waypoint data was ingested into a Postgres database. For the estimation of travel times between two polygons, processing consisted of identifying all GPS points within each polygon and corresponding trip IDs. The operation is performed on-demand for the desired polygon using built-in geographic analysis capabilities (PostGIS). Once relevant points and trip IDs are identified, two types of analyses are conducted:

- Average number of distinct trip IDs observed at each (and both) polygons by time of day during the analyzed period(s).
- Average travel time between selected polygons, by time of day, during the analyzed period(s). Travel time is estimated in both directions of travel by considering the latest timestamp in the origin polygon, and the earliest timestamp in the destination polygon.

The travel-time estimation methodology is very simple, and the reliability of the resulting values depends on the size of the selected polygons. Further, statistical cleaning to eliminate outliers may be required but has not been implemented in the web application.

5.6.1 Manor Road Case Study

The processing workflows described above were used to analyze changes in travel times on Manor Road between two two-week periods of interest to the City of Austin, one in October 2017 and one in March 2018. The corridor is approximately five miles long, but

given that no trips were identified at both ends of the corridors, CTR defined sub-segments for this analysis. Figure 17 presents the corridor, sub-segments, and corresponding number of trips during the two weeks of October considered in this study. Similar data availability was observed for the March 2018 time period.



Figure 17: Manor Road corridor, sub-segments and corresponding number of trips (October 2017)

Travel time estimates by hour for each segment were combined to produce the end-to-end corridor travel time values for each time period (Table 3). The reported standard deviations were computed based on the standard deviation of the data on each sub-segment.

Table 3: Total travel time estimates for Manor Road corridor

	Eastbound				Westbound			
	Oct-17		Mar-18		Oct-17		Mar-18	
	Average	Deviation	Average	Deviation	Average	Deviation	Average	Deviation
7am			11	0.63	11	0.63	13	1.62
8am	13	0.74						
9am	13	1.72						
10am			14	0.63	14	0.63		
11am			13	2.44	13	2.44		
12pm	13	1.43					13	2.15
1pm	12	1.15						
2pm	14	1.10						
3pm	13	1.61	13	1.89	13	1.89		
4pm	13	1.35	14	1.85	14	1.85	12	1.68
5pm	13	0.92	14	1.81	14	1.81		
6pm	13	2.30						

Estimates were produced for segments only where more than three observations were available during the considered time period, resulting in missing data throughout the day. Further, the standard deviation of the data is fairly high when compared to the difference in corresponding average values, suggesting that the comparison of means across time periods may not be significant. For this application, waypoint data was insufficient to analyze total corridor travel time during the selected time periods. CTR shared the study results with representatives from INRIX who agreed with the conclusion.

Due to the need to divide the corridor into multiple segments, the same analysis was not performed using the web application. Polygon information entered through the web interface was stored in a data table, and queries were run manually in the database to extract corresponding travel times; the travel times were later aggregated in a spreadsheet. While the integration of multi-segment analysis into the web application is fairly straightforward, the seemingly sparse nature of available data makes the results unreliable. The City of Austin therefore recommended discontinuing the effort during this fiscal year. However, having manipulated waypoint data in a database environment with the support of a web application for visualization proved to be an efficient methodological approach. The framework is available to be used when new potential uses of the waypoint data are proposed.

Chapter 6. Bond Corridor Performance Analysis Tool

CTR leveraged the implemented data workflows, processing, and analysis techniques described in previous sections to develop the Bond Corridor Performance Analysis Tool (BCPAT) to support corridor performance analysis over time, focusing on facilities identified for improvement under the Mobility Bond. Specifically, the BCPAT leverages Python3, PostGIS, R, and the R Shiny package. Python3 performs back-end workflows to extract data from the City of Austin network or online data resources such as Socrata, perform data pre-processing (e.g., data aggregation prior to ingestion), and perform data ingestion into a PostGIS database. In turn, PostGIS allows efficient data storage and querying, which then feeds into R and its Shiny extension for data analysis and visualization. R Shiny connects and queries the PostGIS database for corridor information, sensor locations, traffic volume reports, Bluetooth, Wavetronix, and APC data. It also enables the development of the interactive user interface.

The BCPAT interface includes a home page (Figure 18) featuring a table, a map that summarizes corridor data availability, and varying tabs that correspond to travel time, traffic volume, speed, and transit analyses. The user can navigate between analyses by clicking on the tab of the desired analysis. The map on the right side of the interface remains in place throughout the user interaction. The spatial information showcased on the map changes depending on the type of analysis selected on the left side so that, for example, only Bluetooth sensors are shown when the user is on the travel time tab, or only Wavetronix sensor locations are visible when the user navigates to the speed tab. Currently, information from Bluetooth and Wavetronix sensors, traffic studies, and APC transit data are incorporated into the tool. Future work will incorporate GRIDSMART sensor data and, potentially, AVL transit information.

City of Austin Bond Corridor Tool

Corridor Summary Travel Time Traffic Volume Speed Transit

ID	Name	Bluetooth Locations	Wavetronix Locations	Bus Routes	Count Locations	GRIDSMA RT Locations
1	Airport	5		24	461	
2	Burnet	6	2	14	139	
3	N. Lamar	4		15	124	2
4	E. Riverside	5		19	44	1
5	E. MLK	2		1	610	
6	S. Lamar	9	4	7	145	
7	Guadalupe	6		22		
8	William Cannon	7		15	2	3
9	Slaughter	2		8	3	

Previous 1 Next

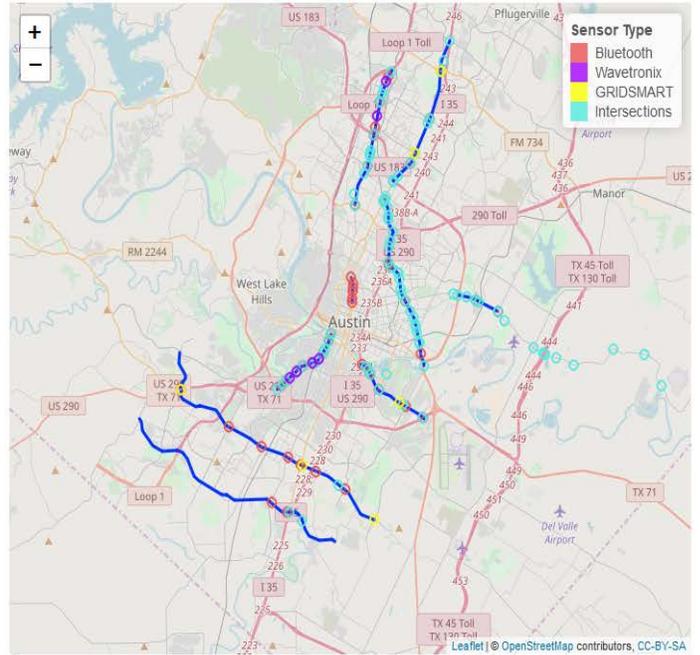


Figure 18. Bond Corridor Performance Analysis Tool homepage

Travel time analysis from Bluetooth sensor information facilitates period comparisons by direction for a chosen corridor. The web tool allows the user to choose a corridor and compare two date ranges. From the user selection, the application provides a visualization that summarizes the median travel time by hour and distinguishes between direction and time period. When applicable, the travel time algorithm implements the segment-by-segment analysis, as shown for Burnet Road in Figure 19, and displays segment travel time as well.

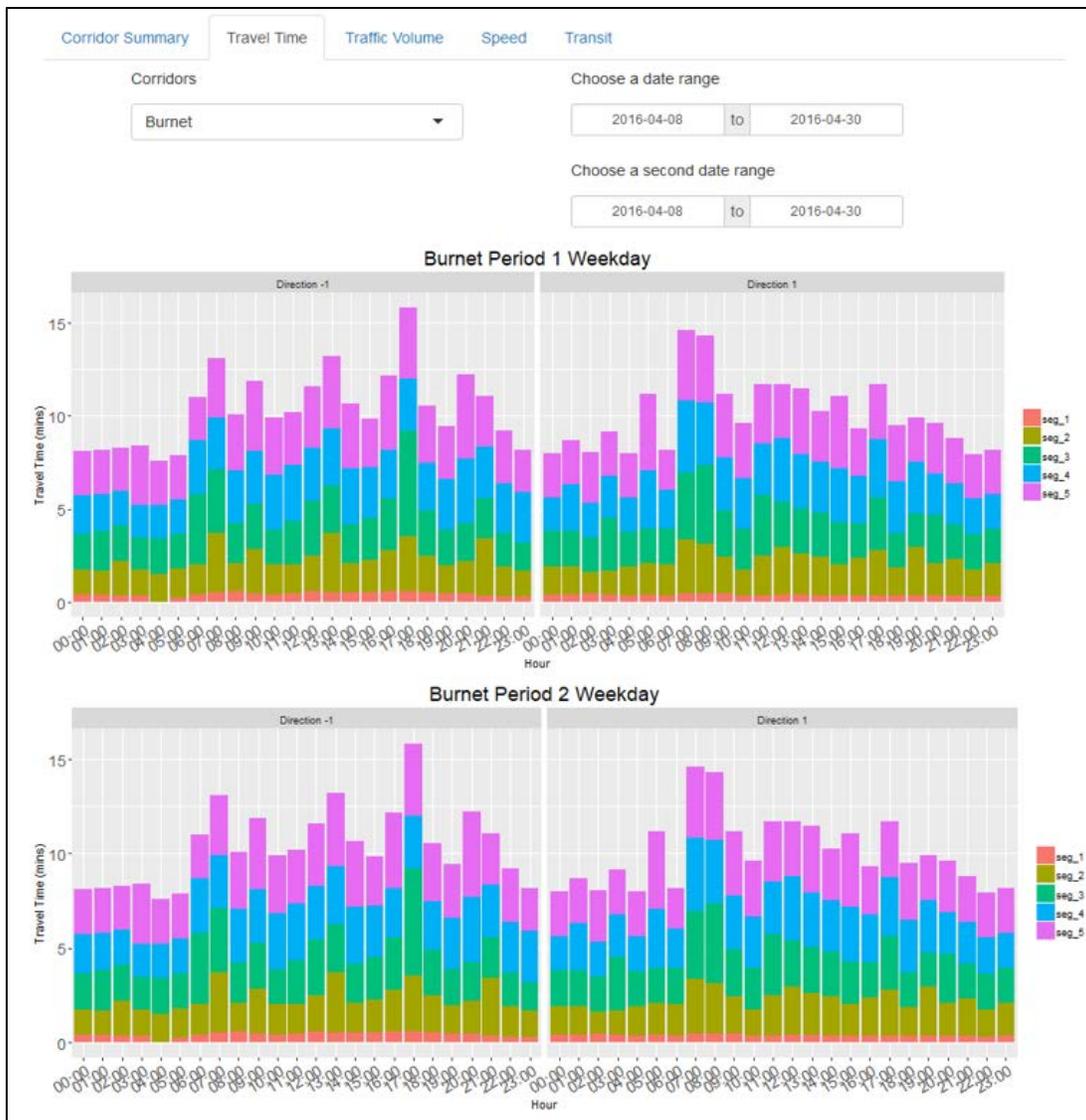


Figure 19. Corridor Travel Time by Hour Analysis for Burnet Road

From the Traffic Counts tab, the user can filter through city count studies, HDR traffic studies, and Bond Corridor studies by date range, time of day, and traffic count type (e.g., bike or automobiles counts). Once filtered, the interface provides a link to each of the relevant studies to view as a PDF or download as a CSV file. Furthermore, the tab includes a “sensor” option which incorporates spot volume analysis from Wavetronix data. Once the “sensor” option is chosen, the user can select a corridor and two distinct dates for analysis. The analysis showcases the average volume by time of day with standard errors for the entire weekday around the dates selected. The “speed” tab includes Wavetronix spot speed information in a similar structure, as shown in Figure 20.

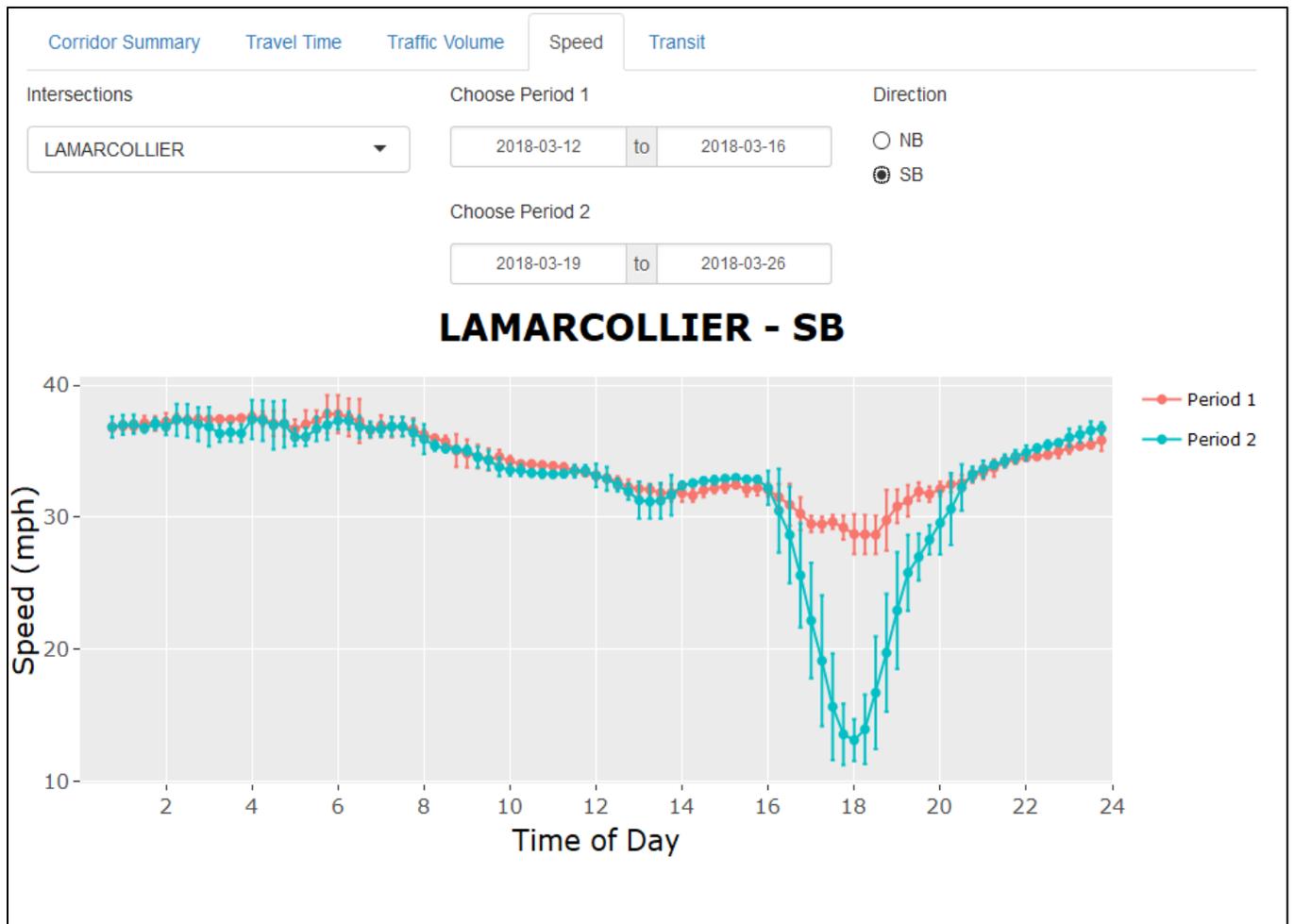


Figure 20. Corridor speed by hour analysis for Lamar Collier intersection

The “transit” tab combines APC and AVL information to compute dwell time analysis and potentially bus route occupancy as well as average transit speed. Dwell time analysis is currently used to summarize average dwell time by corridor for June 2016 and June 2017, as shown in Figure 21, and a visualization of APC, which summarizes total boardings and alightings per corridor across stops for an average day in June 2016 and June 2017.

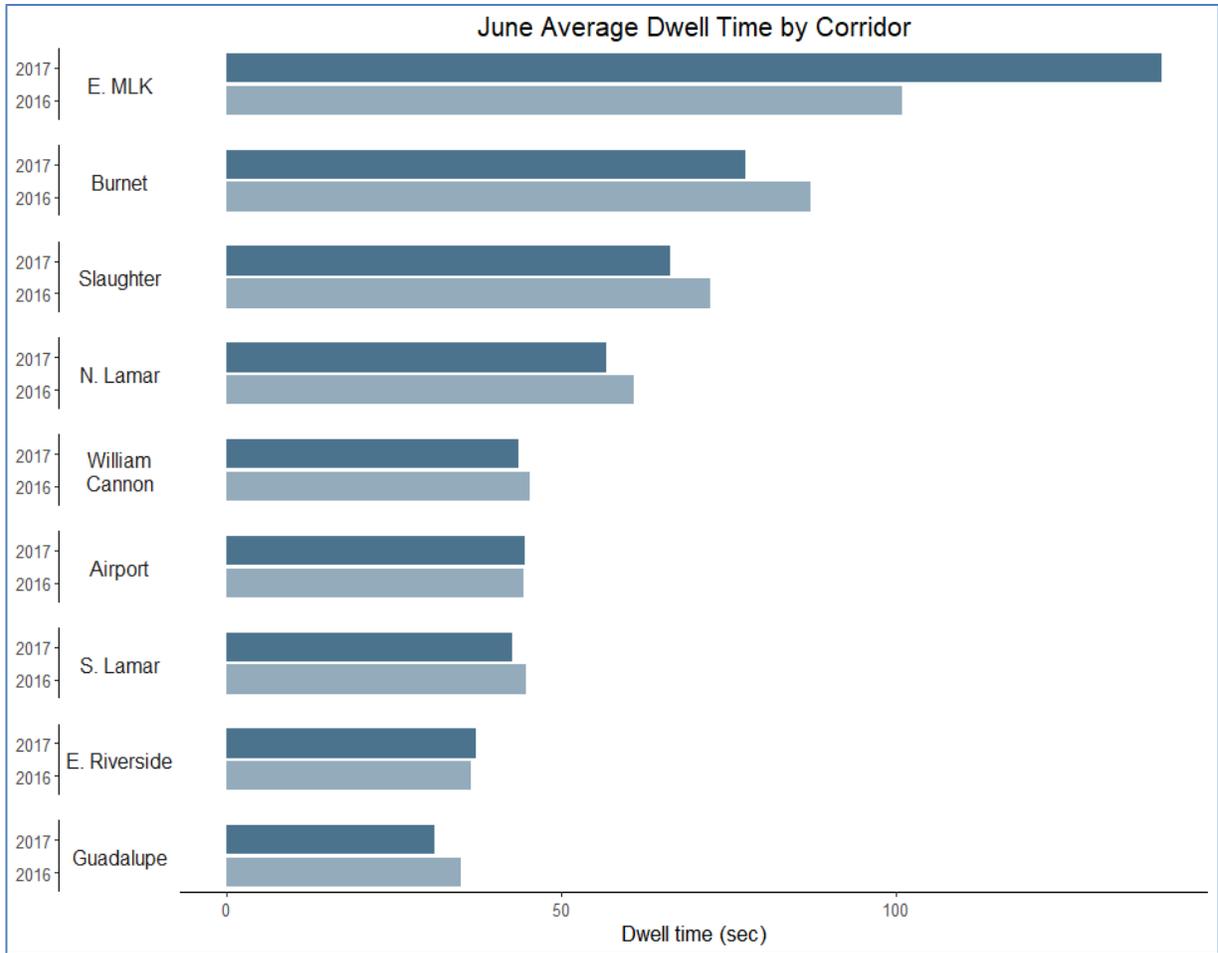


Figure 21. Average dwell time per stop by corridor, June 2016 and June 2017

Future analytics from the data types to be incorporated include additional volume counts, turning movement counts, and intersection speed from the GRIDS MART data as well as transit speed data from AVL information. The application may also incorporate a bus occupancy analysis in the future, depending on data quality. Additionally, overall interface refinement, such as visualization of turning movements from GRIDS MART data, will be finalized once all data types and analyses are incorporated.

Chapter 7. Additional Considerations and Future Direction

The work completed in this fiscal year has resulted in significant progress toward developing an infrastructure through which ATD may further leverage sensor data and maximize its value by analyzing it in combination with data from other sources/agencies. Data workflows and a prototype system architecture are currently in place to systematically access, archive, and process data from multiple sensor types. The use of some data types has been further explored, leading to interesting analyses and a better understanding of the potential value and limitations of the collected data. In addition, the prototype Bond Corridor Performance Analysis Tool exemplifies a valuable use case of the proposed framework, by providing a single-access point to data collected along selected corridors, with the goal of analyzing the evolution of corridor performance over time. The development of the tool, funded through a separate project for the Corridor Project Office (CPO), provided insights into the requirements of a sustainable and scalable system architecture and data workflows. Further, the tool may be extended to analyze other corridors of interest to ATD. The following sections provide additional detail on the nature of the collaboration between CTR, the CPO, and ATD, and the work conducted toward establishing an effective collaboration scheme, and briefly outlines the project direction for Task 2018-12.

7.1 Data Sharing and Collaboration with the Corridor Project Office

The City of Austin Connected Corridors task (2017-10) work has been completed in collaboration with the Corridor Project Office (CPO), the group charged with implementing transportation bond projects. With the approval of the transportation bond, the Austin City Council executed a contract with Austin citizens to improve corridor operating levels for all modes of transportation. CTR has worked with CPO and its consultants to develop the CoA Bond Corridor Performance Analysis Tool (BCPAT) as a means to fulfill the City's contract with Austin citizens. The web tool interface includes a home page with a table, a map that summarizes corridor data availability, and varying tabs that correspond to travel time, traffic volume, speed, and transit analyses (see Figure 12). Currently, information from Bluetooth and Wavetronix sensors, traffic studies, and APC transit data are incorporated into the tool. Future work will incorporate GRIDSMART sensor data and, potentially, AVL transit information.

The plan is for CPO staff to use the BCPAT to obtain statistics and analysis results which would then be made available via a public-oriented interface to be developed and maintained by CPO. Through sharing of data in the BCPAT format, the information will readily available to assess corridor performance levels before, during, and after construction.

The CPO has an additional goal of making information publicly available to Austin citizens, to demonstrate and measure the impact of corridor improvements executed through the transportation bond. During a joint ATD/CPO workshop, attendees discussed holding a hackathon to develop the public-oriented interface for dissemination of data and/or information learned through data analysis with the BCPAT. This idea was

welcomed enthusiastically and CTR will continue to facilitate discussions regarding a hackathon, with the goal of scheduling the hackathon in 2019.

7.2 Coordination of Corridor Data Analysis Efforts across Stakeholders

The coordination of corridor data analysis efforts across these stakeholders has been beneficial in that it has allowed CTR to work on data-flow solutions related to the Connected Corridor project, while at the same time being able to focus on specific corridors (bond corridors) and ensuring the BCPAT and other efforts in gathering, ingesting, processing, and analyzing will be put to use in the near future. CTR has conducted monthly meetings with ATD, with meetings generally occurring on a six-week cycle between October 2017 and September 2018.

7.3 Stakeholder Meetings

Additionally, CTR has met with CPO and HDR (the bond corridor consultant) on numerous occasions. CTR facilitated a joint workshop on May 10, 2018, between CPO and ATD, which resulted in a series of action items related to coordinating ATD and CPO schedules for improvements. The PowerPoint presentation from that workshop is contained in the Appendix 2. CTR continues to facilitate these discussions with another workshop planned for September 21, 2018.

7.4 Future Efforts

Moving forward, work efforts related to data collection, cataloging, access, and storage will be the focus of the contract with ATD. CTR will work with the City of Austin to research and evaluate data storage and distribution technologies and make recommendations based on cost, scalability, and sustainability. CTR will also work with the City of Austin to develop and deploy a data storage system. For each data type:

- If not complete, exploratory analysis/visualization to understand characteristics, coverage, and potential limitations.
- If needed, validate data quality using alternative data sources.
- Document data characteristics, processing workflows and potential data cleaning steps to support desired applications.
- Develop and deploy automated extraction, transformation, and loading (ETL) of source data to the storage system

The Corridor Project Office contract will continue to focus on improving the Bond Corridor Performance Analysis Tool interface and enhancing analytics, as well as developing metrics that combine data sources, to assess corridor performance.

Appendix 1: Source Code List

This section identifies most of the source codes that are responsible for collecting, shuttling, ingesting, and analyzing data. In detail:

- **Collecting** involves acquiring or retrieving data from actual sensors or places where sensor solutions write raw output files. This step runs on the CoA network.
- **Shuttling** delivers these raw output files to their archiving location. Currently, this involves sending raw files to UT CTR, but in the future will involve sending to a CoA-maintained resource.
- **Ingesting** is the process of reading a raw output file and incorporating it into a database. This is necessary to allow the use of the database for analysis purposes. The ingestion process may also reduce data volume by aggregating.
- **Analyzing** is querying, visualizing and manipulating data that's found in the database. The example here is the source code for the Bond Corridor App, which is written in R and runs on a server that hosts a Web-enabling framework called Shiny.

Key examples of source code are listed according to these categories.

Collecting

Name	Description
getPiData.py	Copies the traffic data from the Raspberry Pi to the CoA server
putData.py	Called daily to send the copied files from CoA server to the UT CTR server
wifiTshark.sh	Collects traffic data by using the Tshark package
socrata_wavetronix_call.py	Uses Socrata API to collect Wavetronix data that had been placed there through other processing
gs_getcounts.py	Obtains counts records for one or all GRIDSMART devices for a given date or date range
gs_metadata.py	Obtains metadata for GRIDSMART devices and places it into a preliminary database
city/gridsmart/g_s_tables.py	Database table logic for the preliminary database to keep track of devices and movements
city/gridsmart/log_reader.py	Parser for GRIDSMART counts files
city/db_util.py	Utility class for database access
city/log_util.py	Utility class for log output

Shuttling

Name	Description
gs_exportcounts.py	Ships GRIDSMART counts over a date range to a given destination
...	Additional minimal shell scripts

Ingesting

Name	Description
bt_insert_unmatched.py	Inserts unmatched Bluetooth results from daily log files into a database
ctr/bt/bt_tables.py	Database table logic for the ingester
ctr/bt/log_unmatched.py	Log file parser
wt_insert.py	Inserts Wavetronix results from daily log files into a database
ctr/wt/wt_tables.py	Database table logic for the ingester
ctr/wt/log_wavetronix.py	Log file parser
coa/date_dirs.py	Utility class for managing a directory of files containing dates in the filenames
coa/zip_helper.py	Utility class for managing access to files in compressed archives

Analyzing

Name	Description
server.R	Provides the meat code of the application. It reads information from the database, processes it and creates plots and tables for the UI
ui.R	Provides the graphical layout code of the application
global.R	Global variables for server.R
functions.R	Encompasses code for traffic volume tab
bluetooth.R	Encompasses Bluetooth data processing