SJSU SAN JOSÉ STATE UNIVERSITY

MTI MINETA TRANSPORTATION INSTITUTE

Assessing GTFS Accuracy

Gregory L. Newmark

MINETA TRANSPORTATION INSTITUTE

Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation's transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the Mineta Consortium for Transportation Mobility (MCTM) and the Mineta Consortium for Equitable, Efficient, and Sustainable Transportation (MCEEST) funded by the U.S. Department of Transportation, the California State University Transportation Consortium (CSUTC) funded by the State of California through Senate Bill 1 and the Climate Change and Extreme Events Training and Research (CCEETR) Program funded by the Federal Railroad Administration. MTI focuses on three primary responsibilities:

Research

MTI conducts multi-disciplinary research focused on surface transportation that contributes to effective decision making. Research areas include: active transportation; planning and policy; security and counterterrorism; sustainable transportation and land use; transit and passenger rail; transportation engineering; transportation finance; transportation technology; and workforce and labor. MTI research publications undergo expert peer review to ensure the quality of the research.

Education and Workforce Development

To ensure the efficient movement of people and products, we must prepare a new cohort of transportation professionals who are ready to lead a more diverse, inclusive, and equitable transportation industry. To help achieve this, MTI sponsors a suite of workforce development and education opportunities. The Institute supports educational programs offered by the Lucas Graduate School of Business: a Master of Science in Transportation Management, plus graduate certificates that include High-Speed and Intercity Rail Management and Transportation Security Management. These flexible programs offer live online classes so that working transportation professionals can pursue an advanced degree regardless of their location.

Information and Technology Transfer

MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and works to integrate the research findings into the graduate education program. MTI's extensive collection of transportation-related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

Report 24-15

Assessing GTFS Accuracy

Gregory L. Newmark

August 2024

A publication of the
Mineta Transportation Institute
Created by Congress in 1991

College of Business
San José State University
San José, CA 95192-0219

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 24-15	**2. Government Accession No.**	**3. Recipient's Catalog No.**	
4. Title and Subtitle Assessing GTFS Accuracy		**5. Report Date** August 2024	
		6. Performing Organization Code	
7. Authors Gregory L. Newmark ORCID: 0000-0002-8780-8832		**8. Performing Organization Report** CA-MTI-2017	
9. Performing Organization Name and Address Mineta Transportation Institute College of Business San José State University San José, CA 95192-0219		**10. Work Unit No.**	
		11. Contract or Grant No. 65A0660	
12. Sponsoring Agency Name and Address California Department of Transportation 1120 N Street Sacramento CA 95814		**13. Type of Report and Period Covered**	
		14. Sponsoring Agency Code	
15. Supplemental Notes 10.31979/mti.2024.2017			

16. Abstract

The promised benefits of the General Transit Feed Specification (GTFS) Schedule and Realtime standards are dependent on the underlying quality of the data. Despite this fundamental reliance, there has been relatively little research on techniques and strategies to assess GTFS accuracy. The need for such assessment is growing as federal and state governments increasingly require transit agencies to make these data available to the public. This research fills this gap by presenting a suite of methods and metrics to assess the temporal accuracy of GTFS Realtime and the spatial accuracy of GTFS Schedule feeds. The temporal assessment demonstrates an approach to collect and clean *TripUpdate* messages to identify (and derive) a set of values for measuring the accuracy of the vehicle arrival predictions. These metrics are carefully designed to provide transit agencies insight into the quality of the data they provide to customers in terms of the impact of those inaccuracies on the customer experience. The spatial assessment demonstrates an approach to match scheduled information on the location of transit routes and stops with the actual travel patterns demonstrated in the realtime *VehiclePosition* messages. The measured divergence between the planned and provided transit service yields a series of location accuracy metrics. All of the proposed metrics can be scaled to examine GTFS accuracy from the stop to the systemwide level. All of the proposed metrics can be easily generated from publicly available GTFS feeds without any additional data sources. Finally, all of the proposed metrics can help transit agencies continuously assess and therefore improve the quality of GTFS data they share with the public.

17. Key Words Public transit, Data, Real time information, Problem solving, Statistical analysis.	**18. Distribution Statement** No restrictions. This document is available to the public through The National Technical Information Service, Springfield, VA 22161.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 112	**22. Price**

Form DOT F 1700.7 (8-72)

DOI: 10.31979/mti.2024.2017

Mineta Transportation Institute
College of Business
San José State University
San José, CA 95192-0219

Tel: (408) 924-7560
Fax: (408) 924-7565
Email: mineta-institute@sjsu.edu

transweb.sjsu.edu/research/2017

ACKNOWLEDGMENTS

The author is deeply grateful to Bradley Mizuno of Caltrans and Karen Philbrick and Hilary Nixon of the Mineta Transportation Institute for their excellent and supportive project management. Their organizational acumen, personal kindness, and generous administrative efforts were essential to the success of this project – and tremendously appreciated.

The author thanks Chad Riding, Gillian Gillett, and Elizabeth Sall for shaping the direction of this research at its outset and Chad Riding, Tiffany Chu, Evan Siroky, Jack Walker, and Bradley Mizuno for taking the time to review drafts and provide detailed comments. Caltrans's commitment to putting transit research to practice has been a constant source of inspiration.

Finally, the author thanks Jordan Holt and Scott Traum of WMATA for informal discussions of these issues at the early stage of this project, Raji Rajesh of MTI for providing careful copy editing of this report at the later stage of this project, and Alex Brufsky, Peter Haas, and Michael Trifonov, whose stellar computer science skills made the data analysis at the core of this research possible.

This work is dedicated to the transit riding public for whom accurate data makes a daily difference.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Executive Summary

The emergence of the General Transit Feed Specification (GTFS) standard opens tremendous potential for improving the user experience of public transportation. These benefits can only be realized if the underlying GTFS information is accurate. This research compares publicly available GTFS Static and Realtime feeds for five transit systems in California to generate a suite of techniques and metrics for assessing that accuracy.

This first portion of this research explores the temporal accuracy of the vehicle arrival times presented in the trip update messages of the GTFS Realtime feeds.

This analysis describes two recommended waves of data cleaning to arrive at a canonical data set for calculating metrics. The first cleaning wave removes duplicative records while including all records for all trips that began during the prescribed study period. This latter feature includes trips that began before midnight during the last day but extended into an additional day. The second cleaning wave constrains the trip update messages to those sent at or after the scheduled start time of the trip. This culling removes the many messages that are broadcast prior to the initiation of the trip under consideration. This cleaning wave resolves conflicts with different prediction values having the same timestamp by randomly selecting one and only one prediction for each stop on each trip for each given time. This cleaning wave also removes stop predictions made after the last known vehicle departure event to exclude unnecessary predictions as well as redundant predictions included in the data for stops that have already been served. Finally, this cleaning wave removes unusual predictions, here called continuity errors, for which the anticipated arrival times are listed prior to the timestamp of the message, an impossible outcome given the other cleaning steps. This cleaning reveals underlying data concerns regarding time stamp conflicts and continuity errors and recommends a ratio of these excised values to the final data set to assess their prevalence (which also serve as initial accuracy metrics).

The cleansed data provides the basis for the remaining accuracy metrics. These metrics build on values pulled directly from the GTFS feed as well as additional and easily derived values capturing the time to prediction, prediction error, time to stop, and prediction change. These metrics include the share of trip minutes for which an update is available, plots of prediction error percentiles, interquartile range of predictions scaled by the time to prediction, the likelihood of catching a bus given the prediction, the expected wait time (including to the next bus) given the prediction, the amount of padding necessary to have a 95 percent change of not missing the bus given the predictions, and the prediction inconsistency. The report presents these metrics as numbers and visualizes them as charts (at either the systemwide or route level) to demonstrate how they might be effectively employed to diagnose prediction accuracy.

The second portion of this research explores the spatial accuracy of the GTFS products by comparing the schedule data to the vehicle location messages in the realtime feed. This analysis focuses on the accuracy of the vehicle paths and stop locations.

The paths analysis cuts the shapes within the GTFS static product into segments and then calculates the straight-line distance from each vehicle ping to the nearest point on those paths. This analysis is conducted on a single transit system in California for tractability. A review of the cumulative distribution of these ping-path distances suggests a 14-meter threshold as reasonable for identifying ping-path discrepancies; however, as discussed in the report, a 10-meter threshold could work in many circumstances. The pings exceeding the threshold are flagged to generate route- and segment-level accuracy metrics.

The use of these metrics to flag path inaccuracies is presented. For example, one route with high shares of flagged ping-path distances is found to be poorly coded in the GTFS static data. The shape of the route does not align with the actual road network, which results in the apparent discrepancies. Conversely, at another portion of that same route, which is accurately coded in the GTFS static data, the bus drivers are diverging from the path (by leaving a highway an exit early), which results in actual discrepancies. Similarly, segments with high shares of flagged ping-path distances are identified and explored. These segments reveal specific locations where bus drivers appear to be consistently deviating from the path, for example to facilitate turning movements or to respond to construction detours, or where global positioning signals seem to drift consistently away from the roadway.

A similar approach is used to assess the accuracy of stop locations. This analysis is limited to a different California transit system that includes within its GTFS Realtime feed an indicator that the vehicle was at a stop. This analysis also incorporates an additional data cleaning step to reduce the data to unique stop events (rather than counting separate pings on the same day on the same trip at the same location as different stops). A relatively lenient 30-meter threshold is established to flag ping-stop discrepancies to generate accuracy metrics at the route, trip, and stop level.

The use of these metrics to flag stop inaccuracies is presented. For example, the route-level metrics suggest that commuter and school routes are much more likely to have ping-stop deviations than other types of routes, but overnight routes are more likely to make large ping-stop deviations. Trip-level metrics find that the ping-stop exceedance rates vary substantially on trips for the same route and demonstrate ways to explore this variation further, for example, by time-of-day. Stop-level metrics reveal stops with major disparities between the scheduled location and the actual stopping location. Many of these locations are characterized by congested traffic and the convergence of multiple bus routes, but some appear to reflect situations in which the driver appropriately stops at a bus shelter located some distance from the stop. The ping-stop distances with the greatest magnitude suggest possible equipment or coding errors for the agency to investigate.

Taken together this report provides techniques for transit agencies to assess the temporal and spatial accuracy of the GTFS products they share with the public and to diagnose the causes of any inaccuracy. The goal of this work is to provide both metrics and means of applying them so that transit agencies can methodically improve their GTFS data accuracy and ensure their customers enjoy the benefits of these data feeds.

1. Introduction

The emergence of the General Transit Feed Specification (GTFS) standard opens tremendous potential for improving the user experience of public transportation. GTFS enables users to plan trips in such a way as to account for the full transit schedule and realtime deviations from that schedule. GTFS also enables users to be dynamically apprised of key variables, such as expected vehicle arrival time. The former reduces barriers to using transit in the first place while the latter reduces the costs of such use. Furthermore, although designed for trip planning, the information coded within GTFS offers many alternative applications, from service analysis to performance measurement.

All of these benefits depend on the underlying accuracy of the GTFS Static and Realtime products. To date, there has been little research into the assessment of GTFS accuracy. This research will advance this effort by presenting methods and metrics to explore and quantify GTFS accuracy. A key innovation of this work is to derive these products from the publicly available GTFS feeds themselves. This feature ensures the wide applicability of these tools.

This report is divided into two major sections. The first section focuses on the accuracy of vehicle arrival times presented in the trip update feeds as part of GTFS Realtime. This analysis is primarily temporal and emphasizes the development of metrics to measure prediction accuracy. The second section focuses on the geographic accuracy of the GTFS Schedule information in reflecting actual operations. This analysis is primarily spatial, and in addition to presenting accuracy metrics, offers techniques to explore divergences between what the GTFS Schedule data says should happen and what the GTFS Realtime data says does happen. This section is subdivided to consider the two key spatial components of GTFS—travel paths and stop locations.

1.1 Project Background

This research builds on the extensive standardization efforts that have taken place to develop the GTFS Static and GTFS Realtime feeds.

GTFS Static

GTFS Static codes transit schedules as a set of tables with linkable identifiers (and is also called GTFS Schedule for this reason). At the core of the system is the *trip_id*, which identifies the stops, stop sequence, stop times, and service days that uniquely define the structure and timing of a single transit trip. (In this document, all explicit references to GTFS nomenclature are italicized.) The *trip_id* is also linked to a *route_id* as a series of related trips comprise a transit route.

In practice, transit agencies do not always invest in ensuring the accuracy of GTFS Static information. For example, while GTFS Static allows for the separate coding of stop arrival and

departure times (to reflect the expected dwell times at stops), many systems report these as the same exact time (Wessel & Widener, 2017). Similarly, while agencies generally make an effort to code these stop times accurately for key time points to guide drivers, less effort is aimed at correctly coding the timing of intervening stops (Wessel & Farber, 2019). These shortcuts reflect the reality of transit provision. It can be difficult to predict how long dwell times are likely to be for any given stop at any given time along a route. Similarly, exhorting drivers to reach a few key destinations (often timed transfer locations) at set timepoints is a far more reasonable request (given the cognitive load and traffic vicissitudes) than requiring drivers to reach every single stop on a set schedule. Fortunately, by providing realtime updates, GTFS Realtime can correct for both poor stop time coding and limited schedule adherence endemic to GTFS Static.

GTFS Realtime

Most transit agencies track the positions of their fleets in realtime using automatic vehicle location (AVL) devices. These devices transmit timestamped geocoded locations for each vehicle, typically over a cellular network. While this information was initially sought to manage far-flung fleets, it also can be employed in concert with GTFS Static data to predict stop arrival and departure times. Transit agencies process the AVL information into several GTFS Realtime feeds for the public, usually through outside vendors.

The most relevant GTFS Realtime message sets for the current research are the *TripUpdate* and *VehiclePosition*. The *TripUpdate* feed provides anticipated arrival and/or departure times for stops (as well as the timestamp for when those estimates were made) along upcoming and current trips. These trips are identified by the same *trip_id* in GTFS Static, which enables linking the realtime information to the schedule. The *VehiclePosition* feed relays timestamped coordinates of vehicle locations, also coded by *trip_id*.

1.2 Previous Work

Any accuracy assessment requires sourcing ground truth vehicle location data. Some GTFS Static researchers have ingested external automatic vehicle location (AVL) feeds (Wessel & Farber, 2019) while others have relied on the *VehiclePosition* data from the GTFS Realtime (Abusalim, 2020; Steiner et al., 2015) feeds. While the latter solution seems particularly elegant in application to realtime assessment, Steiner et al. (2015) found the *VehiclePosition* data relatively sparse, noting that such information was only available for 57% of bus stops across the entire Dutch transit network. Similarly Abusalim (2020) jettisoned consideration of bus delays since too few of the *VehiclePosition* records across all bus lines in Boston included an indicator that the bus was at a stop location. The Dutch case might be explained by its use of data from 2014 when GTFS Realtime was quite new, and the Boston case might be addressed by imputing stop status rather than relying on the *VehicleStopStatus* field of the *VehiclePosition* messages. Nonetheless, both cases

demonstrate difficulty using the GTFS Realtime fields as internal sources of ground truth location data.

Any accuracy assessment also requires measuring the difference between the stated and actual transit experience. Wessel and Farber (2019) estimated actual and scheduled travel times between the same set of origins and destinations in a given transit system for every minute during a weekday peak period. They then visualized systemwide accuracy by creating density plots of the percent difference (presented on a log scale) in these values for each scheduled estimate for each city. This approach is unique in considering transit travel times rather than route-by-route performance. More commonly, researchers compared the actual arrival and/or departure times culled from GTFS Realtime feeds to scheduled activities from the GTFS Static feeds. Steiner et al. (2015) estimated average delay per bus stop, which they presented as a histogram to show average delay frequency experienced at stops across the network. Abusalim (2020) similarly estimated delays across all transit modes in Boston, which he presented as a box plot.

Another set of researchers and practitioners has explored the accuracy of the GTFS Realtime feeds themselves, most commonly to assess the accuracy of the *TripUpdate* predictions (Machlab et al., 2017; Steiner et al., 2015). An early effort overlayed line charts of the predicted and actual arrival delays at successive bus stops along a single route using three *TripUpdate* and one *VehiclePosition* streams to show limitations in data availability in GTFS Realtime data, particularly the feature of *TripUpdate* delay predictions defaulting to zero beyond the then ten-minute prediction window (Steiner et al., 2015). While the authors did not directly analyze GTFS Realtime accuracy, their visualization suggests an approach of comparing the delay values at each stop.

Machlab et al. (2017) used such a stop-level comparison to assess the accuracy of a GTFS Realtime feed in Denver prior to its public launch. The authors proposed an approach like that commonly used for determining transit on-time performance. They establish four sequential bins of time leading up to a bus arriving at a stop (30 to 12 minutes, 12 to six minutes, six to three minutes, and three to zero minutes). For each bin, they recommend a threshold of tolerable prediction error (defined as the actual arrival or departure time minus the predicted time). Those thresholds shrink as the arrival time approaches to encourage more accurate predictions. The thresholds for the first three bins are offset slightly to discourage negative predictions, which would place passengers at the stop after the bus had departed. Those threshold bandwidths are, respectively, -4.0 to 6.0 minutes, -2.5 to 3.5 minutes, -1.5 to 2.0 minutes, and -1.0 to 1.0 minute. All predictions that fall within those bandwidths are considered acceptable. The accuracy metric, as for on-time performance, is the share of predictions that are acceptable.

This method has several benefits: it allows for a differential consideration of early and late predictions, it values increasing accuracy as the time to the bus arrival shrinks, and it uses a binary measure of success to make it easy to aggregate the information into a single percentage. In the absence of other approaches to determine prediction accuracy, this method is commonly used by

transit properties to assess a sample of predictions across their system (Swartz, 2020). This method also has several drawbacks: it relies on arbitrary bin widths that make for a disjointed consideration of predictions on either side of those dividers, it relies on arbitrary thresholds for determining prediction acceptability, its determination of acceptability is binary so that a huge error counts the same as a small one—despite the potential impact on a passenger, and its assessment window is limited to a half an hour even though the availability of those predictions might be longer. The metric also does not consider the transit service frequencies, and thus the consequences of a late prediction.

This research builds on these approaches to assessing GTFS accuracy, particularly the complementary comparison of GTFS Static and Realtime feeds to offer an internal ground truthing. This comparison is made possible by the more complete GTFS resources available from selected transit systems in California. This work offers a series of techniques and metrics that allow transit properties to evaluate the quality of the information they share with the public as well as strategies for diagnosing areas for effective intervention.

2. Arrival Prediction Accuracy

Transit agencies seek to attract new passengers and retain and expand the patronage of existing passengers. This objective serves the pressing public policy needs to reduce the carbon intensity of travel (to mitigate climate change), to optimize the use of limited road capacity (to manage congestion and avoid construction costs), to fill unused transit capacity (to reap more public benefit from sunk costs and justify associated subsidies), and to raise revenues (to support the service).

The electronic provision of schedule (i.e., routes, stops, and times) and operational (i.e., delays, service changes) information has been shown to reduce uncertainty regarding transit use and increase both the use and the satisfaction associated with use (Barbeau & Fretheim, 2018). While the development of electronic route planners using static schedule information was already a major advance in transit navigation, the addition of dynamic arrival time predictions has further reduced the uncertainty tied to transit use. Reductions in uncertainty translate into increases in perceived reliability—an important human value. Furthermore, realtime arrival predictions allow customers to minimize transit wait times, typically seen as the most onerous component of a transit trip, and maximize non-transit activity times, such as additional minutes at home or at work locations—a net gain to society. Similarly, when the experience of waiting for transit vehicles is physically uncomfortable due to inclement weather or security concerns, the ability to accurately estimate vehicle arrivals can further reduce stress on travelers. These benefits are dependent on the accuracy of the arrival prediction information. While historically, transit agencies used different approaches to electronically provide these service data, over the last decade and a half, the industry has converged on two related standardized formats: the general transit feed specification (GTFS) for fixed schedule information and its realtime extension (GTFS-RT) for operational deviations from the fixed schedule. While several researchers have explored the accuracy of the former (Abusalim, 2020; Steiner et al., 2015; Wessel & Farber, 2019), there has been surprisingly limited work on the latter (Machlab et al., 2017). This omission is problematic since in theory, GTFS-RT adjusts for the inaccuracies of GTFS Static (Wessel & Farber, 2019). In practice, perfect schedule adherence is almost impossible for transit services operating in mixed traffic and passengers often rely entirely on the GTFS-RT streams for making near-term trip planning decisions (Abusalim, 2020). Understanding the accuracy of those realtime fields is therefore essential for transit agencies that broadcast GTFS-RT data to the public (Machlab et al., 2017). Such assessment is doubly important since most transit agencies contract out the generation of GTFS-RT fields to private companies and have a fiduciary duty to assess the quality of the information they are purchasing with public monies.

This section presents a proposed suite of performance metrics that assess the accuracy of the realtime updates on predicted arrival times. These metrics are designed to reveal accuracy from the perspective of both the transit agencies broadcasting the predictions and the transit customers seeking to use this information to improve their travel experience. These metrics are structured to be easy to understand and useful for tuning both prediction algorithms and transit service delivery.

To facilitate application in practice by any interested party, the proposed metrics can be calculated entirely from the *TripUpdate* messages of the GTFS-RT feed and the GTFS Static transit schedule—both freely available from the transit agency—without reliance on any additional sources of data.

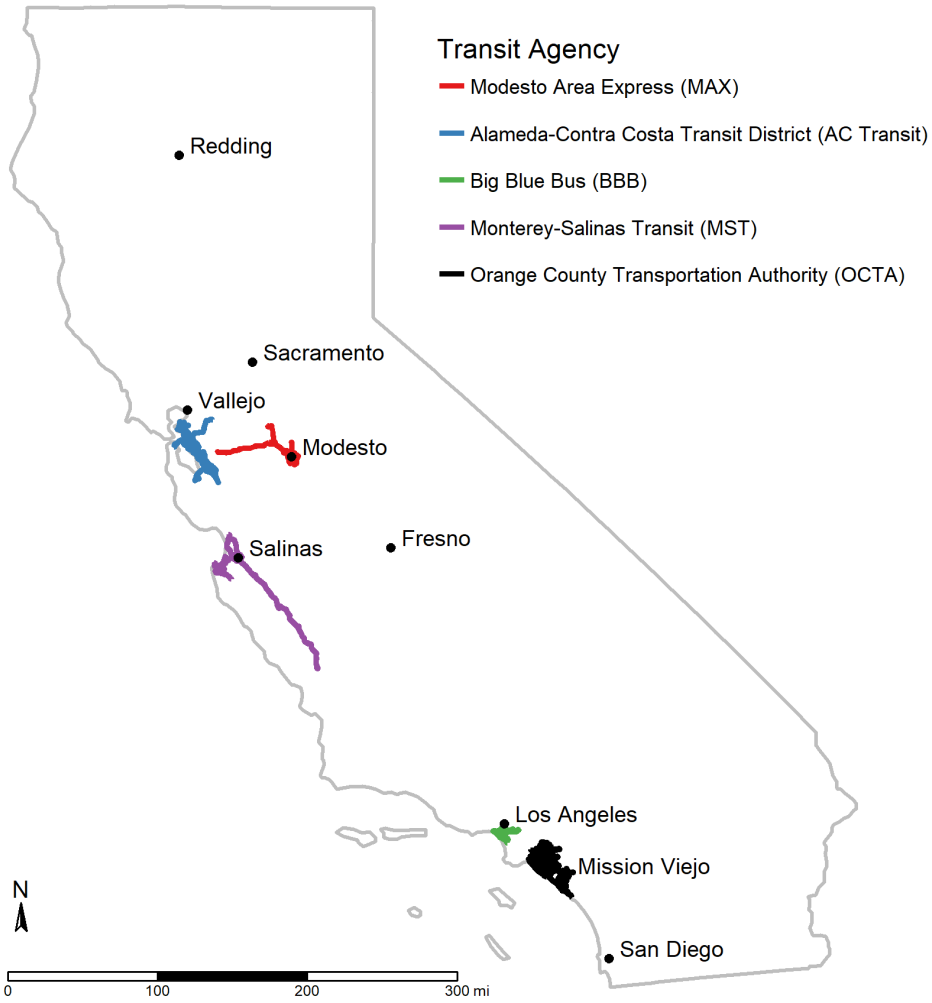
2.1 Methodology

This section explores five full days of GTFS-RT *TripUpdate* data from five California bus systems to generate quantitative metrics for assessing the accuracy of transit arrival predictions. These performance measures are designed to address distinct needs of transit agencies and the traveling public.

2.2 Study Area

The metrics proposed in this section are based on publicly available GTFS-RT feeds from five California transit agencies: Modesto Area Express (MAX), Alameda-Contra Costa Transit District (AC Transit), Monterey-Salinas Transit (MST), Big Blue Bus (BBB), and Orange County Transportation Authority (OCTA). These agencies were chosen to represent differently sized transit operations in various parts of the state. Figure 1 presents the route footprint of studied transit agencies as well as their relative locations within California.

Figure 1. Map of Studied Transit Agencies



It is worth noting that on July 1, 2021, less than a year prior to the data collection for this project, the city-owned MAX merged with the county-owned Stanislaus Regional Transit (StaRT) to form a new transit agency, the Stanislaus Regional Transit Authority (StanRTA). Despite this legal union, many former MAX assets remain distinct, including the GTFS feeds. For this reason, this research refers to MAX, even if that agency no longer exists as a separate legal entity.

2.3 Data Collection and Cleaning

A consistent approach to collecting and cleaning data is central to the proposed GTFS-RT accuracy metrics. This data cleaning approach sought to extract the stop prediction information most relevant for transit riders and therefore also for the transit agencies purchasing and broadcasting the data. The cleaning approach has two waves. The first wave cleans the raw data to remove duplicates, trim the data to the period of interest, and remove any record without stop

predictions to yield a base data set with complete information. The second wave cleans the base data set to remove predictions that are either not useable by the consumer or are impossible to yield the final data set with complete and actionable information. This subsection presents the scope of the data collection and the structure of the data cleaning.

Data Collection

The transit agencies all post their current GTFS-RT files in the open-source protocol buffer format (.pb) on their respective website for users to download. Five full days of GTFS-RT data for all five transit systems were collected from Tuesday, March 1, 2022 through Saturday March 5, 2022 by downloading the available protocol buffer file every ten seconds during that 120-hour period (as well as additional selection before and after the time frame to ensure no missing records for trips that spanned days). This five-day span was selected to provide a large sample of data covering both weekday and weekend operations. The several hundred thousand protocol buffer files downloaded over the study period from the five transit agencies were converted into a single comma-separated (.csv) file for analysis. To provide route and scheduling context, the contemporaneous static GTFS data for the five systems were also downloaded.

Data Cleaning: Wave I

The first cleaning wave creates the base data set of unique rows, all within the study period and all with stop predictions. Table 1 shows the number of rows remaining (and the relative shares) at each stage of the first data cleaning wave.

Table 1. Data Cleaning: Wave I (Download to Base Data)

Removed Rows	MAX	AC Transit	BBB	OCTA	MST
Number of Rows Remaining					
(Nothing Removed)	136,801,210	426,021,170	119,382,333	7,766,917	50,818,594
Duplicates	53,978,293	260,573,419	17,238,678	7,766,917	33,876,802
Null Predictions / Wrong Dates	45,736,942	213,214,340	13,387,144	6,357,630	28,029,902
Share of Rows Remaining					
(Nothing Removed)	100.0%	100.0%	100.0%	100.0%	100.0%
Duplicates	39.5%	61.2%	14.4%	100.0%	66.7%
Null Predictions / Wrong Dates	33.4%	50.0%	11.2%	81.9%	55.2%

The first cleaning step removed duplicative records. This duplication is expected whenever the update rate of the protocol buffers themselves was less frequent than the 10-second data collection download rate. All systems except OCTA, which has high refresh frequency, witnessed substantial data reductions after de-duplication. MST dropped by a third (-33.3%), AC Transit by two fifths (-38.8%), MAX by three fifths (-60.5%), and BBB by six sevenths (-85.6%). These findings suggest some initial processing efficiencies might be found in tailoring the download rate to the upload rate.

The second cleaning step removed records outside the study period or without a prediction. The time filtering introduces a small complication, as some transit trips begin before midnight on one day and continue into the following day. This research followed a protocol of including all trips that originated during the study period even if they continued into the following day. Therefore, some trips that began on February 28 and extended into March 1 were entirely excluded, while other trips that began on March 5 and extended into March 6 were fully included.

GTFS-RT has an experimental *start_date* field, which makes this selection straightforward, when available. All transit agencies in the study sample, except AC Transit, included the optional start date information. To populate this field for AC Transit, the lowest stop sequence number (which corresponds to the first stop) was identified for each trip reported within the *TripUpdate* messages. The date portion of the timestamp for that record was then added to the otherwise blank *start_date* field for all records both sharing that same *trip_id* and taking place over the subsequent twelve hours. This process ensured all records had the appropriate start date upon which to filter the data set.

It is important to note that the program that downloaded the data was set to last for the same span of time once initiated, but that it was initiated at separate times on the same day for each transit system studied. The purpose of this second cleaning step is to ensure the base data sets are for the same period. Since the downloaded data sets are slightly staggered temporally, and since the number of transit trips made vary throughout the day, the reductions from this stage of the data cleaning should not be precisely compared to one another. As noted above, any record without either a stop arrival or a stop departure prediction was also removed at this stage, as these data cannot be used to assess the accuracy of the predictions. In combination, the filtering by time and removing of records without predictions reduced the records by a similar order of magnitude, specifically between 22.2%, the largest drop, for BBB and 15.4%, the smallest drop, for MAX.

Data Cleaning: Wave II

The second cleaning wave creates the final data set from the base data output from the first cleaning wave. Table 2 shows the number of rows remaining (and the relative shares) at each stage of the second data cleaning wave. This section details the decisions to pare the base data to the final data set for analysis.

Table 2. Data Cleaning: Wave II (Base to Final Data)

Removed Rows	**MAX**	**AC Transit**	**BBB**	**OCTA**	**MST**
Number of Rows Remaining					
(Nothing Removed)	45,736,942	213,214,340	13,387,144	6,357,630	28,029,902
Timestamps before Trip Start	7,985,117	90,381,225	4,966,186	6,304,955	10,813,054
Timestamp Conflicts	7,977,617	90,377,255	4,965,557	6,304,955	10,813,054
Timestamps after Trip "End"	7,308,076	90,375,024	4,965,557	5,105,297	10,799,042
Stops Already Passed	4,764,643	89,378,262	4,795,844	269,881	9,986,918
Continuity Errors	4,700,564	89,370,069	4,793,511	0	9,851,628
Share of Rows Remaining					
(Nothing Removed)	100.000%	100.000%	100.000%	100.000%	100.000%
Timestamps before Trip Start	17.459%	42.390%	37.097%	99.171%	38.577%
Timestamp Conflicts	17.442%	42.388%	37.092%	99.171%	38.577%
Timestamps after Trip "End"	15.978%	42.387%	37.092%	80.302%	38.527%
Stops Already Passed	10.417%	41.919%	35.824%	4.245%	35.630%
Continuity Errors	10.277%	41.916%	35.807%	0.000%	35.147%

The first decision was to begin the accuracy assessment with the scheduled start time of the trip. This approach entailed excluding any predictions broadcast earlier. For all the studied systems except OCTA, most of the trip updates were made before there was any actual information on the conditions along a given trip—so these reductions were substantial (as high as 82.5% for MAX). In theory, it is possible (and recommended) to use the block information available in GTFS-RT to identify how delays on one trip might affect a subsequent one on the same block or to incorporate historical information in anticipating travel speed. In practice, however, pre-trip predictions did not seem to account for this information and incorporating these pre-scheduled start predictions only increased the share of errors in summary metrics.

The second decision was to address timestamp conflicts. In the *TripUpdate* feed from MAX, AC Transit, and BBB (but not OCTA or MST) there were thousands of instances where more than one prediction for a single stop shared a timestamp but provided a different predicted arrival time. These predictions would suggest that the bus, like quantum particles, could be in two places at the same time. While timestamp conflicts are likely the result of processing bottlenecks in the generation and transmission of the update, they present a challenge for analyzing prediction accuracy. As an expedient, when faced with a timestamp conflict, it was decided to randomly select one prediction to keep in the data set and to discard any others. (It would also be entirely reasonable to either include or exclude all such conflicts.) The occurrence of such conflicts, however, is a red flag for the accuracy of predictions, as will be discussed in the assessment metric section.

The third decision was to remove all stop predictions broadcast after the last known departure event. This event is referred to as the trip "end" in quotations simply because it is the last record for the progress of any given trip. In many cases, the trip "end" is not the last scheduled stop along the trip. (The MAX data only provides predictions through the penultimate stop, a reasonable practice for *TripUpdate* feeds since no one can board a vehicle at its last stop. The other systems do provide predictions of last stop arrivals since that information can still be useful for trip planning.) These data appear to be records delayed by computation processes internal to the transit vehicle. Such post-trip "end" records are only seen on MAX and OCTA, suggesting hardware concerns to be addressed. (It should be noted that while these post-trip updates are excluded from the generation of the performance metrics, the information contained within them is incorporated when relevant. For example, when a post-trip end update includes actual arrival and departure times that were otherwise missing among the trip updates, which happens for MAX, those were appended appropriately to stop predictions.)

The fourth decision was to remove all stop updates generated after the vehicle had departed that stop. GTFS-RT feeds can include data for all stops along the active trips, even if those stops have already been served. This redundant information does not affect trip planning since the vehicle has already passed those stops. This research selected the last available prediction for any given stop to represent the actual arrival and departure times. All the systems, except MAX, provide identical final arrival and departure times. MAX, by contrast, provides distinct times once the stop has been

passed. The MAX approach enables a more nuanced understanding of travel by accounting for dwell times.

The fifth and final decision was to remove a set of unusual remaining predictions whose predicted time of arrival were for times prior to the timestamp of the prediction. In plain language, an example of this impossibility might read, "the bus will arrive at this stop five minutes ago," even though the vehicle has yet to arrive at the stop based on the sorting from the previous data cleaning stage. These predictions for events that are expected to occur in the past are called "continuity errors." Such continuity errors were present in the GTFS-RT feeds of all five transit systems studied, but particularly predominant at OCTA. The OCTA *TripUpdate* feed was distinct in that it included no predictions. Rather, the feed only provided the actual arrival time at a stop after that event had occurred. Consequently, there were no remaining OCTA predictions in the final data set.

The result of the second wave of data cleaning is the final data set that refers to all reasonable predictions during the active period. This number also serves as the denominator for two accuracy metrics: the Timestamp Conflict Ratio and the Continuity Error Ratio. These ratios are similarly defined as the number of relevant rows removed to each thousand rows in the final data set, as shown in Table 3.

Table 3. Timestamp Conflict and Continuity Error Ratios

Accuracy Metric	MAX	AC Transit	BBB	OCTA	MST
Timestamp Conflict Ratio	1.60	0.04	0.13	0.00	0.00
Continuity Error Ratio	13.63	0.09	0.49	NA	13.73

Ideally, both ratios would be at or close to zero. For the Timestamp Conflict Ratio, MAX has a score of 1.60—more than ten times the next highest ratio of 0.13 at BBB. For the Continuity Error Ratio, MAX and MST share ratios above 13.5, substantially higher than those at either AC Transit or BBB, which are both less than 0.5. The Continuity Error Ratio is undefined for OCTA, since without values for the final data set the denominator of the value is zero.

2.4 Derived Values

The previous section discussed fields downloaded from either GTFS-RT or GTFS Static feeds to structure the final data set. These downloaded fields, shown in Table 4, are used to derive four additional fields, each referring to a different time span necessary for generating the accuracy metrics. This section presents those four derived fields, which are also enumerated in Table 4.

Table 4. Description of Downloaded and Derived Data Fields

Values	**Definition/Formula**
Downloaded	*Definition*
Timestamp	Time of GTFS-RT update
Prediction	Predicted time of arrival/departure at stop
Actual Arrival	Actual time of arrival at stop
Actual Departure	Actual time of departure at stop
Scheduled Trip Start	Scheduled time trip departs first stop (from GTFS Static)
Trip "End"	Actual time of departure at last known stop
Derived	*Formula*
Time to Prediction	Prediction - Timestamp
Prediction Error	If Actual Arrival ≤ Prediction ≤ Actual Departure, then 0
	If Prediction < Actual Arrival, then Actual Arrival - Prediction
	If Prediction > Actual Departure, then Actual Departure - Prediction
Time to Stop	Actual Arrival - Scheduled Trip Start
Prediction Change	$\lvert Prediction_x - Prediction_{x-1} \rvert$

Time to Prediction

Time to prediction is the difference between the predicted arrival time at a given stop and the timestamp of the update. This length represents the interval between the time the prediction is made and the time of the arrival that is predicted. Given the data cleaning approach presented earlier, time to prediction values remaining in the data set are positive. The time to prediction interval is important because it represents the time theoretically available to a traveler until the arrival/departure of the vehicle – time that might be balanced between productive activities and travel to (and waiting at) the transit stop. Time to prediction is also expected to be negatively related to prediction accuracy (i.e., prediction accuracy increases as time to prediction decreases), and therefore it is useful for adjusting relevant metrics.

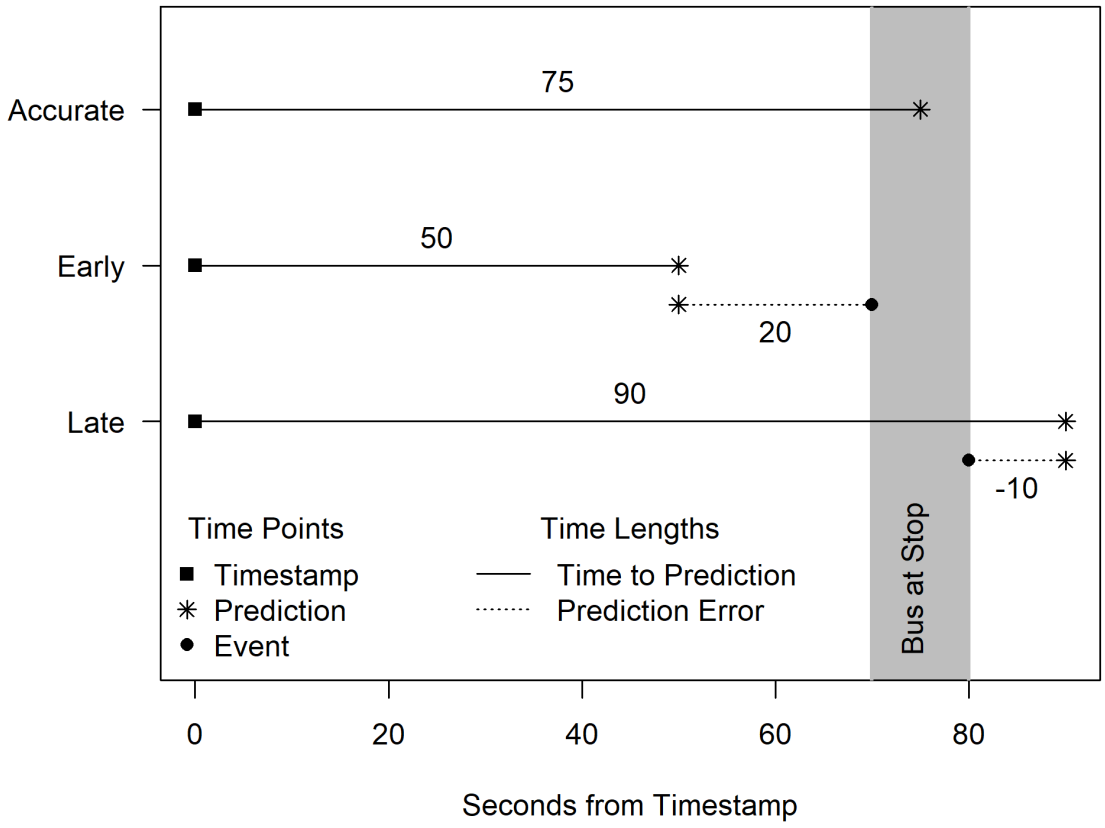
Prediction Error

Prediction error is the difference between the prediction and the actual stop event defined by the actual arrival time and the actual departure time (Machlab et al., 2017). Four of the five studied GTFS-RT feeds report the same exact time for the actual arrival and actual departure time. By contrast, the more fulsome MAX feed reports distinct actual arrival and departure times. While

the calculation of prediction error is the same for all GTFS-RT feeds, this calculation is best illustrated on the MAX data, which enables all three possible forms shown in Table 4.

The prediction error calculation is from the perspective of the traveler who follows the prediction exactly. If the predicted arrival occurs between the actual arrival and the actual departure (or exactly at these times), the prediction error is considered to be zero, an accurate prediction—the traveler experiences the bus exactly when advised it would arrive. If the predicted arrival occurs before the actual arrival, the prediction error is positive, an early prediction—the traveler experiences wait time at the stop before the bus arrives. In the last case, when the predicted arrival occurs after the actual departure, the prediction error is negative, a late prediction—the traveler arrives at the stop after the bus has already left. These three options are presented visually in Figure 2.

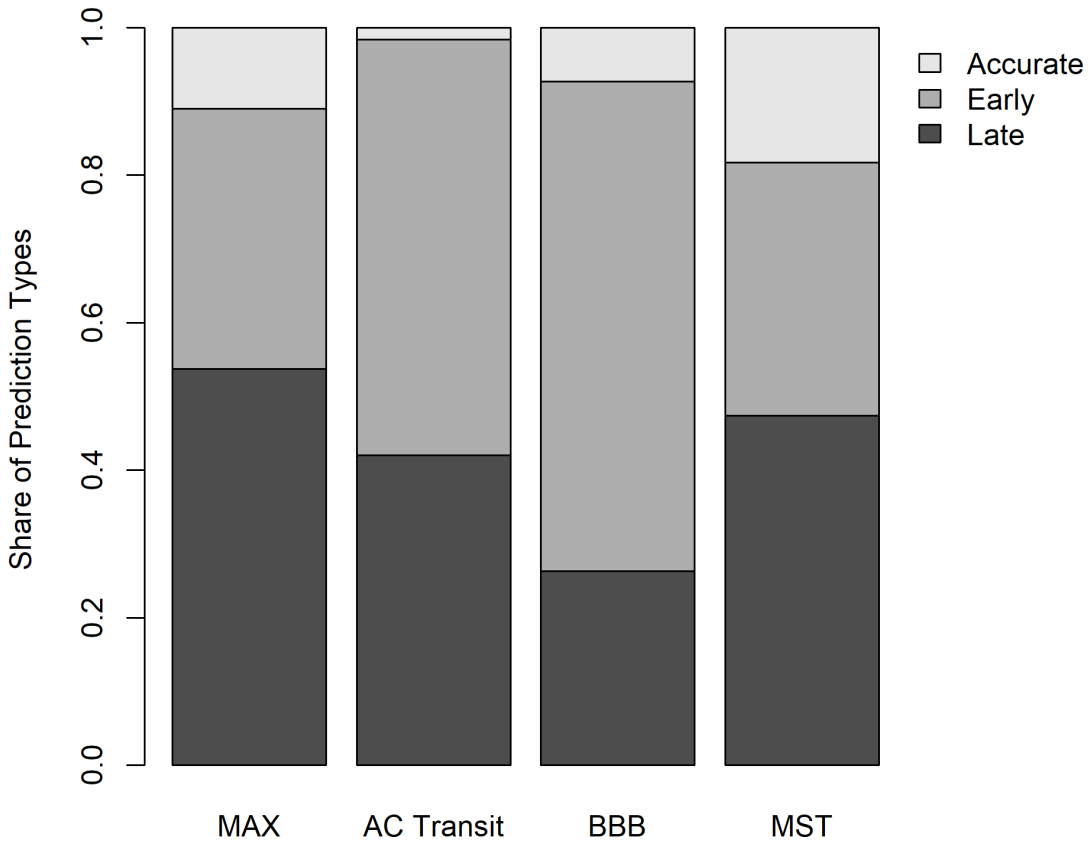
Figure 2. Graphical Depiction of Prediction Types



All three types of prediction errors are witnessed on the data collected from the four studied transit systems, as shown in Figure 3, although the relative shares vary widely. Notably, MST has the most even distribution of the three prediction types. This outcome is likely as all the prediction values provided in the MST *TripUpdate* feed are reported as whole minutes, providing a sixtieth of the granularity of the values in the remaining three systems (which report all values to the

second). Another notable finding from Figure 3 is the low share of late predictions (i.e., negative prediction errors) in BBB. This finding suggests that predictions from BBB, if precisely heeded, are less likely to result in missed buses.

Figure 3. Share of Prediction Types for Studied Transit Agencies



Time to Stop

Time to stop is the difference between the actual arrival and the scheduled trip start. This value represents the time interval for which predictions are available for a given stop on a given trip. Time to stop is used to understand the variation in prediction accuracy over time.

Prediction Change

Prediction change is the absolute value of the difference between one prediction and the subsequent one for a given stop on a given trip. This value is important for understanding the readings experienced by transit customers watching the TripUpdate feed. Large prediction changes frustrate customers.

2.5 Assessment Metrics

The downloaded and derived values enable the consideration of metrics to assess the accuracy of the predictions. The selection of metrics was guided by several overriding principles. Metrics that reflect the transit rider experience were favored (as improving that experience is the primary rationale for public investment in developing GTFS-RT feeds). Metrics were (generally) structured to avoid binary categorizations of stop arrival predictions as either good or bad. Metrics were aimed to measure accuracy directly, and therefore allow a transit property to constantly assess the quality of their predictions. Metrics were chosen to be easily understood without extensive math or statistical training. Metrics were calculated at the stop or trip level but could also be aggregated to one or more component aspects of a transit system, such as mode, route, direction, time of day or area of town.

Table 5. Descriptions of Accuracy Assessment Metrics

Metrics	**Description**
Timestamp Conflict Ratio	Ratio of timestamp conflicts removed to thousand rows of cleaned data
Continuity Error Ratio	Ratio of continuity errors to thousand rows of cleaned data
Update Availability	Share of trip minutes with a prediction update
Prediction Error Percentile Plots	Graphic representation of prediction errors with key percentiles
Scaled Prediction Error IQR	Interquartile range of prediction errors scaled by time to prediction
Bus Catch Likelihood	Share of non-negative prediction errors
Expected Wait Time	Mean wait time to catch vehicle if predictions are followed
Prediction Padding	Absolute value of 5th percentile of prediction errors
Prediction Inconsistency	Sum of prediction changes / sum of time to stop values

All the proposed metrics are presented in Table 5. Since the Timestamp Conflict Ratio and the Continuity Error Ratio were presented previously, this section begins with Update Availability.

Update Availability

Update Availability is defined as the share of minutes during a trip for which GTFS-RT predictions are made. It measures the availability of new prediction information for users. This measure is a fundamental concern of agencies who want to know how often predictions are being broadcast to their riders. There can be no consideration of the accuracy of predictions if they are not being shared in the first place.

To derive the Update Availability, the time span of each trip is calculated from the scheduled start time of the trip in GTFS Static to the departure time from the last stop broadcast via GTFS-RT, i.e., the trip "end," even if that point is not the last (or, in MAX's case, penultimate) stop in that trip's stop sequence. That span is cut into one minute bins based on the integer of the minute value. All the trips in the four-system sample had scheduled start times that began precisely at the top of the minute. This feature ensured that all the bins, except the final one, lasted for a full sixty seconds. The timestamp of each prediction is used to allocate those updates to the minute bins. Any bin with even a single associated update is dummy coded as a success. (Even though the final bin is less than a minute long, it always has at least one associated update.) Update Availability is the percentage of total bins coded as a success.

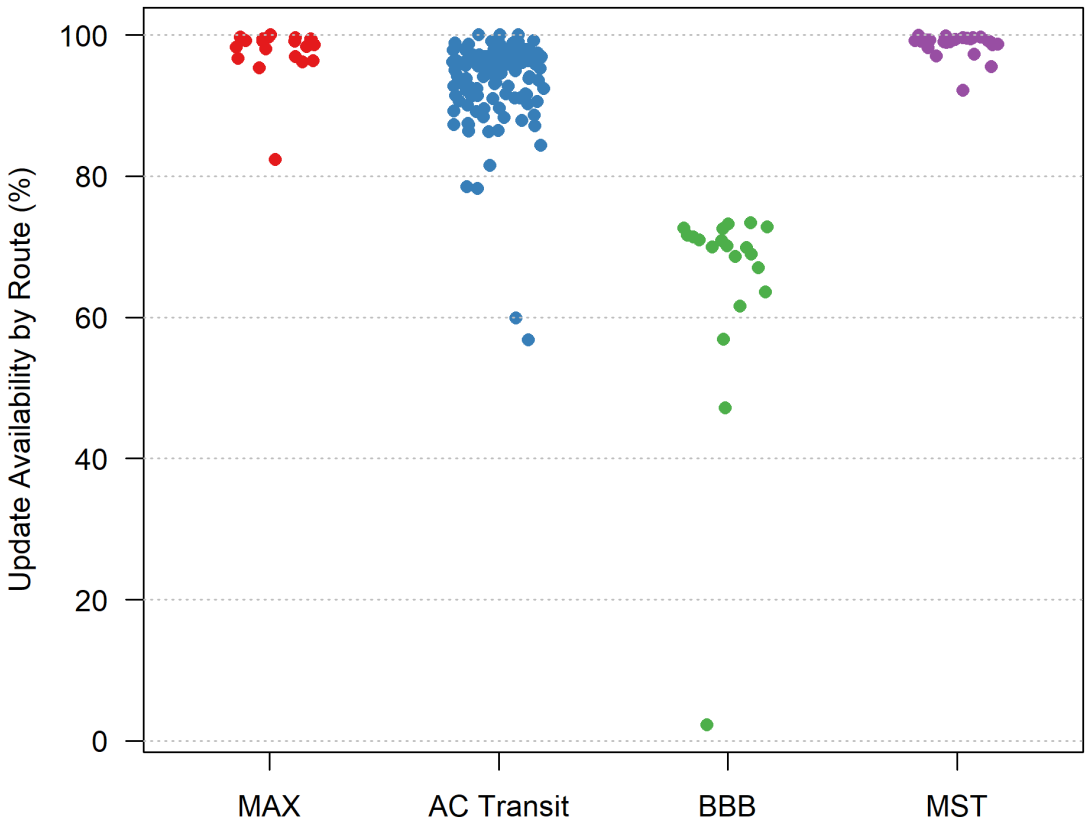
The specific design of this metric raises two key concerns. First, the reliance on the GTFS-RT feed for determining the availability span seems problematic when the goal of the metric is itself to evaluate the availability of the GTFS-RT feed. The specific concern was handling cases when the GTFS-RT feed cut out prior to providing departure information for the last (or, in MAX's case, penultimate) stop in a trip's stop sequence. For example, if actual stop departure information was available for the first five stops of a ten-stop sequence but not further, there is a temptation to revert to the scheduled time of the last (or, in MAX's case, penultimate) stop to capture the entire scheduled time that predictions should be available.

In examining the scenarios, however, it was found that (with only one exception on MAX) whenever the GTFS-RT feed cut out before providing information on the final (or in MAX's case, penultimate) stop in the sequence, the last reported actual arrival time was already later than the scheduled time of the final (or in MAX's case, penultimate) stop. This finding suggested that reverting to the scheduled trip end time in GTFS Static would underestimate the actual span when predictions were available and thus misstate the actual Update Availability. It appeared that the trips whose last reported stops were not the final (or, in MAX's case, penultimate) ones in the sequence were experiencing problems that legitimately caused the vehicles to come offline. For these reasons, the actual arrival time at the last reported stop defined the outer bound of the availability range.

The second concern was the structure of the bins, specifically their size and their impact on metric interpretation. The appropriate bin size to reflect transit user expectations was debated. While customers would prefer more frequent updates, it was felt that a minimum expectation was one update each minute. However, it would not be unreasonable for a transit agency to prefer a smaller bin window for assessing the Update Availability, such as thirty seconds. The use of standard bin intervals across routes of different lengths will cause the same number of missed bins to have differing impacts on Update Availability. Each missed bin has a larger impact on Update Availability for shorter routes than longer routes. Transit agencies should therefore keep route length in mind when interpreting Update Availability.

Figure 4 presents the Update Availability scores by route for each transit system studied as a strip chart. (Points in this chart are randomly jittered horizontally to disambiguate overlapping values, since only the height matters.) For this metric, higher scores are better. These data show that MAX and MST report the highest levels of Update Availability, while BBB reports the lowest. This variation suggests that aiming for Update Availability scores of 95% or higher is attainable; conversely, given this possibility of success, consistently low levels of Update Availability is cause for concern regarding a transit agency's GTFS-RT feed.

Figure 4. Update Availability by Route for Studied Transit Agencies



Note: Each dot represents a single route

The aggregate graphing in Figure 4 is useful for tracking general trends in Update Availability, but a transit agency is likely to want to explore low performing routes more specifically. Table 6 presents the twenty routes from each of the studied systems with the lowest scores of Update Availability. A closer look at the specific routes at the bottom of their transit agency's respective rankings for Update Availability suggests that express routes may present a problem for transmitting GTFS-RT data. (Please note that BBB's Pico Boulevard Express was not running at the time of this study.) Such focused analysis might reveal route geographies or even hardware on a specific vehicle that hamper transmitting *TripUpdate* information.

Table 6. Twenty Routes with Lowest Update Availability (UA) by Transit Agency

MAX	UA	AC Transit	UA	BBB	UA	MST	UA
BART Express	82.4	Crespi Middle - San Pablo Dam	56.8	Pico Blvd Express	2.3	King City - Paso Robles	92.2
Route 35	95.3	Buena Vista - Fruitvale	59.9	Downtown LA Freeway Express	47.1	CHOMP-Monterey	95.5
Stockton Express	96.2	DeJean - Cutting - Macdonald	78.3	Pacific Palisades	56.9	CHOMP- Del Mesa via Carmel	97.0
Route 29	96.3	De Anza - Crespi - Rollingwood	78.5	Barrington Ave	61.6	Santa Rita via Northridge	97.2
Route 38	96.7	De Anza - Crespi - Fairmede	81.5	SMC- 17th St Station- Montana	63.6	Carmel Rancho - Sand City	98.2
Route 41	96.9	MacArthur - Eastmont Transbay	84.3	UCLA - Marina del Rey	67.0	Carmel Valley Grapevine Express	98.6
Route 31	98.0	Korematsu - El Cerrito - Carlson	86.3	Wilshire Blvd/UCLA	68.6	Aquarium/Sand City via Hilby	98.7
Route 44	98.2	Bishop O'Dowd High - Montclair	86.4	26th Street	69.0	Aquarium/Sand City via Broadway	98.9
Route 42	98.3	Skyline High - 35th Ave.	86.5	Main St & Santa Monica Blvd/UCLA	69.9	Veteran's Shuttle	98.9
Route 33	98.6	Newark Memorial - Newark Blvd.	87.1	17th St Sta - SMC Bundy Campus	70.0	Salinas - CSUMB	99.0
Route 23	99.1	Irvington High - Horner Jr. High	87.3	Bundy Dr & Centinela Ave	70.1	Pacific Grove - Carmel	99.0
Route 26	99.1	Montclair - Park Blvd. Transbay	87.3	Wilshire Bl/Bundy Dr-Marina del Rey	70.9	Northridge via Westridge	99.0
Route 25	99.1	Rich - Oak Transbay All Nighter	87.4	Culver City Sta – UCLA	71.0	Watsonville - Salinas via Prunedale	99.1
Route 32	99.3	Mission San Jose - Hopkins Jr.	87.8	Lincoln Blvd/LAX	71.4	Asilomar-Monterey	99.1
Route 36	99.3	St. Mary's College - Montclair	88.3	Ocean Park Blvd & Westwood Bl/UCLA	71.6	Hartnell East Alisal - West Alisal	99.2
Route 30	99.4	Bay Farm - Park St. Transbay	88.3	Olympic Blvd	72.5	Salinas - King City	99.2
Route 22	99.4	High - South Shore Transbay	88.6	Venice/Westwood Sta/UCLA Rapid	72.6	East Salinas-Westridge	99.3
Route 28	99.5	Bishop O'Dowd High - MacArthur	89.1	Pico Blvd Rapid	72.8	Presidio-Marshall Park	99.3
Route 37	99.6	Montera - Oakmore - Seminary	89.2	Pico Blvd	73.2	CSUMB - Marina	99.4
Route 21	99.7	Korematsu - E.C. - No. Richmond	89.5	Lincoln Blvd/LAX Rapid	73.4	Salinas - Alisal - Northridge	99.4

Prediction Error Percentiles Plots

Once predictions are made available, it is possible to assess their accuracy by visualizing the distribution of prediction errors. Due to the tremendous amount of trip update data (as shown in Table 2), it is recommended to separate prediction errors into negative (i.e., late) and non-negative (i.e., accurate and early) predictions, identifying percentiles for each set, and plotting those as reversed percentiles (which places outliers on the bottom of the graph) as shown in Figures 5 and 6.

A useful analogy to understand this visualization is the cross-section of a mound created by a person tossing thousands of beanbags at a target on the ground. A person with incredible coordination and focus might toss all the beanbags close to the target, resulting in an almost perfect column at the target location. Another person with normal coordination and focus will make more of a hill of beanbags, gently sloping down from the center outward. A cross-section of the first mound will reflect the high accuracy of the tosses and hew closely to the target location. A cross-section of the second mound will reflect the lower accuracy of the tosses and demonstrate greater dispersion from the target. For ease of presentation, these cross-sections are split into throws that exceed the target (i.e., negative prediction errors) and throws that met or were shy of the target (i.e., non-negative prediction errors).

This approach not only collapses the data management requirements by reducing the data set to one hundred rows (or fewer), but also scales the data in a consistent and meaningful way that facilitates comparison independent of the number of observations. The curves in Figures 5 and 6, for example, compare the distribution of prediction errors across the four studied systems. This visualization facilitates interpretation. Since more accurate predictions reduce prediction error, the ideal outcome would be for the curves to hug the y-axis at zero.

To complement the visualization with specific numbers that are easier to track over time, specific reversed percentiles can be recognized as discrete performance metrics. Figures 5 and 6 denote median and 10^{th} percentile values of prediction error. The median is always an excellent measure of "typical" experience as the middlemost value. However, since the median is insensitive to extreme values and passengers are disproportionately aggravated by those high prediction errors, transit agencies might also select a percentile that both expressly marks such high values and has a reasonable potential to be affected by intervention. Each transit agency faces its own distinct reliability challenges that complicate prediction accuracy and will need to mark that high value percentile for themselves. This research proposes the 10^{th} percentile as a possible metric, but agencies might also select a more rigorous fifth percentile measure.

Figure 5. Negative Prediction Error Percentiles Plot

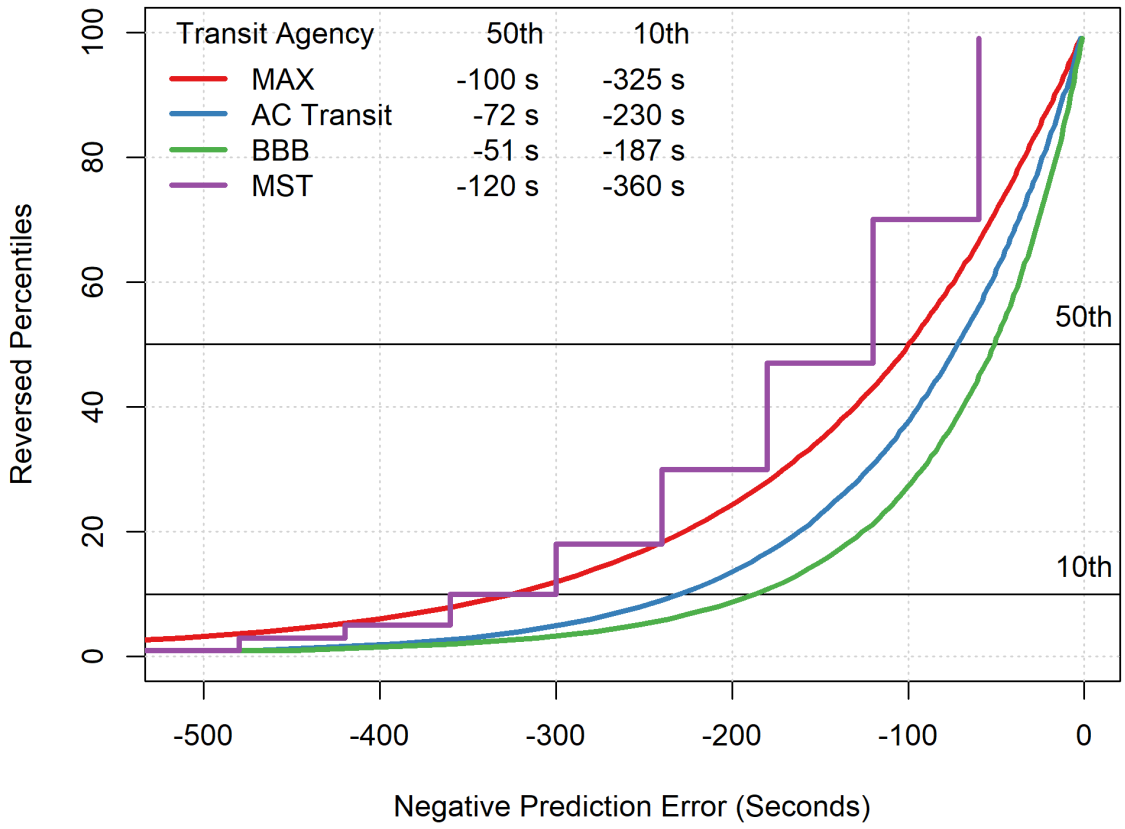


Figure 5 presents the negative prediction error percentiles for the four studied systems. Negative prediction errors refer to predictions that, if followed exactly, would result in a passenger arriving at the stop after the bus has already departed. The curves in Figure 5 demonstrate that BBB reports the smallest negative prediction errors, followed in ascending order by AC Transit, MAX, and, finally, MST. Figure 5 also presents the median and 10th percentile values of the negative prediction error as a discrete metrics of prediction accuracy. The median negative prediction error for BBB is -51 seconds while the same value for MST is more than twice as large at -120 seconds. The 10th percentile for MST is -360 seconds, or six minutes.

MST percentiles are distinct from the other three curves because they are graphed as steps rather than as straight lines between each successive percentile. This distinction is made because the *TripUpdate* messages from MST only provide predictions in whole minutes, rather than in seconds. Consequently, all the quantile breaks for MST occur at minute intervals. For this reason, when using GTFS-RT *TripUpdate* messages exclusively to assess accuracy, there is no negative prediction error between zero and -60 seconds and for the more common values of negative prediction error, the MST data demonstrate relatively slight variation. For example, 30% of messages with negative prediction error are off by exactly a minute, and a fifth are off by exactly

two minutes. By contrast, the other three systems, which have sixty times the granularity of MST, report distinct values for almost every percentile. While it would be most accurate to graph these percentiles as steps as well, they can be graphed simply as lines, which improves readability without unduly compromising interpretation.

Figure 6. Non-Negative Prediction Error Percentiles Plot

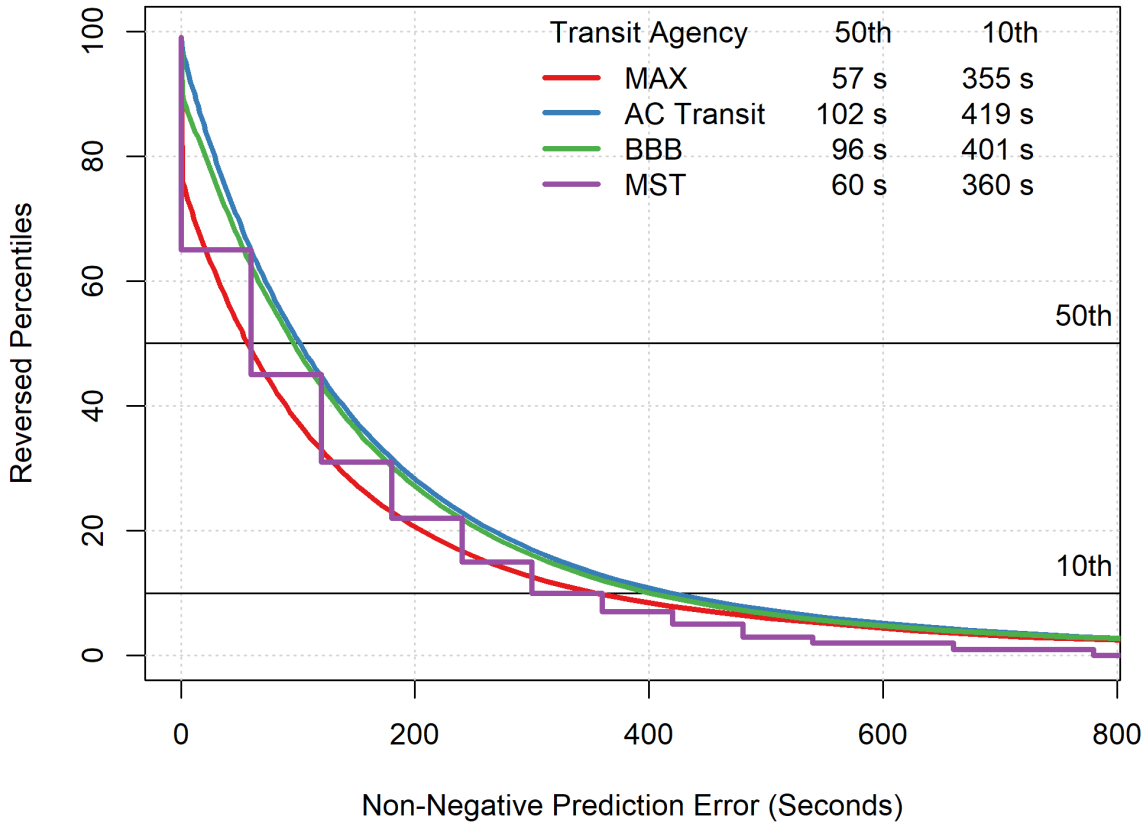
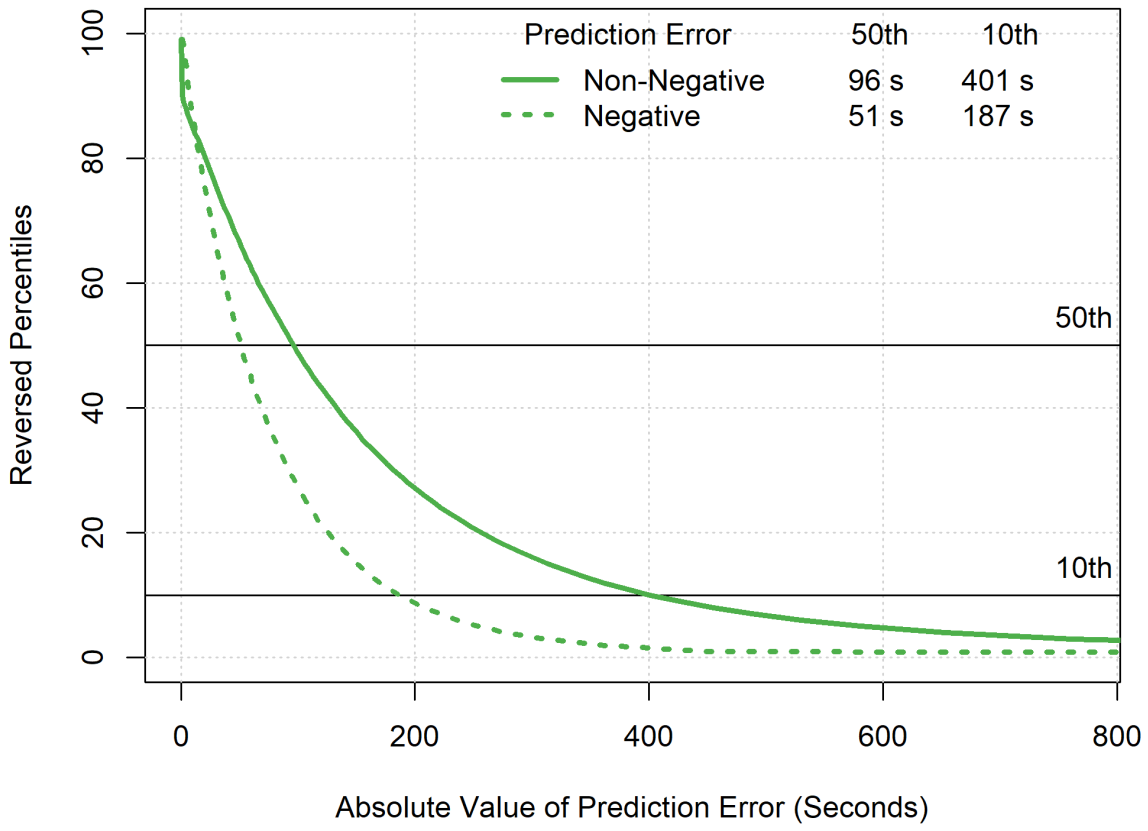


Figure 6 presents the non-negative prediction error percentiles for the four studied systems. These values represent how long riders following predictions will need to wait at the bus stop for a bus to arrive. These predictions would not result in missed vehicles, just potential waits at the stop. Here the curves present a different picture of accuracy, with MAX being the most accurate, followed in ascending order by MST, BBB, and AC Transit. The median non-negative prediction error on MAX was less than a minute, while the same value for AC Transit was almost 80% larger. While all systems have some prediction errors with a value of zero (as discussed earlier for Figure 3), these predictions are particularly visible for MAX (which provides distinct actual arrival and departure times) and MST (which only reports *TripUpdate* information at the whole minute level). Of these two conditions, only the former is likely to result in no waiting time. Taking the ratio of the 10th percentile value to the median value for these curves can offer an additional accuracy metric. Smaller values suggest less slope (and therefore less loss in accuracy) between the two percentiles.

A comparison of those ratios shows that such ratios for the negative prediction error percentiles were typically just over three, while those same ratios for the non-negative prediction percentiles were typically four (although in MAX's case more than six). The finding that negative predictions are more accurate is not surprising, since buses are more likely to lose time along a route than to make it back. However, these differences are useful to understand the GTFS-RT messages riders are receiving.

Figure 7 presents both negative and non-negative prediction error percentiles for BBB on a single chart. This visualization is especially useful for an agency seeking to understand its prediction accuracy. This chart shows that negative predictions errors are smaller than non-negative ones. Transit agencies might track how these curves compare over time. In any case, it is important to understand that the curves represent different absolute numbers of predictions. For BBB, there are approximately three non-negative prediction errors to each negative one. While this context means that BBB riders experience more prediction inaccuracy than if these two curves represented the same number of predictions, it also means that BBB riders are less likely to miss their intended vehicles.

Figure 7. Combined Non-Negative and Negative Prediction Error Percentiles Plot



Scaled Prediction Error Inter-Quartile Range (IQR)

Since it is more difficult to predict events farther into the future, it is reasonable to expect that arrival predictions with longer associated time to prediction values would also have higher prediction error magnitudes. For example, a prediction that a given vehicle will reach a given stop in an hour is more likely to have a higher prediction error than a prediction that the same vehicle will reach the same stop in five minutes. The latter prediction has fewer minutes for events to occur that might affect prediction accuracy. These expectations were confirmed in practice by graphing the data and also undergird the existing practice of considering prediction accuracy (Machlab et al., 2017).

To reflect this impact on prediction accuracy, a scaled error measure was created by dividing the prediction error by the time to prediction value. This measure reinterprets the prediction error as a share of the associated time to prediction. In other words, a prediction that a bus will arrive in half an hour that is off by six minutes has a scaled value of one fifth—the same scaled value for a prediction that a bus will arrive in ten minutes that is off by two minutes.

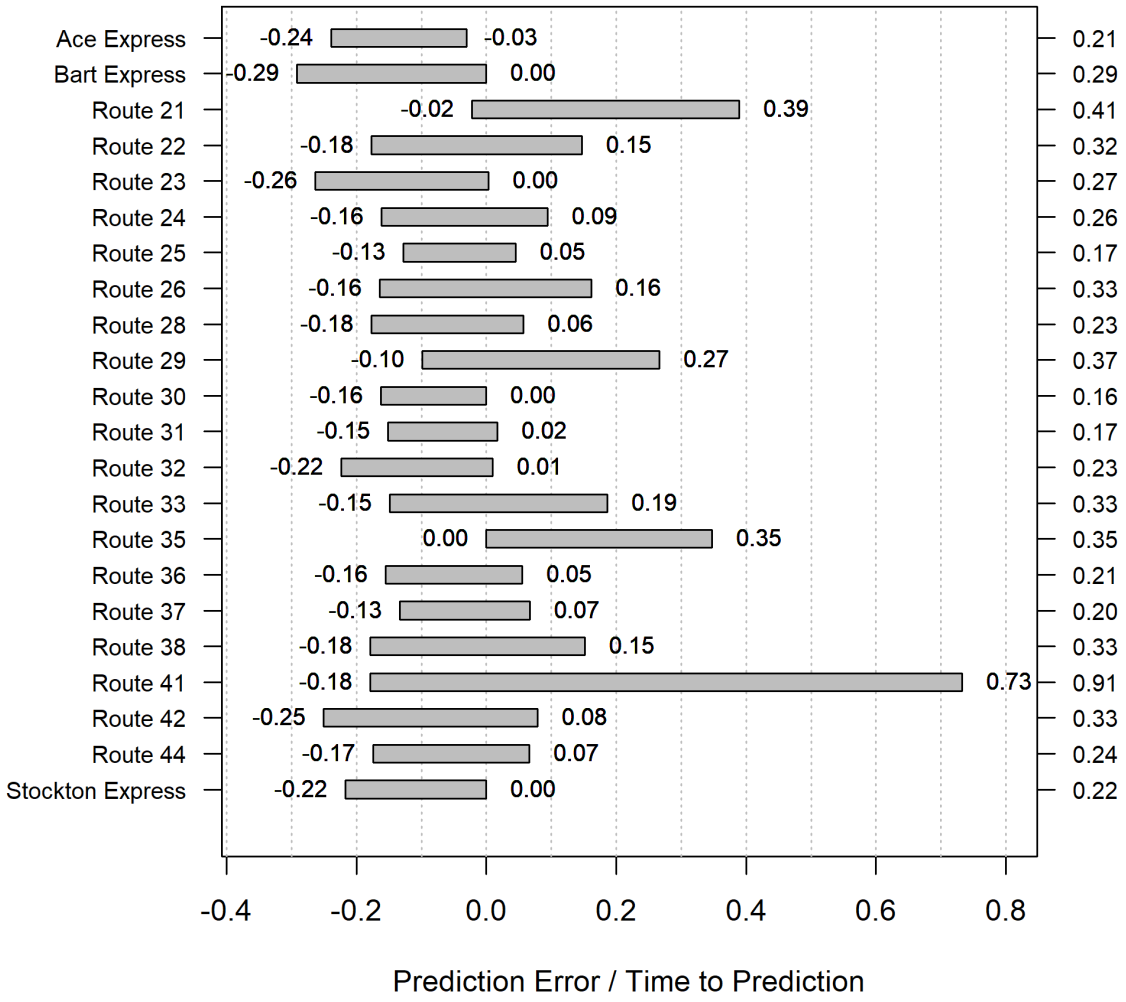
Using a scaled measure of accuracy introduces two problems. First, when the bus is expected to be at the stop for a given timestamp, the time to prediction would be zero; however, a zero value in the denominator of the scaled error measure would yield an undefined value. To avoid this problem, any zero value for the time to prediction is recoded to equal a single second before calculating the scaled error measure. If the bus is in fact in the station, the scaled error would be zero. If the bus is not in the station, the scaled error equals the prediction error. Second, a scaled error measure accentuates extreme values when a prediction with a very short time to prediction has a large prediction error. For example, if the time to prediction is ten seconds, but the bus takes three minutes (180 seconds) to arrive, the scaled prediction error would equal 18. If the time to prediction was five seconds, the scaled prediction error for the same prediction would double to 36, an extremely high scaled prediction error. It is the equivalent for a bus predicted to arrive in ten minutes actually arriving in six hours. To both avoid these extreme values and use a well-known measure, this measure is expressed as the interquartile range (IQR) of the scaled error values. The IQR is defined as the difference between the 25 and 75 percentile values. Half of the total predictions fall within the IQR.

Figures 8 through 11 graph the scaled prediction error IQR as bars stretching between the 25 and 75 percentile values. These charts also include the numeric value of this range along the right-hand axis. Each bar represents the range within which the middle half of scaled prediction errors fall. Accurate predictions produce narrow bars close to zero while wider bars represent reduced accuracy.

Figure 8 displays the scaled prediction error IQR for each of MAX's bus routes. The data along the right axis show that almost half (10 out of 22) of MAX's routes have a scaled prediction error IQR of 0.25 or less. These low values denote good prediction accuracy. A transit agency concerned

about GTFS-RT *TripUpdate* accuracy might explore the IQR values that exceed a given threshold, such as 0.25. One would expect that, over time, this threshold of concern would be lowered as prediction accuracy improved.

Figure 8. Scaled Prediction Error IQR for MAX



Exploring the bus lines with large IQRs might reveal problems with predictions or with the bus routes themselves. For example, MAX Route 21, which has the second largest IQR, is a busy loop route in the heart of Modesto. It is possible that local traffic results in highly variable bus speeds that confound prediction algorithms. Instead of focusing on better prediction accuracy for this route (which may not be easily accomplished), MAX might seek to improve service reliability by investing in additional infrastructure, like dedicated bus lanes or transit signal priority.

The bars in Figure 8 do show that while the 75th percentile is often very accurate, the 25th percentile values are often substantially negative, meaning the user would miss the bus. In many cases those

scaled prediction error magnitudes are around 0.2, meaning that a prediction that is 10 minutes out will cause the user to arrive two minutes too late. MAX may seek to adjust its prediction algorithm to skew less negative.

Figure 9. Scaled Prediction Error IQR for AC Transit

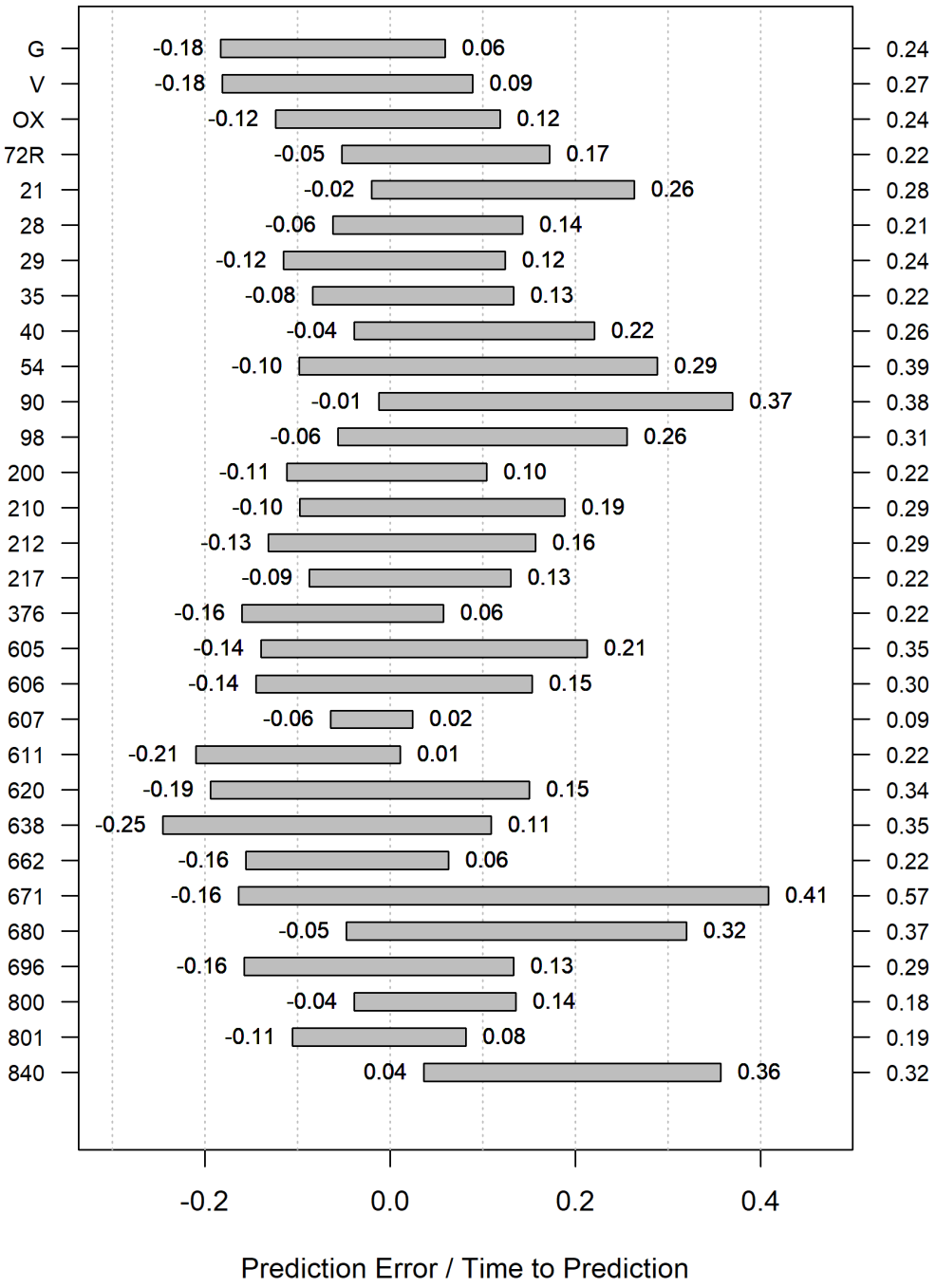


Figure 9 presents 30 randomly selected routes from AC Transit's full set of 130 routes. The bars in this chart are less negatively skewed than for MAX; furthermore, there are no very distinct outliers, as say MAX Route 41 with an IQR of 0.91. Nonetheless, it is important to recognize that these data only represent a sample of lines, and the ideal use of this metric is to explore on a line-by-line basis to identify issues of concern.

Figure 10. Scaled Prediction Error IQR for BBB

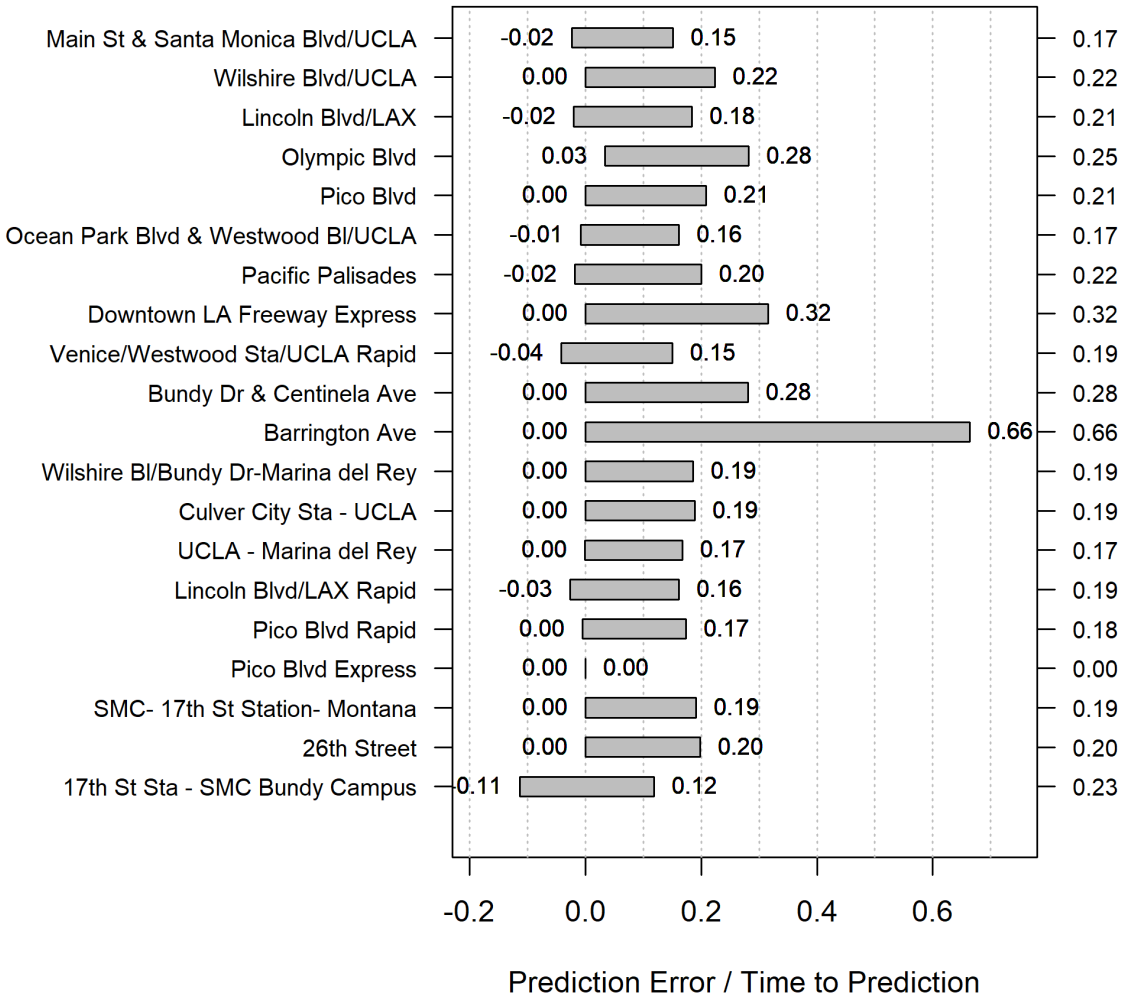


Figure 10 displays the scaled prediction error IQR for each of BBB's 20 bus routes. Compared to MAX and AC Transit, these IQRs are generally much smaller, demonstrating greater accuracy (despite the outlier route on Barrington Avenue). Furthermore, they are skewed positive, which reduces the likelihood that users relying on stop arrival predictions will miss their intended bus. (Please note the Pico Boulevard Express was not in service during the study period, resulting in no prediction error.)

Figure 11. Scaled Prediction Error IQR for MST

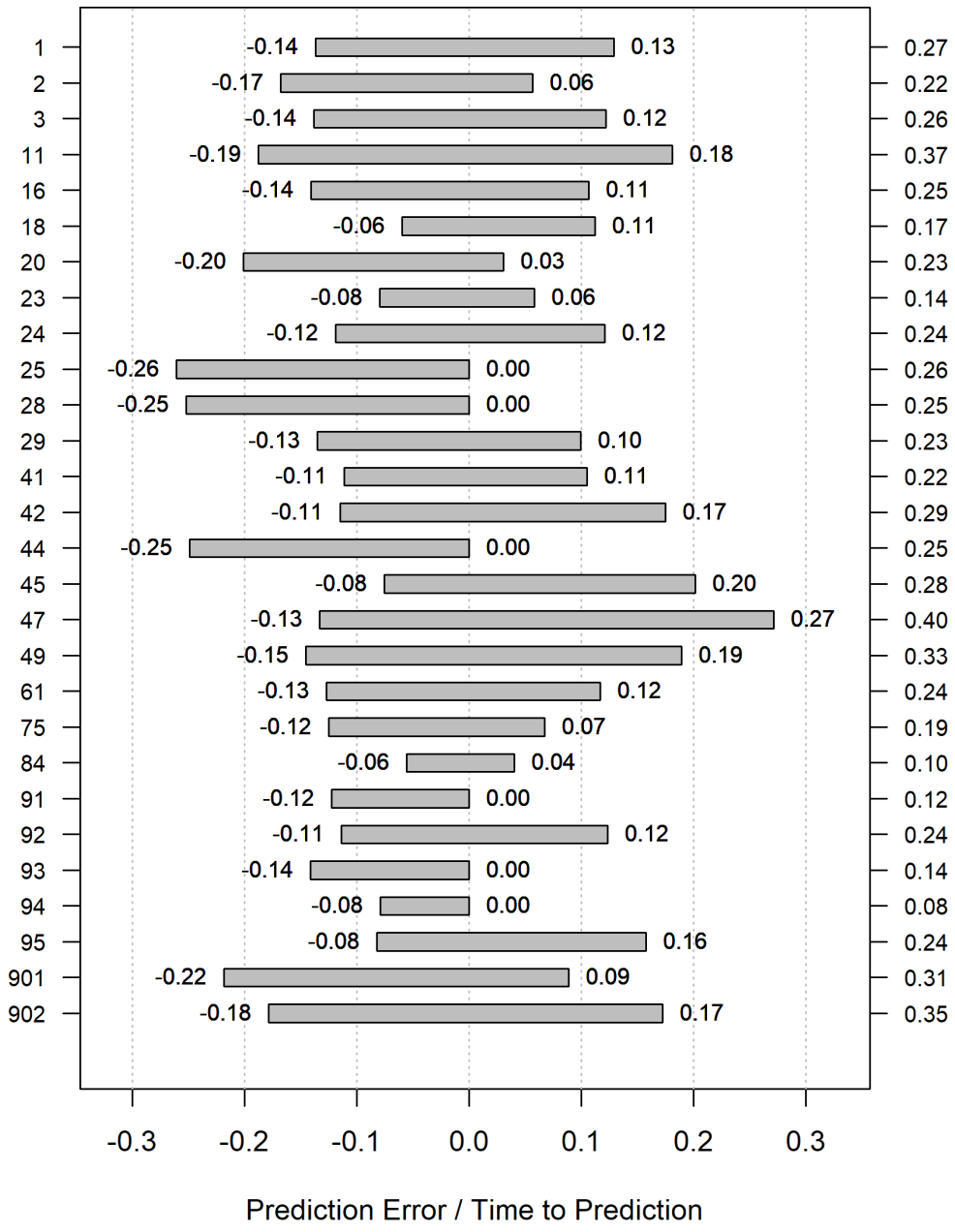
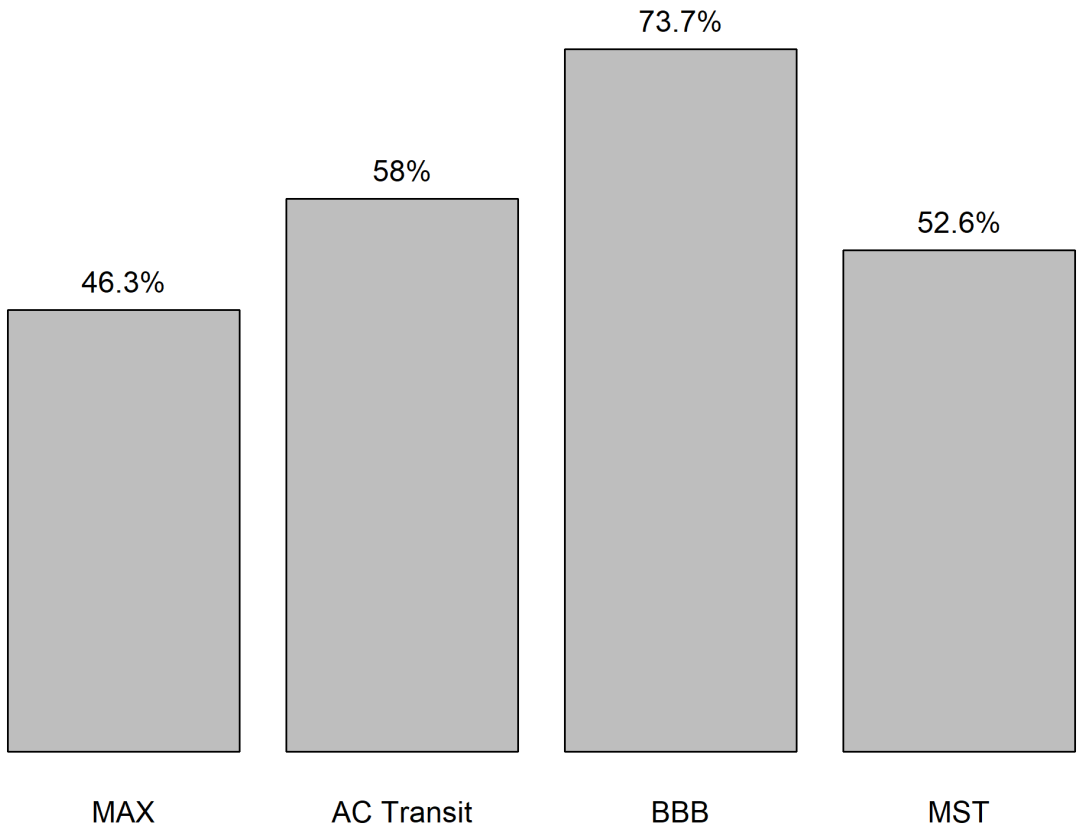


Figure 11 displays the scaled prediction error IQR for each of MST's 28 bus routes. As for MAX, there were several IQR's for which the 75th percentile is close to zero, while the 25th percentile was substantially negative.

Bus Catch Likelihood

The distribution of prediction errors can be better translated into a metric that directly affects riders, namely the Bus Catch Likelihood. This metric measures the percentage of updates that will lead to a traveler making their intended vehicle. This measure is calculated as the share of prediction errors with non-negative values (also presented as accurate or early predictions in Figure 3). By contrast, late predictions with negative prediction errors will lead to a user arriving after the bus has left the stop. This metric is shown aggregated for each transit agency in Figure 12. (A similar graphic might also be generated for each route within a given transit agency.)

Figure 12. Bus Catch Likelihood by System



These data show that in three out of the four systems studied, on average, a traveler following GTFS-RT predictions precisely will make their intended bus. On MAX, however, most *TripUpdate* predictions result in negative prediction errors, which would lead users to arrive at a stop after the bus has left. AC Transit and MST report Bus Catch Likelihoods above 50% but below 60%, while BBB stands out in providing predictions that will result in a user catching the bus almost three-quarters of the time. These findings suggest that it might be reasonable for vendors of GTFS-RT data to adjust their algorithms to encourage early predictions to increase their Bus Catch Likelihood.

Expected Wait Time

Extending the logic of the Bus Catch Likelihood leads to an additional metric, Expected Wait Time. This metric measures the mean wait time at a stop that occurs should each prediction be followed precisely. Wait time is calculated as the time interval between the prediction and the actual arrival time of the first bus on the desired route that is at or after the prediction. For accurate or early predictions, the wait time is simply the prediction error, but for late predictions, the wait time is the time from that prediction until the actual arrival of the next bus making the same stop sequence. Expected Wait Time weights prediction accuracy by bus frequency to reveal the impacts of poor predictions. The measure represents the actual experience of users who follow the GTFS-RT predictions.

When aggregating data from many routes, Expected Wait Time is sensitive to very high frequencies (which reduce waits) and very low frequencies (which increase waits). For this reason, this measure is most productively applied to single routes or clusters of routes of the same type (i.e., locals, express, or shuttles). To demonstrate that these metrics are useful at other aggregations, Figure 13 presents Expected Wait Time by hour for each of the four studied systems. This chart visualizes the average experience if each GTFS-RT message was translated into an actual attempt to board that bus.

Figure 13. Expected Wait Time for Buses by Hour of the Day

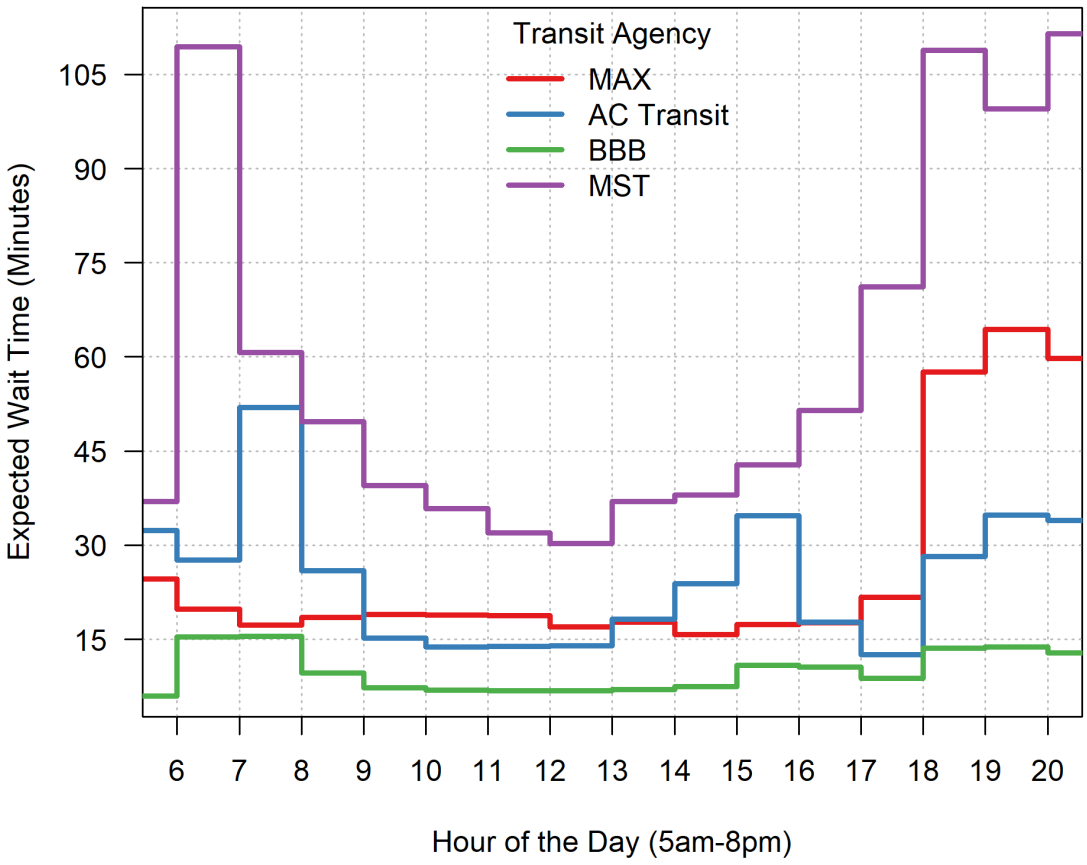


Figure 13 is censored to show only the data from five o'clock in the morning through eight o'clock in the evening, inclusive. The other hours are excluded from this graphic to improve core hour legibility, since the Expected Wait Times for the off-peak and commuter services during those low frequency times are so high as to dwarf the rest of the day. (A full consideration of Expected Wait Time should include these excluded service hours.) There is a clear pattern that Expected Wait Times are high in the morning and evening when the consequences of missing a low frequency bus are high. These values are particularly high for MST, which never reports an hourly average Expected Wait Time of less than 30 minutes. This reflects the impact of a low Bus Catch Likelihood on a low frequency service. Similarly, AC Transit shows a peak Expected Wait Time between seven and eight o'clock in the morning, after which there is a break in service for their many school and commuter bus routes.

To reduce the impact of very low frequency buses (i.e., commuter services), hugely delayed trips (i.e., due to severe external events), or the last trip of the day before service ends, the following is a moderated Expected Wait Time metric that excludes any wait values of more than two hours. This two-hour wait threshold was selected to reflect the outer bound of typical service headways. Transit agencies are encouraged to set cutoffs that are relevant for their services.

Figure 14. Expected Wait Time for Buses by Hour of the Day (Two Hour Limit)

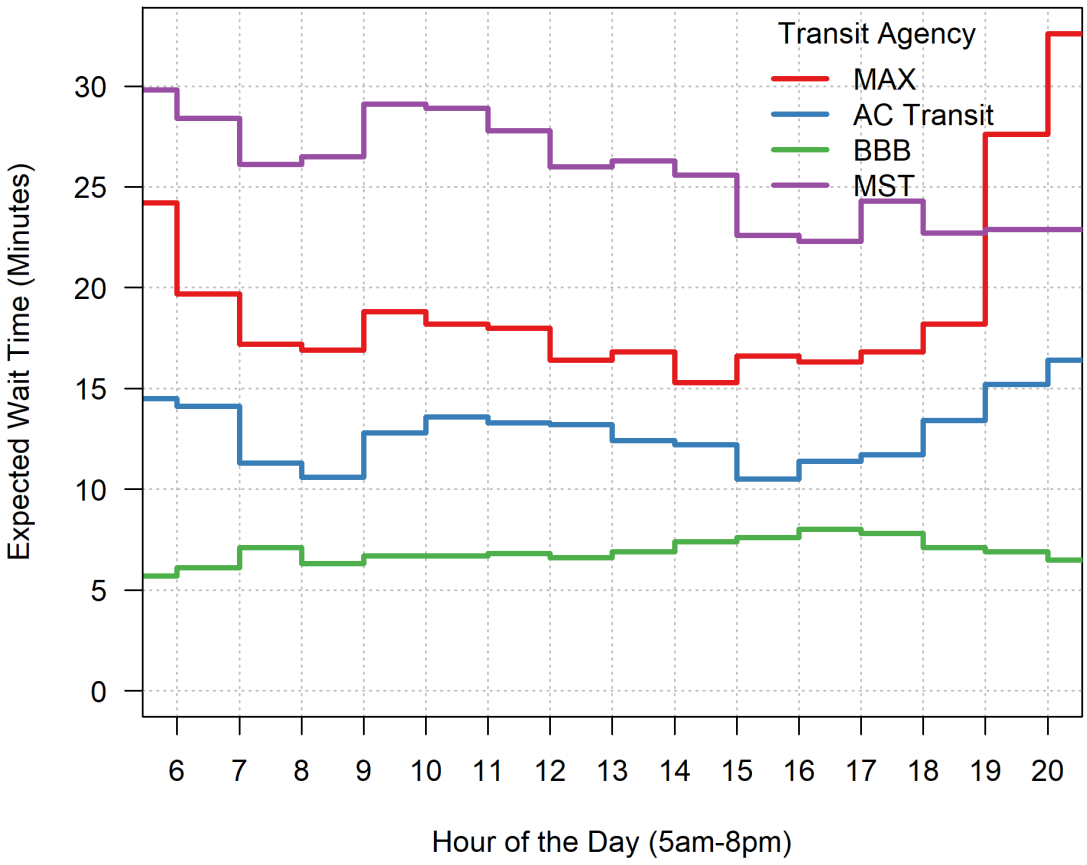


Figure 14 presents the data for this moderated Expected Wait Time using the same set of hours as Figure 13. By excluding any wait time greater than two hours, the resulting visualization has much higher granularity. Surprisingly, there is very little overlap between the systems, with only MAX and MST crossing lines after seven o'clock in the evening. BBB had very low Expected Wait Times between five and ten minutes, with AC Transit just a bit higher between ten and (mostly) 15 minutes. Even without any long waits, MST was alone in reporting no hours with Expected Wait Times less than 20 minutes.

Prediction Padding

Passengers can reduce their own Expected Wait Time by padding the GTFS-RT predictions to increase their chance of making their intended bus. While there has been limited research on how precisely passengers take GTFS-RT predictions, it is likely that regular users naturally begin to handicap the predictions. The size of such a handicap is itself a metric of prediction accuracy, which is systematized as Prediction Padding.

Prediction Padding is defined as size of the buffer (in terms of the share of the time to prediction or the actual number of seconds) that must be added to the GTFS-RT prediction to ensure that the passenger will make their intended bus 95% of the time. Prediction Padding is the absolute value of the fifth percentile (not reversed) of either the scaled or unscaled prediction error for a given stop or route. Prediction Padding is inversely related to prediction accuracy.

Figure 15. Prediction Padding by Route

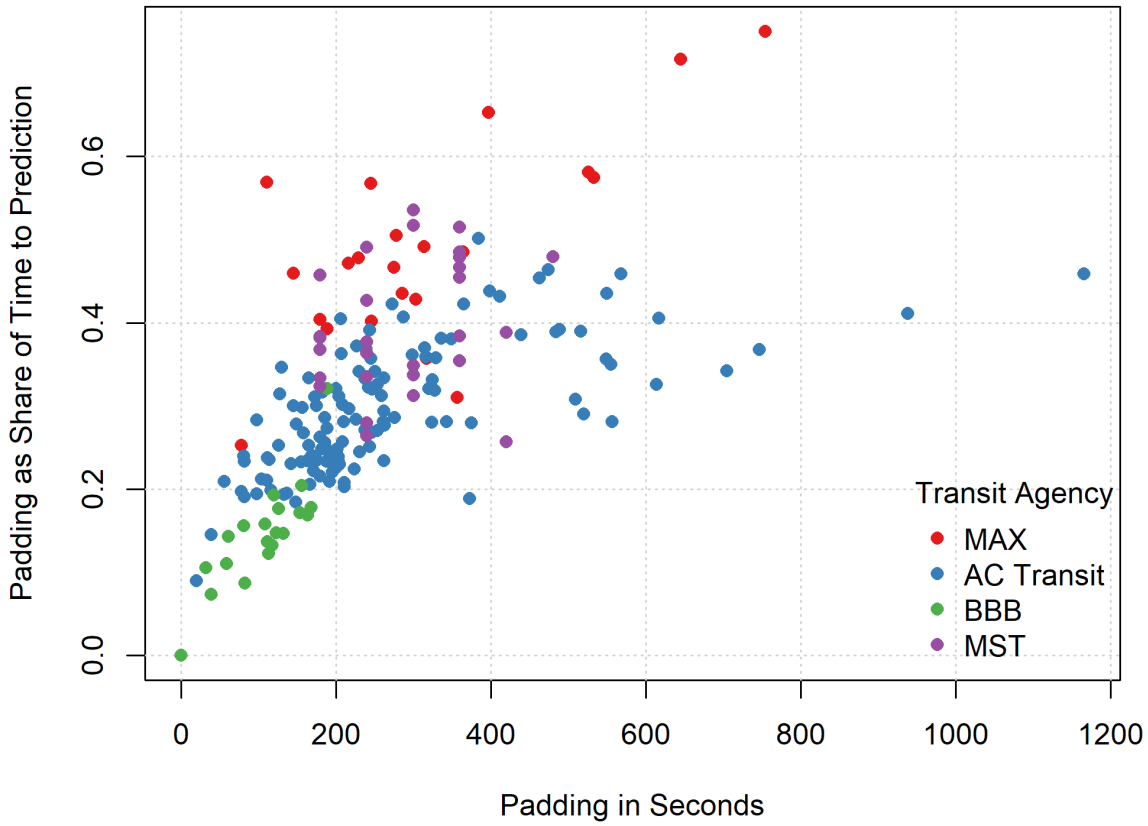
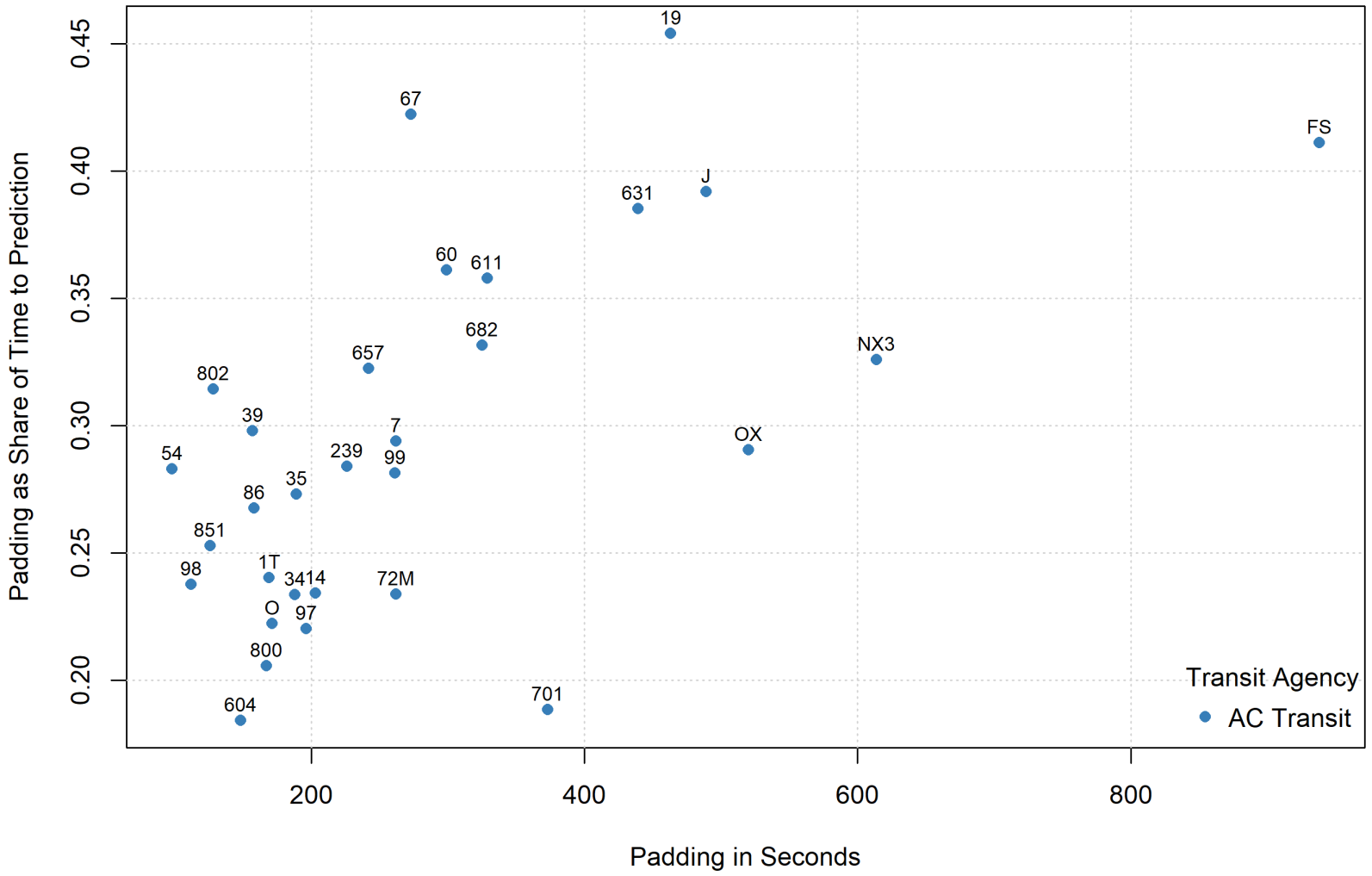


Figure 15 presents a scatterplot of the Prediction Padding values in both relative and absolute terms for all the bus routes among the four studied systems. This graph demonstrates that the predictions for BBB are far more accurate than the other three systems with no absolute Prediction Padding larger than 200 seconds and few relative Prediction Padding values of more than 20% of the time to prediction. By contrast, other systems see the need to add ten minutes or more to be sure to catch certain buses. Similarly, for several routes, passengers should add a buffer to the prediction of half the time to prediction value. To clarify, if the GTFS-RT data reports the bus will arrive in ten minutes, prospective passengers need to be there in five to ensure they will make that vehicle.

While an unlabeled comparison plot such as Figure 15 can be especially useful for understanding how different transit agencies fare in comparison to each other, for intervention purposes it is preferable to label each point by route. Figure 16 plots the Prediction Padding values for 30 routes randomly chosen from among the AC Transit network. These data show that the routes with very high absolute Prediction Padding (i.e., values of eight minutes or more) are entirely the Transbay routes. These routes, referred to by letters and not numbers, are subject to tremendous travel time variation due to their extensive use of highway infrastructure (including the Bay Bridge) where speeds can vary widely based on traffic conditions. This observation highlights the utility of considering Predication Padding (and prediction accuracy in general) by route type.

One important type of route for any transit system are locals, which on AC Transit are denoted by numbers below 400. A clear outlier here is Line 19, which connects downtown Oakland with the Fruitvale BART station via the island community of Alameda. The absolute Prediction Padding is more than seven minutes, while the relative padding is 45% the time to prediction.

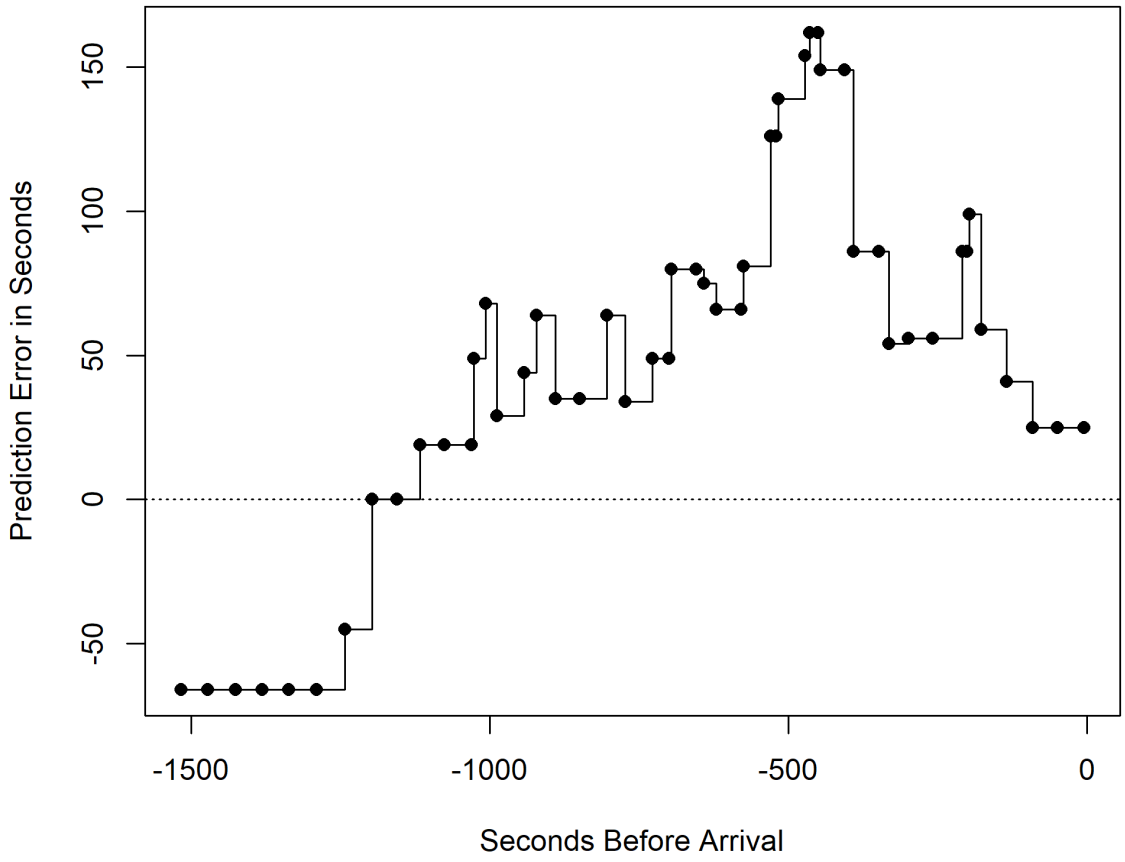
Figure 16. Prediction Padding by Route on a Sample of AC Transit Routes



Prediction Inconsistency

The final proposed metric reflects passenger frustration with vertical arrival predictions that jump around. Figure 17 graphs the prediction error for successive predictions for a single MAX trip on a single stop in Modesto (Route 22 at McHenry and Bowen Avenues). The data demonstrate that in many cases, successive predictions reported significant changes in the expected vehicle arrival time. Such "jumpiness" can be viewed as a form of prediction inaccuracy. A Prediction Inconsistency metric that sums all the prediction changes for a given stop and divides these by the time to stop is proposed. For the example in Figure 17, this metric is calculated as a total of 679 seconds of prediction change over a 1,517 second time to stop, or 0.45. For an entire trip, the sum of all prediction changes for all stops is divided by the sum of all time to stops for all stops.

Figure 17. Consecutive Prediction Error Example from MAX Route 22



Prediction Inconsistency is analogous to guidance for hikes in mountainous areas that advertise the cumulative change in elevation for the horizontal distance traveled. It is technically a measure of slope where the steepness of the incline represents user annoyance with predictions. A Prediction Inconsistency score of zero reflects no slope (i.e., perfect consistency and no annoyance), with increases in slope demonstrating more inconsistency (and more annoyance) for the passenger. A Prediction Inconsistency score of one represents changes in prediction times that equal the times to stop. In hiking terms, that would be an incline of 45 degrees, or climbing stairs.

Figure 18. Prediction Inconsistency by Route

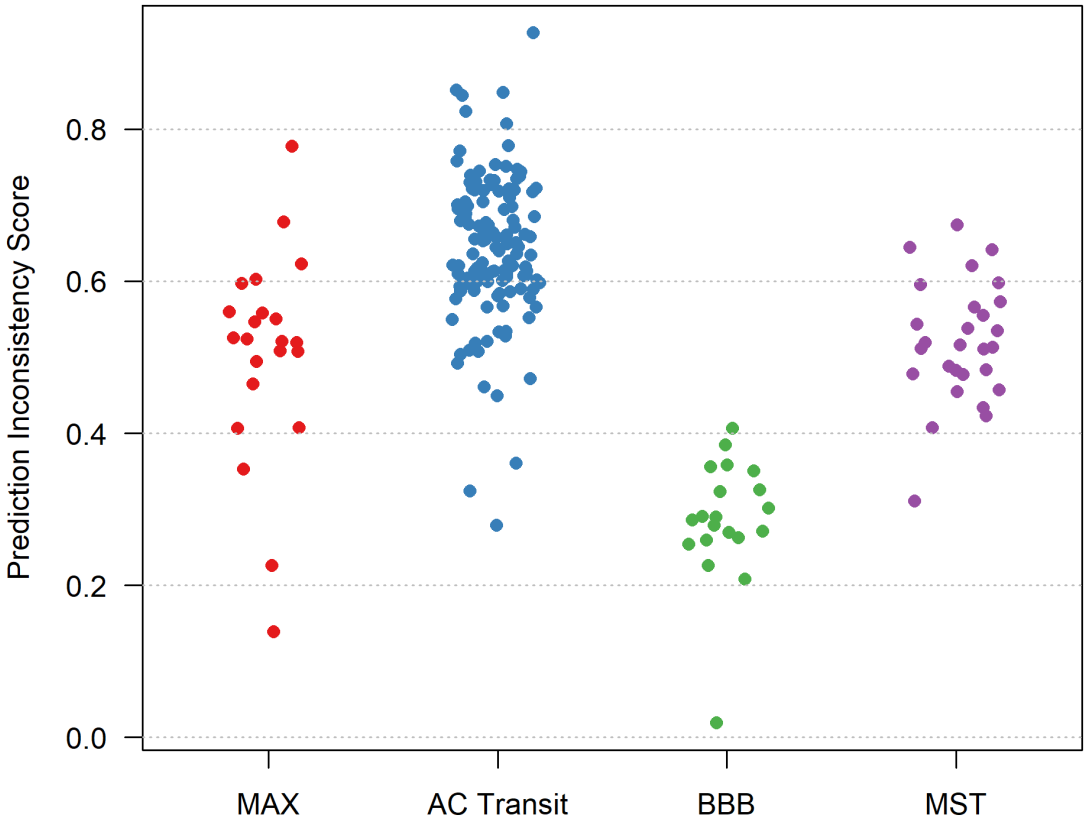


Figure 18 presents the Prediction Inconsistency metric for each of the routes in each of the studied systems. BBB demonstrates more consistency and AC Transit demonstrates less consistency compared to the other systems. While most systems have clustered Prediction Inconsistency scores, MAX's scores are more dispersed.

Figure 19. Prediction Inconsistency by Route for MAX

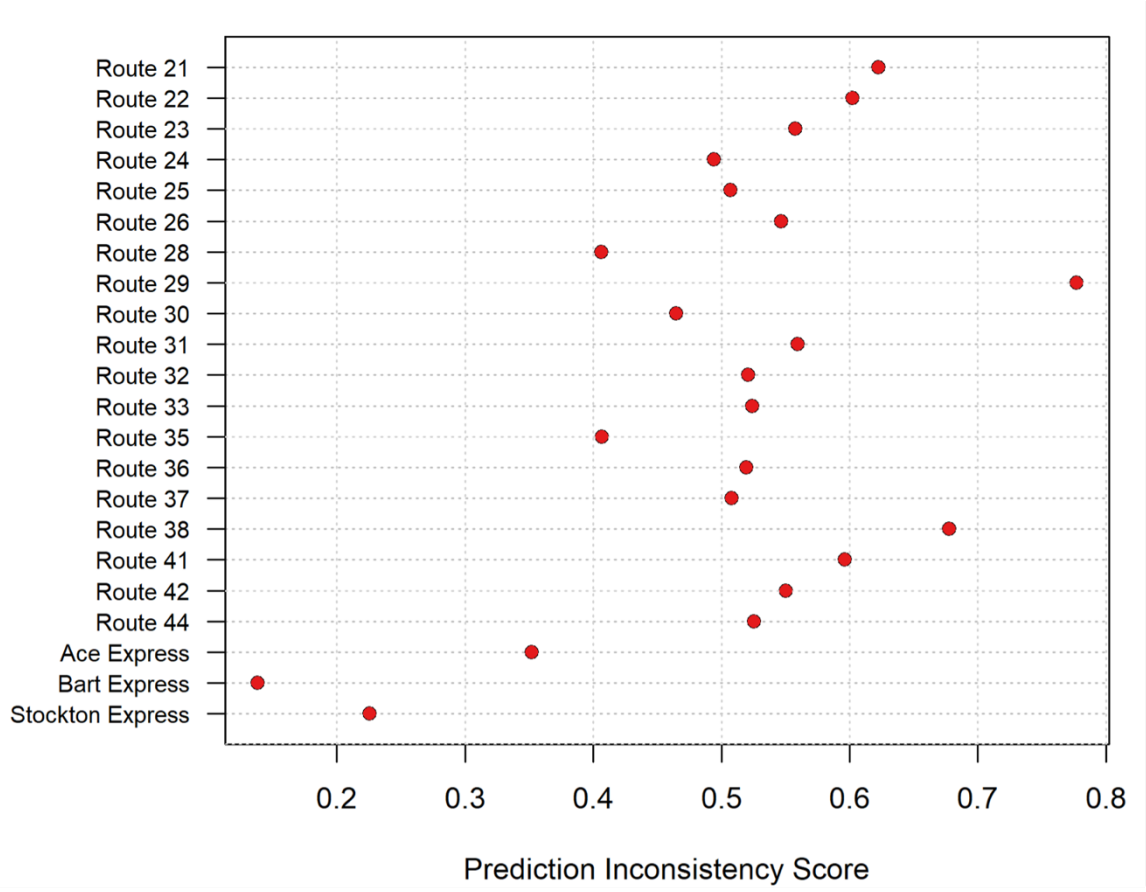


Figure 19 breaks out the Prediction Inconsistency score for each of MAX's routes. This chart turns Figure 18 sideways to better understand the distribution among MAX routes. Notably, the express routes with very few stops and long times to stops show little Prediction Inconsistency, while Route 29, a tremendously complicated route with two (and sometimes three) sub-loops, shows much Prediction Inconsistency.

2.6 Discussion

This research presents systematic approaches to considering GTFS-RT prediction accuracy. This presentation begins with a structured cleaning of the raw data to ensure that appropriate predictions are measured for accuracy. This cleaning has two waves. The first wave eliminates duplicates and establishes a consistent period for assessment. The second wave removes predictions from before the trip is scheduled to start, ensures there is only one prediction for a given stop for a given timestamp, limits predictions to those made prior to the last known departure event, deletes redundant updates for stops that have already been passed, and expunges continuity errors. These cleaning waves are essential for yielding a consistent and comparable data set. Effective accuracy assessment begins with a cleansed data set. This cleaning results in two metrics of GTFS accuracy,

namely the incidence of timestamp conflicts and continuity errors. These measures are useful for identifying problems in the GTFS-RT feed.

The cleansed data set is then used to derive additional values necessary to calculate the proposed accuracy metrics. These derived values are represented in seconds and measure the time to prediction, prediction error, time to stop, and change across subsequent predictions. By convention, negative prediction errors represent late predictions, while positive prediction errors represent early predictions. The derivation of prediction errors presented is robust to GTFS-RT feeds that include distinct (rather than identical) departure and arrival times for a given trip and stop.

The core of this work offers statistical and visual means for transit agencies to identify the accuracy of the predictions they are sharing with the public. These measures are designed to overcome existing reliance on thresholds to encourage continuously variable approaches to assessing accuracy. These measures are also designed to reflect the experiences of the traveling public. This work is not prescriptive over the specific metrics agencies chose to explore their prediction accuracy, rather it seeks to offer myriad options for understanding the implications of prediction accuracy on transit users. Transit agencies are encouraged to choose among these options (and to tailor the options selected) to best meet their needs. This work is prescriptive in emphasizing the importance of a systematic consideration of prediction accuracy. It is incumbent on transit agencies broadcasting this information to establish a program to track its accuracy. These proposed metrics will help agencies with that task.

3. Path Accuracy

The broad availability of General Transit Feed Specification (GTFS) Schedule and Realtime data has transformed transit trip planning. While GTFS was designed expressly for this purpose, the underlying availability of standardized information on transit provision offers many exciting off-label applications, from estimating transit accessibility to mapping service equity to identifying significant bus corridors.

All GTFS use cases, both on-label and off, depend upon the fundamental accuracy of the underlying spatial data. Despite this critical requirement, surprisingly little has been written on practices for assessing whether the locations reported in the GTFS products delivered to the public accord with reality. There are no clear guidelines for flagging errant vehicle locations or poorly coded route shapes. There is no common yardstick for transit agencies and their GTFS vendors to assess the spatial accuracy of these products. Finally, there is no documented understanding of how existing accuracy issues might be most problematic for downstream users.

Traditional approaches to accuracy assessment are to compare data in question to known quantities. Bulk scales are calibrated with test weights. Mechanical watches are calibrated with digital chronometers. Travel demand model outputs are calibrated with traffic counts. Arranging for such a comparison with an entire transit system is more complicated. A spot check in one location provides little insight into accuracy elsewhere. More comprehensive solutions, such as placing additional global positioning system (GPS) receivers on vehicles or accessing existing automatic vehicle location (AVL) feeds, may require express permission from the transit agency and will certainly require a custom, and therefore expensive, data analysis structure. The coordination and cost considerations of acquiring external ground truthing data reduce the ease with which concerned actors can evaluate GTFS spatial accuracy.

This research offers an alternative approach to assessing GTFS spatial accuracy that requires no coordination with a transit agency nor additional data outside of the standard GTFS Schedule and Realtime feeds. This approach simply exploits the complementary structure of the GTFS feeds to pair actual vehicle pings with scheduled route paths to calculate the resulting ping-path distances. Short ping-path distances suggest higher spatial accuracy, while long ping-path distances suggest lower spatial accuracy. Applying a distance threshold to flag pings as potentially problematic enables statistical and spatial assessment of these discrepancies. This approach, which abandons any external ground truthing, relies instead upon the large volume of vehicle location data to identify areas of divergence between the GTFS Schedule paths and the GTFS Realtime pings. This approach assumes, as an expedient, that areas of ping-path convergence are assumed to be accurate. Observations with ping-path distances above the threshold meet one or more of the following conditions:

1. The path is inaccurate (i.e., not the actual transit route)

2. The ping is inaccurate (i.e., not the location of the transit vehicle)

3. The transit vehicle is not following the actual transit route

This research does not automate identifying which of the above conditions are true. Rather, it offers metrics and mapping techniques to guide analysts to visually identify these underlying conditions and make their own conclusions about the resulting impacts on data use or the implications for policy. It is hoped the accuracy assessment presented here will lead to enhanced expectations for GTFS quality and therefore better base data for the array of users who rely on these valuable feeds.

3.1 Methodology

This quantitative study compares several days of GTFS Realtime *VehiclePosition* pings from the Modesto Area Express (MAX) with the associated spatial paths defined in the GTFS Schedule data for the same period. The distances between the pings and the paths are calculated and examined to determine an accuracy threshold. This threshold forms the basis for a series of accuracy metrics and visualizations.

3.2 Study Area

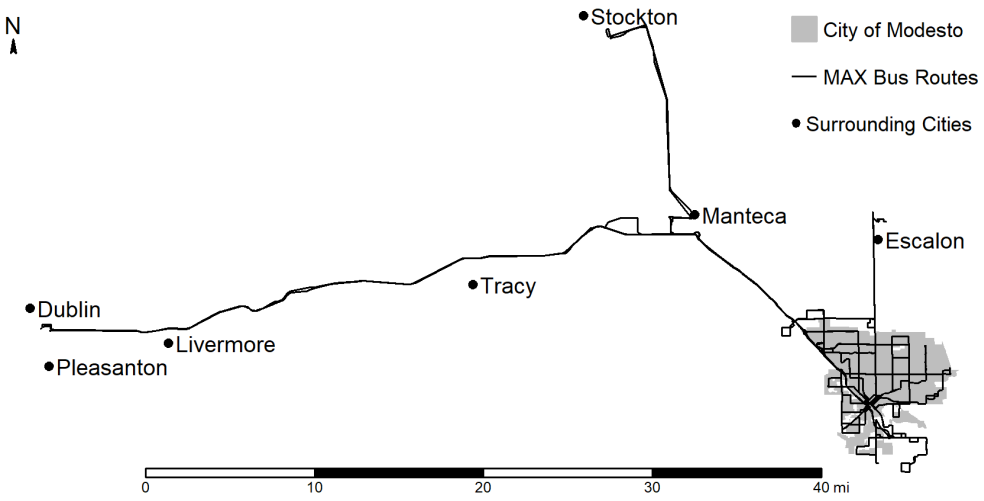
The approach presented in this section is illustrated in reference to the MAX transit system based in Modesto, California—a city of 218,464 residents (U.S. Census Bureau, 2021) in the heart of the state's Central Valley. While relevant data were collected for multiple transit agencies, MAX was chosen to demonstrate the approach as a midsized system with many different route configurations.

It is worth noting that on July 1, 2021, less than a year prior to the data collection for this project, the city-owned MAX merged with the county-owned Stanislaus Regional Transit (StaRT) to form a new transit agency, the Stanislaus Regional Transit Authority (StanRTA). Despite this legal union, many former MAX assets, including the GTFS feeds, remained distinct at the time of data collection in Spring 2022. For this reason, this research refers to MAX even if that agency no longer exists as a separate legal entity.

MAX Transit Service

In March 2022, MAX operated 61 buses across 22 fixed routes within Modesto and neighboring communities, as shown in Figure 20.

Figure 20. Map of MAX Transit Routes



MAX routes included 19 numbered local lines, primarily on half-hour (74%) or hourly (16%) headways during weekdays. One local route (Route 21) offered fifteen-minute headways while the last numbered route (Route 35 or eTrans) made three round trips per day to the nearby community of Escalon. (eTrans service has been subsequently discontinued.) On weekends, only the local routes with half-hour (or shorter) headways continued to operate, often at reduced frequency.

MAX routes also included three named commuter express lines. These routes, respectively, connect Modesto to the Altamont Corridor Express (ACE) commuter rail station just outside Manteca, the Bay Area Rapid Transit (BART) commuter rail station between Dublin and Pleasanton, and to the city of Stockton. The ACE Commuter Express line offers seven round trips per day, while the other two commute express lines offer less than half that level of service.

3.3 Data Collection

This research collected *VehiclePosition* feeds from MAX's public posting of GTFS Realtime data beginning on February 28, 2022, and ending on March 6, 2022. Included in these data were five full days (March 1-5) and two partial days for a total of 251,631 records. This research also collected the contemporaneous static GTFS Schedule data for the same period.

3.4 Ping-Path Distance Thresholds

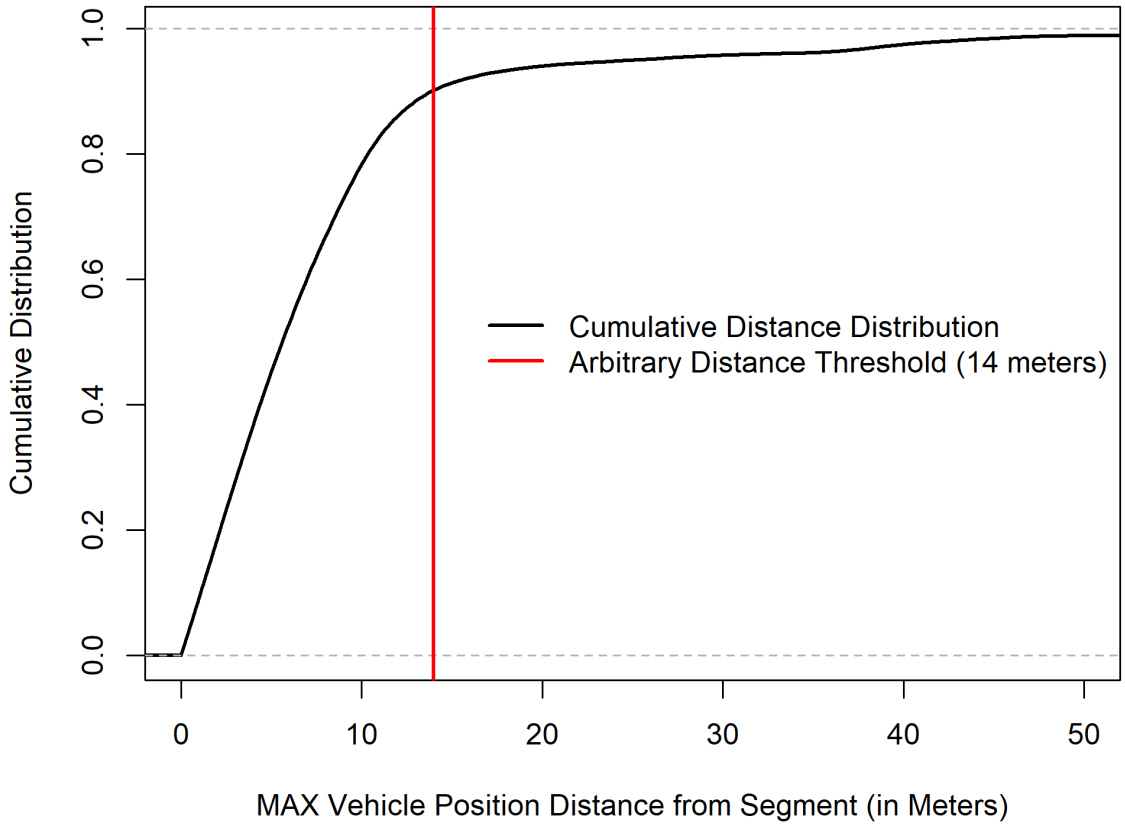
This research applied an approach to parse the shape files within the GTFS Schedule feeds into variants and segments and then to append the pings to these paths. In this nomenclature, a variant refers to those trips on a given route that have the same set of stops in the same order and follow an identical path as defined by the associated vertices. A segment refers to the portion of that path

that connects two consecutive transit stops. Pings refer to the location of the bus provided by the timestamped coordinates within the *VehiclePosition* message sets.

Once the pings are associated with the appropriate path, the shortest straight line distance between the points and the lines are calculated in meters. Figure 21 presents a cumulative distribution function for these distances. The chart is truncated for readability at 50 meters, which accounts for 98.9% of the data, but the full range extends to 149 meters.

The first task was to establish a threshold for determining ping-path distances that signify an accuracy concern. Ping-path pairs beyond this threshold are flagged for further analysis. A visual inspection of these distances showed an inflection point in the cumulative distribution function around the 90% level. While recognizing that these data reflect a single transit system, establishing a threshold for problematic ping-path pairs at the ninth decile is not an unreasonable statistical practice. In this case that decile is 13.8 meters, which is rounded to 14 meters (45.9 feet) for ease of application. The red vertical line in Figure 21 denotes this threshold.

Figure 21. Cumulative Distribution Function of Ping-Path Distances

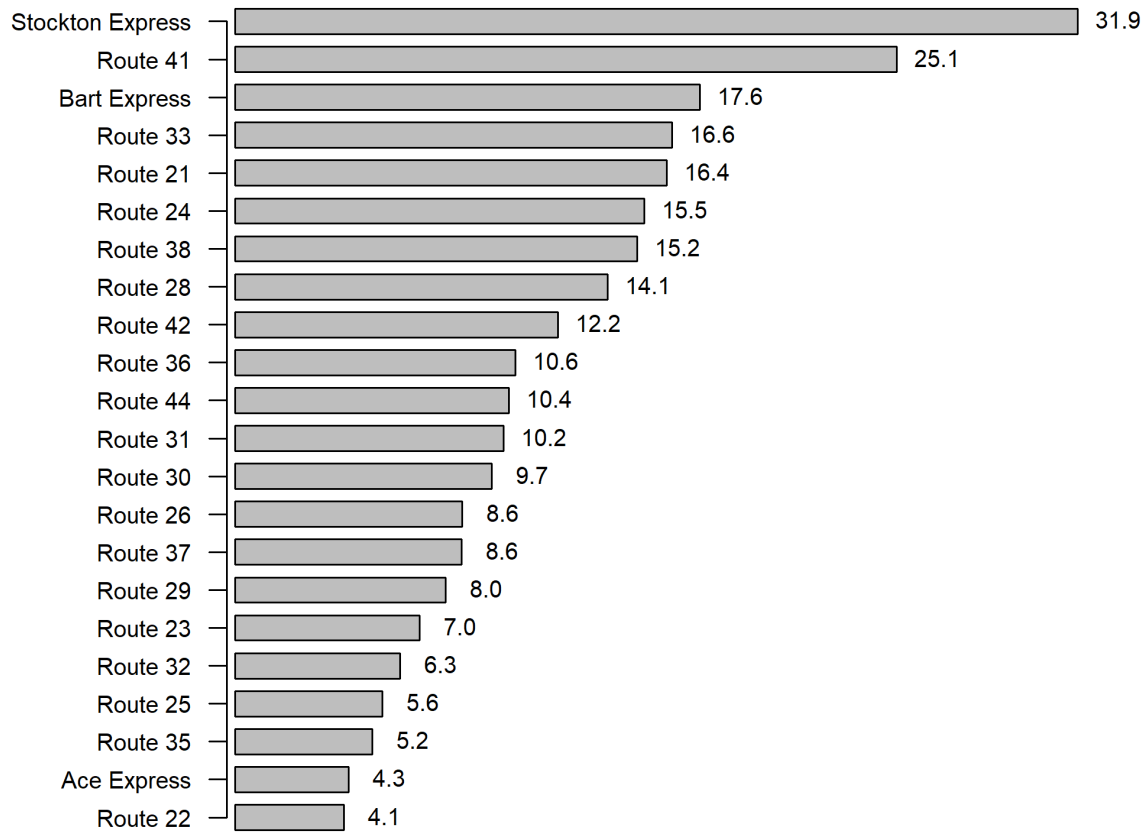


The decision to select this threshold was not simple. Roadway lane widths in the United States vary between ten and 12 feet, with buses typically (but not always) running on roads with at least four lanes. Some geographic information system (GIS) files of roadways code road centerlines rather than actual lanes, and GTFS shape coding practices often follow this simplified convention. Therefore, in a major arterial corridor one might reasonably expect a transit vehicle to have up to a three-lane distance from the centerline, which—depending on the location of the GPS antenna on the bus, the width of the lanes, and the presence of a center median—could easily be 40 feet (and even more on a highway). Alternatively, GTFS guidance calls for a vertex in the shape description that is within 30 feet from the coded location of the transit stop, and those stops should not be placed within roadways. Most buses operate on streets with narrower profiles. Given these concerns, a more rigorous threshold was also considered (and tested) at ten meters (32.8 feet), which was at the 79th percentile of ping-path distances. While this research selected the more permissive threshold of fourteen meters to highlight the most egregious disparities, agencies and analysts should select the threshold that meets their specific needs. In many contexts, particularly for entirely local services, the ten-meter threshold is appropriate.

3.5 Route-Level Metrics

Flagging problematic ping-path pairs makes it possible to readily calculate route-level accuracy metrics. A straightforward metric is the share of total records for a given route that exceed the distance threshold. Figure 22 presents this metric for each MAX route sorted from highest to lowest shares of flagged ping-path distance values.

Figure 22. Percent of MAX Route Ping-Path Distances above 14-Meter Threshold



These route-level metrics offer one avenue for agencies and analysts to prioritize the exploration of GTFS spatial accuracy. For example, the data presented in Figure 22 might prioritize consideration of the Stockton Express and Route 41, both of which report more than a quarter of their GTFS Realtime pings beyond the 14-meter threshold from the GTFS Schedule path.

These route-level metrics might also reveal potential patterns of ping-path disparities for analysts familiar with the transit system. For example, the three routes with the highest shares of flagged records (Stockton Express, Route 41, BART Express) are all routes that have long highway sections. Perhaps GTFS Schedule coding becomes laxer when there is a long distance between stops or the multiple, wide highway lanes are more likely to place pings beyond the threshold distance from paths. (Incidentally, these data also challenge this position as the route with the second smallest share of flagged records, the ACE Express, runs almost entirely on highways.)

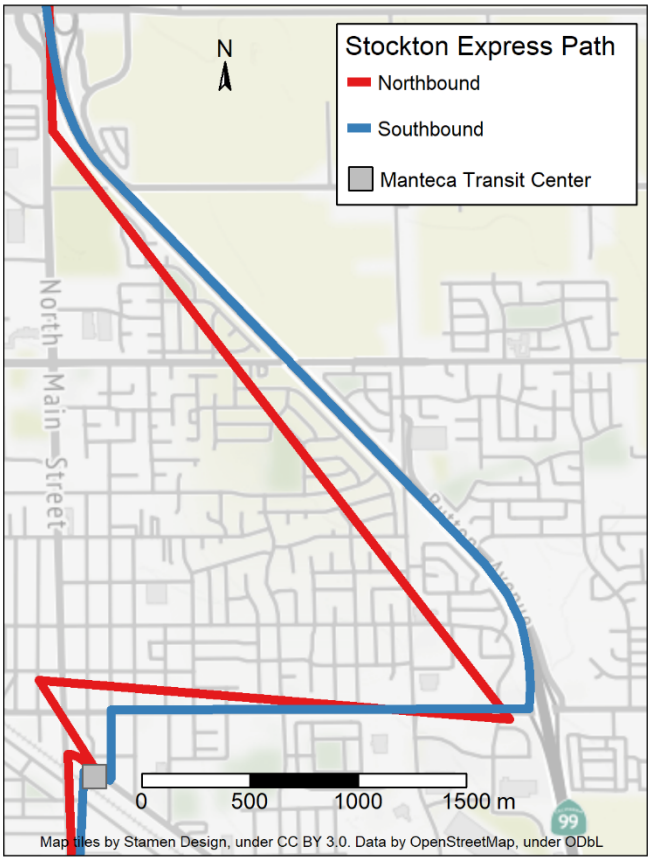
To demonstrate how route-level metrics can foster GTFS spatial analysis, this research focuses on the two most problematic routes: The Stockton Express and Route 41.

Stockton Express

The Stockton Express originates at the Modesto Transit Center, travels north along California State Route 99 (SR 99)—exiting only twice (at the Vintage Faire Mall and the Manteca Transit Center)—before arriving an hour later at the Stockton Transit Center. The route then makes the same trip in reverse. A review of this route reveals two common contributors to GTFS spatial inaccuracy.

First, GTFS schedule data is often coded coarsely along long segments between stops. For example, Figure 23 shows a section of the Stockton Express path coding just north of the Manteca Transit Center. The southbound variant, coded in blue, exemplifies careful coding of a transit path in GTFS schedule data. The path runs south along SR 99 southbound, takes the appropriate exit lane to depart the limited access highway, and then follows surface streets to the Manteca Transit Center, shown in grey. By contrast, the northbound variant, coded in red, exemplifies coarse coding of a transit path in GTFS schedule data. From the moment it departs the transit center, the path does not follow any roadway until it joins North Main Street near the top of the map. Unfortunately, the express route does not in practice traverse that portion of North Main Street. Such imprecise coding leads to a high rate of flagged pings when the coded path is farther than the threshold distance from the actual route on which the buses are traveling. (It is worth noting that in the small section of Figure 23 south of the Manteca Transit Center, the southbound path diverges from any roadway while the northbound variant correctly follows North Main Street before backtracking around a rail right-of-way to reach the station.) Route-level analysis helps diagnose this coarse coding of the path in the GTFS schedule data. Fortunately, the vertices that define the path within the GTFS standard can be easily adjusted to align with the actual roadway. Such a revision to the static data is recommended to remove the Condition #1 problem of inaccurately coded paths.

Figure 23. Inaccurate GTFS Shape Coding for Stockton Express Route



Second, bus drivers do not always follow the scheduled route. While such deviations are not contrary to MAX policy (which empowers drivers to take alternative routes if they will better maintain the time schedule), the frequency with which such deviations occur might be a reason to revisit the official route design.

Driver deviations can be seen in several locations along the Stockton Express. One of the more interesting deviations is on the southbound approach to the Vintage Faire Mall. Instead of following the route shown in Figure 24, in which the express bus continues south on SR 99 before exiting at Beckwith Road just south of the mall and then backtracking to the bus stop, drivers leave the highway at Pelandale Avenue, one exit upstream from the mall, cross over the highway and then travel down Sisk Road, a parallel frontage roadway just east of SR 99, to arrive at the mall bus stop without backtracking.

Figure 24. Driver Deviations along the Stockton Express Route

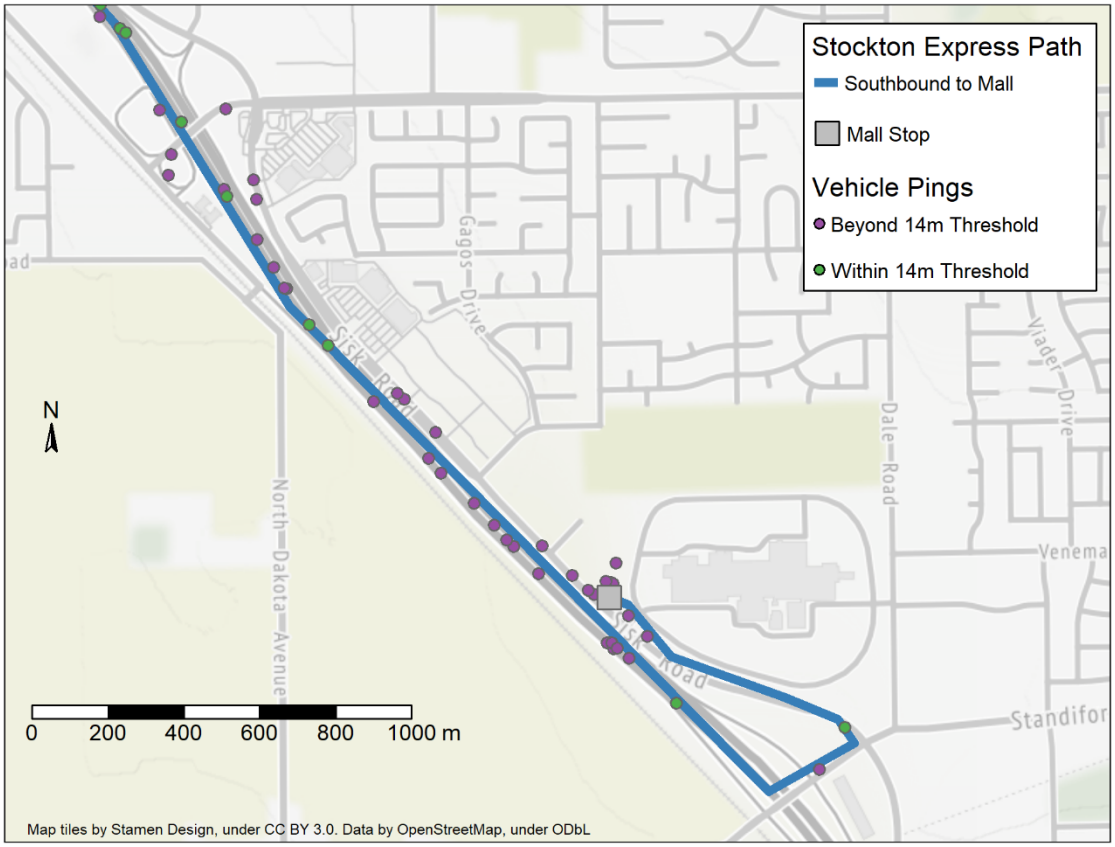


Figure 24 includes the pings of buses that leave the prescribed path (which is itself so coarsely coded that many buses on the appropriate alignment are flagged as problematic). These driver deviation pings are found on the upstream exit ramp, on Pelandale Avenue (unlabeled), and then traveling down Sisk Road parallel to the actual scheduled route on SR 99.

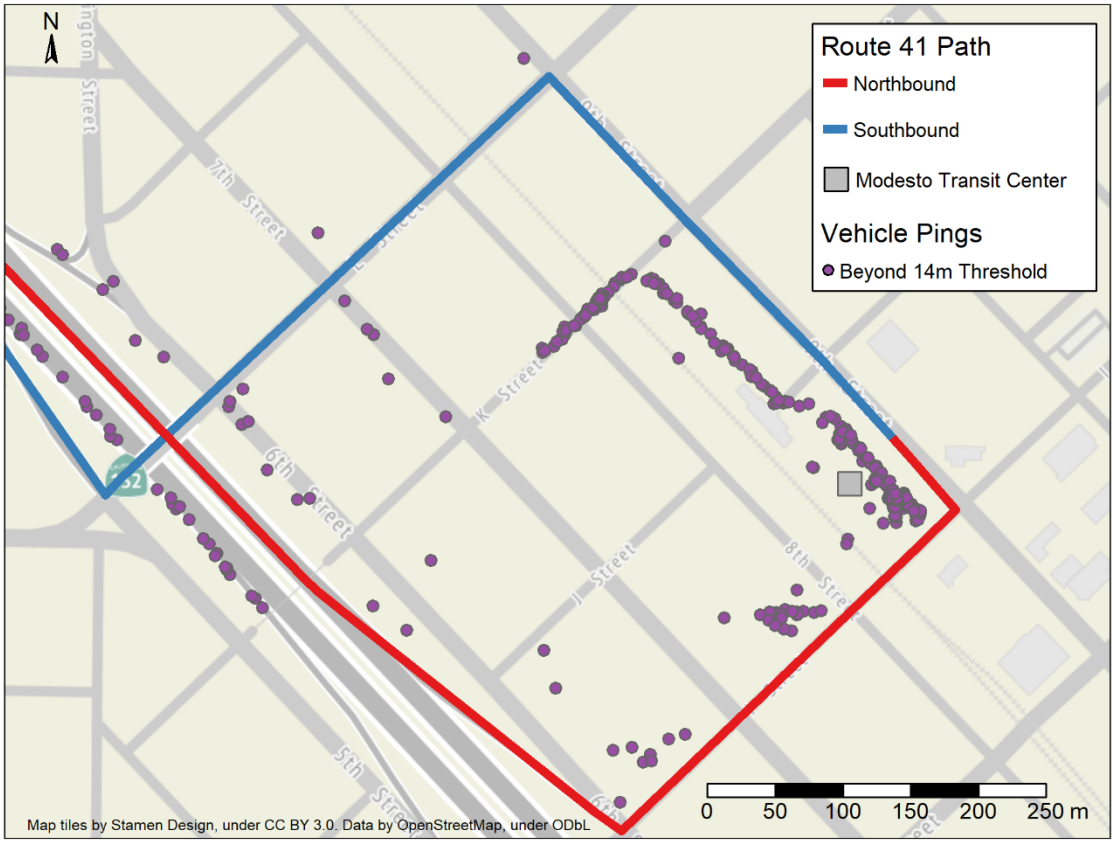
Similarly, most southbound trips departing the Manteca Transit Center (the previous stop) did not follow the scheduled path to take California State Route 120 (SR 120) eastbound to SR 99 southbound; instead, the majority of drivers chose to bypass this highway interchange entirely and connect from SR 120 to SR 99 via Moffat Boulevard, a frontage road that serves as the hypotenuse between the two highways. The prevalence of these Condition #3 driver deviations from the scheduled path on this one segment might trigger a revision of that path to reflect driving practices.

Route 41

The second most problematic line serves as a shuttle between the Modesto Transit Center and the Vintage Faire Mall along SR 99. It overlaps the southernmost segments of the Stockton Express but makes twenty more round trip runs on weekdays and offers service on Saturdays and Sundays. These sources of spatial inaccuracy are particularly visible near the downtown terminus, as shown

in Figure 25 (which only shows flagged pings for simplicity). These inaccuracies illustrate three common concerns.

Figure 25. Route 41 Pings Beyond the Threshold near the Modesto Transit Center



First, exit ramp construction noted in driver dispatch reports is temporarily causing deviations. These deviations are evidenced by the string of purple dots along the highway (rather than along the exit ramp) and along 6th Street (from earlier exiting). Despite the extended nature of this highway construction, MAX did not update the GTFS shapes to reflect the detours. Consequently, many pings were flagged during the period of construction.

Second, while the path defined by the GTFS Schedule data within the downtown is carefully coded to actual roadways, that coding does not reflect the actual bus route. The southbound (blue) path is currently coded to continue along L Street before turning right onto 9th Street, where it terminates some physical distance from the actual bus stop. The actual southbound route, however, turns right earlier onto 7th Street then makes a left onto K Street before making another right into the Modesto Transit Center, which is not technically on any street. The path as currently coded does not follow that course in the downtown, with the result that all the pings that occur as the bus is entering the transit station are flagged as suspect. Compounding this problem, the Modesto

Transit Center is not itself coded on the path defined by the GTFS Schedule data. This Condition #1 path alignment issue could be easily resolved by recoding the shape to the actual roadways used by the route—even if these are not public streets. A comparison of the map shown in Figure 25 with the aerial view shown in Figure 26 illustrates that the Modesto Transit Center stop is coded properly, however. the path to that stop is not.

Figure 26. Aerial View of the Modesto Transit Center in Relation to 9th Street

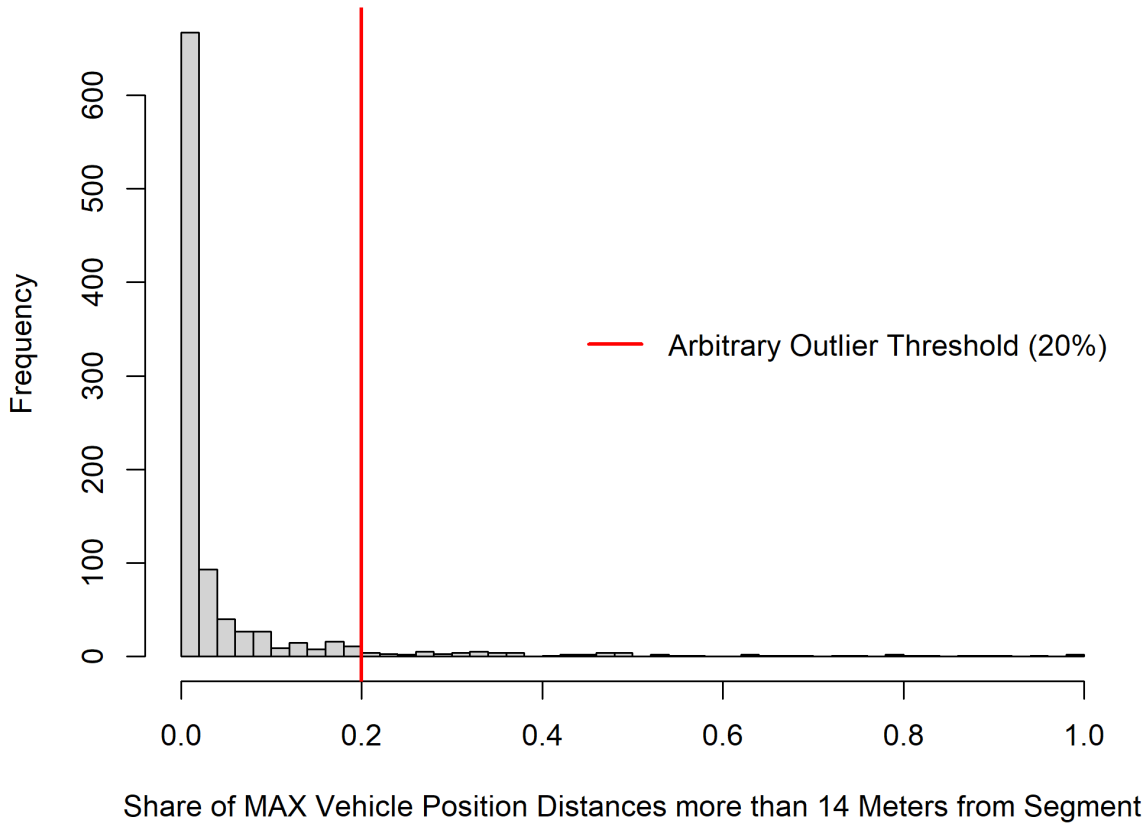
Finally, in certain places GPS pings seem to lose their accuracy. This feature is surprisingly consistent in certain locations. For example, Figure 25 demonstrates that as Route 41 departs the Modesto Transit Center and travels between 8th and 6th Streets along I Street, the associated pings show up not on the roadway, but in the middle of the block. Figure 26 shows those locations are either not suitable or not accessible for buses. These Condition #2 erroneous ping locations can be identified through mapping but are difficult to resolve. In this case, there is not a known street canyon effect that might be blocking the GPS signal, as the surrounding buildings are single-story commercial structures. Despite the difficulty in addressing Condition #2 inaccuracies, it is worth identifying them across a system. Such recognition might lead to some trial and error with equipment to reduce these inaccuracies.

3.6 Segment-Level Metrics

Route-level metrics are important for comparing route-level accuracy, but, in most cases, areas of ping-path discrepancies are clustered only on small portions of routes. A more targeted way to assess (and address) these problematic areas is to calculate segment-level metrics. Since segments, the stop-to-stop portions of route variants, do not have readily intuited identities, it is more effective to map them than to graph them. While it is possible to simply make a choropleth map of all segments that is color coded by the share of flagged ping-path distances, in practice that approach is visually overwhelming and difficult to parse. A simpler mapping flags problematic segments for consideration by introducing a second threshold.

Figure 27 presents a histogram of the distribution of segments based on their share of flagged pings. Once again, there is a need to arbitrarily select a level at which these spatial units are seen as potentially problematic. A visual examination suggests that a useful breakpoint in the distribution might be 20% as demarcated by the vertical red line on the graph. This cutoff identifies the most problematic segments, however, just as analysts may choose a lower threshold for ping-path distances, one might choose a lower threshold for the segment share of flagged distances to identify more areas of concern. In this case, there also appears to be a reasonable breakpoint at the 10% level. Once again, setting the threshold should reflect the local concern for GTFS spatial accuracy.

Figure 27. Histogram of Shares of MAX Segments Beyond the 14-Meter Threshold



Once a threshold is set, it is possible to simply map flagged segments, as shown in Figure 28, for the city of Modesto and its close environs. Such a map offers an easy identification of areas of GTFS accuracy concern and, when incorporated into a GIS that includes a basemap, an easy means to zoom in to visually diagnose their etiologies. For example, in addition to denoting the issues on the Stockton Express and Route 41 discussed earlier, Figure 28 also demonstrates additional issues around the Modesto Transit Center in the downtown (partly obscured by the station symbology), which a more focused analysis would show reflect the same path coding and unexplained GPS drift issues discussed for Route 41.

Figure 28. MAX Segments with 20% of Pings Beyond the 14-Meter Threshold

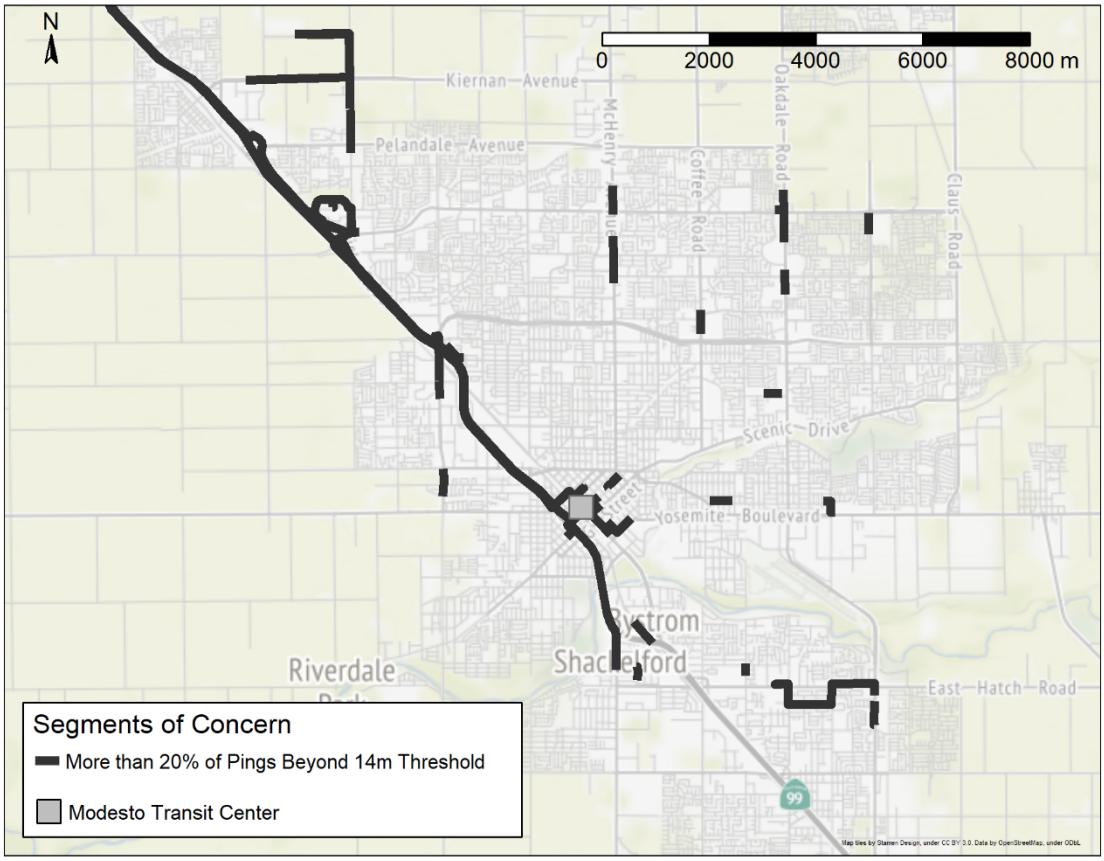


Figure 28 also demonstrates new segments of concern scattered throughout the service area. Since the selected threshold was relatively high, the number of flagged segments is relatively small and checking them all is a manageable task. Many of these segments reflect the common issues noted earlier in the discussion of the Stockton Express and Route 41. For example, the vertical paths along McHenry Avenue in the upper middle section of the map are perfectly coded to the roadway, but demonstrate consistent GPS drift with the pings, suggesting the buses are magically traveling over buildings west of the roadway. Similarly, the vertical paths along Oakdale Road, two miles east of McHenry Avenue, are centerline coded. When the southbound lane shifts westward with the addition of a median during these segments, the pings along the roadway are all flagged as too distant from the unwavering path line. Figure 28 also flags segments of concern that reflect accuracy issues not discussed earlier. Two of these issues warrant additional discussion, as they add insight into the sources of discrepancy between schedule and realtime spatial data.

Unreported Driver Deviations

Exploring these segments of concern led to several cases where the vehicle pings substantially diverged from the route design path. It appears that drivers are consistently deviating from the scheduled route, however none of these deviations were recorded within MAX's dispatch notes—meaning drivers did not report these detours. While in theory it is possible that drivers are avoiding areas of street construction, an open records request from the City of Modesto only identified one project that aligned with the deviations shown in Figure 28. A sewer replacement along Covena Avenue north of Yosemite Boulevard closed the intersection at Miller Avenue. This closure likely resulted in the short flagged segment along Miller Avenue parallel to Yosemite Boulevard in Figure 28. A more probable cause is the decision by drivers to permanently detour to improve on-time performance, which might spark a review of the scheduled route path.

Figure 29. Driver Deviations Along Route 37

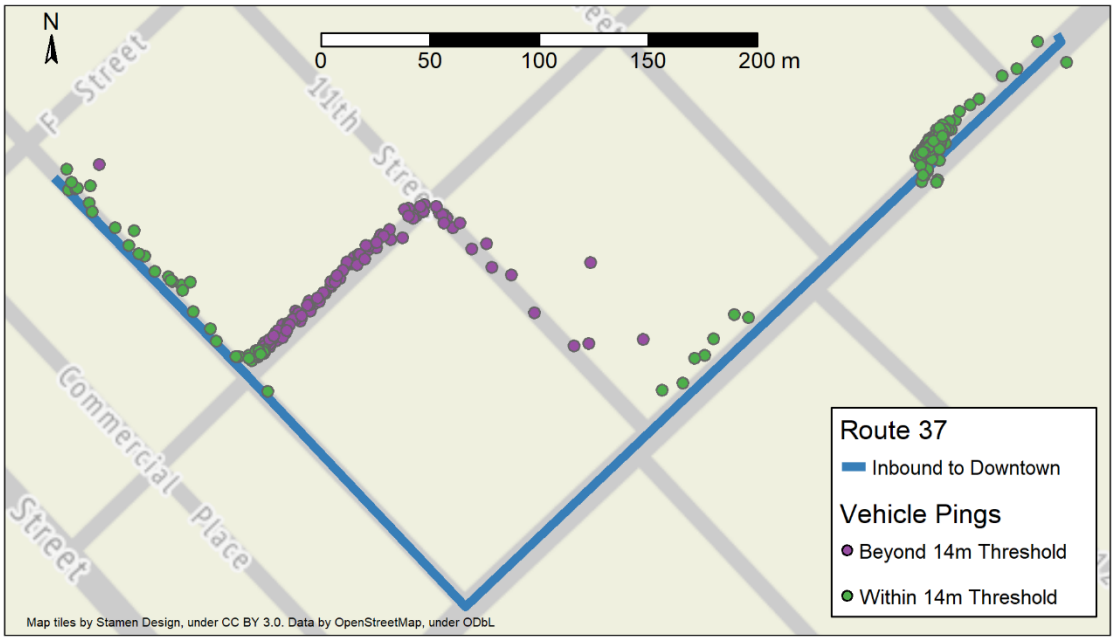


Figure 29 presents one such example where it appears that buses on Route 37 that are approaching Modesto's downtown are all detouring to avoid the corner of 10th and D Streets. (There is one ping that might remotely suggest the use of the scheduled path.) This deviation avoids a traffic light and might offer operational advantages. This segment-level analysis reveals this consistent deviation. MAX might treat this inaccuracy as a Condition #1 coding issue and recode the path, or it might treat this as a Condition #3 driver deviation issue and require adherence to the scheduled route.

Incomplete Path Coding

Exploring the segments of concern revealed another possible source of inaccuracy, namely when the GTFS Schedule data is itself incomplete. This feature can be seen in the outbound Route 44 shown in Figure 30. This route travels eastward along East Hatch Road but is forced to detour through a residential neighborhood between Central Avenue and Moffett Road because there is no room to make an eastbound stop on that section of East Hatch Road. The detour is necessary to place an eastbound stop on Moffett Road to maintain appropriate stop distances along the route. MAX seeks to minimize the nuisance to residents of this neighborhood, so it routes alternating trips along two different parallel roads—but did not include this path difference within its GTFS Schedule feed. The pings that are flagged in Figure 30 are following their intended route—the inaccuracy is due to a Condition #1 error of a missing path in the schedule data.

Figure 30. Incomplete Path Coding Along Route 44

Figure 30 illustrates a second important insight regarding GTFS spatial accuracy assessment, namely pings reported in some GTFS Realtime feeds are not evenly distributed but instead spatially clustered. One would expect the purple pings in Figure 30 to stretch across the entire missing link in the GTFS Schedule path—but these pings only appear at the entrance and exit of the link as if the bus cinematically passed through a tesseract. Unlike most GTFS Realtime feeds which offer temporal snapshots of where the vehicle is at that moment, the MAX pings appear to show when vehicles are in specific places, as if triggered by geocoded gates. This approach leaves

holes in the record of vehicle locations, particularly evident in the example shown in Figure 30 (but also visible in Figure 29). In practice, these clustered pings reduce the ability of the method proposed in this paper to assess spatial accuracy across the entire service area.

3.7 Discussion

The primary goal of this research is to present a method for assessing the spatial accuracy of GTFS Schedule and Realtime products that is easy to implement and scalable across transit agencies of varied sizes. The secondary goal is to identify the types of inaccuracies that are common and consider their implications for downstream users.

Accuracy Assessment Method

The proposed method relies on only two inputs, both of which are publicly available and structured in a standardized format—the GTFS Schedule and associated Realtime feeds. While the former data are altered infrequently (depending on the system) and might only need to be downloaded once, the latter are dynamic and will need to be downloaded at regular intervals ideally, but not necessarily, matching the rate at which the transit agency updates them. Users will also need GIS-enabled software to store, relate, query, and display the data. This research presents a particularly careful procedure to pair pings and paths for each segment of each variant. While this precise approach enables detailed segment-level analysis and increases the likelihood that each ping is matched to the correct path, for many users it will be sufficient to simply join the pings to the associated shape defined by the GTFS Schedule data. This simplified approach reduces the data handling burden and still enables all the route-level analysis. (It is possible but not recommended to simplify data management further by first creating a conjoined path from all the shapes associated with a route and then calculating the distances from the vehicle pings to that conjoined path. This very simplified approach will underestimate actual ping-path distances when pings from one variant match with closer paths inherited from another variant. For example, the poor coding of the northbound path of the Stockton Express shown in Figure 23 may never be flagged if the northbound pings match to the well-coded southbound path. Similarly, the southbound ping deviations shown in Figure 24 when Stockton Express buses leave SR 99 early to take a parallel frontage road east of the highway may not pass the distance threshold if they pair with the closer path of the northbound variant. Analysts who take this highly simplified approach need to recognize that only very gross inaccuracies will be flagged—which may still be sufficient for some purposes.)

The approach as presented here relies on setting thresholds for determining problematic ping-path distances. Establishing thresholds is useful as it allows for a clear and binary identification of whether a given reading is good or bad. Furthermore, the use of thresholds is commonplace within the transit industry. On-time performance, for example, is similarly dummy coded at the disaggregate level and similarly averaged to serve as a route-level measure. The problem with

thresholds is that they are arbitrary. It can be difficult to justify a given value and why values just below and just above that threshold are materially different.

This research explores both a distribution of ping-path distances and a theoretical justification for selecting the appropriate threshold for flagging values of concern. As a practical matter moving forward, it makes sense to select the threshold value a priori rather than attempt to derive it. Such a selection would offer the percentile value of that threshold as a new performance metric. Agencies might stick with a given threshold and see how its associated percentile among the ping-path distances changes over time. Ideally, that percentile would get higher as the ping-path distances became more accurate.

This research advocated for a 14-meter threshold to flag the most problematic readings. However, as more agencies apply this approach, there might be a consensus for a different value, such as ten meters. Similarly, transit agencies might apply different expectations of spatial accuracy, and therefore different thresholds to different portions of their system. For example, express routes might use a larger threshold while local routes on surface roads would use a smaller threshold. While an industry consensus approach would be good for generating (and reporting) comparative metrics, it would still be advantageous for individual agencies to try different (and progressively tighter) thresholds as tools to identify different issues that might be affecting the spatial accuracy of the GTFS data.

The challenge of setting such thresholds extends from flagging problematic ping-path pairs to identifying the share of such flagged values in a segment. This research used a 20% share to identify segments of concern, but that value could also be adjusted downward to catch more nuanced issues of spatial accuracy. While not presented here, it might also make sense to aggregate segment-level data to create new performance measures. One measure could be the share of total weekly service mileage for which a fifth of its ping-path distances are flagged as above the threshold. The value of such metrics is to track improvements over time (and to compare accuracy levels between properties and GTFS vendors).

Impact of GTFS Spatial Accuracy

This research identifies distinct types of common situations that result in ping-path discrepancies above a specified threshold. The value of this identification is in gauging their impacts on different downstream users.

Fortunately, most of the problems identified in this research will have negligible effect on the traditional use of GTFS data for trip planning purposes if buses still stop at every stop according to the stop sequence. That said, the poor coding of paths in the GTFS Schedule data might affect algorithms for predicting stop arrival times to the extent that they rely on the actual GTFS shapes rather than historical data. Similarly, driver deviations from scheduled routes might confuse

patrons who use online applications that show vehicle locations based on the realtime pings. It is common to see a vehicle traveling off the expected path on a trip tracking application.

The spatial accuracy issues presented here are more likely to affect off-label uses of GTFS products. For example, using GTFS Schedule data to identify the paths and service levels of transit to assess traffic or noise or air quality impacts for equity analysis might be confounded by both Condition #1 and Condition #3 errors. Incorrectly coded paths that place buses where they do not actually go (such as the Stockton Express in Manteca in Figure 23 and Route 41 in downtown Modesto in Figure 25), correctly coded paths from which drivers frequently (or always) deviate (such as the Stockton Express southbound to the Vintage Faire Mall in Figure 24 and Route 37 inbound at the intersection of 10th and D Streets in Figure 29), and incomplete paths that do not place buses where they actually go (such as Route 44 outbound between Central Avenue and Moffett Road in Figure 30) all reflect inaccuracies in the GTFS Schedule data that would yield deceptive representations of facts on the ground. This scenario could create a professional nightmare for a city planner who uses the GTFS Schedule information to publicly insist that a given residential neighborhood has no buses driving through it (say if they looked only at the paths but not the pings in Figure 30), while those residents reply that the planner is entirely mistaken and show video evidence to prove it.

Retroactive analysis of GTFS Realtime pings to impute stop arrival and departure times could also be confused if the associated GTFS shapes do not reflect the actual transit paths, such as for the Modesto Transit Center (Figures 6 and 7), which is the hub of the entire MAX system. Similarly, potential future use of GTFS Schedule shapes to direct driverless vehicles could theoretically lead to substantial confusion when the scheduled paths do not physically align with the actual stop locations.

The value of this research is that it can easily screen for discrepancies between what the GTFS Schedule data says is supposed to happen and what the GTFS Realtime data says is actually happening. The sorting algorithms presented here provide an avenue to quickly flag and map such discrepancies, ideally so that they can be resolved in some way that improves future spatial accuracy. A lingering question is the impact of GPS readings that seem to drift from actual roadways and the surprising phenomena of clustered rather than distributed pings found on MAX. This proposed method for identifying spatial inaccuracies is predicated on the availability of a vast number of pings that reflect a random spatial sampling. If those pings are only collected near specific points, they may not identify all the areas of discrepancy in a transit network.

3.9 Conclusion

This research presents an elegant approach to marry GTFS Schedule and Realtime data to identify spatial inaccuracies. A method for flagging problematic ping-path pairs is proposed along with techniques for aggregating that information to the route or segment levels to quantify and explore

areas of concern. Common sources of inaccuracy are discussed, along with their potential implications for traditional and off-label users of GTFS data products. This approach provides a very straightforward method for transit agencies to review and troubleshoot their operations. The goal of this work is to make it easier for transit agencies to improve the accuracy of the data they provide the public and to increase customer satisfaction and confidence in the accuracy of GTFS feeds.

4. Stop Accuracy

The accuracy of the General Transit Feed Specification (GTFS) Schedule and Realtime data is particularly important when it comes to bus stop locations, as these are the locations where users enter and exit the system. Incorrect stop information can increase the physical and cognitive burden for travelers thereby undoing the promised benefits of online trip planning. Furthermore, stop inaccuracies can affect all the downstream uses of GTFS data—from system planning to equity analysis to a future of driverless provision—since stops define transit access points.

Fortunately, the complementary nature of GTFS Schedule data, which represent where the bus is supposed to stop, and GTFS Realtime data, which represent where the bus actually stops, enables an inherent cross-check. To date there have been limited efforts to exploit this self-corrective aspect of GTFS; however, since more accurate GTFS data is inherently beneficial to the myriad transit riders, planners, researchers, and advocates that rely on GTFS products, this research seeks to reverse this situation. Specifically, this work explores strategies to engage GTFS products in the assessment of their own stop accuracy. The goal of this work is to showcase easily implementable techniques to diagnose and resolve problematic stop information.

These techniques are all based on the core identification of divergent location data between designated and actual bus stop locations. Flagging these divergences allows for the generation of different spatial and aspatial visualizations to characterize and address the underlying accuracy problems—primarily undesignated stopping locations, but also incorrectly coded GTFS Schedule data and, depending on the system, errant GTFS Realtime pings—which might warrant adjustment of the transit design, provision, or hardware.

4.1 Methodology

This quantitative study compares several days of GTFS Realtime *VehiclePosition* pings from the Alameda Contra Costa Transit District more commonly known as AC Transit (ACT) with associated stop locations defined in the GTFS Schedule data for the same period. The distances between the pings and the stops are calculated and examined to determine an accuracy threshold. This threshold forms the basis for a series of accuracy metrics and visualizations.

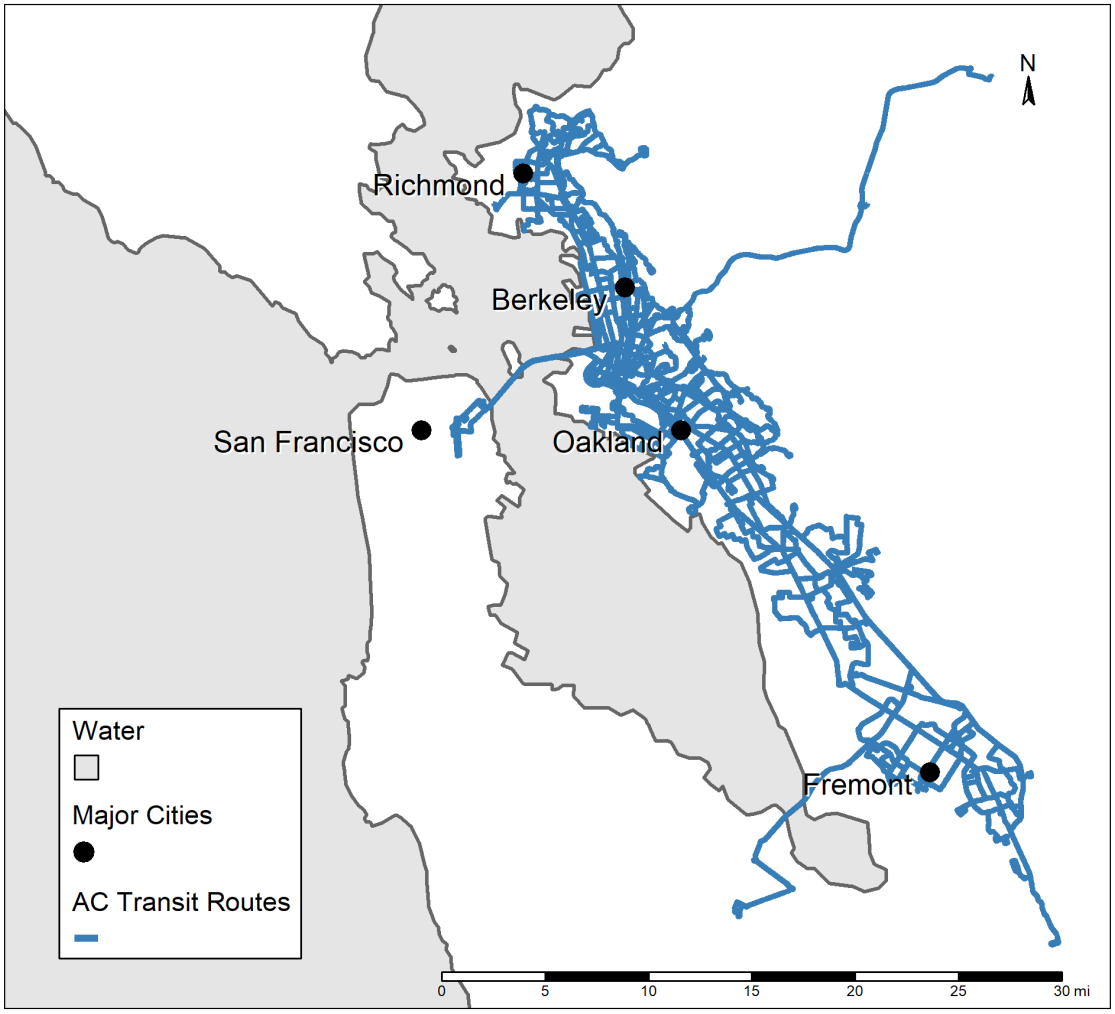
4.2 Study Area

The approach presented in this section was explored on the AC Transit system based in the East Bay of the San Francisco Bay Area. AC Transit was chosen as the only transit provider among the five sampled whose GTFS Realtime feed codes for whether the vehicle is at a stop at the time of the ping. This coding is an optional aspect of the GTFS Realtime standard and therefore not present in feeds from all transit providers.

AC Transit Service

In 2021, the latest federal data available, AC Transit maintained 466 buses, 69 commuter buses, and 27 bus rapid transit buses to serve their fixed-route operations (Alameda-Contra Costa Transit District, 2021). In March of 2022, when these data were collected, this network included 129 uniquely named routes, of which 60 are the main East Bay services (#1-399), 44 are school services that serve arrival and dismissal times (#600-699), three are early bird express services that serve selected Bay Area Rapid Transit District (BART) links before those rail services open (#700-799), six are late night services that serve key corridors after midnight (#800-899), and 16 provide service across the San Francisco Bay. These transbay routes are designated by letters rather than numbers. The full network, many of whose routes overlap, is shown in Figure 31.

Figure 31. Map of AC Transit Routes



4.3 Data Collection and Cleaning

A key aspect of this project is the care in collecting and cleaning the initial data.

Data Collection

This research collected *VehiclePosition* feeds from AC Transit's public posting of GTFS Realtime data beginning on February 28, 2022, and ending on March 6, 2022. Included in these data were five full days (March 1-5) and two partial days. This realtime information was collected by downloading the available data every ten seconds during the study period. The contemporaneous GTFS Schedule data were collected for the same period.

For the purposes of this research, pings refer to the location of the bus provided by the timestamped coordinates within the *VehiclePosition* message sets. AC Transit vehicles have GPS antennas in the front of the vehicle so ping locations should closely represent bus boarding locations. All pings are also coded within the GTFS Realtime messages with the *stop_id* of the approaching stop location from the GTFS Schedule data. That designated location is referred to here as the stop.

Of the data collected, only those pings with the optional *current_status* field with a status of *stopped_at* (which signifies that the ping was sent while the bus was engaged in a stop event) were selected for analysis. A small number of these selected records lacked trip identifiers and were therefore dropped for further consideration.

Data Cleaning

The data cleaning of the selected stop-event pings included multiple steps. First, duplicate records were removed to leave only unique GTFS Realtime data. This ensures that each transit event is recorded only once. Downloading all available data every ten seconds leads to duplicate records whenever the download rate exceeds the GTFS Realtime polling rate as the same information becomes archived more than once. If the number of duplicates were exactly the same for each record, this duplication would increase the data processing demands but not affect aggregate statistics. However, varying rates of duplication would result in incorrectly weighted data for aggregate measures. Removing duplicates ensures the appropriate consideration of the data (and reduces computational demands).

Second, all GTFS Realtime records more than a minute prior to the official start time of the route in the GTFS Schedule data were excluded from consideration. This exclusion is meant to insulate the data analysis from the common practice of bus drivers stopping and opening their doors in the few minutes preceding the trip's designated departure time from the first stop, among other reasons, to accommodate boarding passengers (and transit staff)—often before bringing the vehicle to the scheduled start location. This excluded information can be useful for understanding pre-embarking behaviors, but affects the accuracy analysis by misrepresenting transit activities, particularly for

trips with few stops. The minute buffer was selected to capture stop accuracy at the initial stop—even though it may introduce some pre-trip start activity.

Excluding pings more than a minute prior to the trip start time requires joining the GTFS Realtime records to the GTFS Schedule data. Since the AC Transit GTFS Realtime feeds label many trips with identifiers that are not within the GTFS Schedule data (typically with large negative numbers) and these records cannot therefore be joined to the GTFS Schedule data, removing pre-trip pings also removes trip records with non-standard identifiers. This reduction was seen as a reasonable trade off to exclude pre-trip stopping activities.

Third, pings were consolidated so that each stop event is only represented by a single record in the data. The raw data set reports ping events, which occur based on the polling rate of the GTFS Realtime system. This research, however, focuses on stop events, which inherently include some duration to allow passengers to board and alight. Given that duration, multiple pings are possible during a single stop event. In addition to consistency concerns like those noted above for duplicate records, pings themselves are not the unit of analysis of this research—stops are. Consequently, any ping that shares a date, trip identification, and geolocation (with the latitude and longitude coordinates expressed to fifteen decimal places) to a previous record is removed from the data set. This consolidation results in one ping record representing one stop event, an essential condition for this research.

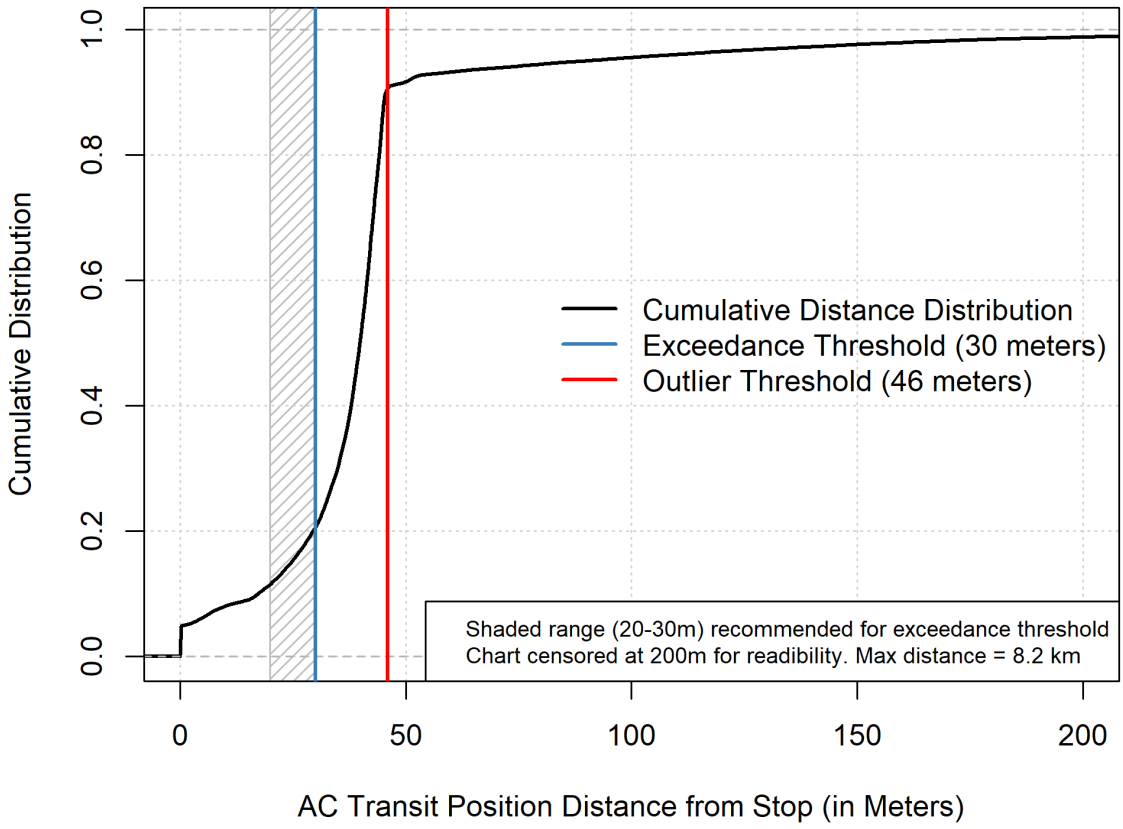
4.4 Ping-Stop Distance Thresholds

The focus of this research is to explore the accuracy of stop data. This analysis requires calculating the distance between the ping location and the scheduled stop location and then determining whether that distance is of concern.

Ping-Stop Distance Calculation

The first step of this analysis is to identify the ping-stop distance between where the bus actually stops and where the bus is supposed to stop. The straight-line distance (in meters) for each ping location to the associated stop location is calculated. Figure 32 presents the cumulative distribution of these distances in black. For readability purposes, the chart is censored at 200 meters, which accounts for 98.8% of the data, but the full range extends to 8.2 kilometers for a ping that was incorrectly coded, as will be discussed later in this report.

Figure 32. Cumulative Distribution Function of Ping-Stop Distances.



This distribution, surprisingly, shows that only about 5% of the stop-event pings occur right at the scheduled stop location (i.e., within one meter). From that inflection point, there is a parabolic curve to a second inflection at 46 meters that accounts for 90% of the stop events. From that second inflection, the curve is more linear and flatter towards the end of the chart.

Some ping-stop distance above zero is expected. Scheduled stop locations are typically geocoded to the sidewalk location of the bus stop. Bus drivers, at best, can pull up only next to that point, not on top of it, which creates some inherent ping-stop distance. Similarly, despite a move towards automation, bus drivers remain human and cannot exactly align the bus with the associated bus stop every single time. Furthermore, since most buses operate in shared rights-of-way, there cannot be certainty that there will not be another vehicle or roadway impediment that forces the bus driver to stop either a little behind or a little in front of the designated stop location. The data in Figure 32 suggest that most stop events occur at a significant distance from the designated stop location. For example, only 11.5% of ping-stop distances are less than 20 meters (denoted by the start of the shaded area in Figure 33). In sports terms, twenty meters (65.6 feet) is the distance between the pitcher's mound and the catcher's box in baseball or from the foul line to the end of the lane in bowling. In transit terms, this distance is just over one and a half standard bus lengths (40 feet)

or one articulated bus length (60 feet). In other words, according to these data, at 88.5% of stop events, AC Transit riders walk (or more likely scurry) more than the distance it takes to throw a strike or to traverse an articulated bus to board their vehicle. (While this distance distribution seems not entirely aligned with the author's multi-year experience as an AC Transit rider, it is directly generated from the information that is presented to the public. Given how unexpected these findings are, the underlying data and associated calculations have been thoroughly reviewed to be used for the remainder of the analysis.)

Exceedance Threshold Determination

The second step of this analysis was to establish a threshold for determining ping-stop distances that signify an accuracy concern. Ping-stop pairs beyond this threshold are considered "exceedances" and are flagged for further analysis.

The visual inspection of these distances, as noted above, reveals two main inflection points, at one and 46 meters, respectively—both of which are imperfect thresholds for determining exceedances. A threshold of one meter is unreasonable given the physical realities of bringing a bus to the curb, and as evidenced by the 5% of ping-stop distances that meet this threshold. A threshold of 46 meters (150.9 feet) offers an evidenced-based threshold for determining the top 10% of outlying values. This distance is, however, half the length of an American football field. Passengers waiting at a bus stop should not be expected to complete a fifty-yard dash to catch a bus. Furthermore, while several buses converging on a single stop is not entirely unusual (given stops shared by multiple routes and the endemic issue of bus bunching where buses on the same route end up at the same stop at the same time), it is rather unusual for a bus in such a queue to be more than three and a half standard (or almost two and half articulated) bus lengths from the scheduled stop. In short, this threshold is too long for determining stop data inaccuracy, although it might be a good measure of outlying distance values.

The lack of a clear break between the two inflection points shifts the threshold determination from an empirical exercise to a professional one. One option, to be as inclusive as possible in identifying ping-stop divergence (while acknowledging the reality of some curb competition at bus stops), is to set the exceedance threshold at 20 meters. This level, discussed earlier, is a restrictive standard that penalizes excessive bus bunching—a reasonable objective for a transit agency. A second option that is more accommodating of operational issues like bus bunching, while still flagging major ping-stop discrepancies, is to set the exceedance threshold at 30 meters (98.4 feet)—slightly longer than the length of a regulation National Basketball Association (NBA) court (94 feet). This level, marked in Figure 32 with a blue line, is a permissive standard that still identifies almost four-fifths of AC Transit stop events as exceedances. While early rounds of research on this project evaluated the tighter 20-meter threshold, the looser 30-meter threshold was ultimately selected to flag exceedances. Nonetheless, transit agencies should feel comfortable selecting any point between these two thresholds, denoted by the shaded area on Figure 32, as a reasonable expectation for

determining GTFS stop accuracy. It is also reasonable (and recommended) to reduce this threshold distance over time as issues are addressed and service improves. Once a given exceedance threshold has been selected, all ping-stop distances beyond that can be flagged. The remainder of this paper demonstrates how these flags enable the creation of metrics to explore stop inaccuracies at the route, trip, and stop levels.

4.5 Route-Level Metrics

Since transit planning and service delivery focuses on routes, route-level metrics represent a key avenue to assess (and address) stop accuracy. These metrics are scaled to the number of stops tied to the route. The scaling is helpful in that it facilitates comparisons across routes with different numbers of stops, but it is problematic in that it inflates the impact of exceedances for routes with few stops. The purpose of route-level analysis is to identify route types that might systematically produce stop inaccuracies as well as to identify specific routes that might benefit from deeper exploration.

Route-Level Exceedance Rate

A straightforward accuracy metric is the route-level exceedance rate defined as the percentage of ping-stop distances for a given route that exceeds the selected threshold. This metric is calculated as the quotient of flagged to total stop events along a route (to determine the proportion) multiplied by one hundred (to express as a percentage).

Figure 33. Route-Level Ping-Stop Distance Exceedance Rate (%)

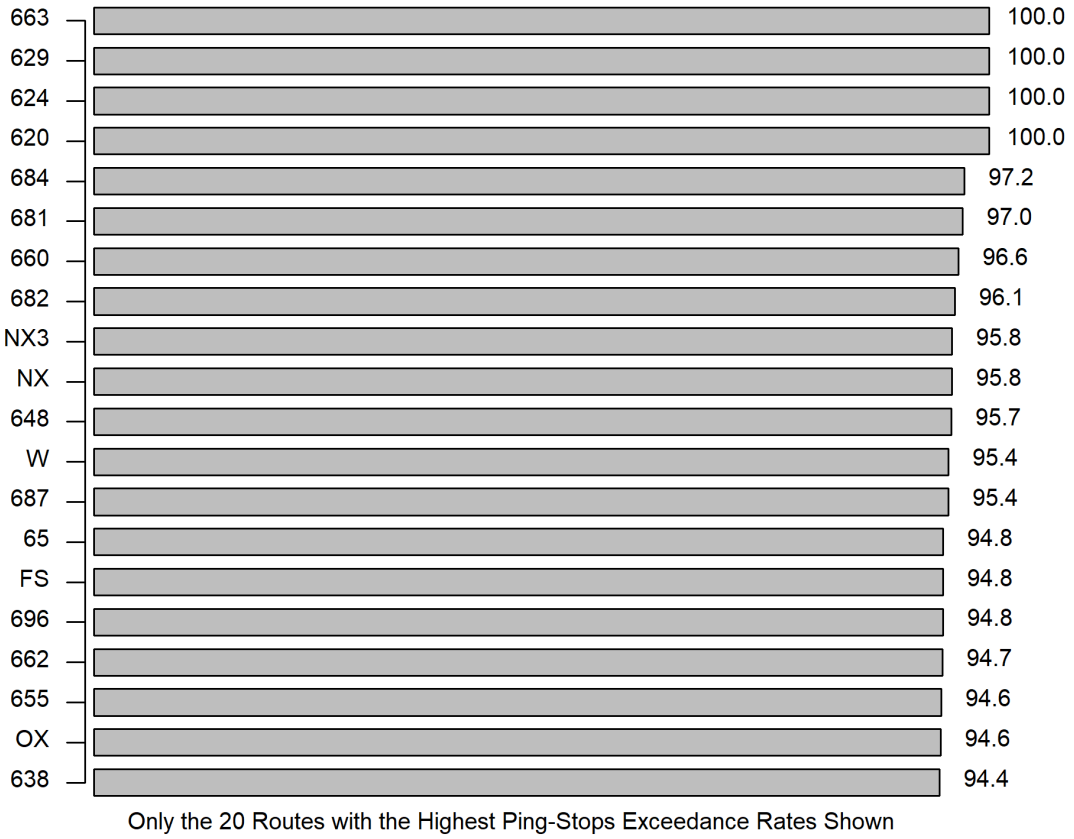


Figure 33 presents this metric for the AC Transit routes in descending order. For illustrative purposes, this presentation includes the twenty routes (regardless of type or frequency of stop events) with the highest exceedance rates. In practice, route-level metrics should be stratified by route type (i.e., all school routes, all transbay routes) for an apples-to-apples comparison and triaged by stop frequency within those strata to maximize the impacts of any intervention. Again, for practical purposes, the complete list for each stratum should be considered (rather than the censored approach presented here) to avoid border case disparities in mitigation policies. This point is especially relevant here when there is such minor variation among the rates.

Table 7. High Exceedance Rate Routes as a Share of All Routes

Routes	Total	Main #1-399	School #600-699	Early Bird #700-799	Late Night #800-899	Transbay Lettered
All Routes	**129**	60	44	3	6	16
Highest Exceedance Rate Routes	**20**	1	14	0	0	5
Shares (%)	**16**	**2**	**32**	**0**	**0**	**31**
Difference of Proportion (p)		0.005	0.017	0.500	0.324	0.111
One-Tailed Direction		-	+	-	-	+

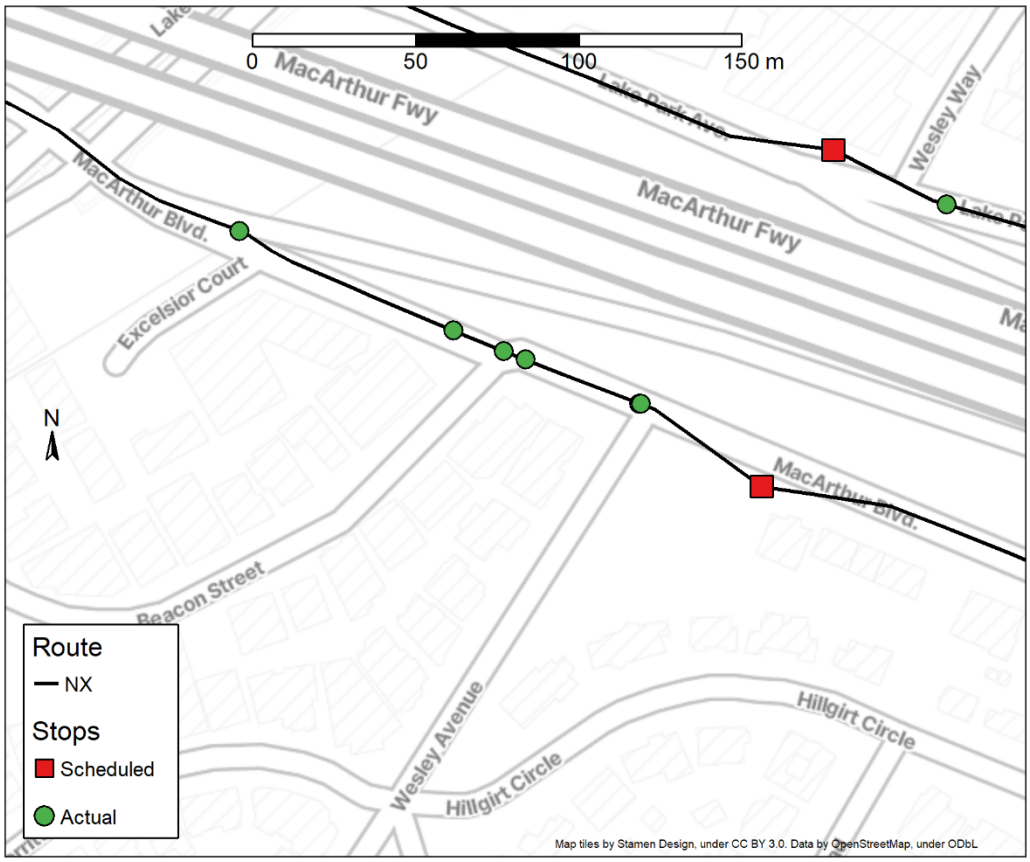
Table 7 shows the distribution of the twenty bus routes with the highest exceedance rates by type. If the type of route had no impact on the exceedance rate, it would be expected that the high exceedance routes would account for around 16% of each type. Instead, the shares of school (32%) and transbay (31%) routes are roughly double this expected number while the shares of main (2%), early bird (0%), and late night (0%) routes are practically (or actually) zero. Since the twenty highest exceedance rate sample is small, the differences between expected and observed proportions are only statistically significant at the 95% confidence level for the main and school routes. However, the transbay routes are statistically significant at an 89% confidence level, which is fairly suggestive.

School and transbay routes are commuter services that make few trips a day—almost entirely in one direction in the morning and the opposite in the afternoon. The infrequency results in many riders scheduling their travel to make the same trips each day, while the directionality results in most stops being used either for boarding or alighting—but not both. Bus drivers come to know their clientele and can tailor the stop locations to that clientele without compromising the route schedule. For example, an afternoon stop on a transbay route is likely to only involve dropping off commuters; boarding passengers would be exceedingly rare. Consequently, a driver of such an afternoon route, knowing that there is unlikely to be anyone waiting at the scheduled bus stop, might be willing to drop off regulars at an earlier point closer to off-transit destinations since that driver would not lose time with another stop at the designated location. This decision might be commonplace or affected by external conditions, such as slow-moving traffic or inclement weather. Such decisions might also reflect greater sensitivity to emerging local conditions than were embedded in the original route design.

For example, the NX/NX3 transbay routes serve commuters who work in San Francisco but live along the MacArthur Boulevard corridor in Oakland. In the eastbound direction, the NX exits the MacArthur Freeway (Interstate 580) earlier to serve riders closer to downtown Oakland, while the NX3 exits the freeway later to serve riders further out along the same corridor. Together, they represent the transbay routes with the highest ping-stop exceedance rate, as shown in Figure 33. In the afternoon, the NX bus exits from the freeway and travels along a narrow one-way section

of MacArthur Boulevard south of the interstate. Figure 34 shows the approach to the second stop in Oakland east of Wesley Avenue. None of the stop events in the data set show a bus stopping at the designated stop. Instead, all stops in the data set are made at upstream locations closer to housing and in better access to connecting streets. (Figure 34 also shows one westbound stop event with a similar pattern demonstrating the ease of waving down a driver.)

Figure 34. AC Transit Route NX in Oakland

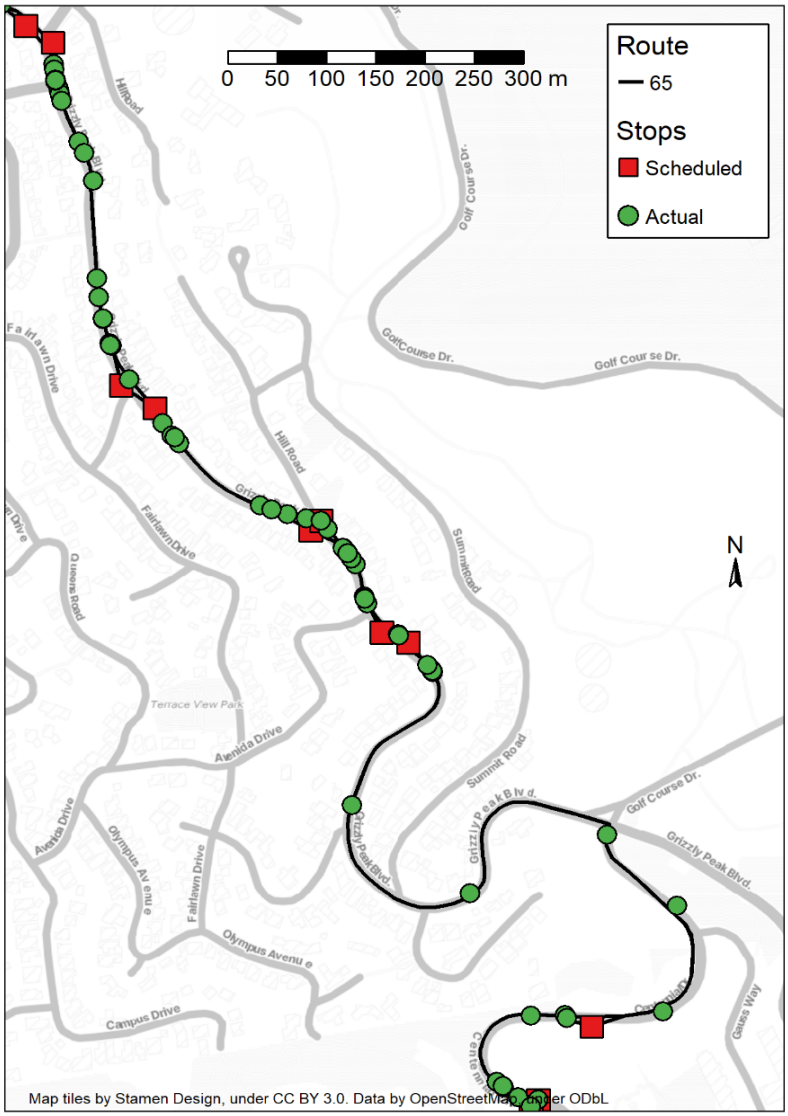


Similarly, school route drivers alter their stopping locations with sensitivity to locations that minimize pedestrian backtracking and facilitate street crossing. This feature can be seen clearly with #663, which serves several schools in Alameda. Figure 35 shows a stop event across from St. Philip Neri School. The scheduled stop in the northwest direction is at the corner of High Street and Sterling Avenue, but the bus instead stops about forty meters before the stop at a location both directly opposite the school and before a marked crosswalk which is serving many pedestrians at that point in the day. The decision to stop here facilitates safe crossing for students and reduces their distance between the bus and the school. Furthermore, it is possible that the bus driver has no choice if the vehicle is stopped anyway at the crosswalk and students request to board it right there.

Figure 35. AC Transit Route #663 Across from St. Philip Neri School

The sensitivity to local conditions extends to the one main bus route that fell into the twenty routes with the highest exceedance rates, AC Transit's Route #65. That route has an unusual path, as it travels from downtown Berkeley up along the narrow and winding Grizzly Peak Boulevard through the Berkeley Hills to the Lawrence Hall of Science. Figure 36 presents a portion of this route near that outbound terminus. While this image shows scheduled and actual stops in both directions, it makes clear that the designated stops are mostly suggestions, and the actual stop locations occur throughout the twisty route. Given the difficult terrain for pedestrians, the apparent bus driver discretion concerning where to pick up and drop off passengers is understandable, even if it is counter to the AC Transit policy to stop only at designated stop locations.

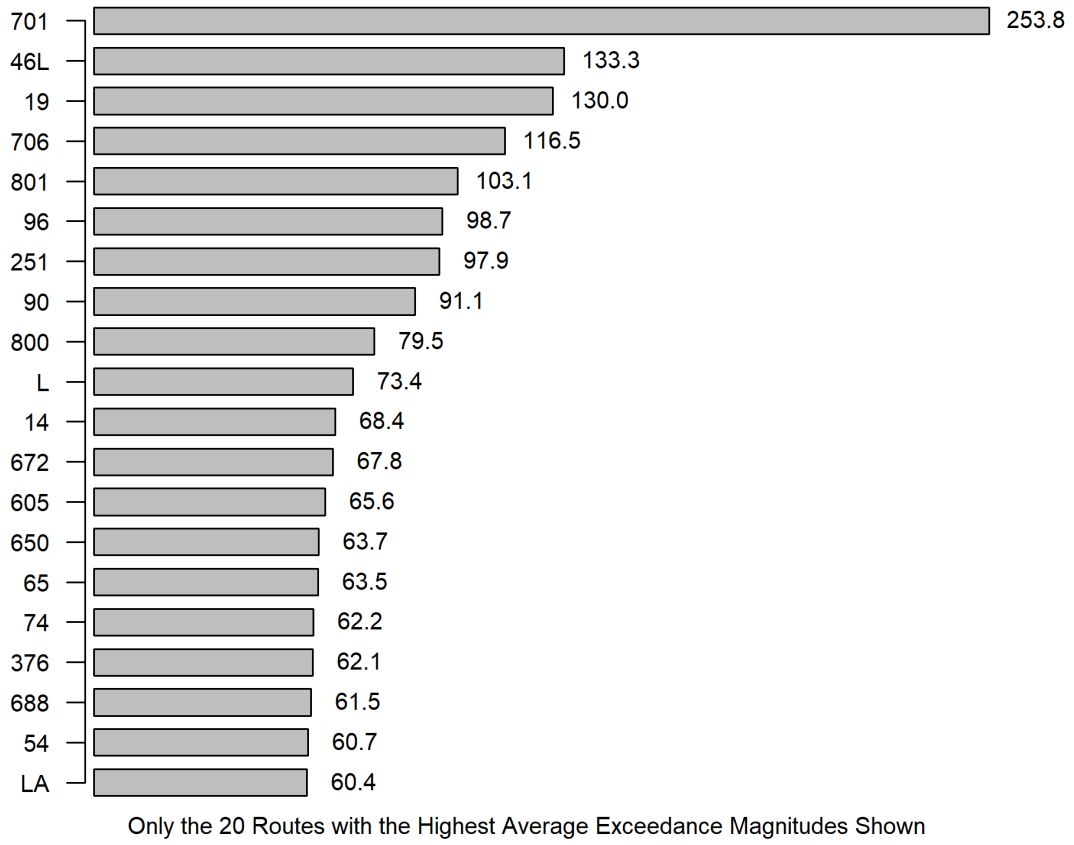
Figure 36. AC Transit Route #65 Near Lawrence Hall of Science



Route-Level Exceedance Magnitude

Another route-level approach is to consider the average magnitude of the ping-stop exceedances. This metric shifts the focus from the incidence of exceedance along a route to the typical distance of those exceedances. These distances are of greater concern to both transit planners and users since they reflect the spatial divergence between actual and designated stop locations. Figure 37 presents the twenty routes with the highest average exceedance distance.

Figure 37. AC Transit Route-Level Ping-Stop Average Exceedance Magnitude (m).



These data represent a different pattern than the exceedance rates presented earlier in Figure 33. First, all route types are represented, including the two route types that were not represented among those with the highest exceedance rates, namely early bird (#701, #706) and late-night routes (#801, #800). Furthermore, these two types are over-represented among high exceedance distance routes, as shown in Table 8, although this finding is only statistically significant at the 95% confidence level for early bird routes. Second, main routes, which were underrepresented among high exceedance rate routes, and transbay routes, which were overrepresented among high exceedance rate routes, demonstrate precisely the statistically expected share of high exceedance distance routes. Third, school routes, which were overrepresented among high exceedance rate routes, appear slightly underrepresented among high exceedance distance routes. These last findings are not statistically significant at a 95% confidence interval, so they should be considered suggestive rather than definitive. The differences in exceedance patterns are also seen in that only one route, #65, is found on both lists. These findings suggest that the special services that serve patrons at night (including pre-dawn) times, when headways are longer and passenger safety is a greater concern, are more likely to make larger deviations from scheduled stops to accommodate passenger needs. These findings also suggest that while school routes deviate from designated stop locations, the magnitude of these deviations are limited.

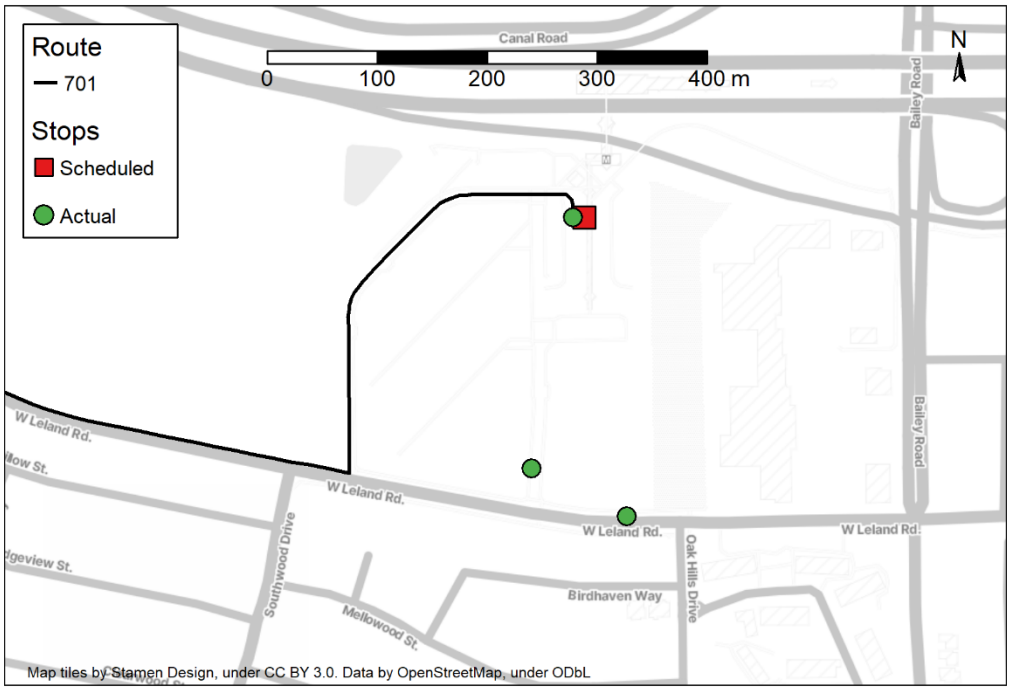
Table 8. High Exceedance Magnitude Routes as a Share of All Routes

Routes	Total	Main #1-399	School #600-699	Early Bird #700-799	Late Night #800-899	Transbay Lettered
All Routes	**129**	60	44	3	6	16
Highest Exceedance Magnitude Routes	**20**	10	4	2	2	2
Shares (%)	**16**	**17**	**9**	**67**	**33**	**13**
Difference of Proportion (p)		0.500	0.209	0.059	0.277	0.500
One-Tailed Direction		+	-	+	+	-

Early bird buses are unusual among AC Transit routes in that they each have only a few trips per day, embark from a large parking lot at a BART station, run in one direction only, and typically have only one designated stop for which GTFS Realtime data are available (since the terminal stop in San Francisco occurs underground). The high exceedance distances seem to come from the practice of picking up passengers at undesignated locations within and around the BART parking lots, or just pausing to relax with the door open before starting the trip. For example, the #701 route boards passengers on weekday mornings for three runs in close succession (3:55, 4:05, and 4:15am) at the Pittsburg/Baypoint BART station with no additional stops before disembarking passengers at the Salesforce Transit Center in San Francisco three-quarters of an hour later. As the Salesforce Transit Center is indoors and there are no pings from the final stop, all the data for the #701 route are tied to the initial stop activity, which accentuates the impact of any deviation.

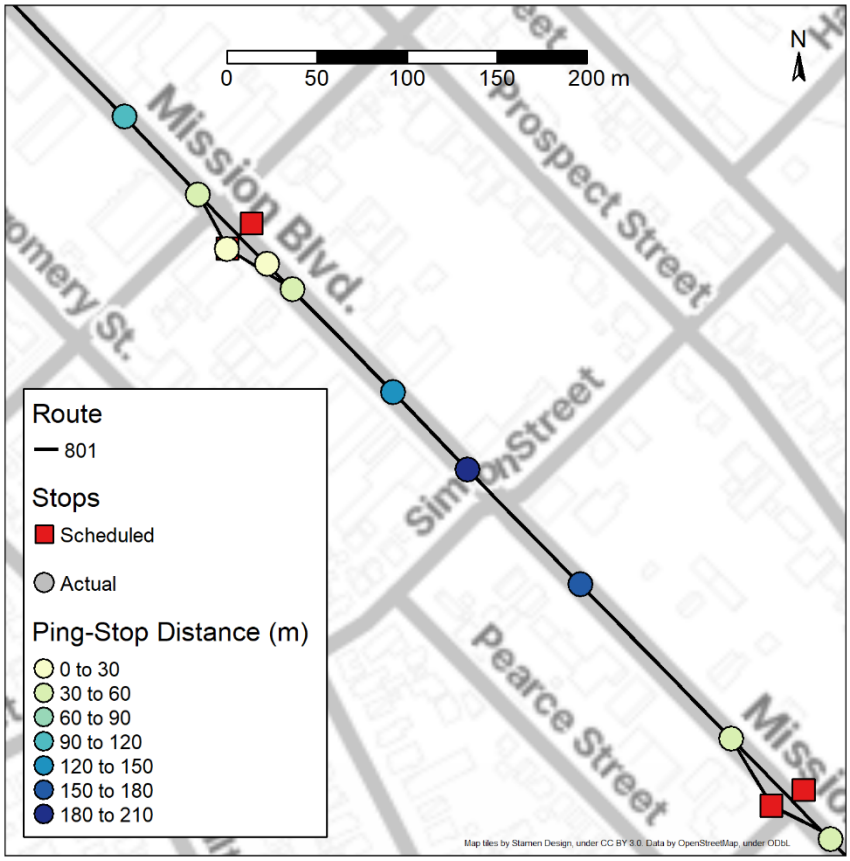
Figure 38 shows registered stop events in GTFS Realtime near the parking lot entrance before arriving at the official boarding location marked with a red square. Given the size of the BART park and ride lot, these stops are quite a distance from their scheduled location. It is difficult to assert with certainty what is driving these stop events in the minute before departure. (One interesting finding from before the data were limited to a minute before departure is that a stop was made on Leland Road further east, between Bailey Road and Oak Hills Drive, where an AC Transit stop does exist, but for different bus lines. Perhaps a confused rider was waiting there and flagged down the bus.)

Figure 38. AC Transit Route #701 at Boarding Stop (Pittsburg/Baypoint BART)



The late-night buses provide a more clear-cut understanding of the cause of high average exceedance distances. These buses provide essential service along key East Bay corridors during the overnight hours. For example, the #801 provides half-hourly service between the San Leandro and Fremont BART stations, primarily along Mission Boulevard (State Route 238/185, the historic El Camino Viejo), between midnight and seven in the morning. This route emulates the BART corridor to serve an additional four intervening BART stations as well as numerous bus stops. This route is also the late-night route with the highest average exceedance distance of 103 meters—more than the length of the shortest Olympic track event. An analysis of the stop events in the data set suggests that bus drivers regularly stop throughout the corridor. For example, Figure 39 presents stop events made along Mission Boulevard between the designated stop locations at Sunset Boulevard and Grace Street. Figure 39 color codes the stop events so that each gradation reflects an additional thirty meters (the threshold for determining an exceedance) distance. These data show that stops are made more than 180 meters or six times the threshold distance from the designated locations in the GTFS Schedule data. The land use along this section is characterized by car-oriented commercial services (i.e., body work, tires, window tinting) where one might want to travel at all times but feel uncomfortable walking at night. Bus drivers may be making stop deviations to minimize the need for riders to walk at night.

Figure 39. AC Transit #801 High Magnitude Exceedances Along Mission Boulevard



4.6 Trip-Level Metrics

Route-level metrics provide an important filtering role to identify routes that have systematic ping-stop exceedances, however they do not capture variation within a route. Since the experience of providing transit trips—even along the same route—changes throughout the day, there is a need for more detailed trip-level metrics. These metrics can reveal problems related to external traffic conditions, specific trip characteristics (such as those trips that have more or fewer stops than typical along a route), or even a specific driver/vehicle. For example, a driver may be particularly lax in their stop location adherence, or a vehicle might have a malfunctioning GPS transponder.

Trip-Level Data Pre-Processing

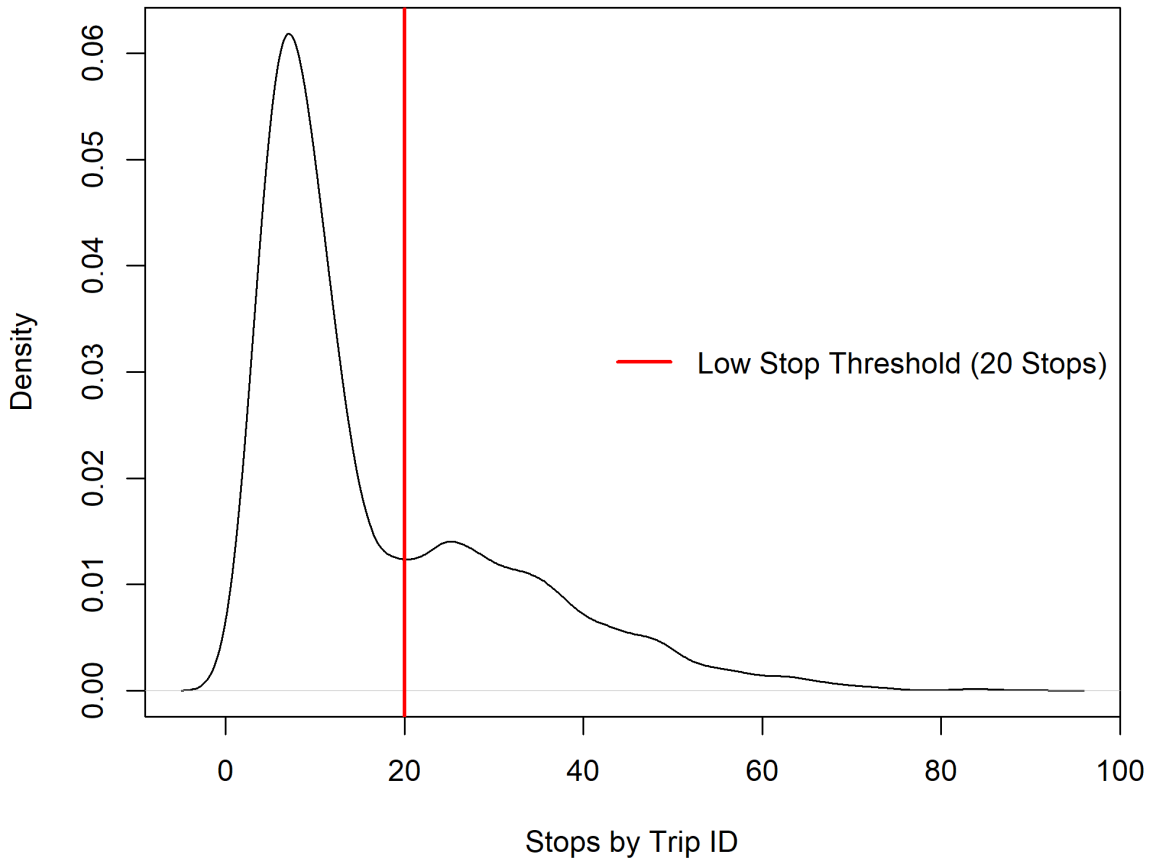
Trip-level analysis requires sufficient data collected so that trips can be compared to one another. This report emphasizes trip-level analysis among different trips for the same route, but trips from different routes could also be compared to one another, for example, to look at time-of-day effects. Similarly, this report emphasizes trip-level analysis aggregated to a single trip identifier, but trips with the same identifier made on different days could be separately considered, for example, to

look at variation throughout a week. The goal of this report is to point to analytic possibilities given sufficient data availability, not necessarily to demonstrate all of them.

To ensure a robust data set for the purpose of demonstrating trip-level analysis, only the ten routes for which at least three hundred unique trip identifiers exist are considered in this report. Furthermore, since trips, even on the same route, might have vastly distinctive characteristics, this report further culls trip identifiers that have few stop events associated with them during the study period. These specific constraints are aimed at selecting the routes with the greatest number of trips and, of those, the trips with the most stops. As noted earlier, these constraints are aimed at fulsome data sets for displaying these metrics. In practice, analysts might apply different (or no) constraints in filtering data for trip-level analysis of GTFS stop accuracy.

Figure 40 presents the distribution of stop events associated with each trip identifier. The chart shows a high peak followed by a small depression and then a much smaller secondary peak. The red vertical line at that depression represents an arbitrary cutoff to distinguish between trip identifiers with low (less than 20) and high (20 or more) associated numbers of stops. To avoid trip-level metrics from being skewed by the high number of trips with few stop events reported in their GTFS Realtime feeds, trip identifiers below this 20-stop cutoff are excluded from further analysis.

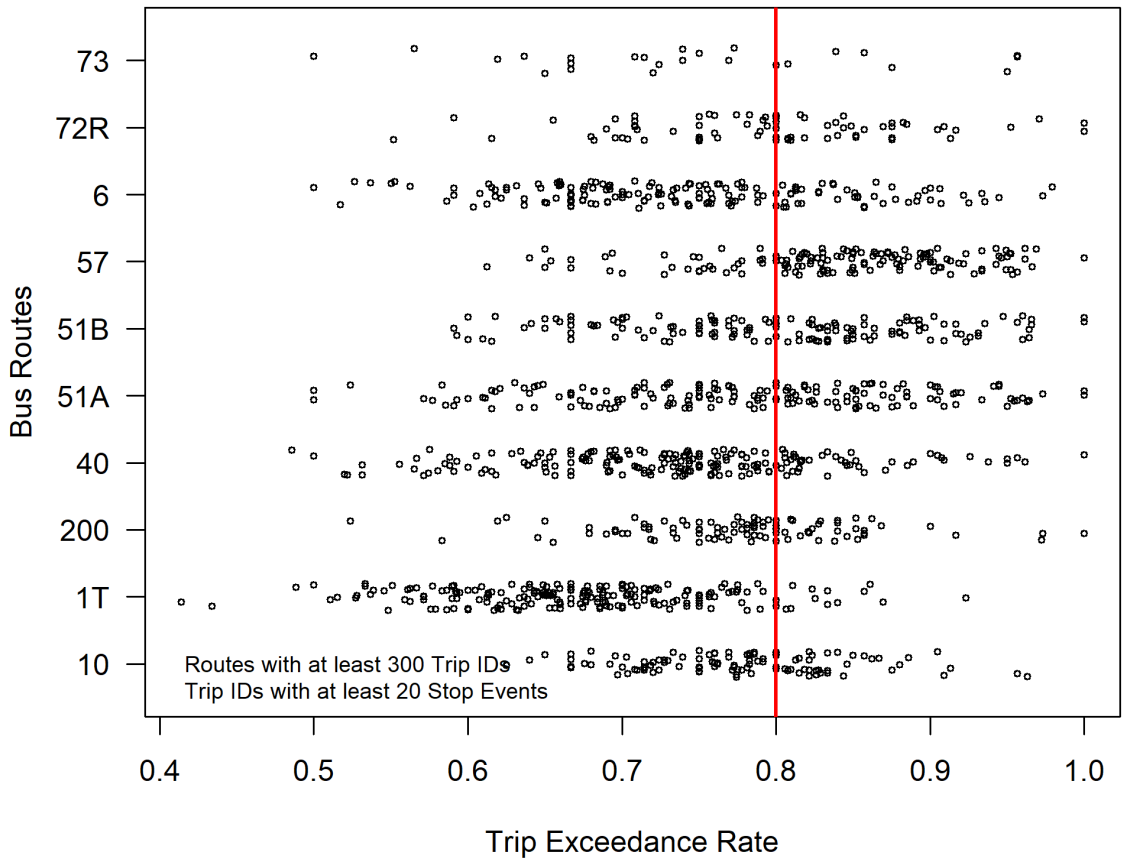
Figure 40. Density Plot of Stop Events by Trip Identifier in the Data Set



Trip-Level Exceedance Rates

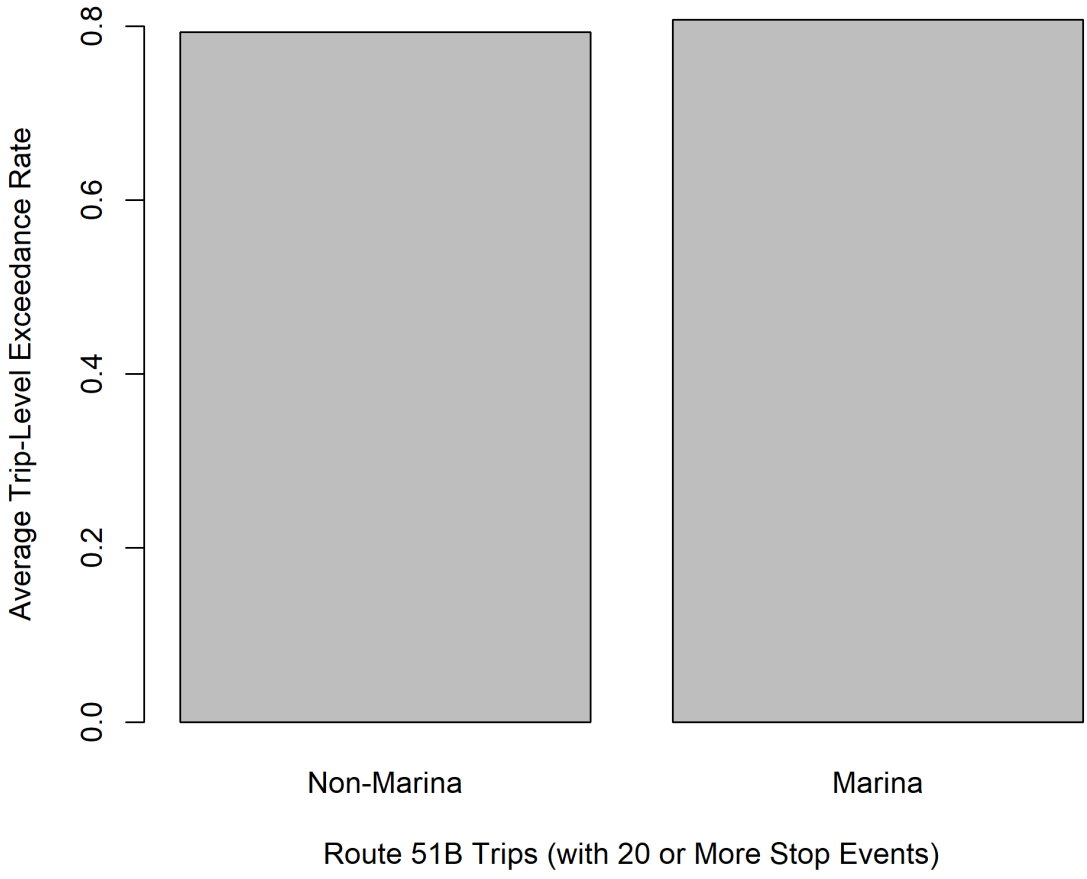
Figure 41 presents trip-level ping-stop exceedance rates for all ten AC Transit routes that have more than three hundred unique trip identifiers during the study period. These data are plotted in a strip chart that jitters points so that trips with the same or similar exceedance rates are randomly nudged up or down rather than simply overplotted. This approach makes it possible to see the exceedance rate for each trip on a given route. A vertical line representing the 80% exceedance rate is superimposed on the chart to denote an arbitrary threshold of concern. This threshold is chosen as it reflects the expected share of stop events with ping-stop distances of thirty meters or more. Marking a threshold helps visually identify trips of highest concern.

Figure 41. Trip-Level Ping-Stop Exceedance Rates by AC Transit Route



A visual analysis of the trip-level data in Figure 41 shows that some routes (i.e., #1T and #40) have few trips whose exceedance rates are above 80%, while other routes (i.e., #51B and #57) have relatively more trips above this level. These observations suggest other systematic uses of trip-level stop accuracy data. For example, Route 51B travels between the Rockridge BART station via the University of California, Berkeley campus to the Berkeley AMTRAK station. Roughly, half of Route 51B trips extend further west beyond the rail station to the Berkeley Marina. One potential area of inquiry is whether the route extension leads to a higher rate of stop inaccuracies.

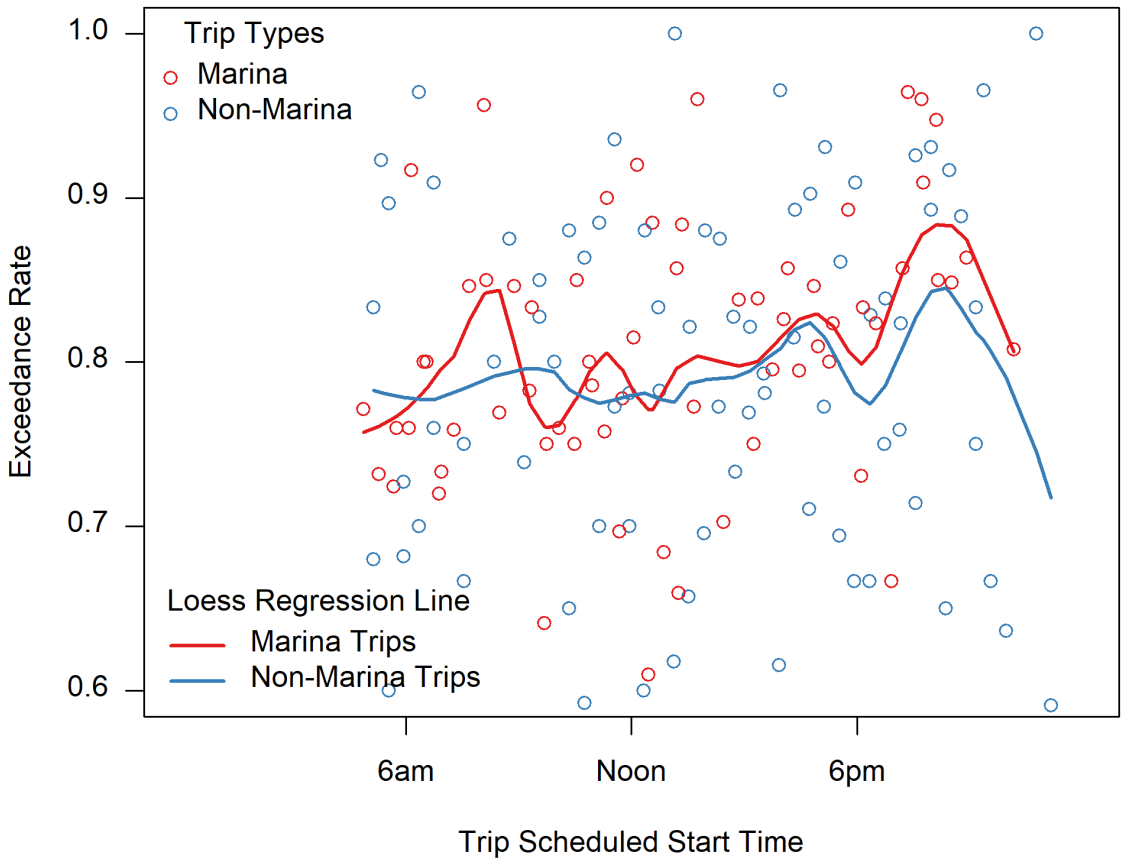
Figure 42. Trip-Level Exceedance Rate Variation by Route Extensions



By splitting 51B trips into those that serve the marina and those that do not, it is possible to compare average exceedance rates to confirm or reject that hypothesis. The data in Figure 42 suggest that, contrary to conjecture, this route extension has no meaningful impact on average ping-stop exceedance rates.

Another use of trip-level data is to explore time-of-day impacts. For example, by sorting trips based on their scheduled start time in the GTFS Static feed, it might be possible to identify systematic impacts of traffic patterns on stop accuracy. Figure 43 presents a scattergram of trip-level exceedance rates by the scheduled trip start time for Route 51B. The separation for marina and non-marina trips is maintained to demonstrate how exploratory techniques can be combined. It is difficult to see any distinct pattern in the raw plot. The addition of locally estimated scatterplot smoothing (LOESS) regression line of bandwidth 0.3 provides some additional insight. During the morning peak and evening periods, exceedance rates for marina-serving trips are higher than non-marina trips. This information might structure an assessment of the stop accuracy along this route.

Figure 43. Trip-Level Exceedance Rates by Time of Day for Route 51B



While these applications of trip-level data focus on exceedance rates, the same ideas easily port to average exceedance distance. The goal of this section is to raise the importance of intra-route variation as a cause for stop inaccuracy and to showcase some simple approaches to using trip-level data to understand those inaccuracies.

4.7 Stop-Level Metrics

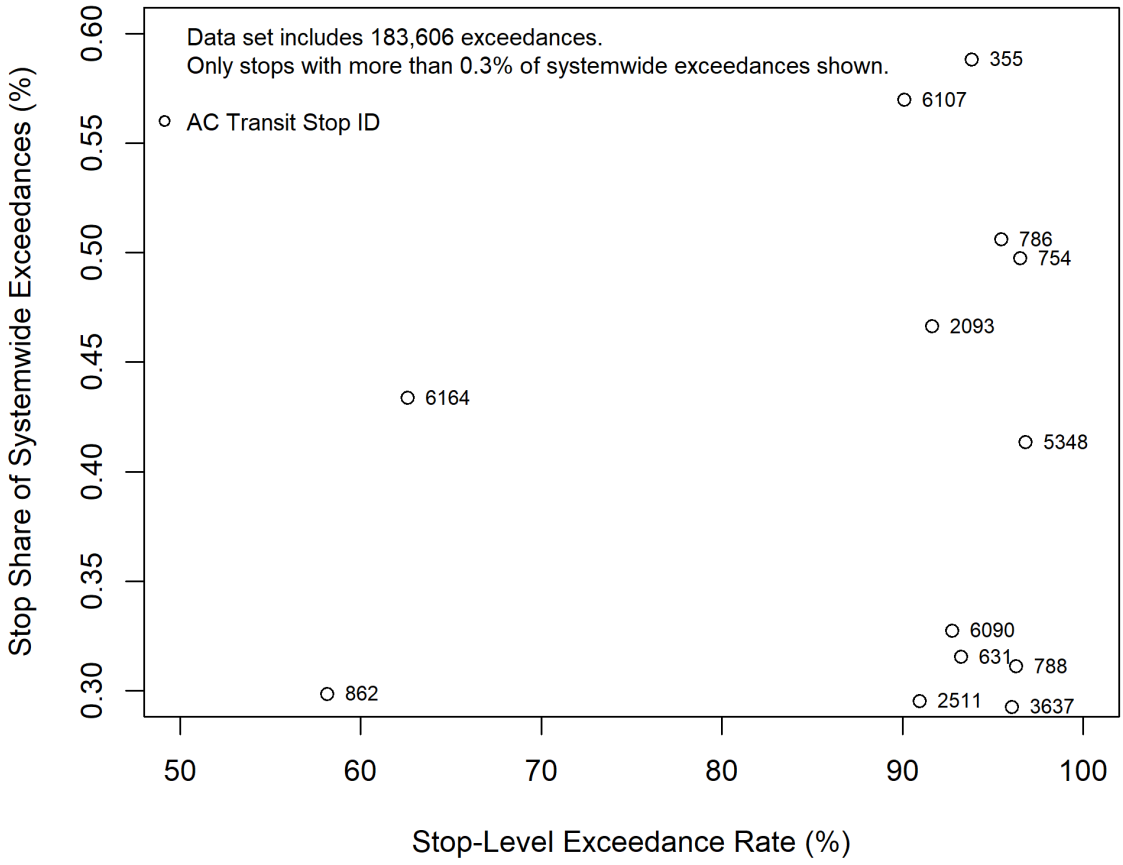
While transit design and provision focuses primarily on routes and the trips that make up those routes, there are many reasons an exploration of ping-stop exceedances might be best addressed at the stop level. First, these inaccuracies might be more closely tied to the unique geography of a specific stop location than to an etiology endemic to the entire route. Second, since many routes share the same stops—and the very act of sharing curb space may be a cause of ping-stop exceedances—it makes sense from a system operations perspective to focus on problematic stops. Finally, and perhaps most importantly, transit agencies should focus troubleshooting efforts on fixing the most problematic ping-stop inaccuracies. This filtering demands some understanding of the magnitude of the ping-stop deviations, which can be more readily seen at the stop level.

Problematic Stop Identification

Figure 44 presents a scatterplot of the AC Transit stops with the most ping-stop exceedances. The x-axis presents the share of all pings for that stop that are flagged, which is analogous to the route-level analysis above. This metric is useful for identifying the stops where pings most often diverge from the scheduled location, but not the stops which have the highest number of exceedances—and therefore are most problematic from the customer perspective. The y-axis presents the share of systemwide exceedances that occur at that stop to offer this critical information on the magnitude of the problem. Only the stops whose exceedances account for at least 0.3% of the systemwide total are shown. This threshold is entirely arbitrary and transit agencies are encouraged to select the cutoff points that make sense for the level of exceedance in their systems. It is expected that, overtime, these thresholds would shift downward as the high exceedance share locations are remedied. (After a certain point, agencies may declare they have reached a sufficient level of accuracy as to require no additional intervention.)

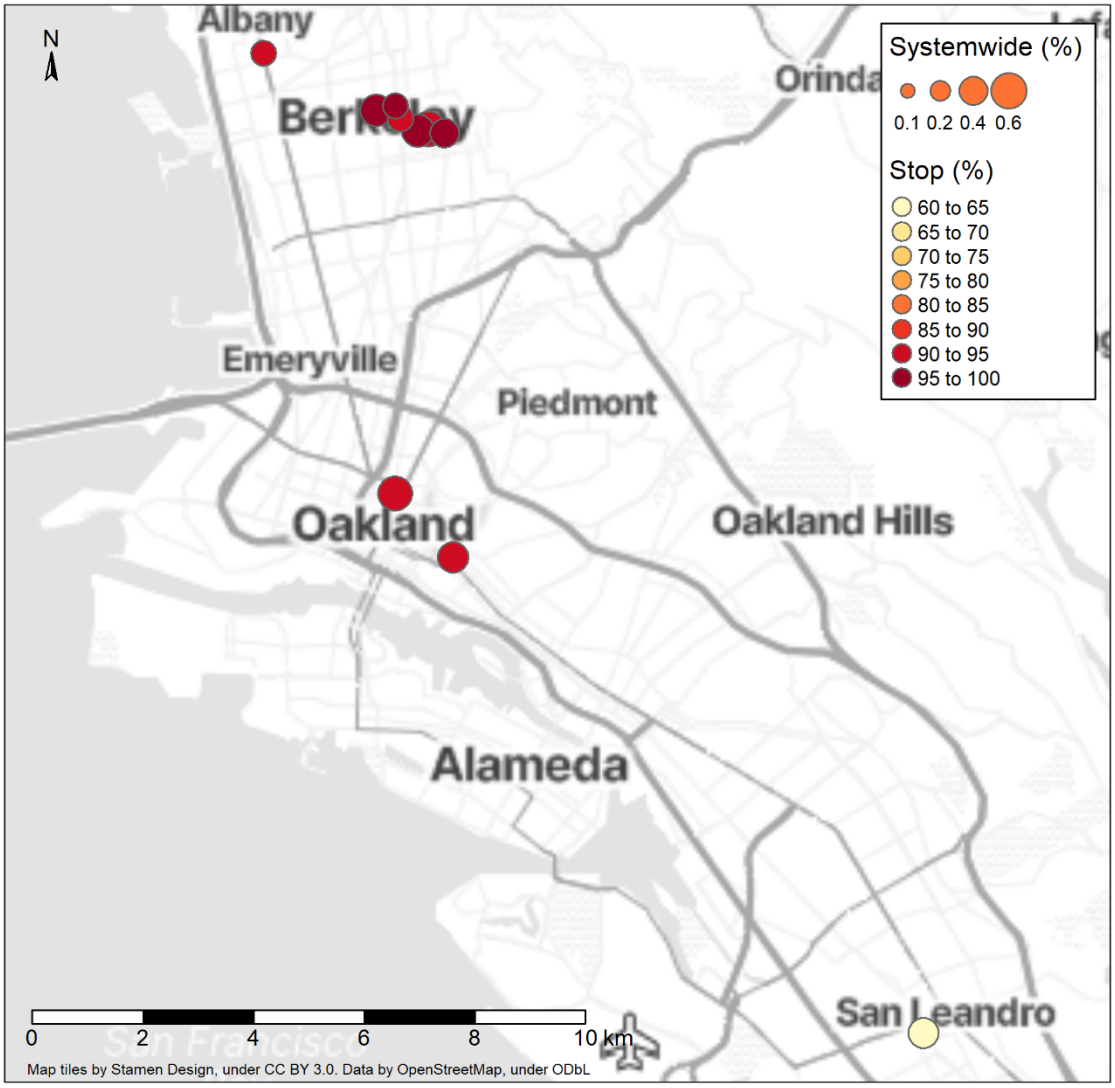
This presentation suggests a methodology for triaging problematic stops. A transit agency might proceed from the stops with the highest to the lowest share of exceedances (i.e., from the top to the bottom of the scattergram). When more than one stop reports a similar share of total exceedances, the agency might proceed from the stops with higher shares of exceedances to those with lower shares (i.e., from the right to the left of the scattergram). For example, on the bottom of Figure 44, there are several stops that each account for roughly 0.3% of systemwide exceedances, however the two stops (#2511, #3637) on the right side of the chart have stop-level exceedance rates over 90% and might be prioritized for intervention over the one stop (#862) on the left side of the chart whose exceedance rate is less than 60%.

Figure 44. AC Transit Stop-Level Exceedance Rates (Stop and Systemwide Shares).



The stops in Figure 44 are labeled with their *stop_id* from the GTFS Schedule data. These codes fit conveniently onto a graph but provide little obvious information to planners or the public on their actual location within the system. For this reason, it is best to map these locations within the transit service area as shown in Figure 45. These data show a cluster of problematic stops around the University of California, Berkeley campus as well as stray problematic stops in North Berkeley (just south of Albany), Oakland, and San Leandro.

Figure 45. AC Transit Stops with Highest Shares of Systemwide Exceedances

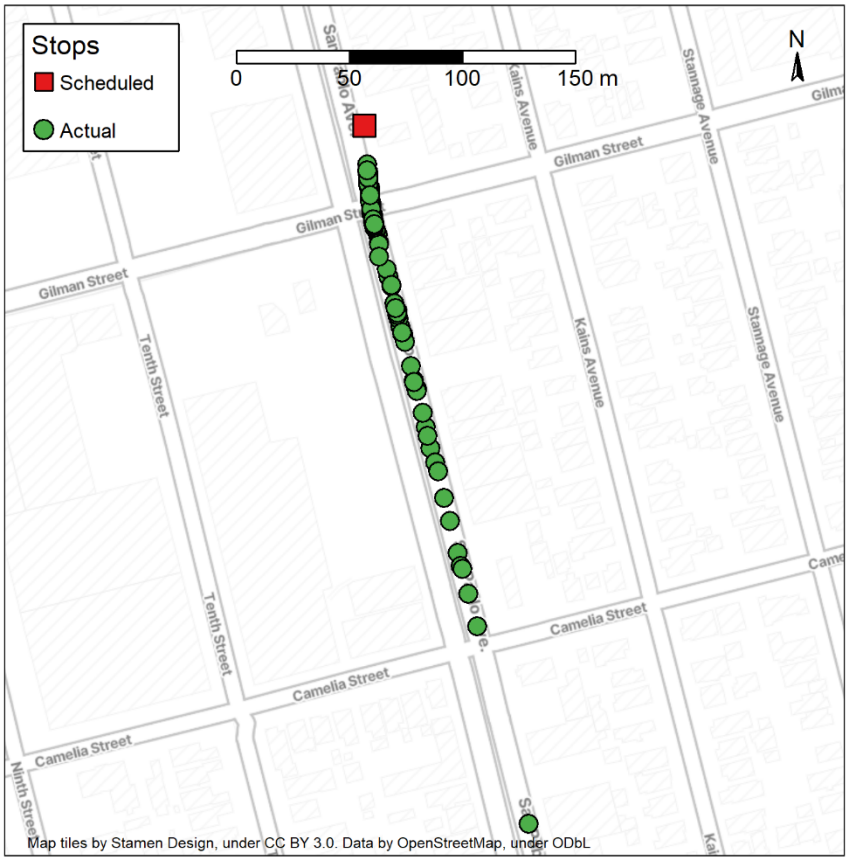


The locations shown in Figure 45 tend to be stops that serve several routes, otherwise they would not account for such a high share of systemwide exceedances. Such locations are prone to situations in which buses from the same or different routes compete for limited space at the curb. This bunching can be exacerbated by local traffic conditions.

Figure 46 illustrates the stop events within the data set that are associated with the northernmost stop in Figure 45. This stop is northbound on San Pablo Avenue (State Route 123) just north of Gilman Street. San Pablo Avenue is served by the #72, which was AC Transit's first branded "Rapid" bus route. The intersection is signalized and a common place where traffic backs up. The actual stop locations suggest patrons seek to alight (and possibly board) substantially upstream from the scheduled stop location. The availability of such information on actual stop locations

might inform a policy review from working with relevant traffic authorities on signal timing to relocating the stop upstream rather than downstream of the intersection.

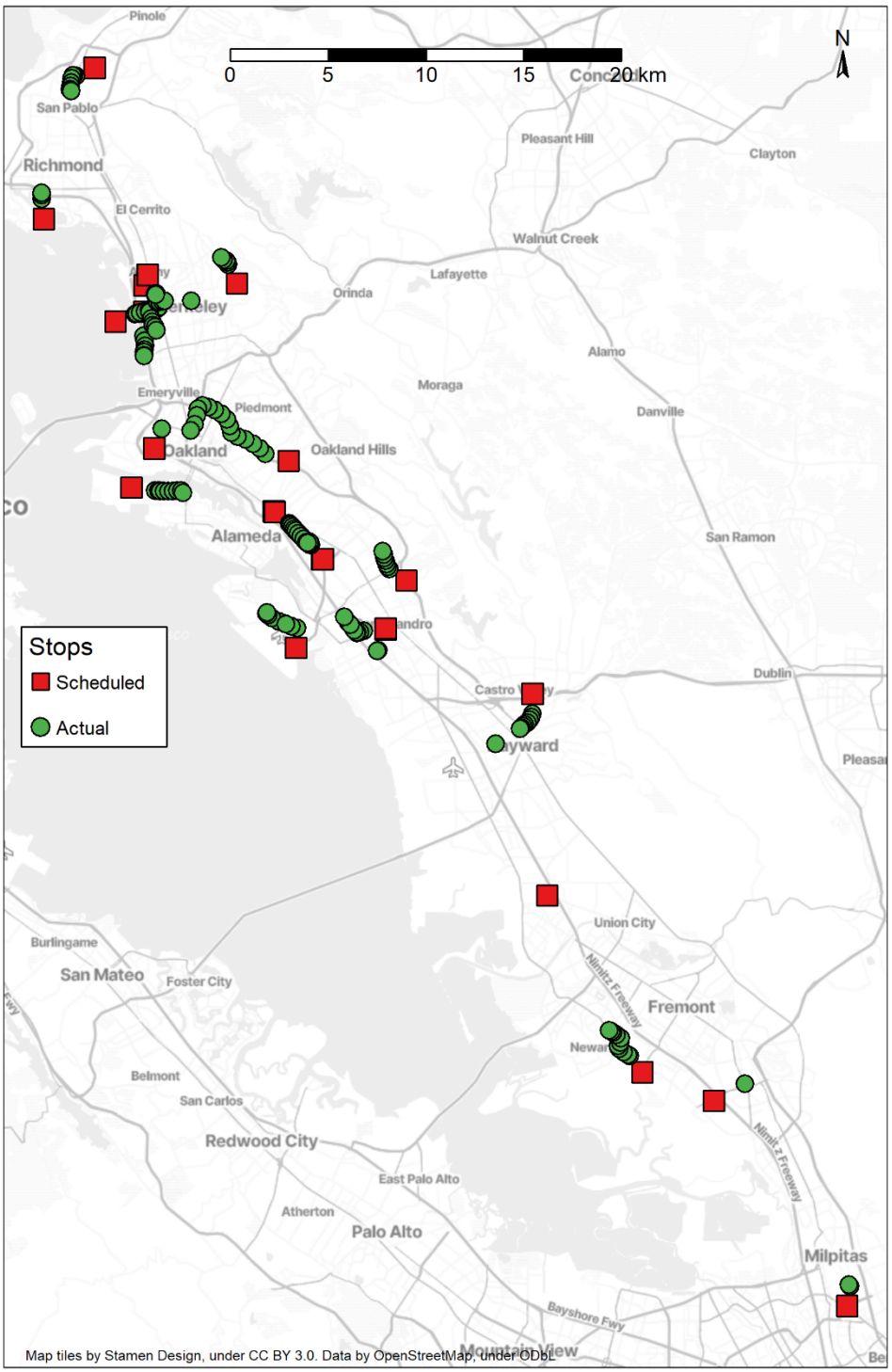
Figure 46. AC Transit Stops on San Pablo Avenue north of Gliman Street



Ping-Stop Distance Outliers

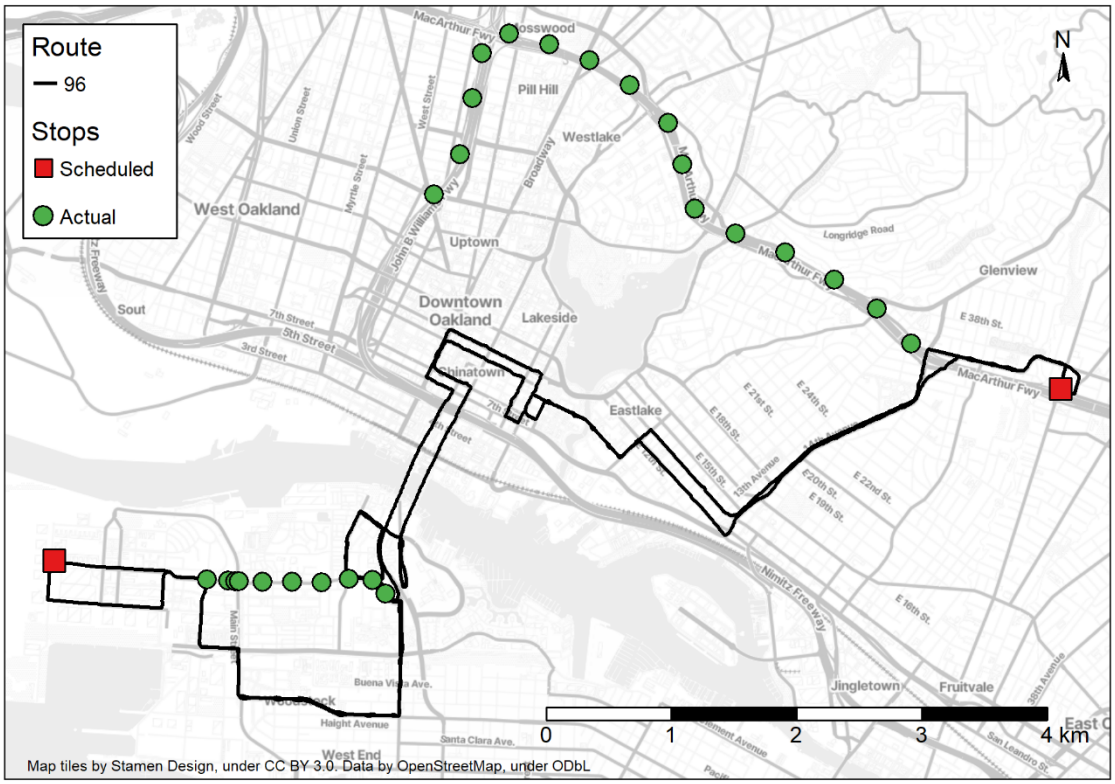
Another important application of stop-level data is to identify those pings that are at the greatest distance from the associated stop. These outliers warrant exploration and, more importantly, explanation. Figure 47 displays all the ping-stop distances greater than a kilometer. This high number was selected to identify the most unusual findings, but the analytical approach would be equally useful at the 46-meter outlier threshold determined in Figure 32.

Figure 47. AC Transit Ping-Stop Distance Outliers (Above One Kilometer)



The data presented in Figure 47 suggest that many high ping-stop distances are associated with a single stop. This finding hints at a systematic problem of the data collection rather than actual stops being made at great distances from the designated stop location. This idea is affirmed with a detailed look at two clusters of ping-stop outliers highlighted in Figure 48. These clusters are both collected from Route 96 on the same day (March 6), and both represent stop locations that are not actually along the path of the bus route. The northern cluster is along two interstate highways where one would not expect any stopping. The associated stop is the first stop in the westbound direction, but the pings are moving away from and not towards that stop. It seems that the bus driver might have been called away to a different location, but that change is not registered with the GTFS Realtime data. Similarly, the southern cluster is associated with the first stop in the eastbound direction, but the bus is traveling away from that stop also along a corridor with no scheduled bus stops but with no stopping the GTFS feeds.

Figure 48. AC Transit Ping-Stop Distances Outliers on Route 96



The regularity of pings for both clusters reflects the polling of bus locations. The larger spacing of these pings along the interstate highways in Oakland compared to the arterial in Alameda reflects the variation in respective traffic speeds, however it seems that there is some error in the signal that suggests the vehicle is at a stop in each of these moments. Stop events would not be expected

to be so regularly spaced, nor would they be expected on routes that have no designated stop locations. These strange results suggest a combination of errors—both registering a trip that is not a trip and registering a stop that is not a stop. The former might be due to a decision to alter the route in the field, while the latter might be due to hardware malfunction, such as a faulty door sensor or a door that does not properly close.

Stop-Shelter Distance

One common feature that emerges through a stop-level analysis of ping-stop disparities is the tendency for buses to stop where bus shelters or street furniture are located rather than where the official stop is designated. Figure 49 demonstrates one such location along the #51A route in Alameda where the actual bus stop is downstream from a nicely appointed shelter with seating. This location (eastbound on Santa Clara Avenue just beyond Chestnut Street) typifies a dilemma often facing bus drivers, namely, to stop where people are already waiting or force them to walk to the official stop. This dilemma can be seen in the stop-level data on ping-stop disparities when pings are consistently at a specific location which is not the official stop. Those data can be helpful for identifying situations, such as that found in Figure 49, where the official stop location is not aligned to the natural waiting location. Transit agencies can therefore use the information revealed through ping-stop disparities to make micro adjustments to their network in order to reduce user confusion and, consequently, to improve GTFS stop accuracy.

Figure 49. Stop-Shelter Distance along the #51A Route (Google Street View Image)

4.8 Conclusions

The goal of this section was to compare GTFS Schedule and Realtime data to assess the accuracy of stop locations. This analysis relied exclusively on data from AC Transit whose GTFS Realtime feed includes the optional *current_status* field that identifies whether the vehicle is at a stop. Strangely, the pings flagged with this status were rarely within even thirty meters from the stop, the distance set as a threshold for determining a ping-stop distance exceedance. This disjunction was surprising and may reflect a deeper issue in the algorithm used to convert AC Transit vehicle location data into the GTFS Realtime feed shared with the public. Those disjunctions did not come from miscoded locations in the GTFS Schedule data as checks of sampled locations within the data aligned with the location of marked bus stops in Google Street View. Those disjunctions also did not come from poor quality geocoding from faulty GPS receivers. No pings were observed to be in places that a bus could not reasonably be. Instead, the stop inaccuracies appeared to come primarily from drivers stopping upstream from designated stops and secondarily from incorrect stop flagging within the GTFS Realtime feed. In both cases, the approaches demonstrated within this research would allow for refinement of policies and software/hardware to reduce these stop inaccuracies.

These approaches to stop accuracy assessment began with a structured cleaning of the raw data. First, only those pings that were coded within the GTFS Realtime data whose *current_status* was coded as *stopped_at* were considered. These records were purged of duplicates and pings that occurred more than a minute prior to the scheduled trip start time in GTFS Static. Finally, the remaining data were consolidated to represent stop events by removing subsequent pings made on the same date for the same trip at exactly the same location.

This work set a ping-stop distance of thirty meters as the threshold for determining an exceedance, and then flagged each ping accordingly. These data were then aggregated at the route, trip, and stop level to aid in the identification of inaccuracies. The analysis of the AC Transit data showed statistical differences in route-level exceedance shares and magnitudes by different route types, time of day differences in trips made along a single route, and particularly problematic stop locations throughout the East Bay. This work also revealed unusual patterns of ping-stop outliers made by incorrect attribution of bus movements outside designated paths, and a more prosaic problem of formal stop locations located at some distance from associated shelters and street furniture.

The techniques presented to compare GTFS Schedule and Realtime data to identify ping-stop inaccuracies also served to monitor the success of interventions to reduce these inaccuracies. It is hoped that transit agencies will implement these straightforward approaches to analyze their operations, identify areas of concern, craft appropriate interventions, and monitor the outcomes of those actions over time.

The intention of this section was to demonstrate the ease by which GTFS products can be used to align formally designated stopping locations with the locations where buses actually stop.

5. Conclusion

The exciting transition to GTFS entails a commitment to the accuracy of its underlying data products. This accuracy is essential to achieve the full benefits promised by GTFS. Unfortunately, the GTFS Static and Realtime feeds are not pristine. They contain inaccuracies that affect their effectiveness for trip planning and other purposes. Fortunately, because these feeds reflect real-world phenomena that are both spatially and temporally bound, there are ways to assess this accuracy. The innovation of the work presented in this report is to make that assessment by comparing the different GTFS products to each other within the constraining context of the built environment and the linear structure of time. This approach triangulates these various sources of data to identify GTFS inaccuracies and suggest the source of this discrepancy. This work can help transit agencies refine the GTFS products they share with the public. The elegance of this approach is that it can be readily replicated by anyone using the publicly available data broadcast by a given transit agency alongside basic geographic information systems software.

This study finds one source of inaccuracy in the GTFS Schedule products that define what the system should do. In many cases, the actual paths defined for the vehicles are poorly coded in the *shape.txt* file. Common problems include flows in two directions snapped to a single roadway centerline, sloppy coding of path vertices that result in shapes unaligned to the actual roadways, and well-coded shapes that do not reflect the paths bus drivers actually use. Such problems are easy to fix simply by updating the vertices in the shapes file within the GTFS Schedule data. This research demonstrates how the preponderance of vehicle point locations in the GTFS Realtime feed can be harnessed to identify the location of these GTFS Static coding inaccuracies by flagging route segments whose ping-path distances exceed a threshold of acceptability.

Conversely, ping-path exceedances can also flag inaccuracies with the GPS systems that report the vehicle positions. For example, this report documents blocks in downtown Modesto in which the GTFS Realtime feeds placed vehicles—despite the reality that these blocks are occupied by buildings and not roadways. The clustering of these errors to specific areas rather than their spread across the system may suggest local conditions affecting the GPS signal. Unfortunately, the Modesto pings always occur in clusters as if they are geographically triggered rather than spread throughout the system as would be expected if they are temporally triggered—and this clustering might result in a limited view of the incidence of inaccuracies throughout the network. The solutions here are less simple than a careful recoding of the GTFS Static resources but are still worth exploring. Pings should be polled based on time, not space intervals, to ensure a representative sample across the systems. Areas in which ping locations for transit vehicles are found to be impossible need to be examined to better assess what is causing such inaccuracies and how to remedy them.

These identified inaccuracies in the path descriptions (including those due to driver deviations from those intended paths) and in the ping locations of vehicles are likely to confound the algorithms that predict stop arrival times in the *TripUpdate* messages. Those algorithms compound this confusion by broadcasting rather than filtering out nonsensical data, such as the continuity errors and multiple predictions with the same timestamp noted in this report.

This research provides an approach to assess the arrival prediction messages that allow users to minimize their transit wait times. That approach entails the careful cleaning of the extensive trip update messages and then the generation of a series of performance measures, including Update Availability, Prediction Error Percentile Plots, Scaled Prediction Error IQR, Bus Catch Likelihood, Expected Wait Time, Prediction Padding, and Prediction Inconsistency. These metrics allow transit agencies to evaluate and refine the information they provide to the public. These metrics also provide a consistent manner to compare the quality of different proposed algorithms for predicting arrival times to facilitate a more complete consideration of what constitutes good prediction accuracy. In short, this research provides a myriad of ways to consider and thus refine the accuracy of the *TripUpdate* messages.

Transit agencies are responsible for providing customers the best arrival information possible. This research offers metrics for calculating the accuracy of these predictions, and more importantly, measuring these inaccuracies in terms of their implications for users. For example, one metric identifies the time padding necessary to buffer the predicted arrival time to ensure catching the desired bus. Another metric provides an accounting of how long the user will wait should they follow the prediction precisely (and possibly miss the desired bus and need to wait for the subsequent vehicle). Some of these metrics, such as the one that tracks the availability of updates, or the one that calculates the absolute change across subsequent predictions for the same stop arrival, represent aspects of the consistency and reliability of prediction information transit agencies share with their customers. While one might argue these metrics are not direct measures of prediction accuracy, they certainly reflect the customer experience of those predictions. In other words, if predictions are not made available or are maddeningly variable, users will discount their value. These measures reiterate a central premise of this work: That the goal of sharing GTFS products with the public is to facilitate transit use.

The findings presented in this report offer techniques to systematically identify inaccuracies within the GTFS Static and Realtime ecospheres. It is hoped transit agencies will adopt these approaches to review the information they share with the public to continually prune problems and thus, over time, reduce the incidence of inaccuracies.

List of Acronyms

Acronyms	Description
GTFS	General Transportation Feed Specification
GTFS-RT	General Transportation Feed Specification (GTFS) Realtime
AVL	Automatic Vehicle Location
MAX	Modesto Area Express
AC Transit	Alameda-Contra Costa Transit District
MST	Monterey-Salinas Transit
BBB	Big Blue Bus
OCTA	Orange County Transportation Authority
StaRT	Stanislaus Regional Transit
StanRTA	Stanislaus Regional Transit Authority
.pb	Protocol Buffer
GPS	Global Positioning Systems

Bibliography

Abusalim, M. (2020). *Accuracy and Effectiveness of GTFS Transit Feeds for Trip Planning in Public Transit Networks*. Carinthia University of Applied Sciences, Villach, Austria.

Alameda-Contra Costa Transit District. (2021). *Alameda-Contra Costa Transit District: 2021 Annual Agency Profile*. 90014 2021. Washington, D.C. USDOT Federal Transit Administration. https://www.transit.dot.gov/sites/fta.dot.gov/files/transit_agency_profile_doc/2021/90014.pdf

Barbeau, S.J., & Fretheim, D. (2018). *Meeting & Exceeding Mobility User Expectations with Real-Time Transit Information*. Transportation Research and Education Center (TREC) Webinar Series. Portland, OR. https://pdxscholar.library.pdx.edu/trec_webinar/31.

Machlab, F., Riegel, L., Sood, R., & Warade, R. (2017). *A customer-focused methodology for determining prediction accuracy using automatically collected data*. Conference Paper presented at the Transportation Research Board 96th Annual Meeting, Washington, D.C.

Steiner, D., Hochmair, H., & Paulus, G. (2015). Quality Assessment of Open Real-time Data for Public Transportation in the Netherlands. *GI_Forum Journal of Geographic Information Science* 1: 579–588.

Swartz, P. (2020). *WMATA GTFS-RT Response. Massachusetts Bay Transportation Authority.*

U.S. Census Bureau (2021). U.S. Census Bureau QuickFacts: Modesto city, California. U.S. Department of Commerce. https://www.census.gov/quickfacts/modestocitycalifornia

Wessel, N., & Farber, S. (2019). On the accuracy of schedule based GTFS for measuring accessibility. *Journal of Transport and Land Use,* 12(1). JSTOR: 475–500.

Wessel, N., & Widener, M.J. (2017). Discovering the space–time dimensions of schedule padding and delay from GTFS and real-time transit data. *Journal of Geographical Systems,*19(1). Springer: 93–107.

About the Author

Gregory L. Newmark, PhD

Dr. Newmark is an Associate Professor in the Department of Design and Planning in the School of Architecture and Planning at Morgan State University. His research focuses on fostering the sustainability of travel, particularly through transit provision and use. Dr. Newmark seeks to make emerging transit data sources meaningful for policy.

MTI FOUNDER

Hon. Norman Y. Mineta

MTI BOARD OF TRUSTEES

Directors