Journal of Public Transportation | scholarcommons.usf.edu/jpt
Vol. 21 No. 2 [2018] pp. 19-40

Journal of Public Transportation

Understanding Transit System Performance Using AVL-APC Data: An Analytics Platform with Case Studies for the Pittsburgh Region

Xidong Pi, Mark Egge, Jackson Whitmore, Zhen (Sean) Qian
Carnegie Mellon University

Amy Silbermann
Port Authority of Allegheny County

Abstract

This paper introduces a novel transit data analytics platform for public transit planning, assessing service quality and revealing service problems in high spatiotemporal resolution for public transit systems based on Automatic Passenger Counting (APC) and Automatic Vehicle Location (AVL) technologies. The platform offers a systematic way for users and decision makers to understand system performance from many aspects of service quality, including passenger waiting time, stop-skipping frequency, bus bunching level, bus travel time, on-time performance, and bus fullness. The AVL-APC data from September 2012 to March 2016 were archived in a database to support the development of a user-friendly web application that allows both users and managers to interactively query bus performance metrics for any bus routes, stops, or trips for any time period. This paper demonstrates a case study using the platform to examine bus bunching in a transit system operated by the Port Authority of Allegheny County (PAAC) in Pittsburgh. It is found that the incidence of bus bunching is heavily impacted by the location on the route as well as the time of day, and the bunching problem is more severe for bus routes operating in mixed traffic than for bus rapid transit, which operates along a dedicated busway. Furthermore, a second case study is presented with a comprehensive analysis on a representative route in Pittsburgh under schedule changes. Suggestions for operation of this route to improve service quality are proposed based on the data analytics results.

Keywords: *Transit system, Automatic Vehicle Location, Automatic Passenger Counting, data analytics platform, performance metrics, bus bunching, service quality*

© 2018 Xidong Pi, Mark Egge, Jackson Whitmore, Zhen (Sean) Qian & Amy Silbermann
http//dx.doi.org/10.5038/2375-0901.21.2.2
ISSN: 1077-291X | Licenced under Creative Commons License Attribution - Noncommercial 4.0

The *Journal of Public Transportation* is published by the Center for Urban Transportation Research at the University of South Florida

Introduction

Public transit plays an important role in sustainable mobility. It provides an affordable way for all residents to travel across the city. It also helps to reduce congestion in urban road networks and reduce energy consumption in the daily passenger transport of cities (Barrero et al. 2008). To attract more riders, public transit agencies have been trying to improve their service efficiency and reliability by reducing passenger waiting time, reducing dwell time, and rebalancing the distribution of service to meet travel needs. This paper develops a transit data analytics platform for systematically understanding transit system performance (efficiency, reliability, and service quality) with the increasingly available Automatic Passenger Counting (APC) and Automatic Vehicle Location (AVL) data. The data platform has been deployed for assessing the service quality of a transit system in the Pittsburgh region operated by the Port Authority of Allegheny County (PAAC).

Prior to the advent of AVL-APC systems, much of the data for the performance metrics used by transit agencies to assess their systems had to be painstakingly collected and tabulated by hand. Transit agencies either conducted passenger satisfaction surveys by asking passengers about their transit experience (Nathanail 2008; Eboil and Mazzulla 2009) or hired observers to record bus punctuality and fullness at bus stops (Badami and Haider 2007). The expenses associated with these processes meant that such data was collected only as needed or as scheduled by the agency in question. Thus, any researcher or member of the public who was interested in looking into the performance of their transit system had access only to whatever data the transit agency made publicly available. Alternatively, video cameras paired with computer vision algorithms can be used by transit agencies to automatically collect data on buses or at bus stops, e.g., the passenger count (Kimpel et al. 2003; Jaiswal et al. 2008). However, this data collection method is costly and generates limited sample sizes.

As AVL-APC systems have become more widespread and the cost of storing data has decreased drastically, agencies have found themselves in possession of much of the data researchers or citizens might be interested in. These AVL-APC systems not only provide much bigger data coverage in spatial and time ranges, but also ensure a high level of data quality if a suitable quality control process is applied when pre-processing the raw data. In fact, in the early 2000s, some transit agencies such as TriMet in Portland, Oregon, collaborated directly with university researchers to have them evaluate the data and identify efficiencies that could be realized via service modifications (Strathman 2002). Moreover, in some state-of-the-art research, AVL-APC data can be used to evaluate bus service reliability, hence improving the schedule and performance of a bus system (Cham 2006). AVL-APC data may also be used to estimate bus running time (Tétreault and El-Geneidy 2010) or to predict arrival and departure times (Shalaby and Farhan 2004) at certain bus stops. In Feng and Figliozzi (2011), the spatiotemporal causes of bus bunching phenomena were studied using a half year's AVL-APC data for a low-performance TriMet route. The on-time performance of certain bus routes can also be improved using AVL data, as presented in Cevallos et al. (2011).

On the other hand, with the development of web technology and the popularity of web applications, web-based open-source urban traffic and transit planning tools have emerged in recent years. These web-based tools can significantly improve the efficiency of data fetching, data analysis, and the decision making of urban traffic and public transit planners. Two websites that specialize in transit planning for transit agencies include UrbanEngines.com and Remix.com. Another web-based application designed for data analysis of the transit system in Singapore is BusViz, which uses passenger fare card data (Anwar et al. 2016). However, BusViz results rely on the assumption that every passenger correctly taps their EZ-Link card when boarding and alighting. Other transportation data analytics platforms designed or developed in recent years include the traffic big data platform by Mian et al. (2014), the Sipresk (Khazaei et al. 2015), and the conceptual platform Godzilla (Shtern et

al. 2014), which were designed for general traffic data or highway traffic data. The platform discussed here was developed specifically for bus transit data analytics.

By combining the advantages of web applications (low cost for development and maintenance, convenience for access anywhere and anytime, customizability, and scalability) and of AVL-APC data (low cost, broad coverage, and high quality), the data analytics platform presented here helps automate the public transit data requests and analysis process with high spatiotemporal resolutions. There are two major advantages when compared to similar platforms such as BusViz and Remix:

1. All data fetching, visualization, and calculations in this platform use archived AVL-APC data. The dataset is able to provide historical bus arrival, departure, and load data accurately down to the trip and stop level, so queries may be performed for a single trip, stop, or route, up to the transit system in the Pittsburgh city area as a whole.
2. The platform offers a systematic way for users and decision makers to examine system performance from many aspects of service quality, such as passenger waiting time, bunching, bus trips that skip stops due to heavy loads (stop-skipping), crowding level, travel time, and on-time performance. Those metrics can be accurately estimated in the platform at any desired spatiotemporal resolution, allowing a better understanding of system performance.

The paper is organized as follows. First, the system architecture of the transit data analytics platform is described, followed by details of the data used. We then introduce the different functionalities of our platform and various transit performance metrics. In the next section, we perform two case studies using the platform to study the causes, patterns, and possible solutions to bus bunching in Pittsburgh, as well as the influence of a schedule change of Route 61B in Pittsburgh. Finally, conclusions are drawn.

System Architecture

AVL-APC systems include an onboard computer, a GPS device to record bus location (AVL), and a laser sensor at each bus door that counts passengers as they board and alight (APC). In combination, these devices comprise a system capable of recording a plethora of statistics regarding each bus trip. We built a transit data analytics platform by combining the AVL-APC data and modern web technology. Transit agencies and other users can access or download the AVL-APC data, or query various kinds of performance metrics and visualize the analytical results on a map or in charts, by simply using a cell phone or personal computer (PC) with a modern web browser.

Figure 1 illustrates the system architecture of this platform. After collecting the AVL-APC data from onboard sensors, a quality control process is applied to the raw data to remove illogical or missing records, including those with unidentifiable route or stop numbers; trips that cannot match the blocks/stops in the correct sequence along the route; and records with invalid on/off/load values, such as a negative load. The processed sensor measurements and the General Transit Feed Specification (GTFS) data, which contains geographical information for all bus stops and schedule information for all bus routes/trips, are stored in our structured query language (SQL) database. Then we use the Django web framework and Nginx HTTP server to set up our website. Django and Nginx are both free and open source, and they work well together to create a clean design, rapid development, and easy deployment. When users start a query on their terminals (cell phones or PCs), the query will pass through the web browser, the Internet, and the Nginx server, and then reach the Django framework. Django processes the query, fetches the requested data from the SQL database, then computes

different performance metrics and sends the data and analytical results back to the user terminals. Finally, the web browser presents the query results to users in the form of data files, data charts, or various types of visualization on maps.

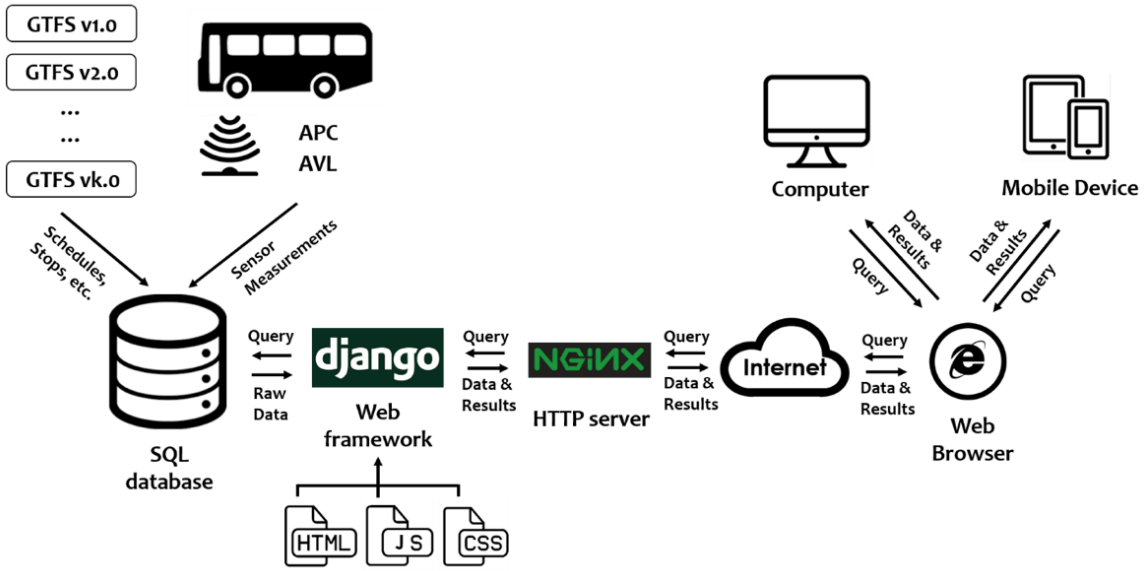


FIGURE 1.
The system architecture of the web platform

Using this data platform, users need only to specify several query conditions, including date and time range, bus routes or bus stop, type of metrics needed, and data format or visualization preferred, on any portable device. They can quickly obtain the desired AVL-APC data or analytical results for any date and time-of-day range for any bus route/trip or bus stop. While this framework works generally for any AVL-APC datasets, this paper demonstrates it using data from the Pittsburgh region.

Data Description

Usually the AVL-APC system is activated when a bus departs its depot and begins a trip. From the beginning of the bus run, the system records basic information such as day of the week, route number, direction (inbound, outbound), trip number, vehicle number, and date. As a bus progresses along its route, the system records information at each stop, which includes the stop number, arrival time, departure time, and number of boarding and alighting passengers. Certain stops are designated as time points. These have a scheduled arrival time for the bus, which allows the system to calculate any deviation of the bus from its planned schedule. All this data can be recorded onto the system's onboard memory, downloaded upon the bus's return to its depot, and then integrated into a data warehouse. To provide a quick overview of the AVL-APC data stored in our platform database, Table 1 lists the main components, including the trip information, stop information, AVL data, and APC data of each data entry.

TABLE 1.

Main Components of Stored AVL-APC Data

Component	Detail Description
Trip Information	Date, day of week, route, direction (inbound or outbound), trip ID number, vehicle number
Stop Information	Stop ID number, stop sequence number, stop name
AVL Data	Arrival and departure times (at this stop), dwell time, scheduled arrival time, schedule deviation
APC Data	Number of passengers boarding and alighting, number of passengers on bus

All calculations in the analytics platform are made using these AVL-APC data. In July 2012, PAAC installed APC and AVL systems on all of its buses.

Metrics and Functionalities

Passenger Waiting Time

Quantifying the excess waiting time experienced by riders at bus stops helps determine the efficacy of a bus system at providing reliable service for customers. Waiting time includes two components: waiting time before and waiting time after the scheduled arrival. Waiting time before the scheduled arrival was generally a function of headway length and passenger arrival time, while waiting time after the scheduled arrival was related to schedule reliability (Salek and Machemehl 1999). However, these approximations do not sufficiently capture the effects of service irregularity on passengers because of the difficulty in acquiring the arrival and departure time data required to compute them. With the advent of AVL-APC technology, the arrival and departure time data of every trip is readily available for transit agencies to analyze (Furth et al. 2006).

Transport for London (TfL), the organization that runs London's transit system, currently computes waiting time in two categories: scheduled waiting time and average actual waiting time (2014). From these two measurements, excess waiting time is calculated by subtracting the scheduled waiting time from the average actual waiting time. For buses running with no variability and as scheduled, excess waiting time would equal zero. Unlike the traditional measures described above, this measurement incorporates historical schedule deviation data to better approximate the effect of the deviations on those using the system. TfL's waiting time measurement is intended for use on buses with headways under 12 minutes, shorter than the headways of the majority of PAAC's routes. To help account for this discrepancy, options for two passenger arrival distributions are offered. The first will be a uniform distribution, indicative of passengers arriving randomly for a high-frequency route, and the second will be a skewed passenger arrival distribution (a shifted Johnson S_B distribution), which will assume people try to time their arrival at the bus stop to coincide with the bus arrival for long-headway routes. Users can choose the passenger arrival distribution they want to apply when performing the query.

Uniform Distribution

The actual waiting time is equal to the sum of the half square of observed headways, divided by the sum of observed headways. The scheduled waiting time is the sum of the half square of scheduled headways divided by the sum of scheduled headways. The excess waiting time is equivalent to actual waiting time minus scheduled

waiting time, which results in the amount of time riders had to wait in excess of the scheduled waiting time. The formulas are as follows:

$$AWT = \frac{\sum_i H_i \cdot \frac{H_i}{2}}{\sum_i H_i} \quad (1)$$

$$SWT = \frac{\sum_i T_i \cdot \frac{T_i}{2}}{\sum_i T_i} \quad (2)$$

$$EWT = AWT - SWT \quad (3)$$

where AWT, SWT, EWT are the actual, scheduled, and excess waiting times, respectively; H_i, T_i are the actual and scheduled headways of the ith bus trip in the selected time period.

Non-uniform Distribution

It is also possible to calculate the wait time using a non-uniform distribution such as a skewed distribution, which internalizes the fact that people do not arrive randomly for bus stops with longer headways, but instead try to time their arrival with that of the bus. According to Luethi et al. (2007), a shifted Johnson S_B distribution is suitable to depict the skewed passenger arrival distribution. The calculations for actual waiting time and excess waiting time for a non-uniform passenger arrival distribution case are generalized from the uniform case. The formulas are as follows:

$$AWT = \frac{\sum_i \int_{T_{i-1,d}}^{T_{i,d}} f(t) \cdot |t - T_{i-1,d}| dt}{\sum_i \int_{T_{i-1,d}}^{T_{i,d}} f(t) dt} \quad (4)$$

$$SWT = \frac{\sum_i \int_{T_{i-1,s}}^{T_{i,s}} f(t) \cdot |t - T_{i-1,s}| dt}{\sum_i \int_{T_{i-1,s}}^{T_{i,s}} f(t) dt} \quad (5)$$

$$EWT = AWT - SWT \quad (3)$$

where AWT, SWT, EWT are the actual, scheduled, and excess waiting times, respectively; $T_{i-1,d}, T_{i,d}$ are the actual departure times of the (i–1)th (preceding) and ith bus, respectively; $T_{i-1,s}, T_{i,s}$ are the scheduled departure times of the (i–1)th (preceding) and ith bus, respectively; and $f(t)$ is the passenger arrival distribution (e.g., a shifted Johnson S_B distribution).

When users query for passenger waiting time, they can choose to query by certain bus routes or by certain bus stops. Figure 2 shows the query by route view (a), in which users select a specific bus route and one or multiple bus stops along that route, and by stop view (b), in which users select a specific bus stop and one or multiple bus routes passing that stop. Users select route, direction, date and time range, and passenger arrival distribution from the left panel, and bus stops of interest from the map. Bar graphs under the map display the waiting time results separately for weekdays, Saturdays, and Sundays (cyan for scheduled, orange for actual, and brown for excess waiting time).

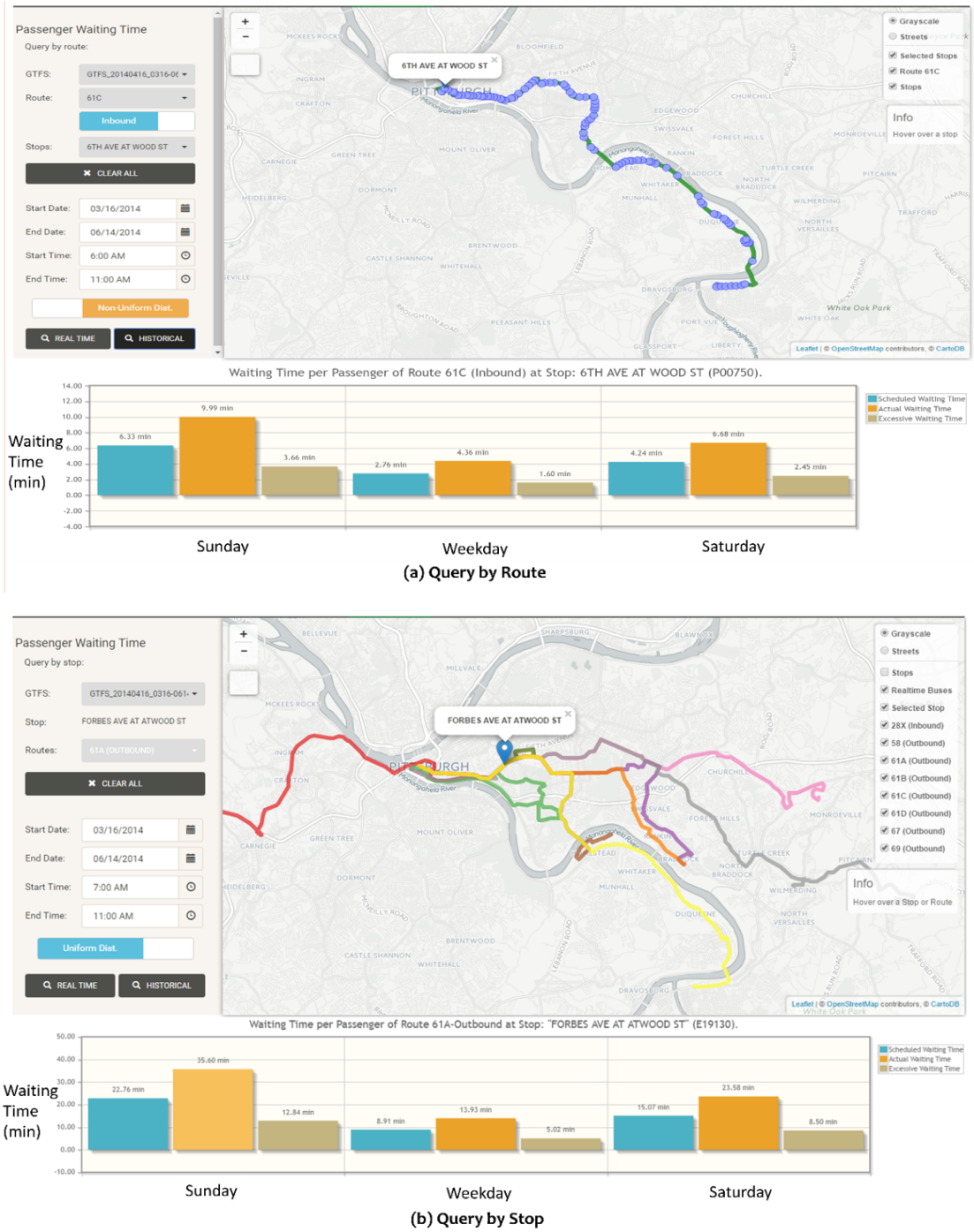


FIGURE 2.

Query by route and query by stop views

Bus Bunching Level

Bus bunching occurs when buses along a route deviate from their scheduled headways and begin to travel closer together forming "bunches." When one transit vehicle follows another at an initial headway, increased delays for the lead vehicle reduce delays for the trailing vehicle, as the lead vehicle services larger-than-average passenger volumes and the trailing vehicle serves smaller-than-average passenger volumes (National Academies 2013). This process typically continues until the trailing vehicle reaches the lead vehicle, after which the vehicles may travel together in tandem as a platoon.

These deviations can be caused by a multitude of factors including traffic congestion, traffic signal timing, traffic incidents, and excess crowding. Bunching decreases the reliability of a bus service, a characteristic that is highly valued by transit passengers and instrumental in decisions regarding modal choice (Strathman et al. 2003). Additionally, since bunching results in an inefficient allocation of rolling stock, it forces operators to purchase additional buses to prevent buses from becoming overloaded with passengers (Strathman et al. 2003). While some level of headway deviation is unavoidable in the unpredictable environment within which bus systems operate, it is possible to set a lower bound on acceptable headway between buses (Feng and Figliozzi 2011). This value can then be used to identify buses that have deviated far enough from their scheduled headways to form a bunch. These calculations will be performed using the archived AVL-APC data and a user-defined headway threshold (e.g., 0.5 times the scheduled headway). Any pair of buses with a headway below the user-defined threshold would have experienced bunching during the defined times. The number of trips identified as part of a bunch will be used to compute a percentage representing the "bunching level" along the route segment in question, which will be returned to the user.

Stop-Skipping Frequency

Stop-skipping happens on two major occasions. The first occasion is when there are no passengers waiting at the bus stop and no passengers wanting to disembark the bus. The second occasion is when a completely full bus skips some stops regardless of whether there are passengers waiting at those stops. Only the second case matters because it prevents passengers from boarding, which leads to extra passenger waiting time and a deterioration of the bus system's service quality. The difficulty in computing the frequency of stop-skipping with a full bus load is that AVL-APC data do not contain information regarding whether there are passengers waiting at a bus stop if the bus does not make a stop. As an approximate solution, our platform filters out the bus trips with full bus loads and zero dwell time at stops, considering all of them as stop-skipping because they all have the potential to prohibit passenger boarding.

Crowding Level/Fullness

Crowding level is a vital approximation of the service quality provided by a bus operator from both the passenger and operator perspectives. The immediate physical inconveniences of crowding experienced by passengers, such as standing for long periods, directly affect customer perception of service quality and significantly affect their decisions about using the service again (National Academies 2013). Outside of wanting to please and retain their rider base, transit operators have a strong incentive to reduce crowding because overcrowded buses have longer dwell times, which leads to decreased service regularity and throughput.

Typically, passenger load levels are measured and then compared against a standard, which the agency considers acceptable from its operational point-of-view as well as customer perception. These standards can be defined relative to the seated capacity of the vehicle in question (e.g., 125% of seated capacity) (National

Academies 2013). The agency defines acceptable levels of crowding, which can vary widely from agency to agency (Li and Hensher 2013). Our platform allows users to analyze crowding levels using the number of passengers as a percentage of seated capacity. It also allows users to define their own thresholds for overcrowding.

For a specific route and a selected date period with D days, each day has M_i (i=1,2,…,D) trips on that route in the selected time interval, and N stops on that route are queried. The average crowding level is computed as:

$$\mathrm{ACL} = \frac{\sum_{i=1}^{D}\sum_{j=1}^{M_i}\sum_{k=1}^{N} C_{i,j,k}}{\sum_{i=1}^{D} M_i \cdot N} \qquad (6)$$

where $C_{i,j,k}$ denotes the number of passengers as a percentage of seated capacity of the jth bus trip in the selected time intervals, at the kth stop of the selected route, on the ith day of the selected date range.

When users query for bus bunching, stop-skipping, or crowding level, they select route, direction, start and end bus stops, date and time range, and finally customize the legend from the left panel. When querying for bunching, users also specify their own criterion for bunching. The result is presented using a heat map with detail pop-ups for any selected stops along the queried route. For bunching or crowding, the pop-ups contain average levels. For stop-skipping, the pop-ups contain detailed information regarding stop-skipping at the selected bus stop, including the total number of trips that skip the stop by trip ID and the percentage of stop-skipping trips. An example is shown in Figure 3. For all three functions, users select a specific route segment and the date and time range. Results are then shown on the map as different colors for each metric. The threshold for colors can be customized by users. Users can also click on stops along the route for pop-ups detailing information on (a) bunching level—the percentage of bus trips that are considered as bunching; (b) stop-skipping—frequency and trip IDs that skip stops; and (c) crowding level—average bus fullness level.

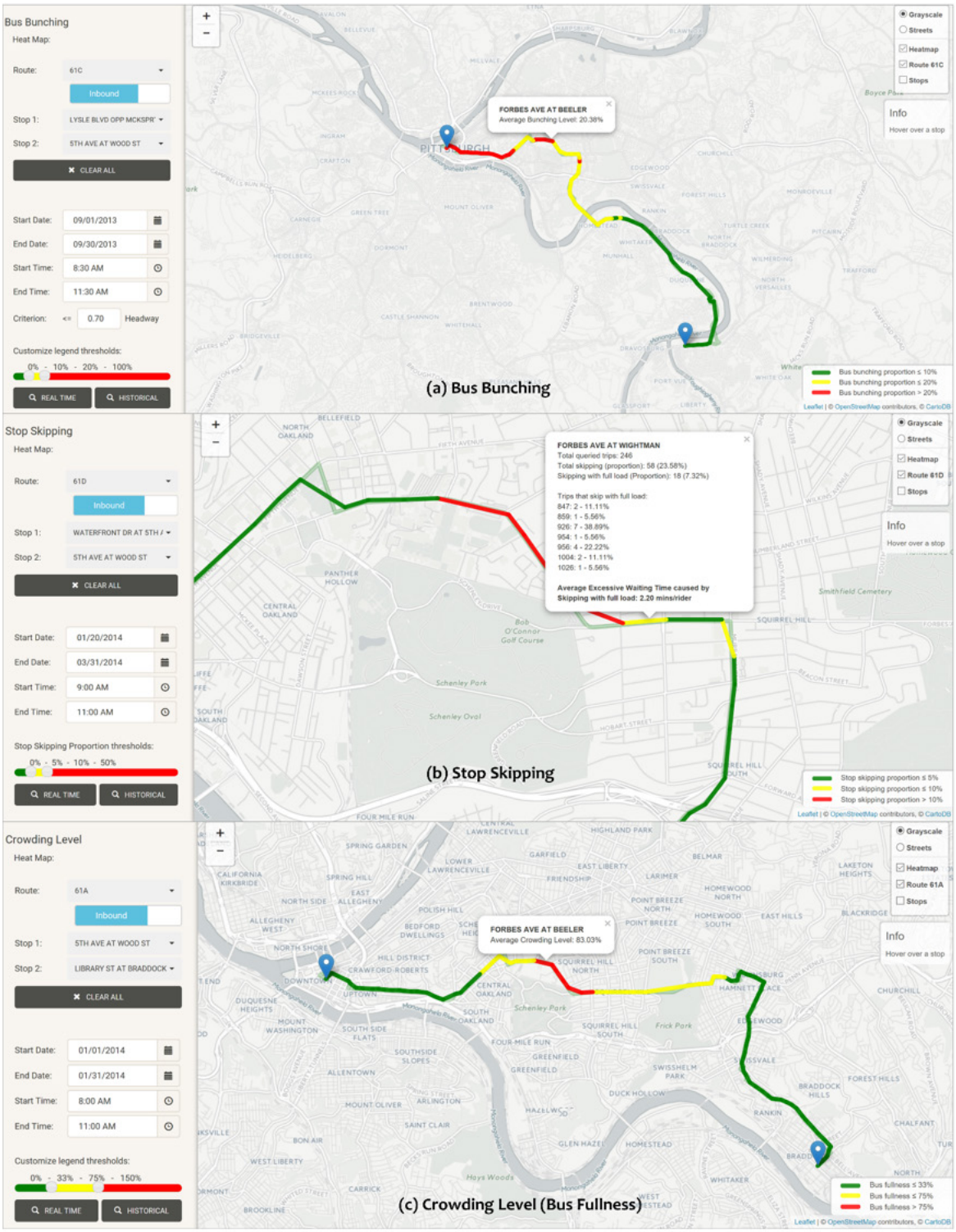


FIGURE 3.

Bus bunching, stop-skipping, and crowding level views

Travel Time

The length and variability of a trip's travel time are two of the major considerations when deciding whether to make a trip via public transit or automobile. Also, transit agencies and researchers can use historical travel times and their deviations from scheduled ones to assess a system's ability to provide the service it advertises and identify problem areas. Our platform will report the average travel time and two measures of variability: the standard deviation of a route's observed travel times across user-selected times and stops and the 90% travel time confidence interval of the distribution of these times (Mazloumi et al. 2009). The platform also produces a chart depicting the distribution of the observed travel times for the stops and times in question to illustrate any skews or abnormalities associated with the selected travel times.

On-time Performance

On-time performance is the most widely used transit reliability measure in North America (Sen et al. 2011). Our web platform uses the commonly accepted definition of on-time performance, which is simply "the percent of schedule deviations that fall within a defined range (e.g., 1 min early to 5 min late)" (National Academies 2013). It can be calculated using the bus schedule and actual bus departure time at stops, which are recorded by the AVL devices. On-time performance is relatively straightforward and is one of the most useful performance metrics for general transit riders since it answers the eternal question: "What are the chances my bus will be on time?" The definition of *on-time*, however, can vary from person to person depending upon the context in which the question is asked. To remedy this issue, users can input a chosen time range to define how early or late a bus can be before being considered not on-time, for example from one minute early to five minutes late (Bates 1986). This flexibility will benefit researchers by allowing them to adjust their definitions of on-time to match varying research questions and to allow for interagency comparisons. The query will return a schedule deviation distribution bar plot with the deviation mean and standard deviation (most distributions are approximately normal with, however, a long tail on the late arrivals) and the percentage indicating the number of trips within the defined on-time range.

Case Studies

Case 1: Bus Bunching in Pittsburgh

The transit data analytics platform is used to analyze AVL-APC data to investigate the geo-temporal occurrence of bus bunching in Pittsburgh. This analysis compares the incidence of bunching between two high-ridership routes with different operational characteristics. Route P1 is a bus rapid transit (BRT) route that operates on a dedicated busway. Route 61C is a "key corridor" route that operates in mixed traffic connecting several of Pittsburgh's business districts, including the central business district (CBD) and Oakland. The key corridor is served by Routes 61A, 61B, 61C, and 61D, which operate along the same path within the key corridor (the CBD to Squirrel Hill), and then branch to serve separate destinations. Bus bunching is frequently observed along the key corridor served by Routes 61. Table 2 summarizes several operational aspects of key corridor Route 61C, the combined routes of 61A, 61B, 61C, and 61D, and BRT Route P1.

TABLE 2.

Operational Aspects of Studied Bus Routes

Component/Route	61C	61A, 61B, 61C, 61D (Combined)	P1 (BRT)
Average Weekday Ridership	6,201	20,193	12,850
Peak Weekday Headway	15 min	4 min	4 min
Average On-time Performance	64%	67%	80%

Using AVL-APC data available through the analytics platform, hourly heat maps of bunching incidence for weekdays in the period of March 1 – March 31, 2016, were created for both Route 61C outbound and Route P1 outbound for the period from 12:00 to 10:00 p.m. By studying the time period from 12:00 to 10:00 p.m., we include the noon-peak hours, off-peak hours, and evening-peak hours for the two outbound (i.e., from downtown to suburbs) routes. A side-by-side comparison of bunching incidence is shown in Figures 4 and 5. To control for the difference in headway between 61C (15-minute peak and off-peak daytime headway) and P1 (4-minute peak headway, 6-minute off-peak daytime headway), Route 61C reports observations as bunched if operating with less than two-thirds of the scheduled headway (namely < 10 minutes), and Route P1 reports observations as bunched if operating with less than one-third of the scheduled headway (namely < 1 minute 20 seconds in peak hours, or < 2 minutes in off-peak hours).

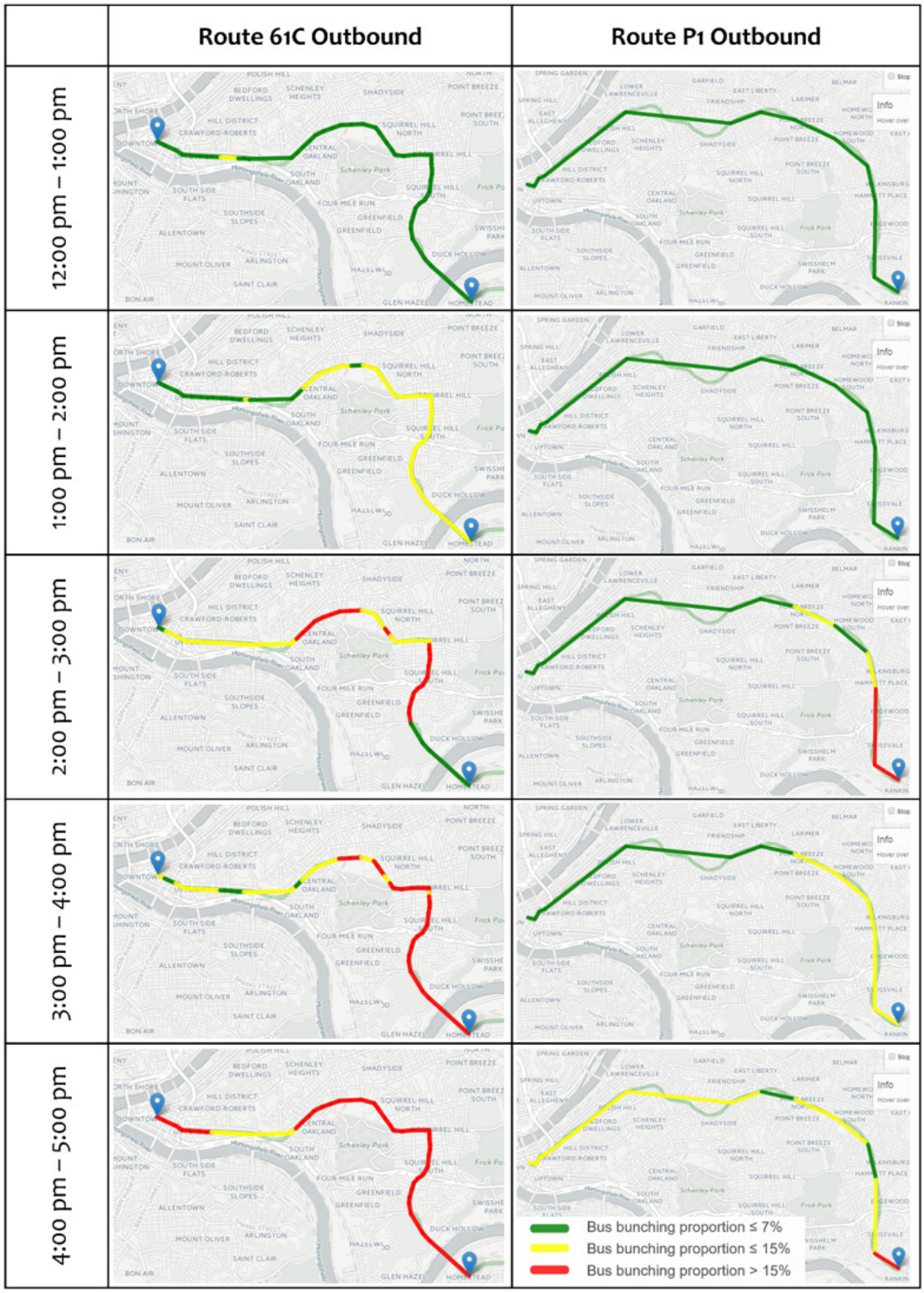


FIGURE 4.

Hourly heat maps (12-5 p.m.) of bunching incidence in March 2016 for Routes 61C and P1

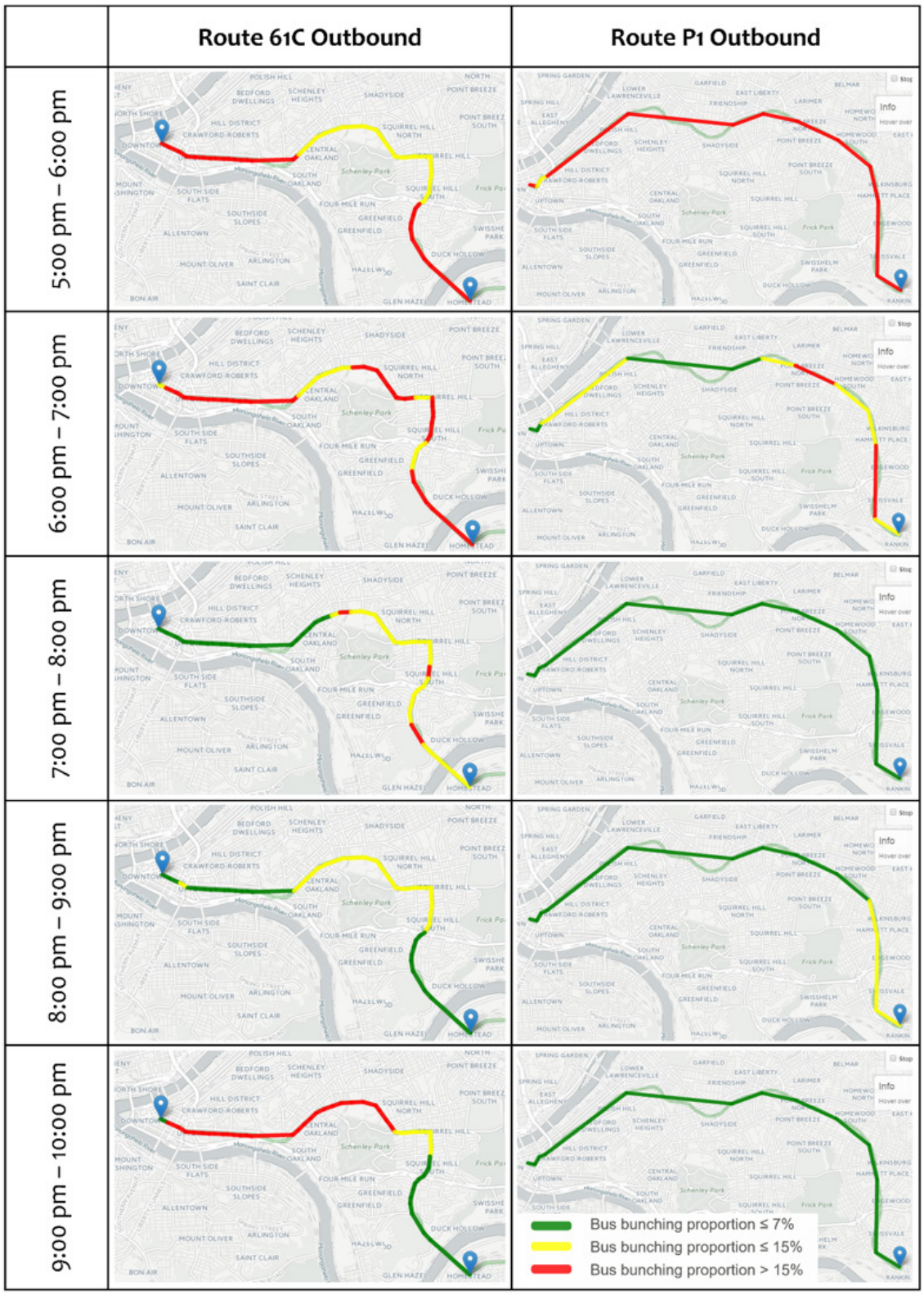


FIGURE 5.

Hourly heat maps (5-10 p.m.) of bunching incidence in March 2016 for Routes 61C and P1

The heat maps in Figures 4 and 5 show that key corridor Route 61C exhibits significant bunching throughout the afternoons. BRT Route P1, by contrast, shows significant bunching only during the 5:00–6:00 p.m. peak and little bunching otherwise. Figure 6 is an alternative chart representation of the incidence of bunching (for Route 61ABCD combined and Route P1). Each row corresponds to 1,000 feet of distance traveled along the route, with neighborhood boundaries marked on the vertical axis. The color of each cell represents the portion of trips that are bunched (pure red = incidence > 80%). Observations summarized in Figure 6 are characterized as bunched when observed within 820.2 feet (i.e., 250 meters) of another bus operating the same route and direction.
The chart view of the incidence of bunching suggests that, regardless of time of day (vertical red bands), some locations (horizontal red bands) are more likely to have bunched trips than others. Route 61 exhibits significant bunching in the downtown, Oakland, and South Squirrel Hill geographies. Route P1, also as shown in Figures 4 and 5, exhibits much less bunching than Route 61 (observations of bunching for Route P1 in Downtown and Swissvale are attributable to buses in close proximity at the beginning and ends of their routes, not headway deviation caused by unscheduled delays).

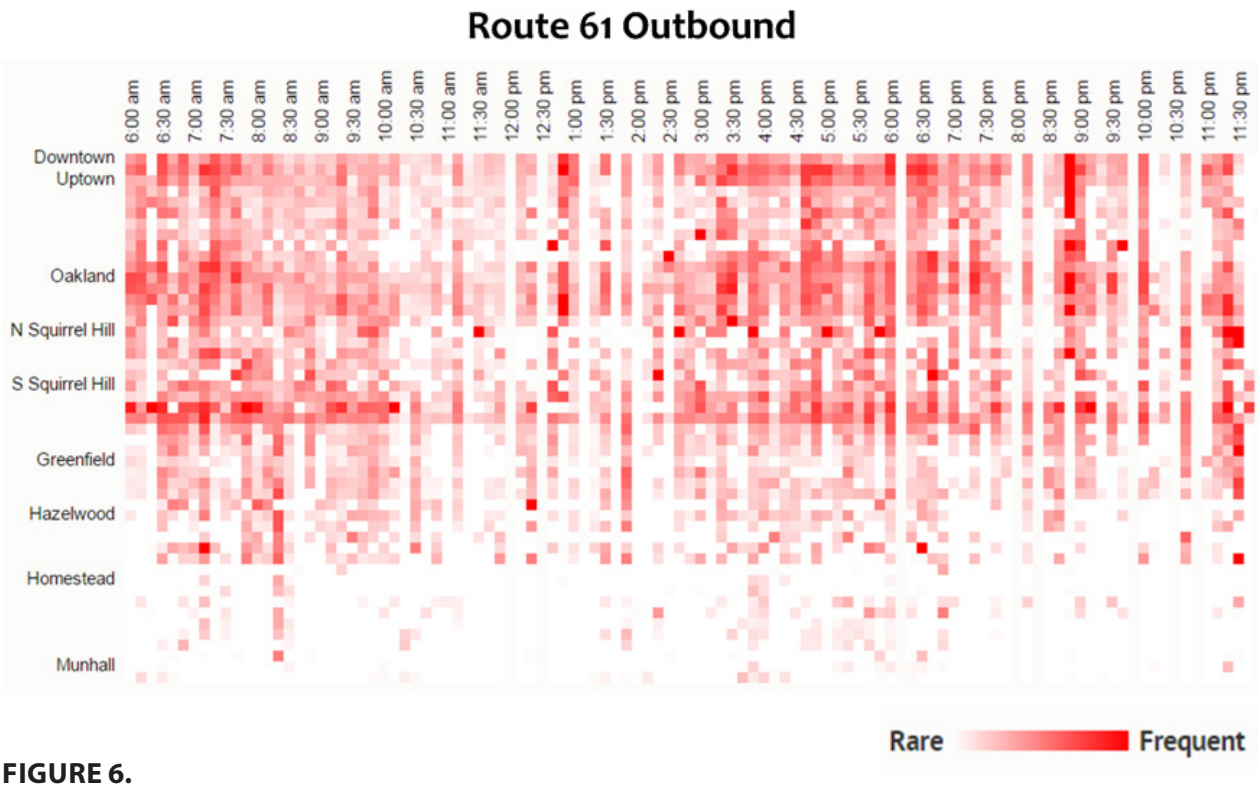


FIGURE 6.
Chart view of bunching incidence for Routes 61 and P1

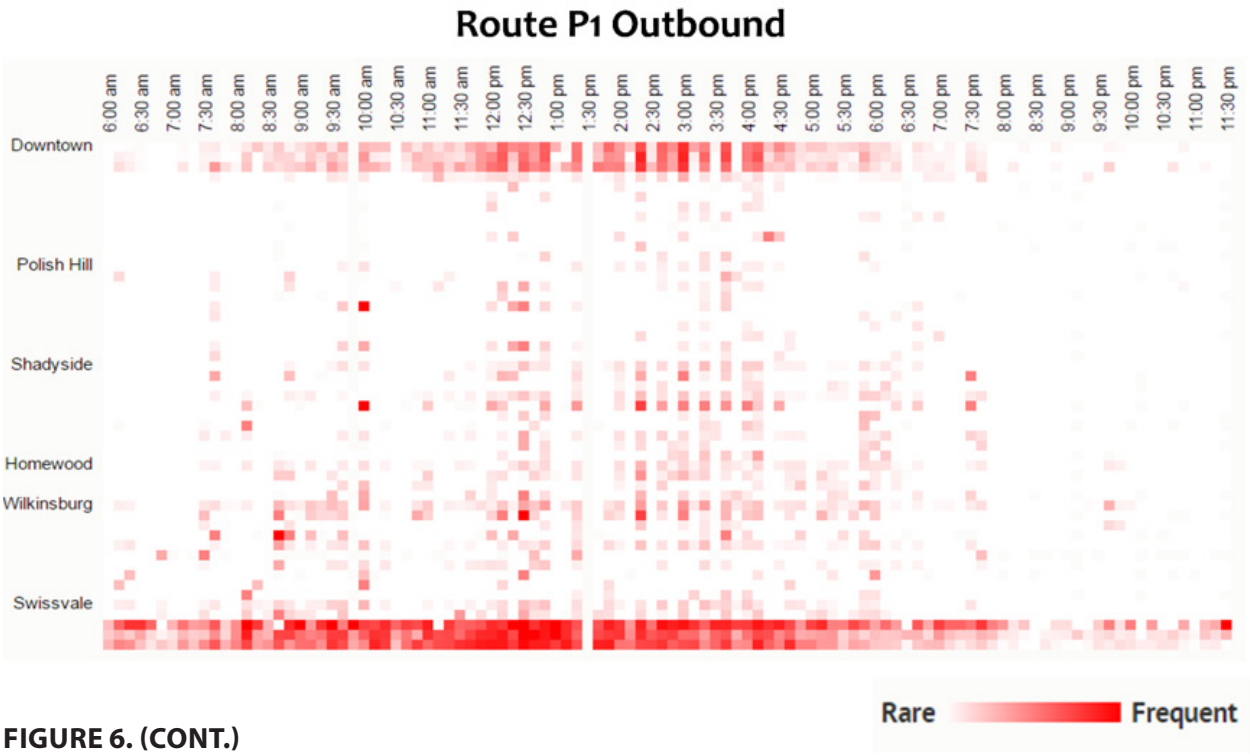


FIGURE 6. (CONT.)

Chart view of bunching incidence for Routes 61 and P1

Comparing the key corridor Route 61C with BRT Route P1 separates the headway deviation caused by peak passenger demand (observed between 5:00 and 6:00 p.m.) from other causes of delay (traffic congestion, signal delay). The frequency of bunching along Route 61C outside of peak demand suggests that solutions to bunching in the key corridor will need to address the delay associated with operating in mixed traffic. Everything else being equal, intuition would suggest that a route with higher ridership and lower headway would exhibit more bunching. However, comparison between Routes 61C and P1 shows the opposite: Route 61C, as a route with lower ridership and higher headway compared to Route P1, exhibits more severe bunching. The reason is because Route 61C operates in mixed traffic, whereas P1 operates as BRT along a dedicated busway. In addition to passenger delay experienced by both routes, Route 61C experiences signal delay, interference with on-street traffic, and lower maximum operating speeds.

What can be done to reduce bunching and increase the reliability of transit on Routes 61? Prior research suggests that adding service will not increase reliability, while worsening resource utilization. When bunching is prevalent, excess waiting time is not resolved by adding additional vehicles, as additional vehicles simply aggregate to platoons (Gershenson and Pineda 2009). Possible solutions to reduce the frequency of bunching for key corridor bus routes such as Route 61 in Pittsburgh include transit signal priority (National Academies 2013), scheduling recovery time between inbound and outbound trips (delay at the point of origin tends to persist across a route), incorporating operator experience and temporal variation in passenger demand to minimize bunching (Strathman et al. 2003), or the creation of a new BRT corridor connecting downtown (CBD), university districts (Oakland), and large residential districts (Squirrel Hill).

Case 2: Schedule Change Analysis for Route 61B

To further demonstrate the practical usage of this data analytics platform, we perform a comprehensive analysis on Route 61B inbound, which travels from Braddock through Oakland to downtown Pittsburgh. Route 61B is a representative bus route serving demand along a busy corridor and carrying a substantial number of riders. Passenger waiting time, bunching level, crowding level, schedule deviation, and stop-skipping under two different bus schedules are studied and compared. Suggestions for scheduling Route 61B inbound are also proposed to improve its service quality based on the performance metrics results.

On November 22, 2015, PAAC published a new version of GTFS in which the schedule of bus Route 61B inbound changed substantially. Specifically, the total number of trips remains the same, but the departure time of all trips on weekdays at the initial bus stop is scheduled seven to nine minutes early. We use the AVL-APC data from Route 61B inbound for two weeks before the change and two weeks after the change to examine the influence of the schedule change on that route on weekdays. The AVL-APC data used before the schedule change cover November 9–20, 2015 (except weekends), and the data after the schedule change cover November 30–December 11, 2015 (except weekends). These four weeks span from November to December, so we can almost rule out the seasonal effects on bus performance, such as the average temperature and daytime length. We do not use the data between November 23 and 29 because it is believed that commuters need time to get used to the new bus schedule and that span covers the Thanksgiving holiday. Since Route 61B inbound travels from suburb to downtown, we mainly study the morning peak hours from 6:00 to 9:00 a.m. In this time period, the trip frequency is about one trip every 15 minutes.

The average passenger waiting time and average absolute schedule deviation for two different bus schedules of Route 61B inbound are presented in Figure 7. Waiting time is calculated using the non-uniform passenger arrival distribution. The average is taken on nine main bus stops along the route that are selected as the "time-point" stops. The results show the new schedule leads to approximately 20% less average absolute schedule deviation, and approximately 10% less average passenger waiting time. Given that the trip frequency is the same, 15 minutes within the queried time range, before and after the schedule change, this schedule change effectively improves both waiting time and on-time performance.

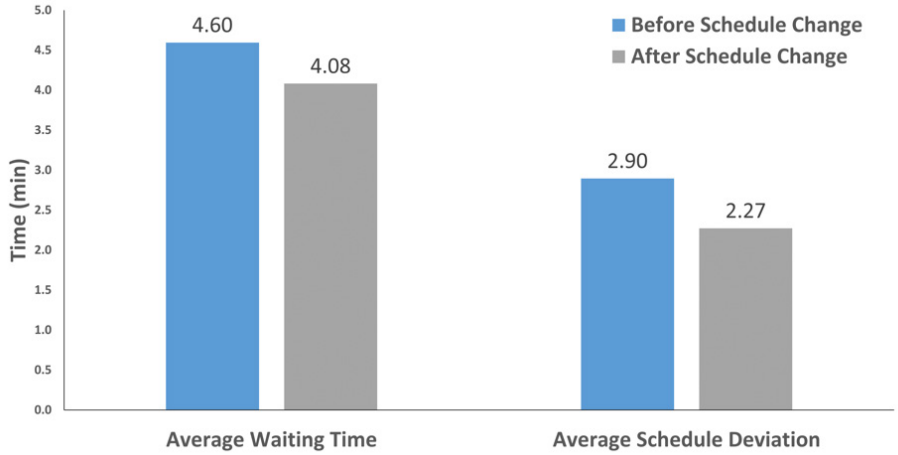
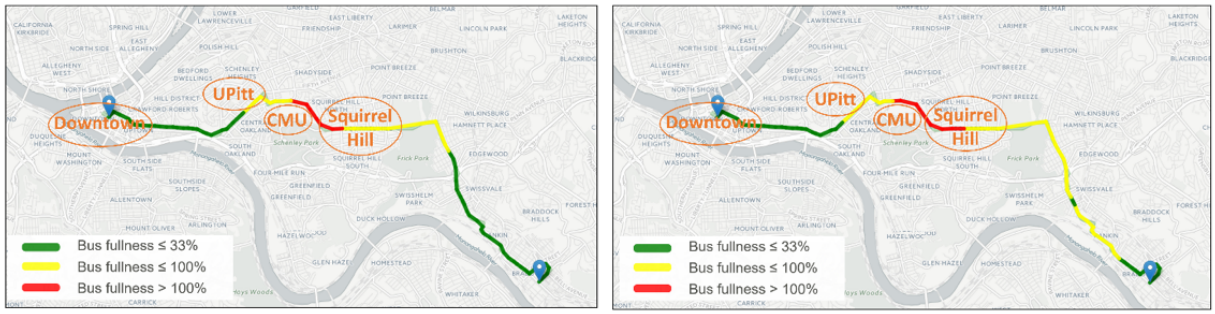


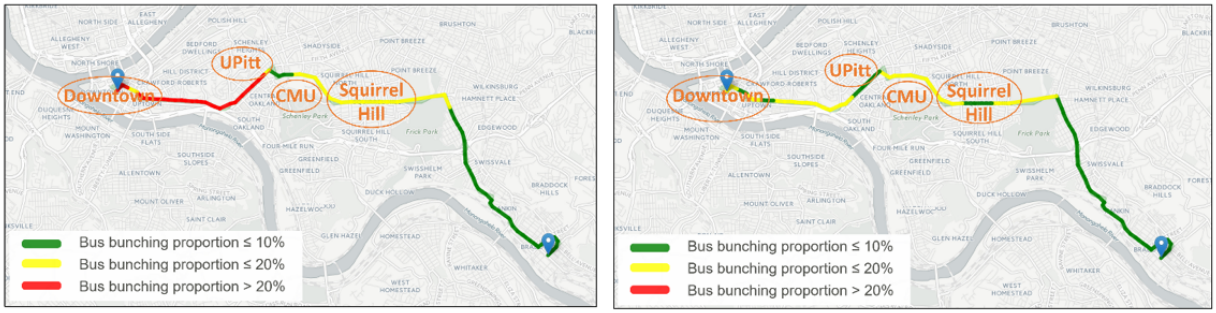
FIGURE 7.

Average passenger waiting time and average schedule deviation for two different bus schedules of Route 61B

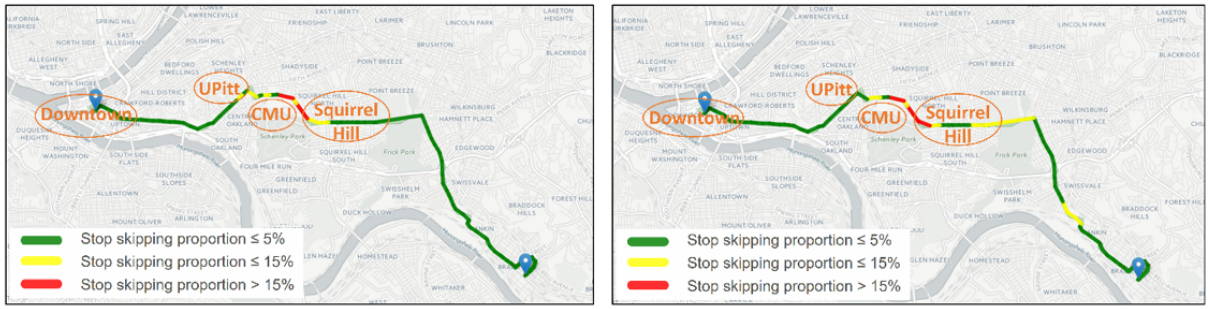
The heat maps of bus crowding level, bus bunching level (bus trip percentage considered as bunching), and the proportion of stop-skipping trips with full load for two different bus schedules of Route 61B inbound are also presented in Figure 8, with the identical color legend for the same function. According to the heat map results, PAAC could easily identify the spatial distribution of overcrowded bus, bus bunching, and stop-skipping. For example, Figure 8(b) shows the bunching problem was relieved after the schedule change, especially in the areas close to downtown, which also matches the result of decreased schedule deviation in Figure 7. However, the bunching still takes place sometimes near the Carnegie Mellon University (CMU), University of Pittsburgh (UPitt), and downtown Pittsburgh, which requires additional operational strategies to alleviate. Figure 8(a) and (c) also reveal a higher crowding level of bus trips in morning peak hours after the schedule change, and hence more stop-skipping occurrences, especially in the CMU-UPitt area and the residential areas ahead of the universities along the bus route (such as Squirrel Hill). PAAC may need to replace the buses running in morning peak hours with larger capacity models or decrease the trip headways if possible.



(a) Bus crowding level



(b) Bus bunching level



(c) Proportion of stop-skipping trips with full load

FIGURE 8.

Heat maps of (a) bus crowding level, (b) bus bunching level, and (c) proportion of stop-skipping trips with full load before (left) and after (right) schedule change

Conclusions

This paper introduces a novel transit data analytics platform for assessing the service quality of public transit systems based on APC and AVL systems. The platform offers a systematic way for users and decision makers to examine the system performance from many aspects of service quality, including passenger waiting time, stop-skipping frequency, bus bunching level, bus travel time, on-time performance, and bus fullness. We archive the AVL-APC data from September 2012 to March 2016 in a database and develop a user-friendly web application that allows users to interactively query bus performance metrics for any bus routes, bus stops, or bus trips for any time period. After data processing and computation, the query results are presented to users in a convenient and easy way to understand the system performance in high spatiotemporal resolutions.

In the case studies, we use the platform to discover the causes, patterns, and possible solutions to the bus bunching problem along several representative routes in Pittsburgh. It is found that bus bunching incidence is heavily impacted by both the location on route as well as the time of day, and the bunching problem is more severe for bus routes operating in mixed traffic than for BRT, which operates along a dedicated busway. Possible solutions to reduce the frequency of bunching for key corridor bus routes, such as Route 61 in Pittsburgh, include transit signal priority, scheduling recovery time between inbound and outbound trips, or the creation of a new bus rapid transit corridor connecting downtown (CBD), university districts (Oakland), and big residential districts (Squirrel Hill). Furthermore, a comprehensive impact analysis of schedule change on a representative bus route in Pittsburgh is performed using this analytics platform. Suggestions for operation of this route are also proposed to improve its service quality based on the performance metrics results before and after the schedule change.

In future work we plan to focus on three aspects. First, real-time access to AVL-APC data will help identify non-recurrent causes of bunching to detect and resolve bunching issues. Currently we only have the archived AVL-APC data in the backend of our web platform. In our future work, we hope to incorporate real-time AVL-APC data into our models to predict and manage bus bunching. Second, each of the current metrics currently works independently. We plan to combine all metrics (e.g., stop-skipping, waiting time, and bus bunching) and provide a holistic method to evaluate and optimize the transit systems. And third, additional data sources other than the AVL-APC data could be incorporated into our models and the platform. Some potential data sources include customer satisfaction data as well as demographic, weather, and economic data. These sources could be further mined to learn behavioral insights in conjunction with the AVL-APC data.

Acknowledgements

This research is funded through the Pennsylvania Infrastructure Technology Alliance. The authors would like to thank the Port Authority of Allegheny County for providing AVL-APC data. This web application was built with the help of Zhangning Hu, a former master student with the Heinz College, Carnegie Mellon University.

References

Anwar, A., A. Odoni, and N. Toh. 2016. "BusViz: Big Data for Bus Fleets." *Transportation Research Record: Journal of the Transportation Research Board* 2544: 102-109. doi: 10.3141/2544-12.

Badami, M. G., and M. Haider. 2007. "An Analysis of Public Bus Transit Performance in Indian Cities." *Transportation Research Part A: Policy and Practice* 41 (10): 961-981.

Barrero, R., J. Van Mierlo, and X. Tackoen. 2008. "Energy Savings in Public Transport." *IEEE Vehicular Technology Magazine* 3 (3): 26-36.

Bates, J. W. 1986. "Definition of Practices for Bus Transit On-time Performance: Preliminary Study." *Transportation research circular 300.* Washington, DC: Transportation Research Board, National Research Council.

Bates, J., J. Polak, P. Jones, and A. Cook. 2001. "The valuation of reliability for personal travel." *Transportation Research Part E* 37 (2-3): 191-229. doi: 10.1016/S1366-5545(00)00011-9.

Cevallos, F., X. Wang, Z. Chen, and A. Gan. 2011. "Using AVL Data to Improve Transit On-Time Performance." *Journal of Public Transportation* 14 (3): 21-40.

Cham, L. C. 2006. "Understanding Bus Service Reliability: A Practical Framework Using AVL/APC Data." Doctoral dissertation, Massachusetts Institute of Technology.

Eboli, L., and G. Mazzulla. 2009. "A New Customer Satisfaction Index for Evaluating Transit Service Quality." *Journal of Public Transportation* 12 (3): 21-37.

Feng, W., and M. Figliozzi. 2011. "Using Archived AVL/APC Bus Data to Identify Spatial-Temporal Causes of Bus Bunching." Paper presented at the Institute of Transportation Engineers (ITE) Western States Annual Meeting, Anchorage, AK.

Furth, P. G., B. Hemily, T. H. Muller, and J. G. Strathman. 2006. *Using Archived AVL-APC Data to Improve Transit Performance and Management.* Transit Cooperative Research Program (TCRP) Report 113. Washington, DC: Transportation Research Board.

Gershenson, C., and L. A. Pineda. 2009. "Why Does Public Transport Not Arrive on Time? The Pervasiveness of Equal Headway Instability." *PLoS ONE* 4 (10): e7292. doi: 10.1371/journal.pone.0007292.

Jaiswal, S., J. M. Bunker, and L. Ferreira. 2008. *Relating bus dwell time and platform crowding at a busway station.* Paper presented at 31st Australasian Transport Research Forum (ATRF), Gold Coast, QLD, Australia.

Khazaei, H., S. Zareian, R. Veleda, and M. Litoiu. 2015. "Sipresk: A Big Data Analytic Platform for Smart Transportation." In Leon-Garcia, A. et al. (eds) *Smart City 360°*, 166: 419-430. Springer International Publishing. doi: 10.1007/978-3-319-33681-7_35.

Kimpel, T., J. Strathman, D. Griffin, S. Callas, and R. Gerhart. 2003. "Automatic Passenger Counter Evaluation: Implications for National Transit Database Reporting." T*ransportation Research Record: Journal of the Transportation Research Board* 1835: 93-100.

Kimpel, T. J., J. G. Strathman, K. J. Dueker, D. Griffin, R. L. Gerhart, and K. Turner. 2000. *Time Point-Level Analysis of Passenger Demand and Transit Service Reliability*. Center for Urban Studies, Portland State University.

Li, Z., and D. A. Hensher. 2013. "Crowding in Public Transport: A Review of Objective and Subjective Measures." *Journal of Public Transportation* 16 (2): 107-134. doi: 10.5038/2375-0901.16.2.6.

Luethi, M., U. A. Weidmann, and A. Nash. 2007. *Passenger Arrival Rates at Public Transport Stations*. Paper No. 07-0635 presented at Transportation Research Board 86th Annual Meeting.

Mazloumi, E., G. Currie, and G. Rose. 2009. "Using GPS Data to Gain Insight into Public Transport Travel Time Variability." *Journal of Transportation Engineering* 136 (7): 623-631.

Mian, R., H. Ghanbari, S. Zareian, M. Shtern, and M. Litoiu. 2014. "A Data Platform for the Highway Traffic Data." *2014 IEEE 8th International Symposium on the Maintenance and Evolution of Service-Oriented and Cloud-Based Systems (MESOCA)* 47-52. doi:10.1109/MESOCA.2014.14.

Nathanail, E. 2008. "Measuring the Quality of Service for Passengers on the Hellenic Railways." *Transportation Research Part A: Policy and Practice* 42 (1): 48-66.

National Academies of Sciences, Engineering, and Medicine. 2013. *Transit Capacity and Quality of Service Manual, Third Edition*. Washington, DC: The National Academies Press. doi: 10.17226/24766.

Saberi, M., A. K. Zockaie, W. Feng, and A. El-Geneidy. 2013. "Definition and Properties of Alternative Bus Service Reliability Measures at the Stop Level." *Journal of Public Transportation* 16 (1): 97-122. doi: 10.5038/2375-0901.16.1.6.

Salek, M. D., and R. B. Machemehl. 1999. *Characterizing Bus Transit Passenger Wait Times*. Center for Transportation Research, University of Texas, Austin.

Sen, L., S. R. Majumdar, M. Highsmith, L. Cherrington, and C. Weatherby. 2011. *Performance Measures for Public Transit Mobility Management*. Texas Transportation Institute, The Texas A&M University System.

Shalaby, A., and A. Farhan. 2004. "Prediction Model of Bus Arrival and Departure Times Using AVL and APC Data." *Journal of Public Transportation* 7 (1): 41-61. doi: 10.5038/2375-0901.7.1.3.

Shtern, M., R. Mian, M. Litoiu, S. Zareian, H. Abdelgawad, and A. Tizghadam. 2014. *Towards a Multi-cluster Analytical Engine for Transportation Data*. Paper presented at 2014 International Conference on Cloud and Autonomic Computing (ICCAC). doi: 10.1109/ICCAC.2014.37.

Strathman, J. G. 2002. *Tri-Met's Experience with Automatic Passenger Counter and Automatic Vehicle Location Systems*. Center for Urban Studies, Portland State University.

Strathman, J. G., T. J. Kimpel, and S. Callas. 2003. *Headway Deviation Effects on Bus Passenger Loads: Analysis of Tri-Met's Archived AVL-APC Data*. Center for Urban Studies, Portland State University.

Tétreault, P. R., and A. M. El-Geneidy. 2010. "Estimating Bus Run Times for New Limited-Stop Service Using Archived AVL and APC Data." *Transportation Research Part A: Policy and Practice* 44 (6): 390-402.

Transport for London. 2014. "London Buses Quality of Service Indicators."

About the Authors

Xidong Pi *(xpi@andrew.cmu.edu)* is a third-year PhD student at the Department of Civil and Environmental Engineering at Carnegie Mellon University. His research interests lie in optimization and control in traffic networks and public transportation.

Mark Egge *(megge@andrew.cmu.edu)* is a second-year MS student of Information Systems at the H. John Heinz III College (Schools of Information Systems and Public Policy), Carnegie Mellon University. He is broadly interested in large-scale data analytics and specifically in the transportation and health care domains.

Jackson Whitmore *(jdwhitmo@andrew.cmu.edu)* is an associate planner of transportation at the Regional Plan Association in New York.

Amy Silbermann *(asilbermann@portauthority.org)* is a senior analyst in the Division of Service Planning and Evaluation at the Port Authority of Allegheny County.

Zhen (Sean) Qian, PhD *(seanqian@cmu.edu)* is an assistant professor jointly appointed at the Department of Civil and Environmental Engineering and H. John Heinz III College (Schools of Information Systems and Public Policy) at Carnegie Mellon University (CMU). He directs the Mobility Data Analytics Center (MAC) at CMU. The primary focus of his research is to manage aging and overcrowded transportation infrastructure systems, and to build sustainable and resilient infrastructure