Public Transit for Special Events: Ridership Prediction and Train Scheduling

Tejas Santanam, Anthony Trasatti, Pascal Van Hentenryck, *Member, IEEE*, and Hanyu Zhang

***Abstract*— Many special events, including sports games and concerts, often cause surges in demand and congestion for transit systems. Therefore, it is important for transit providers to understand their impact on disruptions, delays, and fare revenues. Ridership after large sporting events is distinct from many other ridership patterns due to the high density of ridership localized to a few nearby stations. This paper provides the following novel methodology for long-term planning of large-event, post-game ridership by 1) predicting the total post-game ridership; 2) combining the total prediction with historical trends to forecast the passenger flow curve at nearby stations after the game; and 3) estimating the required train frequencies to serve these customers with minimal passengers left behind by each train. Additionally, this paper proposes a suite of data-driven techniques that together create a data-driven pipeline to exploit Automated Fare Collection (AFC) data for evaluating, anticipating, and managing the performance of transit systems. This paper includes a case study where the proposed pipeline is used to generate an adjusted train schedule for the post-game period and simulated with the rail ridership data from the Metropolitan Atlanta Rapid Transit Authority (MARTA). The simulation results highlight how the proposed schedules based on the estimated required post-game train frequencies could significantly improve post-game congestion and wait time. Furthermore, the results show that the long-term post-game demand forecasts could be an effective tool for tactical planning decisions such as the number of additional trains and operators that are needed during post-game periods compared to the regularly scheduled timetables.**

***Index Terms*— Special events, machine learning, public transportation, smart cards, demand forecasting.**

I. Introduction

SPECIAL events, including sports games, concerts, and festivals, are important for transit providers; they often lead to fundamentally different ridership patterns and bring significant fare revenues. In addition, special events may be the introduction of certain riders to a transit system. Hence, it is critical to ensure that the system is smooth and efficient, the wait time is reasonable, and that the vehicles are not too crowded in order to attract additional recurring ridership.

Manuscript received 9 June 2021; revised 19 January 2022, 2 June 2022, 10 November 2022, and 21 July 2023; accepted 5 February 2024. This work was supported in part by NSF Leading Engineering for America's Prosperity, Health, and Infrastructure (LEAP-HI) under Award 1854684. The Associate Editor for this article was J. Xun. *(Corresponding author: Anthony Trasatti.)*

Tejas Santanam, Pascal Van Hentenryck, and Hanyu Zhang are with the Industrial and Systems Engineering Department, College of Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.

Anthony Trasatti was with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. He is now with Google, Mountain View, CA 94043 USA (e-mail: anthony.j.trasatti@gmail.com).

Digital Object Identifier 10.1109/TITS.2024.3373634

This paper originated as a study of special events for the Metropolitan Atlanta Rapid Transit Authority (MARTA), the transit system of the city of Atlanta in the state of Georgia. In particular, the study aims at addressing two main objectives of MARTA:

1) Is it possible to forecast special event rail ridership based on expected attendance and the type of event using historical Automated Fare Collection (AFC) data;
2) Can scheduling train frequencies based on forecasted ridership significantly improve passenger wait time and congestion following the events?

Ridership after large sporting events is distinct from many other ridership patterns due to the high density of ridership localized to a few nearby stations. Post-game demand surges can negatively impact wait time and congestion for both event and non-event riders. In the case of MARTA, large events, e.g., at the Mercedes-Benz Stadium in downtown Atlanta, can lead to demand that far exceeds the capacity of the usual weekday or weekend train schedules. Therefore, additional trains are required to increase the system frequency and capacity to help mitigate safety, congestion, and performance concerns.

From a high-level perspective, this paper provides a novel methodology for tactical planning of large-event, post-game ridership consisting of the following steps:

1) Obtain an attendance prediction from stadium ticket sales and predict ridership from that attendance using supervised machine learning;
2) Combine the total prediction with historical trends to forecast the passenger flow curve at nearby stations after the game;
3) Based on the forecasted arrivals, estimate required train frequencies to serve the demand with minimal passengers left behind by each train.

This paper includes a case study where the proposed methodology is used to estimate required post-game train frequencies for the post-game period. The proposed schedule is simulated with the rail ridership data from the Metropolitan Atlanta Rapid Transit Authority (MARTA). The simulation results highlight how the using the estimated required train frequencies for tactical planning could significantly improve post-game congestion and wait time. To our knowledge, there are no comparable methods for highly accurate forecasts of the post-game ridership curves.

This paper additionally makes the following technical and analytical contributions for leveraging Automated Fare Collection (AFC) data:

1) This paper proposes a technique that uses unsupervised machine learning models to cluster passengers and

analyze which trains they take, their waiting time, and the departure time of trains without accurate train schedule or train tracking data.

2) This paper presents simple supervised learning models to predict the total ridership for various types of recurring special events that can be used along with historical trends to forecast passenger flows at nearby stations after the game allowing for longer-term planning as opposed to real-time disruption planning.
3) This paper demonstrates a data-driven, end-to-end special event preparation pipeline for transit agencies with all components from ridership prediction to schedule adjustment.

The rest of the paper is organized as follows. Section II reviews prior work on similar topics. Section III presents the case study, analysis of the baseline ridership, and the mean passenger flow for entries and exits at rail stations on weekdays and weekends. It also analyzes the special event ridership and estimates how many riders use the rail to travel to and from the events and where they come from. Section IV first introduces the unsupervised learning to cluster riders by tap-out times to estimate train departures times and proportion of passengers left behind. It then presents supervised learning models to predict the ridership for various types of recurring special events, an algorithm for estimating the required train frequencies based on forecasted demand, and simulation techniques used to validate the methodology. Section V presents the validation results of the clustering method and the evaluation of the proposed schedules created based on the estimates of required train frequency using simulations. The simulation results demonstrate the potential benefits of the proposed methodology on post-game congestion and wait time.

II. Literature Review

The focus of this paper is on providing transit operators with valuable insights for tactical planning decisions surrounding special event ridership, especially for post-game demand surges for large events. Special events can lead to a high density of ridership late at night or on weekends when a transit agency might normally be running a less frequent schedule. To prepare for this, planners often need to schedule additional shifts for operators to handle these influxes of demand. The literature is structured as follows: Section II-A presents prior work around analyzing AFC data to characterize ridership for both event days and non-event days. Section II-B addresses other papers that focus on prediction of special event ridership. Finally, Section II-C presents works that focus on creating and adjusting train timetables.

A. Analysis With AFC Data

Automated Fare Collection (AFC) technologies have enabled more sophisticated analysis of transit ridership. This section focuses on related work that exploits AFC data for evaluating, anticipating, and managing the performance of transit systems.

Station-level analysis of AFC data has provided knowledge on ridership behavior and passenger flow as a result of various characteristics of the station and surrounding area. El Mahrsi et al. [1] use AFC data to extract mobility patterns for different types of passengers and stations, clustering stations based on their activity curves. Moradi and Trépanier [2] use smart card data to identify stable temporal habits of rail riders across the week. Ghaemi et al. [3] also use a clustering approach to analyze smart card data temporally. Shen et al. [4] use clustering methods to study the time-varying pattern of passenger flow at Shanghai rail stations. They find that the pattern of passenger flow is consistent during the work week, as well as weekends. Li et al. [5] use smart card station entry data and the day's train schedule at a station in Shanghai to cluster passengers and determine passenger route choice probability.

Significant work has been done for estimating the passengers left behind by high-frequency trains operating at capacity using AFC transaction data and Automated Vehicle Location (AVL) data. Zhu et al. [6] estimate the probability distribution of the number of left-behind passengers using entry-exit AFC data and train-tracking AVL data. Miller et al. [7] estimate the level of crowding at train stations in real-time using AFC data and train departure times. When historical or real-time AVL data is unavailable, the train departure times can often be estimated with the scheduled times. However, for special events, additional trains are sometimes added off-schedule. Hence there is a need to estimate the arrival and departures of the trains for downstream analysis of wait times and train loads.

This deviation from the schedule due to the variability of special events makes these estimations necessary in many cases. Tan et al. [8] propose a technique to derive train arrival times though clustering of passengers by tap-out time at each station then matching these clusters into trains. Hong et al. [9] focused on determining the connections of passengers by estimating the boarding and alighting time windows based on distributions of platform to gate times and the schedules. From here, they were able to reliably estimate the connections of passengers. They noted the that it was a non-trivial task to derive the time intervals solely by plotting the tap-in and tap-out times.

However, this paper applies an unsupervised learning technique that requires little tuning to reliably cluster passengers based on their train, after first applying a time-adjustment to account for varying destinations. Its effectiveness is demonstrated during peak periods such as post-game demand surges. Simulation is used to validate this method and to estimate metrics such as passenger load and wait time. This method can also be generalized to other entry-exit systems. In addition to being helpful when rails are running without an accurate schedule, the method can also be used to measure on-time performance and deviations from the published schedules.

B. Ridership Forecasting for Special Events

Various data sources have also been used to study special-event ridership including survey, AFC, and web data [10], [11]. Additionally, Pereira et al. [12] use average daily passengers flows to detect overcrowding hotspots and calculate an estimation of event ridership. Special events are an interesting challenge for transit providers because their time and place is known ahead of time, but there tends to be more uncertainty compared to commuting ridership.

Short-term prediction has also been the focus of other ridership models in [13] and [14], which contrasts to the longer time horizon considered in this paper. Noursalehi et al. [15] develop a general methodology for short-term ridership prediction within 15 minutes for planned special events like soccer games in London. Li et al. [16] predict riders at a station half an hour in the future using a one-step-ahead multiscale radial basis function network model for three large Beijing rail stations.

Ni et al. [17] predict ridership from the number of tweets for Mets games and US Open tennis matches using linear regression and established a correlation between tweets related to an event and event ridership flow. King [18] predicts NBA game attendance using random forest models. Karnberger et al. [19] analyze weekly system averages of the Munich public transit system, using AFC data to build a gradient-boosted random forest prediction system with 30-minute resolution for ridership between linked stations using information about type of day (holiday, weekend, etc.) and existence of a few types of events.

Rodrigues et al. [20] focus on the arrival curve of riders, grouped into 30 minute bins, for stations nearby to special events using a Bayesian additive model and Singapore smart card data and 30 minute bins. The Bayesian models allow for decomposing the activity into pieces contributed from habitual activity and each event.

In contrast, this paper focuses on predicting non-habitual demand surges, specifically those during post-game periods. Section III highlights how these congestion peaks can overwhelm the capacity of the standard schedule for MARTA. This method focuses on having an accurate prediction at the total ridership number and leverages consistency in the passenger flow curves after similar events to create a highly accurate (11-12% MAPE) and high resolution (5)-minute intervals) demand forecast. The results of this paper show how transit planners may be able to use the demand forecasts from these proposed methods to mitigate congestion and disruptions from demand surges that exceeded the capacity of the normal weekday or weekend schedules.

C. Train Scheduling

Traditionally, public transit schedules are operated on timetables with relatively consistent headways at each station [21], [22]. Zimmermann and Lindner [23] use a MILP-based approach for rail schedule optimization for railroads with repeating fixed-period schedules. Wang et al. [24] use a genetic algorithm based optimization approach through Hadoop to come up with an optimized rail schedule for a station in Beijing. Guo et al. [25] model the input passenger flow and then uses an adaptive large neighborhood search algorithm to generate an optimized rail timetable. Ceder et al. [26] determine timetables for buses with different capacity sizes using a simulation-based and optimization-based approach with predicted ridership flows to minimize congestion and wait times. Li et al. [27] optimize demand-oriented schedules with short turning and heterogeneous headways.

In the case of disruptions or special events, traditional timetables are usually not as effective. Li et al. [28] allow for non-cyclic timetables, which allow headways to vary in accordance with demand patterns such as special events. Niu et al. [29] have both online and offline scheduling models that can respond to variation of temporal and spatial demand distributions, exploring both buffer time assignment and train skip-stop patterns. Short-turning, skip-stop patterns, and buffer time are also techniques that can be used used to improve service quality for special events. Schettini et al. [30] present a MILP and heuristic for scheduling of a metro line serving a special event and conclude that minimizing passenger wait-time yields a good compromise between their four presented objectives.

Additional factors such as train running area, turnaround times, and operating costs are used in other studies to determine optimal train schedules. In determining ideal train schedules, Zhao et al. [31] consider train turnaround time and train coupling constraints while scheduling in such a way that minimizes operating cost. Wong et al. [32] minimize passenger wait times when scheduling trains based on manipulation of many factors, including headways and train turnaround times. Qi et al. [33] minimize total running distance of unoccupied seats and the total number of stops for all involved trains via skip-stop techniques that consider train running area, turnaround times, and operating cost. Wang et al. [34] consider various stop plans when constructing rail schedules that employ flexible pricing to maximize rail system revenue.

In contrast, this paper focuses on a tactical planning perspective. The simulation results showcase the benefits of utilizing the methods forecasting and estimating the required train frequencies to serve riders with minimal passengers left behind when planning a train schedule for post-game periods. During these periods, demand often far exceeds the supply generated by the normal timetable. To address this increased demand during a period of reduced schedule capacity, transit agencies often add additional trains to the normal timetable to help serve these short periods with high demand from special-event riders. Large special events often occur at night or on weekends when train schedule is not typically running at peak-level capacity. To our knowledge, there are no comparable methods that focus on this combination of forecasting and scheduling methods to more effectively address the demand of post-game ridership.

III. Ridership Analysis

Figure 18 depicts the four MARTA rail lines [35]. The Red and Gold lines run North-South and the Blue and Green lines run East-West, with the two directions intersecting at Five Points. The three event locations considered in this paper are noitemsep

1) The Mercedes-Benz Stadium;
2) The State Farm Arena;
3) The Georgia World Congress Center.

TABLE I
EVENT DATA EXAMPLE

Date	Category	Event	Location	Attendance
01/15/2018 15:00:00	Basketball	Hawks v San Antonio Spurs	State Farm Arena	15,000
01/16/2018 07:00:00	Conference	Mary Kay Leadership Conference	Georgia World Congress Center	7,000
01/16/2018 09:00:00	AmericasMart	Intnl Gift & Home Furnishings	AmericasMart	72,000

TABLE II
THE EVENT DATA OVERVIEW

Primary Event	# of Days	Avg. Attendance	Avg. Post-Game Ridership
Basketball - Hawks	67	15,278	1,425
Football Games	28	69,477	10,846
Soccer	39	52,712	8,037
Day Type	**# of Days**	**Avg. Attendance of Primary Event**	**Avg. Post-Game Ridership**
Single Event	78	38,062	5,487
Two Events	56	36,712	5,082

These three key venues located in downtown Atlanta on the Blue and Green lines are highlighted in the left center of the map. The two closest stops are Dome/GWCC and Vine City. Users coming from the North or South can use the Red or Gold line and transfer at Five Points to get on the Blue or Green line. These locations are important for the subsequent analyses.

A. Event Data

The event data provided by MARTA is a list of many public events in the Atlanta area during 2018 and 2019 containing types of events, locations, date, time, and estimated attendance. An example of event entries is presented in Table I.

The three most popular venues for large events in Atlanta are the Georgia World Congress Center, the Mercedes Benz Stadium (MBS), and the State Farm Arena, which were mentioned earlier as the focus of this paper. In 2018 and 2019, there were 1330 special events with an estimated attendance greater than 500 people in Atlanta. 706 of these 1330 special events were held in the Georgia World Congress Center, the Mercedes Benz Stadium (MBS), or the State Farm Arena. The three locations are geographically close to each other; moreover, the closest two rail stations are the Dome/GWCC and Vine City stations.

This paper focuses on events with the largest impact and occur frequently, in particular, the 200 sporting events (Basketball, Soccer, and Football games). All basketball games were held in the State Farm Arena, while the soccer and football games were held in the MBS. Apart from the sporting events, there were 74 conferences, 53 conventions, and 102 expos & shows at the target locations. However, the conference center events tends to have little impact on congestion as the ridership is smaller and more diffuse, making them less interesting for analysis. The event data is summarized in Table II.

B. Automated Fare Collection Data

To enter or exit the MARTA rail system, customers are required to use a ticket or a reloadable card (the "Breeze Card") at the gates of individual stations. MARTA provided anonymized transaction-level data showing tap-in and tap-out times and locations for the rail network from 2016 to mid-2020.

Trip chaining is performed to turn these individual transactions into Origin-Destination (OD) pairs. Tap-ins and tap-outs, when chained, are sufficient to determine where a rider enters and exits the rail network. For the most part, entries are matched to the following exit and create an OD pair. In some cases, the chained trip is an entry and exit at the same station back to back. For example, when an exit tap occurs, but there is a missing entry tap, then the system records a "forced entry" transaction and an exit transaction at the exit station at the same time. Similarly, if someone tries to re-enter a station, but there is a missing exit transaction or an extended amount of period has elapsed (3-4 hours), the system will add a "forced exit" transaction at the most recent tap-in location. Since the majority of transactions follow the expected pattern (pairs of distinct locations), the analysis focuses on these transactions.

This section first analyzes ridership on non-event days, as average day patterns help identify the effects of special events on the system. The analysis considers baseline days and looks at the mean passenger flow in and out of a station in 15-minute segments over the course of a day. These flows are then used to calculate the ridership that can be attributed to special events. The analysis also assesses the consistency of special-event ridership.

C. Mean Passenger Flows

To create mean passenger flow graphs for baseline days, the daily average passenger flow for each station is calculated from the transaction data and partitioned on the basis of (i) weekday versus weekend, and (ii) entry versus exit.

This partitioned data is used to create four baseline mean passenger flows for each station: weekday entry, weekday exit, weekend entry, and weekend exit.

Since the rail is closed between 1:30am and 4:30am, the day is defined as the 24-hour period starting at 3AM and finishing 3AM to capture passengers returning after midnight. Each day is further partitioned into 15 minute intervals and the number of passengers are counted for each interval for each of the four types of transactions. These flows are consistent within each type.

Figure 1 shows how the mean passenger flows are very similar and consistent month-to-month. This kind of consistency is good for future modeling and planning. In the appendix, Figures 19a–19d depict the four baseline mean passenger flows for the North Avenue and Midtown stations. The shaded region represents the 10th-90th percentile range for each bin. The mean passenger flows show regular commute spikes on weekdays for both entry and exit flows. They also show that passengers follow similar patterns throughout the year. These mean passenger flows are calculated for every day in 2018 and 2019 and then partitioned in the ways described above. Figures 19e–19f report the results for the Vine City station, which does not have many regular commuters. There is still a weekday and weekend difference for the ridership of Vine City station, but the overall magnitude of ridership is low compared to the

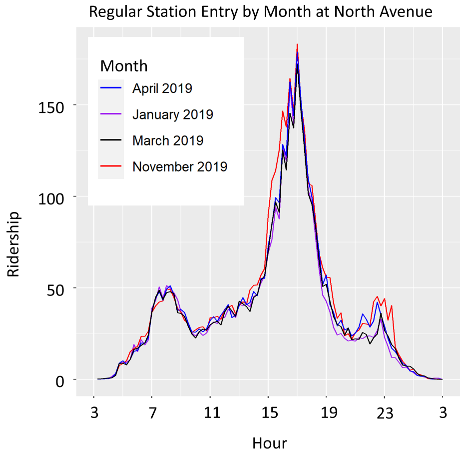


Fig. 1. Monthly baseline mean passenger entries per 15-minute bin for entries at North Ave rail station.

TABLE III

NOTATIONS USED IN THIS STUDY

Name	Descriptions
P	Set of passengers p
S	Set of stations s
I	Set of trains itineraries i
t	A point in time
τ	A time interval
M_t	Mean passenger flow at time t
U_t	90th percentile of the passenger flow distribution at time t
R_t	Actual ridership at time t
t_{start}	Earliest time with $E_t > M_t$ to capture the start of the event ridership
t_{end}	Latest time with $E_t > M_t$ to capture the end of the event ridership
E_t	Number of Rail Riders at time t who attended the event
$\mathbb{E}$	Number of Event Riders
$t^s_{\text{in},p}$	Passenger p's entry time into the rail system at station s
$t^s_{\delta,p}$	Passenger p's arrival time (or boarding time) by train at station s
$t^s_{\text{out,p}}$	Passenger p's exit time out of the rail system at station s
$\tau^{s,s'}$	Train travel time from station s to s', Positive if s is a upstream station of s', Negative other wise.
$\pi^{s_{\text{ref}}}_{\text{in},p}$	Passenger p's reference arrival time for a given reference station s_{ref}
ω_p	Passenger p's wait time in the rail system before boarding the train
t^s_i	Departure time of train i at station s
C_i	Capacity of train i
C^s_i	Remaining capacity of train i at station s
D^s_i	Number of passengers arriving at station s between t^s_{i-1} and t^s_i
R^s_i	Number of passengers who are available to board train i
L^s_i	Number of passengers left behind by train i at station s
ρ^s_i	Proportion of passengers left-behind at station s by train i

North Avenue and Midtown stations. As will become clear, this low ridership changes drastically in presence of an event as it is one of two closest stations to the nearby Mercedes-Benz Stadium and State Farm Arena.

D. Event Ridership Estimation

The last section highlights the consistency of the baseline passenger flow. This section describes how the special event ridership is characterized and measured. The estimation for special event ridership aligns with how Pereira et al. [12] use the 90th percentile to measure *hotspot* impact. This analysis focuses on riders with an origin or destination at the Dome/GWCC and Vine City stations, since the majority of

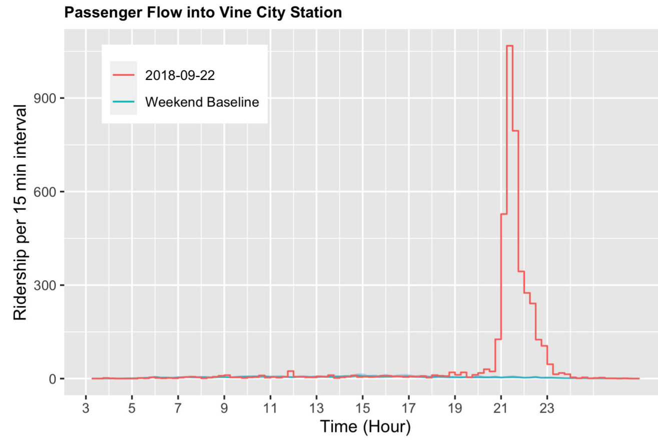


Fig. 2. Vine City station's post-game ridership on September 22, 2018 versus its baseline weekend flow.

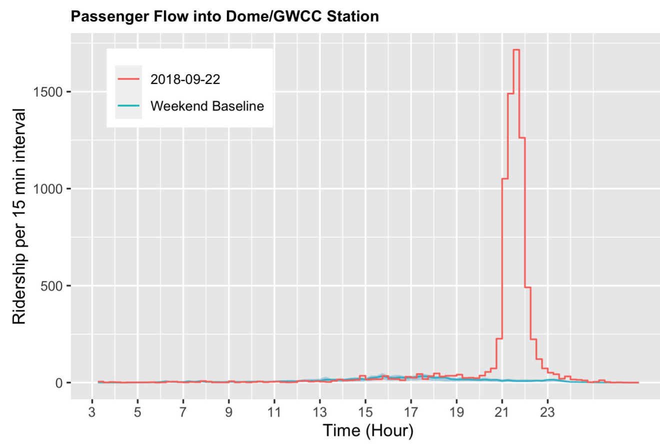


Fig. 3. Dome/GWCC station's post-game ridership on September 22, 2018 versus its baseline weekend flow.

riders use these two stations to get to and from the events at the three venues of focus. This measurement is formalized as follows.

The number of rail riders who attend the event at time t, i.e. E_t, is given by the number of riders $U_t - M_t$ at the points where R_t is strictly greater than the 90th percentile U_t.

$$E_t = \begin{cases} U_t - M_t, & R_t > U_t \\ 0, & R_t \leq U_t \end{cases} \tag{1}$$

The number of rail riders attending the event is given by

$$\mathbb{E} = \sum_{\forall t \in [t_{start}, t_{end}]} E_t. \tag{2}$$

To illustrate the equations above, consider the Atlanta United game on September 22, 2018. Approximately 10,813 people entered the Dome/GWCC and Vine City stations after the game, including 3,787 from Vine City and 7,026 from Dome/GWCC. These values were calculated using Equations 1 and 2. The event had no significant influence on the ridership on any other station. The two station flows for September 22, 2018 are presented in Figures 2 and 3. It can be seen that the two stations are rarely used outside of special events due to the low baseline flow. Thus, the impact of special events on the station passenger flow can be clearly identified.

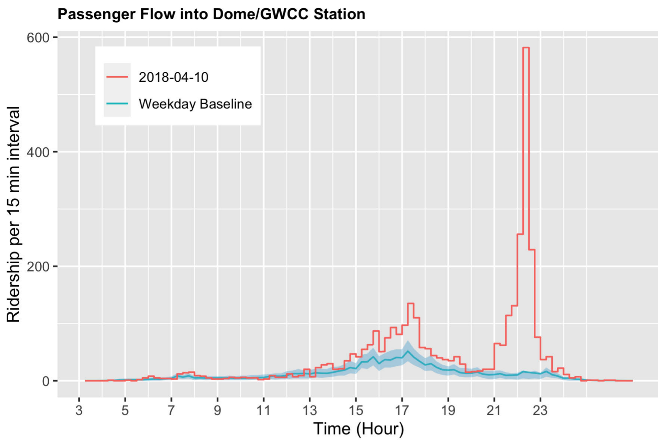


Fig. 4. Illustrating a double event day.

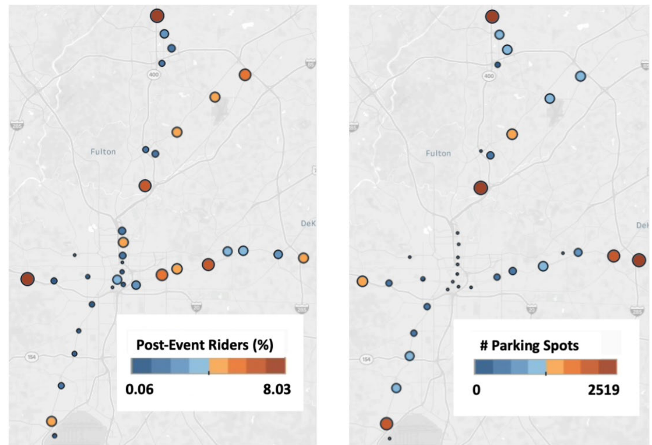


Fig. 5. Heat maps showing the median percentages for post-game destinations (left) and number of parking spots (right).

A double event day is shown in Figure 4. A double event day occurs when two large special events are held on the same day. The later spike corresponds to Hawks v.s. Philadelphia 76ers game in State Farm Arena at 17:30. The earlier spike corresponds to MODEX 2018, an expo in Georgia World Congress Center, which starts at 10:00. It can be observed that the event departure pattern of the riders for the non-sporting event depart is much more spread-out than for sporting events.

The obtained values represent estimates for special-event ridership that are later used to build prediction models. It is assumed that the vast majority of these riders are indeed traffic due to the special events, as the ridership shows large deviations above the normally low baselines at these stations.

E. Event OD Patterns

This section focuses on the OD analysis of all Atlanta United games to understand which areas the special event ridership is coming from. The analysis can be used to understand the distributions for the origins (before the game) and the destinations (after the game) of these riders. Understanding where special-event riders come from can help transit agencies improve their offerings. The destination analysis suggests where riders might live, what forms of transportation they take, and what other factors contribute to the stations they use. Analysis of other types of games give similar results with changes mainly to the magnitude of station ridership.

a) Data: Without loss of generality, this section focuses on the destinations after Atlanta United games, since the later sections will focus on post-game service analysis and simulation. This post-game analysis focuses on riders who enter Dome/GWCC or Vine City stations 1 to 4 hours after the start of the game.

For this analysis, the raw ridership counts for each game are normalized to obtain the percentages of riders alighting at the destination stations. These median percentage for each station are also shown on the left side map in Figure 5 to give a geographical representation of this data.

The right-side heat map in Figure 5 shows number of parking spots available at each station. The heat map suggests that two major factors explain why riders use a particular station for events: noitemsep

1) the proximity of the event location to the East/West line;
2) the parking space availability.

The stations with a high number of riders are on the East/West line or have ample parking or both. The four stations most used by riders (i.e., North Springs station, H.E Holmes station, Lindbergh Center Station, and East Lake Station) are all in the top six for parking spots. A larger number of riders also use the East/West line although the North/South line is closer to larger population centers.

IV. Methods

This section presents multiple techniques that use Automated Fare Collection (AFC) data to evaluate, anticipate, and manage the performance of transit systems during recurring congestion peaks due to special events. Section IV-A presents a technique using unsupervised machine learning to cluster passengers and estimate the departure times of the trains they boarded without an accurate train schedule or train-tracking data during periods with high levels of overcrowding. Section IV-B presents linear regression and random forest models for predicting total event ridership that are used in combination with historical trends to forecast post-game passenger flows at nearby stations after the game. The forecasted demands are used to estimate the required frequencies to serve these customers with minimal passengers left behind by each train. Finally, a simulation algorithm is presented for evaluating a given train schedule in terms of performance metrics such as proportion of left-behind riders and waiting times.

A. Train-Level Clustering

This section presents a clustering method that helps estimate the actual train departure times and proportion of riders left behind at the station when there is not an accurate schedule data or historical vehicle location data. A large event can lead to demand that exceeds the capacity of the the regular train schedules, leading transit agencies to add additional trains compared to the normal weekday or weekend schedules during the post-game peaks. This method can also be helpful to evaluate on-time performance of trains when there is not a record of vehicle location data, such as that produced by AVL technologies.

The train-level clustering analysis identifies the train schedules and which riders were on the same train based on the AFC data. It consists of three steps:

1) *Time Adjustment:* Exit times of riders are adjusted based on scheduled travel time between stations;
2) *Rider Clustering:* Riders are clustered into trains based on their adjusted exit times;
3) *Schedule Estimation:* The train schedule is estimated from the clustering results with a cluster corresponding to group of passengers on the same train.

1) Time Adjustment: To determine the departure time of riders at the event stations, their station exit time is shifted backwards, using the train travel times. Suppose a passenger's origin station is s_1, and the train takes the passenger to their destination station s_3 passing through station s_2 in the middle. Here it is assumed the boarding time is taken into account in the train schedule as MARTA does and that the travel time is defined as the time between departure times. It is also assumed that passengers exit the system at station s at departure time of station s, i.e. $t^{s}_{\text{out},p} = t^{s}_{\delta,p}$. A passenger's travel time in the transit system can be decomposed by the following equation, where τ^{s_1,s_2} is obtained from MARTA train schedule.

$$t^{s_3}_{\text{out},p} = t^{s_1}_{\text{in},p} + \omega_p + \tau^{s_1,s_2} + \tau^{s_2,s_3} \tag{3}$$

Subsequently, if a passenger enters the system at station s_1

$$t^{s_1}_{\delta,p} = t^{s_1}_{\text{in,p}} + \omega_p. \tag{4}$$

From equation 3 and 4, assuming instead of traveling from s_1 to s_3, the passenger travels from s_1 to s_2. Passenger's exit time at station s_2 can be inferred as:

$$t^{s_2}_{\text{out},p} = t^{s_2}_{\delta,p} = t^{s_3}_{\text{out},p} - \tau^{s2,s3}. \tag{5}$$

As an event can have effect on multiple neighboring stations, in order to make the start time comparable, the entry time of a passenger at the event station s_1 from a separate origin station s_0 is defined as:

$$t^{s_1}_{\text{in,p}} = t^{s_0}_{\text{in,p}} + \tau^{s_0,s_1}. \tag{6}$$

The goal of time adjustment is for passengers with different OD. Their entry time and exit time are adjusted based on equation 5 and 6 to the same OD.

2) Rider Clustering: Once the entry times and exit times are adjusted to the same OD, it is possible to apply an unsupervised learning model to cluster riders in trains. Unsupervised learning algorithms are commonly used for analyzing patterns in unlabeled data. In this case, an unsupervised learning algorithm is applied on the times that riders exit the system to learn the patterns of their exit times since riders on the same train and with the same OD have similar exit times. Algorithm HDBSCAN [36], [37] was selected for this task because of its strength in obtaining clusters of varying densities according to the mutual distances between the data points. HDBSCAN is faster than other density based methods and only requires a single parameter for minimum cluster size, unlike other multi-parameter methods. Since the number of trains after any given game are not always known or constant, HDBSCAN works well as it is not needed to specify the number of clusters. The fact that this method uses only a single parameter also makes it very accessible to transit agencies even if there is not someone with extensive experience in machine learning and parameter tuning.

3) Schedule Estimation: The clustering algorithm identifies which passengers are on the same train for the adjusted OD pair. The train departure time at the origin station s is approximated as the latest arrival time of the riders on that train at that station. Given the train schedule of the origin station, the train schedule for all stations can be derived from $\tau^{s,s'}$, using

$$t^{s'}_i = t^s_i + \tau^{s,s'}. \tag{7}$$

B. Predictive Analytics

This section presents methods to train linear regression and random forest models using historical AFC and event data to predict the total number of event riders to help prepare for post-game peaks. Then, it presents a method for creating a new proposed schedule for a future game based on the total event ridership prediction and average passenger flow curve for similar events as well as the simulation used for evaluation.

1) Ridership Prediction Model: This section presents supervised machine learning models to predicted total event ridership. Supervised learning models can be used to create a mapping function from input features to the output. Here, the output is the number of event riders that can be estimated by equation 1 and 2. These models can be used with the mean rider throughput curves (see Figure 9, 10) for each event type to forecast the post-game passenger flows. Table IV lists the input attributes compiled for the predictive models to use as input and outputs T_a, i.e., the total number of riders for the event at the considered station. In the table, Event 1 is the event whose ridership must be predicted and Event 2 is another event on the same day. For days with a single event, the attributes of Event 2 are set to null. Note that, for future events, a prediction of event attendance can be used in place of actual attendance. For example, NBA game attendance has been predicted using random forest models with a 6% MAPE using team/opponent statistics, stadium capacity, local average income, team popularity and other factors in [18].

The first predictive model is a simple linear regression (LR), which uses Attendance from Event 1 as the sole feature.

$$\mathbb{E} = \beta_0 + \beta_1 \times \text{Attendance}, \tag{8}$$

where β_0 and β_1 are parameters to be estimated using minimization of sum-of-squares error. It captures the strong linear dependency between the event ridership and the event attendance, which is highlighted in Figure 11. This is mainly due to the size of Mercedes-Benz Stadium in comparison to the other two venues, leading to much higher attendance numbers. The second predictive model is a random forest (RF) that uses all the attributes in Table IV. The third model (LR+RF) is a two staged approach. The first stage is a linear regression model described in Equation 8 using the Attendance from Event 1 as the sole feature. The second stage is a random

TABLE IV
THE INPUT ATTRIBUTES FOR THE PREDICTIVE MODELS

	Attributes	Type
Event 1	Category	Factor: Soccer, Football Game, Basketball
	Location	Factor: State Farm Arena, MBS, MBS with Upper deck Open
	Attendance	Numeric, Attendance for Event 1
	Win Percentage Difference	Numeric, Home Team win percentage minus Away Team win percentage
	Regularized Margin	Numeric, Margin of victory (loss) divided by standard deviation margin of victory for that league
Event 2	Category 2	Factor (15 Categories)
	Location 2	Factor (GWCC/MBS/SFA/No Location)
	Attendance 2	Numeric Attendance for Event 2)
	Time Difference	Time difference (in minutes) of the two events, 0 if there is no second event
	Two_Event	Binary (True, if there is a second event)
	Week	Binary (True, if the day is weekend)
	Month	Factor (Month of the event)

forest model used to predict this error term using the remaining features in Table IV as described in Equation 9,

$$\epsilon = \frac{1}{B}\sum_{b=1}^{B}\mathcal{T}_b(\mathbf{x}), \tag{9}$$

where B is the number of decision trees, $\mathcal{T}_b$ is the b^{th} decision tree, and $\mathbf{x}$ is the input vector. B is an hyper-parameter obtained by fitting the model over different values and selecting the one minimizing the RMSE. Not only do these methods work well for this scenario, but they also need minimal tuning and are easily interpretable relative to other types of supervised learning models.

2) Proposed Schedule: To prevent the overcrowding of the stations, the key strategy in this paper is to estimate the required train frequencies to serve the forecasted demand with minimal passengers left behind. Transit agencies may select a minimum frequency $f_{\min}$ and maximum frequency $f_{\max}$ requirement. Generating a new train schedule, which we refer to as the *proposed schedule*, for the periods of increased demand can be done by following these steps:

1) For a given direction of a line, scale the total event ridership forecast based on the average percentage of riders that use the line in that direction during the peak period after the game. This set of ridership is the targeted post-event riders P.
2) Divide the post-game period into bins $[B_{\tau_\alpha}, \alpha = \{1 \ldots n\}]$, where $\{\tau_1, .., \tau_n\}$ are consecutive 5-minute time intervals for similar previous events based on event end time.
3) For each event station $s \in S$, compute the forecasted arrivals for each bin B_{τ_α} by multiplying the average percentage of post-game event riders that arrived during B_{τ_α} for similar previous events. For non-event stations, the baseline mean passengers can be used to forecast arrivals.
4) For each station s and bin B_{τ_α}, generate arrivals times $t^{s}_{\text{in},p}$ that are equally distributed in the interval τ_α. Let $\{\pi^{s_{\text{ref}}}_{\text{in},p} = t^{s}_{\text{in},p} + \tau^{s,s_{\text{ref}}}; p \in P, s \in S\}$ be the sorted reference arrival times for passenger set P, station set S, and reference station s_{ref}.

Algorithm 1 Train Scheduling for Post-Games on Reference Station s_{ref}

Input: Set of Trains: I
Train Capacities: $C_i \ \forall i \in I$
Sorted Reference Arrival Times: $\pi^{s_{\text{ref}}}_{in,p} \ \forall p \in P$
$p \leftarrow 0$
for i *in* I **do**
  $c_i \leftarrow 0$
  while $c_i < C_i$ **do**
    $c_i \mathrel{+}= 1$
    if $c_i == C_i$ **then**
      $t^{s_{\text{ref}}}_i \leftarrow \max(\pi_{in,p}, t_{i-1} + f_{\min})$
      break
    end
    else if $\pi^{s_{ref}}_{in,p} > t^{s_{ref}}_i + f_{max}$ **then**
      $t^{s_{\text{ref}}}_i \leftarrow t^{s_{\text{ref}}}_{i-1} + f_{\max}$
      break
    end
    else
      $p \mathrel{+}= 1$
    end
  end
end
Output: Proposed schedule $t^{s_{\text{ref}}}_i \ \forall i \in I$

5) Generate a proposed set of departure times for a reference station using Algorithm 1 by estimating the the required frequencies to serve the forecasted passenger flows at nearby stations with minimal passengers left behind.
6) Finally, align each train in the original time table to the nearest train in the proposed schedule based on the reference station. As long as $f_{\max}$ is equal to or less than the frequency of the normal timetable, then each train will be uniquely mapped to the new proposed schedule.

Algorithm 1 estimates the required frequencies to effectively manage forecasted passenger flows, ensuring minimal passengers are left behind. The notation is presented in Table III. This algorithm focuses on minimizing the number of passengers left behind during periods when demand surpasses the capacity of the normal rail schedule to reduce wait time for both event and non-event riders.

During post-game periods for large events, demand often far exceeds the supply generated by the normal timetable. Large special events often occur at night or on weekends when train schedule is typically reduced compared to peak commuting periods. In the case of MARTA, many additional trains are added to the normal timetable during the post-game periods to service the increased demand with reasonable wait times. This method leverages forecasts of the passenger flow curves for the post-game period to create a proposed post-game schedule for transit agencies to use during their the tactical planning. This method is designed to serve the forecasted demand with minimal passengers left behind. Due to the importance of preventing excessive overcrowding, delays, and disruptions for transit agencies, additional buffer can be added to forecasts potential to account for potential uncertainty.

Algorithm 2 Proportion-Left-Behind Simulation

Input: Set of Trains: I Set of Stations: S
Train Capacities: $C_i \ \forall i \in I$
Train Departure Times: $t_i^s \ \forall i \in I, \forall s \in S$
Passenger Arrival Times : $t_{in,p}^s \ \forall p \in P$
l_0^s=0
$C_i^0 = C_i$
for *i in 1:I* **do**
  for $s \in S$ **do**
    Calculate demand D_i^s from $t_{in,p}^s$
    $R_i^s = L_{i-1}^s + D_i^s$
    $L_i^s = \max(R_i - C_i^s, 0)$
    Calculate wait times ω_p for passengers on train s
    $C_i^{s+1} = \max(0, C_i^s - R_i^s)$
    $\rho_i^s = \frac{L_i^s}{R_i^s}$
  end
end
Output: $\rho_i^s \ \forall i \in I \ \forall s \in S$
$\omega_p \ \forall p \in P$

3) Simulation: This section presents an algorithm, Algorithm 2, used to simulate a first-in-first-out (FIFO) queue of riders at stations in set S. This simulation is used to validate the clustering results and also evaluate system performance in terms of fundamental metrics such as the passenger load per train, the wait times (ω_p) of riders, and the proportion of riders left behind ρ_i^s. For each train i and station s, Algorithm 2 calculates the amount of new demand D_i^s, the total number of passengers trying to board R_i^s, the amount left behind L_i^s, the remaining capacity C_i^{s+1}, and the proportion left behind ρ_i^s.

V. Results

This section presents the results of the techniques used to exploit Automated Fare Collection (AFC) data for evaluating, anticipating, and managing the performance of transit systems during recurring congestion peaks due to special events. First, the results of the unsupervised machine learning algorithm HDBSCAN used to cluster passengers and estimate the departure times of the trains are presented and validated with simulation. Simulation are also used to estimate the average train load during these peak periods with high proportion of riders left behind. Then, in Section V-B, the results of the linear regression and random forest models for predicting total event ridership are shown. Additionally, the consistency of the passenger flows post-game for both Atlanta Falcons and Atlanta United games are shown. Finally, simulations are performed that showcase how the adjusted train schedules based on forecasted demand using the proposed methods may significantly improve the post-game congestion and wait time.

A. Train-Level Clustering

This section presents the results of the unsupervised machine learning algorithm HDBSCAN used to cluster passengers and estimate the departure times of the trains. The clustering results are presented and validated with simulation. First, it goes through an example day starting with the departure time inference step and ending with the simulation estimation of train capacity by simulating various train capacities and comparing the proportion of passengers left behind to that of the clustering results.

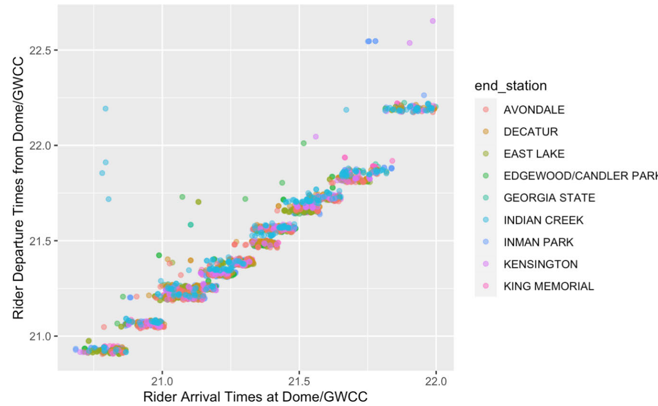


Fig. 6. Arrival and departure time of riders at the Dome/GWCC station.

1) Case Study Data: The analysis in this section focuses on the Atlanta United game on September 22nd, 2018 for concreteness. Due to the nature of the rail system, the primary focus is on riders using the rail to travel in the east direction after leaving the stadium and entering either the Dome/GWCC or the Vine City stations. Some passengers may proceed to switch to another rail line, but the analysis focuses on the subset of passengers who solely use the West to East tracks (either the Blue or Green line).

2) Departure Time Inference: Figure 6 shows departure time inference results with Dome/GWCC as the event station. In the figure, the colors represent a different alighting station for the riders. Observe the horizon clusters that represent sets of riders boarding the same train.

3) Clustering: This section reports the results of this clustering for riders entering the Vine City and Dome/GWCC stations after the game and alighting at Edgewood/Candler Park, East Lake, Decatur, Avondale, and Kensington. The most crowded period is between 20:40:00 and 22:00:00 and is the focus of this section.

To cluster the selected 2,392 riders in selected time interval, HDBSCAN was run with its parameter `MinPts` for minimum cluster size set to 50. An initial run detected 12 clusters and the 21:45 and 21:49 trains were not separated because the 21:45 train was delayed at some stations and hence the departure time inference resulted in some scattered data points. The first HDBSCAN run identified the noisy data points and a second HDBSCAN run was applied to produce the 13 clusters that correspond to the 13 trains that left Dome/GWCC after the game. The cluster results and the estimated train arrival times at Dome/GWCC station are presented on Figure 7. The estimated train departure times at the Dome/GWCC station are plotted with dash lines which represent the latest rider boarding time for each train.

The clustering algorithm assigns each rider to the corresponding train they boarded. These passengers enter MARTA

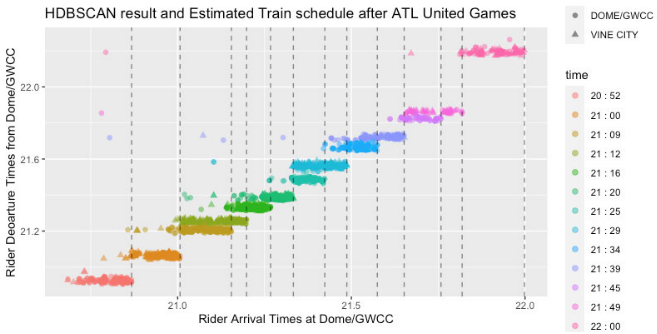


Fig. 7. Train clusters for riders entering the dome/GWCC and vine city stations after the atlanta united game on september 22nd, 2018.

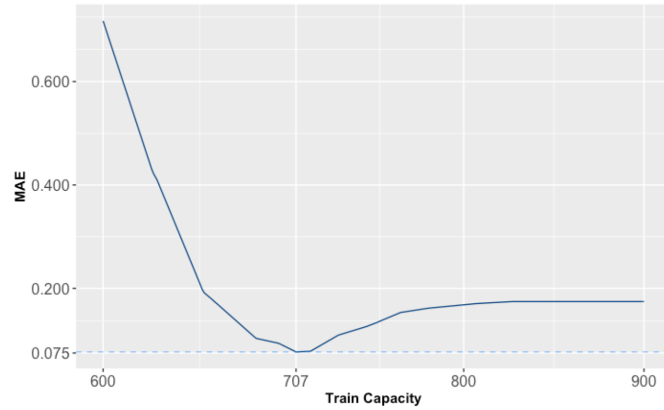


Fig. 8. Minimization of MAE for L_1 loss function when searching for the maximum capacity.

from two stations, Vine City and Dome/GWCC. Because the trains are heading to Dome/GWCC from Vine City, Vine City riders have a priority to board the train. During peak times after the game, some riders cannot board the first "available" train. Moreover, some riders entering Dome/GWCC have to wait up to three trains to board. Hence, an important metric to evaluate MARTA's performance is the proportion of riders left-behind by each train, which is referred to as *proportion-left-behind*. Note that 3 of the 13 trains did not stop at the Vine City station while all 13 trains did stop at Dome/GWCC station. Many of the trains are mostly filled with Vine City passengers, so this decision likely helps improve the wait time for riders using the Dome/GWCC station.

4) Train Capacity Utilization: From the previous clustering, it is possible to tell, for a subset of passengers, which train they were on and identify what was the number and frequency of trains departing the stadium area. This section uses simulation with the arrival data and the train frequencies to match the proportion-left-behind to determine roughly how many people were boarding the trains when the stations are crowded, which will be helpful for later simulations.

To estimate the train capacity, this section presents a simulation that estimates the proportion-left-behind at a station based on the arrival time of passengers and the train capacity. By comparing the proportion-left-behind computed by the simulation and the clustering, it becomes possible to estimate how the trains are utilized. More precisely, the goal is to find a train capacity that minimizes the absolute difference (L1 loss) between the proportion of left-behind passengers from the simulation and clustering.

Algorithm 2 is repeated for train capacities from 600 to 900 (assuming every train has the same capacity $C_i = C_j$ $\forall i, j \in I$) while keeping I and T_i the same. The I and T_i used in this section are estimated from the clustering model. Each simulation on a capacity outputs an estimation of the proportion-left-behind by each train. The L1 loss between the proportion of riders left behind from the simulation and from the clustering is calculated.

Figure 8 depicts how the L1 loss function evolves for different train capacities. When assuming the train capacity is 707, the proportion-left-behind from the simulation best matches the proportion-left-behind from the clustering, suggesting a train capacity of 707. Table V presents the results and reports

TABLE V
SIMULATION RESULTS FOR THE PROPORTION-LEFT-BEHIND BY EACH TRAIN

Train Time	Proportion-left-behind estimated at Dome by the clustering	Proportion-left-behind at Dome in the simulation (Maximum Capacity 707)
20:52	4%	0%
21:00	11%	18%
21:09	8%	19%
21:12	60%	70%
21:16	29%	29%
21:20	16%	0%
21:25	9%	0%
21:29	37%	31%
21:34	35%	17%
21:39	12%	0%
21:45	7%	0%
21:49	0%	0%
22:00	0%	0%

the proportion-left behind by both the clustering and simulation models for a train capacity of 707 at the Dome/GWCC station. The "real" percentages (from clustering) are larger than the simulated percentages when the numbers are small. This is due to the fact that the simulation assumes an orderly first-come-first-served system where a full train has exactly C_i passengers. In reality, the number of people in "full" trains will vary when the station is crowded.

The maximum capacity of 707 is a lower bound estimation because riders already on the trains (approximately a total of 35 people for the whole time period) are not counted here. These trains are 6-car trains post-game which have a recommended maximum capacity of 576 people. From this analysis, however, one can see that the maximum capacity is often exceeded post-game: people often cram together in very close quarters as a result. It is also likely that this over-capacity situation leads to an increased risk of accident, injury, or illness. However, the analysis simply confirms the anecdotal evidence that people have a tendency to "pack it in" after sporting events. Note also that, under the assumption that people left behind end up boarding the next train before the new arrivals, riders wait a maximum of two trains, which corresponds to the case in Figure 7.

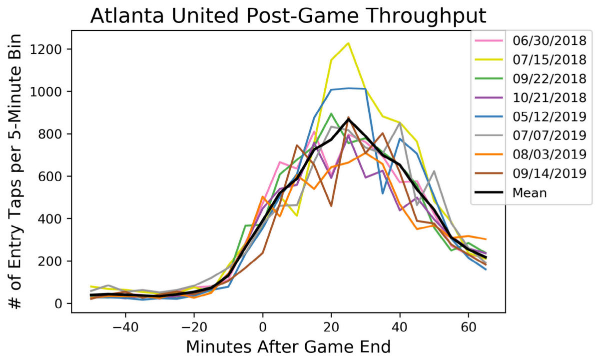


Fig. 9. Post-game rider throughput at Dome/GWCC after Atlanta United games with the upper-deck seating opened.

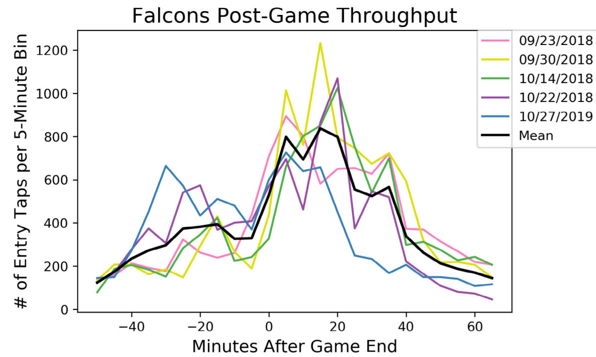


Fig. 10. Post-game rider throughput at Dome/GWCC after falcons games.

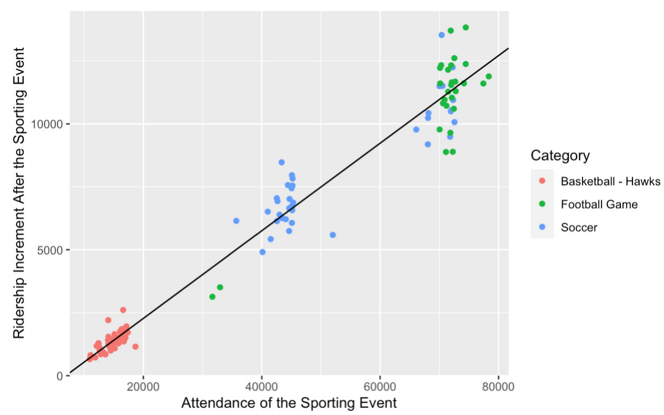


Fig. 11. Linear trend between attendance and ridership increment after the main event.

B. Predictive Analytics

This section presents the results of the demand forecasting and train scheduling with a focus on regular season Atlanta United games with larger attendances. Simulations are performed that showcase how the adjusted train schedules based on forecasted demand using the proposed methods may significantly improve the post- game congestion and wait time.

1) Throughput Consistency: This section highlights the post-game throughput consistency for stations near special events. Due to their larger sizes, events at Mercedes-Benz where the upper-deck seating was open are the focus of this section. In games with an open upper deck, an additional 30,000 seats are available for purchase in the upper deck of Mercedes-Benz Stadium. Due to the additional attendance, these events have a much larger impact on the MARTA rail system, especially compared to Atlanta Hawks games where the average attendance is only around 15,000 people. In this section, the post-game throughput is analyzed from 40 minutes before the end time to 80 minutes after the end time.

In most games, it is assumed that the end time is the average game length after the scheduled start time: 1 hour & 50 minutes for soccer and 3 hours & 10 minutes for football. However, a few of the end times were adjusted in this analysis because it was believed that the end time might have been delayed due to injuries, delayed starts, or overtime. The delay in the actual end time of the game compared to the end time calculated using the average game length is referred to as the *offset*. For each game, the *offset* is estimated by comparing the throughput curves in cases where there was a clear delay to the peak of the throughput. When there are delays to the game, such as overtime for a Falcons game, the train operator waits to make the necessary adjustments to the actual schedule.

Figure 9 shows the entries to Dome/GWCC and Vine City are grouped into 5 minute bins and plotted for analysis. Note that three games had offset adjustments as stated later in Table VII. This highlights the consistency of arrivals to the rail stations Dome/GWCC and Vine City rail stations after Atlanta United games with an open upper deck. The highest number of riders arriving in any bin is almost 1,200, which can be served with less than two trains assuming a train capacity of 707 as estimated in Section V-A.4. Note also that only roughly 8% of people take a train going west, while the rest of the riders wait for trains going east. Figure 10 show a similar, yet distinct consistency for the football games. In some of the Falcons' games, there are some peaks that could align with the end of the third quarter (ex. 10/27/2019 where the Falcons were losing 24-0 at halftime). Falcons games are typically more than an hour longer than Atlanta United games in length, so this could also explain why more fans leave early for Falcons games than for Atlanta United games.

The mean curve gives an relatively accurate estimation of the arrival patterns at the station. The mean curve is converted into percentages by dividing by the average ridership and used later in combination with the predicted ridership to obtain a predicted throughput curve. Given a ridership prediction, it is possible to estimate the arrival distribution over time at the station quite accurately.

2) Ridership Prediction: The results presented in this section focus on sporting events near Dome/GWCC station and Vine City station, i.e., Atlanta Hawks games, Atlanta Falcons games, and Atlanta United games, as there are enough data points for those events to build strong models. These events are also among those that had the largest impact on ridership. Post-game ridership is estimated using Equation 1. The training data consists of 134 event days ranging from January 2018 to December 2019.

After the first stage, the error term ε is calculated from the fitted LR model as shown in Equation 10, with the parameters $\beta_0 = -1201$ and $\beta_1 = 0.1739$.

$$\text{Ridership} = -1201 + 0.1739 \times \text{Attendance} + \varepsilon \tag{10}$$

TABLE VI
PREDICTION RESULT FOR MODEL 1

	MAE	MAPE	RMSE
LR	509	0.1169	509.26
RF	582	0.1356	582.10
LR+RF	**506**	**0.1130**	**506.24**

The LR+RF model recognizes that the error term of the linear model depends on other factors, e.g., whether there is a second event at that day or the win percentage of the home team.

Table VI presents the results obtained using a *leave-one-out cross validation* because of limited sample size. RF uses 1,500 trees and LR+RF uses $B = 800$ trees. The proposed LR+ RF outperforms the other models among all metrics, i.e., MAE, MAPE, and RMSE.

3) Simulation: The goal of this section is to take the results from the previous section and demonstrate how they can be combined into a predicted passenger flow for a future post-game event. To do this, the passenger flow for previous games is predicted and simulated to evaluate how it would have performed versus the actual schedule. Again, please note that the train operators pulled additional trains from reserves during this period to deal with excess demand.

The forecast splits the post-game ridership in five minute bins and can then be used to create a train schedule. The proposed schedule can be compared to the actual (recovered) schedule, giving key insights to help dispatchers improve performance of the rail service during post-game spikes.

The case study is focused on the Atlanta United games with the upper deck open. Two Atlanta United games were excluded from the analysis, one because it was a playoff game and the other because there was an overlapping basketball game and no similar examples to use for the predictions. For each day, the actual schedule is recovered using the methods in Section IV-A. Note manual adjustments are made to add in trains in the case that two trains were close together to make sure that it was a fair comparison.

Due to the results of Section V-A.4, it is assumed that each train will fill up to a max of 707 passengers. Based on the analysis in III-E, the total ridership prediction is scaled by 92% (the percentage of riders using the green/blue lines in the east direction), then by 68% (the percentage of riders leaving during this post-game peak period), then finally by 110% to add a buffer, based on the average MAPE from the previous section IV-B.1. Then, this forecasted demand is used to create the proposed schedule using Algorithm 1.

The vector graph for MARTA's regular eastbound service on weekends can be seen in Figure 12. These times were derived from MARTA's past GTFS schedule. The eastbound direction consists of alternating trains of the Blue Line, which goes all the way from H. E. Holmes to Indian Creek, and the Green Line, which covers some of the middle stops and additionally covers Bankhead station. Figure 13 shows the proposed eastbound train schedule derived from the generated departure times for Dome/GWCC using Algorithm 1. 8 trains are added in the eastbound direction, with the 5 blue line

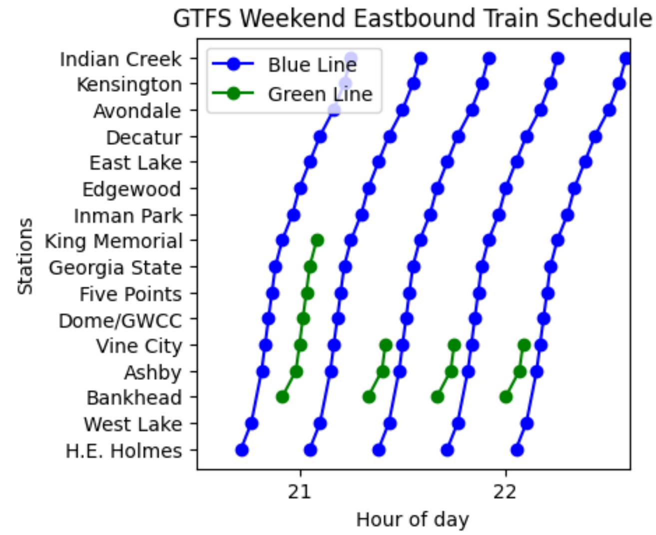


Fig. 12. The normal (non-event) weekend eastbound train schedule based for MARTA.

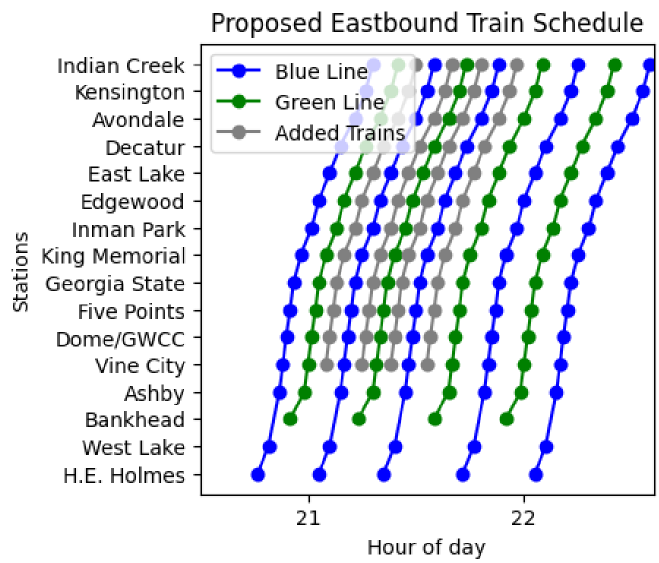


Fig. 13. The proposed eastbound post-game train schedule for the atlanta united game on september 22, 2018 based on the generated departure times for Dome/GWCC from Algorithm 1.

trains staying relatively consistent. However, it is important to note that of the 8 added trains, 4 of them are simply extensions of the 4 green line trains seen in Figure 12 that now go all the way to the Indian Creek station. Thus, only 4 additional trains are needed. The trains are more frequent in the proposed schedule, but meet the minimum headway requirements of MARTA. Figure 14 is the reconstruction from Breeze Card data of the actual stops and frequencies of trains on September 22, 2018. The lines each train belongs to is depicted as well. In comparing Figure 13 and Figure 14, one can see that the proposed schedule uses a reasonable number of trains relative to what actually transpired. Furthermore, the added trains start slightly earlier and are more evenly spaced out. Figure 15 compares the capacity of the actual and composed schedules against the arrival rate for the Atlanta United game on September 22, 2018. It demonstrates how the

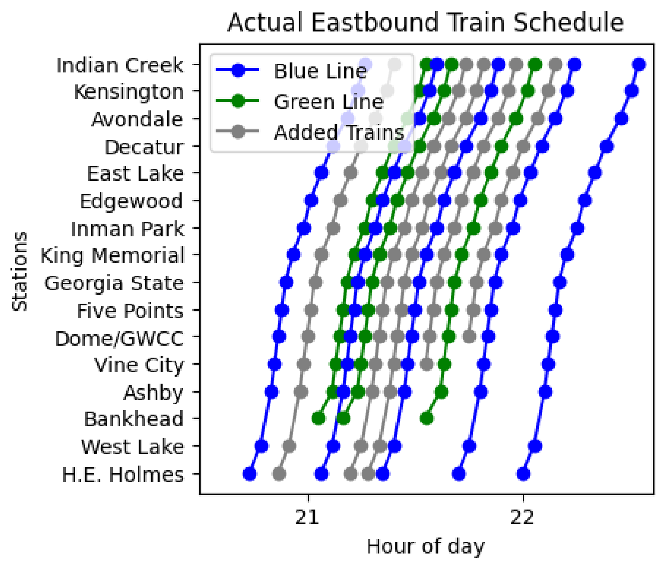


Fig. 14. The actual eastbound post-game train schedule for the Atlanta United game on September 22, 2018 estimated from the Breeze data.

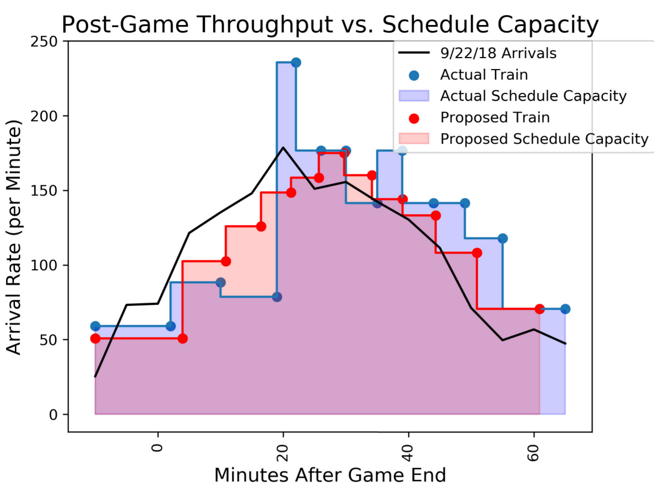


Fig. 15. The arrival rate for the Atlanta United game on September 22, 2018 is depicted along side the "capacity" of the actual and proposed schedules, which have their height represented by the capacity per train (707 passengers) divided by the minutes since the last train.

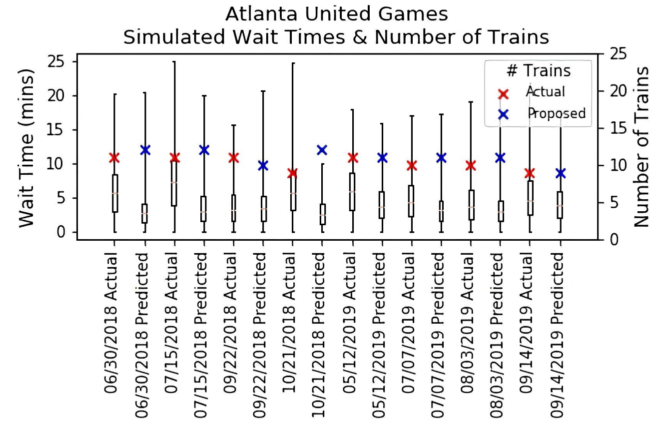


Fig. 16. For each of the eight Atlanta United games with the upper deck open, the two box plots compare simulated actual train schedules vs. simulated proposed train schedules. Note the number of trains is represented with x markers.

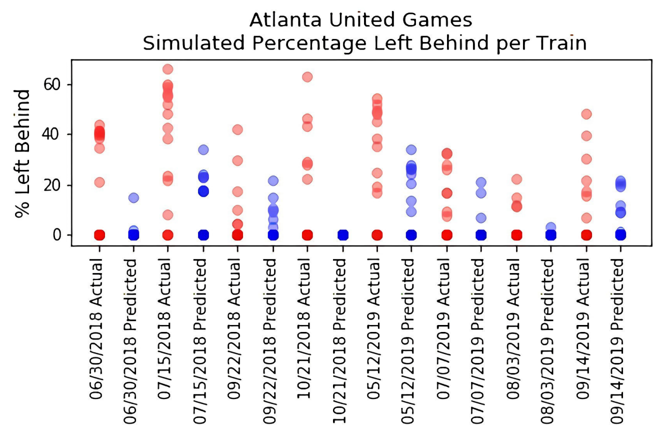


Fig. 17. For each of the eight Atlanta United games with the upper deck open, the *proportion-left-behind* for each train is plotted for both simulations. The results from the simulations of actual (recovered) schedules are in red and results from simulations with proposed schedules are in blue. Note trains near the beginning and end of the schedules tend to have 0 riders left behind.

capacity of the proposed schedule more closely aligns with the actual demand, which leads to better service quality.

The proposed and actual schedules are compared by evaluating their simulation results using Algorithm 1 across the subset of Atlanta United game day with the upper deck seating open while keeping the other inputs the same. The train capacities C_i $\forall i \in I$ are set to 707. I and S are the actual I and S for the specific game day. The simulations focus on the post-game time period, which requires an increased train frequency to service the surge in demand. It is assumed that the schedules return to normal (every 20 minutes per line) following this modified schedule for the peak period.

Table VII displays the results of the 16 simulations: two for each of the eight Atlanta United games of focus where the upper-decks were open. There is significant decreases to the average percent left behind ($\rho_{i,s}$) on average for each train compared to the actual train, which shows the schedule is better matching the increased demand. Additionally, average passenger wait time (ω_p) improves in most cases. In the case of MARTA, the majority of riders during this post-game period are boarding at Vine City and Dome/GWCC stations. There is a less significant impact on wait time and passengers left behind for non-event riders at the following stops, since enough passengers alight and transfer at Five Points, the stop following Dome/GWCC. On average about one more train is used as an extra buffer was added to protect against the times that the ridership is underpredicted. This is the number of trains that depart from Dome/GWCC within the considered time window.

Figure 16 shows boxplots of the simulated wait time using both actual and proposed schedules. The number of trains represented by the $\times$ markers. Note that, in most cases, the maximum wait time is decreased, as well as the 75th percentile and median.

Figure 17 highlights the proportion of people left behind by individual trains. The proposed schedules performed significantly better in this category as the demand matches the proposed train schedule more accurately. In the existing

TABLE VII
RESULTS FROM TWO SIMULATIONS FOR EACH OF THE EIGHT ATLANTA UNITED GAMES USING ACTUAL AND PROPOSED SCHEDULES ASSUMING A MAX OCCUPANCY OF 707 RIDERS PER VEHICLE

Game Data		Ridership		Simulated with Actual Schedule				Simulated with Proposed Schedule			
Date	Offset	Actual	Predicted	# trains	Avg. ω_p	Std.	Avg. $\%\rho_i^s$	# trains	Avg. ω_p	Std.	Avg. $\%\rho_i^s$
6/30/18	0	7456	8292	11	5.7	3.5	22.3	12	3.3	2.8	**1.2**
7/15/18	25	8773	8097	11	7.2	4.3	39.2	12	4	3.8	**12**
9/22/18	0	7693	6765	11	3.8	2.8	8	10	4	3.4	**5.3**
10/21/18	10	6431	7787	9	6.5	4.7	20.2	12	2.8	2	**0**
5/12/19	0	7790	6949	11	6.1	3.7	28.4	11	4.2	3.1	**13.5**
7/7/19	25	7356	7333	10	4.7	3.2	13.6	11	3.6	2.9	**3.3**
8/3/19	0	6517	7250	10	4.4	3.4	4.8	11	3.3	2.5	**0.2**
9/14/19	0	6631	5909	9	5.3	3.9	15.7	9	4.3	3.1	**8**

TABLE VIII
ROBUSTNESS ANALYSIS FOR EACH OF THE EIGHT ATLANTA UNITED GAMES USING VARIOUS PERCENTS OF ACTUAL TOTAL RIDERSHIP AS THE FORECASTED TOTAL RIDERSHIP WITH NO ADDITIONAL BUFFER

Game Data	90% of True Ridership				True Ridership				110% of True ridership			
Date	# trains	Avg. ω_p	Std.	Avg. $\%\rho_i^s$	# trains	Avg. ω_p	Std.	Avg. $\%\rho_i^s$	# trains	Avg. ω_p	Std.	Avg. $\%\rho_i^s$
6/30/18	9	7.1	4.2	29.7	10	4.1	2.8	8.0	11	3.3	2.4	1.9
7/15/18	11	6.1	4.9	25.5	12	3.5	2.9	10.7	13	2.6	2.0	2.0
9/22/18	9	4.9	3.5	17.0	10	3.3	2.5	1.2	11	3.0	2.4	0.0
10/21/18	8	6.5	3.8	24.7	9	4.0	2.6	5.3	9	3.5	2.3	1.5
5/12/19	9	7.1	6.1	29.7	11	3.9	3.1	11.9	12	3.0	2.3	3.4
07/07/19	9	5.3	4.3	14.5	10	3.8	3.0	4.8	11	3.3	2.5	2.6
8/03/19	8	6.3	4.3	19.8	9	4.2	3.2	3.9	10	3.8	2.8	1.6
9/14/19	8	6.0	4.2	19.6	9	4.0	2.8	6.0	10	3.3	2.4	1.8

schedule, less than half the people at the station are able to board the train. This could lead to potential crowding and decreased customer experience as passengers have to wait multiple trains before boarding in some cases.

The robustness of the methods outlined in this paper are validated by a set of experiments that can be seen in Table VIII. For these experiments, various proportions of true ridership were used as the forecasted value when generating the proposed schedule. One can see that improvements on the performance of the actual schedule from Table VII are achieved by using the true ridership value and 110% of the true ridership. With 100% accurate predictions, the wait time and average proportion left behind are improved across the scenarios with further improvements when an additional 10% is added to the forecasted input in the rightmost columns. In the 90% case, where ridership is underpredicted by 10%, the wait times and overcrowding are still sometimes better than the simulated performance of the actual schedule, but overall leads to a significantly poorer performance than the 100% and 110% cases. For transit operators there is significant concern with mitigating the potential negative impact of special events on wait time, congestion, and safety for both event and non-event riders. Therefore, it is justified to slightly overprepare during the tactical planning stages by adding a buffer to forecasted ridership to account to account for the average prediction error. Overall, the proposed methods are fairly robust to forecasts that are slightly off and result in feasible schedules that can achieve reasonable performance even if slightly more people use transit than expected.

VI. DISCUSSION

This paper studies the impact of forecasting and scheduling methods on transit congestion and wait times after large sporting events. The high density of post-game ridership can have negative impacts on congestion and wait times for both event and non-event riders. Transit agencies often have to use additional trains compared to their normal weeknight or weekend schedules. This paper provides key tools for transit operators to use for tactical planning for this disruptive post-game periods to ensure that the system runs smoothly and efficiently.

This paper presents a case study on MARTA, focusing on the large sporting events at the Mercedes-Benz Stadium in downtown Atlanta, GA. The results show that post-game special event ridership can effectively be forecasted by combining total ridership predictions with historical passenger arrival curves. Additionally, the case study highlights how transit operators can significantly improve passenger wait times and congestion following large events by estimating the required train frequencies to serve the forecasted demand with minimal passengers left behind and then generating a post-game schedule. For further improvements, transit agencies can also consider using other existing real-time optimization techniques such as short-turning, skip-stop patterns, or adding buffer time to their schedules. Even though the proposed methods were designed to create a schedule days or weeks in advance for tactical planning purposes, they are very computationally efficient, taking only seconds to run, and could be modified in the future for real-time, operational planning problems.

The analysis and methods are expected to be generalizable to other cities that are both more decentralized and polycentric than Atlanta, since the main focus of the analysis starts with identifying specific line and direction where special events cause a surge in demand beyond the capacity of the normal weekday or weekend schedule. For many transit agencies, special events often occur during normal off-peak hours.

Fig. 18. The MARTA Rail lines and stations [35].

Weeknight and weekend schedules often do not have enough capacity to handle the crowds that swarm to the to the rail lines after a large event. Therefore, advance coordination and tactical planning to increase the total number of trains or cars per train is often possible and will help to alleviate wait times, congestion, and potential safety concerns. In some cases, transit agencies may have system capacity that is insufficient to serve the post-game arrival rate. In these cases, the transit agencies are limited to running their service at maximum frequency and looking at ways to mode shift passengers to buses or other forms of transit.

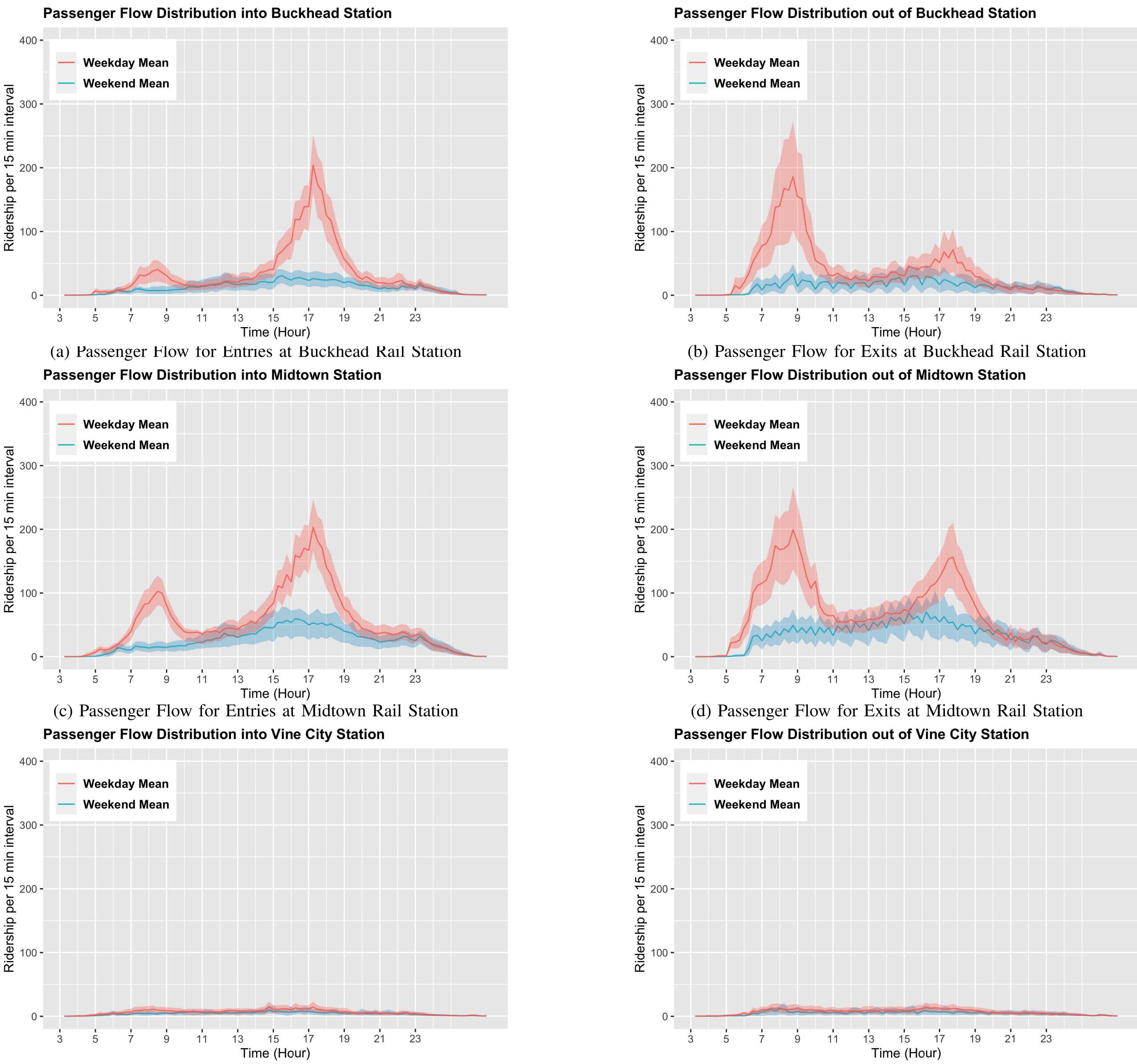
Future work could include a longitudinal study over more years to further improve the accuracy with larger sample sizes. Mercedes-Benz Stadium and Atlanta United did not exist before 2018, thus, no additional historical data is available. Similarly, Atlanta in 2020 was affected by the COVID-19 pandemic, thus restricting the data set to less than two full years. Another direction would be additional case studies on other transit agencies and types of events, including events with lower sample sizes compared to more the more frequent professional sporting events. It would also be interesting to perform related analysis to evaluate potential impacts of any future infrastructure projects such as line extensions on special event riders for transit agencies.

Appendix A
MARTA Rail Station Map

See Fig. 18.

Appendix B
Mean Ridership Flows

See Fig. 19.



(a) Passenger Flow for Entries at Buckhead Rail Station
(b) Passenger Flow for Exits at Buckhead Rail Station
(c) Passenger Flow for Entries at Midtown Rail Station
(d) Passenger Flow for Exits at Midtown Rail Station
(e) Passenger Flow for Entries at Vine City Rail Station
(f) Passenger Flow for Exits at Vine City Rail Station

Fig. 19. Passenger Flow (non-event days) graphs for three stations showing the weekday (red) and weekend (blue) Passenger flow for entries (left) and exits (right). The shaded region represents the 10th-90th percentile range for each bin.

ACKNOWLEDGMENT

Many thanks to Robert Goodwin, the Director of research and analysis with MARTA, for asking the research questions that motivated this research and sharing MARTA's data.

REFERENCES

[1] M. K. El Mahrsi, E. Côme, L. Oukhellou, and M. Verleysen, "Clustering smart card data for urban mobility analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 712–728, Mar. 2017.

[2] M. Moradi and M. Trépanier, "Assessing longitudinal stability of public transport users with smart card data," *Transp. Res. Proc.*, vol. 48, pp. 1364–1375, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146520305834

[3] M. S. Ghaemi, B. Agard, M. Trépanier, and V. Partovi Nia, "A visual segmentation method for temporal smart card data," *Transportmetrica A, Transp. Sci.*, vol. 13, no. 5, pp. 381–404, May 2017, doi: 10.1080/23249935.2016.1273273.

[4] P. Shen, L. Ouyang, C. Wang, Y. Shi, and Y. Su, "Cluster and characteristic analysis of Shanghai metro stations based on metro card and land-use data," *Geo-Spatial Inf. Sci.*, vol. 23, no. 4, pp. 352–361, Oct. 2020, doi: 10.1080/10095020.2020.1846463.

[5] W. Li, Q. Luo, Q. Cai, and X. Zhang, "Using smart card data trimmed by train schedule to analyze metro passenger route choice with synchronous clustering," *J. Adv. Transp.*, vol. 2018, pp. 1–13, Jul. 2018.

[6] Y. Zhu, H. N. Koutsopoulos, and N. H. M. Wilson, "Inferring left behind passengers in congested metro systems from automated data," *Transp. Res. Proc.*, vol. 23, pp. 362–379, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146517302995

[7] E. Miller, G. E. Sánchez-Martínez, and N. Nassir, "Estimation of passengers left behind by trains in high-frequency transit service operating near capacity," *Transp. Res. Record, J. Transp. Res. Board*, vol. 2672, no. 8, pp. 497–504, Dec. 2018.

[8] H. E. Tan, D. W. Soh, Y. S. Soh, and M. A. Ramli, "Derivation of train arrival timings through correlations from individual passenger farecard data," *Transportation*, vol. 48, no. 6, pp. 3181–3205, Dec. 2021.

[9] S.-P. Hong, Y.-H. Min, M.-J. Park, K. M. Kim, and S. M. Oh, "Precise estimation of connections of metro passengers from smart card data," *Transportation*, vol. 43, no. 5, pp. 749–769, Sep. 2016.

[10] A. Kuppam et al., "Innovative methods for collecting data and for modeling travel related to special events," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2246, no. 1, pp. 24–31, Jan. 2011.

[11] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Internet as a sensor: A case study with special events," in *Proc. 91st Transp. Res. Board Annu. Meeting*, Washington, DC, USA, 2012, Paper 12-3365.

[12] F. C. Pereira, F. Rodrigues, E. Polisciuc, and M. Ben-Akiva, "Why so many people? Explaining nonhabitual transport overcrowding with internet data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1370–1379, Jun. 2015.

[13] Y. Zhao, Z. Ma, X. Jiang, and H. N. Koutsopoulos, "Short-term metro ridership prediction during unplanned events," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2676, no. 2, pp. 132–147, Feb. 2022, doi: 10.1177/03611981211037553.

[14] X. Wang, N. Zhang, Y. Zhang, and Z. Shi, "Forecasting of short-term metro ridership with support vector machine online model," *J. Adv. Transp.*, vol. 2018, pp. 1–13, Jun. 2018.

[15] P. Noursalehi, H. N. Koutsopoulos, and J. Zhao, "Real time transit demand prediction capturing station interactions and impact of special events," *Transp. Res. C, Emerg. Technol.*, vol. 97, pp. 277–300, Dec. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X18301797

[16] Y. Li, X. Wang, S. Sun, X. Ma, and G. Lu, "Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks," *Transp. Res. C, Emerg. Technol.*, vol. 77, pp. 306–328, Apr. 2017.

[17] M. Ni, Q. He, and J. Gao, "Forecasting the subway passenger flow under event occurrences with social media," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1623–1632, Jun. 2017.

[18] B. E. King, "Predicting national basketball association game attendance using random forests," *J. Comput. Sci. Inf. Technol.*, vol. 5, no. 1, pp. 1–14, Jun. 2017.

[19] S. Karnberger and C. Antoniou, "Network–wide prediction of public transportation ridership using spatio-temporal link–level information," *J. Transp. Geography*, vol. 82, Jan. 2020, Art. no. 102549.

[20] F. Rodrigues, S. S. Borysov, B. Ribeiro, and F. C. Pereira, "A Bayesian additive model for understanding public transport usage in special events," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2113–2126, Nov. 2017.

[21] C. Liebchen, "The first optimized railway timetable in practice," *Transp. Sci.*, vol. 42, no. 4, pp. 420–435, Nov. 2008.

[22] P. Serafini and W. Ukovich, "A mathematical model for periodic scheduling problems," *SIAM J. Discrete Math.*, vol. 2, no. 4, pp. 550–581, Nov. 1989.

[23] U. T. Zimmermann and T. Lindner, "Train schedule optimization in public rail transport," in *Mathematics—Key Technology for the Future*, W. Jäger and H. J. Krebs, Eds. Berlin, Germany: Springer, 2003, doi: 10.1007/978-3-642-55753-8_53.

[24] Y. Wang, L. Zhu, Q. Lin, and L. Zhang, "Leveraging big data analytics for train schedule optimization in urban rail transit systems," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 1928–1932.

[25] H. Guo, Y. Bai, Q. Hu, H. Zhuang, and X. Feng, "Optimization on metro timetable considering train capacity and passenger demand from intercity railways," *Smart Resilient Transp.*, vol. 3, no. 1, pp. 66–77, May 2021.

[26] A. Ceder, S. Hassold, C. Dunlop, and I. Chen, "Improving urban public transport service using new timetabling strategies with different vehicle sizes," *Int. J. Urban Sci.*, vol. 17, no. 2, pp. 239–258, Jul. 2013, doi: 10.1080/12265934.2013.776287.

[27] S. Li, R. Xu, and K. Han, "Demand-oriented train services optimization for a congested urban rail line: Integrating short turning and heterogeneous headways," *Transportmetrica A, Transp. Sci.*, vol. 15, no. 2, pp. 1459–1486, Nov. 2019.

[28] K. Li, H. Huang, and P. Schonfeld, "Metro timetabling for time-varying passenger demand and congestion at stations," *J. Adv. Transp.*, vol. 2018, pp. 1–26, Jul. 2018.

[29] H. Niu, X. Zhou, and R. Gao, "Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: Nonlinear integer programming models with linear constraints," *Transp. Res. B, Methodol.*, vol. 76, pp. 117–135, Jun. 2015.

[30] T. Schettini, O. Jabali, and F. Malucelli, "Metro scheduling for special events," *Transp. Res. Proc.*, vol. 52, pp. 147–154, 2021.

[31] Y. Q. Zhao, D. W. Li, Y. H. Yin, X. L. Dong, and S. L. Zhang, "Integrated optimization of train formation plan and rolling stock scheduling with multiple turnaround operations under uneven demand in an urban rail transit line," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–6.

[32] R. C. W. Wong, T. W. Y. Yuen, K. W. Fung, and J. M. Y. Leung, "Optimizing timetable synchronization for rail mass transit," *Transp. Sci.*, vol. 42, no. 1, pp. 57–69, Feb. 2008, doi: 10.1287/trsc.1070.0200.

[33] J. Qi, L. Yang, Z. Di, S. Li, K. Yang, and Y. Gao, "Integrated optimization for train operation zone and stop plan with passenger distributions," *Transp. Res. E, Logistics Transp. Rev.*, vol. 109, pp. 151–173, Jan. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1366554517303812

[34] J. Wang, W. Zhao, C. Liu, and Z. Huang, "A system optimization approach for trains' operation plan with a time flexible pricing strategy for high-speed rail corridors," *Sustainability*, vol. 15, no. 12, p. 9556, Jun. 2023. [Online]. Available: https://www.mdpi.com/2071-1050/15/12/9556

[35] MARTA. (2021). *Train Stations & Schedules*. [Online]. Available: https://www.itsmarta.com/train-stations-and-schedules.aspx

[36] L. McInnes, J. Healy, and S. Astels, "Hdbscan: Hierarchical density based clustering," *J. Open Source Softw.*, vol. 2, no. 11, p. 205, Mar. 2017.

[37] C. Malzer and M. Baum, "A hybrid approach to hierarchical density-based cluster selection," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2020, pp. 223–228.

Tejas Santanam received the B.S. degree in industrial engineering from Georgia Institute of Technology and the M.S. degree in analytics from the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. He is a Lecturer with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, and a Research Engineer with Beep, an autonomous mobility company. His research interests include vehicle routing in public transit and autonomous mobility.

Anthony Trasatti received the Ph.D. degree in operations research from the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. His Ph.D. dissertation was on Data-Driven Network Design of On-Demand Multimodal Transit Systems. He is a Software Engineer with Google. His main research areas are optimization, data science, and machine learning for transportation.

Pascal Van Hentenryck (Member, IEEE) is the A. Russell Chandler III Chair and a Professor with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. He is also the Director of the NSF AI Institute for Advances in Optimization (AI4OPT). His current research focuses on the fusion of AI and optimization and their applications to societal challenges in energy, mobility, supply chains and manufacturing, sustainability, and resilience. Several of his optimization systems have been in commercial use for more than 20 years.

Hanyu Zhang received the B.A. degree in mathematics from Oberlin College in 2018 and the M.S. degree in operations research from Georgia Institute of Technology in 2020, where she is currently pursuing the Ph.D. degree in machine learning with the H. Milton Stewart School of Industrial and Systems Engineering. Her research interests include time series forecasting, uncertainty quantification, and hierarchical time series prediction.