Chapter 10

# Transit Data Analytics for Planning, Monitoring, Control, and Information

**Haris N. Koutsopoulos, Zhenliang Ma, Peyman Noursalehi and Yiwen Zhu**

*Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, United States*

## Chapter Outline



## 1 INTRODUCTION

Automated data collection systems are transforming the planning, scheduling, monitoring, and operations control of transit systems. They provide operators extensive disaggregate data about the state of their system and the movement of passengers within the system (Wilson et al., 2008; Pelletier et al., 2011). The main categories of automated data sources include the following:

- Automated vehicle location systems (AVL)
- Automated passenger counting systems (APC)
- Automated fare-collection (AFC, smart cards)

**Mobility Patterns, Big Data and Transport Analytics.** https://doi.org/10.1016/B978-0-12-812970-8.00010-5

In addition to the previously mentioned main data sources, passengers also carry sensors (smartphones) that can provide detailed information about how they use the system (Calabrese et al., 2013; Goulet-Langlois et al., 2016; Zhao et al., 2018b).

Data from these sources have different characteristics both, with respect to the information they convey and their availability in time. AVL and APC data have been available for a long time and used for operations planning and scheduling (e.g., run time distributions, bus loads, etc.). AFC systems are a rather recent development. They are becoming common among transit agencies because of the convenience smart cards offer to the passengers, and the efficiency with respect to other functions (e.g., accounting). AFC systems are in general, open or closed (the functionality, mainly dictated by the agency's fare policy). Open systems require that passengers only tap in when they enter the system (e.g., MBTA in Boston). Closed systems require both, tapping in and tapping out (e.g., the transit system in Seoul, Korea). As such, they provide direct information about the origin-destination (OD) flows. Many systems are hybrid, utilizing an open architecture on the bus side and closed on the subway side (e.g., London).

The previously mentioned technologies are vehicle or station based (i.e., they are part of the agency's infrastructure). AVL data is typically available in real time at the transit control center. However, APC and AFC data is not communicated in real time yet (although technically feasible). The data is stored locally and, usually, uploaded overnight. For this reason, real-time applications of AFC data are only recently emerging.

Passengers and infrastructure are also increasingly interconnected allowing effective communication. The introduction of the mobile internet and the apps ecosystem has changed the way transit systems communicate directly with their customers. These technological advances conveniently link passengers and services. Passengers receive real-time information, for example, about bus arrivals (which alters how waiting times are traditionally estimated), updates about incidents, and provide feedback to operators about the quality of their services. Furthermore, apps and mobile sensors can provide additional information that complements the data collected from smart card systems, enhancing the development of customer centric performance metrics, measures of equity and inclusion to inform policy, and better planning of operations and services. Accelerating the adoption of such technological advances is the foundation for innovation and important means to increase public transportation effectiveness and appeal.

Fig. 1 summarizes the evolution of transit analytics as a function of the technological advances and introduction of new systems and sensors. Solid arrows indicate existing capabilities and dash arrows emerging applications.

The fusion of data from the various sources is a key element in fully capitalizing on the potential contributions to public transport. All major agency

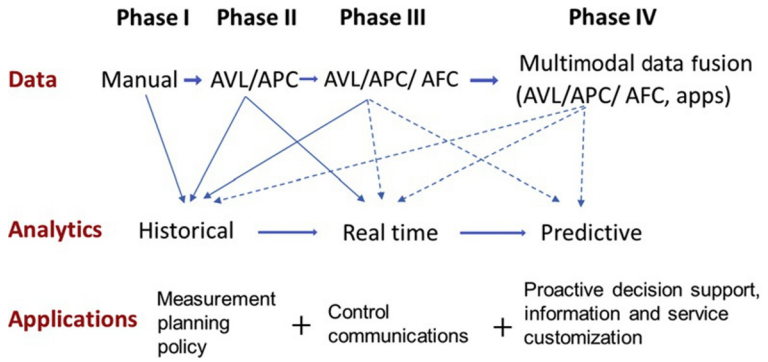


**FIG. 1** Transit analytics evolution.

functions can benefit from such data: planning, performance measurement, operations control and management, and customer information. The latter two functions take place in real time (and hence, require real-time availability of data).

Fig. 2 suggests a framework for planning and managing transit service in light of the availability of automated data. It connects the off-line and real-time functions, recognizing the role of the operator and the user of the system. The transit analytics element refers to fundamental building blocks for analysis, such as monitoring, performance evaluation metrics, and prediction.

Predictive analytics has not yet received much attention. As data from the various sources are becoming increasingly available in real time, prediction is an important capability to design better control strategies, generate

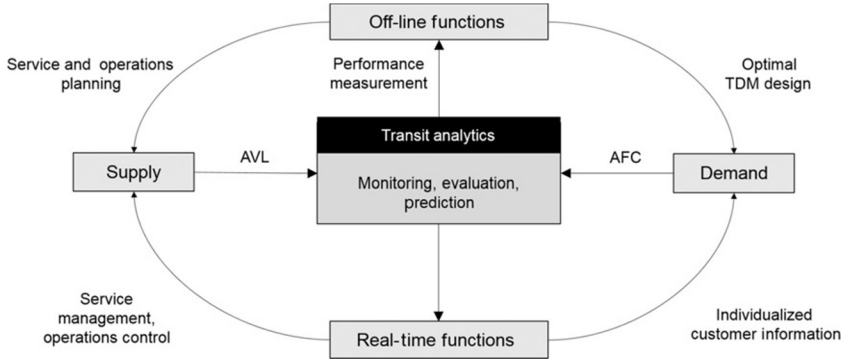


**FIG. 2** Automated data collection and key functions (Koutsopoulos et. al., 2017).

customized passenger information, deploy more dynamic services, and implement proactive transit demand management strategies (Zhao et al., 2018b; Noursalehi et al., under revision). The development of such methods is complicated due to the role of customer behavior and response to information (feedback loop on the demand side in Fig. 2).

An important building block in pursuing the various applications is the inference of trip OD matrices fusing AFC and AVL data (Munizaga and Palma, 2012; Gordon et al., 2013; Sánchez-Martínez, 2017). The complexity of this task varies, depending on the AFC system (open, closed, and hybrid). In this paper, we assume that the OD matrix has been estimated.

The objective of the paper is to provide an overview of the solution of several problems enabled by the availability of extensive transit databases and demonstrates the use of automated data to better understand and deal with the planning, monitoring, and control of transit systems. The discussion is on applications related to closed, urban heavy rail systems (subways) and focuses on the following problems.

**(a)** measuring system performance from a passenger's point of view;
**(b)** making real-time decisions to improve operations and level of service; and
**(c)** designing transit demand management strategies to increase capacity utilization.

These areas represent views of the system from three distinct perspectives with respect to time: the first one uses AFC/AVL data to evaluate past system performance. The second area deals with the problem where historical data and real-time observations are fused to inform proactive operations control. It presents a decision support platform that combines online prediction of passenger demand and simulation-based transit network performance. It is applied to generate customer information with respect to expected levels of crowding. The last area uses available information on how the system is used by passengers to design optimal transit demand management (TDM) strategies. It focuses on promotions (e.g., off-peak discounts) which aim to incentivize users to modify their choices of departure time, route, and transfer stations to reduce crowding and passenger congestion, and hence improve the utilization of available capacity.

## 2 MEASURING SYSTEM PERFORMANCE FROM THE PASSENGER'S POINT OF VIEW

The availability of AFC and AVL data affords the opportunity to monitor system performance and facilitate the development of relevant metrics to measure passenger experience, such as service reliability, crowding levels, excess waiting time due to limited capacity, etc. (Pelletier et al., 2011; Bagchi and White, 2005; Agard et al., 2006; Zhao et al., 2007). We present metrics for

measuring passenger experience with respect to: (a) system reliability; and (b) crowding.

## 2.1 The Individual Reliability Buffer Time (IRBT)

Transit operators' ability to understand and improve service reliability experienced by passengers depends upon their ability to measure it. Traditionally, largely due to lack of data, passengers' experience of reliability has not been measured directly. Several studies have shown that the commonly used on-time performance and headway regularity measurements do not capture effectively passengers' experienced reliability (Abkowitz et al., 1983; Furth and Muller, 2006; Henderson et al., 1990; Ma et al., 2013). These traditional metrics generally measure reliability at the route or line level and focus on deviations from a published timetable. For passengers, however, reliability is experienced at the level of individual journeys between OD pairs, ranging from short trips on a portion of a route, to long trips involving multiple interchanges. Furthermore, especially for high frequency services, passengers' perceptions of reliability are largely determined by the predictability of their journey times, rather than adherence to a published schedule. Passengers are unlikely to consult a schedule before starting their journeys when using high frequency services.

With the increasing availability of automatic data collection systems, it has become feasible to measure passengers' reliability experience at a detailed level. Addressing the limitations of traditional reliability metrics, the reliability buffer time (RBT) has been proposed to quantify passenger-experienced transit service reliability. RBT is the extra time passengers need to budget into their journey time to reduce to an acceptable level the likelihood of late arrival. The RBT is a function of service reliability and is typically defined as the difference between the Nth and 50th (median) percentiles of the distribution of passenger journey times for a specific OD pair and time period. It represents the additional time passengers need to budget for, in order to achieve an $N$-percent likelihood of on time arrival. The median journey time is usually used since it is not sensitive to outliers:

$$\mathrm{RBT}_{\mathrm{OD}} = (\mathrm{TT}_{N\%} - \mathrm{TT}_{50\%}) \tag{1}$$

$\mathrm{TT}_{N\%}$ and $\mathrm{TT}_{50\%}$ indicate the $N$th percentile and median of this distribution, respectively, for a specific OD pair and time period.

Fig. 3 illustrates the concept of the RBT metric. The graph shows the journey time distribution and the values corresponding to the 50th and 95th percentiles. A distribution skewed to the right with a long right-side tail indicates that passengers may experience longer journey times so the RBT increases. If travel times are consistent from day to day (for a given time period), more trips will be concentrated around the median so the difference between percentiles will be small. For $N = 95$th percentile, on average, only 1 in 20 trips (or about

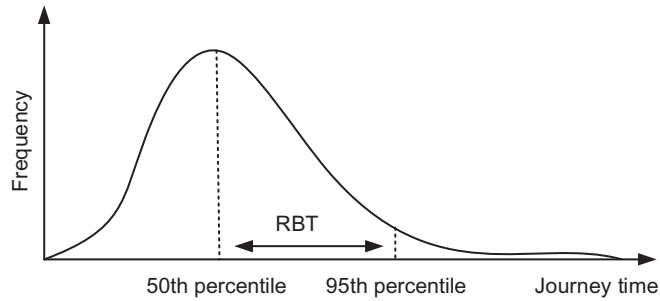


**FIG. 3** The RBT metric.

one trip per month for a commuter) will exceed the allocated time for the trip. The metric captures service reliability from the passengers' perspective: a high value suggests an unreliable service with users experiencing frequent delays (e.g., crowding or incidents) that should be taken into account when scheduling their trips. Low values suggest reliable service with journey times which are consistent from day to day.

AVL and AFC data have been used in the past for the calculation of the RBT. The AVL-based RBT, proposed by Furth and Muller (2006) and later extended by Ehrlich (2010) and Ma et al. (2014), is mainly applied to bus services. It models the journey time distribution indirectly, using vehicle headways and running times from AVL data. The AFC-based RBT was first proposed by Chan (2007) and extended by Uniman (2009). It calculates the journey time distribution directly from gate-to-gate travel times based on AFC transactions. Since it requires both, tap-in and tap-out times it can be used with closed AFC systems.

Typical AFC-based approaches use the journey time distribution across users for calculating the RBT. Since the distribution is estimated based on gate-to-gate travel times, it captures the complete journey variability. It includes operational variability caused by, for example, delays and denied boarding. However, it also includes variability contributed by the interpersonal variation among users (Wood et al., 2018; Wood, 2015). This cross-passenger variation is caused by differences in walking speed, route choices, and access and egress paths within the stations (often influenced by the degree of familiarity with the system). As a result, the AFC-based RBT can be biased as it captures both, the variation of operations and the variation among passengers.

The impact of the cross-passenger variation is illustrated in Fig. 4 using AFC data from a busy subway system for 1 month. Passengers are divided into two groups based on their frequency of using the system. One group had more than 20 trips per passenger in the analysis month, while the other had less than 10 trips in the same period. Fig. 4 compares the RBT for different periods of the day for the two groups using the 95th percentile. The RBTs for frequent riders are lower than for the infrequent ones. The overall RBT is close to the RBT for infrequent users as the demand on any given day is dominated by

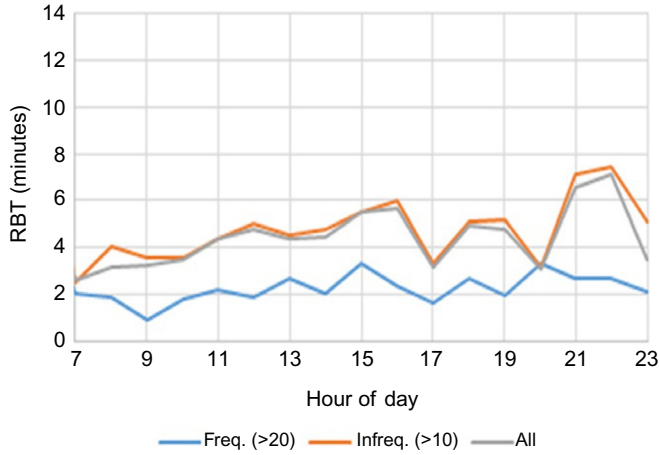


**FIG. 4** RBT for frequent and infrequent users.

infrequent users. Hence, individual variability, if not controlled for, influences the calculation of the RBT. Journey times, calculated as gate to gate times (as reported by the AFC transactions), consist of access and egress times, transfer times, waiting times, and in-vehicle times. Individual characteristics impact access, egress, and interchange times, and interpersonal variations in these components of journey time elements should not contribute to the calculation of the RBT when the metric is used for measuring system performance.

In order to deal with this problem, we have proposed to calculate the RBT at different aggregation levels of user groups: all users, specific groups, individual. Individual RBT (IBT) is, by the definition of RBT, the most accurate measure of reliability as experienced by the users. It is defined as (Wood et al., 2018):

$$\mathrm{IBT} = (\mathrm{TT}_{N\%} - \mathrm{TT}_{50\%}) \tag{2}$$

where $\mathrm{TT}_{N\%}$ and $\mathrm{TT}_{50\%}$ represent the individual's journey time percentiles.

The factors affecting the journey time distributions can be grouped into two categories: service- and passenger-related. Passenger-related factors contribute to the population journey time variability, but for the same individual, they are unlikely to vary significantly across trips. Hence, the passenger-related component of the typical individual's travel time variability should be smaller relative to the service-related component. The IBT reflects mostly the individual's service experience, whereas typical AFC-based RBT values reflect the combination of both, passenger-related and service-related factors. Consequently, RBT values may be higher than IBT values. The RBT, in effect, "over-estimates" the typical individual's buffer time.

The previously mentioned hypothesis was tested by analyzing individuals' actual travel times from the same system for a specific OD pair and for a period

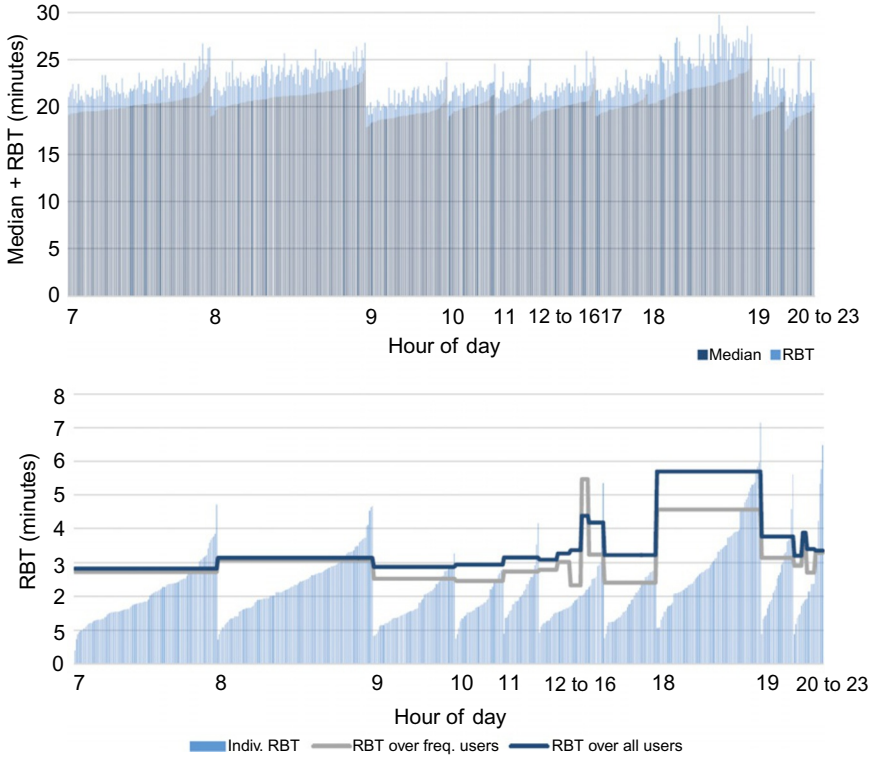


**FIG. 5** Individual RBT and median journey times.

of 2 months (Wood et al., 2018). Fig. 5 compares the individual RBT (IBT) and 95th percentile journey time values (broken down into median time and IBT) for each hour of the day for passengers who had at least 20 trips over the 2-month period. Fig. 5A orders the individuals by the median journey time, with that user's RBT added on top, while in Fig. 5B individuals are ordered by their IBT. The RBT is calculated over those 2 months for all users, and for users with at least 20 trips in the same period. Midday and night hours have been aggregated for the individual RBTs because of the small number of frequent users in these periods (the overall hourly RBTs are still shown hour-by-hour).

The individual RBT for frequent riders is lower than the RBT calculated over all passengers for different groups. The overall RBT seems to be similar to the higher end of the IBT, hence can play the role of an upper bound on buffer time. Fig. 5A further suggests that having a low median travel time does not correlate with a low RBT. While the individuals are ordered by median travel time, the total height is irregular, implying that some users with shorter median

travel times actually have a longer *median+RBT* time. Some passengers are fairly consistent, even if their trip is long, while others may have a wider distribution of journey times even if their median time is low. Those passengers with consistently longer travel times may be in such a group for reasons unrelated to service variability; for example, they could have lower walk speeds.

While the IBT as defined earlier applies at the individual level, in order to quantify the "typical" passenger's reliability experience at the system level, the Individual Reliability Buffer Time (IRBT) is defined as follows:

$$\mathrm{IRBT}_{\mathrm{OD}} = \mathrm{median}\left\{\mathrm{IBT}_{\mathrm{OD}}\right\} \tag{3}$$

$\mathrm{IBT}_{\mathrm{OD}}$ is the IBT for users traveling on the corresponding OD pair.

The IRBT can be calculated, depending on the application, at different levels of spatial aggregation (line segment, direction-line, network, and transfer pattern), in addition to the OD level. At the line level for example, it is calculated as follows:

$$\mathrm{IRBT}_{\mathrm{Line}} = \frac{\sum_{\mathrm{OD}\in\mathrm{Line}} f_{\mathrm{OD}}\mathrm{IRBT}_{\mathrm{OD}}}{\sum_{\mathrm{OD}\in\mathrm{Line}} f_{\mathrm{OD}}} \tag{4}$$

where Line is the set of same-line OD pairs and $f_{\mathrm{OD}}$ is the OD pair demand.

We demonstrate the use of the IRBT metric in a scenario were the demand in the system increased due to external shocks (Wood, 2015). In particular, the system-wide demand increased by about 9%, with one of the lines experiencing more than 25% surge. Furthermore, mostly the increase was concentrated in the peak periods, especially the morning peak. With a system already operating at capacity, this surge in demand led to additional delays for passengers, especially due to denied boarding. The IRBT metric can be used to assess the impact of the demand increase, by comparing its value for a period of 6 weeks before (base) and 6 weeks during the surge. The $\mathrm{IRBT}_{\mathrm{Line}}$ for passengers using the most affected line in the system (i.e., both origin and destination stations belong to the line) increased by 1.5 minutes during the peak periods. The IRBT for passengers transferring to the most impacted line, using one of the most congested stations in the entire system as their transfer station, increased by more than 3 minutes in the peak (about 100%). This is mainly because the transfer station was already very crowded (even before the demand increase) and close to the line's peak load point. Denied boarding and associated delays were the main contributors to this increase. The IRBT during the off-peak periods remained mostly the same.

The results indicate that the IRBT was able to capture the significant impact of the demand surge on passengers' experienced service reliability—for certain journeys and times of day.

## 2.2 Denied Boarding

Increases in ridership are outpacing capacity in many transit systems, such as Hong Kong's Mass Transit Railway (MTR), the London Underground, and the New York subway system (Zhu et al., 2017a). Crowding at stations and on trains is an important issue due to its impact on safety, service quality, and operating efficiency. Various studies have measured passengers' willingness to pay for less crowded conditions (Li and Hensher, 2011) and suggest the incorporation of the crowding disutility in investment appraisals (Haywood and Koning, 2015). Given the interest in dealing with crowding-related problems effectively, developing related measures of performance is important. Denied boarding due to overcrowding has become a major concern for many transit operators. Hence, the number of times passengers are denied boarding and how long they wait before they can board a train are often used as related performance metrics.

However, the problem of measuring the number of times a passenger is denied boarding is not trivial (some agencies conduct manual counts to collect

**TABLE 1** Approaches to Denied Boarding Estimation From AFC/AVL Data

| Approach | Data | Level | Applications | Characteristics |
|---|---|---|---|---|
| Statistical inference | AFC (tap-in and out)<br>AVL<br>Access/ egress distance/ speed | Station | Performance measurement | Needs access/ egress time distributions<br>Unsupervised learning |
| Regression | AFC (tap-in)<br>AVL<br>Denied boarding observations | Station | Performance measurement<br>Prediction | Requires actual observations of denied boarding for calibration<br>Supervised learning |
| Network assignment | OD flows<br>Path choice fractions<br>AVL<br>Capacity | Network | Performance measurement<br>Planning | Applied at the network level<br>Various crowding metrics<br>Requires capacity<br>Deterministic |

the needed data). Since it is not directly observable by current automated data sources, various approaches have been proposed to estimate it by fusing smart card (AFC) and train movement (AVL) data. These methods belong to two broad categories: (a) statistical (inference and regression models); and (b) assignment. Table 1 summarizes their main characteristics. The statistical methods are either based on unsupervised learning or use actual observations for calibration. Recently, we presented a method for the estimation of denied boarding using two data sources: (i) fare transaction records from a closed AFC system (i.e., a system where passengers both tap-in and tap-out), and (ii) train tracking data from the AVL system which provides station arrival and departure times (Zhu et al., 2017b).

We assume a closed AFC system, where the tap-in/out times of passengers are known. Train arrival/departure times at stations are also known from the train control and signaling system (AVL). Fig. 6 shows the movement of a passenger who enters the system at $t^{in}$ and exits at $t^{out}$. The estimation uses data from trips without transfers and route choice. Access time is defined as the time

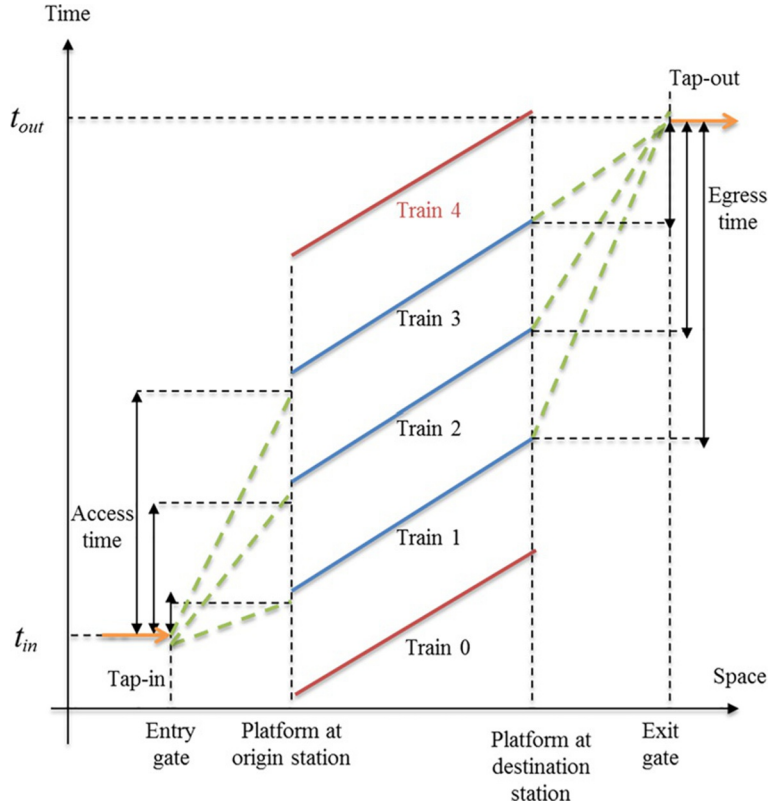


**FIG. 6** Time-space diagram for a passenger and trains (Zhu et al., 2017a).

to walk from the tap-in (entry) gate to the platform; waiting time is the time waiting on the platform; and egress time is the time to walk to the tap-out (exit) gate after alighting.

For each passenger $i$, we define the set of feasible trains. A train $j$ is feasible if it arrives at the origin station after the passenger reaches the platform and at the destination before the passenger taps out. This is a conservative definition, assuming zero access and egress times. For example, the passenger in Fig. 6 can board one of three trains (1, 2, 3).

The following notation is used in the discussion:

$t_i^{\text{in}}$: passenger $i$'s tap-in time.
$t_i^{\text{out}}$: passenger $i$'s tap-in time.
$t_i^a$: passenger $i$'s access time.
$t_i^e$: passenger $i$'s egress time.
$\tau_i^a$: minimum access time for passenger $i$ (set conservatively to zero).
$\tau_i^e$: minimum egress time for passenger $i$ (set conservatively to zero).
$M_i$: number of feasible itineraries for passenger $i$.
$M$: maximum number of feasible itineraries in the group.
$\text{DT}_{ij}$: "relative" departure time from the origin station of the $j$th train in the feasible itinerary set (after setting the tap-in time of passenger $i$ to zero) for $j \leq M_i$.
$\text{AT}_{ij}$: "relative" arrival time at the destination station of the $j$th train in the feasible itinerary set (after setting the tap-in time of passenger $i$ to zero) for $j \leq M_i$.
$\text{JT}_i$: journey time distribution for passenger $i$.
$f_a(t)$: access time distribution.
$f_e(t)$: egress time distribution.
$P_n$: probability of left behind $n$ times.

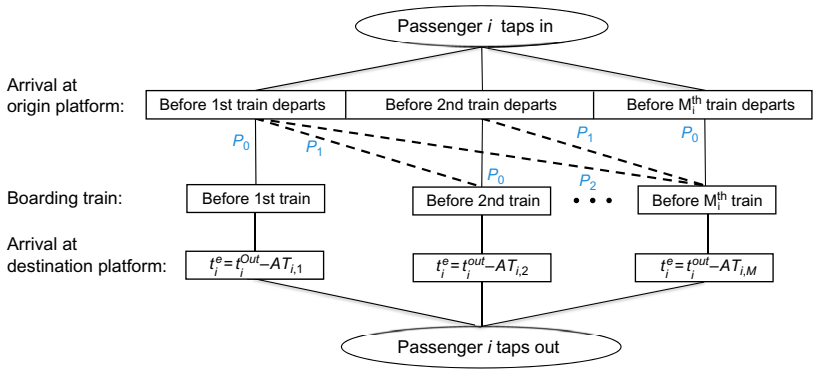


**FIG. 7** Possible passenger trajectories (Zhu et al., 2017a).

Fig. 7 illustrates all possible instances for a passenger with $M_i$ feasible itineraries. Given the feasible set, the passenger may have arrived at the platform in any of the train departure intervals (based on his/her access time) and may board the first train upon his/her arrival at the platform or be left behind if there is no available capacity. The branches with a dashed line represent the denied boarding instances. For example, even if the passenger arrives before the first train, he/she can be left behind and have to board the second (or third) train due to capacity constraints.

We assume that the access/egress speed distributions are known. They can be estimated from manual surveys conducted at stations. Zhu et al. (2017a) discuss how AFC data can be used to estimate walk speeds at stations using observations from trips that only have one feasible train. The method corrects for the bias inherent in such observations, since passengers with one feasible train may not be representative of the population. The approach was shown to be effective when compared to manual observations, and provides an alternative to collecting access/egress times data through expensive (manual) surveys.

Assuming that during a short time period, the probability distribution of denied boarding at one station is constant, the parameters of the denied boarding probability mass function can be estimated by the likelihood function of the observations. As shown in Fig. 7, the probability of passenger $i$ arriving at the origin station platform between the departures of trains $j-1$ and $j$ is as follows:

$$P\left(\mathrm{DT}_{i,j-1} \le t_i^a < \mathrm{DT}_{i,j}\right) = \int_{\mathrm{DT}_{i,j-1}}^{\mathrm{DT}_{i,j}} f_a(t)dt \quad \text{for} \quad 1 \le j \le M_i \tag{5}$$

with $\mathrm{DT}_{i,\,0} = 0$.

It can be shown that the probability of a passenger $i$ tapping out at their observed exit time can be calculated as follows:

$$\begin{aligned} L_i(\mathbf{Z}) &= \sum_{j=1}^{M_i}\sum_{k=j}^{M_i} P\left(t_i^{\mathrm{out}}, \text{board train } k, \mathrm{DT}_{i,j-1} \le t_i^a < \mathrm{DT}_{i,j}\right) \\ &= \sum_{j=1}^{M_i}\sum_{k=j}^{M_i} \int_{\mathrm{DT}_{i,j-1}}^{\mathrm{DT}_{i,j}} f_a(t)dt P_{k-j} f_e(\mathrm{JT}_i - \mathrm{AT}_{i,k}) \\ &= \sum_{j=1}^{M_i} \int_{\mathrm{DT}_{i,j-1}}^{\mathrm{DT}_{i,j}} f_a(t)dt \sum_{k=j}^{M_i} P_{k-j} f_e(\mathrm{JT}_i - \mathrm{AT}_{i,k}) \end{aligned} \tag{6}$$

where $\mathbf{Z} = [P_0, P_1, \ldots, P_{M-1}]^T$, is a vector of the parameters of the denied boarding distribution, $f_e(\mathrm{JT}_i - \mathrm{AT}_{i,k})$ is the egress time probability distribution (derived from the walk speed distribution). We assume that the maximum number of times a passenger is denied boarding equals the maximum number of feasible itineraries in the group minus one, i.e., the length of $\mathbf{Z}$ is equal to $M$.

For the whole group, assuming conditional independence among passengers, the probability of observing the journey times of all passengers in the group is as follows:

$$L(\mathbf{Z}) = \prod_{i=1}^{N} L_i(\mathbf{Z}) \tag{7}$$

The maximum likelihood formulation of the problem is based on Eq. (7) and yields the probability that a passenger, during the corresponding time period, is denied boarding $n$ times.

The model was validated with synthetic data and also applied using an extensive AFC/AVL data set from a congested subway system (Zhu et al., 2017a). The MLE problem was solved using the SciPy optimization package (Jones et al., 2001).

The data used reflect AFC transactions at two busy stations (S1, S2). The heavily used OD pair was used to estimate denied boarding probabilities at station S1. The journey time distribution for the OD pair S1–S2 is shown in Fig. 8. The journey times increase during the peak, reflecting longer waiting times due to denied boarding.

The estimated denied boarding probabilities for passengers boarding at Station 1 are shown in Fig. 9 and compared against manual surveys that took place at the station on the same day as the AFC/AVL data used for the estimation. The estimation results are similar to the survey results. In many cases, the quantity of most interest is the denied boarding rate, defined as the percentage of passengers not able to board the first train. The results from the method

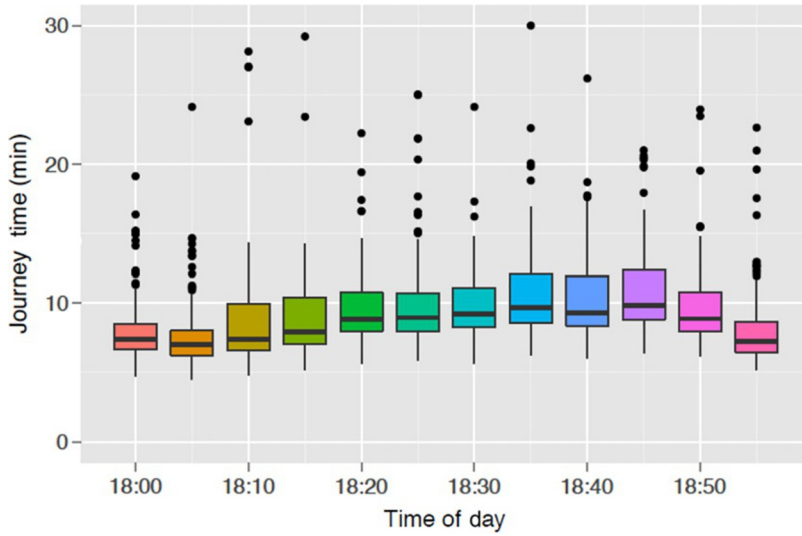


**FIG. 8** Journey time distribution.

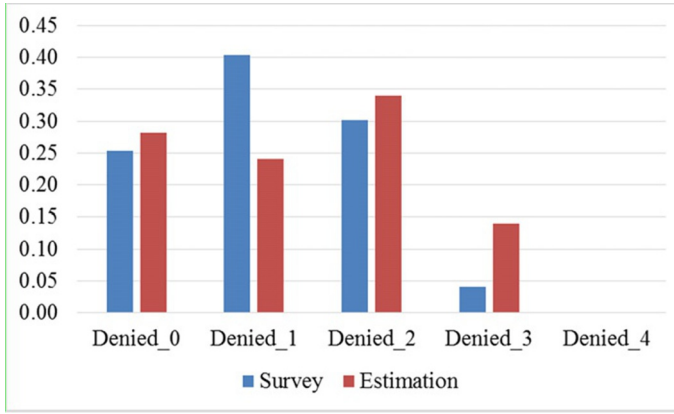


**FIG. 9** Probability of denied boarding.

presented here are consistent with the manual observations with respect to the denied boarding ratio. The differences observed in the detailed distributions can be attributed to both, the method (e.g., assumptions about access/egress speed distributions), as well as the inaccuracy of manually counting the number of passengers during the congested times (especially those waiting for several trains). This is due to the counting process itself, as well as the assumption used in the processing of the manual data, that boarding follows a first come, first served (FCFS) principle.

The agency has used the model for denied boarding analysis and estimated the corresponding probabilities for a period of 2 months. Fig. 10 illustrates the heat map of the denied boarding ratio. While, in general, the performance of the system is consistent from day to day, in a few occasions the denied boarding was higher than expected, for example, on some weekend days. Further analysis revealed that in those days, incidents that took place reduced the train frequency and increased the number of passengers who had to wait for an extra train.

## 3 DECISION SUPPORT WITH PREDICTIVE ANALYTICS

Most of the literature on the use of automated transit data has focused on the retrospective analysis of system performance (evaluation and monitoring, developing performance measures, and understanding how passengers use it). Predictive models, however, enable proactive strategy implementation to deal with abnormal conditions (incidents, surges in demand, etc.). They can also be used to generate information for the users about the upcoming state of the network.

Atypical conditions may be caused demand fluctuations, either due to day-to-day variations, or external factors such as large scale events and incidents (e.g., line or station closures). Short-term demand predictions, less than an hour

| Weekday | Week | 1800–1815 | 1815–1830 | 1830–1845 | 1845–1900 | 1900–1915 | 1915–1930 | 1930–1945 | 1945–2000 |
|---|---|---|---|---|---|---|---|---|---|
| Mon | 1 | 51% | 96% | 94% | 85% | 34% | 35% | 0% | 0% |
| | 2 | 25% | 52% | 60% | 0% | 11% | 0% | 2% | 0% |
| | 3 | 100% | 41% | 1% | 8% | 0% | 0% | 0% | 4% |
| | 4 | 36% | 88% | 95% | 90% | 15% | 0% | 0% | 2% |
| | 5 | 33% | 80% | 90% | 55% | 25% | 0% | 0% | 0% |
| | 6 | 96% | 96% | 93% | 85% | 19% | 0% | 0% | 0% |
| | 7 | 37% | 90% | 94% | 91% | 23% | 8% | 0% | 22% |
| | 8 | 25% | 77% | 95% | 92% | 0% | 0% | 0% | 4% |
| | 9 | 45% | 87% | 94% | 65% | 14% | 10% | 0% | 2% |
| | 10 | 28% | 87% | 78% | 47% | 0% | 0% | 0% | 6% |
| Tue | 1 | 3% | 1% | 0% | 2% | 12% | 0% | 0% | 0% |
| | 2 | 19% | 78% | 74% | 22% | 0% | 0% | 4% | 2% |
| | 3 | 20% | 53% | 86% | 46% | 0% | 11% | 3% | 0% |
| | 4 | 29% | 72% | 93% | 45% | 25% | 0% | 0% | 0% |
| | 5 | 42% | 80% | 80% | 37% | 12% | 0% | 0% | 4% |
| | 6 | 22% | 72% | 86% | 47% | 9% | 0% | 0% | 0% |
| | 7 | 27% | 89% | 93% | 96% | 24% | 0% | 4% | 0% |
| | 8 | 58% | 100% | 98% | 96% | 97% | 42% | 0% | 0% |
| | 9 | 39% | 95% | 94% | 85% | 37% | 39% | 1% | 0% |
| Wed | 1 | 22% | 76% | 65% | 37% | 30% | 1% | 0% | 2% |
| | 2 | 30% | 64% | 77% | 36% | 0% | 0% | 0% | 1% |
| | 3 | 30% | 79% | 95% | 47% | 0% | 0% | 0% | 0% |
| | 4 | 34% | 88% | 94% | 80% | 0% | 1% | 0% | 0% |
| | 5 | 34% | 83% | 94% | 82% | 48% | 0% | 0% | 0% |
| | 6 | 31% | 89% | 94% | 100% | 46% | 0% | 0% | 1% |
| | 7 | 39% | 92% | 93% | 93% | 28% | 0% | 0% | 0% |
| | 8 | 62% | 91% | 93% | 99% | 24% | 0% | 0% | 0% |
| | 9 | 44% | 93% | 92% | 83% | 39% | 0% | 3% | 3% |
| Thu | 1 | 25% | 78% | 83% | 43% | 0% | 5% | 3% | 4% |
| | 2 | 28% | 74% | 81% | 23% | 0% | 0% | 1% | 0% |
| | 3 | 32% | 81% | 87% | 81% | 0% | 6% | 0% | 3% |
| | 4 | 49% | 93% | 94% | 94% | 45% | 27% | 0% | 0% |
| | 5 | 48% | 94% | 94% | 92% | 51% | 0% | 0% | 0% |
| | 6 | 37% | 88% | 90% | 87% | 22% | 0% | 0% | 0% |
| | 7 | 42% | 95% | 93% | 95% | 29% | 0% | 0% | 0% |
| | 8 | 37% | 92% | 94% | 77% | 18% | 0% | 0% | 7% |
| | 9 | 36% | 95% | 100% | 95% | 49% | 0% | 0% | 0% |
| Fri | 1 | 29% | 84% | 94% | 89% | 33% | 0% | 0% | 0% |
| | 2 | 36% | 93% | 94% | 93% | 59% | 0% | 0% | 0% |
| | 3 | 62% | 94% | 96% | 95% | 80% | 43% | 0% | 0% |
| | 4 | 69% | 95% | 96% | 100% | 83% | 51% | 0% | 0% |
| | 5 | 85% | 93% | 95% | 97% | 73% | 44% | 0% | 0% |
| | 6 | 41% | 95% | 96% | 97% | 90% | 36% | 0% | 0% |
| | 7 | 51% | 94% | 94% | 96% | 93% | 34% | 0% | 0% |
| | 8 | 59% | 97% | 96% | 100% | 78% | 34% | 0% | 0% |
| | 9 | 99% | 100% | 100% | 97% | 98% | 67% | 0% | 0% |
| Sat | 1 | 0% | 0% | 4% | 2% | 6% | 0% | 1% | 0% |
| | 2 | 3% | 0% | 0% | 0% | 1% | 0% | 2% | 0% |
| | 3 | 3% | 1% | 0% | 0% | 0% | 0% | 0% | 0% |
| | 4 | 11% | 20% | 1% | 17% | 6% | 0% | 7% | 10% |
| | 5 | 0% | 48% | 60% | 41% | 1% | 0% | 9% | 18% |
| | 6 | 0% | 12% | 0% | 4% | 0% | 6% | 0% | 0% |
| | 7 | 4% | 0% | 6% | 2% | 2% | 1% | 3% | 0% |
| | 8 | 4% | 8% | 0% | 0% | 0% | 1% | 0% | 0% |
| | 9 | 15% | 15% | 70% | 11% | 16% | 9% | 1% | 4% |
| Sun | 1 | 0% | 0% | 0% | 0% | 0% | 1% | 0% | 4% |
| | 2 | 0% | 0% | 0% | 0% | 0% | 0% | 1% | 0% |
| | 3 | 0% | 3% | 0% | 0% | 0% | 0% | 0% | 7% |
| | 4 | 0% | 2% | 8% | 0% | 3% | 0% | 0% | 3% |
| | 5 | 3% | 2% | 2% | 0% | 7% | 0% | 1% | 4% |
| | 6 | 17% | 2% | 6% | 0% | 0% | 5% | 8% | 14% |
| | 7 | 1% | 1% | 9% | 0% | 0% | 0% | 2% | 0% |
| | 8 | 0% | 0% | 0% | 0% | 0% | 0% | 8% | 5% |
| | 9 | 6% | 13% | 0% | 9% | 4% | 2% | 0% | 1% |

**FIG. 10** Denied boarding rate (% of passengers unable to board the first train).

into the future, are thus important for developing anticipatory dynamic control strategies and providing useful customer information. Predictive control and information provide the opportunity for improving passenger experience by adjusting service and possibly influencing passengers' trip making choices. Operators can foresee upcoming, and most importantly, unexpected demand patterns at stations and proactively adjust service or implement crowd management strategies.

There is also a growing demand from passengers to be provided with information on the near-future service conditions of the system. The prevalence of smartphones facilitates the delivery of such information to users in real time. This dissemination of information provides the opportunity to incite cooperative behavior from the passengers while they make informed travel decisions. For example, passengers whose origin stations are predicted to experience overcrowding and are unlikely to board the first arriving train, can be advised to delay their arrival, or use an alternate route. This may alleviate the pressure on the transit system and reduce congestion on platforms and trains, resulting in better utilization of the available capacity and improved passenger experience as well.

## 3.1 Framework

We discuss a predictive decision support platform that addresses both, operations control and customer information needs. Fig. 11 illustrates the main structure of the proposed framework. The platform consists of two modules: (i) the short-term *demand prediction engine*, and (ii) the *on-line mesoscale simulation engine* (performance prediction). The framework accounts for demand–supply interactions, especially taking into account passenger response to information about the state of the system (if available). The inputs include real-time AFC transactions and train position data (and train car loads if available, e.g., from trainload sensors), as well as timetables, historical AFC transactions, and information about exogenous events.

### *3.1.1 Demand Prediction Engine*

The short-term demand prediction engine has two components: prediction of arrivals at stations, and OD prediction. Arrivals at stations are predicted in real time for the next few time periods (for example, each time period may be 15 minutes long). Information about major (planned) events that are known

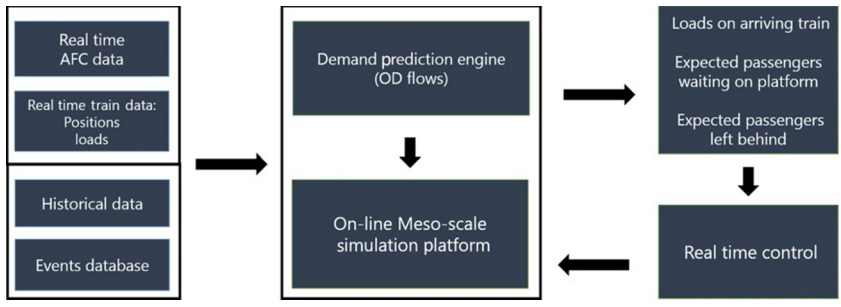


**FIG. 11** Predictive decision support platform.

to happen on that day (e.g., football games) are input to the system, and their effects are reflected in predictions.

Model specification for each station can be a time-consuming and tedious task and station-specific models do not easily account for the possible relationships between demands at other stations. For practical purposes, it is desirable to have models which capture such interactions intrinsically and deal with a large number of stations simultaneously. Dynamic factor models (DFM) are an effective way of capturing such effects intrinsically. The main characteristic of DFMs is their ability to model a large number of time series simultaneously, through a few common factors. Fig. 12 shows the 1-step (15 minutes) ahead prediction for arrivals at a busy station in the London Underground network, and compares it to the true, observed demand, as well as the historical average values (the model was estimated using data from 27 weekdays, and the leave-one-out method was used for validation).

Major social and sports events can increase demand for public transportation significantly over a short period of time. The method is able to deal with planned events, assuming enough days with such cases are available in the historical database (training set). Fig. 13 compares the 1-step ahead prediction of arrivals during a soccer game night with the observed demand. The model predicts the demand surge, as opposed to responding to it with a lag (see Noursalehi, 2017 for details).

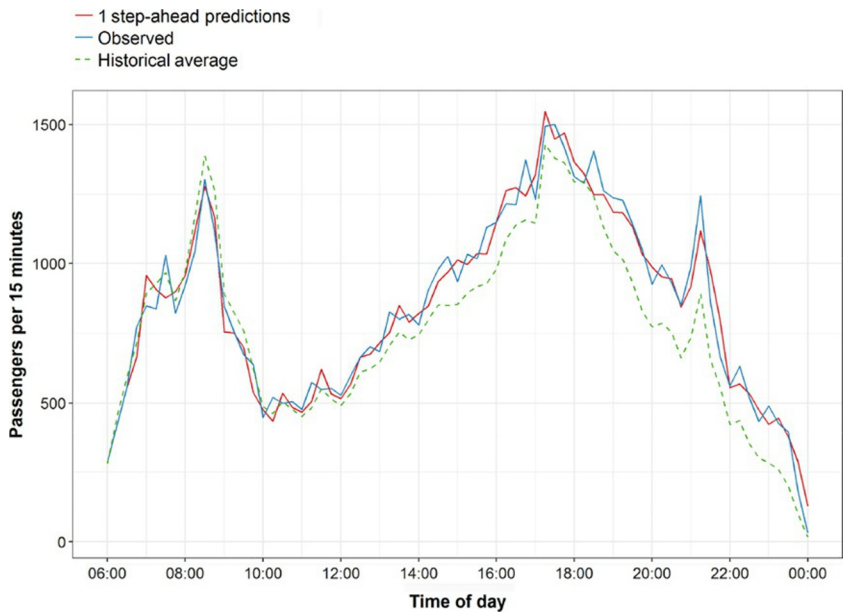


**FIG. 12** One-step ahead arrival predictions (Koutsopoulos, et al., 2017).

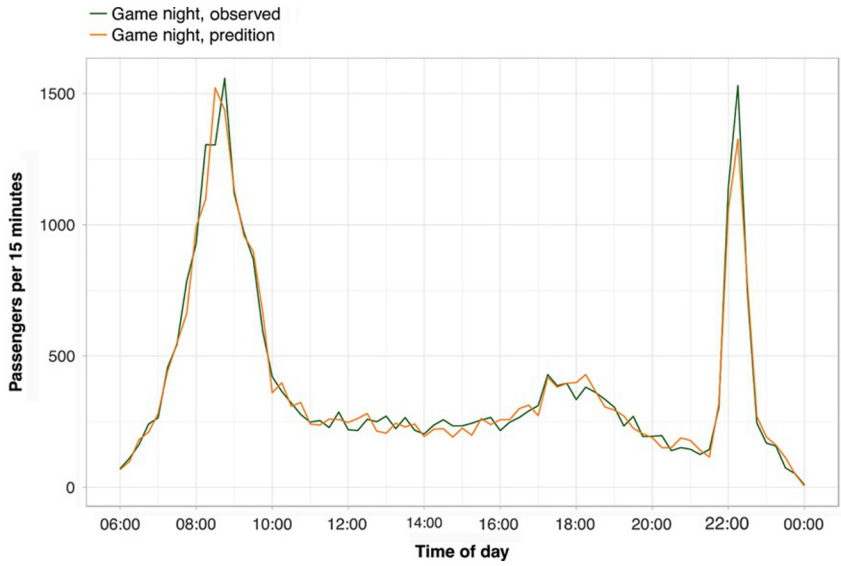


**FIG. 13** One-step ahead arrivals prediction vs observed demand during a special event (Koutsopoulos et al., 2017).

While station demand is useful for station-specific crowd management, OD flows are important as they drive many decisions related to crowd management for the whole network and customer information. An interesting characteristic of the OD prediction problem is that the true OD demand cannot be observed until all passengers entering a station at time $t$ have finished their trips at their destination station at time $t'$, $t' > t$. The observation lag, $\Delta t = t' - t$, is a function of the travel time from the origin to destination, which itself depends on many other factors, such as dwell times, other OD demands, train capacities and speeds, etc. Therefore, at each time period $t$ observations include arrivals at the origin station and the number of trips to the destination station that have been completed by that time.

OD demands from consecutive time intervals are likely correlated. There is also a correlation between OD demand and passenger arrivals at the origin station. In addition, OD flows often exhibit a nonsmooth behavior, with fluctuations in consecutive time steps. This is in part due to lower demand per time interval, as opposed to the station arrivals. Demand patterns during the peak hours are typically different from the rest of the day. The learning model should be able to capture both of those patterns.

Because of these characteristics, tree-based ensemble methods are used for this prediction task. There are two major types of tree-based ensemble methods. Random forests, which are based on the idea of bagging, and gradient boosted trees, which use boosting for combining the trees (Breiman, 2001). Fig. 14

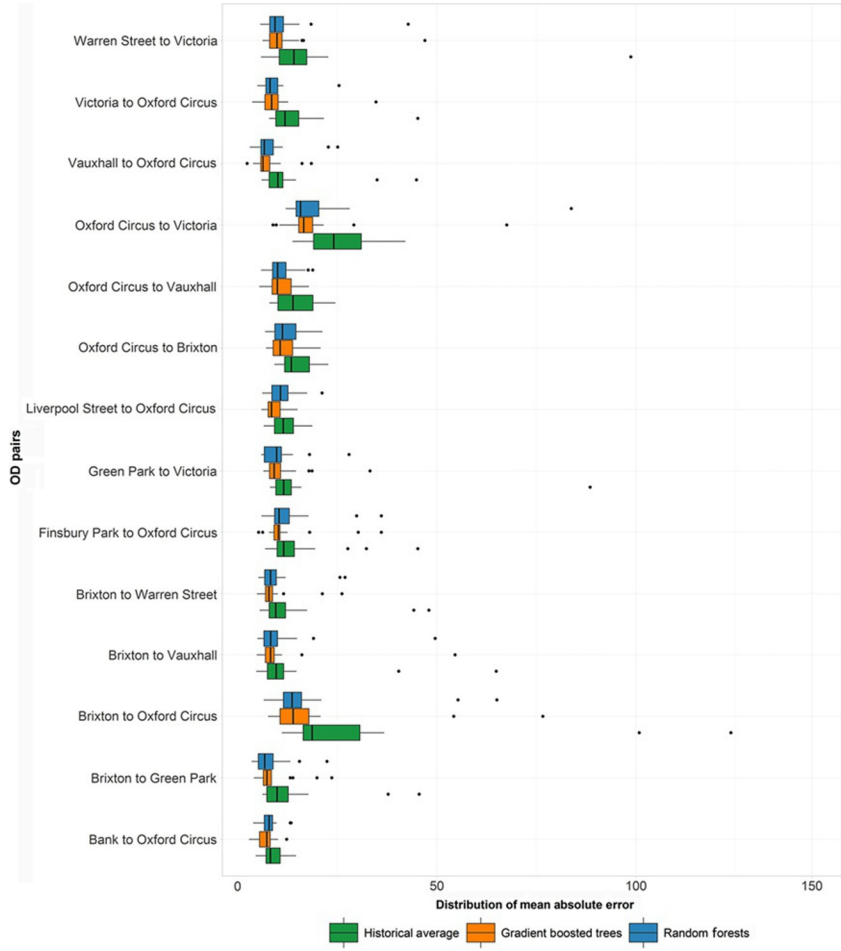


**FIG. 14** MAE distribution of 1-step ahead predictions for a few major OD pairs.

compares the predictive performance of the two models for a number of busy stations in the London Underground using the Minimum Absolute Error (MAE) as the performance metric. It also shows predictions based on the historical average. The MAE has been calculated over 27 days using the leave-one-out cross validation method (i.e., the model is estimated on 26 days, and used to predict demand for the other day).

Fig. 15 compares the 1-step ahead prediction of the demand for a busy OD pair with the historical averages. The predictions are generated using gradient boosted trees. The model is able to capture the demand fluctuations, even during the peak period, when the actual demand is much higher than the historical average.

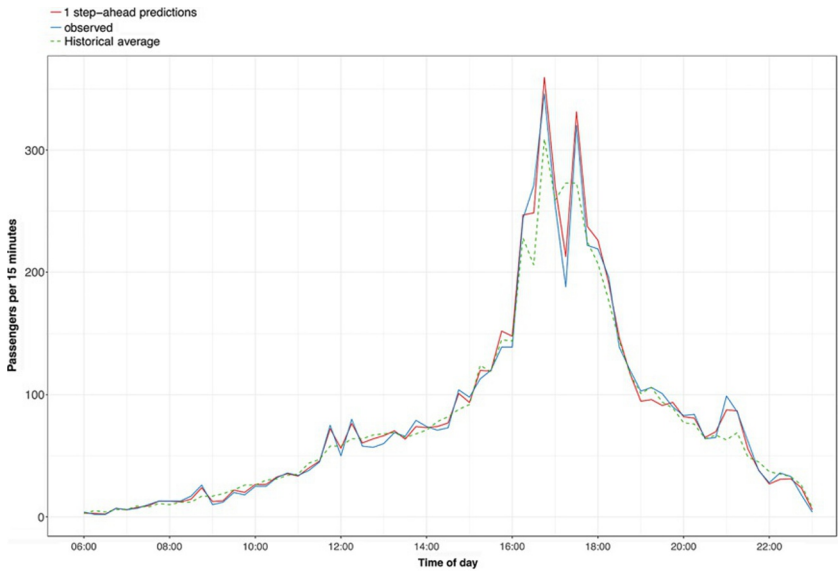


**FIG. 15** One-step ahead demand prediction for a busy OD pair.

### *3.1.2 Online Simulation Engine (Performance Prediction)*

The second building block of the predictive decision support platform is the online simulation engine for performance evaluation. It simulates train movements based on link-specific speed distributions and station-specific dwell times, taking into account a minimum buffer between consecutive trains. The model also simulates passenger arrivals at stations and assigns them to specific direction, line, and train, according to the OD predictions, and available train capacity. A passenger, if there is information at the platforms related to crowding in upcoming trains, may defer boarding the first train and wait for the next one if the information indicates more favorable conditions.

The simulation model is designed to be computationally efficient, as it is used on-line (real time). Another important design feature is its ability to self-correct based on real-time train position data that may be available (usually from the signaling system).

### *3.1.3 Demand, Supply, Information Loop*

The decision support platform captures explicitly the demand–supply interactions, especially in the presence of information. Information influences the travel decisions of passengers, impacting their pretrip decisions, or path choices while in the system. If for example, information of crowding on upcoming trains is provided to passengers waiting at the platform, the information affects the passengers' decision about boarding the current train or not. This, in turn,

changes the train loads and their available space upon arrival at the next station, which may change the predicted boarding likelihood. As such, it is important for the decision support system to incorporate the passenger response to information in its prediction of train loads. If predictions ignore this feedback loop, information may be unreliable. Unreliable information can result in erosion of trust, with users eventually ignoring it.

The problem is treated as a fixed point problem and an iterative algorithm is developed to capture passenger's response to information in the prediction framework (a similar problem also exists in the context of generating traffic information (Ben-Akiva et al., 2010). Details of the predictive information logic implemented in the decision support system can be found in Noursalehi (2017).

### 3.1.4 Implementation

In a typical implementation, the decision support system model will be used on a continuous basis, constantly updating prediction and information as data becomes available. At 7:00 am, for example, the demand prediction module outputs OD predictions for the next 30 min, in 15-minute intervals. Based on the available data the simulation engine simulates passenger arrivals and train movements for the same time interval. As new data becomes available, the simulation corrects its state based on the most recent information. At 7:05 am, it compares actual train position data (e.g., from the signaling system) with the predicted ones for that time, and updates them accordingly. With its corrected state, the simulation engine then models the system for the prediction horizon.

## 3.2 Application: Provision of Crowding Predictive Information

We use the decision support model in an application to provide passengers who are waiting on platforms with information about the upcoming trains, including their predicted arrival time and available space for passengers to board. For the purposes of this application, we assume that information about crowding is communicated to the passengers at stations using the design shown in Fig. 16. The color coding scheme translates the expected residual capacity of a train upon arrival at the stations to the likelihood of boarding. Passengers

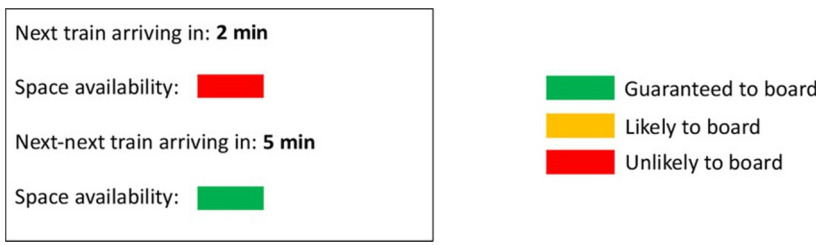


**FIG. 16** Information displayed to passengers waiting on platforms

see the predicted arrival times and predicted available space for the next two upcoming trains. Based on this information, they may defer boarding the first train if the information indicates that a subsequent train is expected to give a better experience.

Passengers make their boarding decisions based on the state of the train at the platform (observed), and the displayed predictive information for the next arriving train. It is assumed that if the passenger has been denied boarding the previous train, or had decided to not board it, will attempt to board the current train regardless of the information about upcoming trains. If there is enough space available on the train for all the passengers who are waiting at the platform to board (guaranteed boarding), then he/she will always do so. If boarding is not guaranteed, then the passenger consults the information on upcoming trains. If there is an upcoming train that arrives in less than the tolerance time (e.g., 5 minutes), and the predictive information about its crowding state is "Green" (i.e., guaranteed boarding) the passenger may decide to wait for the next train with some probability $p$. Otherwise, the passenger will always attempt to board the current one.

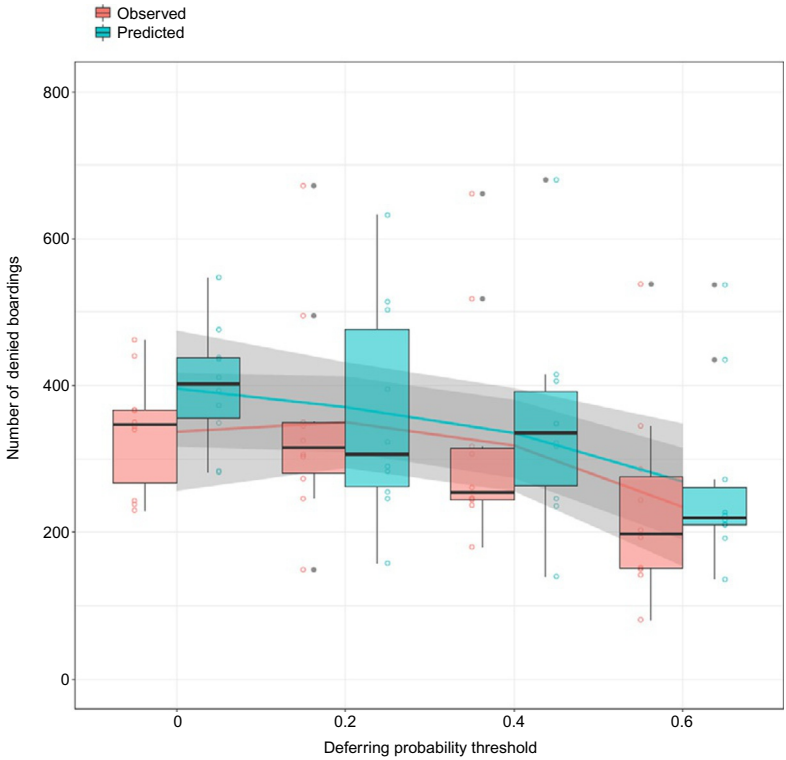


**FIG. 17** Number of left-behind passengers vs deferring probability threshold.

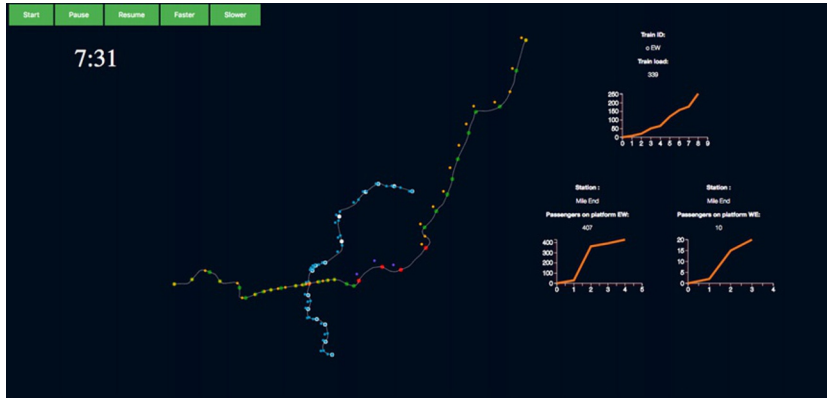


**FIG. 18** Data-driven dashboard, for station and train crowding levels.

The study area consists of 46 stations on Central and Victoria line, for a total of 2070 OD pairs. The simulation is performed for the peak morning hours, from 6 am till 11 am. Different scenarios are considered based on the assumed probability of deferring boarding. $p = 0$ corresponds to the case where there is no information. As $p$ increases, passengers become more responsive to the information, and make their boarding decisions accordingly. Fig. 17 compares the predicted number of left-behind passengers with the "observed" ones, with respect to the deferring probability $p$. As expected, as $p$ increases, meaning that passenger decisions are more responsive to the information about the status of upcoming trains, the number of left behind decreases. The run time for simulating 30 minutes of operations ranges from 40 to 70 seconds, on a computer with 16 GB of RAM and 3.4 GHz CPU. The runtime depends on the number of concurrent computations (number of updates, simultaneous train loadings, etc.), making the platform suitable for real-time applications.

An interactive visualization module has been developed. It provides animation of train movements and crowding information. Stations predicted to experience overcrowding are shown in red (otherwise in green or white, depending on the line). Trains operating at capacity are shown in purple. By selecting any train, a time series of its load to that point is illustrated. Similarly, selecting any station displays the current number of people waiting on each platform, as well as crowding levels for the previous time period (Fig. 18).

## 4 OPTIMAL DESIGN OF TRANSIT DEMAND MANAGEMENT STRATEGIES

While control strategies, like the ones described in the previous section, are used to mitigate crowding situations in real time, many agencies also deploy transit demand management (TDM) strategies to divert passengers to less crowded

routes and time periods. Well-structured transit TDM strategies can help agencies better manage the available system capacity when the opportunity and investment to expand are limited. Many transit TDM strategies currently implemented, use off-peak discounts to incentivize users to switch from peak periods. Within this context, various implementations offer station-based promotion schemes, such as the 'early bird' promotion in Hong Kong, where passengers exiting the designated stations between 7:15 am and 8:15 am receive a 25% discount (Halvorsen et al., 2016). The discount serves as an incentive for passengers to shift their travel times to an earlier period, hence resulting to less crowding during the peak, especially on lines that operate at capacity (critical links).

An issue with such discount strategies is that the promotion benefits many passengers, beyond the ones that actually switch from the peak period. Let us assume a promotion that gives a discount to passengers travelling in the prepeak period, and denote period I as the early morning period, period II the prepeak promotion period, and period III the peak period. The main groups of passengers involved in such a promotion are:

- G1: Passengers who usually exit in time period I (off-peak) and shift to time period II (prepeak) to receive the discount. The change in behavior of the passengers in this group has no impact on the crowding levels at the critical links during the peak period.
- G2: Passengers who typically exit in time period II. These passengers receive the discount without having to alter their behavior, and with no benefit to the peak hour crowding levels.
- G3: Passengers who usually exit in time period III and shift to period II but do not contribute to the congestion during the peak period because they do not use the most congested paths or critical links.
- G4: Passengers who usually exit in time period III and shift to time period II, travel through the critical links and hence, their behavioral change results in reduction of crowding on the critical links during the peak period III.

Passengers in group G4 are the ones actually targeted by the promotion (effective passengers), while passengers in groups G1–G3 are not targeted since they are not contributing to the load on critical links, but still benefit from the promotion (ineffective passengers). Hence, the cost of implementing such promotions includes the lost revenue due to ineffective passengers, in addition to the other implementation costs. An effective TDM strategy should try to minimize the passengers who receive the "free lunch" and maximize the effective passengers. However, transit systems are complex and the design of a TDM scheme, deciding when, where, and how much discount or surcharge is implemented, is not trivial.

We discuss a general framework for the optimal design of TDM promotion schemes for urban heavy rail (subway) systems. The approach is enabled by the

availability of AFC data that inform the spatiotemporal characteristics of the users, as well as accurate AVL data.

## 4.1 Framework and Problem Formulation

The TDM strategies developed by various systems vary in their design. For example, Singapore, Hong Kong, and Melbourne offer free travel or discounts to trips entering/exiting designated stations at the specified time periods. These strategies may succeed in mitigating crowding, but are not necessarily cost-efficient as discussed earlier. The congestion in urban rail systems is usually unbalanced with only several links overcrowded (critical link). Furthermore, users, based on their travel patterns and sociodemographic characteristics, may have different response to promotion strategies. Halvorsen et al. (2016) in their analysis of the Hong Kong promotion grouped passengers in six groups, based on spatiotemporal trip characteristics. The response to the promotion clearly varied among the groups. Therefore, TDM strategies can be more effective and efficient by targeting passengers who use the congested links during the peak and are sensitive to promotion.

The main design parameters of a TDM strategy include:

- *Spatial structure*: It refers to the entry/exit stations, OD pairs, route, and links or a combination of these that are targeted. Passengers entering or exiting the designated stations, or travelling between specified OD pairs, or using the designated links or routes, transfer stations, or combination of these, during the promotion time period receive a discount.
- *Temporal structure*: It refers to the time periods during which the promotion is effective. Discount may be provided during the prepeak and/or after peak periods.
- *Discount structure*: It refers to the nature of the discount and timing. A flat structure uses the same discount level across the discount time period, while a step-wise one has varied discount levels (e.g., in 15 minutes interval) within the discount time period.

The main design parameters include the place (e.g., station, OD pair, routes, and links), time and duration (e.g., discounted time period), and pricing (e.g., discount level). These parameters influence where, when, and how much discount to be offered in order to better target effective passengers who contribute to congested links during the peak time periods.

The general approach consists of two main components: assignment and optimization. The assignment updates the OD demand in response to the promotion and assigns the OD demand to the network. It outputs the load on critical links. The optimization component has as inputs the system's network topology, operational characteristics (schedule), fare-table, and the OD demand by time period in response to a specific TDM policy. It uses the various decision variables discussed earlier (promotion structures) to

formulate alternative objective functions. It identifies optimal designs considering the trade-off between performance and cost and by better targeting users (e.g., users contributing to the network performance of interest and sensitive to the TDM strategy). In the case of a discount promotion, the output of the approach is for example, users exiting station 1 between 7:15 am and 8:15 am get a 25% discount; users exiting station 2 between 7:30 am and 8:30 am get a 30% discount, etc.

For the optimal TDM design problem formulation, the following notation is used:

$X = \{x_{j\hat{t}\tau}\}$: Set of binary decision variables $x_{j\hat{t}\tau}$ ($x_{j\hat{t}\tau} = 1$ if station $s_j$ is eligible for discount $\tau$ in time period $\hat{t}$; 0 otherwise).
$\theta_{lh}$: Minimum acceptable load reduction (% of base case) of critical link $l$ in time period $h$.
$f_{lh}$: Base case (with no promotion) passenger load on link $l$ in time period $h$.
$f_{lh}^{X}$: Passenger load on link $l$ in time period $h$ given a promotion design $X$.
$\nu_x$: Cost of the promotion scheme (e.g., fare revenue loss).
$f_{lh}^{\max}$: Maximum acceptable passenger load on critical link $l$ in time period $h$.
$S = \{s_1, s_2, \ldots, s_{N_s}\}$: Set of network stations.
$L = \{l_1, l_2, \ldots, l_{N_l}\}$: Set of network links.
$T = \{t_1, t_2, \ldots, t_{N_t}\}$: Set of time periods, where the time period $t = (t, t+\Lambda]$, e.g., 7:00–7:15 am.
$B$: Budget constraint

Given the previously mentioned notation, the problem of minimizing the total load on critical links, subject to budget and minimum performance constraints, is formulated as a 0–1 integer program.

$$
\begin{aligned}
&\underset{x}{\text{Minimize}} && \sum_{x} \sum_{l \in L_c} \sum_{h \in H_c} f_{lh}^{x}, \\
&\text{subject to} \\
&&& \sum_{x} v_x x \le B, && \forall x \in \mathcal{X}, \\
&&& f_{lh}^{X} \le f_{lh}^{\max}, && \forall l \in L_c, \forall h \in H_c, \\
&&& \sum_{\hat{t}} \sum_{\tau} x_{j\hat{t}\tau} = 1, && \forall j \in S, \forall \hat{t} \in T, \forall \tau \in \Gamma, \\
&&& x \in \{0, 1\}, && \forall x \in \mathcal{X}.
\end{aligned}
\tag{8}
$$

The constraints guarantee that the load on all critical links is less than a maximum acceptable load, cost (lost revenue) does not exceed the available budget, and only one strategy is selected for each station (for details of the formulation see Ma and Koutsopoulos, 2017). Other objective functions may also be considered, for example, minimization of the cost subject to constraints on the acceptable loads at critical links (maximum acceptable load).

## 4.2 Application: Prepeak Discount Design

We apply the previously mentioned methodology to the design of a morning peak discount strategy based on exit time using the MTR subway system in Hong Kong as an example. The network consists of 90 stations and 4 of the links are operating close to capacity. The objective is to maximize the load reduction on the critical links under different budget constraints. The behavioral response, measured by the fraction of the passengers shifting to the discount periods, is based on Halvorsen (2015). Fig. 19 shows an example of the expected demand shifts assuming a discount level of 25%. Passengers who regularly exit between 8:15 am and 8:30 am and 8:30 am and 8:45 am and switch to an earlier time period, typically shift to the 8:00–8:15 am, which is the latest time period they can switch to and still receive the discount.

The problem was solved assuming that the load at the critical links should be less than 98.5% of the base load (no promotion). Different levels of budget constraints (lost revenue due to ineffective passengers) were considered. The Gurobi solver was used to solve the resulting integer optimization problem in Python (Gurobi, 2017).

Table 2 summarizes the results showing the expected load reduction as a function of the budget and the design of the promotion (discount structure and timing). Each cell in the table is the optimal solution of the corresponding design structure and budget level. The table therefore provides a portfolio of schemes that can be adopted based on given budget constraints and implementation considerations. For example, with a budget of $18 million/year, the reduction in the load of the critical links ranges from 1.59% to 2.03% compared to the base case of no TDM. By targeting a specific performance level, e.g., 1.8%, different strategies can be implemented, however at different budget levels.

Based on the results, and given the behavioral response assumptions used in the study, as expected, the performance improvement will not exceed 2.10%

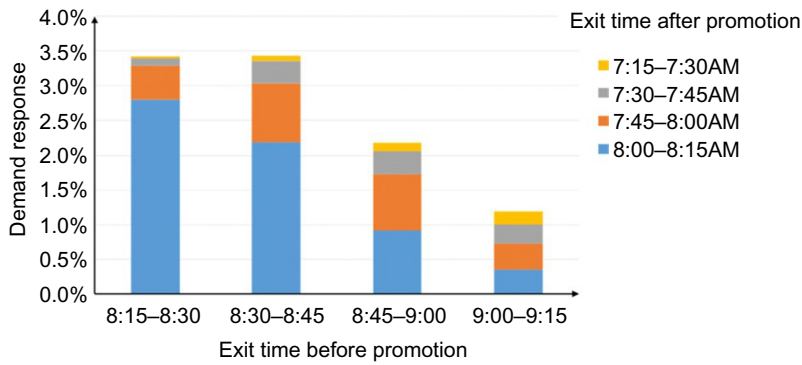


**FIG. 19** Behavioral response: % users shifting from the peak to the promotion time period.

**TABLE 2** Comparison of Promotion Design Effectiveness (Load Reduction) of Different Strategies

| | Budget Level (USD Million/Year) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Strategies** | ***8*** | ***10*** | ***12*** | ***14*** | ***16*** | ***18*** | ***20*** | ***22*** |
| Step_VT_VD | 1.71% | 1.83% | 1.92% | 2.00% | 2.01% | 2.03% | 2.05% | 2.06% |
| Flat_VT_VD | NA | NA | 1.62% | 1.71% | 1.80% | 1.82% | 1.83% | 1.83% |
| Step_FT_VD | 1.62% | 1.74% | 1.81% | 1.85% | 1.86% | 1.87% | 1.87% | 1.87% |
| Flat_FT_VD | NA | NA | 1.43% | 1.48% | 1.60% | 1.61% | 1.61% | 1.61% |
| Step_FT_FD | 1.58% | 1.67% | 1.78% | 1.79% | 1.80% | 1.80% | 1.80% | 1.80% |
| Flat_FT_FD | NA | NA | NA | 1.49% | 1.56% | 1.59% | 1.59% | 1.59% |

*Note: VT(VD), discount time (level) varies by station, FT(FD), same discount time (level) at all stations, NA, infeasible given the budget level and constraints.*

regardless of the budget invested. However, transit agencies can accomplish the maximum possible load reduction more efficiently if they design their TDM scheme carefully. The results also show that discount based promotions are not enough to reduce crowding during the peak periods. They should be considered in combination with other strategies, such as reward programs, in order to improve the overall effectiveness of transit TDM.

## 5 CONCLUSION

Automated data collection systems have the potential to better inform the critical functions of transit agencies. The fusion of AFC and AVL data enables measurement of system performance that best captures actual passenger experience, reveals the way individual customers use the system, supports predictive customer information and operations control, and informs planning functions, such as effective and cost efficient design of demand management strategies.

An important advantage of using AFC/AVL data in developing metrics for performance measurement is that such metrics capture performance from the passenger's point of view. The RBT has been used before as a service reliability measure but suffers from a number of drawbacks since it does not separate the impact of variability in operations from inter-passenger variability. In contrast, the IRBT uses individual AFC transaction data. It is based on the reliability buffer time calculated for frequent passengers separately, based on their individual records of travel times. It controls for the impact of personal variability and is effective in monitoring service reliability, measuring the impact of various operating factors (e.g., incidents).

With many systems experiencing increased demand, near capacity operations result in crowding at stations and trains. An important metric, from the passenger's point of view, related to crowding, is the probability of a passenger denied boarding at busy subway stations. The paper discusses methods that can be used to estimate this probability from AFC/AVL data. The estimated probabilities compare favorably with manual surveys, and provide crowding information at a very detailed level.

While AFC and AVL data have traditionally been used to measure past system performance, there is potential to develop predictive decision support systems for proactive real-time control of operations and information dissemination. We presented an on-line, self-correcting simulation-based decision support platform. The platform consists of a demand prediction module (based on historical and real-time AFC data) and a simulation performance module that uses AVL data to update its state representation of the transit network. The simulation models the interaction between supply and demand and captures the impact of information, predicting the near-future state of the transit network. A case study illustrates the use of the decision support system for generating information about expected crowding levels on upcoming trains, taking into account passenger response to information as well.

On the planning side the problem of designing transit TDM strategies to deal with crowding benefits from the availability of detailed AFC data. Such strategies, especially when they are based on incentives, typically suffer from inefficiencies introduced by the fact that many users may be rewarded without actually changing their behavior. We presented a framework that can be used to optimally design TDM strategies incorporating a wide range of TDM structures, as well as diverse response from various user groups (which can be identified based on their spatiotemporal characteristics as revealed by the AFC data). The case study demonstrates the applicability of the proposed method. The results also show that discount-based promotions are not enough to reduce crowding during the peak periods. They should be complemented by other strategies, such as reward programs, in order to improve the overall effectiveness of transit TDM.

## ACKNOWLEDGMENTS

The authors would like to thank the various transit agencies for their support and data sharing. We would also like to thank Anne Halvorsen and Daniel Wood for their work on the individual RBT metric and colleagues in the Transit Lab for many helpful discussions.

## REFERENCES

Abkowitz, M., Slavin, H., Waksman, R., Englisher, L., Wilson, N.H.M., 1983. Transit service reliability. Tech. report, U.S. Dept. of Transportation.

Agard, B., Morency, C., Trépanier, M., 2006. Mining public transport user behaviour from smart card data. IFAC Proc. Vol. 39 (3), 399–404.

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transp. Policy 12 (5), 464–474.

Ben-Akiva, M., Koutsopoulos, H.N., Antoniou, C., Balakrishna, R., 2010. Traffic Simulation with DynaMIT (Chapter 10). In: Barcelo, J. (Ed.), Fundamentals of Traffic Simulation. Springer-Verlag, New York, pp. 363–398.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transp. Res. Part C: Emerg. Technol. 26, 301–313.

Chan, J., 2007. Rail Transit OD Matrix Estimation and Journey Time Reliability Metrics Using Automated Fare Data Matrix. Master of Science in Transportation thesis, Massachusetts Institute of Technology.

Ehrlich, J.E., 2010. Applications of Automatic Vehicle Location Systems Towards Improving Service Reliability and Operations Planning in London. Master of Science in Transportation thesis, Massachusetts Institute of Technology.

Furth, P., Muller, T., 2006. Service reliability and hidden waiting time: insights from AVL data. Transp. Res. Rec. 1955, 79–87.

Gordon, J., Koutsopoulos, H.N., Wilson, N.H.M., Attanucci, J., 2013. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transp. Res. Rec. J. Transp. Res. Board 2343, 17–24.

Goulet-Langlois, G., Koutsopoulos, H.N., Zhao, J., 2016. Inferring patterns in the multi-week activity sequences of public transport users. Transp. Res. Part C: Emerg. Technol. 64 (Suppl. C), 1–16.

Halvorsen, A., 2015. Improving Transit Demand Management With Smart Card Data: General Framework and Applications. Master of Science in Transportation thesis, Massachusetts Institution of Technology.

Halvorsen, A., Koutsopoulos, H.N., Lau, S., Au, T., Zhao, J., 2016. Reducing subway crowding: analysis of an off-peak discount experiment in Hong Kong. Transp. Res. Rec. J. Transp. Res. Board 2544, 38–46.

Haywood, L., Koning, M., 2015. The distribution of crowding costs in public transport: new evidence from Paris. Transp. Res. Part A: Policy Pract. 77, 182–201.

Henderson, G., Adkins, H., Kwong, P., 1990. Toward a passenger-oriented model of subway performance. Transp. Res. Rec. 1266, 221–228.

Jones, E., Oliphant, E., Peterson, P., SciPy: open source scientific tools for Python. 2001, http://www.scipy.org/, Accessed 18 February 2018.

Koutsopoulos, H.N., Noursalehi, P., Zhu, Y., Wilson, N.H.M., 2017. Automated data in transit: Recent developments and applications.Proceedings of the 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS2017), pp. 604–609.

Li, Z., Hensher, D.A., 2011. Crowding and public transport: a review of willingness to pay evidence and its relevance in project appraisal. Transp. Policy 18 (6), 880–887.

Ma, Z., Koutsopoulos, H.N., 2017. In: Optimal design of transit demand management strategies.IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan.

Ma, Z., Ferreira, L., Mesbah, M., 2013. A framework for the development of bus service reliability measures.Proceedings of Australian Transport Research Forum.

Ma, Z., Ferreira, L., Mesbah, M., 2014. Measuring service reliability using automatic vehicle location data. Math. Prob. Eng. 2014, 1–12.

Munizaga, M.A., Palma, C., 2012. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago Chile. Transp. Res. Part C: Emerging Technol. 24, 9–18.

Noursalehi, P., 2017. Decision support platform for urban rail systems: real-time crowding prediction and information generation. (Ph.D. dissertation), Department of Civil and Environmental Engineering, Northeastern University.

Noursalehi P., Koutsopoulos H.N., Zhao J., Real time transit demand prediction capturing station interactions and impact of special events, *Transp. Res. C*, under revision.

Gurobi Optimization, I, 2017. Gurobi Optimizer Reference Manual. http://www.gurobi.com.

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. Part C: Emerg. Technol. 19 (4), 557–568.

Sánchez-Martínez, G.E., 2017. Inference of public transportation trip destinations by using fare transaction and vehicle location data. Transp. Res. Rec. J. Transp. Res. Board 2652, 1–7.

Uniman, D., 2009. Service Reliability Measurement Framework Using Smart Card Data: Application to the London Underground. Master of Science in Transportation thesis, Massachusetts Institute of Technology.

Wilson, N.H.M., Zhao, J., Rahbee, A., 2008. The potential impact of automated data collection systems on urban public transport planning. In: Schedule-Based Modeling of Transportation Networks: Theory and Applications, Operations Research/Computer Science Interface Series. Springer, Boston, MA, pp. 75–97.

Wood, D., 2015. A Framework for Measuring Passenger-Experienced Transit Reliability Using Automated Data. Master of Science in Transportation thesis, Massachusetts Institute of Technology.

Wood, D., Halvorsen, A., Koutsopoulos, H.N., Wilson, N.H.M., 2018. In: Measuring passengers' reliability experience from AFC data.INSTR2018 (extended abstract), 7th International Conference on Transport Network Reliability.

Zhao, J., Rahbee, A., Wilson, N.H.M., 2007. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Comput. Aided Civil Infrastruct. Eng. 22, 376–387.

Zhao, Z., Koutsopoulos, H.N., Zhao, J., 2018b. Individual mobility prediction using transit smart card data. Transp. Res. Part C: Emerging Technol. 89, 19–34.

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H.M., 2017a. A probabilistic passenger-to-train assignment model based on automated data. Transp. Res. Part B: Methodol. 104, 522–542.

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H.M., 2017b. Inferring left behind passengers in congested metro systems from automated data. Transp. Res. Part C: Emerg. Technol. 94, 323–337.

## FURTHER READING

Goulet-Langlois, G., Koutsopoulos, H.N., Zhao, Z., Zhao, J., 2017. Measuring regularity of individual travel patterns. IEEE Trans. Intell. Transp. Syst. 99, 1–10.

Zhao, Z., Koutsopoulos, H.N., Zhao, J., 2018a. Detecting pattern changes in individual travel behavior: a bayesian approach. Transp. Res. Part B: Methodol. 112, 73–88.