Transportation Research Record

DATA DRIVEN REAL-TIME PLATFORM CROWDING PREDICTION USING AUTOMATED FARE COLLECTION AND VEHICLE LOCATION DATA IN URBAN RAILWAY SYSTEMS

--Manuscript Draft--

Full Title:	DATA DRIVEN REAL-TIME PLATFORM CROWDING PREDICTION USING AUTOMATED FARE COLLECTION AND VEHICLE LOCATION DATA IN URBAN RAILWAY SYSTEMS
Abstract:	Growing urbanization causes increase in crowding in many urban railway systems. Providing real-time crowding information would enable informed travel decisions and encourage cooperative behavior of passengers, as well as improve operating efficiency and safety. However, the problem of real-time crowding prediction is not trivial due to the unavailability of ground-truth crowding data. This paper proposes a data-driven approach for real-time platform crowding prediction in urban railway systems using automated fare collection (AFC) and automated vehicle location (AVL) data. The methodology aims to use the existing data-driven denied boarding estimation method to generate ground truth information and thus, extending the denied boarding framework into having predictive capabilities. The distribution of denied boarding probability is estimated by adopting the structured mixture model proposed in Ma et.al (19) using AFC and AVL data. Using these estimates, the real-time crowding prediction is formulated as a supervised learning problem with the denied boarding estimate as the response variable and explanatory variables extracted from AFC and AVL data, including characteristics of demand, operations, and disruptions. A Case study using Mass Transit Railway (MTR) data in Hong Kong is conducted to illustrate the effectiveness of the proposed methodology. The model is able to provide accurate platform crowding prediction in short-term using explanatory variables such as transfer demands and headways.
Manuscript Classifications:	Data and Information Technology; Urban Tranportation Big Data ABJ30SC; Big Data; Transformative Transit Data AP000; Smartcard; Transit Capacity and Quality of Service AP015; Capacity; Customer; Measuring; Service; Public Transportation; Rail Transit Systems (not Light, Freight, Commuter, or Highspeed) AP065; Passenger Information; Subway; Performance
Manuscript Number:	
Article Type:	Presentation and Publication
Order of Authors:	Kerem Sinan Tuncel, PhD.
	Haris Koutsopoulos, PhD
	Zhenliang Ma, PhD

DATA DRIVEN REAL-TIME PLATFORM CROWDING PREDICTION USING AUTOMATED FARE COLLECTION AND VEHICLE LOCATION DATA IN URBAN RAILWAY SYSTEMS

Kerem Sinan Tuncel
Department of Mechanical and Industrial Engineering
Northeastern University, Boston, MA, 02115
Email: tuncel.k@husky.neu.edu

Haris N. Koutsopoulos
Department of Civil and Environmental Engineering
Northeastern University, Boston, MA, 02115
Email: h.koutsopoulos@northeastern.edu

Zhenliang (Mike) Ma
Institute of Transport Studies
Department of Civil Engineering
Monash University, Clayton, VIC, Australia, 3800
Email: Mike.Ma@monash.edu

Word Count: 5102 words + 1 tables = 5252 words

Submitted 08/01/2019

ABSTRACT
Growing urbanization causes increase in crowding in many urban railway systems. Providing real-time crowding information would enable informed travel decisions and encourage cooperative behavior of passengers, as well as improve operating efficiency and safety. However, the problem of real-time crowding prediction is not trivial due to the unavailability of ground-truth crowding data. This paper proposes a data-driven approach for real-time platform crowding prediction in urban railway systems using automated fare collection (AFC) and automated vehicle location (AVL) data. The methodology aims to use the existing data-driven denied boarding estimation method to generate ground truth information and thus, extending the denied boarding framework into having predictive capabilities. The distribution of denied boarding probability is estimated by adopting the structured mixture model proposed in Ma et.al (19) using AFC and AVL data. Using these estimates, the real-time crowding prediction is formulated as a supervised learning problem with the denied boarding estimate as the response variable and explanatory variables extracted from AFC and AVL data, including characteristics of demand, operations, and disruptions. A Case study using Mass Transit Railway (MTR) data in Hong Kong is conducted to illustrate the effectiveness of the proposed methodology. The model is able to provide accurate platform crowding prediction in short-term using explanatory variables such as transfer demands and headways.
Keywords: real-time crowding prediction, urban railway systems, AFC and AVL, supervised learning, denied boarding probability

INTRODUCTION

Increases in ridership are outpacing capacity in many large urban rail transit systems, such as Hong Kong's Mass Transit Railway (MTR), the London Underground, and the New York subway system (*1; 2*). Crowding at stations and on trains is a concern due to its impact on safety, service quality, and operating efficiency. Various studies have measured passengers' willingness to pay for less crowded conditions (*3*) and suggested incorporating the crowding disutility in investment appraisals (*4*). Compared to congestion management in traffic, including adaptive traffic control, ramp metering, congestion pricing, etc. (*5; 6*), crowding management in transit is still evolving. Adding capacity, such as building new lines and shortening headways, to deal with the increased demand is often difficult, especially in the short term. Travel demand management (TDM) leading to better utilization of available capacity is a promising alternative to deal with this challenging issue.

Many agencies have implemented or tested TDM strategies in transit systems, which usually take the form of of incentives and penalties, such as free trips, off-peak discounts, peak surcharge (e.g. Hong Kong, London, Melbourne, Singapore, Sydney, Tokyo, Washington, D.C.); working with employers to encourage company-specific programs (e.g. Singapore); lottery/rebate rewards (e.g. San Francisco's PERKS program, Singapore's Travel Smart program) (*7; 8*). Some studies considered tactical planning methods to reduce the uneven distribution of passenger loads on trains, including optimizing the train stop positions at stations considering passenger distribution on platforms and station entries (*9*), and using one-way gates on platforms to control passenger car boarding choices (*10*).

With the prevalence of advanced technologies in automated data collection in transit systems, such as automated fare collection (AFC) and automated vehicle location data (AVL), many cities have been providing real-time vehicle arrival time information at stations, on websites or in mobile applications. Many studies have proposed methods to predict bus and train arrival times using various available data (*11;12*), and some studies have shown the positive effect of real-time information on passengers' perceived waiting times, safety and security, impacts of service disruptions, and general satisfaction (*13, 14, 15*). Several cities also provide crowding information (e.g. BART in San Francisco, JR East in Tokyo) (*16; 17*). 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. However, very few studies have been reported on real-time crowding prediction. Recently, Jenelius (*18)* formulates the car-specific metro train crowding prediction problem using real-time train load data.

The paper focuses on the crowding prediction on platforms in urban railway systems. One of the challenges for crowding prediction is that platform crowding cannot be directly observed Examples of ways to collect crowding data are manually counting the number of passengers on the platform or video processing. Both methods are expensive both in terms of time and resources. Also, both methods are not reliable since manual counting is prone to human error especially in crowded systems and video processing requires setting cameras in a way that the whole platform can be viewed which is not always possible. This research applies the method presented in (*19*) to generate the denied boarding distribution using AFC and AVL data, and then formulates a

supervised learning model to predict the crowding situation (e.g. average waiting time) on platform. The methodology is computationally efficient. Thus, it can be implemented by metro system operator to provide real time crowding information. For example, for every fifteen minutes, the model can generate crowding information for the next fifteen minutes.

Rest of the paper is organized as follows: The Methodology section describes the crowding prediction framework using denied boarding information as well as the formulation of the supervised learning problem. It also discusses various explanatory variables used for prediction. The Case Study section presents a real world application that is used to validate the proposed framework. Finally, the Conclusion section provides the closing remarks.

METHODOLOGY

A. Crowding Prediction Framework

The proposed methodology uses of data-driven methods for denied boarding estimation *(2;12)* in order to generate the training data for real-time crowding prediction. These methods are able to estimate the probability distribution of the number of times passengers are denied boarded for a given discrete time interval. These methods are validated against survey results. Various performance measures for crowding such as; denied boarding rate (the probability of not being able to board the first train upon arrival at the platform), expected number of trains to wait, expected waiting time can be derived from the denied boarding probabilities. Therefore, the prediction framework utilizes these denied boarding estimates in place of direct observations since they are able to convey crowding information. This methodology can be seen as the extension of the existing data-driven denied boarding estimation methods which completes the loop by adding predictive capability to the existing tools.

When the denied boarding information is used as the training set for platform crowding, the crowding prediction problem becomes a supervised learning problem where the denied boarding estimates serve as the response variable. On the other hand, various station and operations related features, such as demand and headways, serve as the predictor variables which will be used to extract the useful relationships and patterns required for accurate predictions. The predictor variables to be used should be readily available, for example, attributres like station demands can be easily extracted from AFC and AVL data. The framework of the proposed methodology is shown in Figure 1. Time is divided into time intervals of length Δt (e.g 15 minutes). Predictions take place in discrete time intervals for example, every fifteen minutes. At time t, the problem becomes predicting the value of the response variable (e.g expected waiting time) for the time interval $[t, t + \Delta t]$ by using only observed (recent) smart card activity and train movement information. Using past denied boarding information allows to predict platform crowding without relying on direct observations.

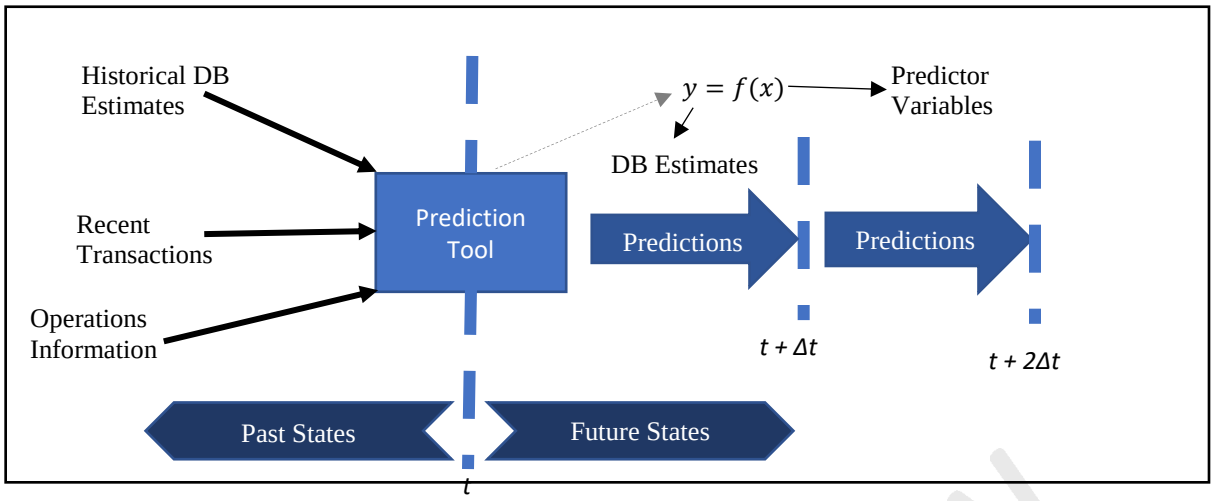

Figure 1. Framework of the crowding prediction methodology

B. Problem Formulation

The setting of the crowding prediction methodology has three major steps; feature selection, choosing the response variables and model selection. The feature selection part involves identifying the correct set of explanatory variables for the problem. This is an important step since gaining insights about the factors which have an impact on crowding is another motivation for this work. Similarly, the selection of the response variable involves identifying which measure to use as the indicator for crowding. Since denied boarding estimates are being used, it is possible to derive multiple performance metrics for crowding. For the purposes of this work, denied boarding rate and expected waiting time are considered as potential response variables. The model selection part involves selecting the model and training its parameters to achieve highest accuracy.

Feature Selection

For the feature selection part, the explanatory variables considered can broadly be categorized in three categories; demand-related features, operations-related features and incident related features. Note that, all of the explanatory variables are system attributes that are known and directly observable from AFC and AVL data. Demand and operations related features are obvious choices but since incidents occurring in the system may have large impact on the crowding levels some variables indicating such occasions are also considered.

Demand related features involve the recent transactions regarding passengers entering in the station of interest as well as upstream stations within the same line and in transferring lines using the station under consideration as a transfer place, feed in the crowding at the station of interest. In addition to transfer demand, the demand in the upstream stations of the same line also have a role in the crowding as they serve as indicators of used capacity. The higher the demand in upstream stations, the higher the train load when the train arrives at the platform. This, in turn, will result in higher probability of denied boarding and therefore higher crowding at the platform. The upstream demands can be included either as aggregated or disaggregated features. If they are included as aggregated features, then each upstream branch is represented as a single feature as the sum of individual demands at each station in that branch. On the other hand, if disaggregated features are used, each upstream station is represented as a separate explanatory variable. The disaggregated features may generate a high number of explanatory variables, which may impact the models predictive accuracy. On the other hand, some of the station specific variability can be

lost if aggregated features are used. Both versions are tested in the case study in order to asses the impact of aggregate/disaggregate demand features.

For the operations related features, the aim is to identify a set of features which are able to capture recent system performance levels. *Average headway* and the *number of trains* arriving at the platform (both from the same line and transferring lines) in the current time interval ([t - Δt, t]) are used to capture the recent service level. Number of trains are also used because *Average Headway* by itself is not enough to indicate the service level. For example, the average headway can be the same for seven trains arriving at the platform in the past fifteen minutes and two trains arriving at the platform. But the service level for these two scenarios would be different and the impact on crowding would also be very different. Furthermore, a set of features referred to as *left-over demand*, is used to capture the amount of demand that may not have been served in the last time interval and will be carried over to the next time interval. The *left-over demand* is calculated both for the station of interest and for upstream stations as follows;

$$L_{S,t} = A_{S,t} * \tau \quad (1)$$

$$L_{S,t_{upstream}} = A_{S,t_{upstream}} * H_{last} \quad (2)$$

where, $L_{S,t}$ and $L_{S,t_{upstream}}$ are the left-over demands for the station of interest and the upstream station, $A_{S,t}$ and $A_{S,t_{upstream}}$ are the average passenger arrival rates for the station of interest and the upstream station, H_{last} is the headway of the last train arriving from transfer line and τ is the time between the arrival of the last train and the beginning of the new prediction interval. $L_{S,t_{upstream}}$ aims to represent the number of passengers transferring to the station of interest with the last train that arrives at the station. These passengers have the highest probability of being carried over to the next time interval. Similarly, $L_{S,t}$ aims to capture the number of passengers arriving at the platform after the last train left the platform. Note that, the left-over demands and the average arrival rates are time-specific and are calculated separately for each time interval. These passengers are the passengers who will be added to the demand of the next time interval. The definitions used in left-over demand calculations are shown graphically in Figure 2.

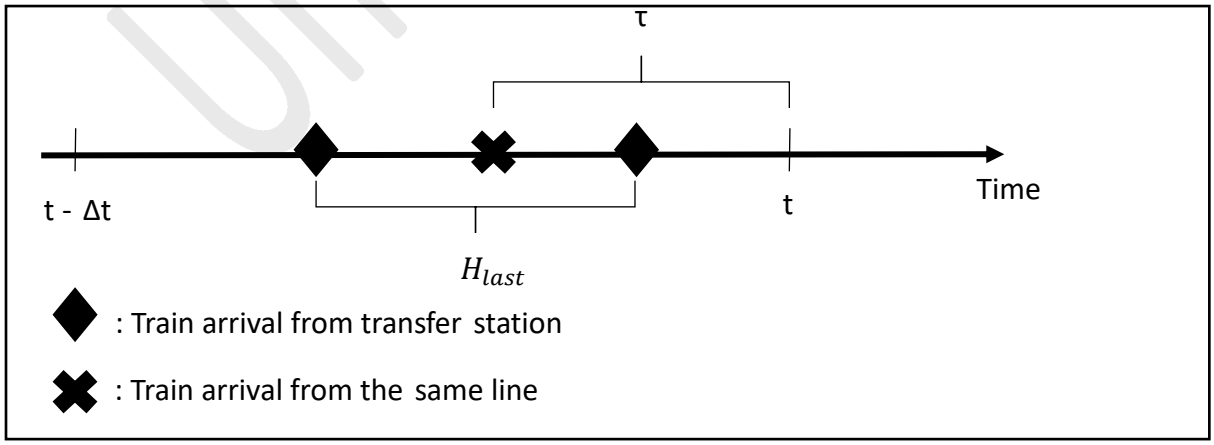

Figure 2. Explanation of left-over demands

The incidents occuring in the system are also a factor to consider within the prediction framework since these disruptions have impact on the crowding. Therefore, it is necessary to add predictor variables into the model to capture the effect of incidents. For that purpose, a predictor variable called *accumulation* is introduced. *Accumulation* is defined as the total number of passengers within the system at a given point in time which is calculated as follows;

$$A(t) = \int_0^t a(t)dt - \int_0^t e(t)dt \quad (3)$$

where, $A(t)$ is the accumulation at time t, $a(t)$ is the arrival rate of passengers (at all staions in the system) at time t and $e(t)$ is the exit rate of passengers at all stations in the system.

It is expected that the number of passengers in the system will be increased to atypical levels during a disruption. Thus, the accumulation is able to capture such distruptions within the system.

Selection of Response Variables

It is possible to derive different performance measures for crowding using denied boarding estimates. Some examples are; denied boarding rate, expected waiting time, expected number of trains to wait and denied boarding probabilities. These performance metrics provide information about the crowding at the platform in different ways, and all can possibly be used as the response variable for the prediction model. Expected waiting time and the denied boarding rate are considered as the potential response variables in the proposed methodology, since both of these metrics provide univariate response variables (as opposed to using denied boarding probabilities directly which is a vector of response variables, adding an extra layer of complexity to the problem). Moreover, both of the proposed response variables are practical and easy to interpret variables as opposed to expected number of trains to wait which may be useful for planning purposes but may not be very informative for communicating to passengers.

As mentioned earlier, denied boarding rate at a time interval refers to the probability that a passenger will experience denied boarding at least once during that time interval. Using the denied boarding probabilities, it is straightforward to calculate the denied boarding rate using the following equation;

$$DB\ Rate = 1 - P(No\ denied\ Boarding) = 1 - P_0 \quad (4)$$

Similarly, the expected waiting time for a time interval can be calculated using the denied boarding estimates and the average headway within the time interval. The expected waiting time is can be calculated by **Equation 5**, where H refers to the average headway and P_k refers to the probability of being denied boarded k times.

$$E(Waiting\ Time) = \frac{H}{2} * P_0 + \left(\frac{H}{2} + H\right) * P_1 + \left(\frac{H}{2} + H + H\right) * P_2 + \cdots + \quad (5)$$

Even though variables convey useful information about crowding, expected waiting time is more representative of the whole denied boarding probability distribution whereas the denied boarding rate provides information that is more compressed. This may cause issues, especially in situations where there are high probabilities of denied boarding multiple times. For example, two scenarios; a scenario where the probability of denied boarding once is equal to one and another

scenario where the probability of denied boarding twice is equal to one. Both scenarios would have a denied boarding rate equal to one, although they represent very different situations from the point of view of passenger experience. Thus, the denied boarding rate feature would not be able to capture the difference between these two cases whereas the difference would be represented by the expected waiting time metric. This phenomenon is illustrated in Figure 3. The figure shows two scatterplots for demand versus denied boarding rate and expected waiting time respectively. Both plots show a behavior similar to a typical queueing system; after the demand exceeds a certain threshold, its performance deteriorate dramatically. For the denied boarding rate, it becomes harder to differentiate as denied boarding rate gets closer to one since there are many observations hitting the upper bound or are close to the upper bound. However, for the expected waiting time, the relationship is much clear. Considering that it is more representative of the entire denied boarding distribution, expected waiting time is selected as the response variable for the proposed prediction framework.

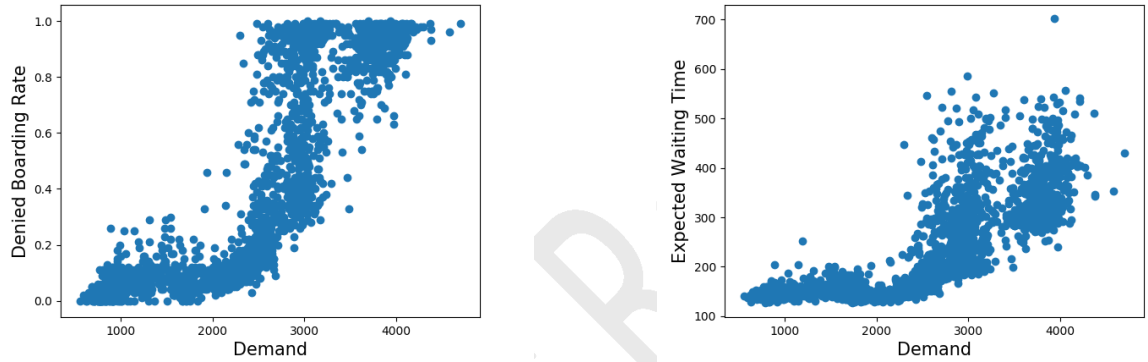

Figure 3 Scatterplot of Demand vs Denied Boarding Rate (left) and Expected Waiting Time (right)

Model Selection

The proposed methodology allows to formulate the crowding prediction problem as a typical supervised learning problem. More specifically, it becomes a regression problem where the aim is to predict a continuous response variable (in this case, expected waiting time) by modeling it as a function of a group of explanatory variables. Note that, the response variable for the time interval $[t, t + \Delta t]$ is represented using the explanatory variables observed for the time interval $[t - \Delta t, t]$. With this formulation, any kind of generic supervised learning algorithm can be used for prediction. But the model selection part for any supervised learning problem involves finding the model that best suits the needs and structure of the data set and fine tuning the parameters to get optimal results in terms of accuracy and robustness.

For this purpose, prediction results from three models, namely, Gradient Boosting Trees (XGB), Random Forests (RF) and Support Vector Machines (SVM) are considered and compared. These models are selected because they are efficient and have shown good performance with high number of explanatory variables. Also, each of these models can be used for regression problems and are able to produce complex, non-linear decision boundaries that is necessary to capture complex patterns. Moreover, the output of the models can be tested for accuracy using well-known accuracy metrics (e.g mean squared error) and for robustness using methods like leave-one-out cross validation. Also, the hyperparameters of these models can be optimized using a grid search through many different combinations of parameter setups.

CASE STUDY

The case study examines the application of the proposed prediction framework using real-life data from the Mass Transit Railway (MTR) operating in Hong Kong. MTR is responsible for operating the urban railway transit network serving the urbanized areas in Hong Kong Island, Kowloon, and the New Territories. Currently, the MTR network consists of 11 lines, serving 159 stations (91 heavy rail and 68 light rail stations) with 218.2 km of rail. MTR uses a smart card fare collection system called Octopus which records more then 5 million trips on an average weekday. For the heavy rail lines, the smart card data records both enry and exit to the system which allows for a complete record of each trip.

The denied boarding estimation model which is utilized in the proposed methodology was validated using manual survey data provided by MTR at some key stations where the crowding levels are high *(2,19*). Following the same idea, the prediction methodology is tested at one of these key stations within the network. The selected platform, is one of the busiest platforms in the entire network. In addition to its own demand, which is one of the highest in the network, it is at the intersection of three lines and receives transfer passengers from three directions. Figure 4 illustrates the network and the stations that are used in the case study.

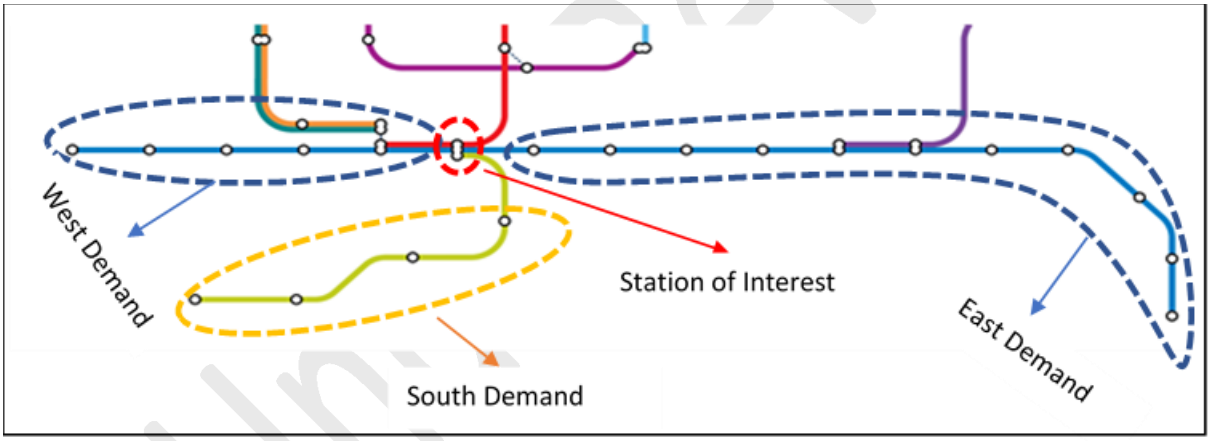

Figure 4 Network used in the case study

Denied boarding probabilities are estimated using the mixture distribution framework described in Ma, et al. *(19)*. For a given OD pair in a closed system, the methodology treats the denied boarding estimation problem as a clustering problem based on the journey time distribution. The denied boarding probabilities in the origin platform are estimated using Gaussian Mixture Models to differentiate journey time distribution into different groups. Then, the weights and probabilities associated with each group are used to represent the denied boarding probabilities. The paraemeters of the model is calibrated using actual data from the MTR system and estimation output for the station of interest is used as training data for the case study.

The available data covers a period spanning from January 2017 to July 2018. Denied boarding probabilities are estimated in discrete time intervals of fifteen minutes for the evening peak (6:00 PM to 8:00 PM), so every day consists of eight independent observations which follows

the methodology in Ma, et al. (19). After removing the missing and abnormal days within the data, 295 days are fully processed which results in 2360 observations that can be used to train and test the prediction model.

For the formulation of the problem, the expected waiting time is calculated for all the observations using **Equation 5**. Furthermore, the demand and headway related explanatory variables are calculated for the station of interest and the transfer stations. As shown in Figure 4, the upstream stations are separated into three parts; west, east and south representing incoming demand and trains from different directions. Moreover, both aggregate and disaggregate versions of upstream demand are calculated. The aggregate version consists of three upstream demand features for east, west and south. On the other hand, the disaggregate version consists of twenty different features representing each upstream station. This results in 11 total features for the aggregated version and 27 for the disaggregate version. Both versions are used in prediction for comparison purposes.

Three models (XGB, RF and SVM) are trained and results from each model is acquired for comparison. Results from the models are compared to the historical averages as well. Historical average method is treated as the most basic (naïve) method. The historical average value specific to each day and time interval (e.g Monday 5:00-6:15) is used as the prediction for all the future values of that day of the week and that time interval. Three prediction performance measures are reported for each model; mean absolute error (MAE), mean squared error (MSE) and, mean absolute percentage error (MAPE). Each model is fine tuned using a grid search through ten thousand different combinations of hyperparameters. Parameter tuning is the longest process within the modeling framework. But it is done only once for each model and then the same optimized parameters are used throughout the process. Therefore, it's impact on computational efficiency is limited. A leave-one-out cross validation scheme is applied in order to utilize all of the observations and mitigate the impact of outlier observations.

Results

Prediction results for various scenarios and models are provided in Table 1 including results for the naïve model. All three models outperform the historical average model significantly.

Table 2. Model Performances

	Model With Aggregate Features			**Model With Disaggregate Features**			**Historical Average**
	SVM	**RF**	**XGB**	**SVM**	**RF**	**XGB**	
MSE	989.925	967.33	940.56	997.33	1068.67	911.77	5442.45
MAE	19.313	19.60	19.10	19.56	20.42	18.65	48.65
MAPE	7.68%	7.85%	7.47%	7.77%	8.08%	7.20%	25.56%

For both aggregate and disaggregate features, the XGB model provides the best results even though the differences are very small. Small differences in accuracy for the three models indicate that the performance depends more on the selected features rather than the preferred supervised learning model. Once the parameters are fine-tuned to the set of explanatory features, the additional gain resulting from the best model is marginal. Similarly, using aggregate versus

disaggregate demand features for the upstream stations also makes small difference in terms of accuracy. The XGB model performs better with disaggregate features while the RF and SVM models perform better with aggregated demand features. Considering these observations and since the model with aggregated demand features has less number of predictors, it is selected as the final set of features. It is unnecessary to use more explanatory variables if the difference in the accuracy is insignificant. Also, since XGB model performs the best, it is selected as the final model.

Figure 5 shows the scatter-plot of predicted versus the actual observations of all test observations (on the right side) and the distribution of the errors (on the left side). Note that, the plotted results are reported from the XGB model with optimized parameters and using a leave-one-out train/test split strategy. The scatter-plot shows that all the observations are distributed around the 45-degrees line. The inner dashed lines indicate plus/minus one-minute distance from the 45-degress line and the outer dashed lines indicate plus/minus two-minutes distance. The mean absolute error for the plotted results is 19.6 seconds which means that on the average prediction is within twenty seconds proximity of the actual waiting time value. This corresponds to around 7% of the average waiting time and it is less than 7% during the so called peak of the peak (6-7 PM), since the waiting times are higher during that period. Moreover, 97% of all the observations are within one-minute of the actual value and 99.25% of all the observations are within two-minutes to the actual waiting time. This means that 2289 observations out of 2360 are predicted with an error less than one minute and 2344 out of 2360 are predicted with error smaller than two minutes. The observations with large errors are either outlier observations in terms of waiting time or cannot be differentiated by the current set of features. One possible solution to these issues may be using external information such as weather or event (e.g concert) to supplement the existing predictors. Currently the model only uses explanatory variables that are observable using AFC and AVL data.

Figure 5b. shows the histogram of errors. The errors are distributed evenly around zero. This result show that the model does not make any methodological errors since there is no bias in the error distribution. The low variability shows the model performs well since all the results are distributed around zero. All in all, the results show that the predictions that are made using the denied boarding estimates are accurate and robust. This is an important outcome since the high accuracy of these results shows that it is possible to provide accurate real time crowding information in a completely data-driven manner.

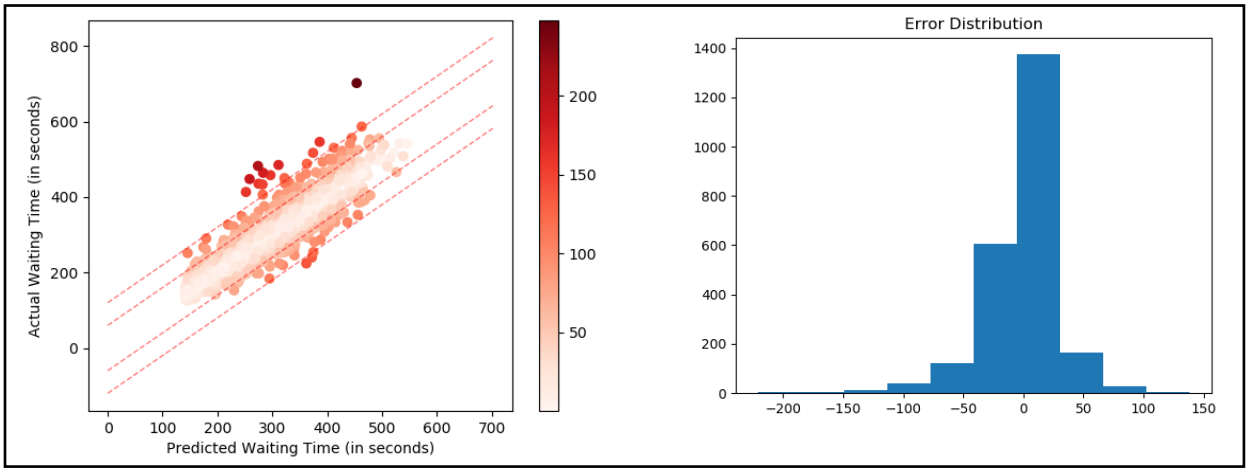

Figure 5 Scatterplot of Actual vs Predicted Waiting Time (a) and Histogram of errors (b)

The application of these methodology is coded in Python. For the XGB application, a Python package called 'xgboost' is used. This package also provides a metric to assess the contribution of each predictor variable to the model. The feature importance metric gives each predictor variable a percentage score that corresponds to the ratio of gain of the variable to the overall gain. The gain in this context refers to the improvement made in the objective function of the model. Note that, this performance metric is valid only for the XGB model. Therefore, the importance scores provided by the model cannot be considered as definitive results. However, they do provide some useful insight as to the usefulness of the proposed feature set within the selected model.

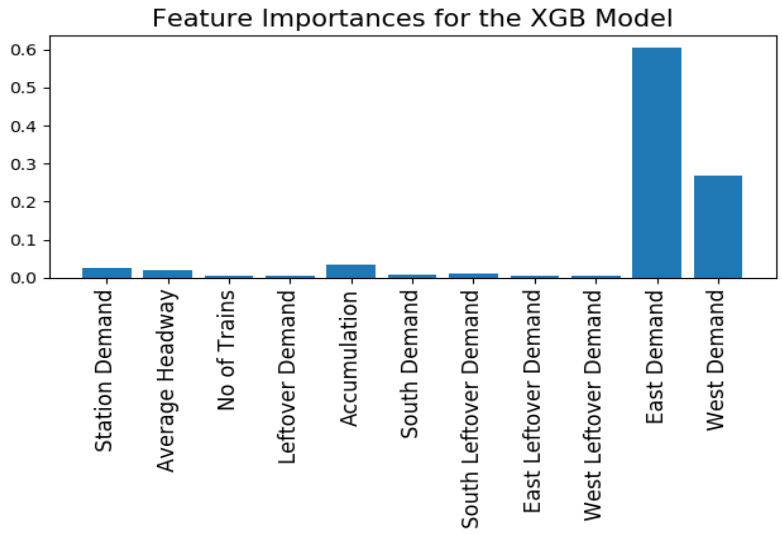

Figure 6 Feature Importances for the XGB model trained on disaggregate demand features

Figure 6 shows the feature importance for the XGB model with aggregated demand features. The transfer demand dominates the model since east and west demand constitute a total of 90% of the total gain. This is sensible since east and west sides of the station of interest has the highest transfer load. Another reason is due to using fifteen-minute time intervals. The average waiting time (for 15-minutes interval) in the platform of interest is at most 6-7 minutes. Since the model looks back into the last fifteen-minute information, the passengers who are tapping in at the station of interest would be most probably served in the past interval. Whereas the passengers transferring from other lines would have a delayed impact on crowding. Therefore, it makes sense that the model is more sensitive to the variability in transfer demand based on information with a fifteen-minute delay. Also, both the service and the arrival rates in MTR system are very regular, especially in the peak periods. So, the headway distribution has a low standard deviation and the passenger arrival rate is almost constant. This means there is low variability in operations related features such as headways and left-over demands which is another explanation for the prevalence of the transfer demand features. Based on these observations, another model has been built using only the demand related features in order to test the impacts on prediction accuracy. The new model resulted in a MAE score of 21.56 seconds and MAPE score of 8.43% which shows lower accuracy than the original model. This result indicates that although these parameters have low performance measurement scores, they are not redundant and they have impact on the accuracy of the model.

CONCLUSION

Real time prediction for platform crowding in major urban rail lines is useful to both operators and customers alike, since it has an impact on operating efficiency, safety and is a useful information for the riders that improves their customer experience. The paper proposes a methodology for real time platform crowding prediction using denied boarding probabilities as ground truth information. Denied boarding estimates can be acquired using data-driven methods that have been proposed in the literature and validated with actual observatios. The proposed methodology aims to extend this data-driven framework to a real-time prediction methodology. Denied boarding estimates are used to calculate the expected waiting time information which is used as the response variable in a supervised learning problem. Moreover, demand, operations, and incident related information are utilized as the predictor variables.

The proposed methodology is validated through a case study conducted on the MTR network in Hong Kong. Denied boarding probabilities are estimated for nearly 1.5 years of data and several supervised learning models are trained on the data set. The results from the case study show that the proposed methodology is able to provide accurate predictions using only AFC and AVL data. The robustness and high accuracy of the results also support the development of customer information tools that communicate this information in real time without relying on expensive and time consuming data collection processes.

ACKNOWLEDGMENTS

The authors would like to thank MTR for their support and provision of data used in this paper.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: all authors contributed to the study conception and design. K. Tuncel prepared the analysis in the paper and all authors contributed to the interpretation of the results. All authors contributed in the preparation of the manuscript and reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] Zhu, Y., H. N. Koutsopoulos, and N. H. M. Wilson. A probabilistic Passenger-to-Train Assignment Model based on automated data. *Transportation Research Part B: Methodological*, 2017.

[2] Zhu, Y., H. N. Koutsopoulos, and N. H. M. Wilson. Inferring left behind passengers in congested metro systems from automated data. *Transportation Research Part C: Emerging Technologies*, 2017.

[3] Li, Z., and D. A. Hensher. Crowding and public transport: A review of willingness to pay evidence and its relevance in project appraisal. *Transport Policy,* Vol. 18, No. 6, 2011, pp. 880-887.

[4] Haywood, L., and M. Koning. The distribution of crowding costs in public transport: New evidence from Paris. *Transportation Research Part A: Policy and Practice,* Vol. 77, No. Supplement C, 2015, pp. 182-201.

[5] Dixit, V. V., R. C. Harb, G. W. Harrison, D. M. Marco, M. S. Mard, A. Essam, E. E. Radwan, and M. P. Schneider. Review of Congestion Pricing Experiences.In, Report to the United States Federal Highway Administration, 2010.
[6] Kachroo, P., S. Gupta, S. Agarwal, and K. Ozbay. Optimal Control for Congestion Pricing: Theory, Simulation, and Evaluation. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18, No. 5, 2017, pp. 1234-1240.
[7] Halvorsen, A., H. N. Koutsopoulos, Z. Ma, and J. Zhao. Demand Management of Congested Public Transport Systems: A Conceptual Framework and Application Using Smart Card Data (DOI: 10.1007/s11116-019-10017-7). *Transportation*, 2019.
[8] Koutsopoulos, H. N., Z. Ma, P. Noursalehi, and Y. Zhu. Transit Data Analytics for Planning, Monitoring, Control and Information.In *Mobility Patterns, Big Data and Transport Analytics*, Elsevier, 2018. p. 448.
[9] K. Sohn, "Optimizing train-stop positions along a platform to distribute the passenger load more evenly across individual cars," IEEE Trans. Intell. Transp. Syst., vol. 14, no. 2, pp. 994–1002, Jun. 2013.
[10] J. C. Muñoz, J. Soza-Parra, A. Didier, and C. Silva, "Alleviating a subway bottleneck through a platform gate," Transp. Res. A, Policy Pract., vol. 116, pp. 446–455, Oct. 2018.
[11] Y. Liu, T. Tang, and J. Xun, "Prediction algorithms for train arrival time in urban rail transit," in Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC), Oct. 2017, pp. 1–6.
[12] R. Zhang et al., "WiFi sensing-based real-time bus tracking and arrival time prediction in urban environments," IEEE Sensors J., vol. 18, no. 11, pp. 4746–4760, Jun. 2018.
[13] K. Dziekan and K. Kottenhoff, "Dynamic at-stop real-time information displays for public transport: Effects on customers," Transp. Res. A, Policy Pract., vol. 41, no. 6, pp. 489–501, Jul. 2007.
[14] K. E. Watkins, B. Ferris, A. Borning, G. S. Rutherford, and D. Layton, "Where is my bus? Impact of mobile real-time information on the perceived and actual wait time of transit riders," Transp. Res. A, Policy Pract., vol. 45, no. 8, pp. 839–848, Oct. 2011.
[15] O. Cats and E. Jenelius, "Dynamic vulnerability analysis of public transport networks: Mitigation effects of real-time information," Netw. Spatial Econ., vol. 14, nos. 3–4, pp. 435–463, Dec. 2014
[16] Halvorsen, A., H. N. Koutsopoulos, S. Lau, T. Au, and J. Zhao. Reducing Subway Crowding: Analysis of an Off-Peak Discount Experiment in Hong Kong. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 2544, 2016, pp. 38-46.
[17] Greene-Roesel, R., J. Castiglione, C. Guiriba, and M. Bradley. BART Perks: Using Incentives to Manage Transit Demand. Presented at 97th Transportation Research Board Annual Meeting, Washington D.C. United States, 2018.
[18] Jenelius, E. Data-Driven Metro Train Crowding Prediction Based on Real-Time Load Data. *IEEE Transactions on Intelligent Transportation Systems,* 2019, pp. 1-12.
[19] Ma, Z., H. N. Koutsopoulos, Y. Chen, and N. H. M. Wilson. estimation of denied boarding in urban rail systems: alternative formulations and comparative analysis. *Transportation Research Record,* 2019.