Car-Specific Metro Train Crowding Prediction Based on Real-Time Load Data

Erik Jenelius
Department of Civil and Architectural Engineering
KTH Royal Institute of Technology
Stockholm, Sweden
jenelius@kth.se
April 25, 2018

***Abstract*—The paper formulates the car-specific metro train crowding prediction problem based on real-time load data and evaluates the performance of several prediction methods (stepwise regression, lasso, and boosted tree ensembles). The problem is studied for multiple stations along a metro line in Stockholm, Sweden. Prediction accuracy is evaluated with respect to absolute passenger loads and predefined discrete crowding levels. When available, predictions with real-time load data significantly outperform historical averages, with accuracy improvements varying in magnitude across target stations and prediction horizons.**

***Index Terms*—public transport, metro, crowding, prediction, load data, stepwise regression, lasso, boosted tree ensemble**

I. Introduction

Many urban public transport systems are experiencing increasing congestion and crowding. In-vehicle crowding may have many negative effects on traveler satisfaction and well-being, including stress, anxiety, threat to personal safety and security, and loss of productivity due to lack of seating space [1], [2]. Stated-preference and revealed-preference studies show that crowding may significantly increase travelers' value of time savings [3]–[6]. Crowding also affects vehicle dwell times at stations as well as passenger waiting times, which in turn increases variability in headways and reduces reliability [7]–[9].

Studies show that passenger loads can be highly unevenly distributed between the cars of trains and metros even during peak hours [10], [11]. This implies that the average experienced crowding level per passenger is higher. Uneven passenger distribution between train cars also means that the effective capacity of the trains is significantly lower than the nominal capacity based on all cars being equally utilized. As a consequence, more vehicles than necessary are required to serve the demand, which leads to significant costs for the operator.

Kim et al. (2014) investigate the factors that determine whether travelers choose a specific train car in the Seoul metro [12]. In the study, 54% of the respondents reported choosing a specific car intentionally to minimize walking distance at the destination station, and 13% sought to minimize walking distance at the origin station; 10% stated that they sought to maximize comfort during the trip.

Zhang et al. (2017) report a case study where car-specific real-time crowding information (RTCI) for the next arriving train was provided to waiting passengers [11]. Crowding conditions were measured at departure from the station immediately preceding the target station. The real-time information reduced the share of passengers boarding the front, most crowded car by 4.3% points for trains that were crowded on arrival, and increased the share of passengers boarding the middle, less crowded car by 4.1% points. Thus, provision of car-specific RTCI may increase the utilization of available train capacity and reduce in-vehicle crowding.

RTCI based only on observed loads at the preceding station restricts the timeliness of the information and may limit the potential for travellers to adjust their route, train and car choices. Providing crowding information earlier generally requires that passenger loads be predicted several stations ahead from the current train locations. In simulation settings, both simple prediction schemes based on the crowding of the one or two most recent train runs [13] and more complex schemes involving running the simulation model forward to a fixed point solution [14] have been proposed and evaluated. Both studies demonstrate that predictive RTCI may equalize crowding among vehicle trips and reduce passengers' experienced travel time. As far as we are aware, however, no studies have tackled the train crowding prediction problem in an empirical context. Also, no study has considered car-specific crowding prediction.

The aim of this paper is to formulate the car-specific metro train crowding prediction problem based on real-time load data and evaluate the performance of several prediction methods (stepwise regression, lasso, and boosted tree ensembles). The focus is on RTCI for provision to waiting, arriving or transferring passengers. The problem is studied for multiple target stations along a metro line in Stockholm, Sweden. Prediction accuracy is evaluated with respect to absolute passenger loads as well as predefined discrete crowding levels.

II. Methodology

This section introduces the crowding prediction problem and proposes several solution methods. The notation used

The work was funded in part by TRENoP Strategic Research Area and Stockholm City Council (SLL).

TABLE I
NOTATION

i	Train car index, $i = 1, \ldots, N$
j	Station index
k	Train run index
l	Day index
q_{ijkl}	Train car load at departure
$\mathbf{q}_{jkl}$	Train car loads, $\mathbf{q}_{jkl} = (q_{1,jkl}, \ldots, q_{Njkl})$
Q_{jkl}	Total train load, $Q_{jkl} = \sum_{i=1}^{N} q_{ijkl}$
$\bar{q}_{ijk'l}$, $\bar{\mathbf{q}}_{jk'l}$	Mean station load, $\bar{q}_{ijk'l} = \sum_{k=1}^{k'} q_{ijkl}/k'$
h_{ijk}^{run}, $\mathbf{h}_{jk}^{\text{run}}$, H_{jk}^{run}	Historial mean load for run k
h_{ijl}^{wday}, $\mathbf{h}_{jl}^{\text{wday}}$, H_{jl}^{wday}	Historial mean load for weekday of l
$\hat{q}_{ijkl}$, $\hat{\mathbf{q}}_{jkl}$	Predicted load
c^{sit}	Seat capacity per train car
r_{ijkl}, $\mathbf{r}_{jkl}$	1 if $q_{ijkl} > c^{\text{sit}}$, 0 otherwise.

throughout the paper is shown in Table I.

Consider the task of predicting the passenger load q_{ijkl} in car i of train run k on day l at departure from station j. In general, the load is predicted based on a combination of current-day and historical data.

First, the prediction can utilize historical average loads at target station j for the same train run k and the same weekday as current day l. Specifically, we consider the $1 \times 2(N+1)$ vector of historical predictors

$$\mathbf{x}_{jkl}^{\text{hist}} = \left(\mathbf{h}_{jk}^{\text{run}}, H_{jk}^{\text{run}}, \mathbf{h}_{jl}^{\text{wday}}, H_{jl}^{\text{wday}}\right) \tag{1}$$

Second, the prediction can utilize load information for previous train runs departing from target station j on current day l. Let $k' < k$ denote the most recent train that has departed from station j at the time for the prediction. Load predictions are based on the average load across all train runs departing from target station j within a time window (in this study, one hour is used) up to train k', collected in the $1 \times (N+1)$ vector of current-day target-station predictors,

$$\mathbf{x}_{jk'l}^{\text{stn}} = \left(\bar{\mathbf{q}}_{jk'l}, \bar{Q}_{jk'l}\right) \tag{2}$$

A weighted average may be used where higher weights are given to more recent train runs.

Third, given that the target train k has departed from the terminal, load data up to the most recent station $j' < j$ from which train k has departed on current day l at the time for the prediction can be utilized. Train car loads are predicted based on load measurements from potentially all stations between the start of the line and station j'. Specifically, we consider the $1 \times (2N+1)j'$ vector of train-specific current-day predictors,

$$\mathbf{x}_{j'kl}^{\text{run}} = (\mathbf{q}_{1,kl}, \ldots, \mathbf{q}_{j'kl}, Q_{1,kl}, \ldots, Q_{j'kl}, \mathbf{r}_{1,kl}, \ldots, \mathbf{r}_{j'kl}) \tag{3}$$

In the following we fix the target station j, source station j' and run k' and omit the indices for simplicity of notation. All predictors for run k and day l are collected in the $1 \times p$ vector

$$\mathbf{x}_{kl} = \left(\mathbf{x}_{kl}^{\text{run}}, \mathbf{x}_{l}^{\text{stn}}, \mathbf{x}_{kl}^{\text{hist}}\right), \tag{4}$$

with elements x_{klm}, $m = 1, \ldots, p$.

The following subsections propose three prediction methods for the crowding prediction problem: stepwise linear regression, lasso regularized regression, and boosted regression tree ensembles.

A. Stepwise Regression

The first predictive model allows for interaction terms and quadratic terms of the predictors in $\mathbf{x}$ and is linear in coefficients. Thus, for each train car $i = 1, \ldots, N$, loads are modelled as outcomes of the regression

$$q_{ikl} = \beta_{00}^{i} + \sum_{m=1}^{p} x_{klm}\beta_{m,0}^{i} + \sum_{m=1}^{p}\sum_{n=m}^{p} x_{klm}x_{kln}\beta_{mn}^{i} + \epsilon_{ikl}, \tag{5}$$

where the errors ϵ_{ikl} are assumed to be independent between cars, runs, stations and days.

Model coefficients are estimated using stepwise regression with bidirectional elimination. At each step, the procedure searches for predictors to add to or remove from the model based on the p-value for an F-test of the change in the sum of squared errors. Excluding a predictor is equivalent to setting the corresponding coefficient to 0. The sparseness of the model is controlled by the p-values for adding and removing predictors.

B. Lasso

The second prediction model is also linear in coefficients but does not include higher than linear terms of $\mathbf{x}$. Estimation uses the lasso regularization [15], i.e., parameters are selected to minimize

$$\frac{1}{2}\sum_{k}\sum_{l}\left(q_{ikl} - \beta_0^i - \sum_{m=1}^{p} x_{klm}\beta_m^i\right)^2 + \lambda^i \sum_{m=1}^{p}\left|\beta_m^i\right| \tag{6}$$

where λ^i is a regularization coefficient that penalizes large parameter values. Larger λ^i enforce sparser solutions, i.e., more parameters equal to zero. The λ^i value is calibrated to minimize the cross-validation mean squared error.

C. Boosted Regression Tree Ensemble

A regression tree ensemble is a model composed of a weighted combination of multiple regression trees, i.e., decision trees with binary splits for regression [15]. In this study, the tree ensemble is fitted using the LSBoost algorithm with shrinkage. In each step, a new tree is fitted to the difference between the observed response q_{ikl} and the aggregated prediction of all trees fitted previously to minimize the mean squared error. Each tree is grown from a template with design coefficients which are calibrated to minimize the cross-validation mean squared error, including the minimum number of observations per leaf and per branch node, and the maximal number of decision splits. Surrogate splits are used to improve the accuracy of predictions for data with missing values.

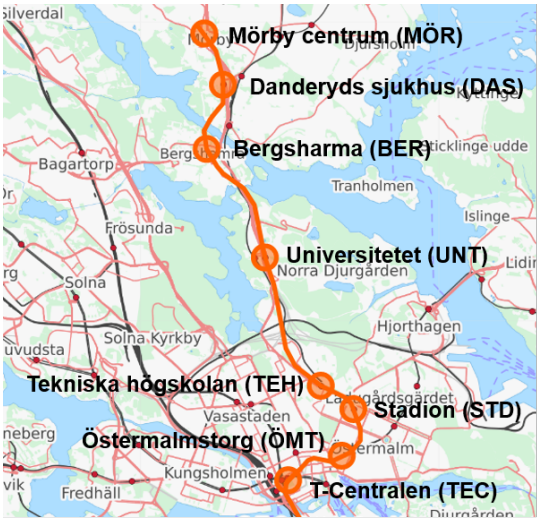

Fig. 1. Map of the studied metro line. Studied direction is from Mörby centrum (MÖR) towards T-Centralen (TEC). Map source: OpenStreetMap.

TABLE II
CROWDING LEVELS DEFINITION

Crowding level	Passenger load per train car
Low	0–149
Medium	150–249
High	250–

III. CASE STUDY

The load prediction methodology is applied to line 14 of the metro network in Stockholm, Sweden. The southbound direction of the north section of the line is considered, starting at terminal Mörby centrum (MÖR) and ending at station T-Centralen (TEC) in the city center. The geography of the line is shown in Figure 1.

The study considers the morning peak from 6:00 am to 9:00 am. During this period, the metro line runs with an average planned headway of 5 min. and there are 34 train runs per day departing from MÖR.

The seat capacity of a standard 3-car metro train is 378 passengers (126 per car). The practical capacity (used by the Stockholm public transport authority) is 650 passengers (217 per car), while the technical capacity (obtained from the train manufacturer) is 1200 passengers (400 per car). Stockholm public transport authority has defined three levels of in-vehicle crowding based on the utilization of available standee areas, common for trains, metro and buses. For the purposes of this study, the crowding levels are expressed in terms of the total number of passengers in a standard metro car, based on typical distributions between sitting and standing passengers in the metro (Table II).

A. Load Data

Load data from October 2016, Monday–Friday 6:00–9:00 am, are used. The load data are obtained from weight measurements in the air suspension system of the train cars. The number of passengers in each car is estimated based on an average weight of 78 kg per passenger including luggage.

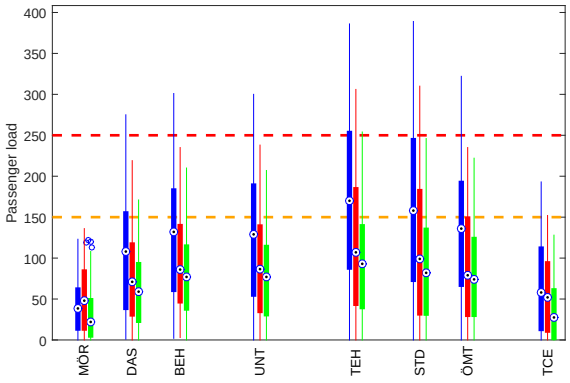

Fig. 2. Box plots of passenger loads at departure in the front (blue), middle (red), and rear (green) train car. Circles indicate medians, boxes indicate interquartile ranges (IQR), whiskers extend 1.5 IQR from nearest quartiles. Orange and red dashed lines indicate lower limits of medium and high crowding levels. Distances between stations indicate departure time differences.

There are in total 714 unique train runs in the data set. Of these, 345 runs have recorded load data. The other 369 runs are primarily runs with older vehicles without load measuring equipment, and are excluded from the analysis. Of the 345 trips with recorded load data, 272 have load data available for all considered stops between MÖR and TCE, while 73 trips have data missing for at least one stop. Model fitting and evaluation utilize data from all trips where data are available for all variables in the model; hence, the number of observations varies with the target station and model specification.

Figure 2 shows the distribution of loads at departure for each train car and station across the data set as box plots. With the exception of terminal station MÖR, crowding is consistently highest in the front car and lowest in the rear car. Crowding gradually increases and reaches the highest levels at stations TEH and STD and then decreases again.

B. Model Fitting

Out of the 21 days, 10 days (157 train trips) are randomly selected as the test set while the remaining 11 days (172 train trips) are used as training set. Historical mean loads are calculated across all 21 days. Current-day target-station mean loads are calculated over all departures with load data available during a one-hour time window up to the time of the prediction.

For the stepwise regression model, we use $p < 0.01$ and $p > 0.05$ as criteria for adding and removing a predictor, respectively. At initialization, each model contains historical predictors and train-specific current-day predictors but no interaction or quadratic terms. For the lasso regression, we use 10-fold cross-validation based on the mean squared error to select the regularization coefficient λ for each model. Similarly, we use 10-fold cross-validation based on the mean squared error to select tree template design coefficients in each

TABLE III
STEPWISE REGRESSION MODEL FOR UNT, HISTORICAL DATA

Train car Predictor	Estimate	p-value
Car 1 (front)		
Intercept	-2.0199	0.84086
$h^{\text{run}}_{1,\text{UNT}}$	1.0589	$5.5945 \cdot 10^{-35}$
Num. obs.	169	
adj. R^2	0.597	
Car 2 (middle)		
Intercept	-97.699	0.013806
$h^{\text{run}}_{2,\text{UNT}}$	1.1201	$3.071 \cdot 10^{-35}$
$h^{\text{wday}}_{2,\text{UNT}}$	0.91714	0.016165
Num. obs.	169	
adj. R^2	0.608	
Car 3 (rear)		
Intercept	-130.98	0.0022371
$h^{\text{run}}_{3,\text{UNT}}$	1.1207	$2.2245 \cdot 10^{-23}$
$h^{\text{wday}}_{3,\text{UNT}}$	1.4897	0.0027199
Num. obs.	169	
adj. R^2	0.464	

TABLE IV
STEPWISE REGRESSION MODEL FOR UNT, SOURCE STATION MÖR

Train car Predictor	Estimate	p-value
Car 1 (front)		
Intercept	-0.5528	0.94835
$q_{1,\text{MÖR}}$	1.2894	$4.0149 \cdot 10^{-28}$
$H^{\text{run}}_{\text{UNT}}$	0.29254	$8.4713 \cdot 10^{-23}$
Num. obs.	146	
adj. R^2	0.788	
Car 2 (middle)		
Intercept	134.82	0.0071146
$q_{2,\text{MÖR}}$	1.6714	$9.6514 \cdot 10^{-14}$
$Q_{\text{MÖR}}$	-1.5891	$3.3883 \cdot 10^{-6}$
$h^{\text{run}}_{2,\text{UNT}}$	0.46903	$1.0227 \cdot 10^{-14}$
$h^{\text{wday}}_{2,\text{UNT}}$	-1.3056	0.0080362
$q_{2,\text{MÖR}} \cdot Q_{\text{MÖR}}$	-0.0048924	$2.9748 \cdot 10^{-5}$
$Q_{\text{MÖR}} \cdot h^{\text{wday}}_{2,\text{UNT}}$	0.011744	0.00019998
$Q^2_{\text{MÖR}}$	0.0029229	$4.9846 \cdot 10^{-8}$
Num. obs.	146	
adj. R^2	0.894	
Car 3 (rear)		
Intercept	-8.3064	0.24156
$q_{3,\text{MÖR}}$	0.51858	0.025178
$Q_{\text{MÖR}}$	0.20456	$2.0639 \cdot 10^{-21}$
$q^2_{3,\text{MÖR}}$	0.0083406	0.00042193
Num. obs.	146	
adj. R^2	0.778	

boosted regression tree ensemble. New trees are fitted with a shrinkage learning rate of 0.005.

IV. RESULTS

We first focus on the stepwise regression method applied to target station UNT. Table III shows the estimated stepwise regression model for each car based on historical and current-day target-station data at the time of prediction just before the target train departs from the terminal; thus, train-specific current-day measurements are not available. The historical mean load for the same train run is found to be highly statistically significant for all train cars, and the historical weekday mean load is included for cars 2 and 3. Notably, current-day target-station mean loads do not further improve the predictive power significantly and are not included in any of the models.

Table IV shows estimated stepwise regression models based on current-day train-specific measurements from source station MÖR, and historical and current-day target-station data. As expected, the predictive power in terms of the adjusted R^2 is higher for all cars than without train-specific data (cf. Table III). For each car, the observed load of the same car at MÖR is highly statistically significant. The models for cars 1 and 2 also include historical data, and the models for cars 2 and 3 incorporate quadratic and interaction terms of certain predictors. Meanwhile, indicators $\mathbf{r}_{kl}$ for train cars exceeding the seat capacity are not included in any model.

Figure 3 shows predicted loads for target station UNT plotted against measured loads for the test data set when loads are predicted based historical data (models according to Table III) and with current-day train-specific data from source station MÖR (Table IV) and source station BEH, respectively. The dashed horizontal and vertical lines indicate the thresholds of the crowding levels. Prediction accuracy in terms of both absolute loads and crowding levels increases with current-day train data, as evident from the lower scatter around the diagonal. Historical data tend to underestimate the front car load on crowded runs, but overestimate the load on the least crowded runs.

Figure 4 shows the prediction accuracy for each train car as a function of time to departure from UNT; left diagram shows load root mean squared error (RMSE), right diagram shows accuracy in crowding level (%). More than 5.5 minutes before departure, prediction is based on historical data (Table III). Once the train departs from terminal station MÖR, predictions are updated according to the model for source station MÖR (Table IV). The train departs from the next station DAS 1.5 minutes later, and predictions are updated according to the model for source station DAS (not shown here). The process is repeated 1.5 minutes later as the train departs from station BEH.

Prediction accuracy increases steadily for all cars as the time to departure decreases. Crowding level prediction accuracy is close to 100% for cars 2 and 3 less than 2.5 minutes to departure. Accuracy in both absolute load and crowding level is generally lowest for the front (most crowded) car.

Dashed lines indicate departure times of preceding trains from UNT given the headway 5 minutes. It can be seen that when the current train departs from UNT, the load of the next train at departure from UNT can be predicted based on data from source station MÖR, while the load of the second next train may be predicted based on historical data.

We now compare the accuracy of the three prediction methods for several target stations, focusing on the front car which is generally the most crowded and most challenging to predict. Loads are predicted for the next and second next train

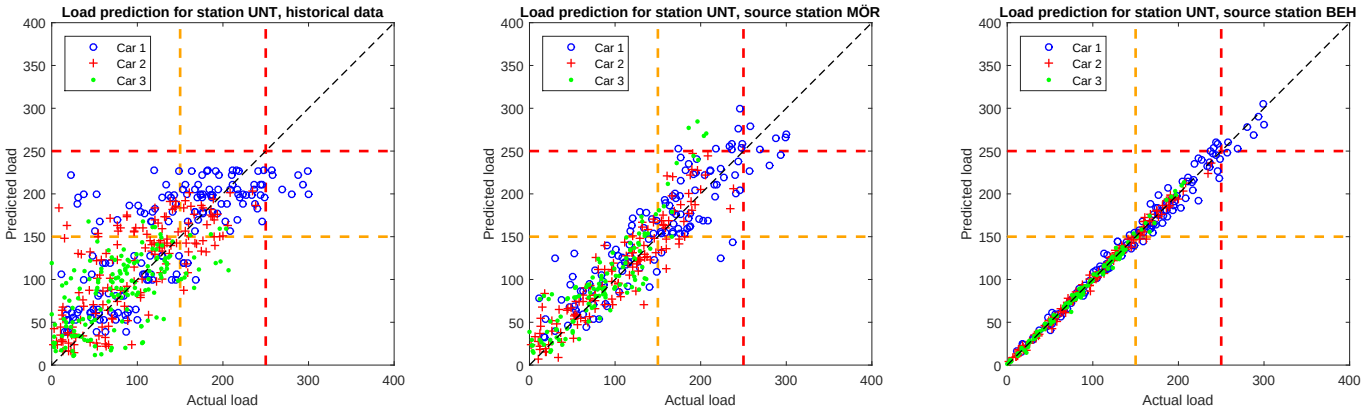

Fig. 3. Target station Universitetet (UNT), stepwise regression. Predicted vs. actual loads for test data set not used for model calibration. Left: Prediction based on historical data. Middle, right: Prediction with current-day train-specific data from source station MÖR and BEH, respectively. Orange and red dashed lines indicate lower limits of medium and high crowding levels.

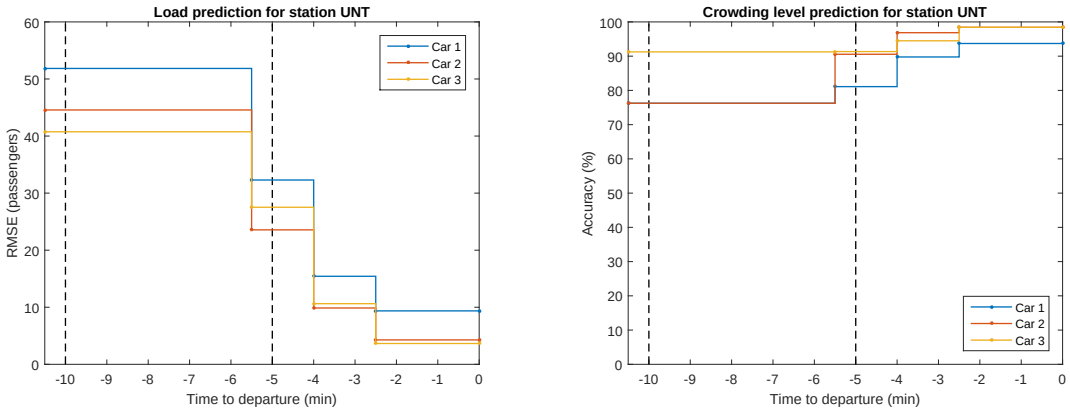

Fig. 4. Target station Universitetet (UNT), stepwise regression. Crowding prediction accuracy as function of time to departure. Left: Passenger load RMSE. Right: Crowding level (low, medium, high) accuracy (%). Dashed lines indicate cumulative headways between train runs.

(5 and 10 minutes ahead, respectively) at the time when the current train departs from the target station. As reference, the prediction using stepwise regression on only historical data is also included.

Figure 5 shows load prediction errors and crowding level prediction accuracy for target station UNT. For the next departing train, the three methods utilizing real-time measurements outperform the prediction based on historical data in terms of RMSE, while the increase in crowding level accuracy is less pronounced. For the second next train which has not yet left the terminal, only historical data are available, and all methods perform similarly.

Figure 6 shows load prediction errors and crowding level prediction accuracy for target station STD. For the next train, current-day train-specific measurements are available up to source station UNT, and the three methods utilizing real-time measurements greatly reduce prediction errors compared to using only historical data. Stepwise regression performs best of the methods; load RMSE is reduced by more than half, and crowding level accuracy is increased from ca. 67% (2 out 3 runs correctly predicted) to more than 80% (4 out of 5 runs correctly predicted) compared to using only historical data. For the second next train, current-day measurements are available from source station MÖR further upstream on the line, and prediction accuracy is lower. Overall, the stepwise regression and the lasso perform similarly while the boosted tree ensemble performs worse than the other two methods.

V. Conclusion

The paper formulates the car-specific metro train crowding prediction problem based on real-time load data and evaluates the performance of several prediction methods. When available, predictions with current-day train-specific load data significantly outperform historical averages, with accuracy improvements varying in magnitude across target stations and prediction horizons. Meanwhile, current-day target-station data for previous runs are found to be less important.

Combined with conclusions from earlier pilot studies and simulation experiments, the results suggest that real-time crowding information can be an effective means to influence travellers' route, train and car choices, and reduce in-vehicle crowding and passengers' experienced travel time. In metro networks with significant numbers of transfers between lines,

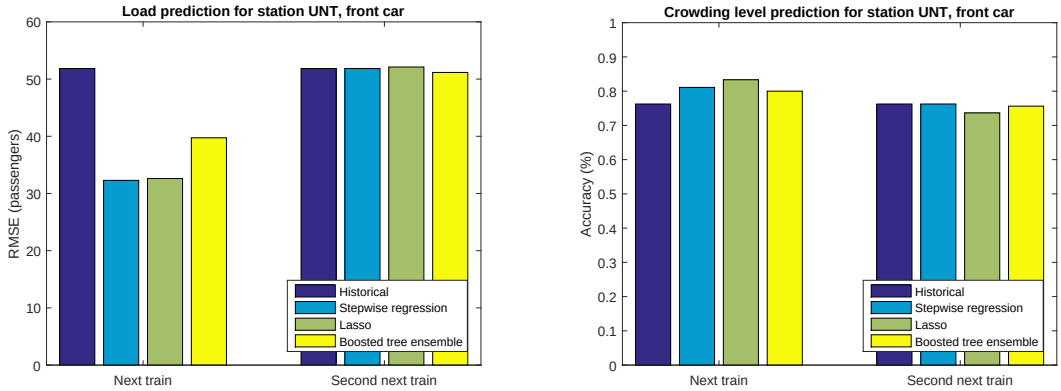

Fig. 5. Target station Universitetet (UNT), different prediction methods. Crowding prediction accuracy at time of departure for next and second next departing train, front car. Left: Passenger load RMSE. Right: Crowding level (low, medium, high) accuracy (%).

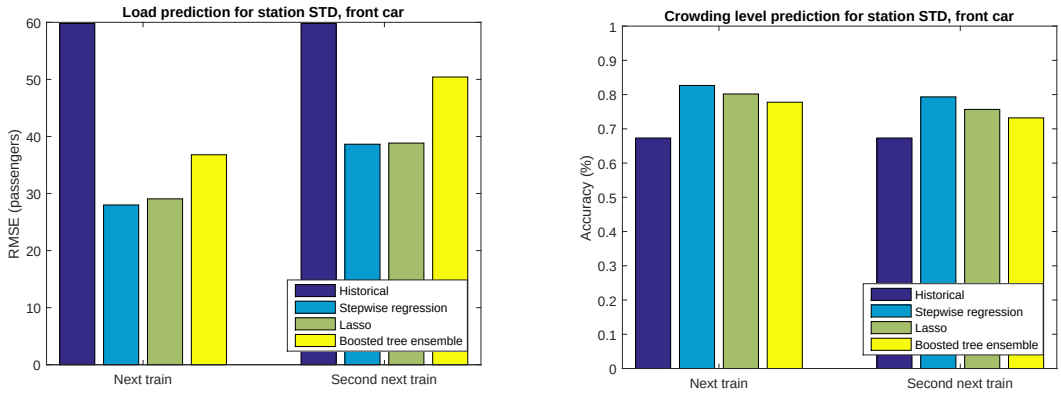

Fig. 6. Target station Stadion (STD), different prediction methods. Crowding prediction accuracy at time of departure for next and second next train, front car. Left: Passenger load RMSE. Right: Crowding level (low, medium, high) accuracy (%).

an interesting research question is whether predictions can be further improved by considering passenger loads on other lines. Another direction for further research is to extend the crowding prediction framework to other public transport modes, e.g., high-frequency buses.

ACKNOWLEDGMENT

The author would like to thank MTR Nordic for kindly providing the metro train car load data.

REFERENCES

[1] G. Beirão and J. A. Sarsfield Gabral, "Understanding attitudes towards public transport and private car: A qualitative study," *Transport Policy*, vol. 14, pp. 478–489, 2007.

[2] A. Tirachini, D. A. Hensher, and J. M. Rose, "Crowding in public transport systems: effects on users, operation and implications for the estimation of demand," *Transportation Research Part A*, vol. 53, pp. 36–52, 2013.

[3] M. Wardman and G. A. Whelan, "Twenty years of rail crowding valuation studies: evidence and lessons from British experience," *Transport Reviews*, vol. 31, no. 3, pp. 379–398, 2011.

[4] D. A. Hensher, J. M. Rose, and A. T. Collins, "Identifying commuter preferences for existing modes and a proposed Metro in Sydney, Australia with special reference to crowding," *Public Transport*, vol. 3, pp. 109–147, 2011.

[5] S. Raveau, Z. Guo, J. C. Munõz, and N. H. M. Wilson, "A behavioural comparison or route choice on metro networks: time, transfers, crowding, topology and socio-demographics," *Transportation Research Part A*, vol. 66, pp. 185–195, 2014.

[6] K. M. Kim, S.-P. Hong, S.-J. Ko, and D. Kim, "Does crowding affect the path choice of metro passengers?" *Transportation Research Part A*, vol. 77, pp. 292–304, 2015.

[7] T. Lin and N. H. M. Wilson, "Dwell time relationships for light rail systems," *Transportation Research Record*, vol. 1361, pp. 287–295, 1992.

[8] W. H. K. Lam, C.-Y. Cheung, and C. F. Lam, "A study of crowding effects at the Hong Kong light rail transit stations," *Transportation Research Part A*, vol. 33, pp. 401–415, 1999.

[9] Z. Qi, H. Baoming, and L. Dewei, "Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations," *Transportation Research Part C*, vol. 16, pp. 635–649, 2008.

[10] TRB, "Transit capacity and quality of service manual," Transit Cooperative Highway Research Program (TCRP) Report 165, Transportation Research Board, 2014.

[11] Y. Zhang, E. Jenelius, and K. Kottenhoff, "Impact of real-time crowding information: A Stockholm metro case study," *Public Transport*, vol. 9, no. 3, pp. 483–499, 2017.

[12] H. Kim, S. Kwon, S. K. Wu, and K. Sohn, "Why do passengers choose a specific car of a metro train during the morning peak hours?" *Transportation Research Part A*, vol. 61, pp. 249–258, 2014.

[13] A. Drabicki, R. Kucharski, O. Cats, and A. Fonzone, "Simulating the effects of real-time crowding information in public transport networks," in *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2017, pp. 675–680.

[14] A. Nuzzolo, U. Crisalli, A. Comi, and L. Rosati, "A mesoscopic transit assignment model including real-time predictive information on crowding," *Journal of Intelligent Transportation Systems*, vol. 20, no. 4, pp. 316–333, 2016.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2008.