Data-Driven Metro Train Crowding Prediction Based on Real-Time Load Data

Erik Jenelius
Division of Transport Planning
KTH Royal Institute of Technology
Stockholm, Sweden
jenelius@kth.se

May 1, 2019

***Abstract*—The paper formulates the car-specific metro train crowding prediction problem based on real-time load data and evaluates the performance of several data-driven prediction methods (lasso, stepwise regression, and boosted tree ensembles). Two variants of the prediction problem are considered: (1) train-centered prediction, focusing on in-vehicle crowding information provision, and (2) station-centered prediction focusing on crowding information provision at stations. The methodology is applied to a metro line in Stockholm, Sweden. Prediction accuracy is evaluated with respect to absolute passenger loads and 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 train cars depending on load variability. The results suggest that real-time crowding information can be provided sufficiently early to influence travellers' route, train and car choices, in order to reduce in-vehicle crowding.**

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

I. Introduction

Growing urban populations cause many public transit systems to experience increasing congestion and crowding. Crowding is associated with negative effects on traveler satisfaction and wellbeing, including stress, anxiety, threat to personal safety and security, and loss of productivity due to lack of seating space [1], [2]. Studies show that travelers' perceived travel times increase with the level of crowding [3]–[5]. Crowding also affects vehicle dwell times at stations as well as passenger waiting times, which increase the variability in headways and reduces reliability [6], [7]. As a consequence, more vehicles are required to serve the demand, which leads to significant costs for the operator.

Passenger loads can be highly unevenly distributed between the cars of trains and metros even during peak hours, which contributes negatively to crowding problems [8], [9]. Uneven passenger loads implies that the effective capacity of the trains is significantly lower than the nominal capacity based on all cars being equally utilized. The skewness of passengers waiting on the platform and boarding different cars depends on a number of factors, including the distance between the platform entrance and the waiting position, the capacity of each waiting position, the crowding distribution between train cars and the exit location at the destination station [9]–[11].

A. Public Transit Crowding Mitigation

Compared to congestion management in urban road transportation, which includes advanced technologies such as adaptive traffic control, metering and congestion pricing, crowding management in public transit is still at an early stage of development. Some studies have considered tactical planning methods to reduce the skewness of passenger loads, including varying the stop positions of trains at stations considering passenger flows and locations of station entrances [12], and the installation of one-way gates on platforms to control passenger flows and influence car boarding choices [13], but real-time crowding management is uncommon.

In recent years, the emergence of advanced public transportation systems (APTS) has spurred the use of technologies for automatically collecting data on vehicle locations (AVL), passenger counts and loads (APC), and fare collection transactions (AFC). Many cities have started to provide real-time vehicle arrival time information (RTI) at stations or in mobile applications. Several methods for short-term arrival time prediction based on various assumptions about data availability have been proposed, e.g., [14]–[16]. RTI systems have been shown to have positive effects on perceived waiting times, safety and security, impacts of service disruptions, and general satisfaction [17]–[20].

Systems providing real-time crowding information (RTCI) are still uncommon in practice; one of the first systems was introduced in Singapore in 2018 [21]. If accurate, at-station RTCI provision allows travellers to make better informed decisions about whether to board a vehicle or not, as well as which car to board in cases where the vehicles have multiple cars, based on their preferences for crowding in relation to waiting time, walking distance, etc [22]. Further, in-vehicle RTCI provision would allow passengers to decide where to alight depending on foreseen downstream crowding conditions. It should be noted that even if only a limited share of the passengers adjust their decisions, this has a large social value for the passengers already in the train and for those boarding at downstream stations.

B. Crowding Prediction

To allow travellers to adjust their route, train and car choices it is important that RTCI is provided in a timely manner.

The work was funded in part by TRENoP Strategic Research Area.

This requires prediction of passenger loads, generally several stations downstream from the current train locations.

In road transportation, short-term traffic flow prediction for traffic management and information provision has long been an active field of research, with prediction methods including time series analysis, regression modeling, clustering and pattern recognition algorithms [23]. In public transit, passenger load prediction is still at an undeveloped stage. In simulation settings both simple prediction schemes based on the crowding of the one or two most recent train runs [24] and more complex schemes involving running the simulation model forward to a fixed point solution [25] have been proposed and evaluated. Both studies demonstrate that predictive RTCI may equalize crowding among vehicle trips and reduce passengers' experienced travel time.

A recent study utilizes real-time AFC and AVL data from trains combined with mesoscopic simulation to predict demand patterns and in-vehicle crowding, focusing on the risk of denied boarding [26]. Another approach proposes a combination of historical pattern recognition and Kalman filtering for short-term prediction of bus loads based on real-time AFC and AVL data [27]. As far as we are aware, however, no studies have tackled the train crowding prediction problem based on passenger load data, which are less privacy-sensitive, require significantly less processing and may be more manageable to collect in real-time than APC data for many transit systems. Also, no study has considered car-specific crowding prediction, which is required for influencing the passenger distribution among cars.

The aim of this paper is formalize the car-specific metro train crowding prediction problem based on real-time load data. With a focus on efficient traveller RTCI provision, a data-driven prediction approach is taken and the performance of several methods (lasso, stepwise regression, and boosted tree ensembles) is evaluated. Two perspectives of the problem are considered: (1) Train-centered prediction focusing on in-vehicle RTCI provision, and (2) station-centered prediction focusing on at-station RTCI provision; it is shown that the latter can be built on the former and utilize the same prediction model. The problem is studied for a metro line in Stockholm, Sweden. Prediction accuracy is evaluated with respect to absolute passenger loads as well as predefined discrete crowding levels (low, medium, and high crowding) across target stations, time of day and prediction horizons.

The remainder of the paper is organized as follows. Section II introduces the prediction methodology. Section III presents the case study, with results reported in Section IV. Section V concludes the paper.

II. Methodology

This section introduces the crowding prediction problem and proposes several solution methods. The notation used throughout the paper is shown in Table I.

We assume that the following sets of information are available:

1) Departure times τ_{jkl} from station j for train run k on day l from a set of consecutive stations J, daily train runs K and days L.

TABLE I
Notation

i	Train car index, $i = 1, \ldots, I$
j	Station index, $j \in J$
k	Daily train run index, $k \in K$
l	Day index, $l \in L$
c^{sit}	Seat capacity per train car
τ_{jkl}^{plan}	Train scheduled departure time
τ_{jkl}	Train departure time
q_{ijkl}	Train car load at departure
$\mathbf{q}_{jkl}$	$(q_{1,jkl}, \ldots, q_{I,jkl})$
Q_{jkl}	Total train load, $\sum_{i=1}^{I} q_{ijkl}$
$\bar{q}_{ijkl}$	Day-station specific mean train car load
$\bar{\mathbf{q}}_{jkl}$	$(\bar{q}_{1,jkl}, \ldots, \bar{q}_{I,jkl})$
$\bar{Q}_{jkl}$	$\sum_{i=1}^{I} \bar{q}_{ijkl}$
r_{ijkl}	Crowding indicator, 1 if $q_{ijkl} > c^{\text{sit}}$, 0 otherwise
$\mathbf{r}_{jkl}$	$(r_{1,jkl}, \ldots, r_{I,jkl})$
h_{ijk}^{run}	Historical run-specific mean load
$\mathbf{h}_{jk}^{\text{run}}$	$(h_{1,jk}^{\text{run}}, \ldots, h_{I,jk}^{\text{run}})$
H_{jk}^{run}	$\sum_{i=1}^{N} h_{ijk}^{\text{run}}$
h_{ijl}^{wday}	Historical weekday-specific mean load
$\mathbf{h}_{jl}^{\text{wday}}$	$(h_{1,jl}^{\text{wday}}, \ldots, h_{I,jl}^{\text{wday}})$
H_{jl}^{wday}	$\sum_{i=1}^{I} h_{ijl}^{\text{wday}}$
h_{ijl}^{week}	Historical week-of-year-specific mean load
$\mathbf{h}_{jl}^{\text{week}}$	$(h_{1,jl}^{\text{week}}, \ldots, h_{I,jl}^{\text{week}})$
H_{jl}^{week}	$\sum_{i=1}^{I} h_{ijl}^{\text{week}}$

2) Scheduled departure times τ_{jkl}^{plan} from station j for train run k on day l from the same sets J, K and L.
3) Passenger loads q_{ijkl} in car i at departure from station j for train run k on day l for all cars $i = 1, \ldots, I$ and the same sets J, K and L.

We consider two closely related prediction problems: (1) train-centered, and (2) station-centered prediction. Train-centered prediction focuses on the downstream crowding of a particular train for in-vehicle information provision, while station-centered prediction focuses on the crowding of subsequent departures from a specific station for at-station information provision. Fig. 1 illustrates the difference between the two perspectives.

A. Train-Centered Prediction

The train-centered prediction problem is formulated as follows: At time τ on current day l, predict the passenger load $q_{ij^{\text{t}}k^{\text{t}}l}$ in car i of target train run k^{t} when it departs from target station j^{t}.

Three types of data are potentially available for the prediction: (1) real-time target station data, (2) real-time target train data, and (3) historical data.

1) Target-Station Predictors: Given that at least one train has departed from the target station at time t, the prediction can utilize load information for previous train runs departing from the target station on the same day. This may capture systematic differences in passenger loads between the current day and the historical average.

Let $K_{j^{\text{t}}l}(\tau)$ denote the set of train runs that has departed from target station j^{t} on day l at time τ, defined as

$$K_{j^{\text{t}}l}(\tau) = \{k \in K \mid \tau_{j^{\text{t}}kl} < \tau\}. \tag{1}$$

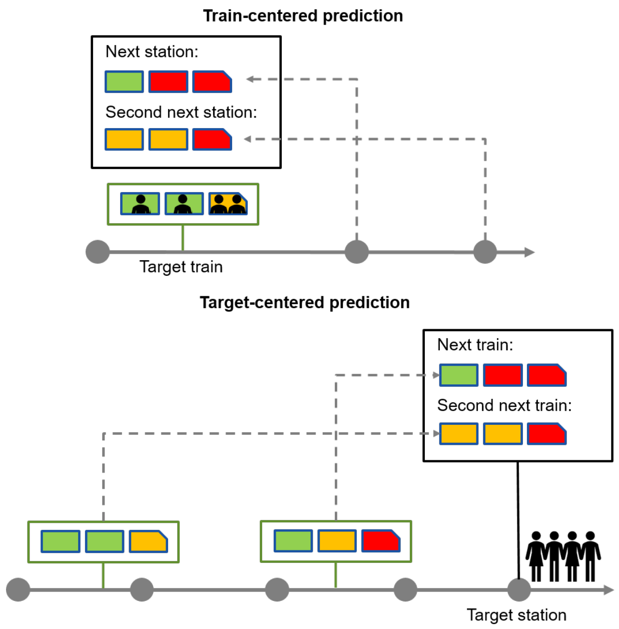

Fig. 1. Train-centered (top) and target-centered (bottom) crowding prediction and RTCI provision.

If $K_{j^{\mathrm{t}}l}(\tau) \neq \emptyset$, the source train run k^{s}, i.e., the most recent train that has departed from target station j at the time for the prediction t, is identified as

$$k^{\mathrm{s}} = \mathrm{argmax}_{k \in K_{j^{\mathrm{t}}l}(\tau)}\, \tau_{j^{\mathrm{t}}kl}. \tag{2}$$

Load predictions are based on the average load across the $K' \leq k^{\mathrm{s}}$ most recent train runs departing from target station j^{t} up to train k^{s},

$$\bar{q}_{ij^{\mathrm{t}}k^{\mathrm{s}}l} = \frac{1}{K'} \sum_{k=k^{\mathrm{s}}-K'+1}^{k^{\mathrm{s}}} q_{ij^{\mathrm{t}}kl}. \tag{3}$$

If $K_{j^{\mathrm{t}}l}(\tau) = \emptyset$, no real-time target-station load data are available; in this case we define $k^{\mathrm{s}} = 0$ which gives $K' = 0$.

The mean loads for each individual car and in total for the whole train are collected in the $1 \times (I+1)$ vector of real-time target-station predictors,

$$\mathbf{x}^{\mathrm{stn}}_{j^{\mathrm{t}}k^{\mathrm{s}}l} = \left(\bar{\mathbf{q}}_{j^{\mathrm{t}}k^{\mathrm{s}}l}, \bar{Q}_{j^{\mathrm{t}}k^{\mathrm{s}}l}\right). \tag{4}$$

2) Target-Train Predictors: Given that the target train (provided as input for train-centered prediction or identified as the next departing train for station-centered prediction) has departed from the terminus, load data up to the most recent station from which the target train has departed on the current day can be utilized.

Let $J_{k^{\mathrm{t}}l}(\tau)$ denote the set of stations from which target train k^{t} has departed on day l at time τ, defined as

$$J_{k^{\mathrm{t}}l}(\tau) = \{j \in J \mid \tau_{jk^{\mathrm{t}}l} < \tau\}. \tag{5}$$

If $J_{k^{\mathrm{t}}l}(\tau) \neq \emptyset$, the source station j^{s}, i.e., the most recent station from which target train k^{t} has departed at the time for prediction t, is identified as

$$j^{\mathrm{s}} = \mathrm{argmax}_{j \in J_{k^{\mathrm{t}}l}(\tau)}\, t_{jk^{\mathrm{t}}l}. \tag{6}$$

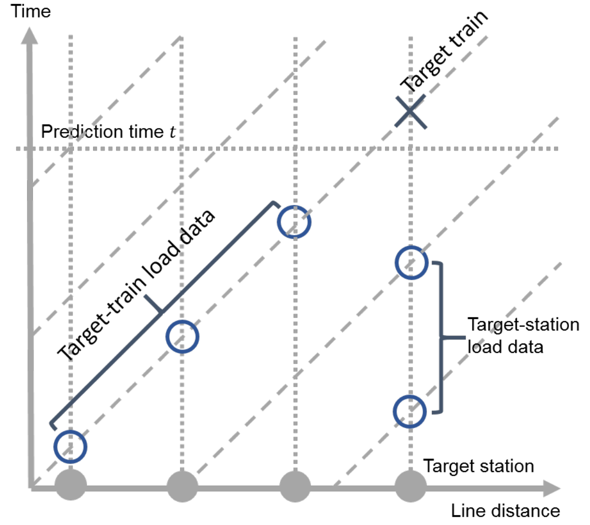

Fig. 2. Real-time target-train and target-station load data. Dashed diagonal lines indicate train trajectories.

The train car load is predicted based on load measurements $q_{ijk^{\mathrm{t}}l}$ from $J' \leq j^{\mathrm{s}}$ stations up to and including j^{s}. If $J_{k^{\mathrm{t}}l}(\tau) = \emptyset$, no real-time target-train load data are available; in this case we define $j^{\mathrm{s}} = 0$ which gives $J' = 0$.

To allow possible effects of seat availability on passengers' boarding choices, we also include seat availability indicators r_{ijkl} equal to 1 if $q_{ijkl} \leq c^{\mathrm{sit}}$ and 0 otherwise. All in all, we consider the $1 \times (2I+1)J'$ vector of train-specific current-day predictors,

$$\begin{aligned}\mathbf{x}^{\mathrm{run}}_{j^{\mathrm{s}}k^{\mathrm{t}}l} = (&\mathbf{q}_{j^{\mathrm{s}}-J'+1,k^{\mathrm{t}}l}, \ldots, \mathbf{q}_{j^{\mathrm{s}}k^{\mathrm{t}}l},\\ &Q_{j^{\mathrm{s}}-J'+1,k^{\mathrm{t}}l}, \ldots, Q_{j^{\mathrm{s}}k^{\mathrm{t}}l},\\ &\mathbf{r}_{j^{\mathrm{s}}-J'+1,k^{\mathrm{t}}l}, \ldots, \mathbf{r}_{j^{\mathrm{s}}k^{\mathrm{t}}l}).\end{aligned} \tag{7}$$

Fig. 2 illustrates the generation of real-time target-train and target-station load data in a space-time diagram.

3) Historical Predictors: Regardless of whether real-time target-station or target-train data are available, prediction can utilize recurring patterns in passenger loads related to target station j^{t}, target train run k^{t} and characteristics of day l extracted from historical data.

It can be expected that passenger loads vary systematically with the time of day, day of week, time of year, etc. Specifically, we consider the historical mean load on vehicle run k^{t} and the historical mean load for the same weekday and week of the year as current day l, for each individual car and in total for the train, collected in the $1 \times 3(I+1)$ vector of historical predictors

$$\mathbf{x}^{\mathrm{hist}}_{j^{\mathrm{t}}k^{\mathrm{t}}l} = \left(\mathbf{h}^{\mathrm{run}}_{j^{\mathrm{t}}k^{\mathrm{t}}}, H^{\mathrm{run}}_{j^{\mathrm{t}}k^{\mathrm{t}}}, \mathbf{h}^{\mathrm{wday}}_{j^{\mathrm{t}}l}, H^{\mathrm{wday}}_{j^{\mathrm{t}}l}, \mathbf{h}^{\mathrm{week}}_{j^{\mathrm{t}}l}, H^{\mathrm{week}}_{j^{\mathrm{t}}l}\right). \tag{8}$$

4) Data-Driven Load Model: All predictors are collected in the $1 \times p$ vector

$$\mathbf{x}_{j^{\mathrm{s}}j^{\mathrm{t}}k^{\mathrm{s}}k^{\mathrm{t}}l} = \left(\mathbf{x}^{\mathrm{stn}}_{j^{\mathrm{t}}k^{\mathrm{s}}l}, \mathbf{x}^{\mathrm{run}}_{j^{\mathrm{s}}k^{\mathrm{t}}l}, \mathbf{x}^{\mathrm{hist}}_{j^{\mathrm{t}}k^{\mathrm{t}}l}\right). \tag{9}$$

In cases where there is no real-time target-station data ($k^{\mathrm{s}} = 0$) or real-time target-train data ($j^{\mathrm{s}} = 0$) available the corresponding vectors $\mathbf{x}^{\mathrm{stn}}_{j^{\mathrm{t}}k^{\mathrm{s}}l}$ and $\mathbf{x}^{\mathrm{run}}_{j^{\mathrm{s}}k^{\mathrm{t}}l}$, respectively, are empty.

Algorithm 1 Station-centered prediction

```
procedure FindTargetTrain($\tau, l, j^{\mathrm{t}}$)
    $K_{j^{\mathrm{t}}l}(\tau) \leftarrow \{k \mid \tau_{j^{\mathrm{t}}kl} < \tau\}$
    $j' \leftarrow j^{\mathrm{t}} - 1$
    $\Delta K_{j'j^{\mathrm{t}}l}(\tau) \leftarrow \emptyset$
    while $j' \geq 1 \wedge \Delta K_{j'j^{\mathrm{t}}l}(\tau) = \emptyset$ do
        $K_{j'l}(\tau) \leftarrow \{k \mid \tau_{j'kl} < \tau\}$
        $\Delta K_{j'j^{\mathrm{t}}l}(\tau) \leftarrow K_{j'l}(\tau) \setminus K_{j^{\mathrm{t}}l}(\tau)$
        $j' \leftarrow j' - 1$
    if $\Delta K_{j'j^{\mathrm{t}}l}(\tau) \neq \emptyset$ then
        $k^{\mathrm{t}} \leftarrow \mathrm{argmax}_{k \in \Delta K_{j'j^{\mathrm{t}}l}(\tau)} \tau_{j'kl}$
    else
        $K'_l(\tau) \leftarrow \{k \mid \tau^{\text{plan}}_{kl} < \tau\}$
        $k^{\mathrm{t}} \leftarrow \mathrm{argmax}_{k \in \Delta K'_l(\tau)} \tau^{\text{plan}}_{1,kl}$
    return $k^{\mathrm{t}}$
```

A passenger load model is defined for each car $i = 1, \ldots, I$ by specifying and calibrating a $\mathbb{R}^p \to \mathbb{R}$ prediction function $f_{ij^{\mathrm{s}}j^{\mathrm{t}}}(\mathbf{x})$ for every feasible combination of target station j^{t} and source station j^{s}. Including the case $j^{\mathrm{s}} = 0$, this gives at most $I\,|J|\,|J+1|\,/2$ combinations. The predicted load is

$$\hat{q}_{ij^{\mathrm{t}}k^{\mathrm{t}}l} = f_{ij^{\mathrm{s}}j^{\mathrm{t}}}\left(\mathbf{x}_{j^{\mathrm{s}}j^{\mathrm{t}}k^{\mathrm{s}}k^{\mathrm{t}}l}\right). \tag{10}$$

Three different prediction models are proposed in Section II-C.

B. Station-Centered Prediction

Station-centered prediction is formulated as follows: At time τ on current day l, predict the passenger load in car i of the next train departing from target station j^{t}.

Station-centered prediction can be divided into two steps:

1) Identify the target train run k^{t} that is next to depart from target station j^{t} at time τ on day l.
2) Apply train-centered prediction for train k^{t}.

Thus, station-centered prediction can be built on train-centered prediction and the same prediction model can be utilized for both purposes. The first step can be performed with the procedure in Algorithm 1. The algorithm first seeks to find the nearest train that has departed from any station but not yet departed from j^{t} by iteratively searching upstream from j^{t}. If no such train exists the algorithm finds the next scheduled train run. It is straightforward to generalize station-centered prediction to the next $n \geq 1$ trains departing from the target station.

C. Prediction Methods

In the following we fix the target station j^{t}, source station j^{s} and source run k^{s}, and omit the indices for simplicity of notation. All predictors for run k and day l are collected in the $1 \times p$ vector $\mathbf{x}_{kl}$ with elements x_{klm}, $m = 1, \ldots, p$. We consider three methods for the crowding prediction problem: lasso regularized regression, stepwise linear regression, and boosted regression tree ensembles.

1) Lasso: The first prediction model is linear in coefficients. Estimation uses the lasso regularization [28], i.e., parameters are selected to minimize

$$\frac{1}{2}\sum_k \sum_l \left(q_{ikl} - \beta_{i,0} - \sum_{m=1}^{p} x_{klm}\beta_{im}\right)^2 + \lambda_i \sum_{m=1}^{p} |\beta_{im}|\,, \tag{11}$$

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.

Note that the same set of predictors is used for each individual car, but parameters are car-specific. Given a new vector of predictors $\mathbf{x}^*$, the passenger load for car i is predicted as

$$\hat{q}_i = \beta_{i,0} + \sum_{m=1}^{p} x^*_m \beta_{im}. \tag{12}$$

2) Stepwise Regression: The second predictive model is also linear in coefficients but allows for interaction terms and quadratic terms of the predictors in $\mathbf{x}$. Model coefficients are estimated using stepwise regression with bidirectional elimination separately for each train car $i = 1, \ldots, I$. 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,

$$\begin{aligned}\sum_k \sum_l \Bigg(q_{ikl} &- \beta_{i,00} - \sum_{m=1}^{p} x_{klm}\beta_{im,0} \\ &- \sum_{m=1}^{p}\sum_{n=m}^{p} x_{klm}x_{kln}\beta_{imn}\Bigg)^2 .\end{aligned} \tag{13}$$

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.

3) 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 [28]. 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.

D. Computational Complexity

The main computational requirement lies in the fitting of the prediction model for each feasible combination of train car, source station and target station. The calibration can be run off-line at relatively infrequent intervals (e.g., once every few months), which means that computational time is not a constraint.

Predictions are performed efficiently in real-time by applying the most recent load data and historical predictors to the appropriate model. As shown in Section IV, the number of predictors in each model is typically low (e.g., never above 10 for the lasso models in this case study), and computation times are negligible.

E. Extension: Transferring Load Predictors

In metro networks with significant numbers of transfers between lines, predictions may be further improved by considering passenger loads on other lines. In particular, the load at departure from target station j^{t} can be influenced by transfers to the considered metro line at all stations $j = j^{\mathrm{s}} + 1, \ldots, j^{\mathrm{t}}$ downstream of the source station where other lines stop. From vehicle load data it is not possible to observe the number of transferring passengers. However, it can be assumed that the number transferring passengers is related to the passenger load arriving to the station on the other lines just before the source train departs.

The total arriving passenger load equals the sum of passenger loads at departure from the preceding stations on the other lines. These loads can be predicted by applying the proposed prediction method separately on each of the other lines. Thus, predicted loads at stations on other lines upstream of common transfer stations can be used as predictors for the target station on the considered line. This principle can be extended recursively to incorporate all transfer stations in the network without significantly increasing the number of predictors in each models.

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 terminus Mörby centrum (MÖR) and ending at station T-Centralen (TEC) in the city center. Both TEC and nearest upstream station ÖMT are transfer stations. The geography of the line is shown in Fig. 3.

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 between departing from MÖR 6:00 am and 9:00 am.

The seat capacity of a standard 3-car metro train is 378 passengers (126 per car). The practical capacity (used by the Stockholm public transportation authority) is 650 passengers (217 per car), while the technical capacity (obtained from the train manufacturer) is 1200 passengers (400 per car). Stockholm public transportation 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).

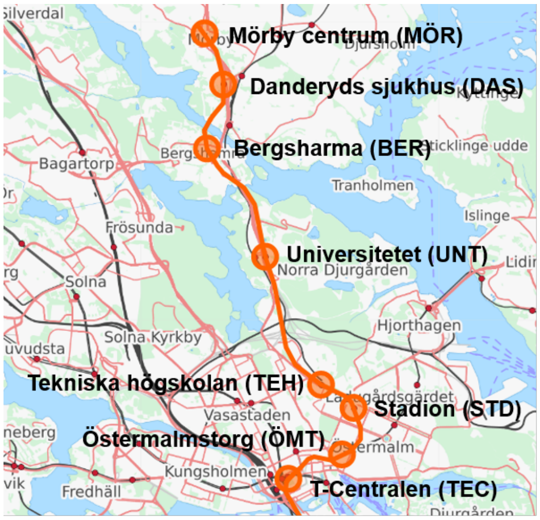

Fig. 3. Map of the studied metro line. Studied direction is from Mörby centrum (MÖR) towards T-Centralen (TEC). Map source: OpenStreetMap.

TABLE II
Crowding levels definition

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

A. Load Data

Load data from October 1–30 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 data is currently collected in batches of several days at a time but is used here to evaluate the potential of having load data available in real-time. The number of passengers in each car is estimated based on an average weight of 78 kg per passenger including 2 kg luggage.

There are in total 680 unique train runs in the data set. Of these, 329 runs (48.4%) have recorded load data. The other 351 runs are primarily runs with older vehicles without load measuring equipment, and are excluded from the analysis. Of the 329 trips with recorded load data, 269 have load data available for all considered stops between MÖR and TCE, while 60 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.

Fig. 4, top left, shows the distribution of loads at departure for each train car and station across the data set; dark color bars show mean values while light color bars show the 10th and 90th percentiles (all bars start at 0 and end at the indicated level). 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. The load variability across days and train runs, indicated by the range between the 10th and 90th percentiles, is generally highest for the front car and for stations TEH and STD which also have the highest mean loads.

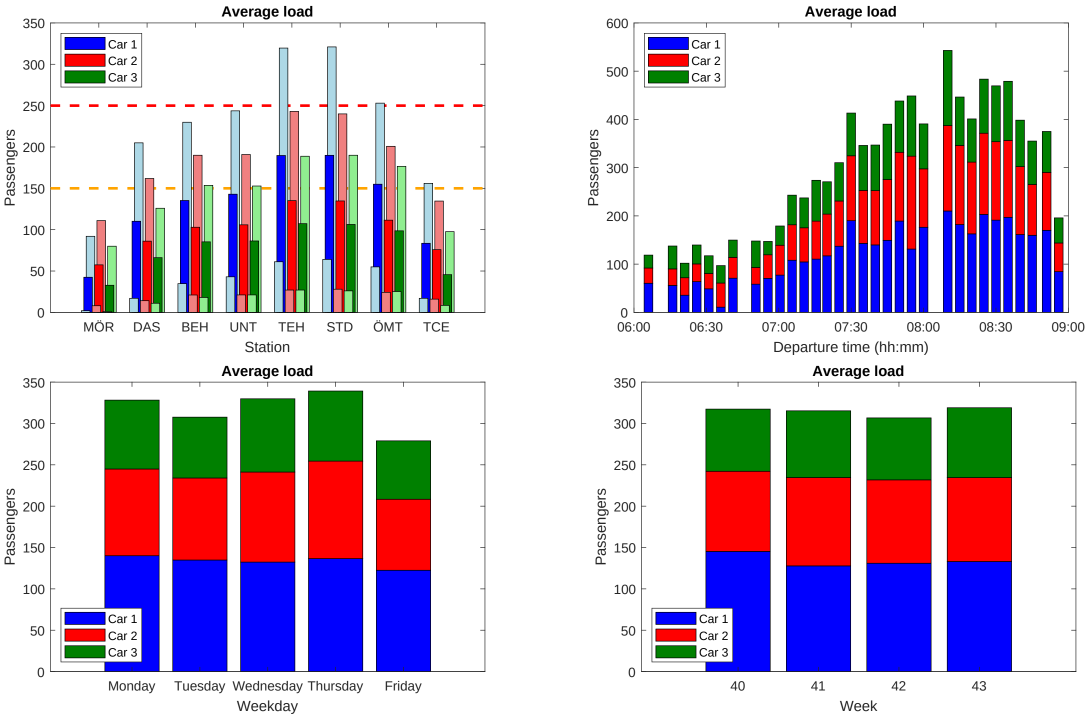

Fig. 4. Average passenger load at departure in the front (blue), middle (red), and rear (green) train car. Top left: station; dark colors indicate mean values, light colors indicate 10th and 90th percentiles (bars extend from 0 to indicated levels); orange and red dashed lines indicate lower limits of medium and high crowding levels. Top right: scheduled departure time from terminus MÖR (bars are stacked). Bottom left: weekday. Bottom right: week of the year.

Fig. 4 also shows that passenger loads vary significantly across different train runs, to some extent across weekdays and to a small extent across weeks of the year.

B. Load Measurement Error Analysis

When evaluating the performance of the prediction methods it will be assumed that the actual passenger loads are measured without error. While validation data are not available, this section models and analyses the measurement errors of the weight-based load data considering two sources of variability: (a) the distribution of individual passenger weights, and (b) measurement error in the train car air suspension system.

Percentiles of the weight distribution of the Swedish population are obtained from Statistics Sweden.[1] The percentiles reveals that the weight distribution is moderately skewed to the right. A lognormal distribution is fitted to the percentiles and the mean weight including luggage (78 kg), yielding distribution parameters $\mu_{\log} = 4.34$ and $\sigma_{\log} = 0.20$. Individual passenger weights w_n are assumed to be independently drawn from this distribution.

Train car weight measurement errors ε are assumed to be independently drawn from a normal distribution with mean 0 and standard deviation σ. Two values are considered: $\sigma = 0$, i.e., perfect measurements, and $\sigma = 1000$ kg, which corresponds to an error of 1.5% relative to the train car curb weight.

Given the true passenger load N, the measured load q is modeled as

$$q = \max\left\{ \text{round}\left(\frac{1}{78} \left(\sum_{n=1}^{N} w_n + \epsilon \right) \right), 0 \right\} \tag{14}$$

Fig. 5 shows the RMSE (top) and crowding level accuracy (bottom) of the measured load q as a function of the true passenger load N, calculated across 100 independent samples for each N. Without car weight measurement errors ($\sigma = 0$), passenger load errors are less than 4 across the whole range of loads. Crowding is correctly categorized for all loads except in the near vicinity of the crowding level thresholds. With 1.5% car weight measurements errors ($\sigma = 1000$ kg), passenger load errors are around 10 for all loads. Crowding level accuracy is reduced in a wider neighborhood around each crowding level threshold.

Overall, the analysis suggests that the deviations between true and measured train car loads are relatively moderate. Thus, measured car loads are considered relevant for crowding prediction and information provision.

C. Model Fitting

Out of the 20 days, 10 days (153 train trips with recorded load data) are randomly selected as the test set while the remaining 10 days (176 train trips with load data) are used as training set. Historical mean loads are calculated across all 20 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.

[1] Statistics Sweden (SCB), The Swedish Living Conditions Surveys (ULF/SILC). BMI, height and weight – percentiles 2010–2011. https://www.scb.se/hitta-statistik/statistik-efter-amne/levnadsforhallanden/levnadsforhallanden/undersokningarna-av-levnadsforhallanden-ulf-silc/pong/tabell-och-diagram/halsa/halsa--fler-indikatorer/, accessed February 2, 2019.

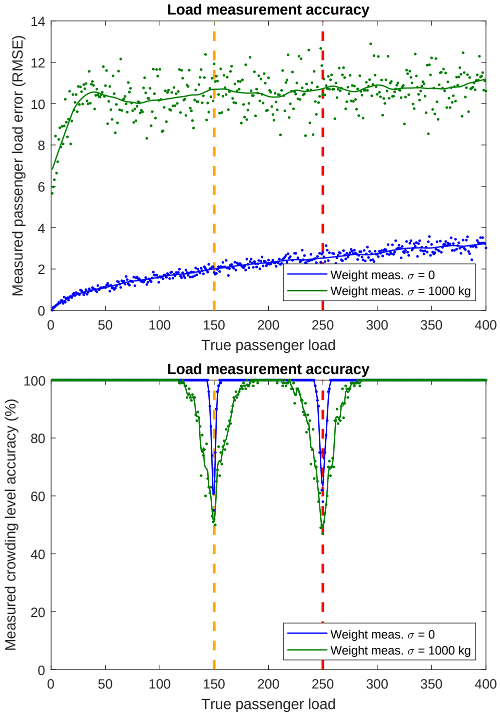

Fig. 5. Measured passenger load accuracy based on measurement model as function of true passenger load. Top: Passenger load RMSE. Bottom: Crowding level accuracy (%). Blue lines: No train car load measurement error. Green lines: Train car load measurement error $\sigma = 1000$ kg.

Before model fitting, the gross number of predictors (not counting quadratic and interaction terms in the stepwise regression models) varies from 12 in the models based only on historical data to 51 in the models with real-time target train data for $J' = 5$ stations up to and including the source station.

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 boosted regression tree ensemble. New trees are fitted with a shrinkage learning rate of 0.005. 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 only linear historical predictors (i.e., no interaction or quadratic terms and no real-time data predictors).

IV. Results

A. Lasso Regression for Specific Target Station

We first analyse the lasso method applied to specific target station UNT at various times of prediction. Table III, column 1, shows the estimated lasso regression model for each car just before the target train departs from the terminus MÖR; thus, historical and real-time target-station data are available but real-time target-train data are not available. The historical train run mean load for the own car is included in the model for all train cars. Interestingly, the train run mean load for car 3 is also included in all three models. All three models include some weekday mean load variables, and cars 2 and 3 include week-of-year mean load variables. Notably, current-day target-station mean loads do not significantly improve the predictive power and are not included in any of the models.

Table III, column 2, shows estimated regression models at the time just after the target train departs from terminus MÖR; thus, historical and real-time target-station data as well as real-time target-train data from MÖR are available. The explanatory power in terms of RMSE is considerably higher for all cars than without real-time target-train data (cf. left column). Thus, the number of boarding passengers at the terminus gives important information about the downstream crowding of each train car. As expected, the observed load of the same car at MÖR is included in the model for each car; the models for cars 2 and 3 also incorporates the total train load at MÖR. All models include historical data for the same vehicle run and, for cars 2 and 3, the same weekday. Current-day target-station mean loads are not included in any model. Interestingly, the crowding indicator for car 1 at MÖR is included in all models, suggesting that passengers adjust their car boarding choice based on in-vehicle crowding.

Table III, column 3, shows estimated regression models after the target train departs from station BEH just upstream of UNT. The explanatory power of the models is considerably higher for all cars than with real-time information from MÖR only. This is not unexpected since the average changes in passenger loads between BEH and UNT are small (Fig. 4, top left). For each car, the observed load of the same car at BEH is included. Compared to source station MÖR, fewer historical variables are included. Meanwhile, the crowding indicator for car 1 at MÖR is still included in all models.

Real-time target station predictors are consistently not included in the models, which suggests that the within-day correlation of train loads is low. This may partly be an effect of missing data since only about half of all train runs have recorded load data. Thus, in the following, only two types of predictors are considered: historical data and real-time target-train data.

Fig. 6 shows predicted loads for target station UNT plotted against measured loads for the test data set. Loads are predicted based historical data (models according to Table III column 1), with current-day train-specific data from source station MÖR (column 2) and from source station BEH (column 3), respectively. The dashed horizontal and vertical lines indicate the thresholds of the crowding levels.

As evident from the lower scatter around the diagonal, prediction accuracy in terms of both absolute loads and crowding levels increases with recent current-day train data. Historical data tend to underestimate the front car load on crowded runs, but overestimate the load on the least crowded runs. Thus, historical patterns alone cannot explain why some train cars are highly crowded. The classification accuracy for high crowding increases with real-time data from the terminus MÖR, and is almost perfect with real-time data from BEH.

B. Train-Centered Prediction

We now extend the analysis to train-centered prediction, i.e., from the perspective of an operator or for providing real-time crowding information inside the vehicle.

TABLE III
LASSO REGRESSION MODELS FOR TARGET STATION UNT.

Train car	(1) Historical data	(2) Source station MÖR	(3) Source station BEH
Predictor	Estimate	Estimate	Estimate
Car 1 (front)			
Intercept	-1.2769	-66.7076	6.1200
$q_{1,\text{MÖR}}$	—	1.1298	—
$q_{1,\text{BEH}}$	—	—	0.9860
$q_{2,\text{MÖR}}$	—	1.1298	—
$r_{1,\text{MÖR}}$	—	64.2670	8.8322
$r_{2,\text{MÖR}}$	—	5.0066	—
$h^{\text{run}}_{1,\text{UNT}}$	0.3756	0.0667	—
$h^{\text{run}}_{3,\text{UNT}}$	0.0011	—	0.0398
$H^{\text{run}}_{\text{UNT}}$	0.2245	0.2011	—
$h^{\text{wday}}_{1,\text{UNT}}$	—	—	—
$H^{\text{wday}}_{\text{UNT}}$	0.0582	—	—
$h^{\text{week}}_{1,\text{UNT}}$	—	0.5444	—
$H^{\text{week}}_{\text{UNT}}$	—	0.0161	—
λ_1	4.9440	2.4058	0.7408
Num. preds.	5	9	4
Num. obs.	148	121	148
RMSE	49.3752	36.7271	8.3270
Car 2 (middle)			
Intercept	-752.5725	-23.8056	2.7105
$q_{1,\text{MÖR}}$	—	0.0062	—
$q_{2,\text{MÖR}}$	—	1.1023	—
$q_{2,\text{BEH}}$	—	—	0.9517
$Q_{\text{MÖR}}$	—	0.0288	—
Q_{DAS}	—	—	0.0081
$r_{1,\text{MÖR}}$	—	81.6558	29.0171
$h^{\text{run}}_{2,\text{UNT}}$	1.1046	0.2627	—
$h^{\text{run}}_{3,\text{UNT}}$	0.3200	0.2355	0.0209
$h^{\text{wday}}_{1,\text{UNT}}$	0.0910	—	—
$h^{\text{wday}}_{3,\text{UNT}}$	3.5372	0.2316	—
$h^{\text{week}}_{1,\text{UNT}}$	2.0309	—	—
$h^{\text{week}}_{2,\text{UNT}}$	4.6421	—	—
λ_2	0.1451	1.8759	0.5907
Num. preds.	7	8	6
Num. obs.	148	121	144
RMSE	44.1209	22.8746	5.4004
Car 3 (rear)			
Intercept	-269.9233	-21.5141	2.9889
$q_{3,\text{MÖR}}$	—	0.8256	—
$q_{3,\text{BEH}}$	—	—	0.9653
$q_{1,\text{DAS}}$	—	—	0.0014
$Q_{\text{MÖR}}$	—	0.1718	—
$r_{1,\text{MÖR}}$	—	38.0247	11.4984
$h^{\text{run}}_{2,\text{UNT}}$	0.0324	0.0410	—
$h^{\text{run}}_{3,\text{UNT}}$	0.9372	0.3948	—
$H^{\text{run}}_{\text{UNT}}$	—	0.0097	—
$h^{\text{wday}}_{3,\text{UNT}}$	1.2567	0.2080	—
$H^{\text{week}}_{\text{UNT}}$	0.8478	—	—
λ_3	0.7934	2.0051	0.5258
Num. preds.	5	8	4
Num. obs.	148	121	146
RMSE	40.8684	24.8951	3.8853

Fig. 7 shows the prediction accuracy for each train car as a function of time to departure from UNT. More than 5.5 minutes before departure, prediction is based on historical data (Table III, column 1). Once the train departs from terminal station MÖR, predictions are updated according to the model for source station MÖR (Table III, column 2). 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 (Table III, column 3). Prediction errors decrease steadily for all cars as the time to departure decreases. Accuracy is generally lowest for the front (most crowded) car and highest for the rear (least 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.

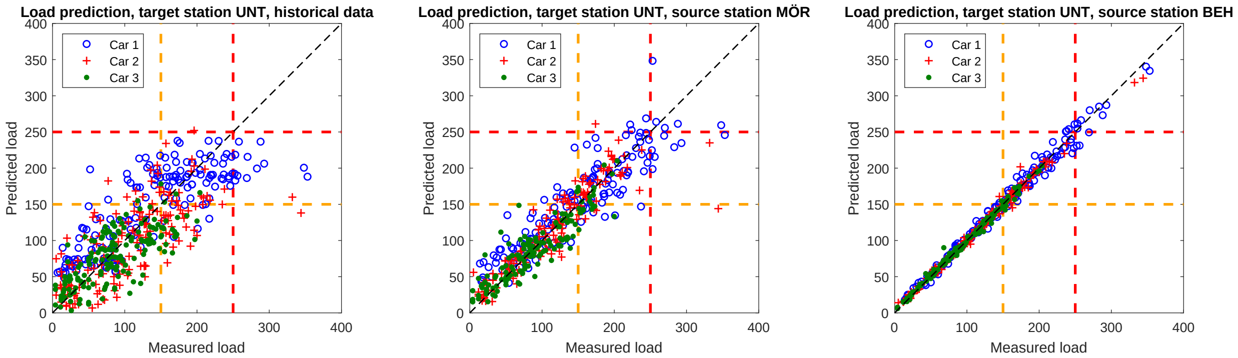

Fig. 6. Predicted vs. actual loads for test data set not used for model calibration. Target station UNT, lasso regression. Left: Prediction based on historical data. Middle, right: Prediction with current-day train-specific data from source station MÖR and BEH, respectively. Orange and red dashed lines indicate lower limits of medium and high crowding levels.

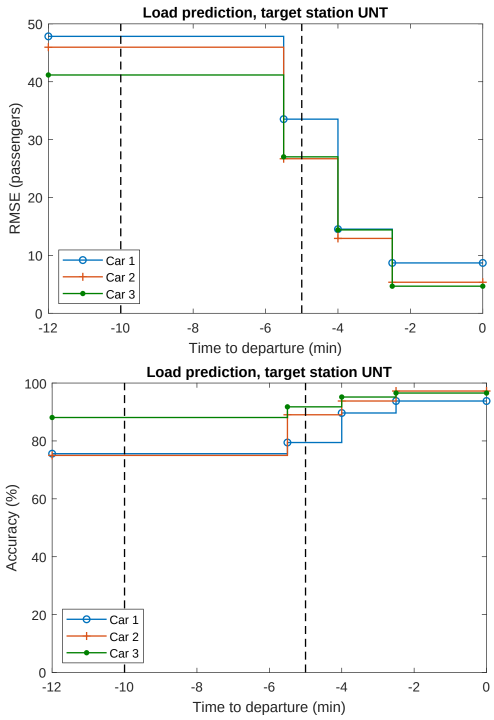

Fig. 7. Train-centered crowding prediction performance as function of time to departure. Target station UNT, lasso regression. Top: Passenger load RMSE. Bottom: Crowding level (low, medium, high) accuracy (%). Dashed lines indicate cumulative headways between train runs.

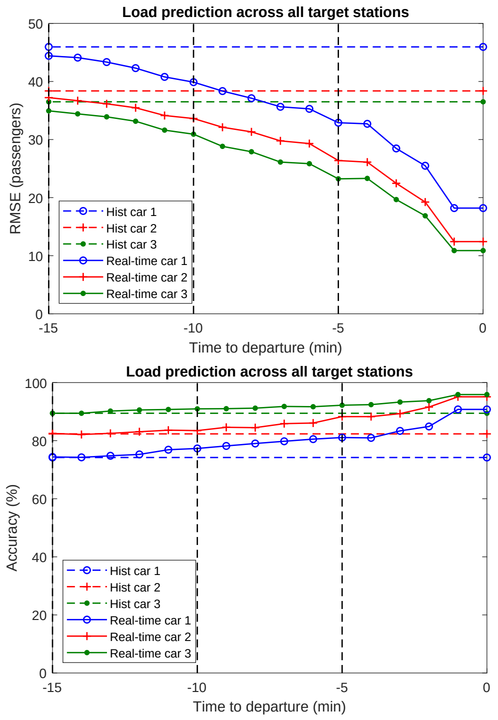

Fig. 8. Train-centered crowding prediction performance as function of time to departure. Average across all target stations, lasso regression. Top: Passenger load RMSE. Bottom: Crowding level (low, medium, high) accuracy (%). Dashed vertical lines indicate cumulative headways between train runs.

Fig. 8 shows the prediction accuracy for each train car as a function of time to departure from the target station, averaged over all stations from DAS to TCE. The horizontal dashed lines shows the prediction accuracy using only historical data while solid lines show prediction accuracy with the full set of real-time and historical data. Note that for a given prediction horizon, predictions are in general based on real-time data for some stations and historical data for some stations.

The results show that real-time load data significantly improves prediction accuracy. Prediction errors steadily decrease as the time to departure decreases. 15 minutes before departure the average RMSE is between 35 and 45 passengers depending on the car; 5 minutes before departure the average RMSE has dropped to between 25 and 35 passengers. Meanwhile, crowding level accuracy increases from between 70% and 90% to between 80% and 95%. Prediction errors are consistently highest for car 1 and lowest for car 3, which reflects the relative variability of car passenger loads (cf. Fig. 4, top left).

The reduction in load prediction errors is more dramatic than the increase in crowding level accuracy. This is mainly because prediction errors do not influence crowding level accuracy as long as true and predicted loads both belong to the same crowding level. This occurs most frequently at stations where loads rarely exceed the lowest level of crowding, in particular TCE (cf. Fig. 4, top left).

Fig. 9 shows the prediction accuracy of all three prediction

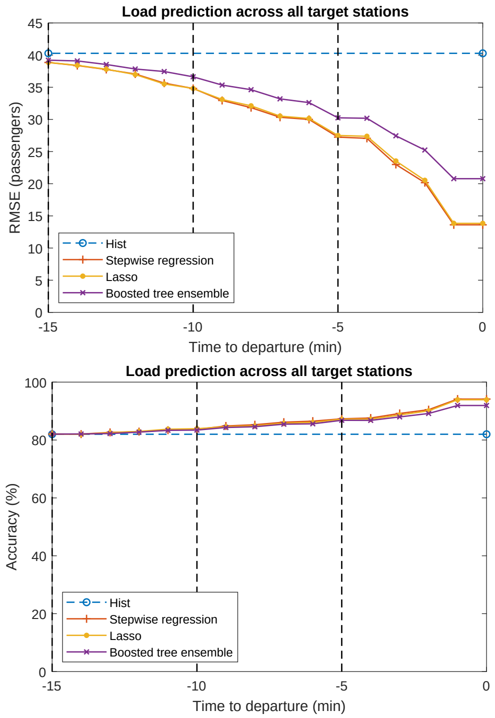

Fig. 9. Train-centered crowding prediction performance as function of time to departure. Average across all train cars and target stations, different prediction methods. Top: Passenger load RMSE. Bottom: Crowding level (low, medium, high) accuracy (%). Dashed vertical lines indicate cumulative headways between train runs.

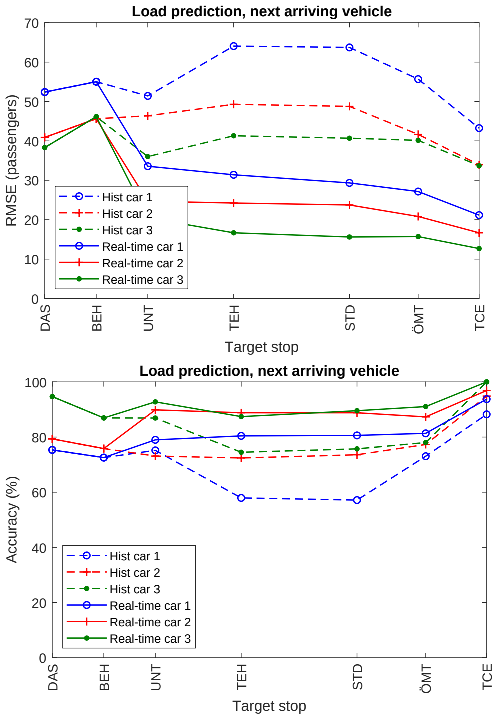

Fig. 10. Station-centered crowing prediction performance as a function of target station. Average across all times of day, lasso regression. Top: Passenger load RMSE. Bottom: Crowding level (low, medium, high) accuracy (%).

methods averaged over all train cars. The dashed line shows the prediction accuracy using stepwise regression only on historical data. Prediction errors are highly similar for stepwise regression and lasso while somewhat higher for the boosted tree ensemble. This model has a comparative disadvantage, however, since the same tree template was used for all model instances. Prediction accuracy would likely improve if specific tree templates were calibrated for every problem instance. The small difference between stepwise regression and lasso indicates that quadratic and interaction terms are of limited importance for prediction performance. This suggests that the value of using higher-order non-linear models may be limited.

C. *Station-Centered Prediction*

We now consider station-centered prediction accuracy, i.e., from the perspective of providing real-time crowding information about the next arriving train at a station.

Fig. 10 shows the average prediction performance for the next arriving train for each target station from DAS to TCE, evaluated at the departure time of the current train from the target station (i.e, on average 5 minutes prediction horizon). For stations DAS and BEH, the next train has not yet departed from terminus MÖR, which implies that only historical data are available. From station UNT and onward, real-time target-train data are available, and prediction errors decrease sharply.

The relative benefit of real-time data for prediction accuracy is the largest for stations TEH and STD where load variability between days and train runs is the highest (cf. Fig. 4, top left). Towards the end of the line crowding level accuracy increases even though absolute load errors remain stable, since most trains are at the low crowding level. Notably, prediction accuracy is the highest for stations ÖMT and TCE even though both are transfer stations, which implies that the influence of transferring passenger flows is limited.

Fig. 11 shows the prediction accuracy across the day for the next arriving train averaged over all target stations. Crowding level accuracy (Figure 11, bottom), in particular, show a similar temporal pattern as the average load (cf. Figure 4, top right). During the early morning crowding is generally low and prediction is highly accurate. During the morning peak both average loads and variability between days increase, and prediction is more challenging. Real-time data significantly improves prediction accuracy during peak hours compared to historical data only.

V. Conclusion

The paper contributes to the small literature on public transit crowding predicting by formulating the car-specific metro train crowding prediction problem based on real-time load data and evaluating the performance of several data-driven prediction methods. Two perspectives of the problem, utilizing the same prediction model, are considered: (1) train-centered prediction, of value for in-vehicle RTCI provision, and (2) station-centered prediction, useful for RTCI provision at stations.

The application to a metro line in Stockholm shows that three data-driven prediction methods, stepwise regression, lasso and boosted tree ensembles, perform similarly. When available, predictions with current-day train-specific load data

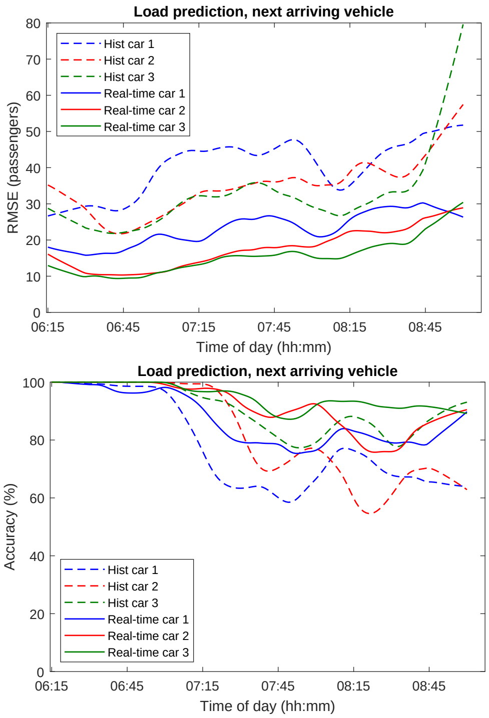

Fig. 11. Station-centered crowing prediction performance as a function of time of day. Average across all target stations, lasso regression. Smoothed using local linear regression with bandwidth 30 minutes. Top: Passenger load RMSE. Bottom: Crowding level (low, medium, high) accuracy (%).

significantly outperform historical patterns, 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, partially due to missing data. Passenger load prediction errors are smallest for train cars, target stations and times of the day with the lowest variability in loads between train runs and days.

The results show that accurate RTCI at stations can be provided long before the trains depart from the nearest upstream station. This implies that travellers have sufficient time to re-evaluate their route, train and train car choices in response to the crowding conditions that will be encountered. Considering the positive results on train car load balancing from earlier pilot studies with less timely information [22] and simulation experiments [24], [25], the results suggest that real-time crowding information can be an effective means to reduce in-vehicle crowding and passengers' discomfort, increase service performance and reduce operation costs also in practice.

RTCI would also allow operators to continuously monitor crowding and take appropriate actions if they reach high levels. Examples of real-time strategies for mitigating crowding could include passenger metering, stop-skipping, short-turning and real-time stop position selection. Exploring these possibilities are directions for future research.

It is common in many metro networks to use articulated trains, which permit passengers to move between the cars and help improving the distribution of passengers in the cars. Given that load data at departure from stations are available for each car the proposed prediction method can be applied without modification also for such trains, although distinct model parameters need to be calibrated for each vehicle type. More generally, the framework is applicable to other multi-unit vehicles such as commuter trains, trams and light rail transit. For trains without car-specific load data the prediction method can easily be adapted to directly predict the total in-vehicle passenger load.

This paper focused on the problem of predicting absolute passenger loads, although results are also shown regarding the ability to predict discrete crowding levels. In some settings predicting discrete crowding levels may be the main focus; in such cases, formulating the task as a classification problem and solving it accordingly may further increase prediction accuracy. Regarding the load prediction problem considered here, an interesting area of further work is to evaluate whether more sophisticated methods can further increase accuracy. For the current case study, however, interaction and quadratic terms of the predictors do not significantly improve prediction accuracy, which suggests that the added benefits of higher-order non-linear models may be limited.

The proposed methodology is focused on providing accurate RTCI under the typical range of traffic conditions based on historical and real-time load measurements. Under anomalous conditions such as special events and incidents, data-driven approaches require that similar events are captured in the historical data set. An interesting research question is whether other data sources, e.g., weather conditions such as precipitation and temperature, can improve predictions. It is straightforward to extend the proposed framework to include such data auxiliary data sources.

An interesting research question to explore further is to what extent predictions can be improved by considering transferring passenger loads from other lines. Another direction for further research is to extend the crowding prediction framework to other public transit modes, e.g., high-frequency buses.

Acknowledgment

The author would like to thank MTR Nordic for kindly providing the metro train car load data. Comments and suggestions from the editor and several anonymous referees have helped to significantly improve the paper.

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] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] S. Peftitsi, E. Jenelius, and O. Cats, "Determinants of passengers' metro car choice revealed through automated data sources: A Stockholm metro case study," CASPT 2018, Brisbane, Australia, 23-25 July, 2018.

[10] 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.

[11] Z. Christoforou, P.-A. Collet, B. Kabalan, F. Leurent, A. de Feraudy, A. Ali, T. J. Arakelian-von Freeden, and Y. Li, "Influencing longitudinal passenger distribution on railway platforms to shorten and regularize train dwell times," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2648, pp. 117–125, 2017.

[12] K. Sohn, "Optimizing train-stop positions along a platform to distribute the passenger load more evenly across individual cars," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, 2013.

[13] J. C. Muñoz, J. Soza-Parra, A. Didier, and C. Silva, "Alleviating a subway bottleneck through a platform gate," *Transportation Research Part A*, vol. 116, pp. 446–455, 2018.

[14] E. Mazloumi, G. Rose, G. Currie, and M. Sarvi, "An integrated framework to predict bus travel time and its variability using traffic flow data," *Journal of Intelligent Transportation Systems*, vol. 15, no. 2, pp. 75–90, 2011.

[15] Y. Liu, T. Tang, and J. Xun, "Prediction algorithms for train arrival time in urban rail transit," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–6.

[16] R. Zhang, W. Liu, Y. Jia, G. Jiang, J. Xing, H. Jiang, and J. Liu, "Wifi sensing-based real-time bus tracking and arrival time prediction in urban environments," *IEEE Sensors Journal*, vol. 18, no. 11, pp. 4746–4760, 2018.

[17] K. Dziekan and K. Kottenhoff, "Dynamic at-stop real-time information displays for public transport: effects on customers," *Transportation Research Part A: Policy and Practice*, vol. 41, pp. 489–501, 2007.

[18] 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," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 8, pp. 839–848, 2011.

[19] O. Cats and E. Jenelius, "Dynamic vulnerability analysis of public transport networks: mitigation effects of real-time information," *Networks and Spatial Economics*, vol. 14, no. 3–4, pp. 435–463, 2014.

[20] L. Eboli and G. Mazzulla, "Relationships between rail passengers' satisfaction and service quality: a framework for identifying key service factors," *Public Transport*, vol. 7, pp. 185–201, 2015.

[21] Land Transport Authority, "Factsheet: Passenger load information system piloted on downtown line for smoother boarding," 2018.

[22] 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.

[23] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where were going," *Transportation Research Part C: Emerging Technologies*, vol. 43, Part 1, pp. 3–19, 2014.

[24] 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.

[25] 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.

[26] H. N. Koutsopoulos, 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, 2019, pp. 229–261.

[27] J. Zhang, D. Shen, L. Tu, F. Zhang, C. Xu, Y. Wang, C. Tian, X. Li, B. Huang, and Z. Li, "A real-time passenger flow estimation and prediction method for urban bus transit systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3168–3178, 2017.

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

Erik Jenelius is Associate Professor in the Division of Transport Planning at KTH Royal Institute of Technology in Stockholm. He is the manager of the Urban Mobility Group, which aims to tackle the challenges of congestion, crowding, and environmental impacts in urban transport systems through the development of methods based on statistical analysis, network modelling, computer simulation and optimization approaches. He is the Director of the iMobility Lab, focused on the use of emerging sensor technologies to address urban mobility problems. His research interests include data-driven traffic and mobility management, transportation network and resilience analysis, and transit systems planning and operations.