Abstract

The paper proposes a methodology for providing personalized, predictive in-vehicle crowding information to bus travellers. Three crowding metrics are considered: (1) the probability of getting a seat on boarding, (2) the expected travel time standing, and (3) the excess perceived travel time compared to uncrowded conditions. The traveller can request the information for a specific trip, customized to the traveller’s crowding preferences, and disseminated back via mobile applications or bus stop displays. The methodology combines a bottom-up framework for prediction of passenger loads and alighting counts based on lasso regression with a probabilistic model of in-vehicle travel conditions in order to predict the crowding metrics. Depending on data availability, the prediction method can use a combination of historical passenger counts, real-time vehicle locations and real-time passenger counts to predict passenger counts. We evaluate the effect on prediction performance of different levels of data availability in a real-world case study for a bus line in Stockholm, Sweden. The results indicate that personalized, predictive crowding information that is robust to varying data availability can be provided sufficiently early to be useful to travellers. The methodology is of value for agencies and operators in order to increase the attractiveness and capacity utilization of public transport.

1 Introduction

Efficient, convenient and reliable public transport is a vital part of a sustainable urban transport system. Crowding inside vehicles has negative effects on all these aspects. Traveler satisfaction and comfort are known to decrease at higher crowding levels [2, 18]. For buses in particular, dwell time impacts of crowding due to longer boarding and alighting times are substantial [27].
There is also a negative feedback loop between crowding and arrival time irregularity, which leads to deteriorating passenger waiting times and experienced crowding [6, 1]. All in all, the negative consequences of crowding inhibit the transition from private to public transport in urban areas.

As the urban populations grow, the need increases to utilize available capacity in the transport system more efficiently. There are multiple supply-side methods for combating in-vehicle crowding, ranging from the strategic planning level (increased service frequency, vehicle capacities, network expansions, etc.) to the real-time control level (holding, stop-skipping, extra departures, etc.). Measures that target the travel demand, meanwhile, are less developed.

The ongoing digitalization of the public transport system has lead to the emergence of several valuable data sources, including automated vehicle locations (AVL), automated passenger counts (APC) and automated fare collection (AFC) data. In recent years, many cities have started to utilize AVL data for providing real-time bus arrival time information at stops and in mobile applications. Real-time arrival time information systems have positive effects on traveller satisfaction [10], perceived waiting times [33], safety and security [9], and service disruption impacts [3]. In this context several methods for short-term arrival time prediction have been proposed [22, 4, 35].

Systems providing real-time crowding information (RTCI) for buses are still rare. For urban rail transit systems, meanwhile, RTCI is being introduced in many cities by public transport agencies and private mobility information providers. RTCI provision allows travellers to make better informed decisions about whether to board a vehicle or not based on their preferences for crowding, waiting time, walking distance, etc. RTCI may thus directly increase passenger satisfaction as well as indirectly increase satisfaction and service quality by reducing negative crowding externalities [36].

Providing timely crowding information to passengers waiting at or arriving to a stop generally requires prediction of crowding conditions based on the most recent information about the system state. Zhang et al. consider the problem of predicting the passenger loads on buses downstream of their current locations based on real-time AVL and AFC data [34]. The prediction is done in a sequence of two steps, where the first step involves identifying the historical day most similar to the current day and predicting downstream loads based on observed loads from the historical day. In the second step the passenger loads are updated by combining real-time and historical data in an extended Kalman filter. The methodology is evaluated
on data from a bus line in Shenzhen, China.

In the context of trains, Khomchuck et al. propose a Bayesian approach to predicting the passenger loads in individual train cars downstream of their current locations based on APC data [17]. Passenger OD flows are assumed to follow a Poisson prior distribution with parameters estimated from historical data, and the posterior distribution is updated based on current-day observed loads. The approach, which also predicts boarding and alighting counts, is evaluated in a simple synthetic case study. Jenelius uses several regression models combining real-time and historical APC data to predict individual train car loads, and evaluates the performance on a metro line in Stockholm, Sweden [16]. Pasini et al. propose a long short-term memory (LSTM) encoder-predictor neural network to predict train loads based on temporal features (day-of-week, time-of-day, etc.) and current-day load measurements from previous train departures at the same station. The method is applied to a railway line in Paris, France using boarding and alighting data from on-board radar sensors.

Simple prediction schemes based on the crowding of the one or two most recent vehicle runs [7] as well as more complex schemes involving running the simulation model forward to a fixed point solution [23] have been proposed and evaluated using simulation models. Both studies demonstrate that predictive RTCI facilitates a more even crowding distribution among metro runs and reduces passengers’ experienced travel time.

Ling et al. predict incoming passenger flows to metro stations, aggregated in 15-min intervals, based on AFC data from preceding time intervals [21]. A gradient boosted regression trees model is trained using historical AFC data from Shenzhen, China. This work does not consider passenger loads or crowding onboard trains. Some papers consider public transport ridership prediction for short-term planning rather than real-time information applications. Tsai et al. use several neural network architectures to predict daily ridership on a train line in Taiwan based on temporal features, but do not consider real-time data or load variations along the line [30]. van Oort et al. use a combination of AFC data, network assignment and elasticity parameters to predict passenger load changes due to strategic decisions such as frequency and route changes [31]. Pereira et al. predict the number of public transport trips to special event locations (e.g., stadiums) based on AFC data and local event information obtained from the Internet. The developed model is designed to predict arrivals several days in advance [24].

The aforementioned studies provide valuable knowledge about the ability to predict passenger loads based on combinations of historical and real-time data and the effectiveness of different prediction techniques. However, some
remaining open research questions can be identified. First, existing studies have focused on predicting passenger loads (in one case also boarding and alighting counts [17]). From a passenger’s perspective, however, the load alone does not fully capture the encountered crowding conditions throughout the trip. There is a need to predict personalized crowding indicators that consider the origin and destination stops of the particular trip and that better reflect the individual traveller’s crowding experience.

Second, all previous papers have each considered only one set of available data sources. However, it is of interest to develop a framework that can handle heterogeneous data availability, and to compare prediction performance under different data scenarios. This facilitates crowding prediction when sensor equipment varies among vehicles, and allows operators and agencies to assess the benefits of investing in additional sensors or communication technology that may be required.

Third, there is still a lack of studies that focus on crowding prediction for buses rather than trains, and further knowledge is needed on the dynamics and predictability of bus crowding.

The aim of this paper is to address the identified research gaps above. It makes the following contributions:

- The paper moves beyond passenger load as crowding indicator and considers more relevant indicators for passengers. Three crowding metrics are considered: (1) the probability of getting a seat on boarding, (2) the expected travel time standing, and (3) the excess perceived travel time compared to uncrowded conditions. Metrics (2) and (3) consider an aggregate of the entire passenger journey from origin stop to destination stop. Metric (3) further incorporates the subjective perception of crowding, which can be customized to the individual traveller.

- The paper proposes a bottom-up approach that reduces the combinatorial complexity of predicting traveller-specific crowding metrics by building predictions from stop-level loads and alighting count predictions. The framework utilizes the probabilistic seat allocation model in [14].

- The paper proposes a lasso regularized linear regression prediction method of passenger loads and alighting counts that facilitates multiple combinations of data sources: historical load data, real-time AVL data and real-time APC data. The approach builds on the load prediction method in [15].
We evaluate the effect on prediction performance of different levels of data availability in a real-world case study for a bus line in Stockholm, Sweden.

The paper is organized as follows. Section 2 presents the methodology including the proposed personalized crowding metrics, the probabilistic seat allocation model, the proposed prediction method, and the considered crowding predictors. Section 3 introduces the real-world case study including the utilized data and the model fitting. Section 4 presents the results of the case study, and Section 5 concludes the paper.

2 Methodology

This section first introduces a set of passenger-oriented crowding metrics, and proposes a way to compute the metrics based on automatic passenger count (APC) data of loads and alighting counts. The section then shows how the crowding metrics for a particular bus trip can be predicted based on predicted passenger loads and alighting counts. A prediction method based on lasso regression is proposed, and sets of predictors based on historical APC data and real-time automatic vehicle location (AVL) and APC data are introduced. The notation used in the paper is listed in Table 1.

2.1 Personalized crowding metrics

A range of crowding metrics have been proposed in the literature, some targeted towards planners and operators and others towards travellers. The focus here is on personalized metrics, i.e., metrics that are relevant for an individual traveller making a journey between a particular origin stop and destination stop. Furthermore, we consider both objective metrics and metrics that incorporate subjective valuations of varying crowding conditions. The metrics are also chosen to be as easy as possible to interpret for the traveller.

Probability of seat on boarding

A clear distinction can be made between travelling seated and travelling standing. The value of having a seat varies from being a necessity for some travellers, to a minor comfort factor for others. In particular for the first category, a relevant indicator of in-vehicle crowding conditions is the chance
of getting a seat immediately when boarding the vehicle, denoted

\[ \pi_{o}^{brd} \]

for any journey origin stop \( o = 1, \ldots, K - 1 \).

**Travel time standing**

For travellers who prefer sitting over standing but do not strictly require a seat the whole trip, a relevant metric of in-vehicle comfort is the expected travel time standing, i.e., the time until getting a seat or until alighting, whichever occurs first. This crowding metric depends on both the origin and the destination of the traveller. Given the probability of getting a seat on boarding \( \pi_{o}^{brd} \), the in-vehicle probability \( \pi_{k}^{inv} \) of getting a seat at stop \( k \), the probability that the probe traveller is standing on line segment \( k = o, \ldots, d - 1 \) is equal to the probability that the traveller does not get a seat.
seat at stop $k$ or any preceding stop,

$$P_{k|o} = (1 - \pi_o^{brd}) \prod_{k'=o+1}^{k} (1 - \pi_{k'}^{inv}). \tag{2}$$

Given the in-vehicle travel time $\tau_k$ on segment $k$, the expected travel time standing is computed as

$$T_{od}^\text{std} = \sum_{k=o}^{d-1} \tau_k P_{k|o} \tag{3}$$

for any journey between origin stop $o = 1, \ldots, K - 1$ and destination stop $d = o + 1, \ldots, K$.

**Excess perceived travel time**

Travellers experience time differently depending on the circumstances under which the time is spent [26, 19]. Several studies have estimated travellers’ willingness to pay for shorter travel times under different crowding conditions, typically expressed as time multipliers to in-vehicle time under nominal conditions [32, 20, 28, 12]. We assume here that crowding conditions are discretized into $M$ levels based on the load factor $q_k/c^{sit}$. The indicator variable $I_{km}$ is 1 if the experienced crowding conditions are at level $m$ and 0 otherwise,

$$I_{km} = \begin{cases} 1 & \phi_m \leq q_k/c^{sit} < \phi_{m+1}, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

for a sequence of thresholds $\phi_m$ such that $\phi_1 = 0$, $\phi_m < \phi_{m+1}$ for all $m = 1, \ldots, M$.

Let $\beta_m^{sit}$ and $\beta_m^{std}$ denote the time multipliers for in-vehicle time seated and standing, respectively, at crowding level $m$. The parameters $\beta_m^{sit}$ and $\beta_m^{std}$ could be taken from planning standards, or customized by the individual traveller. The latter approach yields a highly personalized crowding metric, capturing the characteristics of the particular journey as well as the attitudes of the traveller.

The *excess perceived travel time* $T_{od}^{exs} \geq 0$ is the difference between expected perceived and nominal in-vehicle journey time,

$$T_{od}^{exs} = \sum_{k=o}^{d-1} \tau_k \sum_m I_{km} \left( P_{k|o}^{std} \beta_m^{std} + (1 - P_{k|o}^{std}) \beta_m^{sit} \right) - T_{od} \tag{5}$$

for any journey between origin stop $o = 1, \ldots, K - 1$ and destination stop $d = o + 1, \ldots, K$. 


2.2 Probabilistic seat allocation

The seated status of the traveller is modelled probabilistically consistent with observed passenger loads and boarding and alighting counts. The model makes the same seat priority assumptions as [25], [13] and [14]. Thus, sitting passengers are guaranteed a seat until alighting, and standing passengers are given priority over boarding passengers to the seats that become available when sitting passengers alight.

The number of boarding and alighting passengers, respectively, at stop $k$ are denoted $b_k$ and $a_k$, and the passenger load on the line segment between stops $k$ and $k+1$ is denoted $q_k$. The boarding and alighting counts and on-board loads are related as $q_k = q_{k-1} + b_k - a_k$. It is convenient to define $q_0 = 0$ as the load prior to departure from the first stop.

Seat availability at boarding depends on the existing passenger load on the bus and the number of boarding passengers. The probability that the traveller receives a seat is 0 if the passenger load before boarding, $q_{k-1} - a_k$, exceeds the seat capacity $c_{sit}$, and is 1 if the passenger load after boarding, $q_k$, is lower than $c_{sit}$. Otherwise, the probability of receiving a seat is modelled as the ratio of available seats to the number of boarding passengers,

$$\pi_{brd}^k = \begin{cases} 
0 & q_{k-1} - a_k > c_{sit} \\
\frac{c_{sit} - q_{k-1} + a_k}{q_k - q_{k-1} + a_k} & q_{k-1} - a_k \leq c_{sit}, q_k > c_{sit} \\
1 & q_k \leq c_{sit}.
\end{cases}$$

(6)

Inside the vehicle, seats that become available as passengers alight are filled by standing passengers. Conditional on that the traveller is standing on segment $k-1$, the probability of getting a seat on segment $k$ is 1 if the passenger load before boardings is lower than the seated capacity. Otherwise, alighting passengers are randomly selected among all on-board passengers. Thus, the number of seated passengers alighting is drawn among all passengers according to a hypergeometric distribution. The seats that become available are filled by passengers randomly selected among all remaining standing passengers. All in all, the probability of getting a seat is

$$\pi_{inv}^k = \begin{cases} 
\sum_{a=0}^{a_k} \min \left( \frac{a}{q_{k-1} - a_k - c_{sit} + a}, 1 \right) \frac{c_{sit}}{a_k} \left( \frac{q_{k-1} - c_{sit}}{a_k - a} \right) & q_{k-1} - a_k > c_{sit} \\
1 & q_{k-1} - a_k \leq c_{sit}
\end{cases}$$

(7)

Figure 1 illustrates the different crowding indicators for a particular passenger trip on bus line 4 in Stockholm, Sweden based on AVL and APC data.
Figure 1: Crowding indicators for a particular trip on bus line 4, Stockholm (26 February 2016, departure 15:52) from origin stop 18 (Västerbroplan) to destination stop 31 (Gullmarsplan).

The solid black line shows that the probability of getting a seat on boarding is around 0.2. The seat probability then gradually increases as passengers alight and seats become available. The dashed black line shows the cumulative expected travel time standing, which remains constant from the stop where the passenger is sure to get a seat to the destination stop. The dotted black line shows the cumulative excess perceived travel time based on time multipliers from [32]. This crowding indicator continues to increase even as the passenger is seated, since the perceived in-vehicle time also depends on the load factor.

2.3 Predictive crowding information

Consider a traveller who intends to board the bus line at origin stop o at time $t$ and travel to destination stop $d$. We let $s$ denote the source stop, i.e., the most recently visited stop from which real-time data are available for the relevant bus at the time $t' < t$ when the traveller requests the crowding information. To provide information to the traveller prior to boarding about expected crowding conditions, the crowding metrics must be predicted based on information available at time $t'$. 
For crowding metrics that depend on the origin and destination of the trip a combinatorial challenge arises as the number of combinations of source, origin and destination stops grows cubically with the number of stops on the line. This paper takes a bottom-up approach which utilizes the fact that all proposed crowding metrics can be computed based on passenger loads and alighting counts at relevant stops. Thus, the crowding metrics are predicted by substituting predicted passenger loads and alighting counts into the crowding metric definitions, which reduces the number of required predictions to quadratically proportional to the number of stops. The approach allows multiple crowding metrics to be calculated without the need for multiple predictions. Furthermore, the predicted excess perceived travel time metric can easily be personalized to different traveller’s crowding perceptions by combining predicted load factors with customized time multipliers.

Figure 2 illustrates the main concepts of the approach. The earlier that the traveller requests the crowding information, the longer distance between the source stop $s$ and origin stop $o$. The predicted load and alighting count at stop $k$ given information from source stop $s$ are denoted $\hat{q}_{k|s}$ and $\hat{a}_{k|s}$, respectively. Here, stop $k$ is referred to as the target stop for the prediction.

The predicted probability of getting a seat upon boarding is

$$\hat{\pi}_{o|s}^{brd} = \pi_{o}^{brd} (\hat{q}_{o-1|s}, \hat{q}_{o|s}, \hat{a}_{o|s})$$

As can be seen, the predictive crowding information requires three predictions: the loads on the previous segment $\hat{q}_{o-1|s}$ and after departing from the origin stop $\hat{q}_{o|s}$, and the alighting count at the origin stop $\hat{a}_{o|s}$.

Similarly, the predicted travel time standing is

$$\hat{T}_{od|s}^{std} = T_{od}^{std} (\hat{q}_{o-1|s}, \ldots, \hat{q}_{d-1|s}, \hat{a}_{o|s}, \ldots, \hat{a}_{d-1|s})$$

and the predicted excess perceived travel time is

$$\hat{T}_{od|s}^{exs} = T_{od}^{exs} (\hat{q}_{o-1|s}, \ldots, \hat{q}_{d-1|s}, \hat{a}_{o|s}, \ldots, \hat{a}_{d-1|s})$$.
Both metrics require that loads and alighting counts are predicted for all stops between the origin stop and the destination stop, in total \(2(d - o) + 1\) predictions. Further, both metrics require predicted segment travel times \(\tau_k\) between the origin stop and the destination stop. Multiple methods for bus travel time prediction have been proposed, e.g., [22, 4, 35]. Since the focus of this paper is on the prediction of crowding conditions rather than travel times, the historical mean travel time on each segment is used to compute the crowding metrics.

A drawback of the bottom-up prediction approach is that the aggregated crowding metrics may be biased from the nonlinear transformations of predicted loads and alighting counts. In this paper we consider a simple correction method for the bias by subtracting the mean error on the training set from the baseline prediction for each combination of source, origin and destination stop. Thus, the bias corrected predictions of the travel time standing and the excess perceived travel time are

\[
\hat{T}_{\text{std}|s} = \hat{T}_{\text{std}|s} - \delta_{\text{std}}^{\text{ods}} \tag{11}
\]

\[
\hat{T}_{\text{exs}|s} = \hat{T}_{\text{exs}|s} - \delta_{\text{exs}}^{\text{ods}} \tag{12}
\]

where \(\delta_{\text{std}}^{\text{ods}}\) and \(\delta_{\text{exs}}^{\text{ods}}\) are the mean errors across the training set,

\[
\delta_{\text{std}}^{\text{ods}} = \frac{1}{N_{\text{trn}}} \sum_{i \in I_{\text{trn}}} \left( \hat{T}_{\text{std}|i,s} - T_{\text{std}|i} \right) \tag{13}
\]

\[
\delta_{\text{exs}}^{\text{ods}} = \frac{1}{N_{\text{trn}}} \sum_{i \in I_{\text{trn}}} \left( \hat{T}_{\text{exs}|i,s} - T_{\text{exs}|i} \right). \tag{14}
\]

Here \(I_{\text{trn}}\) is the training set of observations and \(N_{\text{trn}}\) the number of training observations.

### 2.4 Prediction method

This section presents a method for predicting the passenger loads and alighting counts based on lasso regularized linear regression [11]. The method has been previously applied to bus load prediction [15]. Thus, the passenger load \(\hat{q}_{ik}|s\) and alighting count \(\hat{a}_{ik}|s\) at target station \(k\) are assumed to be linear functions of a common set of predictors \(x_{isk}\) whose values depend on the particular bus trip \(i\). In order to represent actual numbers of passengers, predicted alighting counts \(\hat{a}_{ik}|s\) are rounded to the nearest non-negative integer. Further, the
alighting counts are restricted to not exceed the predicted load at the preceding stop,

\[ \hat{a}_{ik|s} = \min \left\{ \max \left\{ \text{round} \left( x_{isk} v_{sk}^T \right) , 0 \right\} , \hat{q}_{i,k-1|s} \right\} \]  

(15)

Here \( v_{sk} \) is a parameter vector, specific to each \((s,k)\) combination, which is estimated so as to minimize the regularized mean squared error,

\[ \frac{1}{2} \sum_{i \in I_{trn}} \left( a_{ik} - x_{isk} v_{sk}^T \right)^2 + \lambda_{sk} \| v_{sk} \|_1 , \]  

(16)
on the training data set. The regularization hyper-parameter \( \lambda_{sk} \geq 0 \) penalizes large parameter values. Larger \( \lambda_{sk} \) enforce sparser solutions, i.e., more parameters equal to zero. The \( \lambda_{sk} \) value is calibrated separately for each \((s,k)\) combination to minimize the cross-validation mean squared error.

Similarly, the predicted passenger load at departure \( \hat{q}_{ik|s} \) is rounded to the nearest non-negative integer, and restricted to not be below the load after alighting,

\[ \hat{q}_{ik|s} = \max \left\{ \text{round} \left( x_{isk} w_{sk}^T \right) , \hat{q}_{i,k-1|s} - \hat{a}_{ik|s} \right\} \]  

(17)

Here, \( w_{sk} \) is a parameter vector which is estimated using lasso regularization,

\[ \frac{1}{2} \sum_{i \in I_{trn}} \left( q_{ik} - x_{isk} w_{sk}^T \right)^2 + \gamma_{sk} \| w_{sk} \|_1 , \]  

(18)
where \( \gamma_{sk} \geq 0 \) is the regularization hyper-parameter.

2.5 Crowding predictors

Figure 3 shows bus trajectories in space-time diagrams from two sample days out of the data set used in the case study in Section 3. For the subset of bus trips with available APC data the passenger load and alighting counts are indicated. The trajectories show some evidence of systematic load and alighting variations among stops, correlations with upstream loads, and correlations with the headway to the preceding bus.

The prediction method is versatile in terms of the type of variables that are used to predict the passenger loads and alighting counts. We consider three main categories of predictor variables: variables based on (i) historical passenger count data, (ii) real-time vehicle location data, and (iii) real-time passenger count data. The specific variables used in each category are described below. These variables represent a gross list of potential predictors; as explained above the lasso regression maintains only a subset of relevant variables in the final regression model.
Historical APC data

The prediction can utilize temporal patterns in passenger counts at the target stop extracted from historical data. This may be the only available information unless buses are equipped with real-time AVL and APC systems. Even with such systems installed, only historical information are available before the buses depart from the first stop on the line. Further, historical averages provide a robust baseline from which real-time data can
It is well known that passenger counts tend to vary systematically with the time of day, day of week, time of year, etc. Specifically, for each bus departure \(i\) we consider the historical mean load and alighting count for departures during the same time-of-day interval \((\bar{q}_{\text{tod}}^{i,k}, \bar{a}_{\text{tod}}^{i,k})\), weekday \((\bar{q}_{\text{wday}}^{i,k}, \bar{a}_{\text{wday}}^{i,k})\) and month \((\bar{q}_{\text{mnth}}^{i,k}, \bar{a}_{\text{mnth}}^{i,k})\) at target station \(k\). We also include two interaction terms. The historical predictors are collected in the \(1 \times 8\) vector

\[
\begin{align*}
\mathbf{x}^{\text{hist}}_{ik} = \left( \bar{q}_{\text{tod}}^{i,k}, \bar{q}_{\text{wday}}^{i,k}, \bar{q}_{\text{mnth}}^{i,k}, \bar{a}_{\text{tod}}^{i,k}, \bar{a}_{\text{wday}}^{i,k}, \bar{a}_{\text{mnth}}^{i,k}, \bar{q}_{\text{tod}}^{i,k} \cdot \bar{q}_{\text{wday}}^{i,k} \cdot \bar{q}_{\text{mnth}}^{i,k}, \bar{a}_{\text{tod}}^{i,k} \cdot \bar{a}_{\text{wday}}^{i,k} \cdot \bar{a}_{\text{mnth}}^{i,k} \right)
\end{align*}
\]  

\((19)\)

**Real-time AVL data**

Vehicle location data are currently more commonly available in real time than passenger counts. Given that the target bus has departed from the terminus, AVL data up to the source stop \(s\) can be utilized. Although AVL data contain no load data in themselves, they may contain useful information for load prediction. We consider two types of variables.

First, the more passengers boarding a bus, the longer the dwell time at the stop [8]. Thus, the total run time \(T_{i,1,s}\) from the start of the line to the source stop is an indicator of the passenger load on the bus.

Second, for high-frequency bus services, passenger arrival times are known to be well approximated by a Poisson process [5]. Under such conditions, the expected number of passengers boarding a bus at a given stop is proportional to the headway to the preceding bus. Headway is therefore an indicator of passenger load. Here, we include the headway \(h_{ik}\) at the \(K_s\) most recently visited stops upstream of and including source stop \(s\). To allow for nonlinear effects we also include the squared version of all variables, which gives a \(1 \times (2K_s + 2)\) vector of predictor variables,

\[
\begin{align*}
\mathbf{x}^{\text{avl}}_{is} = (T_{i,1,s}, h_{i,s-K_s+1}, \ldots, h_{is}, h_{i,s-K_s+1}^2, \ldots, h_{is}^2)
\end{align*}
\]  

\((20)\)

**Real-time APC data**

With real-time APC data, the bus load is predicted based on the load \(q_{is}\), boarding count \(b_{is}\) and alighting count \(a_{is}\) at the source stop. Further, to utilize the fact that passengers’ destinations may vary systematically with their origins, we include the boarding counts \(b_{ik}\) and alighting counts \(a_{ik}\) at
the $K_s - 1$ most recently visited stops upstream of $s$. To allow for nonlinear effects we also include the squared version of all variables. In total, we consider the $1 \times (4K_s + 2)$ vector of real-time passenger count predictors,

$$x_{isk}^{\text{apc}} = (q_{is}, a_{i,s-K_s+1}, \ldots, a_{is}, b_{i,s-K_s+1}, \ldots, b_{is}, q_{is}^2, a_{i,s-K_s+1}^2, \ldots, a_{is}^2, b_{i,s-K_s+1}^2, \ldots, b_{is}^2)$$

(21)

3 Case study

This section describes a real-world application of the personalized crowding prediction methodology, including the considered scenarios of data availability, a description of the bus line, an account of the data sources and the model fitting procedure.

3.1 Levels of data availability

Three scenarios of data availability are considered. The most basic scenario is where only historical load data are available. Thus, predictions must be based fully on historical temporal patterns. In this case, the potential predictors $x_{isk}$ are the $1 \times 9$ vector

$$x_{isk} = \left(1, x_{isk}^{\text{hist}}\right).$$

(22)

In the second scenario, historical load data are complemented with real-time AVL data. In this case, $x_{isk}$ is the $1 \times (2K_s + 11)$ vector

$$x_{isk} = \left(1, x_{isk}^{\text{hist}}, x_{isk}^{\text{avl}}\right).$$

(23)

In the third scenario, representing the highest level of data availability, historical load data and real-time AVL data are complemented with real-time APC data. In this case, $x_{isk}$ is the $1 \times (6K_s + 13)$ vector

$$x_{isk} = \left(1, x_{isk}^{\text{hist}}, x_{isk}^{\text{avl}}, x_{isk}^{\text{apc}}\right).$$

(24)

3.2 Bus line characteristics

The crowding prediction is applied to the north-to-south direction of high-frequency bus line 4 in Stockholm, Sweden, shown in Figure 4. Line 4 is ca. 12.4 km long and has 31 stops in the north-to-south direction. The travel time from start to end is typically around 60 minutes. Parts of the route
are equipped with dedicated bus lanes and/or transit signal priority. To a large extent, however, the buses run in dense mixed traffic. The line is serviced with 4–6 minutes planned headway during work hours. The operations during this time period are regularity-based as opposed to following a fixed schedule, i.e., each bus driver seeks to maintain equal headways to the preceding and following buses on the same line. According to the public transport planning guidelines for Stockholm [29], the seated capacity of the type of articulated buses used on line 4 is $c_{\text{se}} = 45$ passengers.

The line has around 30,000 boarding passengers per day and per direction, and it is common particularly during peak hours to not get a seat. Figure 5, top left, shows the distribution of passenger loads and alighting counts along the line. There are substantial variations across stops and three peaks dispersed along the line. The other three diagrams illustrate the distributions of the three crowding metrics considered in this paper. The distributions differ substantially as the probability to get a seat on boarding (top right) reflects local crowding conditions at each stop, the travel time standing (bottom left) is affected by crowding conditions in a neighborhood of downstream segments, while the excess perceived travel time (bottom right) reflects cumulative crowding conditions over the entire trip.
Figure 5: Passenger-oriented crowding metrics, percentiles across all trips in the data set. Top left: Passenger load and alighting count across stops. Top right: Probability of getting a seat on boarding across origin stops. Bottom left: Expected travel time standing across origin stops on trip to end stop Gullmarsplan. Bottom right: Perceived excess travel time across origin stops on trip to end stop Gullmarsplan.

3.3 Data and model fitting

The study considers the afternoon peak from 15:00 to 17:30, Mondays through Fridays during 2016. Periods with lower service frequency, including summer, holidays and weekends, are excluded. AVL data are available for all buses and all stops on line 4 during this period. APC data are available only for about 20% of all bus trips. The sensors counting boarding and alighting passengers are installed on a random sample of all buses in Stockholm, and are moved between vehicles at regular intervals. Since the vehicles equipped with APC sensors are selected at random, there is no systematic bias in the selection of days for the evaluation of prediction performance. The data is
currently collected in batches, but is used here to evaluate the potential of having AVL and APC data available in real-time.

There are in total 5861 unique bus trips across 191 days in the data set. Of these, 1200 trips (20.5%) have recorded APC data. The other runs are excluded from the analysis; however, headways are computed based on the complete set of trips.

Out of the 190 days, 95 days (471 bus runs with recorded load data) are randomly selected as the test set while the remaining 95 days (475 bus runs with load data) are used as training set. Historical mean loads and alighting counts are calculated across all 190 days. We use 10-fold cross-validation based on the mean squared error to select the regularization coefficient $\lambda$ for each model. The number of stops upstream of and including the source stop $s$ to include is set to $K_s = \min\{s, 6\}$.

4 Results

4.1 Predictor selection analysis

Table 2 shows the relative frequency with which each predictor is included (has a non-negative coefficient) in the estimated regression models. The frequencies are relative to the maximum possible number of occurrences considering that predictors with longer lags can be included in fewer models. The bottom row of the table shows the average number of predictors included in each model relative to the maximum possible number.

Among the historical predictors, the interaction variables are by far the most frequently included. Load-related predictors are sometimes included in models targeting alighting counts and vice versa. It is notable that historical predictors are frequently included even when real-time AVL and APC predictors are available. This suggests that historical temporal patterns provide an important baseline for predictions.

Among the AVL predictors the source stop headway is the most frequently included predictor for both loads and alighting counts. Headways with longer lags are often included but the frequency tends to decline with the lag. Total run time is included more often than not when real-time APC predictors are not available, but appears to be less important in general than the headway. Squared predictors are less commonly included except, interestingly, for the longest lag headway, which may serve as a collector of circumstances further upstream. Not surprisingly, AVL predictors are less frequently included when APC predictors are available. Thus, AVL serves partially as a proxy for APC.
Table 2: Relative selection frequency of model predictors. Values above 0.5 are indicated in bold.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Hist</th>
<th>Hist+AVL</th>
<th>Hist+AVL+APC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$q_{ik}$</td>
<td>$\hat{q}_{ik}$</td>
<td>$\hat{q}_{ik}$</td>
</tr>
<tr>
<td></td>
<td>$\bar{q}_{tod}$</td>
<td>0.133</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>$\bar{q}_{wday}$</td>
<td>0.333</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>$\bar{q}_{mnth}$</td>
<td>0.267</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td>$\bar{q}<em>{tod} \cdot \bar{q}</em>{wday} \cdot \bar{q}_{mnth}$</td>
<td>1.00</td>
<td>0.200</td>
</tr>
<tr>
<td></td>
<td>$\bar{a}_{tod}$</td>
<td>0.233</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td>$\bar{a}_{wday}$</td>
<td>0.367</td>
<td>0.333</td>
</tr>
<tr>
<td></td>
<td>$\bar{a}_{mnth}$</td>
<td>0.267</td>
<td>0.233</td>
</tr>
</tbody>
</table>

| AVL       | $T_{1,s}$ | 0.589 | 0.568 | 0.283 | 0.202 |
|           | $T_{1,s}^2$ | 0.391 | 0.347 | 0.149 | 0.193 |
|           | $h_{1,s}$ | 0.975 | 0.878 | 0.917 | 0.676 |
|           | $h_{1,s}^2$ | 0.421 | 0.267 | 0.140 | 0.106 |
|           | $h_{1,s-1}$ | 0.707 | 0.603 | 0.419 | 0.345 |
|           | $h_{1,s-1}^2$ | 0.335 | 0.249 | 0.148 | 0.118 |
|           | $h_{1,s-2}$ | 0.545 | 0.447 | 0.233 | 0.225 |
|           | $h_{1,s-2}^2$ | 0.400 | 0.320 | 0.246 | 0.143 |
|           | $h_{1,s-3}$ | 0.496 | 0.453 | 0.180 | 0.188 |
|           | $h_{1,s-3}^2$ | 0.319 | 0.282 | 0.242 | 0.160 |
|           | $h_{1,s-4}$ | 0.443 | 0.431 | 0.206 | 0.197 |
|           | $h_{1,s-4}^2$ | 0.440 | 0.348 | 0.345 | 0.212 |
|           | $h_{1,s-5}$ | 0.430 | 0.410 | 0.363 | 0.293 |
|           | $h_{1,s-5}^2$ | 0.533 | 0.480 | 0.450 | 0.280 |

| APC       | $q_{i,s}$ | 0.752 | 0.632 | 0.283 | 0.202 |
|           | $q_{i,s}^2$ | 0.235 | 0.269 | 0.455 | 0.453 |
|           | $b_{i,s}$ | 0.202 | 0.212 | 0.258 | 0.212 |
|           | $a_{i,s}$ | 0.260 | 0.278 | 0.318 | 0.350 |
|           | $a_{i,s}^2$ | 0.209 | 0.185 | 0.278 | 0.278 |
|           | $a_{i,s-1}$ | 0.247 | 0.224 | 0.219 | 0.224 |
|           | $a_{i,s-1}^2$ | 0.288 | 0.318 | 0.288 | 0.318 |
|           | $a_{i,s-2}$ | 0.228 | 0.220 | 0.206 | 0.206 |
|           | $a_{i,s-2}^2$ | 0.270 | 0.236 | 0.279 | 0.228 |
|           | $a_{i,s-3}$ | 0.328 | 0.302 | 0.328 | 0.302 |
|           | $a_{i,s-3}^2$ | 0.234 | 0.188 | 0.234 | 0.188 |
|           | $a_{i,s-4}$ | 0.231 | 0.259 | 0.231 | 0.259 |
|           | $a_{i,s-4}^2$ | 0.259 | 0.215 | 0.259 | 0.215 |
|           | $a_{i,s-5}$ | 0.308 | 0.314 | 0.262 | 0.231 |
|           | $a_{i,s-5}^2$ | 0.237 | 0.293 | 0.237 | 0.293 |
|           | $a_{i,s}^2$ | 0.273 | 0.213 | 0.380 | 0.317 |
|           | $a_{i,s}^2$ | 0.260 | 0.220 | 0.260 | 0.220 |

Total: | 0.438 | 0.345 | 0.492 | 0.433 | 0.291 | 0.268
As expected, the most frequently included APC predictor is the source stop passenger load. Boarding predictors are included somewhat more frequently than alighting predictors. This is intuitive since passengers boarding have a more direct effect on downstream loads and alighting counts. Selection frequencies are quite stable across lags.

4.2 Passenger load and alighting

Figure 6 shows predicted against observed loads for the test data set and a specific target station (Västerbroplan). Loads are predicted for all three levels of data availability considered in Section 3.1. The dashed horizontal and vertical lines indicate the seated capacity. The top diagram shows the results using only historical predictors, the middle diagrams add AVL predictors and the bottom diagrams further add APC predictors. In the middle and bottom rows, the left side shows results with 10-min prediction horizon, i.e., predictions from the most recent stop 10 min before departing from the target stop. Due to travel time variations, the source stop may vary between bus trips. The right side shows results with 1-min prediction horizon.

Prediction performance increases with real-time bus data. Historical data tend to underestimate the load on crowded runs, but overestimate the load on the least crowded runs. Thus, historical patterns alone cannot explain well why some buses are highly crowded while others are almost empty. Real-time AVL data improve the bias but show high variance, and performance does not improve much from 10-min to 1-min horizon. Real-time APC data substantially improve the prediction performance, particularly with 1-min horizon.

Figure 7 shows the mean absolute errors (MAE) and mean errors (ME) of predicted loads across all stops with 10-min and 1-min prediction horizon. Compared with Fig. 5, top left, MAE follow the passenger load magnitude across stops. ME values fluctuate around zero and show no apparent signs of bias. With 10-min horizon, real-time AVL and APC data do not bring benefits before stop 7, since the bus has not yet departed from the first stop. For subsequent stops, real-time AVL predictors bring a substantial reduction in MAE, and real-time APC predictors improve performance somewhat further. With 1-min horizon, real-time APC predictors drastically increase prediction performance compared to the other levels of data availability.

Figure 8 shows predicted against observed alighting counts according to the same principles as in Fig. 6. Due to smaller magnitudes the discrete nature of the numbers is more apparent than for the loads. The contour plots indicate the frequency with which different values occur. The impacts
Figure 6: Predicted vs. observed load for test data set, target stop 18 (Västerbroplan). Top: Historical data. Middle: Historical data and real-time AVL data, 10 min. (left) and 1 min. (right) prediction horizon. Bottom: Historical data, real-time AVL and APC data, 10 min. (left) and 1 min. (right) prediction horizon. Horizontal and vertical dashed lines indicate bus seated capacity.
of data availability and prediction horizon follow similar trends as for the loads. Overall, however, prediction performance is lower for alighting counts than for loads.

Figure 9 shows the same metrics as Fig. 7 for predicted alighting counts. Compared to Fig. 7, real-time AVL and APC data are less effective in increasing prediction performance.

4.3 Probability of seat on boarding

Figure 10 shows the accuracy and ME of the predicted probability to get a seat on boarding across all origin stops. Accuracy is evaluated as the share of observations for which the prediction correctly classifies the crowding with respect to three categories: no available seat certain ($\pi_{o}^{brd} = 0$), seat availability uncertain ($\pi_{o}^{brd} \in (0, 1)$), and available seat certain ($\pi_{o}^{brd} = 1$).

Prediction accuracy is high for the first few stops even with only historical data, since the load rarely reaches the seated capacity. Compared with Fig. 5, top right, accuracy tends to drop at stations where the seat probability is lower. Real-time AVL and APC data increase accuracy, in particular with 1-min prediction horizon where accuracy is around 90% for most stations. With 10-min horizon the ME tends to be positive which suggests that the predictions are biased towards being too optimistic. This is an effect of the non-linear relation with loads and alighting counts. With 1-min horizon there is no evidence of bias.
Figure 8: Predicted vs. observed alighting counts for test data set, target stop 18 (Västerbroplan). Top: Historical data. Middle: Historical data and real-time AVL data, 10 min. (left) and 1 min. (right) prediction horizon. Bottom: Historical data, real-time AVL and APC data, 10 min. (left) and 1 min. (right) prediction horizon.

4.4 Travel time standing

Figure 11 shows the MAE and ME of the predicted travel time standing across all origin stations for a trip to the final stop (Gullmarsplan). The top
row shows the direct predictions according to Eq. (9) and the bottom row shows the predictions after bias correction according to Eq. (11).

Compared to Fig. 5, bottom left, prediction performance is lower for origin stops with higher chances of long travel times standing. The reduction in MAE with real-time AVL and APC data is moderate for 10-min horizon, but substantial for 1-min horizon. Without bias correction, predictions tend to be overly optimistic particularly with 10-min horizon. The bias correction manages to reduce ME to practically zero, at the cost of a slight increase in MAE.
4.5 Excess perceived travel time

Figure 12 shows the MAE and ME of the predicted excess perceived travel time across all origin stations for a trip to the final stop. MAE values are highest for early origin stops due to larger possible variability in trip travel times, and gradually decrease for shorter trips. Compared to the travel time standing, real-time AVL and APC data provide fairly moderate improvements in prediction performance. This is due to the fact that prediction errors grow with the distance between the source and the target station, and the excess perceived travel time is sensitive to crowding conditions during the entire trip.

Without bias correction, predictions tend to be overly optimistic for both prediction horizons. The bias is fairly proportional to the length of the trip in terms of the number of stops. The bias correction reduces ME to practically zero with virtually no effect on the MAE.
Figure 12: MAE and ME of predicted excess perceived travel time across origin stops for trip to end stop Gullmarsplan. Prediction horizon 10 min (left) and 1 min (right). With (top) and without (bottom) bias correction.

5 Conclusions

The paper has proposed a methodology for providing personalized, predictive in-vehicle crowding information to bus travellers. Three crowding metrics, (1) the probability of getting a seat on boarding, (2) the expected travel time standing, and (3) the excess perceived travel time, are examined. The traveller can request the information for a specific trip, customized to the traveller’s crowding preferences, and disseminated back via mobile applications or bus stop displays. The information can assist the traveller in deciding whether to board a vehicle or choose another travel option.

The methodology combines a bottom-up framework for prediction of passenger loads and alighting counts based on lasso regression with a probabilistic model of in-vehicle travel conditions in order to predict the crowding metrics for the traveller’s requested journey. Depending on data availabil-
ity, the prediction method can use a combination of historical passenger counts, real-time vehicle locations and real-time passenger counts to predict passenger counts.

The methodology is applied to a case study for a bus line in Stockholm, where all vehicles transmit AVL data but only ca. 20% are equipped with passenger counters. The results show that systematic temporal variations captured in historical load data are useful to provide baseline predictions. Real-time AVL and APC predictors provide substantial improvements in prediction performance, in particular for crowding metrics 1 and 2 and for short prediction horizons. AVL predictors serve partially as complements and partially as proxies for APC predictors. The bottom-up prediction approach of aggregated crowding metrics creates some optimistic bias, which can be removed with a simple correction term at little or no cost to mean absolute errors.

The results indicate that personalized, predictive crowding information that is robust to varying data availability can be provided sufficiently early to be useful to travellers. The results are of value for agencies and operators in order to increase the attractiveness and capacity utilization of public transport. Experience with RTCI in bus systems is still limited, but pilot studies [36] and simulation experiments [23, 7] in metro systems suggest that RTCI can be an effective means to reduce in-vehicle crowding and passengers discomfort, increase service performance and reduce operation costs.

The presented research can be extended in several directions. An area of further work is to evaluate whether more elaborated prediction methods can further increase performance. Prediction accuracy for the probability of getting a seat may be increased by formulating it as a classification problem, although this would double the predictions needed if the other crowding metrics are also to be computed. Another interesting topic for investigation is how to design interfaces for RTCI to disseminate the information to travellers in an understandable and useful way.

Acknowledgements

The author would like to thank the Region Stockholm Transport Administration for kindly providing the data used in the case study. The work was financially supported in part by Trenop Strategic Research Area.
References


[22] Ehsan Mazloumi, Geoff Rose, Graham Currie, and Majid Sarvi. An integrated framework to predict bus travel time and its variability us-


