Data-Driven Bus Crowding Prediction Based on Real-Time Passenger Counts and Vehicle Locations

Erik Jenelius
Division of Transport Planning
KTH Royal Institute of Technology
Stockholm, Sweden
jenelius@kth.se
January 5, 2019

***Abstract*—The paper addresses the bus crowding prediction problem based on real-time vehicle location and passenger count data and evaluates the performance of a data-driven lasso regression prediction method. The problem is studied for a high-frequency bus line in Stockholm, Sweden. Prediction accuracy is evaluated with respect to absolute passenger loads and predefined discrete crowding levels. When available, predictions with real-time vehicle location and, in particular, passenger count data significantly outperform predictions based only on historical data, with accuracy improvements varying in magnitude across target stations and prediction horizons.**

***Index Terms*—public transport, bus, crowding, prediction, AVL data, APC data**

I. Introduction

Public transport is a vital part of the urban transport system in order to mitigate environmental problems such as pollution, noise and climate change. However, as urban populations grow, many public transport systems are experiencing increasing congestion and crowding. In-vehicle crowding has negative effects on traveler satisfaction and wellbeing that may inhibit the transition from private to public transport [1]–[3].

For bus services, in particular those running at high frequency, there is also a negative feedback loop between crowding and arrival time irregularity, which leads to longer passenger waiting times. Irregular arrival times typically lead to uneven passenger loads across vehicles, which in turn increases the irregularity of arrival times and the frequency of bus bunching [4]–[6].

In recent years, many cities have started to provide real-time bus arrival time information at stops or in mobile applications. Such information typically relies on automatic vehicle location (AVL) data collected either at fixed locations (e.g., through sensors at every stop) or continuously (e.g., through GPS devices installed in the vehicles). Several methods for short-term arrival time prediction based on various assumptions about data availability have been proposed [7]–[9]. Real-time arrival time information systems have been shown in several studies to have positive effects on perceived waiting times, safety and security, impacts of service disruptions, and general satisfaction [10]–[13].

The research is funded in part by TRENoP Strategic Research Area.

Systems providing real-time crowding information (RTCI) about arriving buses require online automatic passenger counting (APC) and are less common in practice. RTCI provision allows travellers to make better informed decisions about whether to board a bus or not based on their preferences for crowding, waiting time, total travel time, etc. RTCI may thus directly increase passenger satisfaction as well as indirectly increase satisfaction and service quality by reducing the negative crowding externalities, which would be attractive for public transport authorities and operators [14].

Providing timely crowding information generally requires that passenger loads be predicted several stops ahead from the current bus locations. In the context of metro trains, accurate real-time crowding information based on passenger loads in individual metro train cars can be provided sufficiently early for travellers to consider in their route, train and car choices [15]. In simulation settings, both simple prediction schemes based on the crowding of the one or two most recent vehicle runs [16] and more complex schemes involving running the simulation model forward to a fixed point solution [17] have been proposed and evaluated. Both studies demonstrate that predictive RTCI may equalize crowding among metro runs and reduce passengers' experienced travel time. As far as we are aware, however, no studies have tackled the bus crowding prediction problem in an empirical context.

The aim of this paper is to formulate the bus crowding prediction problem based on real-time vehicle location and passenger count data and evaluate the performance of a data-driven lasso regression prediction method . The focus is on RTCI for provision to waiting, arriving or transferring passengers. The problem is studied for the busiest high-frequency bus line in Stockholm, Sweden. Prediction accuracy is evaluated with respect to absolute passenger loads as well as predefined discrete crowding levels.

II. Methodology

This section introduces the bus crowding prediction problem and proposes a data-driven solution method. Consider the task of predicting the passenger load q_{ijk} of vehicle run i on day j at departure from stop k of a specific bus line. In general, the load is predicted based on a combination of real-time

TABLE I
NOTATION

i	Daily vehicle run index, $i \in I$
j	Day index, $j \in J$
k	Stop index, $k \in K$
t_{ijk}	Departure time
h_{ijk}	Headway to preceding vehicle
b_{ijk}	Number of boarding passengers
a_{ijk}	Number of alighting passengers
q_{ijk}	Bus load at departure
$\bar{q}^{\text{tod}}_{ik}$	Time-of-day-specific historical mean load
$\bar{q}^{\text{wday}}_{jk}$	Weekday-specific historical mean load
$\bar{q}^{\text{mnth}}_{jk}$	Month-specific historical mean load
$\hat{q}_{ijk}$	Predicted bus load
c^{sit}	Vehicle seat capacity
r_{ijk}	1 if $q_{ijk} > c^{\text{sit}}$, 0 otherwise.

and historical data. The notation used throughout the paper is shown in Table I.

A. Crowding Predictors

1) Historical APC Data: The prediction can utilize recurring patterns in passenger loads at the target stop 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 for vehicle runs departing during the same time-of-day interval, weekday and month as the target run, collected in the 1×3 vector of historical predictors

$$\mathbf{x}^{\text{hist}}_{ijk} = \left(\bar{q}^{\text{tod}}_{ik}, \bar{q}^{\text{wday}}_{jk}, \bar{q}^{\text{mnth}}_{jk}\right) \tag{1}$$

2) Real-Time AVL Data: At the present, AVL data are more commonly available in real time than APC data. Given that the target bus j has departed from the terminus, AVL data up to the most recent stop k^{s} from which vehicle i has departed on current day j at the time for the prediction can be utilized. Although AVL data contain no load data in themselves, they may contain useful information for load prediction. For high-frequency bus services, passenger arrival times are known to be well approximated by a Poisson process [18]. 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 may therefore be used as a proxy for passenger load. Here, we include the headway h_{ijk} and the squared headway h^2_{ijk} of the target vehicle at the K most recent stops up to and including source stop k^{s},

$$\begin{aligned}\mathbf{x}^{\text{avl}}_{ijk^{\text{s}}} = (&h_{ij,k^{\text{s}}-K+1}, \ldots, h_{ijk^{\text{s}}},\\ &h^2_{ij,k^{\text{s}}-K+1}, \ldots, h^2_{ijk^{\text{s}}})\end{aligned} \tag{2}$$

3) Real-Time APC Data: With real-time APC data, the bus load is predicted based on load measurements q_{ijk} from K stops up to and including k^{s}. To allow possible effects of seat availability on passengers' boarding choices, we also include seat availability indicators $r_{ijk} = 1$ if $q_{ijk} \leq c^{\text{sit}}$, 0 otherwise. Further, to capture possible correlations between boardings and alightings at different stops, we include the number of boardings b_{ijk} and alightings a_{ijk} at the source stop. In total, we consider the $1 \times 2(K+1)$ vector of vehicle-specific current-day predictors,

$$\begin{aligned}\mathbf{x}^{\text{run}}_{ijk^{\text{s}}} = (&q_{ij,k^{\text{s}}-K+1}, \ldots, q_{ijk^{\text{s}}}, b_{ijk^{\text{s}}}, a_{ijk^{\text{s}}},\\ &r_{ij,k^{\text{s}}-K+1}, \ldots, r_{ijk^{\text{s}}})\end{aligned} \tag{3}$$

B. Levels of Data Availability

In the following we fix the target stop k and source stop k^{s} and omit the indices for simplicity of notation. All predictors for vehicle run i and day j are collected in the $1 \times p$ vector $\mathbf{x}_{ij}$. Three scenarios of data availability are considered, each described in turn below.

1) Historical APC Data Only: The most basic scenario is where only historical load data are available. Thus, predictions must be based fully on historical patterns related to the time of day, day of the week, and month of the year. In this case, the potential predictors $\mathbf{x}_{ij}$ are the 1×3 vector

$$\mathbf{x}_{ij} = \mathbf{x}^{\text{hist}}_{ij} \tag{4}$$

2) Historical APC Data and Real-Time AVL data: In the second scenario, historical load data are complemented with real-time AVL data. In this case, $\mathbf{x}_{ij}$ is the $1 \times (2K+3)$ vector

$$\mathbf{x}_{ij} = \left(\mathbf{x}^{\text{hist}}_{ij}, \mathbf{x}^{\text{avl}}_{ij}\right) \tag{5}$$

3) Real-Time and Historical APC data, Real-Time AVL data: 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, $\mathbf{x}_{ij}$ is the $1 \times (4K+5)$ vector

$$\mathbf{x}_{ij} = \left(\mathbf{x}^{\text{hist}}_{ij}, \mathbf{x}^{\text{avl}}_{ij}, \mathbf{x}^{\text{run}}_{ij}\right) \tag{6}$$

C. Prediction Method

This section presents the lasso regression method for the crowding prediction problem. The method have been successfully applied to metro car load prediction [15]. The prediction model is linear in coefficients. Estimation uses the lasso regularization [19], i.e., parameters are selected to minimize

$$\frac{1}{2}\sum_i \sum_j \left(q_{ij} - \beta_0 - \sum_{l=1}^{p} x_{ijl}\beta_l\right)^2 + \lambda \sum_{l=1}^{p} |\beta_l| \tag{7}$$

where λ is a regularization coefficient that penalizes large parameter values. Larger λ enforce sparser solutions, i.e., more parameters equal to zero. The λ value is calibrated to minimize the cross-validation mean squared error. Given a new vector of predictors $\mathbf{x}^*$, the passenger load is predicted as

$$\hat{q} = \beta_0 + \sum_{l=1}^{p} x^*_l \beta_l \tag{8}$$

III. CASE STUDY

A. Bus Line Characteristics

The case study applies the real-time crowding prediction methodology to the southbound direction of high-frequency bus line 4 in Stockholm, Sweden, shown in Fig. 1. Line 4 is ca.

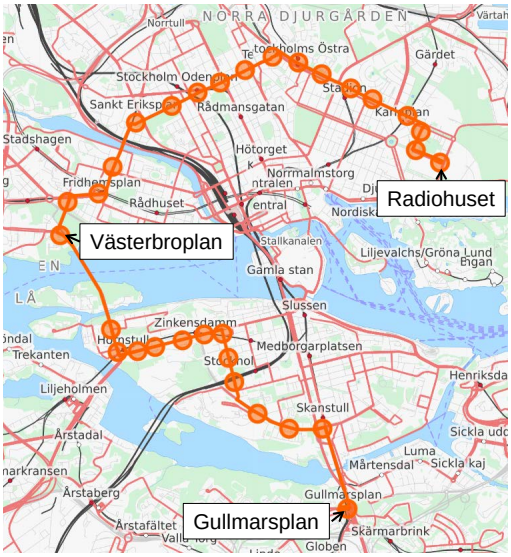

Fig. 1. Route of bus line 4. Studied direction is from Radiohuset to Gullmarsplan. Map source: OpenStreetMap.

12.4 km long, and consists of 31 stops in one direction (north-south) and 30 stops in the other direction (south-north). The typical run time from start to end in one direction is around 60 minutes. Parts of the route are equipped with dedicated bus lanes and/or transit signal priority. Between 7:05 and 18:56 the line is serviced with 4-6 minutes planned headway. The operations during this time period are regularity-based, i.e., each bus seeks to maintain equal headways to the previous and subsequent vehicles on the same line as opposed to following a fixed schedule. Line 4 is the busiest bus line in Stockholm with around 60,000 boarding passengers per day, and in-vehicle crowding is common.

The seated and standing capacities of the buses are taken from the public transport planning guidelines for Stockholm [20]. According to these, the seated and total capacity of the articulated buses used on line 4 is $c^{\text{sit}} = 45$ and $c^{\text{tot}} = 120$ passengers, respectively.

B. Vehicle Location and Passenger Count Data

The study considers the afternoon peak from 15:30 to 18:30, 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 for the same time period and all stops, but only for about 20% of all buses. The sensors detecting 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 runs across 191 days in the data set. Of these, 1200 runs (20.5%) have recorded APC data. The other runs are excluded from the analysis. 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 model specification.

For dissemination purposes crowding can be represented as a few discrete levels rather than absolute numbers. We define three quality of service levels of in-vehicle crowding; low crowding (0–45 passengers) corresponds to available seats for all passengers, medium crowding (46-75 passengers means that up to 40% of all passengers must stand, while high crowding (76-120 passengers) implies that between 40% and 62.5% of all passengers must stand.

Fig. 2 shows how passenger loads vary along four dimensions: bus stops, departure times from the terminus (in 10-minute intervals), weekdays, and months. The figure shows the median as well as the 10th and 90th percentile loads. As can be seen, there is significant variability between different bus runs, partially explained by systematic spatial and temporal patterns.

C. Model Fitting

Out of the 191 days, 96 days (612 bus runs with recorded load data) are randomly selected as the test set while the remaining 95 days (588 bus runs with load data) are used as training set. Historical mean loads are calculated across all 191 days. 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. For each model the number of stops upstream of and including the source stop k^{s} to include is set to $K = \min\{k^{\text{s}}, 4\}$.

IV. Results

A. Prediction Accuracy for Specific Target Stop

Fig. 3 shows predicted loads for specific target stop Västerbroplan plotted against measured loads for the test data set. Loads are predicted for all three levels of data availability considered in Section II-B from source stop Fleminggatan. 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 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. Real-time AVL data reduces the variance of the prediction but does not improve the bias. Real-time APC data, meanwhile, significantly improves the prediction accuracy across low, medium and high crowding levels.

Fig. 4 shows the prediction accuracy at target stop Västerbroplan as a function of source stop. The interstop spacing indicates the average travel time between consecutive stops, which gives an indication of the prediction accuracy as a function of time until departure from the target stop. Real-time data achieve a reduction of prediction errors compared to historical data. The benefit of real-time APC data in addition

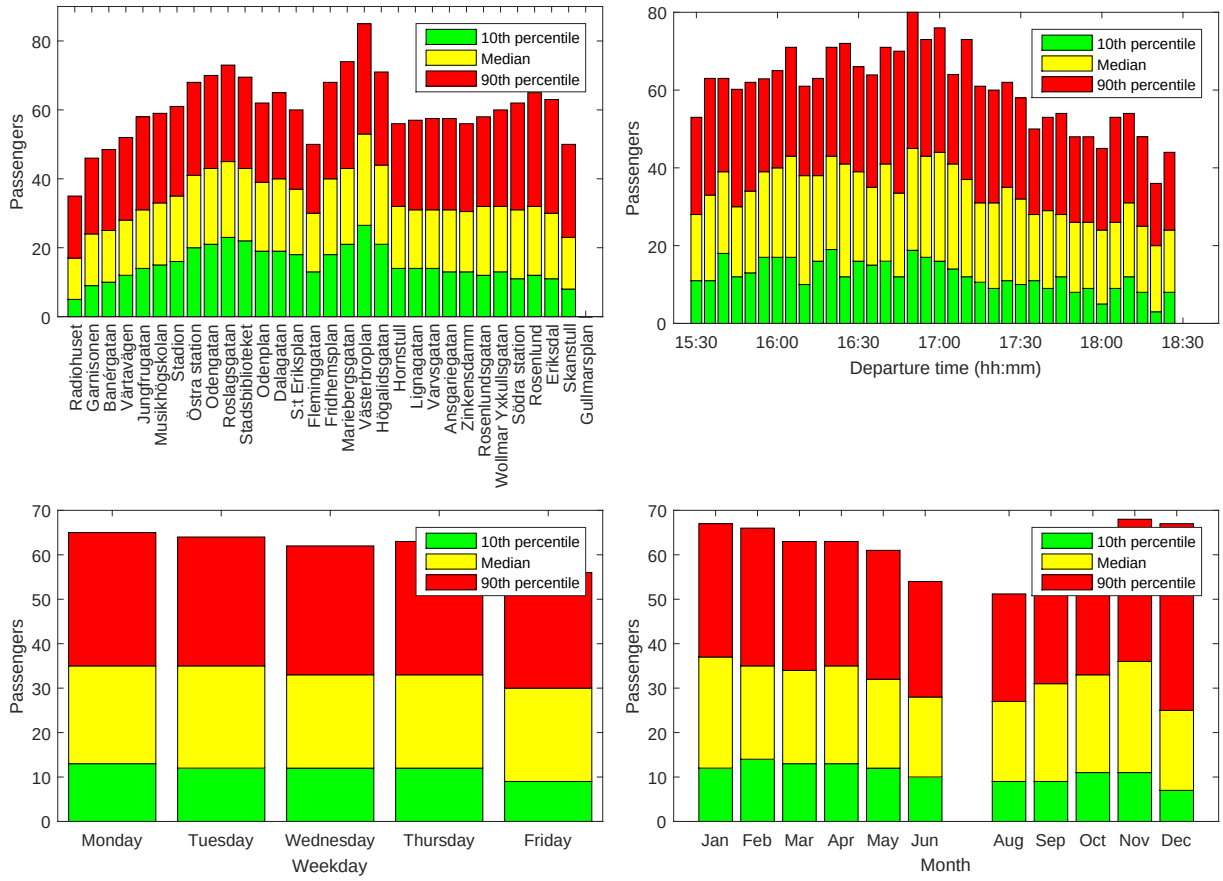

Fig. 2. Passenger load variations for bus line 4, 2016. Top left: stops. Top right: time of day. Bottom left: weekday. Bottom right: month.

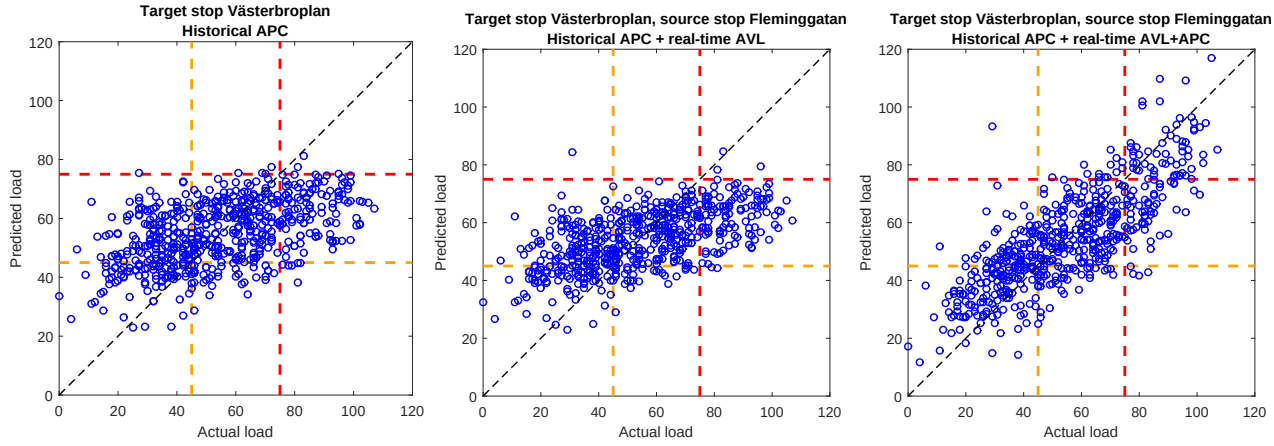

Fig. 3. Predicted vs. actual loads for test data set, target stop Västerbroplan. Left: Prediction based on historical data. Middle, right: Prediction with real-time AVL and APC data, respectively, from source stop Fleminggatan. Orange and red dashed lines indicate lower limits of medium and high crowding levels.

to AVL data becomes increasingly signficant as the bus gets closer to the target stop.

Dashed lines indicate departure times of preceding buses from the target stop given the average headway (ca. 5 minutes). It can be seen that when the current bus departs from Västerbroplan, the load of the next bus at departure from this stop can typically be predicted based on data from source stop Fleminggatan, while the load of the second next bus may be predicted based on data from source stop S:t Eriksplan.

B. Prediction Accuracy Across all Stops

Fig. 5 shows the prediction accuracy as a function of time to departure, averaged over all target stops from Garnisonen to Skanstull. Note that for a given prediction horizon, predictions are in general based on real-time data for some stops towards

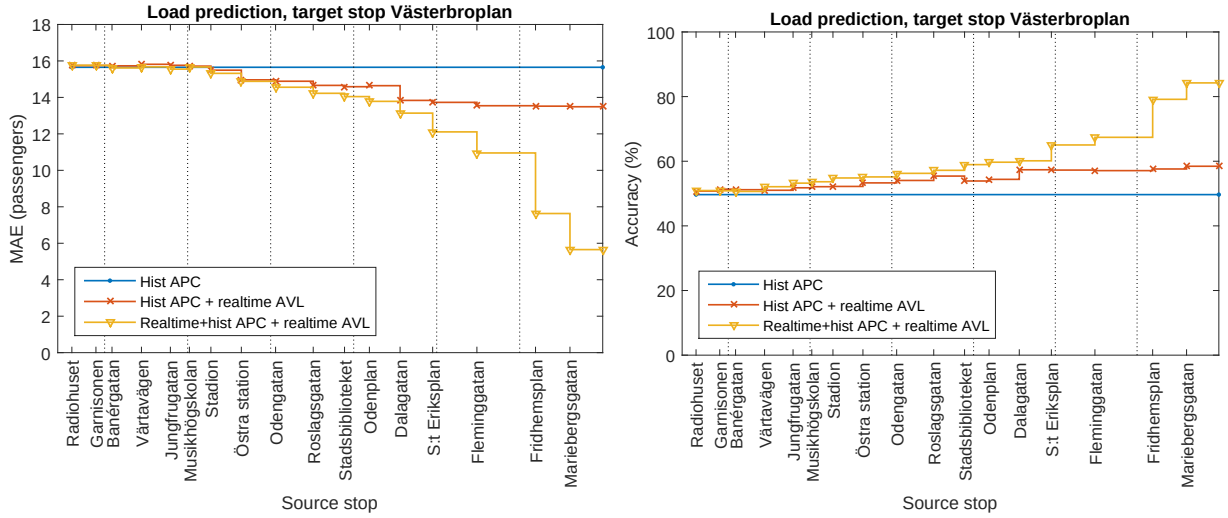

Fig. 4. Crowding prediction performance as function of source stop, target stop Västerbroplan. Left: Passenger load MAE. Right: Crowding level (low, medium, high) accuracy (%). Dashed lines indicate cumulative average headways between bus runs.

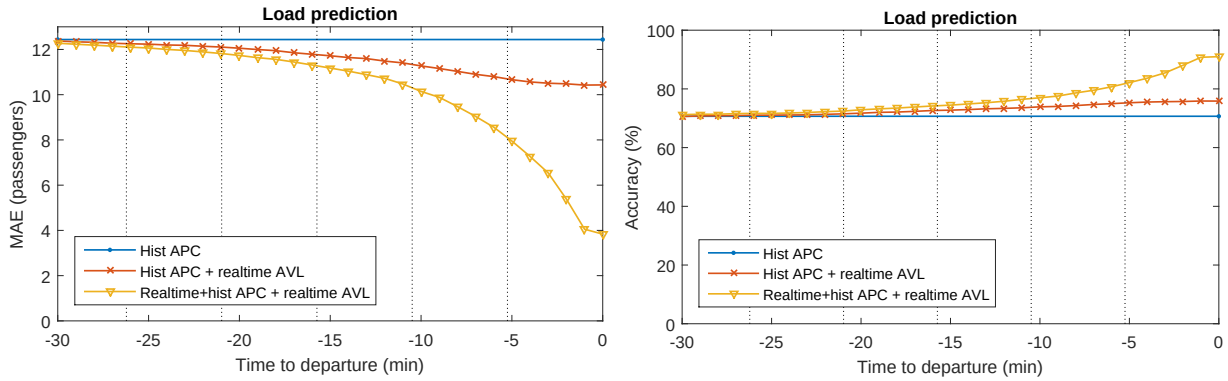

Fig. 5. Crowding prediction performance as function of time to departure, average across all target stops. Left: Passenger load MAE. Right: Crowding level (low, medium, high) accuracy (%). Dashed vertical lines indicate cumulative headways between bus runs.

the end of the line and historical data for some stops towards the beginning of the line.

The results show that the observations for particular target stop Västerbroplan hold more generally. Real-time AVL data and, particularly, APC data significantly improve prediction accuracy. Prediction errors steadily decrease as the time to departure decreases. 15 minutes before departure the average mean absolute error (MAE) is around 11 passengers with real-time APC data; 5 minutes before departure the average MAE has dropped to ca 8 passengers. Meanwhile, crowding level accuracy increases from ca. 75% to more than 80%.

V. Conclusion

The paper considers the bus crowding prediction problem based on real-time vehicle location and passenger count data and evaluates the performance of a data-driven lasso regression prediction method. The application to a high-frequency bus line in Stockholm shows that predictions with real-time data, particularly APC data, significantly outperform historical estimates, with accuracy improvements varying in magnitude across target stops and prediction horizons. The results show that passenger loads can be accurately predicted based on real-time data several stops upstream the target stops. This implies that travellers as well as operators have sufficient time to take actions in response to the forecasted crowding conditions.

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.

An interesting area of further work is to evaluate whether more sophisticated methods can further increase accuracy. For the current case study, however, experiments show that interaction and quadratic terms of the predictors do not significantly improve prediction accuracy, which suggests that the benefits of higher-order nonlinear models may be limited. Meanwhile, in bus networks with significant numbers of transfers between lines, an interesting research question is whether predictions

can be further improved by considering passenger loads on other lines.

ACKNOWLEDGMENT

The author would like to thank Stockholm City Council Transport Administration for kindly providing the AVL and APC data.

REFERENCES

[1] 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.

[2] M. Cantwell, B. Caulfield, and M. O'Mahony, "Examining the factors that impact public transport commuting satisfaction," *Journal of Public Transportation*, vol. 12, no. 2, pp. 1–21, 2009.

[3] 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.

[4] C. F. Daganzo, "A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons," *Transportation Research Part B*, vol. 43, pp. 913–921, 2009.

[5] J.-D. Schmöcker, W. Sun, A. Fonzone, and R. Liu, "Bus bunching along a corridor served by two lines," *Transportation Research Part B: Methodological*, vol. 93, pp. 300–317, 2016.

[6] M. Andres and R. Nair, "A predictive-control framework to address bus bunching," *Transportation Research Part B*, vol. 104, pp. 123–148, 2017.

[7] 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.

[8] O. Cats and G. Loutos, "Evaluating the added-value of online bus arrival prediction schemes," *Transportation Research Part A*, vol. 86, pp. 35–55, 2016.

[9] 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.

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

[11] 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*, vol. 45, no. 8, pp. 839–848, 2011.

[12] 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.

[13] 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.

[14] 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.

[15] E. Jenelius, "Car-specific metro train crowding prediction based on real-time load data," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 78–83.

[16] 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.

[17] 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.

[18] A. Ceder, *Public Transit Planning and Operation: Theory, Modelling and Practice*. Butterworth-Heinemann Ltd, 2007.

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

[20] Trafikförvaltningen, *Riktlinjer Planering av kollektivtrafiken i Stockholms län*, Stockholms Läns Landsting, 2016, sL-S-419761, in Swedish.