The Day-Ahead Forecasting of the Passenger Occupancy in Public Transportation by Using Machine Learning

Atilla Altıntaş[1(✉)], Lars Davidson[1], Giannis Kostaras[2], and Maycel Isaac[2]

[1] Division of Fluid Dynamics, Department of Mechanics and Maritime Sciences, Chalmers University of Technology, SE-412 96, Gothenburg, Sweden
altintas@chalmers.se

[2] Synteda AB, Skånegatan 29, 412 52 Göteborg, Sweden

Abstract. Public transport is one of the main infrastructures of a sustainable city. For this reason, there are many studies on public transportation which mostly answer the question of "when my next bus will arrive?". However now when the public is under the restrictions of the Covid-19 pandemic and learning to live with new social rules such as "social distance" a new yet crucial question arise on public transportation: "how crowded my next bus will be?" To prevent the crowdedness in public transportation the traffic regulators need to forecast the number of passengers the day ahead. In this study, in cooperation with Synteda, we suggest a machine learning algorithm that forecasts the occupancy in a bus or tram the day ahead for each stop for a route. The input data is past passenger travel data provided by the Västtrafik AB which is the public transportation company in Gothenburg, Sweden. The hourly data for the precipitation and temperature also has been added to the forecasting method; the database of precipitation and temperature is obtained by the SMHI, Swedish Meteorological and Hydrological Institute.

Keywords: Artifical intelligence · SVR · Machine learning · Forecasting · Public transport

1 Introduction

The reliability of the public transport system, especially in terms of travel time and space availability, greatly affects the quality of life of travelers [12]. However many researchers have proposed to predict bus arrival times such as [8,18], only a few previous studies have focused on predicting space availability. References [10,13,17] tried to develop an effective solution to solve these type of difficulties with the introduction of AI and ML. A research team from the University of Pittsburgh has studied forecasting bus passenger capacity in the whole urban bus transit system by using a random forest machine learning algorithm and obtained good approximations [4].

A. L. Martins et al. (Eds.): INTSYS 2021, LNICST 426, pp. 3–12, 2022.
https://doi.org/10.1007/978-3-030-97603-3_1

The day-ahead forecasting of the occupancy of the public transport recently become an important area of study. The Covid-19 pandemic made this problem much more difficult to solve for the traffic regulators. Jenelius and Cobecauer [9] studied the Covid-19 related passenger occupancy differences in the three most populated regions of Sweden, namely, Stockholm, Västra Götaland and Skåne. The results show a large amount of decrease in 2020, which is changing in between 40% to 60% depending on the regions.

For forecasting purposes, a number of strategies have recently been built, which can be split into two categories: traditional mathematical statistics and machine learning methods. Regression analysis [2] and time series analysis [7] are employed. Lu et al. [11] applied a deep learning algorithm and they suggested that a sample selection algorithm might improve the prediction accuracy. Novikov et al. [14] selected a number of criteria that effect the number of passengers and suggested a multicriteria optimization problem to design a transport infrastructure. A review and comparison of the methods are given in Ref [16], where ten different state-of-the-art forecasting methods are applied to predict the traffic flow. They used two real-world datasets, first dataset describes the traffic flow in the city center of Lyon (France), while the second is from Marseille that describes the traffic in the city outskirts. As a result, they suggest that the Lasso and Support Vector Regressions (SVR) methods are superior to the other approaches that they have used.

In this study, we forecast the occupancy of the tram or bus route by using the SVR. The present method has also been used for forecasting accuracy of highway tollgates traveling time [3]. The results show that SVR can successfully forecast the day-ahead occupancy for a route for each stop for each trip.

The paper is organized as follows. First, the method is given followed by the application of the method to database. The results are summarized and addressed in the following section, and some concluding remarks are given in the final section.

2 SVR Method

The Support Vector Regression (SVR) is an algorithm for machine learning, which is a variant of Support Vector Machine (SVM) [6,15]. SVR has widely been applied to forecasting problems. Consider a time-series data,

$$D = (X_i, y_i), 1 \leq i \leq N,$$

where X_i represents the ith element and y_i corresponds the target output data. The SVR function, f, is a linear function which is issued to formulate the nonlinear relation between input and output data as: $f(X_i) = \omega^T \phi(X_i) + b$, where ω, b and $\phi(X_i)$ are the weight vector, bias and function that maps the input vector X into a higher dimensional feature space, respectively. ω and b are obtained by solving the optimization problem:

$$min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \tag{1}$$

subject to:

$$\begin{aligned} y_i - \omega^T(\psi(x)) - b &\leq \epsilon + \xi_i \\ \omega^T(\psi(x)) + b - y_i &\leq \epsilon + \xi_i \\ \xi_i, \xi_i^* &\geq 0. \end{aligned} \tag{2}$$

The first term of Eq. 1 measures the flatness of the function. The parameter C balances the trade-off between the complexity of the model and its generalization ability. The cost of error is measured by the variables, ξ_i and ξ_i^*.

The final SVR function is obtained as:

$$y_i = f(X_i) = \sum_{i=1}^{N}((\alpha_i - \alpha_i^*)K(X_i, X_j)) + b \tag{3}$$

where $K(X_i, X_j)$ is the Kernel function [15] and α_i and α_i^* are the Lagrange multipliers.

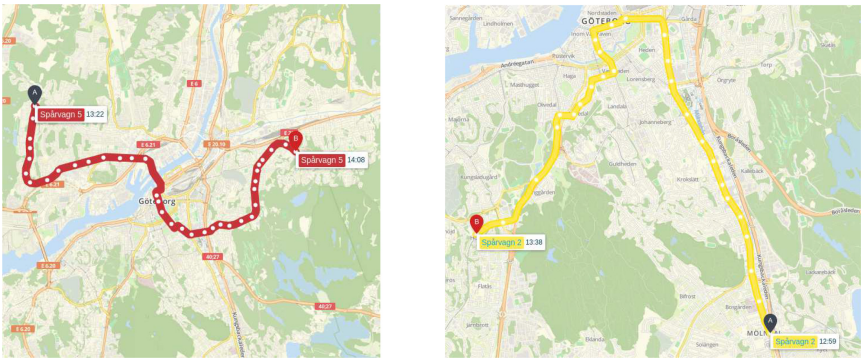
2.1 Application to Västtrafik database

The data are provided by Västtrafik AB [1] which is the company responsible for public transport services involving buses, ferries, trains, and the Gothenburg tram network in the Västra Götaland region, Sweden.

Table 1. Dictionary for the Vasttrafik data.

Operating Day Date	**Departure Date Actual**
- The start of an operating day is usually around 5 AM and can end as late as 3AM of the next day.	- Actual date change at midnight.
	Departure Time Actual
Stop Area Name	- Its the time in HHMM when the tram left the first stop.
- Name of the stop area.	**Stop Area Name**
Sequence Number	- Name of the stop area. - Identifies the route.
- The order in which the stops comes for the specific trip.	**Boarding**
Departure Load	- Number of travelers getting on.
- Number of travelers when departing from the stop.	**Alighting**
Arrival Load	- Number of travelers getting off.
- Number of travelers when arriving at the stop.	**Seating Capacity**
Stop Route Variant ID	- Number of seats.
- Route that the tram is taking.	**Total Capacity**
Comfort Capacity	- Maximum number of people allowed onboard.
- Estimation of the number of people that can comfortably get on the train, seated and standing.	

The data consist of a list of records of actual passenger counting data in Gothenburg area (Table 1). Number of trips measured varies by the route. It varies by 15%–100% depending on the line. The data include *OperatingDayDate*, which is the start of an operating day. *DepartureDateActual*, is the actual date change at midnight. *DepartureTimeActual*, is the time when the tram leaves the first stop. However, if the time starts with a zero it's not included. *StopAreaNumber*, is the identifier for a stop (i.e. station) name. *StopAreaName*, is the name of the stop. *TechnicalLineNumber*, identifies the route. In this database all trams in Gothenburg start with 50 followed by the actual number of the tram.



(a) Route map for tram number 5. (b) Route map for tram number 2.

Fig. 1. Maps for the routes that have been studied in this work. The figures are taken from the Västtrafik website.

Example: 5005 is tram number 5. *SequenceNumber*, denotes the order of the stops for the specific trip. *Boarding*, number of travellers getting on and *Alighting*, number of travellers getting off. *DepartureLoad*, number of travellers when departing from the stop and *ArrivalLoad*, number of travellers when arriving at the stop. *StopRouteVariantId*, informs which route the tram is taking. A route can vary depending on time, day and season. *SeatingCapacity*, is the number of seats of the vehicle. *TotalCapacity*, is the maximum number of people allowed on the tram and *ComfortCapacity*, estimation of the number of people that can comfortably get on the tram, seated and standing.

The travel time data of the vehicles are divided into two hours of time-windows from 08 : 00 in the morning to 18 : 00 in the evening. That means, for example, trip that starts after 08 : 00 and finishes before 10 : 00 will be in the 08:00–10:00 time-window. Therefore five time-windows have been used for each day. There are missing records which means that the variant of the route was not operating in that time-window.

Table 2. The data points used in the forecasting.

l—Number of stops	Tram no: 5		Tram no: 2	
	36		27	
	Number of data points for training	Number of data points for forecasting	Number of data points for training	Number of data points for test
08.00–10:00	1440	36	1620	81
10:00–12:00	1692	72	2430	81
12:00–14:00	1820	72	2322	81
14:.00–16:00	1656	72	2403	54
16:00–18:00	2340	72	2997	81

The prediction is performed for tram line 5 (see Fig. 1(a)) and for the tram line 2 (see Fig. 1(b)) for the hours (08:00–18:00). For the tram number 5 and 2, the route variant that we have used in this study has 36 and 27 stops, respectively. The prediction is for every 2 h time-window between 08:00–18:00 and therefore there are five time-windows to forecast. The *StopAreaNumber* and *SequenceNumber* have been taken from Västtrafik database. Together with these data, an hourly database for precipitation and temperature has been downloaded and included to the forecasting process. The database has been obtained via the SMHI, the Swedish Meteorological and Hydrological Institute, website. The *DayType* is coded as 1,2,3,4,5,6 and 7 for the days of the week starting with the monday. If it is a national holiday then that day is coded as zero. The data predicted are *ArrivalLoad*.

The data are split into two data sets as training for the period of 01-01-2020 (day-month-year) to 22-02-2020 and test for the last day, 23-02-2020 (see Table 2). The aim here is to study the next day forecasting ability of the SVR method by using the aforementioned features.

The data scaled are applied by min-max scaling method to an interval of $[-1, 1]$ before the SVR process. The radial basis function (RBF) is chosen as the kernel function, which is written as:

$$K(X_i, X_j) = exp(-\gamma \|X_i - X_j\|^2), \tag{4}$$

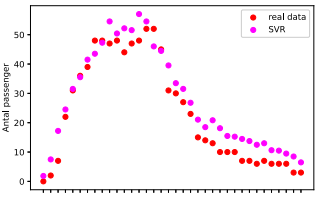
where the parameter γ, intuitively defines the degree to which the effect of a single example of training reaches. In this study $\gamma = 0.96$. We have obtained the best predicitons for $C = 4.3$ and $\epsilon = 0.2$ and kept same for all predictions. The values here agree with the study of Ref. [5].

3 Results and Discussion

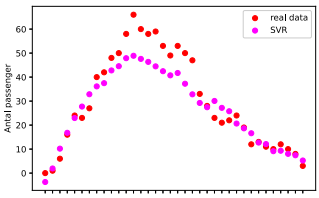
54 days have been used in the study which start at 1st of January to 23rd of February 2020. The 23rd of February has been randomly chosen. The trips of last day (23-02-2020) are predicted by using the trips of the previous days. We would like to clarify that all the parameters in SVR are kept the same for all predictions.

A total of five two-hour time-window predictions for the hours, 08:00–18:00, for the day 23-02-2020 are given in Figs. 2 and 3, and the mean square (MSE) and root mean square error (RMSE) are given in Tables 3 and 4, for the tram line 5 and 2, respectively. In the tables the best approximation is given in a separate column and also highlighted in red.

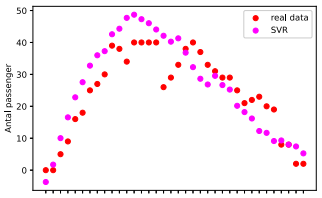
For tram number 5, there are two trips for the time-windows, 10:00–12:00, 12:00–14:00, 14:00–16:00, 16:00–18:00 for 23rd of February 2020, whereas there was only one trip in the time-window 08:00–10:00 (see Fig. 2). For tram number 2, there are three trips for the time-windows, 08:00–10:00, 12:00–14:00, 14:00–16:00, 16:00–18:00 for 23rd of February 2020, there was only two trips in the time-window 08:00–10:00 (see Fig. 3).



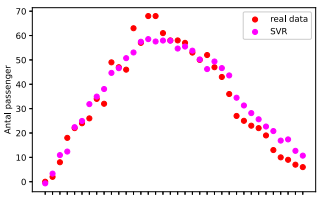
(a) Trip, 08 : 00 − 10 : 00.



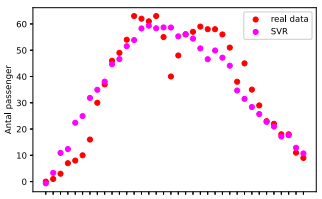
(b) Trip number 1, 10 : 00 − 12 : 00.



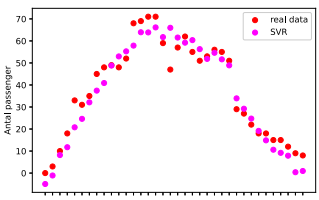
(c) Trip number 2, 10 : 00 − 12 : 00.



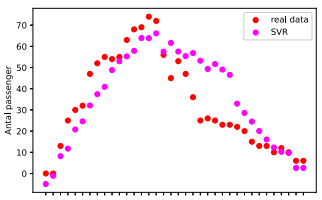
(d) Trip number 1, 12 : 00 − 14 : 00.



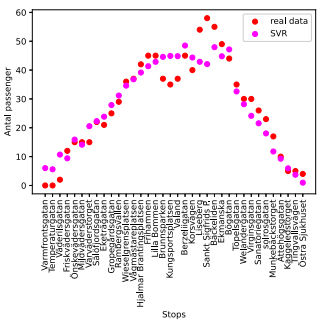
(e) Trip number 2, 12 : 00 − 14 : 00.



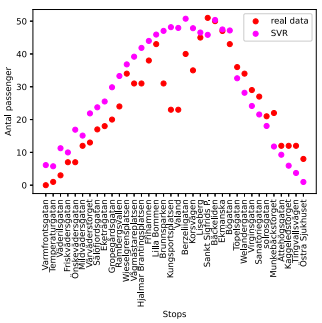
(f) Trip number 1, 14 : 00 − 16 : 00.



(g) Trip number 2, 14 : 00 − 16 : 00.

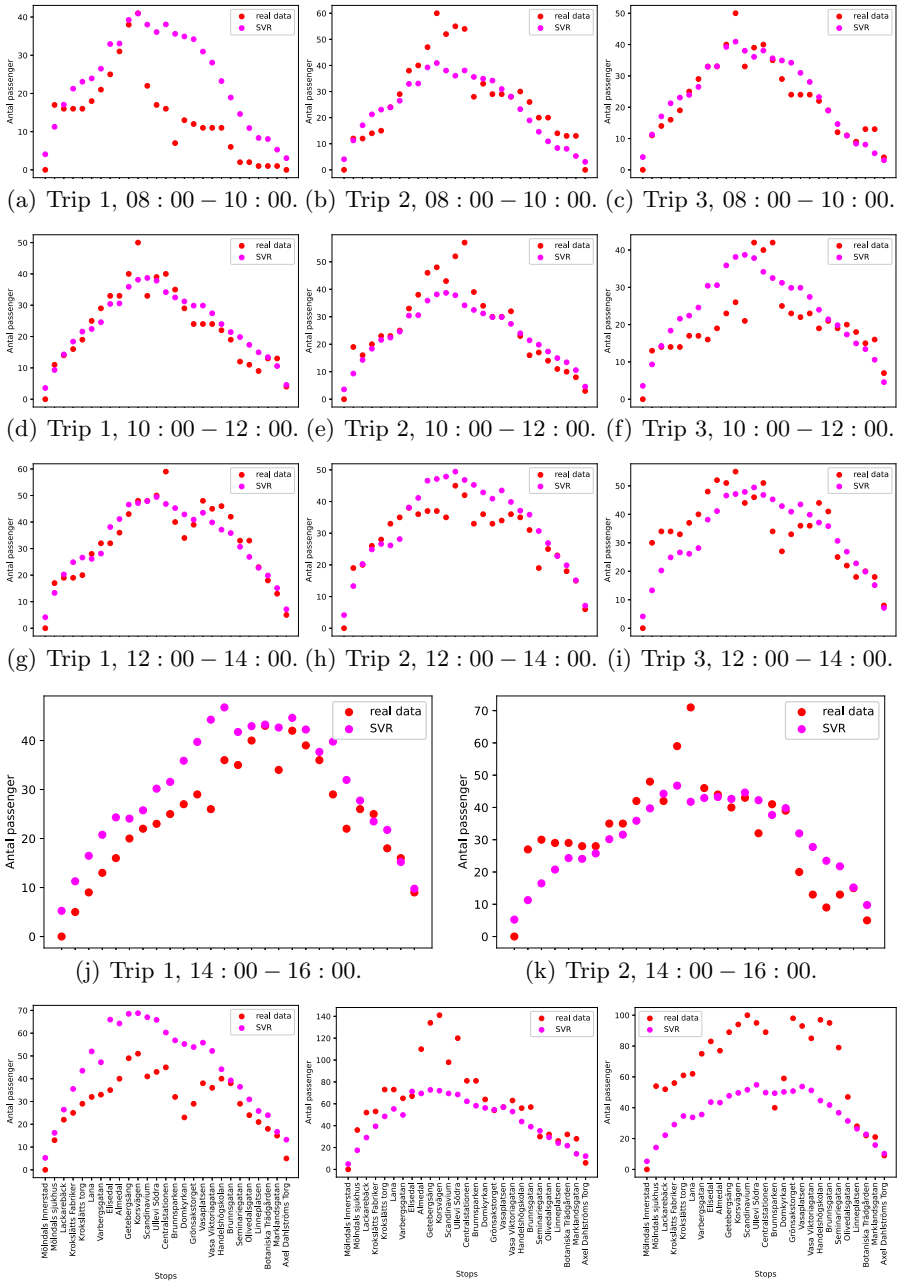


(h) Trip number 1, 16 : 00 18 : 00.



(i) Trip number 2, 16 : 00 − 18 : 00.

Fig. 2. The passenger occupancy predictions for tram number 5 for the time-windows, 08:00–10:00 to 16:00–18:00, each trip and for each 36 stops. The data are for the date 23-02-2020. The number of the figures for each time-window is the number of the trips for that time interval.



(a) Trip 1, 08 : 00 – 10 : 00. (b) Trip 2, 08 : 00 – 10 : 00. (c) Trip 3, 08 : 00 – 10 : 00.

(d) Trip 1, 10 : 00 – 12 : 00. (e) Trip 2, 10 : 00 – 12 : 00. (f) Trip 3, 10 : 00 – 12 : 00.

(g) Trip 1, 12 : 00 – 14 : 00. (h) Trip 2, 12 : 00 – 14 : 00. (i) Trip 3, 12 : 00 – 14 : 00.

(j) Trip 1, 14 : 00 – 16 : 00. (k) Trip 2, 14 : 00 – 16 : 00.

(l) Trip 1, 16 : 00 – 18 : 00. (m) Trip 2, 16 : 00 – 18 : 00. (n) Trip 3, 16 : 00 – 18 : 00.

Fig. 3. The passenger occupancy predictions for tram number 2 for the time-windows, 08:00–10:00 to 16:00–18:00, each trip and for each 27 stops. The data are for the date 23-02-2020. The number of the figures for each time-window is the number of the trips for that time interval.

The features that are used as input data in the forecasting are, *StopAreaNumber*, *SequenceNumber*, *Temperature*, *Precipitation* and *DayType* abbreviated as S1, S2, T, P, D, respectively. For the tram number 5, for the time-window 08:00–10:00, a combination of S1+S2+T (Fig. 2(a)), for the time window 10:00–12:00, S1+S2+T+P (Figs. 2(b), 2(c)), for the time-window 12:00–14:00 and 14:00–16:00, S1+S2+T+D (Figs. 2(d), 2(e) and Figs. 2(f), 2(g), respectively) and finally for the time window 16:00–18:00, S1+S2+T+P+D (Figs. 2(h), 2(i)) gives a better approximation (see also Table 3).

For tram number 2, for the time-windows 08:00–10:00 and 12:00–14:00, a combination of S1+S2 (Figs. 3(a), 3(b), 3(c)) and (Figs. 3(g), 3(h), 3(i), respectively), for the time windows 10:00–12:00, 14:00–16:00 and 16:00–18:00, S1+S2+D, (Figs. 3(d), 3(e), 3(f)), (Figs. 3(j), 3k and Figs. 3l, 3m, 3n, respectively) gives a better approximation (see also Table 4).

Table 3. The passenger occupancy prediction errors for tram number 5, for five time-windows between, 08:00–18:00 (see Figs. 3(a)–3(n)). MSE = mean square error, RMSE = root mean square error.

Tram no: 5	S1+S2		S1+S2+T		S1+S2+P		S1+S2+D		S1+S2+T+P		S1+S2+T+D		S1+S2+P+D		S1+S2+T+P+D		Best approximation
	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	
08.00–10:00	0.0688	0.2623	0.1333	0.3652	0.0599	0.2448	0.0847	0.2910	0.0998	0.3160	0.1423	0.3773	0.0625	0.2501	0.1233	0.3512	S1+S2+P
10:00–12:00	0.1268	0.3561	0.1239	0.3520	0.1224	0.3498	0.1681	0.4100	0.2655	0.5153	0.1657	0.4071	0.1239	0.3520	0.1768	0.4205	S1+S2+T+P
12:00–14:00	0.0896	0.2993	0.1077	0.3282	0.0943	0.3071	0.0851	0.2917	0.098	0.3137	0.0759	0.2755	0.0911	0.3018	0.1301	0.3608	S1+S2+T+D
14:.00–16:00	0.3431	0.5857	0.2610	0.5109	0.3410	0.5839	0.1617	0.4021	0.2654	0.5152	0.1371	0.3703	0.1555	0.3943	0.1414	0.3761	S1+S2+T+D
16:00–18:00	0.3723	0.6102	0.2213	0.4704	0.4187	0.6471	0.0504	0.2246	0.2473	0.4973	0.0595	0.2440	0.0546	0.2338	0.0489	0.2211	S1+S2+T+P+D

Table 4. The passenger occupancy prediction errors for tram number 2, for five time-windows between, 08:00–18:00 (see Figs. 3(a)–3(n)). MSE = mean square error, RMSE = root mean square error.

Tram no: 2	S1+S2		S1+S2+T		S1+S2+P		S1+S2+D		S1+S2+T+P		S1+S2+T+D		S1+S2+P+D		S1+S2+T+P+D		Best approximation
	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	
08.00–10:00	0.2063	0.4542	0.2111	0.4594	0.3458	0.5880	0.3458	0.5880	0.2117	0.4601	0.4389	0.6625	0.3303	0.5747	0.4468	0.6684	S1+S2
10:00–12:00	0.2227	0.4720	0.1761	0.4197	0.2725	0.5220	0.1297	0.3602	0.1902	0.4361	0.1690	0.4111	0.1615	0.4018	0.1788	0.4229	S1+S2+D
12:00–14:00	0.1065	0.3264	0.1269	0.3563	0.1108	0.3329	0.2616	0.5115	0.1284	0.3583	0.1325	0.3641	0.1444	0.3801	0.1359	0.3687	S1+S2
14:.00-16:00	0.2184	0.4673	0.1482	0.3850	0.2150	0.4637	0.1058	0.3253	0.1561	0.3951	0.1072	0.3274	0.1186	0.3444	0.1169	0.3419	S1+S2+D
16:00-18:00	0.7865	0.8868	1.2030	1.0968	0.8834	0.9399	0.5428	0.7367	1.3768	1.1733	1.1572	1.0757	0.6284	0.7927	0.9229	0.9606	S1+S2+D

4 Conclusion

In this study, a Support Vector Regression based machine learning algorithm has been applied to the actual passenger occupancy data for two different routes in Gothenburg, Sweden. The database provided by the Västtrafik Gothenburg is supported by the hourly precipitation and temperature data obtained by the website of SMHI, the Swedish Meteorological and Hydrological Institute. Furthermore, day type features are also added to the forecasting. The prediction results are compared for the combination of all the features, two features always kept in the forecasting, *StopAreaNumber*, *SequenceNumber*. The Västtrafik data

have been used for the hours 08:00–18:00. The data set has been split into five, two-hours time windows. All the parameters are kept for all predictions.

The results shows that by using the previous two months passenger occupancy data, it is possible to predict the next day's occupancy at each stop a tram route. The day that the passenger occupancy is forecasted has been chosen randomly, which is the 23rd of February 2020. The previous data, starting with the 1st of January 2020 has been used in training the forecasting method. The data set used here is relatively small number. However, the approximations follows the real data very close, an exception is the 16:00–18:00 time interval of tram number 2. When we analyze the data we saw that in the 23rd of February 2020, in that time interval, the occupancy was extremly high which did not occur in any previous days.

As an overall result, with the method we have used here the next day's passenger occupancy has been predicted as close as a 5% of mean-square error (MSE) error. For tram number 5, for the day 23-02-2020, we have achieved an approximation to the real data with less than an MSE error of 8% for the three time-windows out of five. For the other two time-windows the errors are 13% and 27%. For tram number 2, for three out of five time-windows, real data have been predicted with less than an MSE error of 13%. The other two are 20% and an extreme occupancy time-window has been estimated with an error of 54%. Therefore we claim that the machine learning algorithm used in this study could be beneficial in public transportation regulators and also forecasting studies to obtain better approximations on next day passenger occupancy in public transportation.

Acknowledgement. We would like to thank to Västtrafik for the data provided. The precipitation and the temperature data downloaded from the SMHI, Swedish Meteorological and Hydrological Institute website. First author thanks to Jonatan Petterson from Västtrafik for the contribution on the dictionary of Västtrafik data.

References

1. Västtrafik. https://www.vasttrafik.se/en/about-vasttrafik/vasttrafik-ab/. Accessed 15 Sept 2021
2. Alam, I., Farid, D.M., Rossetti, R.J.F.: The prediction of traffic flow with regression analysis. In: Abraham, A., Dutta, P., Mandal, J.K., Bhattacharya, A., Dutta, S. (eds.) Emerging Technologies in Data Mining and Information Security, pp. 661–671. Springer, Singapore (2019)
3. Altıntaş, A., Davidson, L.: EMD-SVR: a hybrid machine learning method to improve the forecasting accuracy of highway tollgates traveling time to improve the road safety. In: Martins, A.L., Ferreira, J.C., Kocian, A., Costa, V. (eds.) Intelligent Transport Systems, From Research and Development to the Market Uptake, pp. 241–251. Springer, Cham (2021)
4. Arabghalizi, T., Labrinidis, A.: How full will my next bus be? A framework to predict bus crowding levels (2019). https://doi.org/10.13140/RG.2.2.12969.75368
5. Dahl, M., Brun, A., Kirsebom, O.S., Andresen, G.B.: Improving short-term heat load forecasts with calendar and holiday data. Energies **11**(7), 1678 (2018)

6. Faraj, M.I., Bigun, J.: Synergy of lip-motion and acoustic features in biometric speech and speaker recognition. IEEE Trans. Comput. **56**(9), 1169–1175 (2007)
7. Ghosh, B., Basu, B., O'Mahony, M.: Multivariate short-term traffic flow forecasting using time-series analysis. IEEE Trans. Intell. Transp. Syst. **10**(2), 246–254 (2009)
8. Hong, W.C., Dong, Y., Zheng, F., Lai, C.Y.: Forecasting urban traffic flow by SVR with continuous ACO. Appl. Math. Model. **35**(3), 1282–1291 (2011)
9. Jenelius, E., Cebecauer, M.: Impacts of Covid-19 on public transport ridership in Sweden: analysis of ticket validations, sales and passenger counts. Transp. Res. Interdisciplinary Perspect. **8**, 100242 (2020)
10. Liu, L., Chen, R.C.: A novel passenger flow prediction model using deep learning methods. Transp. Res. Part C Emerg. Technol. **84**, 74–91 (2017)
11. Lu, W., Ma, C., Li, P.: Research on sample selection of urban rail transit passenger flow forecasting based on SCBP algorithm. IEEE Access **8**, 89425–89438 (2020)
12. Lunke, E.B.: Commuters' satisfaction with public transport. J. Transp. Health **16**, 100842 (2020)
13. Ma, Z., Xing, J., Mesbah, M., Ferreira, L.: Predicting short-term bus passenger demand using a pattern hybrid approach. Transp. Res. Part C Emerg. Technol. **39**, 148–163 (2014)
14. Novikov, A., Eremin, S., Kulev, A.: Methodology of passenger public transport organization within the context of long-term territorial development of a city. In: MATEC Web of Conferences, vol. 341, p. 00064. EDP Sciences (2021)
15. Qiu, X., Suganthan, P.N., Amaratunga, G.A.: Short-term electricity price forecasting with empirical mode decomposition based ensemble kernel machines. Procedia Comput. Sci. **108**, 1308–1317 (2017)
16. Salotti, J., Fenet, S., Billot, R., El Faouzi, N.E., Solnon, C.: Comparison of traffic forecasting methods in urban and suburban context. In: 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 846–853. IEEE (2018)
17. Samaras, P., Fachantidis, A., Tsoumakas, G., Vlahavas, I.: A prediction model of passenger demand using avl and apc data from a bus fleet. In: Proceedings of the 19th Panhellenic Conference on Informatics, PCI 2015, pp. 129–134. Association for Computing Machinery, New York, NY, USA (2015)
18. Yu, B., Lam, W.H., Tam, M.L.: Bus arrival time prediction at bus stop with multiple routes. Transp. Res. Part C Emerg. Technol. **19**(6), 1157–1170 (2011)