Machine learning framework to estimate ridership loss in public transport during external crises

Case study of bus network in Stockholm

Movaghar, Mahsa; Jenelius, Erik; Hunter, David

DOI
10.1186/s12544-025-00722-z

Publication date
2025

Document Version
Final published version

Published in
European Transport Research Review

Movaghar *et al.*
European Transport Research Review (2025) 17:37
https://doi.org/10.1186/s12544-025-00722-z

European Transport Research Review

ORIGINAL PAPER

Open Access

Machine learning framework to estimate ridership loss in public transport during external crises: case study of bus network in Stockholm

Mahsa Movaghar[1*], Erik Jenelius[2] and David Hunter[3]

Abstract

Recent technologies for recording and storing data, as well as advancements in data processing techniques, have opened up novel possibilities for urban planners to design a more optimal public transport network. This study aims to initially develop a robust framework for making an insightful understanding of already recorded and available data sets using machine learning approaches. This will give transportation planners a powerful framework to use great recorded datasets to understand the network better and make datasets more meaningful for transport planners. And then introduces an approach to use Machine Learning algorithms and extract hidden patterns for predicting financial loss during any crisis, which is a novel perspective and application. To do this, seven alternative machine learning algorithms were developed to predict ridership: Multiple Linear Regression, Decision Tree, Random Forest, Bayesian Ridge Regression, Neural Networks, Support Vector Regression, and k-Nearest Neighbors. The developed framework was applied to the available 10 years of historical recorded data from the blue bus line number 4 in Stockholm, Sweden. The best model, kNN, with an average R-squared of 0.65 in 10-fold cross-validation, was accepted as the best model. This model is then used to estimate the financial loss of the network during the pandemic in 2020 and 2021. Results reveal a decline of 49% in 2020 and 82% in 2021 in the studied line. Finally, the results were validated with a similar study that analyzed the ticket validations and passenger counts during the spring of 2020.

Keywords Public transport, Ridership, Data-driven prediction, Machine learning, Regression, Financial loss

1 Introduction

Public transport has both direct and indirect effects on environment and economy. Their significant impact on traffic congestion and emissions is inevitable. Recent fuel and energy challenges have brought more interest in designing more attractive networks for both residents and governments. The environmentally friendly features of public transportation lead governments, city planners, and policymakers to optimally design the network which attracts more residents toward public transport.

Demand and supply imbalances are one of the biggest problems in public transportation [20]. This may result in longer trip times, delays, reduced comfort, customer dissatisfaction, and ultimately, a change in users' behavior toward modes of transportation that are more unsustainable [36]. On the other hand, this mismatching may lead to underutilized supply and energy waste, congested roads, and delays, operating the system ineffective from an environmental and financial perspective [34].

Therefore, for the first stage, transportation planners need to accurately estimate the ridership in order to optimally design the network in various scenarios, including

*Correspondence:
Mahsa Movaghar
m.movaghar@tudelft.nl
[1] Department of Transport and Planning, Delft University of Technology, Delft, The Netherlands
[2] Department of Transport Planning, KTH Royal Institute of Technology, Stockholm, Sweden
[3] Sweco AB, Stockholm, Sweden

social crises. However, since there are so many inherent uncertainties in estimating ridership in public transport designing a network that is both user-friendly and cost-effective is challenging for transportation planners, and policymakers [27]. The following characteristics stand out when examining the ridership for public transportation, all of which make planning difficult [2, 3]:

- Ridership patterns fluctuate significantly within various spatial and temporal scales.
- There are an excessive number of independent variables, such as travel time, headway, reliability, cost, weather, and other network features affecting the ridership [11].

Today, more developed, accurate, and automated techniques for planning the public transportation network, together with a greater level of service, are anticipated thanks to a variety of data sources, vast recorded data over time, and more advanced analytical algorithms [27]. This will help transportation planners first design a system that is more economical and environmentally friendly and later be prepared for unseen scenarios and crises.

This study seeks to uncover any hidden patterns in the bus network's recorded journey history data in order to create forecasting models. The findings can provide decision-makers and travel planners with a useful tool to better understand the network's behavior, particularly under extreme and unexpected circumstances where there may not be enough data or records.

During the pandemic years (2020 and 2021), due to lockdowns and social distancing challenges, people preferred to work from home and/or were less willing to use public transport for commuting. Therefore, this caused not only a notable change in travel patterns but also, more significantly, a great financial loss for public authorities. This study aims to initially develop a robust framework for making an insightful understanding of already recorded and available data sets using machine learning approaches. This will lead transportation planners to have a robust framework to use great recorded datasets for a more precise understanding of the network and make datasets more meaningful for transport planners. And then introducing an approach to use machine learning algorithms and extracted hidden patterns for predicting financial loss during any crisis.

To achieve these two goals, the structure of the next sections is as follows: Firstly, an overview of the literature review joint with highlighted contributions of this study is provided. Then in Sect. 3, the steps and the framework is elaborated. Results of different models are compared and discussed in Sect. 4. Estimating the financial loss during pandemic as a practical application of study and the validation of the results is the ultimate goal of this section. The summary of the study and directions for future research are discussed in Sect. 5.

2 Literature review

The latest published articles were analyzed to compile a comprehensive list of both external and internal factors influencing the demand for public transportation. Subsequently, various forecasting methodologies were assessed. To accomplish this, a literature review methodology utilizing various databases was employed to gather information. A thorough search was conducted across Web of Science, Scopus, and Google Scholar databases. This process yielded a significant number of references related to public transport demand. Subsequently, the results were filtered for the past recent years (from 2015), excluding others for reasons such as duplications, irrelevance to the study's scope, or outdated information.

In summarizing the literature review results, the existing literature on predicting public transportation demand can be categorized into six distinct groups, as identified from previously published works. These categories are as follows:

1. **Prediction horizon**: which indicates the length of prediction.
   - *Long-term models* are utilized with a prediction horizon spanning a year to explore the effects of substantial changes within the system and its environment [28].
   - *Short-term models*, operating with a prediction horizon of days or hours, are employed to regulate supply in accordance with immediate needs [21].
2. **Data sources:** The data utilized for data-driven models stems from various sources and methodologies.
   - *Surveys* conducted by Chakrabarti in 2017 [4] and research by Hensher & Rose in 2007 contribute insightful primary data [8].
   - *Historical data* plays a significant role, including studies by Y. Li et al. in 2017 [15], research conducted by Oort, Brands, & Romph in 2015, studies by Oort, Drost, & Yap in the same year [19, 20], and investigations by Xue et al. also in 2015 [35]. This diverse array of data sources enables a comprehensive understanding of the factors influenc-

ing both long-term and short-term models in forecasting public transportation demand.

3. **Methods:** Up until now, numerous projects have relied on past data and records to forecast short-term traffic. These forecasting methods encompass:

   - *Parametric techniques* including historical average [37], smoothing technique [6], and the autoregressive integrated moving average (ARIMA) model [38]. The ARIMA model is specifically employed for predicting traffic flow, travel time, speed, and occupancy.
   - *Non-parametric techniques* encompassing non-parametric regression [23], Kalman filtering models [32], support vector machines [33], the Shepard model [16], neural networks [31], and deep neural networks, which are deep learning algorithms [27]. These methods collectively contribute to short-term traffic forecasting by leveraging a diverse array of modeling approaches and computational strategies.

4. **Spatial Level:** Research endeavors have encompassed diverse spatial levels, ranging from a systemic viewpoint down to the individual vehicle-stop passenger level.
5. **Temporal Level:** Various studies have employed a range of time intervals, spanning from monthly aggregates to observations per passenger per station. For example, Zhou et al. (2017) opted for a narrower focus, utilizing specific hours and weekdays instead of monthly averages to scrutinize intraday trends and patterns in ridership [39].
6. **Affecting Features:**

   - *Temporal features:* Extensive literature has delved into the impact of time and date [35]. Addressing demand seasonality, researchers have employed various methodologies, such as calibrating separate models for peak and off-peak hours [26], devising distinct models for each season [14, 15], integrating dummy variables to denote day types and time periods [18], and employing moving average methods for time series analysis on ridership data [29].
   - *Spatial features:* While some studies have omitted spatial elements in their models, recent research has identified and incorporated several variables influencing ridership. For instance, passenger counts from bus feeder services near metro stations have been utilized to forecast short-term metro ridership [5]. Literature also explores built environment features affecting stop or route attractiveness, including intermodal station connectivity, adjacent business or residential areas, station types, and demographic variables like local population demographics [13].
   - *Other features:* Additional features that impact ridership encompass:

     (a) *Weather:* Extensive literature has explored the influence of weather variables on ridership [7, 12, 14, 39]. Findings indicate a complex relationship as weather variables can indirectly affect travel experiences [7]. However, the impact of seasons appears to be less pronounced compared to the effects of specific weather conditions.

     (b) *Special Events:* Events contribute to increased demand within the transportation network. Some models consider event-related information such as event type/category, proximity to the next event, and dummy variables indicating the presence of an event [21, 22]. Social media data is utilized to gauge event popularity for predicting passenger flow [17]. Conversely, certain studies choose time spans devoid of specific events or holidays to mitigate their influence [39].

     (c) *Holidays:* Ohler et al. (2017) incorporated dummy variables signifying various holiday types into their models [18]. Kalkstein et al. (2009) investigated demand fluctuations before and after holidays [12].

     (d) *Public transport characteristics:* Utilizing clustering analysis, studies have established three clusters based on factors like average headway, route length, number of bus stops, route type, and congestion levels [14]. Subsequently, distinct forecasting models were developed for each cluster. Brakewood et al. (2015) examined the impact of introducing real-time travel information [3]. Other variables such as distance between stations [13], centrality measure evaluating average travel time to all other stations [30], seating and maximum capacity [20], travel time [4, 19], gasoline prices, and bridge tolls have also been investigated in the literature.

     (e) *Socioeconomic factors:* The relationship between the number of cars per household and public transport ridership has been explored in studies conducted by Chakra-

barti (2017) [4] and Spears et al. (2013) [24].

2.1 Literature gaps and contribution

The reviewed literature, categorized into six groups, literature gaps, and the contribution of this study is depicted in Fig. 1.

The main scopes of this study can be summarized as follows:

1. Most already published studies have predicted ridership with time series analysis without including demographic variables. This study regresses the ridership toward different network and demographic variables, including population.
2. Extensive studies have investigated machine learning techniques for predicting ridership in public transport. However, they implement their models on smaller datasets, such as six months of records. This study aims to include all the past records from 2012 to the end of 2021. Although this will bring some new challenges regarding analyzing big data, it opens new opportunities for seasonal analysis and observing the effects of pattern fluctuations within different years and at different levels.
3. What significantly distinguishes this study from other studies using data-driven models for predicting ridership for public transport is the comparison and discussion done over seven different parametric and non-parametric machine learning algorithms. This will give an insightful understanding of the application of these models for similar case studies.
4. Last but not least, the framework developed in this study is a novel perspective and application of machine learning algorithms to estimate the financial loss of systems in case of unexpected crises and outbreaks, such as pandemics.

3 Methods

This project can be categorized as a supervised regression data analysis problem.To create a model for estimating the number of boardings, boardings are regressed toward time (one feature), station (31 features), month (11 features), and population (one feature) using a total number of 44 features. The overview of the framework and steps used in this study is visualized in Fig. 2.

3.1 Data preprocessing

In the available dataset, boarding is recorded as a monthly (excluded July) average number of persons per bus arrival according to the timetable. Since the timetable was not fixed during the study period and throughout the year, the boardings were aggregated per hour in the preprocessing phase.

Then all the data were standardized and normalized to ensure that all the features are scaled in a similar range, and all the machine learning algorithms are precisely applied to the dataset.

Moreover, to deal with categorical dependent variables (i.e. stations and months), dummy encoding (one-hot) was used. To avoid the co-linearity in the input matrix, one of the features was also dropped.

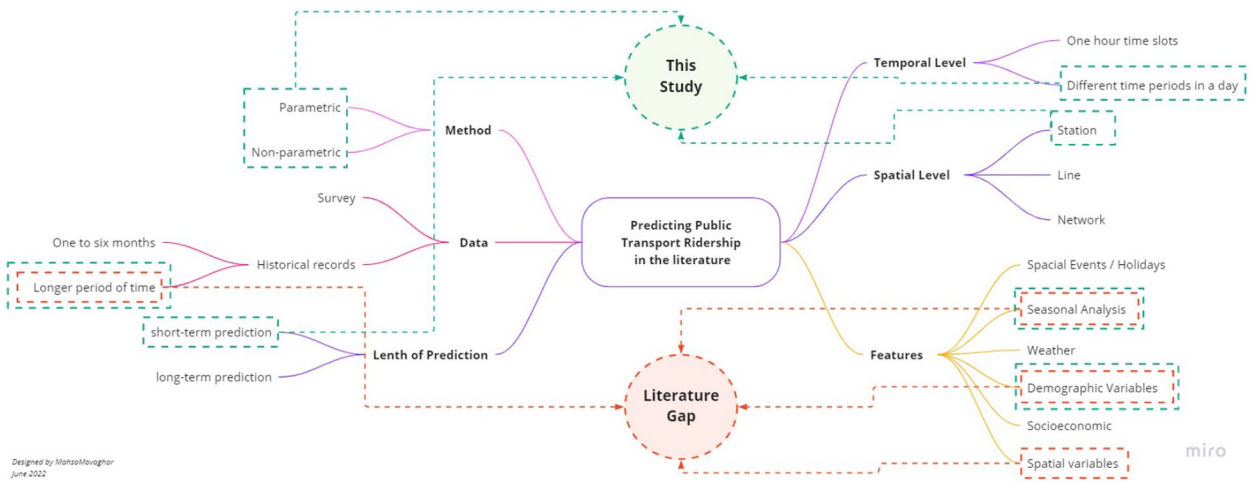

Fig. 1 Summary of the literature for predicting public transport ridership, highlighting research gaps and the contribution of the current study

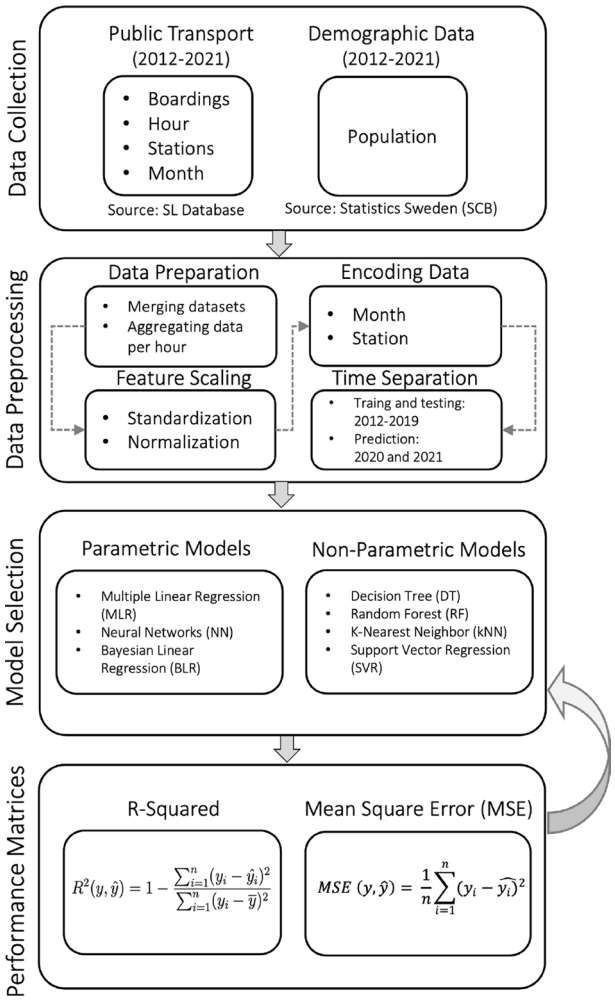

Fig. 2 Overview of the framework developed in this study

3.2 Model selection and tuning

Each of the seven selected models tuned for their parameters to achieve the best performance: (1) Linear Regression (LR): with and without coefficient. (2) Decision Tree (DT): Maximum depth. (3) Random Forest (RF): Maximum depth and number of estimators. (4) Bayesian Linear Regression (BLR): with and without intercept. (5) One layer Neural Network (NN): Number of hidden layers and activation functions (i.e. identity, logistic, tanh, and relu). (6) Support Vector Regression (SVR): kernel (i.e. linear, polynomial, RBF, sigmoid), and (7) k-Nearest Neighbour (kNN): number of neighbors.

Although ridership is inherently a time-series dataset, traditional time-series models such as ARMA or ARIMA were not selected for this study due to their limitations in incorporating additional influential variables [9], such as station characteristics, population, and time-of-day effects. Other machine learning models were chosen instead as they offer greater flexibility in capturing complex and understanding, non-linear relationships between these external factors and ridership patterns, including demographic data and spatial characteristics.

Using the coefficient of determination (R-Squared) and Mean Squared Error (MSE) as the performance metrics, the accuracy of the models have been examined. To ensure that every observation has the chance to show up in both the training and test sets, k-fold cross-validation was used. Using one fold as the test set and the remaining sets as the training set in each run, the process was repeated ten times. The final model performance was reported using the average achieved in each run.

As explained earlier in the introduction, to find the inherent patterns within the recorded data, all of the models were developed, trained, and tested using data collected prior to 2020 when the Covid pandemic began.

Based on the performance indices and other criteria due to specific characteristics of the data, the best model was chosen. This model reveals the best understanding of the hidden patterns within the recorded data set and can be the most reliable model for prediction. Finally, assuming that there was no change in travel patterns, the model was used to predict the ridership in 2020 and 2021. Comparing the predicted values and recorded data implies the financial loss of the whole network during the pandemic. This comparison reveals a quite reasonable estimation of the public authorities' revenue if no pandemic had occurred.

4 Results and discussion

In this section, initially, a brief overview of the case study and data sets used in this study will be given. Then, some descriptive analysis of the data will be presented. The section will be followed by performance indices of different developed machine learning models. A discussion over criteria for choosing the best model will be provided. Finally, the predicted results for the years 2020 and 2021 will be explained. This section ends with results validation with other published studies.

4.1 Case study

The case study focuses on public transport in Stockholm, the capital and largest city in Sweden as well as the largest urban area in Scandinavia. Stockholm, with a current population of 987,661 (June 30, 2023) is estimated to reach a population of almost 3 million by 2045, which highlights the importance of increasing the urban facilities, public transport, and transportation infrastructure in the next 20 years. Public transport in Stockholm consisting of 8 different modes (i.e. bus, metro, commuter rail, inner-city rail, regional rail, light rail, tram, and commuter ferries) is authorized by SL (Storstockholms Lokaltrafik). The dense bus network in Stockholm is operating

within 502 bus lines and 6710 stops. Five main blue bus lines in the inner city, with express character, higher frequencies, shorter distances between two consecutive stops, moving on the inner city main roads, commute the most passengers of the bus network in Stockholm. Among all these bus lines, blue bus line number 4, with the longest length and the busiest bus line in the entire of Sweden, is chosen for detailed exploration within this study [25]. This line has been operating with 28 stations only since 2021. Therefore, the removed three stations have not been removed from the dataset.

Besides the important role of public transport in Stockholm, what makes this case study even more interesting is the quite different strategy taken by public authorities in Sweden during the pandemic. Unlike many other countries, Sweden decided to limit the actions recommendations rather than obligation and full lockdowns. So, residents were recommended to stay at home if they prefer or if they feel sick. Therefore, no restrictions were implied on public transport services, and the supply remained almost unchanged, despite many other countries. Hence, any decline in public transport ridership can be related to residents' own choices [10].

4.2 Datasets

As discussed earlier, ridership in public transport is affected by numerous variables, both within public transport networks and other factors such as population, weather, quality of the network, and so on. Having these many variables raises one of the most important challenges of this Machine Learning study which is the overfitting trap. Adding too many features and variables may result in more accurate training results. However, it is likely to decrease the accuracy of predictions on new data sets. Moreover, adding more variables may increase the model's accuracy, but it reduces the explainability of the model. Therefore, in order to avoid overfitting traps and try to make the results interpretable for transportation planners, the affecting variables are limited to time (one feature), month (11 features), population (one feature), and station (31 features) with a total of 44 features. To ensure transparency, it is important to note that the data for July was not included in the dataset provided by SL for any of the years, and therefore, was excluded from the analysis.

A list of all variables and datasets is introduced in Table 1.

Table 1 Variables and datasets

Variables	Type	Time period	Source
Population	Quantitative	Jan 2012 - Dec 2021	SCB
Year	Qualitative	Jan 2012 - Dec 2021	SL
Time	Quantitative	Jan 2012 - Dec 2021	SL
Month	Qualitative	Jan 2012 - Dec 2021	SL
Station	Qualitative	Jan 2012 - Dec 2021	SL
Boardings	Quantitative	Jan 2012 - Dec 2021	SL

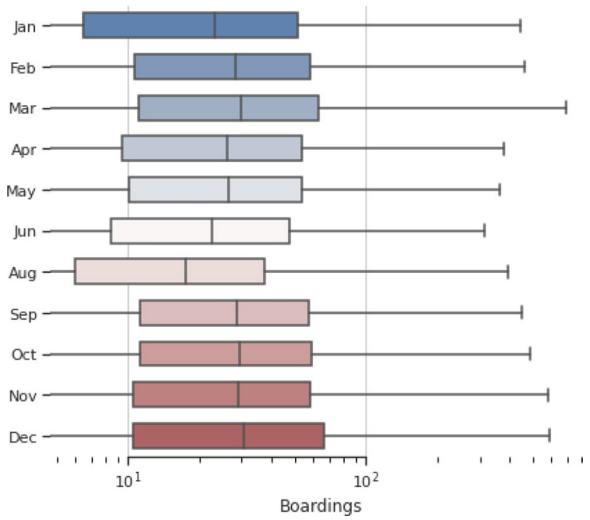

Fig. 3 Box plots for boardings per month (2012-2021)

4.3 Results

To start with, three different patterns were recognized in the 10 years of operation of the network: (1) Throughout a year (Fig. 3), (2) Throughout a route (Fig. 4), and (3) Throughout a day (Fig. 5). According to Fig. 3, August and January exhibit the lowest and highest ridership during the study period, respectively. Although the exact causal relationship behind these patterns requires further investigation [1], these fluctuations may be linked to the number of holidays and vacation periods in these two months. Figure 4 highlights the boarding patterns throughout the line and at various stops. The stops with higher boardings are those located near metro stations and/or other modes of transportation, such as trams and commuter trains. Fig. 5 demonstrates the morning (7–9) and afternoon (15–17) peak hours for the stations with the highest demand.

The predicted daily average boardings per year comparing different models are depicted in Fig. 6. In this figure, the blue line represents the real observations. This blue line shows the significant ridership decrease after 2019, which was the obvious impact of covid pandemic. Each dashed line shows the predicted values with a trained machine-learning model using data from 2012 to 2019. It is worth noting that predicted values for 2020 and 2021 are the estimation of average daily boardings if no pandemic had occurred since the recorded data

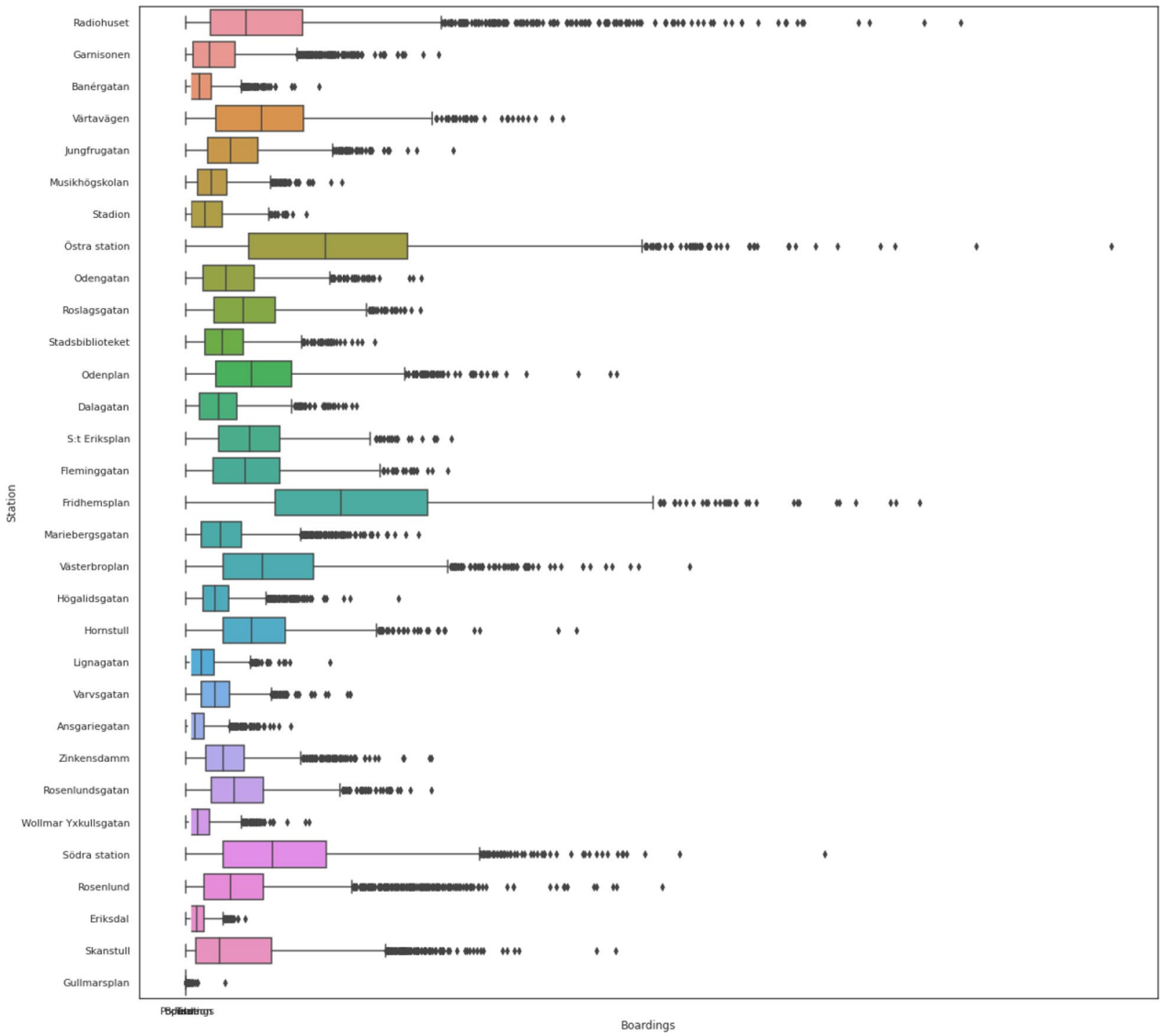

Fig. 4 Box plots for boardings per station (2012-2021)

for these two years have been removed from the training dataset. Therefore, the difference between the blue line (real observations) and the dashed lines (prediction based on trained models) reveals the loss of the network system during the pandemic due to the change in residence travel behavior in 2020 and 2021.

The results of the seven developed models, along with their hyperparameters that yielded the best performance for the 2012-2019 data, are summarized in Table 2. To choose the best reliable model and to mitigate the risk of overfitting, the results are reported using k-fold cross-validation with 10 randomized folds. Results reveal that Random Forest (RF), Neural Network (NN), and k-Nearest Neighbor (kNN) have the best accuracy. According to average accuracy, Random Forest (RF) has the highest value of 0.74 and a promising standard deviation of 0.072. However, Random Forest (RF), as well as all tree-based models, are not suggested for predicting values out of the range of values used in the training dataset. The second-best model, Neural Network (NN), with an average R-squared of 0.71, has also the highest standard deviation of 0.27. This high standard deviation highlights the undesirable possibility of this model overfitting. Ultimately, the third-best model, kNN, with an average R-squared of 0.65, is accepted as the best-trained model. The low standard deviation of 0.05, which highlights the stability of the model on new datasets,

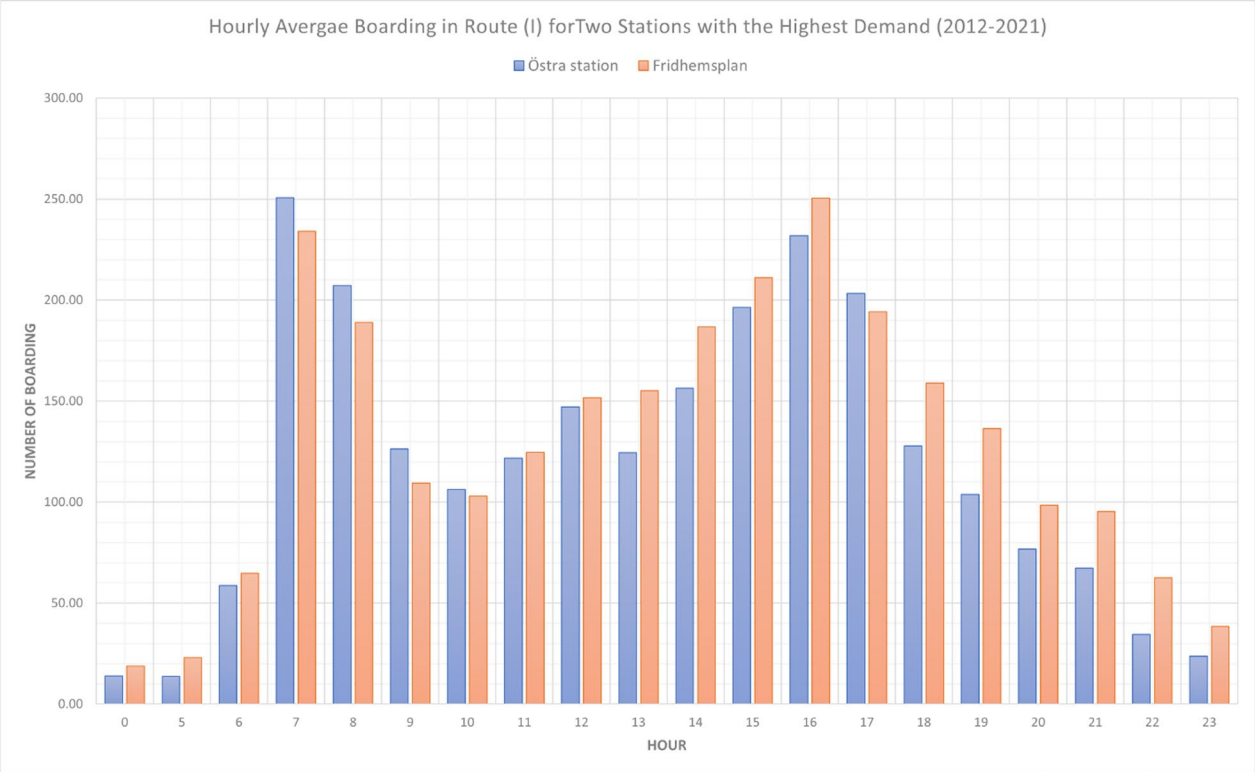

Fig. 5 The average number of boardings for two stations with the highest demand (2012-2021)

besides the explainability and low running time, adds more value to this model within the scope of this study.

To achieve the final goal of this, i.e. estimating the financial loss of the network during the pandemic year, the developed kNN model with 5 neighbors is used to predict the average annual ridership (Table 3).

Results reveal that using the developed KNN model for predicting the values for years 2012-2019, an average error of 1.37 % with a maximum of 4.87% in 2018 is achieved. Now the same model is used for predicting the numbers in 2020 and 2021. It is worth highlighting that the predicted values in these two years are interpreted as the expected ridership for public transport if no pandemic outbreak had happened based on the inherent pattern in the past recorded data. The results showed a decrease of 49% in 2020 and 82% in 2021. It is then estimated that this bus line, over two years of the pandemic, lost on average, 12,667.34 passengers daily in only one direction. This reveals a significant financial loss for the whole system and operational organizations in Stockholm during the pandemic.

4.4 Results validation

The decline in public transport ridership in Stockholm has been studied in another study. Jenelius and Cebecauer (2020) analyzed the ticket validations, sales, and passenger counts data during the spring of 2020 to investigate changes in travel patterns during the pandemic. Since no restrictions were implied by public authorities on public transport services and supply, the declined ridership is referred only to travelers' choices [10]. Their results show a 40%–60% reduction across all the regions and all modes of transportation in Stockholm. Their separate analysis of different modes of public transport in the Stockholm region reveals the deduction of almost 40% for buses during three months of their study period (March to June 2020) [10]. This is quite an interesting comparison with the predicted values for the entire of 2020 with the developed kNN machine learning model based on historical data and input features in the current study (Table 3). Using the developed kNN model, the estimation of declined ridership is depicted in Fig. 7 per month in 2020. Results reveal a decrease of 45% comparing the predicted results from the model and the real observations in May. This finding is in line with Jenelius and Cebecauer (2020) using the ticket validation data. This highlights the importance and accuracy of data-driven models for predicting future conditions based on the inherent patterns in historical data. This will open up

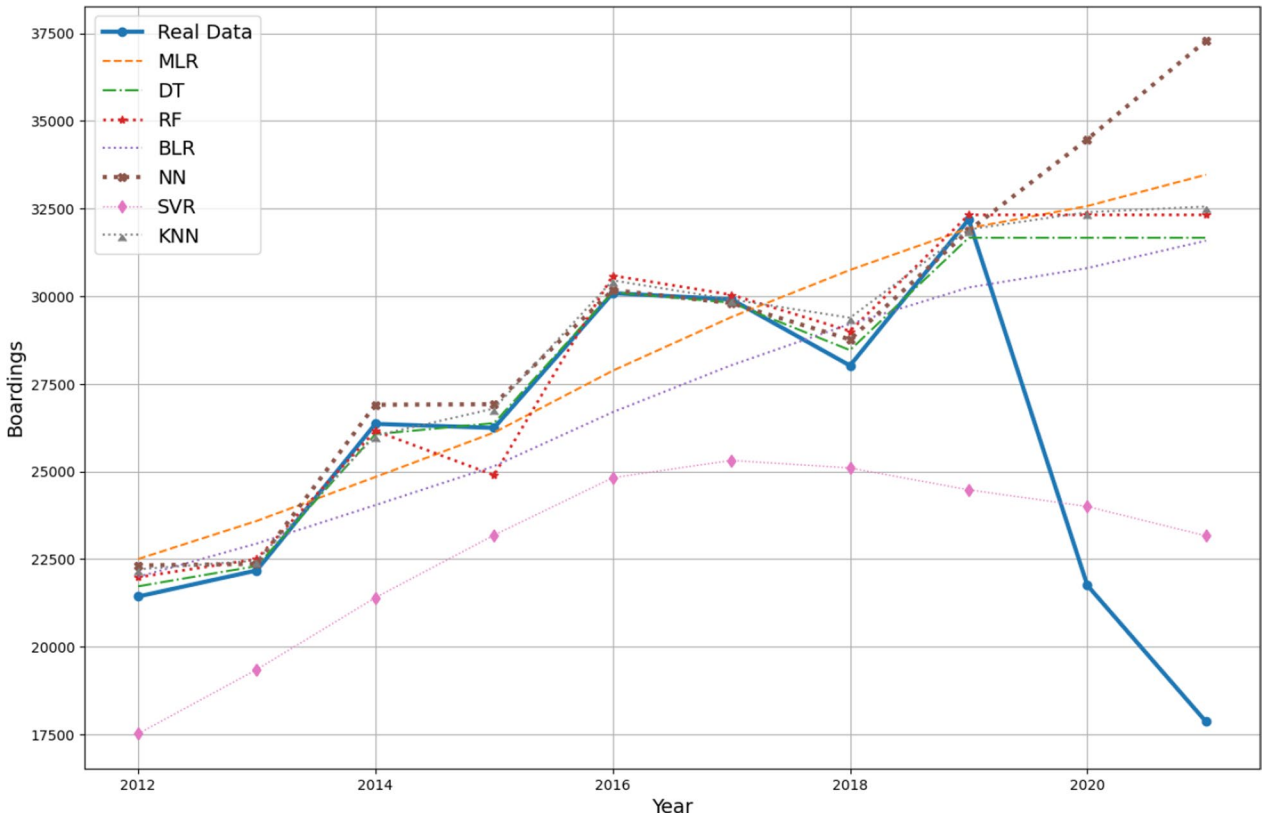

Fig. 6 Comparison between all developed machine learning models (dashed lines) and real observations (blue line)

Table 2 Summary of developed models and their performance

Model	Parameters	Value	Separate train/split			Cross validation		
			R^2	Mean squared error	Training time	k	Average	std
MLR	Intercept	yes	0.395	1545.2367	182 ms	10	0.3703	0.064
DT	Max depth	30	0.7083	744.902	522 ms	10	0.6139	0.0877
RF	Max depth	50	0.8459	393.6823	32.5 s	10	0.7367	0.0724
	# of estimators	150						
BLR	Intercept	yes	0.34804	1665.1382	383 ms	10	0.288	0.1826
NN	Hidden layer	1	0.8911	278.0442	10 min 31 s	10	0.7140	0.2758
	Hidden nodes	150						
	Activation function	tanh						
SVR	Kernel	rbf	0.5687	1101.6712	3 min 22 s	10	−66.86	22.56
kNN	# of neighbors	5	0.7613	609.7122	38.3 s	10	0.6511	0.0477

Table 3 Predicting ridership in different years using the developed kNN

Year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Real observation	21,432.90	22,177.01	26,358.02	26,242.52	30,085.24	29,909.08	28,019.93	32,186.60	21,752.63	17,863.63
Predicted	22,211.03	22,459.00	26,021.75	26,794.92	30,450.12	29,907.22	29,383.70	31,904.22	32,395.45	32,555.48
Difference	778.13	281.99	−336.27	552.41	364.88	−1.86	1,363.77	−282.38	10,642.82	14,691.85
Percentage %	3.63	1.27	−1.28	2.11	1.21	−0.01	4.87	−0.88	48.93	82.24

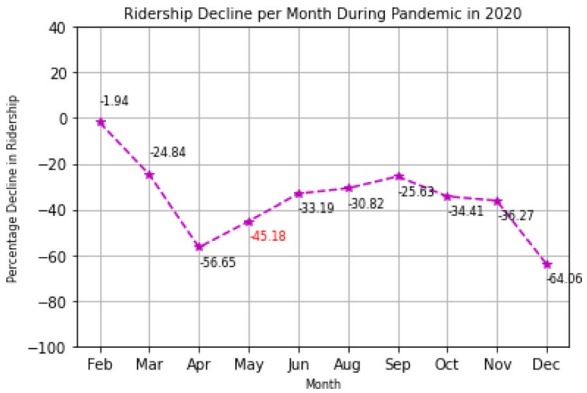

Fig. 7 Declined in ridership in 2020 per month due to pandemic estimated based on the kNN developed model

a significant application of data-driven models for predictions during future pandemics or crises.

5 Conclusions

Designing an optimum public transport while avoiding any demand and supply mismatching is of high importance for city planners due to the population increase in urban areas. Advances and techniques in recording data, on the one hand, and analyzing big datasets, on the other hand, have made researchers interested to make available big recorded datasets meaningful and insightful for transportation planners and policymakers. This study was initiated to fulfill two main goals: (1) First and foremost, to find inherent patterns and insights within available datasets using machine learning algorithms. To do this, the number of boardings in blue bus line 4 in Stockholm regressed toward time, month, station, and population, using seven different algorithms that were trained and tested over data from 2012 to 2019. (2) Then, the best model was used to estimate the financial loss of the system during the pandemic years and lockdowns. In order to achieve this, the third best model based on 10-fold cross-validation, kNN with 5 neighbors, was picked for predicting the values for 2020 and 2021. Results revealed a financial loss of 49% in 2020 and 82% in 2021.

Unlike some other models, such as Neural Networks (NN), kNN did not exhibit signs of overfitting or instability, which made it more reliable for the studied dataset. The simplicity and interpretability of the kNN model also made it a favorable choice for the study, given the various and non-linear input features.

The proposed framework in this paper is transferable to other cities and regions, especially when historical ridership data is available. However, the performance of different models may vary in any new city, as local input factors, such as population, infrastructure, and socio-economic conditions, can significantly influence the results. The ability to predict financial losses and ridership patterns during unexpected events, using models trained on historical data with features selected based on local factors, along with the use of cross-validation for model validation and the process for handling and preprocessing historical data, could provide valuable insights for public transport planners and policymakers.

The currently available dataset was aggregated monthly data. Therefore, no detailed studies have been done on the effect of working days, weekends, and big events. Different models for different types of day, coupled with new datasets such as weather, station distances to the city centers, and big business and residential areas, are interesting topics for further exploration. Future work could also explore the integration of simulated data to supplement real-world datasets, particularly in scenarios with limited historical records, to enhance model performance and reliability.

Acknowledgements
The authors would like to acknowledge the infrastructure and support of SWECO AB, Stockholm, during this project.

Authors contributions
Mahsa Movaghar: conceptualized and designed the study, collected and analyzed the data, and drafted themanuscript. Erik Jenelius: supervised the entire research and critically revised the manuscript. David Hunter: contributed to thestudy design and participated in data collection.

Funding
This research was conducted without external funding. All aspects of the study, including data collection, analysis, and manuscript preparation, were supported by the resources and facilities available at Sweco AB, Stockholm. The authors declare no financial or material interests that could have influenced the research.

Availability of data and material
The data and materials used in this study are available upon reasonable request to ensure transparency and facilitate scientific collaboration. We are committed to sharing our research findings while adhering to ethical and legal considerations, including data privacy and confidentiality.

Declarations

Competing interests
The authors declare no competing interests.

Received: 16 October 2023 Accepted: 26 March 2025
Published online: 28 July 2025

References

1. Almlöf, E., Rubensson, I., Cebecauer, M., & Jenelius, E. (2021). Who continued travelling by public transport during covid-19? socioeconomic factors explaining travel behaviour in stockholm 2020 based on smart card data. In European transport research review, 13 , 31. https://doi.org/10.1186/s12544-021-00488-0
2. Berrebi, S. J., Watkins, K. E., & Laval, J. A. (2015). A real-time bus dispatching policy to minimize passenger wait on a high frequency route. *Transportation Research Part B: Methodological, 81*, 377–389. https://doi.org/10.1016/j.trb.2015.05.012

3. Bordagaray, M., dell'Olio, L., Ibeas, A., & Cecín, P. (2014). Modelling user perception of bus transit quality considering user and service heterogeneity. *Transportmetrica A: Transport Science, 10*, 705–721. https://doi.org/10.1080/23249935.2013.823579
4. Chakrabarti, S. (2017). How can public transit get people out of their cars? an analysis of transit mode choice for commute trips in los angeles. *Transport Policy, 54*, 80–89. https://doi.org/10.1016/j.tranpol.2016.11.005
5. Ding, C., Wang, D., Ma, X., & Li, H. (2016). Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability (Switzerland), 8*. https://doi.org/10.3390/su8111100
6. Gong, M., Fei, X., Wang, Z. H., & Qiu, Y. J. (2014). Sequential framework for short-term passenger flow prediction at bus stop. *Transportation Research Record, 2417*, 58–66. https://doi.org/10.3141/2417-07
7. Guo, Z., Wilson, N. H., & Rahbee, A. (2007). Impact of weather on transit ridership in Chicago, Illinois. *Transportation Research Record*. https://doi.org/10.3141/2034-01
8. Hensher, D. A., & Rose, J. M. (2007). Development of commuter and non-commuter mode choice models for the assessment of new public transport infrastructure projects: A case study. *Transportation Research Part A: Policy and Practice, 41*, 428–443. https://doi.org/10.1016/j.tra.2006.09.006
9. Hightower, A., Ziedan, A., Guo, J., Zhu, X., & Brakewood, C. (2024). A comparison of time series methods for post-covid transit ridership forecasting. *Journal of Public Transportation, 26*, 100097. https://doi.org/10.1016/j.jpubtr.2024.100097
10. Jenelius, E., & Cebecauer, M. (2020). Impacts of covid-19 on public transport ridership in Sweden: Analysis of ticket validations, sales and passenger counts. *Transportation Research Interdisciplinary Perspectives*. https://doi.org/10.1016/j.trip.2020.100242
11. Kacprzyk, J. (2017). Intelligent transport systems and travel behaviour (Vol. 505; G. Sierpinski, Ed.). Springer International Publishing.
12. Kalkstein, A. J., Kuby, M., Gerrity, D., & Clancy, J. J. (2009). An analysis of air mass effects on rail ridership in three us cities. *Journal of Transport Geography, 17*, 198–207. https://doi.org/10.1016/j.jtrangeo.2008.07.003
13. Kuby, M., Barranda, A., & Upchurch, C. (2004). Factors influencing light-rail station boardings in the united states. *Transportation Research Part A: Policy and Practice, 38*, 223–247. https://doi.org/10.1016/j.tra.2003.10.006
14. Li, L., Wang, J., Song, Z., Dong, Z., & Wu, B. (2015). Analysing the impact of weather on bus ridership using smart card data. *IET Intelligent Transport Systems, 9*, 221–229. https://doi.org/10.1049/iet-its.2014.0062
15. Li, Y., Wang, X., Sun, S., Ma, X., & Lu, G. (2017). Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks. *Transportation Research Part C: Emerging Technologies, 77*, 306–328. https://doi.org/10.1016/j.trc.2017.02.005
16. Long, S., & Dhillon, B.S. (Eds.) (2019). *Man-machine-environment system engineering* (vol. 527). Singapore: Springer.
17. Ni, M., He, Q., & Gao, J. (2017). Forecasting the subway passenger flow under event occurrences with social media. *IEEE Transactions on Intelligent Transportation Systems, 18*, 1623–1632. https://doi.org/10.1109/TITS.2016.2611644
18. Ohler, F., Krempels, K. H., & Möbus, S. (2017). *Forecasting public transportation capacity utilisation considering external factors* (pp. 300–311). SciTePress. https://doi.org/10.5220/0006345703000311
19. Oort, N. V., Brands, T., & Romph, E. D. (2015). Short-term prediction of ridership on public transport with smart card data. *Transportation Research Record, 2535*, 105–111. https://doi.org/10.3141/2535-12
20. Oort, N.V., Drost, M., Yap, M. (2015). Data-driven public transport ridership prediction approach including comfort aspects. Conference on advanced systems in public transport (caspt2015).
21. Pereira, F. C., Rodrigues, F., & Ben-Akiva, M. (2015). Using data from the web to predict public transport arrivals under special events scenarios. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, 19*, 273–288. https://doi.org/10.1080/15472450.2013.868284
22. Rodrigues, F., Borysov, S. S., Ribeiro, B., & Pereira, F. C. (2017). A Bayesian additive model for understanding public transport usage in special events. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*, 2113–2126. https://doi.org/10.1109/TPAMI.2016.2635136
23. Smith, B. L., Williams, B. M., & Oswald, R. K. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies, 10*, 303–321. https://doi.org/10.1016/S0968-090X(02)00009-8
24. Spears, S., Houston, D., & Boarnet, M. G. (2013). Illuminating the unseen in transit use: A framework for examining the effect of attitudes and perceptions on travel behavior. *Transportation Research Part A: Policy and Practice, 58*, 40–53. https://doi.org/10.1016/j.tra.2013.10.011
25. Stockholm, R. (2021). Information om rapport effektivare hållplatsstopp - påstigning alla dörrar på stombusslinje 4 (Tech. Rep.). Stockholm, Sweden: Author.
26. Stopher, P. R. (1992). Development of a route level patronage forecasting method. *Transportation, 19*, 201–220. https://doi.org/10.1007/BF01099977
27. Tang, T., Fonzone, A., Liu, R., & Choudhury, C. (2021). Multi-stage deep learning approaches to predict boarding behaviour of bus passengers. *Sustainable Cities and Society*. https://doi.org/10.1016/j.scs.2021.103111
28. Toqué, F., Khouadjia, M., Come, E., Trepanier, M., & Oukhellou, L. (2017). Short & long term forecasting of multimodal transport passenger flows with machine learning methods. In *2017 ieee 20th international conference on intelligent transportation systems (itsc)* (pp. 560–566). https://doi.org/10.1109/ITSC.2017.8317939
29. Tsai, T. H., Lee, C. K., & Wei, C. H. (2009). Neural network based temporal feature models for short-term railway passenger demand forecasting. *Expert Systems with Applications, 36*, 3728–3736. https://doi.org/10.1016/j.eswa.2008.02.071
30. Upchurch, C., & Kuby, M. (2014). Evaluating light rail sketch planning: Actual versus predicted station boardings in phoenix. *Transportation, 41*, 173–192. https://doi.org/10.1007/s11116-013-9499-9
31. Vlahogianni, E. I., Golias, J. C., & Karlaftis, M. G. (2004). Short-term traffic forecasting: Overview of objectives and methods. *Transport Reviews, 24*, 533–557. https://doi.org/10.1080/0144164042000195072
32. Wang, Y., Papageorgiou, M., & Messmer, A. (2007). Real-time freeway traffic state estimation based on extended Kalman filter: A case study. *Transportation Science, 41*, 167–181. https://doi.org/10.1287/trsc.1070.0194
33. Wang, Z., Yang, C., & Zang, C. (2018). *Short-term passenger flow prediction on bus stop based on hybrid model*. Atlantis Press. https://doi.org/10.2991/ecae-17.2018.74
34. Wu, W., Liu, R., Jin, W., & Ma, C. (2019). Stochastic bus schedule coordination considering demand assignment and rerouting of passengers. *Transportation Research Part B: Methodological, 121*, 275–303. https://doi.org/10.1016/j.trb.2019.01.010
35. Xue, R., Sun, D. J., & Chen, S. (2015). Short-term bus passenger demand prediction based on time series model and interactive multiple model approach. In *Discrete dynamics in nature and society*. https://doi.org/10.1155/2015/682390
36. Yao, E., Liu, T., Lu, T., & Yang, Y. (2020). Optimization of electric vehicle scheduling with multiple vehicle types in public transport. *Sustainable Cities and Society*. https://doi.org/10.1016/j.scs.2019.101862
37. Zhang, J., Shen, D., Tu, L., Zhang, F., Xu, C., Wang, Y., & Li, Z. (2017). A real-time passenger flow estimation and prediction method for urban bus transit systems. *IEEE Transactions on Intelligent Transportation Systems, 18*, 3168–3178. https://doi.org/10.1109/TITS.2017.2686877
38. Zhou, C., Dai, P., & Li, R. (2013). *The passenger demand prediction model on bus networks* (pp. 1069–1076). IEEE Computer Society. https://doi.org/10.1109/ICDMW.2013.20
39. Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., & Cao, R. (2017). Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation Research Part C: Emerging Technologies, 75*, 17–29. https://doi.org/10.1016/j.trc.2016.12.001

Publisher's Note