https://doi.org/10.3311/PPtr.22252

Periodica Polytechnica Transportation Engineering, 51(4), pp. 357–374, 2023

Benchmarking Travel Time and Demand Prediction Methods Using Large-scale Metro Smart Card Data

Iyad Zimmo[1], Daniel Hörcher[1*], Ramandeep Singh[1], Daniel J. Graham[1]

[1] Transport Strategy Centre, Department of Civil and Environmental Engineering, Faculty of Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom
[*] Corresponding author, e-mail: d.horcher@imperial.ac.uk

Received: 23 March 2023, Accepted: 26 April 2023, Published online: 21 June 2023

Abstract

Urban mass transit systems generate large volumes of data via automated systems established for ticketing, signalling, and other operational processes. This study is motivated by the observation that despite the availability of sophisticated quantitative methods, most public transport operators are constrained in exploiting the information their datasets contain. This paper intends to address this gap in the context of real-time demand and travel time prediction with smart card data. We comparatively benchmark the predictive performance of four quantitative prediction methods: multivariate linear regression (MVLR) and semiparametric regression (SPR) widely used in the econometric literature, and random forest regression (RFR) and support vector machine regression (SVMR) from machine learning. We find that the SVMR and RFR methods are the most accurate in travel flow and travel time prediction, respectively. However, we also find that the SPR technique offers lower computation time at the expense of minor inefficiency in predictive power in comparison with the two machine learning methods.

Keywords

machine learning, urban public transport, demand prediction, smart card data

1 Introduction

Information and Communication Technologies (ICT) are widely applied in the public transport sector; common examples include vehicle automation, electronic fare collection, and safety and security processes. These digital solutions in public transport operations generate vast volumes of data. In this paper we focus on the utilisation of smart card ticketing data. Smart card data (SCD) is a popular source of information in the emerging literature of transport planning and economics, primarily because it records many characteristics of the observed demand patterns and user behaviour almost immediately, which enables short-term or even real-time decision making. By capturing

1. the volume of travellers at every entry and exit point of the public transport network, and
2. the speed of movement between the tap-in and tap-out locations,

analysts will be able to predict travellers flows and travel times in real-time with high precision, which can be extremely valuable for both the operator (e.g., see Drabicki et al. (2021); Hörcher and Graham (2021); Sun et al. (2014)) and the traveller in preparation for unexpected events such as sudden demand shocks or delays. The literature documents many quantitative prediction methods in the context of public transport demand and travel speed, but very few of them are utilised by transit agencies. At the same time, operators are challenged in dealing with big data arise from issues related to storage, transmission, management, processing, analyses, visualisation, security, privacy and many others (Bahsoon et al., 2017).

The short-run prediction of demand characteristics has recently become increasingly relevant in the well-known context of the COVID-19 pandemic (Tirachini and Cats, 2020). The risk of infection in public transport is determined by two travel attributes: the duration of the trip and the occupancy rate of the vehicle taken, assuming that the occupancy rate is inversely proportional to the average distance between passengers (Hörcher et al., 2022). Naturally, risk-averse passengers intend to reduce the trip duration as well as the prevailing level of crowding as much as possible. If a social distancing regulation

Cite this article as: Zimmo, I., Hörcher, D., Singh, R., Graham, D. J. (2023) "Benchmarking Travel Time and Demand Prediction Methods Using Large-scale Metro Smart Card Data", Periodica Polytechnica Transportation Engineering, 51(4), pp. 357–374. https://doi.org/10.3311/PPtr.22252

is in place, operators are also obliged to monitor the level of crowding and keep it within a predetermined threshold – and even if such regulation is not in place, operators might be interested in keeping occupancy rates low to limit the dissatisfaction of customers. Based on our experience from metro benchmarking activities managed by the Transport Strategy Centre (TSC) at Imperial College London, most urban rail operators are currently unable to make network-level, real-time demand predictions and monitor the likely violation of social distancing rules using automated data sources such as SCD.

The core objective of this study is to provide a guide for metro operators and researchers assisting them on readily available demand and travel time prediction methods that can be utilised without substantial investment in software tools or computing capacity. Using a comprehensive smart card dataset from a large Asian metro operator, we test the performance of four demand prediction methods:

- semiparametric regression,
- random forest regression,
- support vector machine regression, and
- multivariate linear regression.

Simple implementations of these quantitative prediction methods are available in *R*, the open-source statistical coding environment. The contribution of this study is benchmarking the performance of these well-known methods for the specific case of demand and travel time prediction in urban rail transport. This way we synthesise isolated attempts to develop prediction methods with SCD and form recommendations for their direct implementation in practical forecasting exercises. One of the main messages of this work is that there is no universal prediction method for public transport-specific decision support; the present paper concludes with specific recommendations for ridership and travel speed estimation.

1.1 The surrounding literature

The empirical literature of transport science is rich in prediction-oriented research, especially since the introduction of large-scale data sources. Comprehensive reviews of data mining methods in the context of public transport operations are available in Ghofrani et al. (2018), Koutsopoulos et al. (2019), and Pelletier et al. (2011). In this paper we focus on short-term prediction models; for long-term demand forecasting, the four-stage approach to traffic modelling (de Dios Ortúzar and Willumsen, 2011) and the use of demand elasticites (Blainey and Preston, 2013) are the commonly accepted methods in transport science.

For short-term prediction, in the past decade there has been a surge of methodological developments in nonparametric artificial intelligence techniques, including neural networks, genetic algorithms, Kálmán filtering, and clustering methods.[1] In general, the literature on machine learning (ML) techniques relies on data on historical passenger flows, travel costs, and temporal factors such as the time of the day. In Section 1.1, we first review the existing ML studies on travel *flow* prediction, and then we turn to travel *time* or *speed* forecasting.

Bai et al. (2017) develop a deep fusion approach to predict short-term bus passenger flow, by fusing deep belief networks corresponding to flow patterns differentiated by behavioural, temporal, and transport management factors. Using bus data from Guangzhou, China, their approach outperforms nonparametric models. A series of papers focus on flow prediction during special events when demand is more volatile than under regular conditions (see Chen et al. (2020); Li et al. (2017); Ni et al. (2017); Pereira et al. (2015); Xue et al. (2022) among others). Guo et al. (2019) propose fusioning support vector regression (SVR) and short-term memory neural networks for the prediction of abnormal passenger flows, in general. Ding et al. (2016) extend the analysis to multimodal public transport provision. Their gradient boosting decision trees approach predicts metro demand in combination with data on transfer behaviour at adjacent bus stops. A more recent incarnation of interpretable tree-based methods is the XGBoost model of Zou et al. (2022).

Among the most refined nonparametric regression methods, artificial neural networks (ANN) appear as an agile and adaptive method for forecasting with enhanced function mapping capabilities (Zhang et al., 1998). However, a comparison study later concludes that the performance of Gaussian maximum likelihood methods is still comparable to ANN (Tang et al., 2003). Ling et al. (2018) analyse predicted passenger flows on the Shenzen subway network under different traffic conditions extrapolating back-propagation in a multilayer perception model and then compare this against the following: SVR model with a linear kernel function, gradient boosted regression tree model training data in parallel, and historical average model. Their study concludes a better prediction performance at relatively stable passenger flows with the SVR model, however, the multi-layer perception model is more reliable with irregular patterns in passenger flows.

1 Beside an abundance of nonparametric methods, deterministic prediction approaches such as the route choice model of Lee et al. (2022) are still active research areas.

Similarly, Sun et al. (2015) propose Wavelet-SVR model which is then compared to both Wavelet-ANN and EMD-BPN models. The Wavelet-SVR model yields better forecasting performance on a data set from the Beijing subway system. Yang et al. (2021b) show the superiority of a long short-term memory network (LSTM) and wavelet over ARIMA, nonlinear regression and traditional LSTM models.

Several studies use ML methods to predict travel times (or, inversely, travel speed) in a transport network.[2] Bin et al. (2006) estimate bus arrival times with bus movement data from Dalian, China. A SVR model with three-variable representing the route, current route travel time and nearest next route travel time was found to be superior to a constructed three-layer standard ANN. Methods from both queuing theory and ML are merged by Gal et al. (2017) to improve the accuracy of travel time predictions, using historical and real-time data from the Dublin bus network. Their results reveal the accuracy gains of fusions between the snapshot method originating from queuing theory and ML methods.

Based on the rapid evolution of ML methods, it may be argued that more elementary statistical and econometric methods have been surpassed.[3] Naturally, the linearity assumption in standard linear regression modelling causes severe limitations in predictive power. Fully parametric statistical models may include nonlinear relationships with a predefined functional form, however in the volatile environment of spatio-temporal demand fluctuations in public transport, it is unlikely that predefined and static functional forms can achieve high performance. The modelling technique of semiparametric regression presents a viable framework in the context of transit prediction applications enabling the modelling of explanatory variables without a pre-specified functional form. Furthermore, the method has been adapted to enable fast analysis of large data sets Ruppert et al. (2009).

2 Note that in this paper our focus is restricted to travel time prediction under regular (undisrupted) conditions. For a comprehensive coverage of disruption detection, the interested reader is referred to Zhang et al. (2022).

3 Autoregressive (integrated) moving average (i.e., ARMA and ARIME) were introduced very early for transport prediction exercises, first to provide short-term forecasts of traffic flows (see, for example, Ahmed and Cook (1979)). To account for periodic variations of traffic states, a seasonal ARIMA (SARIMA) was developed (Williams and Hoel, 2003). This particular model proved to outperform the historical average model and the Kálmán filter model which were introduced by Okutani and Stephanedes (1984) to the urban traffic control system of Minneapolis-St Paul and Nagoya City.

As a summary, Table A1 in Appendix provides an overview of the strengths and weaknesses of the previously discussed families of models used for transport demand prediction in the literature. A key point is that we find the semiparametric regression and ML methods' performance is unknown in the context of public transport applications. The remainder of this paper addresses this particular gap in the literature, thus providing a practical aid for transport operators seeking effective prediction methods for short-term demand and travel time forecasting.

1.2 Structure of the paper

The rest of the paper is organised as follows: Section 3 presents the model frameworks for four types of models that are trialled, namely:

1. semiparametric regression,
2. random forest regression,
3. support vector machine (SVM) regression, and
4. multivariate linear regression.

We give an overview of the dataset used in the analysis in Section 2 and present the results in Section 4. Finally, Section 5 concludes with a summary of our main findings, and provides suggestions for future research directions.

2 Data

The data utilised in this paper are obtained from a major metro system in Asia. This metro system operates in one of the cities with the highest rates of public transport usage. In the analysis, SCD that tracks individual trips made on the metro network from 2013 and 2014 are used. Over this period, 11 lines and 87 stations were in operation. The SCD include locations and timestamps of the origin and destination taps for each trip a passenger has performed, but transfers are not recorded. Two time periods are selected for analysis: one is 3 months beginning July 2013, and the other is 3 months beginning July 2014. All train journeys from 2013 are used in training the models, while all train journeys in 2014 are used to test the models.

The timestamps of smart card records are accurate to one second. The spatial and other socio-economic dimensions are indicated by time-invariant properties, including:

- anonymous card ID,
- entry and exit station codes,
- transaction type (entry/exit),
- passenger type,
- train direction, and
- fare prices, with or without discounts.

The data sets have been converted from a disaggregate individual level to OD-based aggregates by averaging for 15-minute intervals. Table 1 provides descriptive statistics for the dataset we use for prediction. The mean travel time (tap-in/tap-out) is 34.4 minutes, and its distribution is surprisingly symmetric between the 1st and 3rd quartiles. The mean travel demand within 15-minute is 1448 passengers per OD pair. This variable is right-skewed, meaning that a significant proportion of total ridership is concentrated in a limited number of spatio-temporal markets. This observation is consistent with regular demand patterns in metro systems. The distribution of the fares paid is also right-skewed. The proportion of adult ticket holders is 74.9% in the average 15-minute block, but their share takes very low or very high values at the extrema of the distribution.

3 Methods

Sections 3.1 to 3.3 present the methods used in the development and comparison of the models. First, the theoretical frameworks for each model are presented. Second, we outline the model development process involving data cleaning, explanatory variable selection, and data selection for training, validation, and testing. Third, we present the model evaluation criteria.

All calculations have been performed using *R* statistical analysis software on an Intel Core i7 CPU at 2.8 GHz, 16 Gb of RAM. Details of the specific *R* packages used are given in the discussion of the model frameworks.

3.1 Model frameworks

In the following presentation of model frameworks, the response variable is denoted as y_{ij}. Since two types of analyses will be undertaken, the response y_{ij} refers to

1. passenger flows per time period *i* along origin-destination route *j* and
2. mean journey times per time period *i* along origin-destination route *j*.

The time periods *i* are defined in 15-minute intervals. The explanatory variables are denoted by x_{ij}; further details on the definition of these is given in Section 3.2.2.

Table 1 Descriptive statistics

Key variables	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Travel time (s)	179	1257	1902	2063	2710	9971
Travel demand	128	849	1341	1448	2027	7400
Price	0	5.0	7.7	9.3	10.9	84.6
Proportion of adults	0	0.635	0.818	0.749	1	1

3.1.1 Semiparametric regression

Semiparametric regression is a statistical regression method that enables the modelling of flexible non-linear relationships between dependent and independent variables without the need to define the relationships parametrically a priori. The flexible relationships are generated via basis functions that take the form of thin plate regression splines. The splines are interpolation functions that fit the data points, with a "wiggliness" penalty that represents the trade-off between alignment with data points and smoothness. The regression models adhere to a generalised additive mixed model (GAMM) structure, as the spline functions possess a structure that represent the sum of a linear predictor and random effects. Model fitting is based on the penalised iteratively reweighted least squares (PIRLS) fitting technique, and restricted maximum likelihood (REML) optimisation is used to estimate the model parameters. Further information on the underlying theory of semiparametric modelling can be found in Wood (2017) and Wood et al. (2015).

In the context of our models, the general equation is Eq. (1):

$$g\left(y_{ij}\right)=\alpha+\sum_{k=1}^{K}f_k\left(x_{np,ij}\right)+\sum_{m=1}^{M}\beta_m x_{p,ij}+\epsilon_{ij}, \tag{1}$$

where $g(\cdot)$ is a link function, α is the model constant, f_k are the smooth basis functions based on penalised thin plate regression splines of the independent variables $x_{np,ij}$ modelled non-parametrically, β_m are the coefficient values for the independent variables $x_{p,ij}$ modelled parametrically with a linear form, and ϵ_{ij} is the model error term such that $\epsilon_{ij}\sim\mathcal{N}\left(0,\sigma_\epsilon^2\right)$.

In generating the final model form, we investigate different specifications for the underlying family distribution, the link functions, the number of knots (or turning points) for the regression splines, and we also trial different combinations of linear and spline forms for the independent variables (refer to Section 4.1 for further details). Once the optimal values have been determined using the training and validation data, we use these to generate predictions for observations of x_{ij} in the test data set. The "bam" package in *R* (Wood et al., 2015) is used to perform the semiparametric regression modelling.

3.1.2 Random forest

Random forest is a ML method based on the regression tree algorithm. This algorithm involves partitioning the original data set based on the values of the independent variables. Partitions are undertaken such that the sum of squared

errors is minimised, and the number of partition levels is dependent on a trade-off between fitting and over-fitting the data. The final form of the regression tree is attained when the smallest feasible final partition regions are generated; these regions are termed leaves. The predicted value of the dependent variable in each leaf is a constant calculated as the mean value of all observations in the leaf.

The random forest algorithm involves taking bootstrap samples of the data and fitting regression trees to each sample. For continuous forms of the dependent variable, as is the case in our analysis, the predicted value of the dependent variable is calculated by taking the mean of the predicted values generated by the bootstrap sample of regression trees. The algorithm can be summarised as follows (Breiman, 2001; Hastie et al., 2017):

- Take a bootstrap sample of size N_b from the training data;
- Generate a random forest tree T_b from the bootstrap sample as follows:
  1. For each terminal node, randomly select m independent variables from the total number of independent variables p;
  2. Determine the optimal variable/split point;
  3. Spilt the node into two further nodes;
  4. Repeat steps 1. to 3. at each terminal node until the minimum number of nodes for the tree is reached.
- Repeat 1st main step "Take a bootstrap …" and 2nd main step "Generate a random …" for all bootstrap samples $b = 1…B$.

The random forest prediction for a new point x_{ij} in the test data set is then generated as follows:

$$\hat{f}_{RF}^{B}\left(x_{ij}\right)=\frac{1}{B}\sum_{b=1}^{B}T_b\left(x_{ij}\right). \tag{2}$$

According to Hastie et al. (2017), a default value for m should be $p/3$ where p is the number of independent variables, and the minimum node size which sets the depth of the regression tree should be 5. We trial different values of m and n_{tree} to arrive at a final model form; further details are given in Section 4.1. The "randomForest" (Liaw and Wiener, 2002) and "ranger" (Wright and Ziegler, 2017) packages are used to run the algorithms.

3.1.3 Support vector machines

Support vector machines are ML algorithms for classification or regression applications. Since the dependent variables are continuous in our analysis, we adopt the methodology for support vector machine regression. In simple linear regression, the objective is to minimise error in the model, i.e. to minimise the sum of the squared errors between the fitted and observed data. In support vector machine regression, the objective is to minimise the coefficient vector and the error term is specified to have an allowable error margin. To further improve the model fit, slack variables can be incorporated; these are defined as values that fall outside the error margin, which are to be minimised. The general form of the support vector machine regression optimisation problem is summarised as follows (Drucker et al., 1996):

$$\begin{aligned}&\min_{w,b,\xi,\xi^*} 0.5w^Tw+C\sum_{ij=1}^{N}\xi_{ij}+C\sum_{ij=1}^{N}\xi_{ij}^*\\ &\text{subject to } w^T\phi\left(x_{ij}\right)+b-y_{ij}\le\epsilon+\xi_{ij},\\ &\qquad y_{ij}-w^T\phi\left(x_{ij}\right)-b\le\epsilon+\xi_{ij}^*,\\ &\qquad \xi_{ij},\xi_{ij}^*\ge 0,\ ij=1,\ldots,N,\end{aligned} \tag{3}$$

where $\phi(\cdot)$ is a kernel function that maps the independent variables into a higher dimensional space for separation (if necessary), w is a vector of parameters for the independent variables to be minimised, b is the intercept, ξ_{ij} and ξ_{ij}^* are slack variables, ϵ is the allowable error margin such that $\epsilon > 0$, and C is a tuning or "cost" parameter for the slack variables such that $C > 0$. To generate the final model form, we investigate different kernel function types, and different values of ϵ and C. Further details are given in Section 4.1. The optimal regression parameters w and b generated from the final form are then used to calculate predictions for observations of x_{ij} in the test data set. The support vector machine regression algorithms are implemented using the "e1071" package in *R*.

3.1.4 Multivariate linear regression

As a baseline model type for comparison, we also generate linear regression models using ordinary least squares fitting. Under this framework, the relationship between dependent and independent variables is linear. The general equation is Eq. (4), as follows:

$$y_{ij}=\alpha+\sum_{k=1}^{K}\beta_k x_{ij}+\epsilon_{ij}, \tag{4}$$

where α is the model constant, β_k are the coefficients for $k = 1…K$ dependent variables modelled linearly, and ϵ_{ij} is the random error term such that $\epsilon_{ij}\sim\mathcal{N}\left(0,\sigma_\epsilon^2\right)$. The linear regression models are generated using the "base" package in *R*.

3.2 Model development

3.2.1 Data cleaning

A summary of the data cleaning process prior to developing the prediction models is outlined in Fig. 1. Observations were removed in cases of duplicate entries, erroneous records, outliers, and those with missing values in more than 60% of the following categories: data on temporal information, cost of trip, type of user, entry and exit stations, erroneous records comprised those with 0 or negative travel times.

For observations with less than 60% missing information and those with travel times greater than 10 times the average travel time but less than 10 hours, the missing values in the independent and/or dependent variables were calculated via mean imputation. As a result of the data cleaning process, approximately 1.2% of the observations were removed, and an approximately equivalent number of observations were restored via imputation.

It should be noted that feature engineering was also performed to standardise the variables that possessed large numeric ranges and mean values, however, this yielded minimal improvements, and so the original variable values were retained.

3.2.2 Independent variables

We included a number of independent variables to improve the predictive power of the models. The selection of the variables was based on the literature on passenger demand and travel time prediction and data availability. A summary of the variables is given in Table 2. Initially all variables were included in the models. Further refinement of variable selection was undertaken based on variable statistical significance for the linear and semiparametric models, and importance scores for the machine learning models. Further details are given in Section 4.1.

3.2.3 Training, validation, and test sets

The cleaned data were split into training, validation, and test data sets. The percentage of data in each set was 80%, 10%, and 10%, respectively. This framework of data partitioning is widely accepted in the literature on predictive algorithms, for example, Wang et al. (2018). A summary of the counts in each data set is given in Fig. 2.

3.3 Model evaluation criteria

We use the conventional goodness-of-fit metrics used in machine learning applications to evaluate the predictive performance of the models including the root mean square

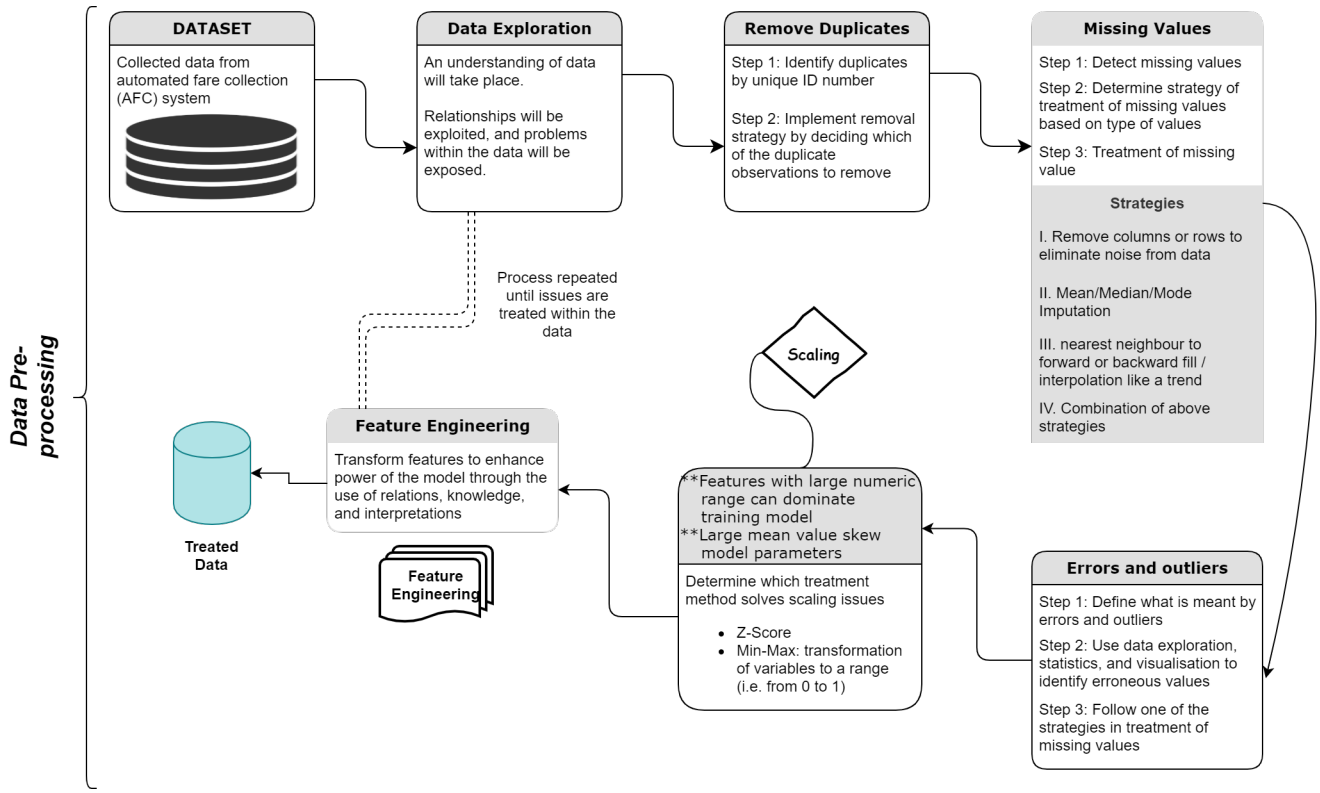

Fig. 1 Flow chart describing strategy implemented throughout the pre-processing of the data

Table 2 Explanatory variables accounted for in travel demand/flow and time prediction models

Variables	Symbol	Description	Type
Entry and exit stations	entst, extst	Species the stations where each individual user entered and then exited.	Categorical
Time intervals (15 min)	entrytime_ 15_double	Specifies the period at which trips between different O-D pairs occurred. Each day is divided into 96 periods of which each period constitutes of 15 minutes.	Categorical
Ticket price	price mean, price median	The average and median price of tickets are included as variables	Numerical
Travel time	traveltime adjusted mean, traveltime adjusted median	The average and median travel times	Numerical
Type of user	per ADL, per CHD, per DIS, per SEN, per STD, per TRS	Users are grouped into adults, child, disabled, seniors, students, and retired persons	Percentage
EBD	EBD	Introduced in 2014 which offers a reduced price for early purchasers	Categorical
Peak time periods	per peak period	Peak periods in weekdays are distributed over the periods from 7:00 AM to 8:30 AM and 5:30 PM to 7:00 PM	Percentage
Day of week	week day	Species the day of the week (i.e. weekday, weekend)	Categorical
Historical observations	ntickets H1	This variable accounted for the previously observed travel flow in the past 15 min, past day, and past week at the same period of trips made	Numerical
Meteorological data	Temperature Weather condition Wind Humidity	Temperature Weather condition: sunny, rainy, storm, etc Wind Humidity	Numerical Categorical Numerical Numerical
Additional explanatory variables accounted for in the travel time prediction model			
No. of tickets	Ntickets	The number of users at each period travelling between each OD-pair (passenger flows or travel demand)	Numerical
Historical observations	Traveltime adjusted H1	This variable accounted for including the previous observed travel time between different OD pairs in the past 15 min, past day, and past week at the same period of trips made	Numerical

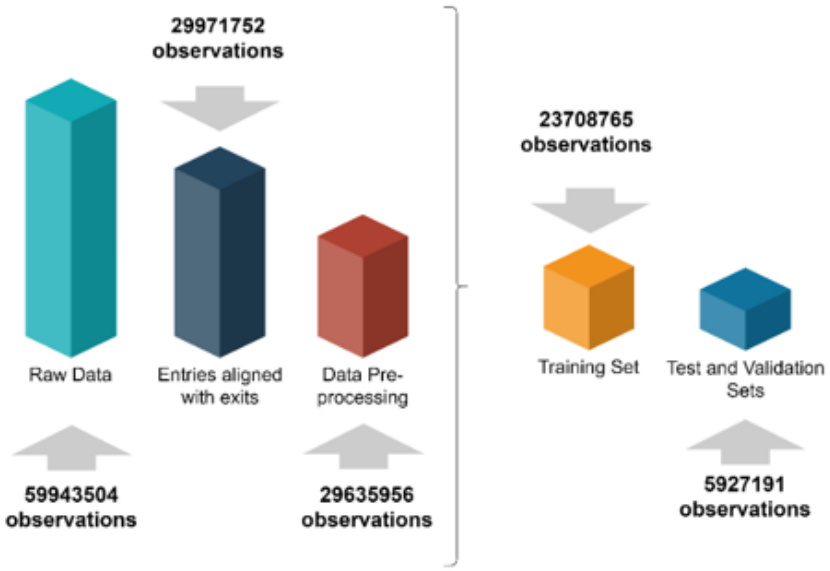

Fig. 2 Training, validation, and test sets

error (RMSE), the mean absolute percentage error (MAPE), the coefficient of determination (R^2), and likelihood-based criteria including the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Models with lower values of RMSE, MAPE, AIC, and BIC, and higher values of R^2 indicate a better fit. We also adopt two additional extensions of the RMSE and MAPE metrics termed the symmetric MAPE (SMAPE) and the normalised RMSE (NRMSE) to minimise the impact of outliers and generate more comparable evaluations (Li et al., 2018);

$$\text{SMAPE} = \frac{1}{n}\sum_{ij=1}^{n}\frac{\left|y_{ij}-\hat{y}_{ij}\right|}{0.5\left(y_{ij}+\hat{y}_{ij}\right)}\times 100, \quad (5)$$

and

$$\text{NRMSE} = \left[\sum_{ij}\left(y_{ij}-\hat{y}\right)^2/n\right]^{0.5}\left(\sum_{ij}y_{ij}/n\right)^{-1}, \quad (6)$$

where y_{ij} is an observation of the dependent variable, $\hat{y}_{ij}$ is the predicted value of the observation, and n is the total number of observations. Each metric represents different aspects of model fit, and thus evaluating a range of metrics provides a better assessment of overall model performance.

4 Results

In the following presentation of the model results, we refer to each model by acronyms as follows: MVLR - multivariate linear regression, SPR - semiparametric regression, RFR - random forest regression, and SVMR - support vector machine regression. We additionally refer to the SPR, RFR, and SVMR models as the proposed models and the MVLR as the baseline model.

4.1 Final model forms

4.1.1 MVLR baseline models

Despite being the simplest model form, the baseline MVLR models performed well with adjusted R^2 values of > 0.95. It should be noted that the performance of the models is greatly improved through the inclusion of lagged temporal variables, namely travel flows and travel times in the past day, month, and year in the travel demand and travel time models, respectively. The results for the final model forms of the baseline MVLR models are given in Tables 3 and 4.

4.1.2 SPR models

The development of the final form for the SPR models involved trialling different family distributions, link functions, and knot values for the regression splines, and trialling different combinations of linear and non-parametric

Table 3 Results for linear regression model of travel demand

Covariate	Coefficient value	Std error	*P*-value	Sig
entrytime_15_double	1.11E-01	4.78E-03	<2E-16	***
entst (dummy)	1.0E0 -- 5.0E1	Varies	<2E-16	***
exst (dummy)	1.0E0 -- 2.5E1	Varies	<2E-16	***
price_mean	−1.99E-01	4.02E-02	7.30E-07	***
price_median	2.86E-02	3.21E-02	3.70E-01	
traveltime_adjusted_mean	−2.26E-03	4.39E-04	2.60E-07	***
traveltime_adjusted_median	3.50E-04	4.38E-04	4.20E-01	
per_ADL	2.57E+00	6.92E-01	2.10E-04	***
per_CHD	1.09E+01	1.09E+00	<2E-16	***
per_DIS	1.31E+00	1.19E+00	2.70E-01	
per_SEN	1.29E+00	7.33E-01	7.70E-02	.
per_STD	8.63E-01	7.11E-01	2.20E-01	
per_TRS	1.29E+00	5.35E+00	8.10E-01	
EBD	8.24E+00	2.83E-01	<2E-16	***
per_peak_period	6.89E+00	1.72E-01	<2E-16	***
week-day1	2.50E+01	4.11E+00	<2E-16	***
week-day2	−3.31E+01	1.11E+00	<2E-16	***
ntickets_H1 (past 15-min)	9.74E-01	3.82E-04	<2E-16	***
temperature	7.93E-01	9.24E-02	<2E-16	***
weather_condition1	−1.44E+01	2.46E-01	<2E-16	***
weather_condition2	−1.61E+01	4.19E-01	<2E-16	***
weather_condition3	−2.22E+01	5.15E-01	<2E-16	***
wind	2.67E-01	2.84E-02	<2E-16	***
humidity	5.78E-01	2.58E-02	<2E-16	***
Adj. *R*-squared	0.9557			

Significance notation: *p*-values 0, "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1, " " 1.

Table 4 Results for linear regression model of travel time

Covariate	Coefficient value	Std error	*P*-value	Sig
entrytime_15_double	3.73E-01	1.87E-02	<2E-16	***
entst (dummy)	−1.5E1 -- 2.0E1	Varies	<2E-16	***
exst (dummy)	−1.5E1 -- 2.0E1	Varies	<2E-16	***
price_mean	−1.65E+00	1.58E-01	<2E-16	***
price_median	1.98E+00	1.25E-01	7.00E-01	
traveltime_adjusted_mean	9.54E-01	5.32E-04	<2E-16	***
traveltime_adjusted_median	3.26E-02	5.25E-04	5.50E-01	
per_ADL	−7.29E+01	2.70E+00	<2E-16	***
per_CHD	−8.08E+01	4.24E+00	<2E-16	***
per_DIS	−6.73E+01	4.66E+00	<2E-16	***
per_SEN	−6.95E+01	2.87E+00	<2E-16	***
per_STD	−5.73E+01	2.78E+00	<2E-16	***
per_TRS	−5.48E+01	2.09E+01	8.80E-03	**
EBD	−1.01E+01	1.11E+00	<2E-16	***
per_peak_period	1.66E+01	6.71E-01	<2E-16	***
week-day1	3.08E+01	4.11E+00	<2E-16	***
week-day2	4.56E+01	6.11E+00	<2E-16	***
ntickets	−5.88E-02	1.50E-03	<2E-16	***
ntickets_H1 (past 15-min)	3.26E-02	5.25E-04	<2E-16	***
temperature	7.36E+00	3.61E-01	<2E-16	***
weather_condition1	−1.48E+01	9.62E-01	<2E-16	***
weather_condition2	−1.58E+01	1.64E+00	<2E-16	***
weather_condition3	−3.99E+01	2.02E+00	<2E-16	***
wind	−5.24E-01	1.11E-01	2.40E-06	***
humidity	1.13E+00	1.01E-01	<2E-16	***
Adj. *R*-squared	0.9782			

Significance notation: *p*-values 0, "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1, " " 1.

spline forms for the independent variables. The results of the trials are given in Table 5. As shown in the table, the Poisson family distribution with a log function link generates the best performing model for travel demand, and the Gamma family distribution with a log function link generates the best performing model for travel times.

The results for the final model forms of the SPR models is given in Tables 6 and 7. The variables that are significant in the SPR model of travel demand are also significant in the corresponding baseline MVLR model. For the SPR model of travel times, there are fewer significant variables than the corresponding baseline MVLR model;

Table 5 Family distributions, model fit, and residual scores

Family	Link	Travel flows		Travel times	
		R-sq. (adj.)	Residuals	*R*-sq. (adj.)	Residuals
Gaussian	Identity	0.957	2.00	0.988	2.20
Gamma	Log	0.91	6.00	0.99	2.50
Poisson	Log	0.959	4.00	0.987	2.30
Quasi	Identity, const. var.	0.957	2.00	0.989	2.25
Quasipoisson	Log	0.958	4.00	0.987	2.30
Adj. *R*-squared	0.9782				

Significance notation: *p*-values 0, "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1, " " 1.

Table 6 Results for semiparametric regression model of travel demand

Covariate	Form	No. of knots	Coefficient value	Std. error	*P*-value	Sig
entrytime_15_double	Linear	-	1.15E-01	4.80E-03	<2E-16	***
entst	Dummy	1.0E0 -- 6.0E1	Varies		<2E-16	***
exst	Dummy	1.0E0 -- 3.5E1	Varies		<2E-16	***
price_mean	Splines	3	-	-	9.00E-02	.
price_median	Splines	3	-	-	<2E-16	***
traveltime_adjusted_mean	Splines	9	-	-	6.00E-09	***
traveltime_adjusted_median	Splines	9	-	-	0.009	**
per_ADL	Splines	3	-	-	6.60E-09	***
per_CHD	Splines	3	-	-	8.50E-02	.
per_DIS	Splines	3	-	-	3.70E-01	
per_SEN	Splines	3	-	-	7.60E-09	***
per_STD	Splines	3	-	-	1.30E-01	
per_TRS	Splines	3	-	-	1.50E-02	*
EBD	Linear	-	7.63E+00	2.87E-01	<2E-16	***
per_peak_period	Splines	4	-	-	<2E-16	***
week-day	Splines	4	-	-	<2E-16	***
ntickets_H1 (past 15-min)	Splines	4	-	-	<2E-16	***
temperature	Splines	3	-	-	1.50E-01	
weather_condition	Linear	-	−1.51E+01	4.17E-01	<2E-16	***
wind	Splines	3	-	-	1.20E-01	
humidity	Splines	3	-	-	5.50E-02	*
Adj. *R*-squared	0.961					

Significance notation: *p*-values 0, "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1, " " 1.

namely, the variables capturing the user types of adults, child, disabled, seniors, students, and retired persons drop in significance.

4.1.3 RFR models

We first undertook trials to establish whether the models should be estimated using the "ranger" or "randomForest" package. As per Table 8, the "ranger" package outperformed the "randomForest" package in terms of processing time. Furthermore, we faced issues with the "randomForest" package in handling categorical predictors with more than 53 categories. Therefore, we adopted the "ranger" package for subsequent modelling. In the development of the final RFR model form, we investigated different combinations of the number of trees n and the number of independent variables to be selected in the generation of trees (m). The results of the trials are given in Fig. 3 (a)–(d). The optimal values of the parameters are as follows: $n = 250$ and $m = 10$.

In RFR model frameworks, the improvement in RMSE each time an independent variable is used as a node split is able to be quantified. This quantity is converted to an impurity score and reported as the variable's "importance". The top five variables with the highest importance in the RF flows prediction model are *nticketsH1*, *type of user*, *travel time*, *price* and *entry station*. For the RF travel times model, the top five variables are *nticketsH1*, *type of user*, *travel time*, *price* and *entry station*.

4.1.4 SVMR models

In the generation of the SVMR models, different values of the following attributes were trialled: kernel types, the error margin ϵ, and the cost parameter C. The kernel types trialled included linear, polynomial, radial basis, and sigmoid kernels. The optimal form was found to be radial.

To determine the optimal ϵ and C values, we undertook a grid search with 10-fold cross-validation sampling, with the objective of minimising the RMSE scores. The darker coloured areas of the grid in Fig. 4 show the optimal values. For the travel demand model, the optimal values are $\epsilon = 0.26$ and $C = 8$. For the travel times model, the optimal values are $\epsilon = 0.2$ and $C = 80$. It is worth noting that if the default values of $\epsilon = 0.1$ and $C = 1$ were used, the goodness-of-fit performance of the SVMR model would fall far below the SPR and RFR models.

Table 7 Results for semiparametric regression model of travel time

Covariate	Form	No. of knots	Coefficient value	Std. error	*P*-value	Sig
entrytime_15_double	Linear	-	2.67E-01	1.86E-02	<2E-16	***
entst	Dummy	−5.0E0 -- 2.5E1	Varies		<2E-16	***
exst	Dummy	0.0E0 -- 2.0E1	Varies		<2E-16	***
price_mean	Splines	3	-	-	9.00E-02	.
price_median	Splines	3	-	-	<2E-16	***
traveltime_adjusted_median	Splines	9	-	-	9.30E-03	**
per_ADL	Splines	3	-	-	2.60E-01	
per_CHD	Splines	3	-	-	8.60E-01	
per_DIS	Splines	3	-	-	5.50E-01	
per_SEN	Splines	3	-	-	3.10E-01	
per_STD	Splines	3	-	-	1.70E-01	
per_TRS	Splines	3	-	-	4.60E-01	
EBD	Linear	-	−2.02E+01	1.11E+00	<2E-16	***
per_peak_period	Splines	4	-	-	<2E-16	***
week-day	Splines	4	-	-	<2E-16	***
ntickets	Splines	3	-	-	<2E-16	***
ntickets_H1 (past 15-min)	Splines	4	-	-	<2E-16	***
temperature	Splines	3	-	-	5.90E-01	
weather_condition	Linear	-	−1.08E+01	1.62E+00	2.70E-11	***
wind	Splines	3	-	-	5.90E-01	
humidity	Splines	3	-	-	9.50E-02	.
Adj. *R*-squared	0.99					

Significance notation: *p*-values 0, "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1, " " 1.

Table 8 Processing time (s) for the "randomForest" and "ranger" packages

	"randomForest"	"ranger"
User	10596.325	221.838
System	396.215	8.506
Elapsed	10658.365	108.293

4.2 Comparison of performance

Table 9 summarises goodness-of-fit and computational performance when comparing the final model forms. For both travel demand and travel time models, the training and prediction computational time for the baseline MVLR models is shortest compared to other forms, however, the SPR, RFR, and SVRM models outperform the baseline MVLR model in terms of predictive power, as anticipated.

For the travel demand models, the SPR, RFR, and SVRM models have R^2 values that range 0.55–2.43% higher than the baseline MVLR model; the SVMR model has the highest score and SPRM the lowest. For the RMSE metric, the proposed models' error ranges between 1–2 times less than the baseline MVLR model, with the SVMR again performing best and SPRM the worst. For the NRMSE performance metric, the improvement in error is similar to that of the RMSE metric, however, the SPR model outperforms the others. The MAPE and SMAPE performance metrics reveal an error reduction of 2–3.5 times when compared to the baseline MVLR model, and the SVMR model performs best under these metrics. In terms of computation time for the proposed models, the SPR model is fastest overall compared to the RFR and SVMR machine learning models.

For the travel time models, the R^2 of the proposed models ranges from 1.08–1.19% higher than the baseline MVLR model, with the SPR model demonstrating the highest R^2 and SVMR and RFR the lowest. For the RMSE and NRMSE, the error reduction in the proposed models range between 2.5–9 times, with the RFR model performing best. The MAPE and SMAPE scores show improvements to a similar degree as the RMSE and NRMSE scores. In terms of computational time, the SVMR and RF models are faster than the SPR model.

Overall, in terms of prediction accuracy, the results show that the semiparametric and machine learning methods outperform the baseline MVLR models. The SVMR

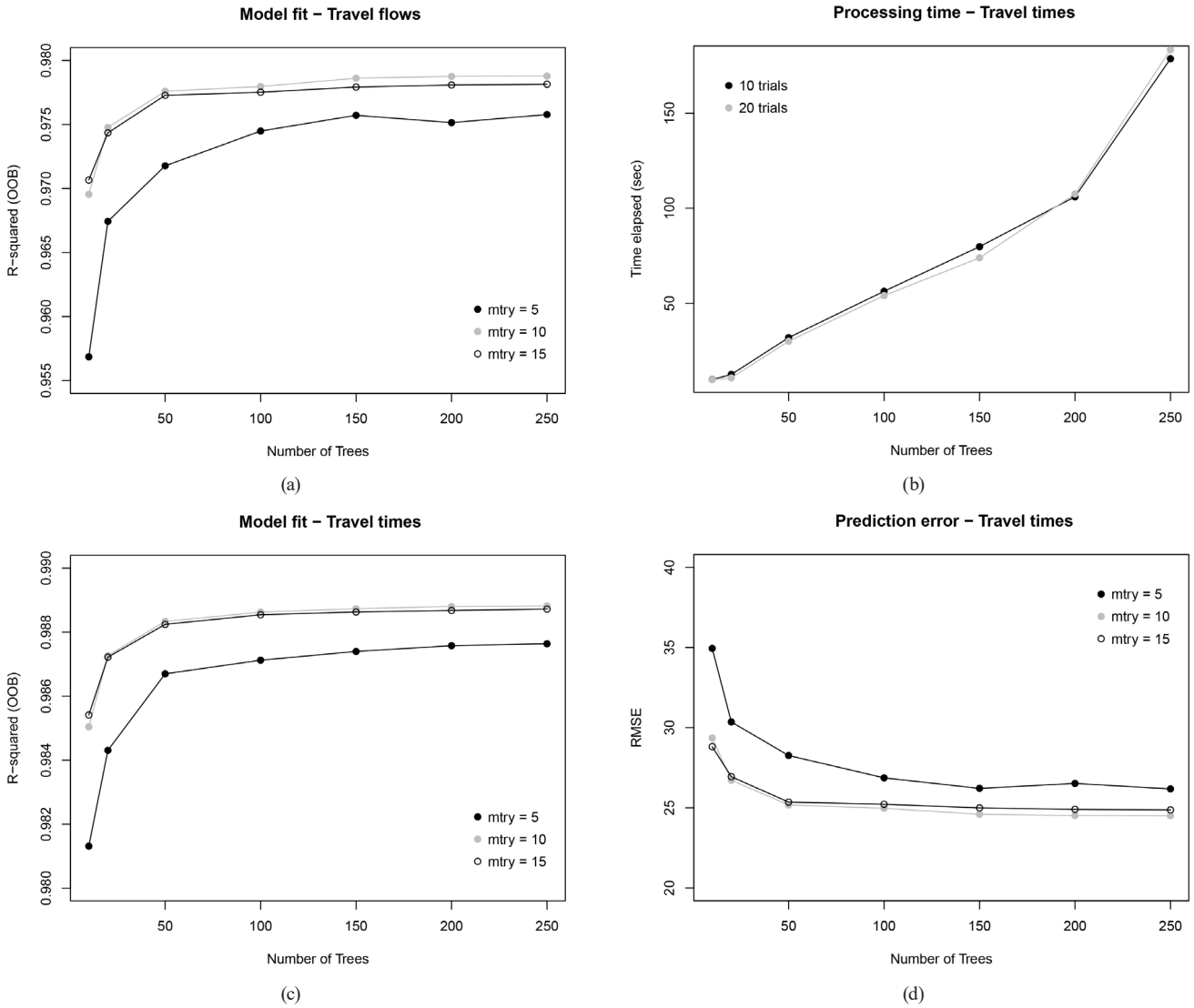

Fig. 3 Effect of the number of trees on (a) model fit – travel flows, (b) processing time – travel times, (c) model fit – travel times and (d) prediction error – travel times

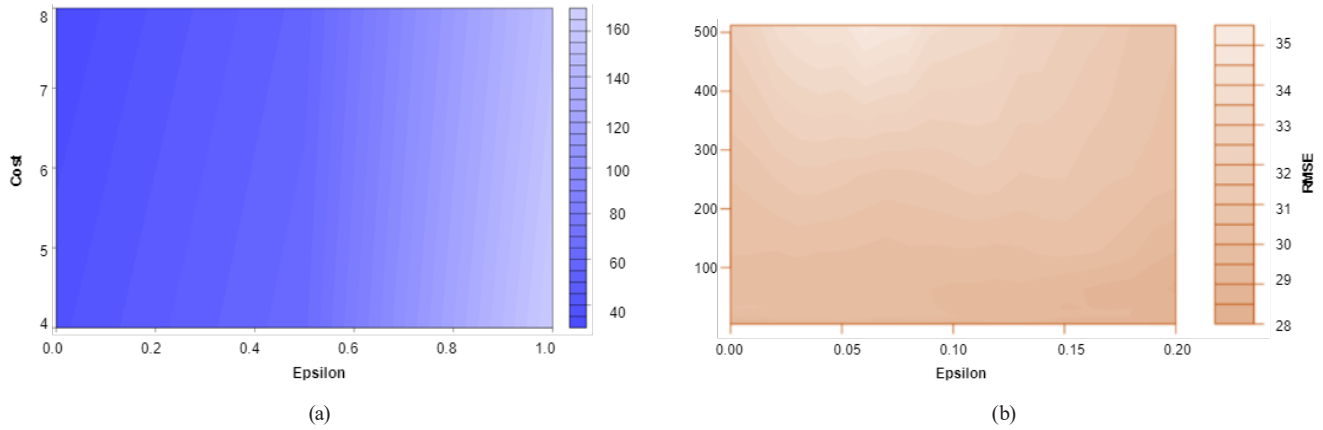

Fig. 4 Hyperparameter optimisation grid plot: (a) travel flow and (b) travel time

model yields the most accurate predictions of travel flows, and the RFR model performs best for travel time predictions. It should be noted that the SPR models demonstrate comparable performance to the machine learning models in terms of prediction accuracy. Moreover, the SPR models generally possess faster computation times than the machine learning models.

Table 9 Performance comparison of travel demand and travel time models

	Travel demand model				Travel time model			
	MVLR	SPR	RFR	SVMR	MVLR	SPR	RFR	SVMR
R-squared	0.9557	0.961	0.97867	0.97948	0.9782	0.99	0.98883	0.98895
RMSE (passengers/seconds)	620.9099	324.884	206.4752	196.65759	103.7764	28.89861	12.01055	15.21659
NRMSE (%)	0.4407	0.16161	0.23447	0.21816	0.33854	0.0815	0.03409	0.0654
MAPE (%)	52.019	18.914	12.715	11.348	12.262	2.668	1.212	1.36
SMAPE (%)	42.821	16.645	10.046	9.977	12.812	2.757	1.203	1.264
Training time (s)	11.44	146.02	183.36	461.15	11.29	139.75	178.64	463.17
Prediction time (s)	1.1	54	11.98	55.04	0.58	51.52	11.19	68.75
Capacity requirement (Mb)	608.63	81.3	1184.5	190.38	603.24	78.71	950.13	211.91

4.3 Trip analysis

As a further extension to the analysis, we undertake a case study to compare the performance of the models for two selected routes differing in length. We analyse a short route, from station (#1) to station (#3), and a longer route, from station (#1) to station (#33). The length of the short route is 3.4 km with 1 intermediate stop, while the longer route is 7.4 km with 7 intermediate stops. The average peak/off-peak passenger flows of the short route are 2400/1280 passengers per 15 minutes, and 785/375 passengers per 15 minutes for the longer route. The average peak/off-peak travel times of the short route are 11/12 minutes, and 20/21.5 minutes for the longer route.

The predictions plotted against the actual travel flows and times for each route are presented in Figs. 5–8. As shown in Figs. 5–8, the models are effective for both route lengths, however, the models possess better performance for the longer route. This could be attributed to a lower degree of variance in travel times and travel flows for the longer route compared to the shorter route. A potential future area of research is to extend this case study to include additional routes to ascertain whether there are differences in prediction accuracy between shorter and longer routes in general. Besides, we observe that the performance of the travel time prediction models is varying at different times of the day and this can also be revisited in future research.

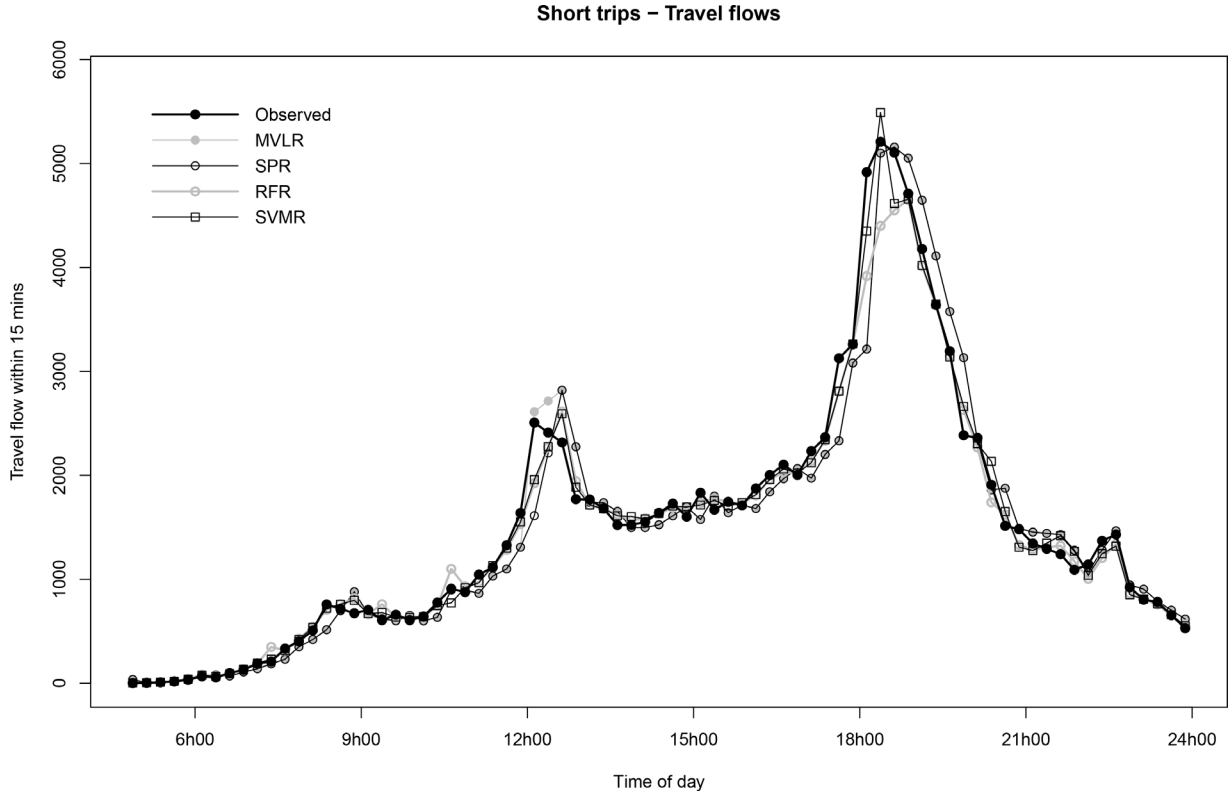

Fig. 5 Actual vs. predicted travel flows of various regression methods between stations #1 and #3

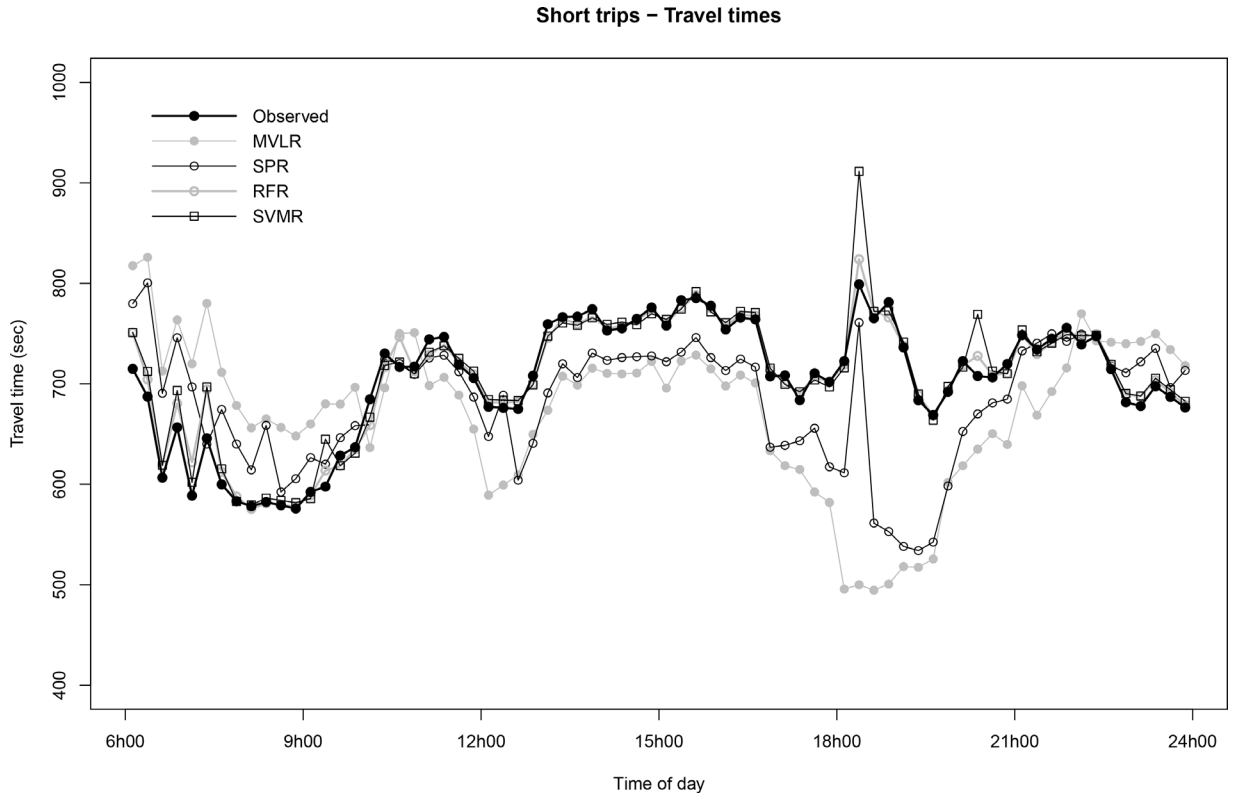

Fig. 6 Actual vs. predicted travel times of various regression methods between stations #1 and #3

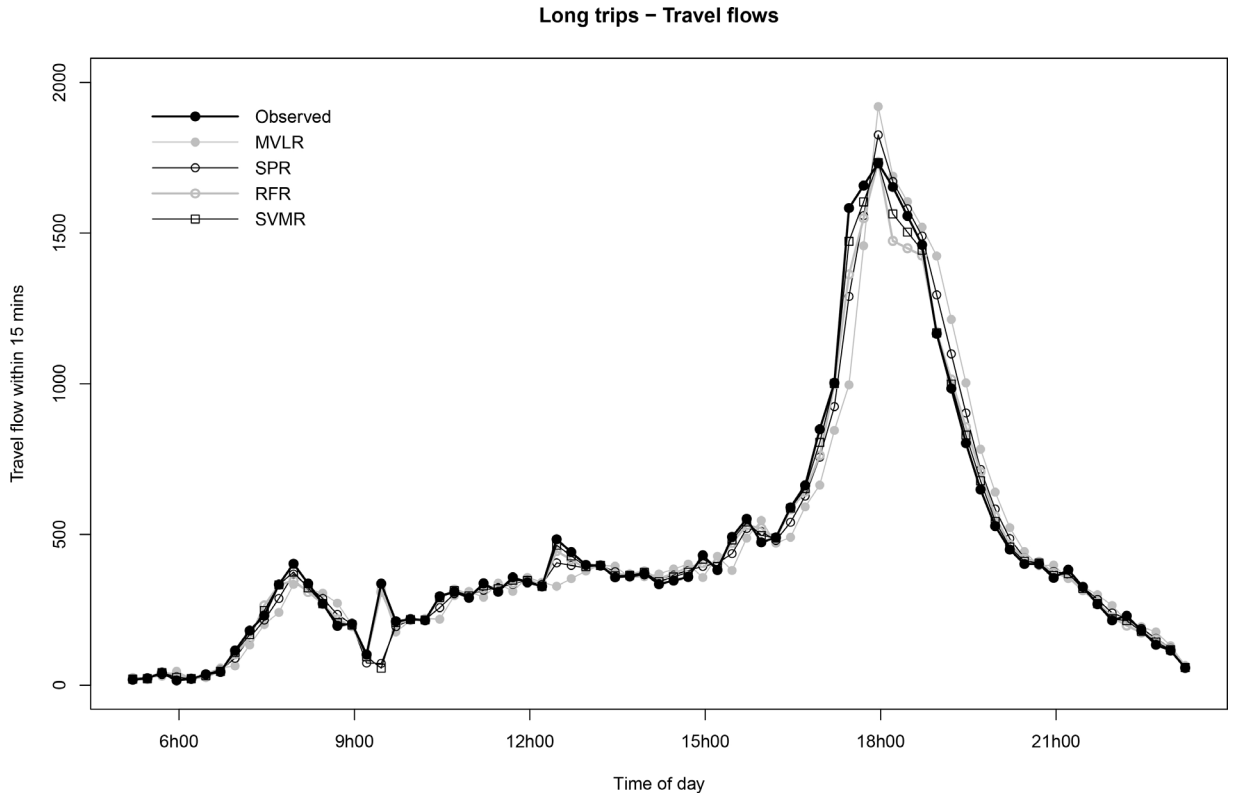

Fig. 7 Actual vs. predicted travel flows of various regression methods between stations #1 and #33

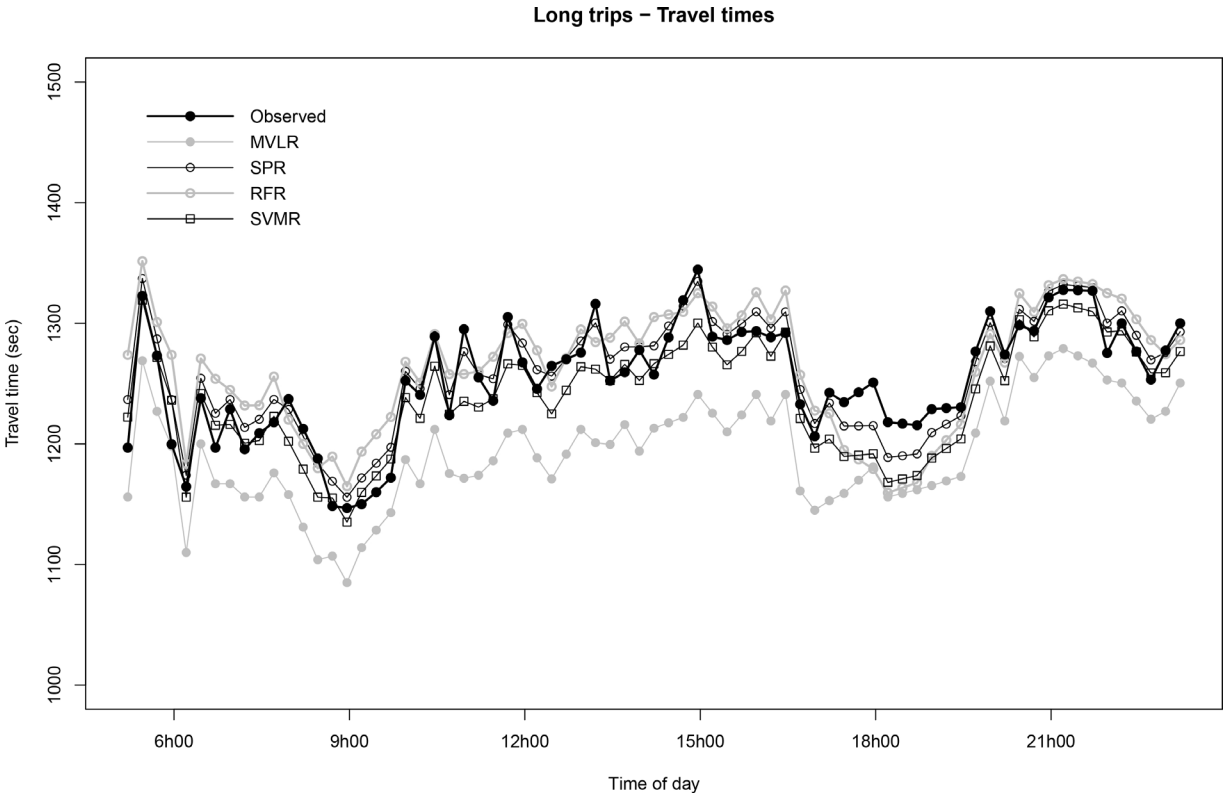

Fig. 8 Actual vs. predicted travel times of various regression methods between stations #1 and #33

5 Conclusions

The core aim of this paper is to document a benchmarking exercise of easily accessible econometric and machine learning methods and compare their performance in travel demand and travel time prediction in an urban rail system. Our analysis is motivated by the observation that smart card data and other automated data sources have become available for most public transport operators around the globe. However, in the absence of local expertise and/or published case studies in the academic sphere, most of them have to make considerable effort to turn these data sources into relevant sources of information. This paper delivers guidance in terms of what methods are available in open-source software packages and how their performance can be ranked in travel demand and travel time prediction.

The paper reports that among semiparametric regression (SPR), random forest regression (RFR), support vector regression (SVMR), and multivariate linear regression (MVLR), the SVMR and RFR models are the most accurate in travel flow and travel time prediction, respectively. Thus, these are the preferred models if accuracy is the most important criteria. We note that the SPR model demonstrates comparable performance in travel time prediction, and this method showcased shorter computation time, which might be a relevant factor when the predictions are used to support real-time operational decisions. This illustrates that none of the four methods can be nominated as a dominant method in the context of public transport analysis.

Numerous operational strategies can be built on accurate real-time ridership and travel time predictions, with the aim of improving the overall efficiency of public transport provision. Amid unexpected demand shocks or headway deviations, real-time dispatching protocols can be designed to control the instantaneous speed of trains, shorten or increase the dwell times at stations, and designing entirely new service temporary timetables to react to fluctuating conditions. These tools have great potential to enhance performance within the network, thus reducing costs for both travellers and the public purse.

Our models can be improved by fusing socio-demographic data from census surveys and trip distances, to capture the foreseeable impact of social and economic trends on travel habits. For future work, the methodology utilised in this paper can be extended to account for crowding effects and traffic incidents within the network, which are expected to be important determinants of both travel demand and speed. Finally, other machine learning models could be added to our pool of benchmarked methods, for example deep learning (see Yang et al. (2021a), for a recent example) and ANN with backpropagation.

Acknowledgement

The authors would like to thank MTR as a supporter of this research through data provision.

References

Ahmed, M. S., Cook, A. R. (1979) "Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques", Transportation Research Record, 722, pp. 1–9.

Bahsoon, R., Ali, N., Heisel, M., Maxim, B., Mistrik, I. (2017) "Chapter 1 - Introduction. Software architecture for cloud and big data: An open quest for the architecturally significant requirements", In: Software Architecture for Big Data and the Cloud, Morgan Kaufmann, pp. 1–10. ISBN 978-0-12-805467-3 https://doi.org/10.1016/b978-0-12-805467-3.00001-6

Bai, Y., Sun, Z., Zeng, B., Deng, J., Li, C. (2017) "A multi-pattern deep fusion model for short-term bus passenger flow forecasting", Applied Soft Computing, 58, pp. 669–680. https://doi.org/10.1016/j.asoc.2017.05.011

Bin, Y., Zhongzhen, Y., Baozhen, Y. (2006) "Bus arrival time prediction using support vector machines", Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, 10(4), pp. 151–158. https://doi.org/10.1080/15472450600981009

Blainey, S. P., Preston, J. M. (2013) "Extending geographically weighted regression from points to flows: a rail-based case study", Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 227(6), pp. 724–734. https://doi.org/10.1177/0954409713496987

Breiman, L. (2001) "Random forests", Machine Learning, 45(1), pp. 5–32. https://doi.org/10.1023/a:1010933404324

Chen, E., Ye, Z., Wang, C., Xu, M. (2020) "Subway passenger flow prediction for special events using smart card data", IEEE Transactions on Intelligent Transportation Systems, 21(3), pp. 1109–1120. https://doi.org/10.1109/tits.2019.2902405

de Dios Ortúzar, J., Willumsen, L. G. (2011) "Modelling Transport", John Wiley & Sons. ISBN 978-1-119-99352-0

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, 8(11), 1100. https://doi.org/10.3390/su8111100

Drabicki, A., Kucharski, R., Cats, O., Szarata, A. (2021) "Modelling the effects of real-time crowding information in urban public transport systems", Transportmetrica A: Transport Science, 17(4), pp. 675–713. https://doi.org/10.1080/23249935.2020.1809547

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., Vapnik, V. (1996) "Support vector regression machines", In: Jordan, M. I., Petsche, T. (eds.) NIPS'96: Proceedings of the 9th International Conference on Neural Information Processing Systems, Denver, CO, USA, pp. 155–161.

Gal, A., Mandelbaum, A., Schnitzler, F., Senderovich, A., Weidlich, M. (2017) "Traveling time prediction in scheduled transportation with journey segments", Information Systems, 64, pp. 266–280. https://doi.org/10.1016/j.is.2015.12.001

Ghofrani, F., He, Q., Goverde, R. M. P., Liu, X. (2018) "Recent applications of big data analytics in railway transportation systems: A survey", Transportation Research Part C: Emerging Technologies, 90, pp. 226–246. https://doi.org/10.1016/j.trc.2018.03.010

Guo, J., Xie, Z., Qin, Y., Jia, L., Wang, Y. (2019) "Short-term abnormal passenger flow prediction based on the fusion of SVR and LSTM", IEEE Access, 7, pp. 42946–42955. https://doi.org/10.1109/access.2019.2907739

Hastie, T., Tibshirani, R., Friedman, J. (2017) "The elements of statistical learning: Data mining, inference, and prediction", Springer. ISBN 9780387848570

Hörcher, D., Graham, D. J. (2021) "The Gini index of demand imbalances in public transport", Transportation, 48(5), pp. 2521–2544. https://doi.org/10.1007/s11116-020-10138-4

Hörcher, D., Singh, R., Graham, D. J. (2022) "Social distancing in public transport: mobilising new technologies for demand management under the Covid-19 crisis", Transportation, 49(2), pp. 735–764. https://doi.org/10.1007/s11116-021-10192-6

Koutsopoulos, H. N., Ma, Z., Noursalehi, P., Zhu, Y. (2019) "Chapter 10 - Transit data analytics for planning, monitoring, control, and information", In: Antoniou, C., Dimitriou, L., Pereira, F. (eds.) Mobility Patterns, Big Data and Transport Analytics: Tools and Applications for Modeling, Elsevier, pp. 229–261. ISBN 978-0-12-812970-8 https://doi.org/10.1016/b978-0-12-812970-8.00010-5

Lee, M., Jeon, I., Jun, C. (2022) "A deterministic methodology using smart card data for prediction of ridership on public transport", Applied Sciences, 12(8), 3867. https://doi.org/10.3390/app12083867

Li, Y., Huang, C., Jiang, J. (2018) "Research of bus arrival prediction model based on GPS and SVM", In: 2018 Chinese Control and Decision Conference (CCDC), Shenyang, China, pp. 575–579. ISBN 978-1-5386-1245-3 https://doi.org/10.1109/ccdc.2018.8407197

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, pp. 306–328. https://doi.org/10.1016/j.trc.2017.02.005

Liaw, A., Wiener, M. (2002) "Classification and Regression by randomForest", R News, 2(3), pp. 18–22.

Ling, X., Huang, Z., Wang, C., Zhang, F., Wang, P. (2018) "Predicting subway passenger flows under different traffic conditions", PLoS ONE, 13(8), e0202707. https://doi.org/10.1371/journal.pone.0202707

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(6), pp. 1623–1632. https://doi.org/10.1109/tits.2016.2611644

Okutani, I., Stephanedes, Y. J. (1984) "Dynamic prediction of traffic volume through Kalman filtering theory", Transportation Research Part B: Methodological, 18(1), pp. 1–11. https://doi.org/10.1016/0191-2615(84)90002-x

Pelletier, M.-P., Trépanier, M., Morency, C. (2011) "Smart card data use in public transit: A literature review", Transportation Research Part C: Emerging Technologies, 19(4), pp. 557–568. https://doi.org/10.1016/j.trc.2010.12.003

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, 19(3), pp. 273–288. https://doi.org/10.1080/15472450.2013.868284

Ruppert, D., Wand, M. P., Carroll, R. J. (2009) "Semiparametric regression during 2003–2007", Electronic Journal of Statistics, 3, pp. 1193–1256. https://doi.org/10.1214/09-ejs525

Sun, L., Jin, J. G., Lee, D.-H., Axhausen, K. W., Erath, A. (2014) "Demand-driven timetable design for metro services", Transportation Research Part C: Emerging Technologies, 46, pp. 284–299. https://doi.org/10.1016/j.trc.2014.06.003

Sun, Y., Leng, B., Guan, W. (2015) "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system", Neurocomputing, 166, pp. 109–121. https://doi.org/10.1016/j.neucom.2015.03.085

Tang, Y. F., Lam, W. H. K., Ng, P. L. P. (2003) "Comparison of Four Modeling Techniques for Short-Term AADT Forecasting in Hong Kong", Journal of Transportation Engineering, 129(3), pp. 271–277. https://doi.org/10.1061/(asce)0733-947x(2003)129:3(271)

Tirachini, A., Cats, O. (2020) "COVID-19 and public transportation: Current assessment, prospects, and research needs", Journal of Public Transportation, 22(1), pp. 1–21. https://doi.org/10.5038/2375-0901.22.1.1

Wang, D., Zhang, J., Cao, W., Li, J., Zheng, Y. (2018) "When Will You Arrive? Estimating Travel Time Based on Deep Neural Networks", Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), pp. 2500–2507. https://doi.org/10.1609/aaai.v32i1.11877

Williams, B. M., Hoel, L. A. (2003) "Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results", Journal of Transportation Engineering, 129(6), pp. 664–672. https://doi.org/10.1061/(asce)0733-947x(2003)129:6(664)

Wood, S. N. (2017) "Generalized additive models: An introduction with R", CRC Press, Taylor & Francis Group. ISBN 9781498728331

Wood, S. N., Goude, Y., Shaw, S. (2015) "Generalized additive models for large data sets", Journal of the Royal Statistical Society Series C: Applied Statistics, 64(1), pp. 139–155. https://doi.org/10.1111/rssc.12068

Wright, M. N., Ziegler, A. (2017) "ranger: A fast implementation of random forests for high dimensional data in C++ and R", Journal of Statistical Software, 77(1), pp. 1–17. https://doi.org/10.18637/jss.v077.i01

Xue, G., Liu, S., Ren, L., Ma, Y., Gong, D. (2022) "Forecasting the subway passenger flow under event occurrences with multivariate disturbances", Expert Systems with Applications, 188, 116057. https://doi.org/10.1016/j.eswa.2021.116057

Yang, X., Xue, Q., Ding, M., Wu, J., Gao, Z. (2021a) "Short-term prediction of passenger volume for urban rail systems: A deep learning approach based on smart-card data", International Journal of Production Economics, 231, 107920. https://doi.org/10.1016/j.ijpe.2020.107920

Yang, X., Xue, Q., Yang, X., Yin, H., Qu, Y., Li, X., Wu, J. (2021b) "A novel prediction model for the inbound passenger flow of urban rail transit", Information Sciences, 566, pp. 347–363. https://doi.org/10.1016/j.ins.2021.02.036

Zhang, G., Patuwo, B. E., Hu, M. Y. (1998) "Forecasting with artificial neural networks: The state of the art", International Journal of Forecasting, 14(1), pp. 35–62. https://doi.org/10.1016/s0169-2070(97)00044-7

Zhang, N., Graham, D. J., Bansal, P., Hörcher, D. (2022) "Detecting metro service disruptions via large-scale vehicle location data", Transportation Research Part C: Emerging Technologies, 144, 103880. https://doi.org/10.1016/j.trc.2022.103880

Zou, L., Shu, S., Lin, X., Lin, K., Zhu, J., Li, L. (2022) "Passenger flow prediction using smart card data from connected bus system based on interpretable XGBoost", Wireless Communications and Mobile Computing, 2022, 5872225. https://doi.org/10.1155/2022/5872225

Appendix

Table A1 Strengths and limitations of modelling methods

General class of model	Input considerations	Outputs	Strengths/Limitations
Mathematical analytical models	Geographical and temporal data, trends, outliers, seasonality, station interdependencies, peak and off-peak travelling, elasticities	Travel demand and travel times	*Strengths* Provides information on the dependence of results on parameters Provides direct performance sensitivity information *Limitations* Mathematical formulation of complex models can be intractable Mathematical analysis that establishes estimates of certain quantities cannot be used for different ones in some cases found challenging to automate
Simulation models	Similar to above	Traffic flow predictions	*Strengths* Can be used to model complex systems Hypothetical configurations Accounts for dynamic behaviour, stochasticity Full control of experimental conditions *Limitations* Simplification for analytical solution might be excessive Postulation of components of a system without further analysis Little information on the dependence of results on parameters No direct performance sensitivity information
Machine learning methods	All of the above in addition to more information concerning mode attributes (travel times, travel time variability, costs, crowding levels). Image (photos and videos) analysis, e.g. satellite photos, street photos, video surveillance; Sentiment analysis of passengers of particular transport modes, e.g. social media data (hashtags, textual and image analysis); Financial transactions analysis and expenditure, e.g. financial statements.	Travel demand and travel times	*Strengths* Typically, better at predictions under particular conditions. Greater spatial and temporal coverage (depends on the data). Capable of absorbing and processing more data (big data). Capable of absorbing a greater variety of data. Capable of capturing non-linear effects. *Limitations* Uncertain performance in forecasting structural/step changes/unless such were included in the training data. Data and processing power hungry. Often lack transparency ("black box" problem). Privacy considerations associated with using large volumes of high-resolution data.