Transportation Research Interdisciplinary Perspectives 2 (2019) 100027

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: https://www.journals.elsevier.com/transportation-research-interdisciplinary-perspectives

Modeling bus passenger boarding/alighting times: A stochastic approach

Taqwa AlHadidi [a,b,d], Hesham A. Rakha [a,c,d,*]

[a] Charles E. Via, Jr. Department of Civil and Environmental Engineering, Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States of America
[b] Civil Engineering Department, Al-Ahliyya Amman University, Amman 19328, Jordan
[c] Bradley Dept. of Electrical and Computer Engineering, Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States of America
[d] Center for Sustainable Mobility, Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States of America

ARTICLE INFO

Article history:
Received 30 March 2019
Received in revised form 3 July 2019
Accepted 25 July 2019
Available online 27 August 2019

Keywords:
Bus loading time
Transit
Bus stop
Transit passengers
Stochastic modeling
Bootstrap
Cholesky decomposition

ABSTRACT

This paper introduces two multilinear models that compute bus passenger boarding/alighting (BA) times at bus stops using empirical data (8341 empirical observations) from King County Metro in Seattle, Washington. The first model is a classical frequentist model, while the second is a stochastic model developed using bootstrapping together with a Cholesky decomposition. Three variables are considered in the aforementioned models to estimate passenger BA times, namely; the number of alighting passengers, the number of boarding passengers, and the number of passengers on board the bus. The models show that passenger BA times increase with an increase in all three variables (i.e. positive model coefficients). The Cholesky decomposition technique was applied to bootstrapped data to capture the model coefficient correlations with the use of only nine parameters (coefficient means, variances, and correlations) while capturing the stochasticity observed in the empirical data.

1. Introduction

Intelligent Transportation System (ITS) applications attempt to enhance the efficiency of the transportation system through the use of emerging technology. One of many intelligent transportation system applications, advanced traffic management and information systems attempt to enhance the efficiency and resilience of the transportation system using real-time data collected from probe vehicles and stationary sensors, such as loop detectors, video cameras, or radar sensors (Kay, 1990 #2).

However, predicting and estimating multimodal travel times is a challenging task due to the complexity of the factors affecting urban network travelers. These factors include day-to-day and within-day demand fluctuations and traffic demand compositions (including passenger cars, non-motorized passengers, and transit vehicles). Predicting transit vehicle travel time is necessary to allow travelers to avoid traffic congestion and travel more efficiently. Using reliable travel time data within Advanced Traveler Information Systems (ATISs) and Advanced Traffic Management Systems (ATMSs) helps transportation system users acquire better information to avoid traffic congestion. Using reliable travel times also allows users to estimate a more reliable arrival time at their destination.

Transit vehicle travel time on any link depends on different factors, including the link's geometric features (i.e., dedicated transit lane versus non-dedicated transit lane), bus stop locations (far-side, near-side, and midblock), bus stop configurations (number of loading areas), adjacent traffic conditions (volume and composition), and traffic control systems (fixed traffic signal, actuated traffic control, or transit signal priority in an actuated traffic control system). Transit vehicle travel times consist of three components: transit vehicle running times along roadways, dwell times (DTs) at bus stops, and signalized intersection delay.

Typically, bus travel time variability stems from the variations in passenger demands (day-to-day and bus-to-bus), traffic signal delays after stopping to serve passengers, wheelchair lift and bicycle rack usage, differences in bus operator experience, route length, and the number and location of stops (TRB, 2013).The time elapsed when serving passenger demand at any station, known as the boarding/alighting (BA) time, constitutes the largest portion of DT.

Transit data collection requires a large database. Data types include operational, spatial, and temporal data. Ultimately, the collected data are used in improving transit quality of service, and in evaluating a transit agency's operational performance. Using Automatic Vehicle Location (AVL) and Automatic Passenger Count (APC) systems help in gathering transit data,

* Corresponding author at: Charles E. Via, Jr. Department of Civil and Environmental Engineering, Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States of America.
E-mail addresses: Taqwaa1@vt.edu, (T. AlHadidi), hrakha@vt.edu. (H.A. Rakha).

http://dx.doi.org/10.1016/j.trip.2019.100027

enhance transit scheduling, and improve service planning and quality monitoring (Furth et al., 2006).

2. Study objectives and paper layout

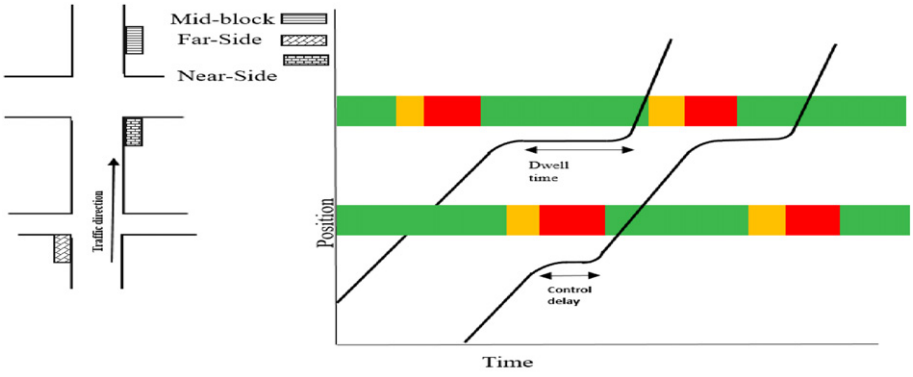
The objective of this study is twofold. First, it attempts to identify critical variables and model bus passenger BA times using a deterministic frequentist regression modeling approach. Second, it develops a stochastic modeling approach by combining bootstrapping and a Cholesky decomposition approach to capture empirically observed stochastic behavior with the use of only nine parameters.

The paper is laid out as follows. Initially, previous research findings are presented. Subsequently, the modeling approach is described followed by a description of the data gathering and validation procedures. Next a discussion of the frequentist regression statistical model that was used to model bus BA times is presented. Subsequently, the bootstrap approach to modeling the loading time is presented, followed by a comparison of the two approaches. The modeling of stochasticity in the model parameters using the Cholesky decomposition is presented next. Finally, the conclusions and recommendations of the paper are presented.

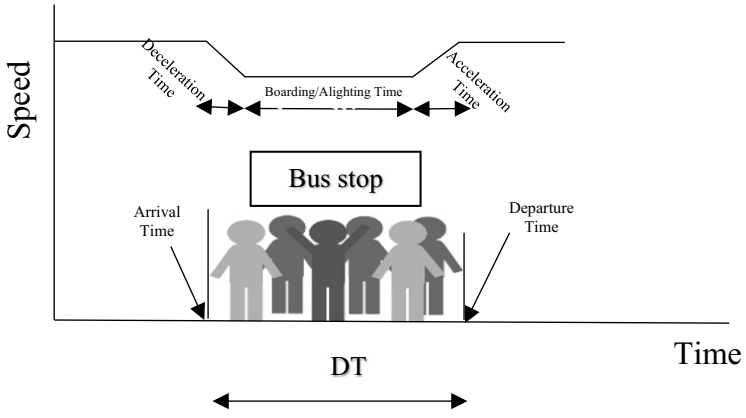
3. Previous findings

Dwell time is defined as the amount of time a transit vehicle remains stopped at a bus stop to serve passengers boarding and/or alighting. DT consists of the time it takes passengers to alight and board the bus (BA time), the time it takes to open and close the door(s), and the time needed to merge with traffic, as shown in Fig. 1 (TRB, 2013). These combined times are critical stochastic factors that affect bus travel times. This variability arises from the fluctuation in the passenger demand between stops for different times of the day, different days of the week, vehicle load factors, passengers with mobility issues, and passengers loading bicycles, luggage, and so forth. Moreover, passenger BA times depend on passenger demand and the bus configuration (e.g., number of doors, floors, and number of units/articulated versus single unit). Consequently, the *Transit Capacity and Quality of Service Manual* (TCQSM) suggests three methods to estimate bus DTs, including field measurements using archived AVL and APC data, default values used by planners, and use of models based on field data considering BA counts, time for fare payment, passenger demand, vehicle configuration, and stop configuration (TRB, 2013). Using archived AVL/APC data helps in designing the collection of real-time data, allowing more accurate computation of the time difference between the transit vehicle arrival time and the departure time at any bus stop (Furth et al., 2006).

DT has been modeled for decades, and can be categorized into two main groups: ordinary-least squares regression models and probabilistic models. Several linear models have been used to estimate DTs, taking into consideration the number of boarding and alighting passengers (Guenthner and Sinha, 1983; Levinson, 1983; Guenthner and Hamat, 1988; Levine and Torng, 1994; Fricker, 2011; TRB, 2013). The first attempt to predict DTs was made by Levinson in 1983 using transit survey data collected from U.S. cities in 1980. Mainly estimating DT as a function of the number of passengers boarding, Levinson found that that DT constitutes about 9–29% of transit vehicle travel time (Levinson, 1983). Considering other DT determinants like fare collection and bus configuration (i.e., number of doors and floors), other authors found that the fare method did not produce significant differences in DTs (Guenthner and Hamat, 1988). Lower bus floors



(a) Transit travel time along link



(b) Transit travel time at bus stop

Fig. 1. Transit vehicle travel time.

were found to decrease DT by 13–15% (Levine and Torng, 1994). Increasing the number of doors and faster fare collections was found to provide faster service, reducing DTs (Kraft, 1975). Ordinary least squares models for estimating DTs consider the number of served passengers as the most relevant factor (Levine and Torng, 1994, and Guenthner and Hamat, 1988). Another study showed that DT is affected by the fare method and the number of doors, but not time-of-day (Kraft, 1975). A later study found that time of day and service type do not show a significant effect on bus dwell time estimation, and adding to the number of standing passengers results in a higher model coefficient of determination (Rajbhandari et al., 2003). Subsequent studies developed several nonlinear and stochastic models as well as a non-stationary, high dimensional time series models to estimate bus travel times (Chen et al., 2007, Mazloumi et al., 2011, Gurmu and Fan, 2014, Meng and Qu, 2013).

A study conducted in Auckland, New Zealand estimated DTs using manual field collected data from 22 bus stops. The study compared Gene Expression Programming (GEP) and Multiple Linear Regression (MLR) results to estimate and model DTs. The results indicated that the GEP model results were better than the MLR in modeling and accurately estimating DTs (Rashidi and Ranjitkar, 2015). Another study was conducted to estimate DTs in traffic congestion using a total of 895 records. The collected data were analyzed using the Compound Poisson Service Time estimation Model (CPSTM) that takes into consideration the interaction between the arriving buses and number of boarding and/or alighting passengers. Different scenarios were used to estimate the total service time. The study results indicated that the accurate service time estimation does not require collecting real-time data (Bian et al., 2015).

Another study was conducted in China to estimate bus DTs and stop lost serving time, it expresses the dwell time as a function of dead time, acceleration and deceleration time, and Boarding/Alighting (B/A) times. The model was validated using data from 7 common stops. Results of model validation show that the model is able to describe >82% of the on-peak and off-peak stop times (dwell times and lost times), with a Mean Absolute Percentage Error (MAPE) of approximately of 13% (Wang et al., 2016).

Using different data sources to estimate DTs a number of studies were recently conducted. Using video collected data a model was developed for Bust Rapid Transit (BRT) systems using a total of 877 observations from a full day service. These observations include the Bus Lost Time (BLT) as a component of the DTs. The study provides an accurate estimation of DT by incorporating the BLT into DT estimation model. Using the developed model it was possible to reduce the BRT capacity by 11% (Kathuria et al., 2016). While a combination of stop-level and GPS data were used to prevent a bias estimation of DT, results of the study showed that the average time per passenger boarding or alighting decreases as the number of passenger boarding or alighting increases. Also, results indicated that DT for nearside or far-side stops is affected by the traffic queuing, or signalized intersection delay; however, no evidence shows the impact of intersection delay or the traffic queuing on the midblock stop DT (Glick and Figliozzi, 2017).

4. Methodology

This section describes the two methodological approaches that were used to develop the proposed models, namely: the frequentist approach and the stochastic approach using the Cholesky decomposition.

4.1. Frequentist modeling

As described in Wikipedia, "*In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.*" Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Frequentist statistical models assume a constant variance of the error term. However, the error distribution depends on the distribution of the model variables and may not be necessarily normally distributed and have a constant variance. Consequently, model transformations may be needed in order to ensure that the error is both normally distributed and that the variance is constant. The tests on the residual error and the procedures used to transform the empirical data are described later in this paper.

4.2. Stochastic modeling

The nature of explanatory variables in field measurements is stochastic because the explanatory variable in some studies may act like the response variable in another study. Several resampling methods have been discussed in the literature, including cross-validation, jackknifing, bootstrapping test, and randomization tests (which are widely used in different statistical inferences also it is known as permutation tests).

According to asymptotic theory (law of large numbers), as the sample size grows, the expected value of the variable coefficient converges to its mean because the variance converges to zero. Thus, a large sample of data is needed in order to estimate population distributions (i.e., mean, variance, and confidence intervals) accurately. For population distribution estimation, repeating large numbers of the sample size is an efficient way to express the population distribution (Singh and Xie, 2008; Fox, 2002).

Nonparametric bootstrapping estimates the sampling distribution without the need to assume the population distribution. This assumption arises for the condition where a sample is drawn from the population with replacement $S_1^* = \{X_{11}^*, X_{12}^*, \ldots\ldots\ldots, X_{1n}^*\}$; the reason for sampling with replacement is to produce a different sample of the original data set to generate a more accurate estimation of the model's statistics. Broadening the data through resampling it is possible to fit models and save the model coefficient values from each bootstrap sample; thus, not only can the coefficient distribution be generated, but the confidence intervals, covariance, quantiles, and error can be also be easily generated. Moreover, resampling allows for successful estimation of the model's uncertainty and the precision of the selected model (Singh and Xie, 2008; Fox, 2002). Based on the normality assumption, the amount of bootstrapping resampling affects the model's confidence intervals; therefore, the minimum bootstrapping resampling is recommended to be 1000 (Fox, 2002). The direct use of bootstrapped data is data intensive because it requires the storage of large amounts of sample realizations. To overcome this drawback, the Cholesky decomposition can be used.

Cholesky decomposition is used widely as a quick mathematical numerical solution to a Monte Carlo simulation. The idea of using Cholesky is to solve the problem stochastically by using the law of large numbers, which uses the mean and the variance for the variable in order to draw a distribution function of the variable. In particular, Cholesky decomposition is used to decrease the effect of the multi-collinearity between variables. The reason for using the Cholesky decomposition is that this technique mainly depends on breaking a symmetric positive-definite matrix into the product of a lower triangular matrix and its conjugate transpose. Moreover, using the positive-definite matrix yields exactly one solution for the matrix set.

Let **A** be a positive definite matrix. Then **A** can be decomposed in exactly one way into a product: $\mathbf{A} = \mathbf{R}^{\mathrm{T}}\mathbf{R}$ such that **R** is an upper triangular matrix and has all main diagonal entries; this is called the Cholesky factor of **A**.

Generally, Cholesky decomposition can be used to estimate β, which here is called the revised Cholesky least-squares estimator, and can be obtained using Eq. (1).

$$\beta(\mu) = S(\mu)^{-1} X^T \tag{1}$$

To elaborate on this, supposing that $Z_0, \ldots, Z_3$ are independent and identically distributed standard normal vectors (i.e., $Z_i \sim N(0,1)$), then the

problem can be formulated using Eqs. (2) and (3).

$$\begin{aligned} \mu_{\beta_0} + c_{00}Z_0 &= \beta_0 \\ \mu_{\beta_1} + c_{10}Z_0 + c_{11}Z_1 &= \beta_1 \\ \vdots &\quad \vdots \\ \mu_{\beta_4} + c_{40}Z_0 + c_{41}Z_1 + \cdots + c_{44}Z_4 &= \beta_4 \end{aligned} \tag{2}$$

$$\begin{pmatrix} \mu_{\beta_0} \\ \mu_{\beta_1} \\ \vdots \\ \mu_{\beta_4} \end{pmatrix} + \begin{pmatrix} c_{00} & 0 & 0 & 0 \\ c_{10} & c_{11} & 0 & 0 \\ \vdots & & \ddots & 0 \\ c_{40} & c_{41} & \cdots & c_{44} \end{pmatrix} \begin{pmatrix} Z_0 \\ Z_1 \\ \vdots \\ Z_4 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_4 \end{pmatrix} \Rightarrow \mu + CZ = \beta \tag{3}$$

Here C is a lower triangular matrix that represents the coefficients that need to be determined. By taking the variance of both sides, and given that μ is constant (Var(μ) = 0) and Z is standard normal (Var(Z) = I) one can compute the variance of the model coefficients. Eq. (4) shows the results of taking the variance of the left side of Eq. (3). In Eq. (5) the variance of the right-hand side of the equation is equated to that of Eq. (4) to solve for the variance matrix Σ. Matrix Σ can be calculated easily given the correlations between the parameters and their standard deviations.

$$\mathrm{Var}(\mu + CZ) = CC^T \tag{4}$$

$$\mathrm{Var}(\beta) = \Sigma = \begin{pmatrix} \sigma^2_{\beta_0} & \cdots & \rho_{ij}\sqrt{\sigma^2_{\beta_i}\sigma^2_{\beta_j}} \\ \vdots & \ddots & \vdots \\ \rho_{ij}\sqrt{\sigma^2_{\beta_i}\sigma^2_{\beta_j}} & \cdots & \sigma^2_{\beta_4} \end{pmatrix} \Rightarrow \Sigma = CC^T \tag{5}$$

5. Data

The data used in the analysis were obtained from King County Metro, Seattle's main transit operator. King County Metro is the eighth-largest transit operator in the U.S., serving on average, >395,000 passengers daily from >200 routes. AVL and APC data were obtained from 163 buses out of a fleet of >1540 buses. These buses are all single unit transit vehicles with two doors: one front door and one back door. The raw dataset contained >12,000 records. The data did not provide passenger characteristics (i.e., age, gender) nor the fare collection method; however, King County Metro uses mainly smart cards as a payment method. The operating transit vehicles analyzed in this research effort were one-floor vehicles with two doors: one front door for boarding passengers and one rear door for alighting passengers. The transit vehicles had a total capacity of 82 persons, with room for 39 seated passengers and 43 standees. The explanatory variables considered in the models are thus limited to the number of passengers boarding, passengers alighting, and passengers on board. The time-of-day factor and stop type factor (i.e., far-side, near-side and midblock) were eliminated based on previous studies in addition to the need to build a generalized model in which time and location (i.e. far side, midblock and near-side) do not matter(Kraft, 1975).

5.1. Data validation and reduction

Data records were reduced to 8341 data points, data were validated using a sequence of validation standards. Given that the data points come from the combination of AVL and APC data, the validation steps for APC data differs from the AVL data. The adopted steps for validating each data source are presented.

APC data used in this research recorded several parameters including the number of passengers boarding, alighting and passenger on board counts, stop name, stop ID, trip ID and route name. The validation of the APC data included the following sequence of steps:

1. Remove any demand response trip. This was done by checking the route ID, stop name and the stop ID to ensure that any stop belongs to the correct route. If this condition was not met, a trip was considered a demand responsive trip and removed from the analysis.
2. For each transit trip, the difference between the total number of the boarding passengers and the total number of the alighting passengers' must be <15% as recommended in the literature (Strathman and Hopper, 1989).
3. The sum of the boarding passengers and on board passengers at any stop could not be higher than the bus capacity.

AVL data included other parameters, some of which matched the APC data parameters, while others provided additional data including the actual bus arrival time, the scheduled bus arrival time, the bus departure time, the bus dwell time, and the bus loading time. The AVL data were only used within the vicinity of the bus stop.

Validation of the AVL data followed these procedures:

1. Check the reasonability of arrival and/or departure times (e.g., departure time record was later than the arrival time).
2. Remove records with missing APC data for that bus stop and stop ID.
3. Remove observations for which the distance between stops was inaccurate by >15%.
4. Remove trips for conditions in which the actual arrival time differed by >20 min from the scheduled arrival time.
5. Remove outliers where the DT was >210 s, which is the recommended time for serving passengers with disabilities, or where DT was <4 s (i.e., less than the minimum recommended time for vehicles to open and close the door) (TRB, 2013).

6. Model development

This section describes the procedures used to develop the proposed models. In order to develop a generalized model that can be used for different bus capacities; the explanatory variables were normalized by dividing by the capacity of the bus. These normalized variables included the number of boarding passengers divided by the bus capacity, the number of alighting passengers divided by the bus capacity and the number of riders divided by the bus capacity. In this case all the explanatory variables are dimensionless and thus produce a more generalizable model.

6.1. Deterministic model development

This section presents the development of the deterministic bus BA time model using the King County Metro data. Data was split randomly into two sub-sets: 88–90% of the data points were used for model calibration and the remaining data set was used to validate the model. Model building has two conflicting objectives. The model has to be informative by including as many regressors as needed to predict the dependent variable, y; however, the model has to decrease the variances in the measured predicted value $\hat{y}$ by involving the least number of regressors. Consequently, finding a compromise between these objectives results in the best regression model.

Choosing the most relevant candidate of regressors was the first step in the model development. Specifically, three variables for estimating the BA time were considered, namely: the number of boarding passengers, the number of passengers onboard the bus, and the number of alighting passengers. All three variables were normalized by dividing by the bus capacity to produce dimensionless explanatory variables (X_1, X_2, and X_3, respectively).

Development of the BA time was done in three stages. First, data were used to fit some existing models. Then, the model was built based on an iterative transformation of the response and the regressors. Finally, the best model was evaluated based on model criteria selection and the minimum residual errors.

Over 100 models were considered and tested in the analysis, including existing and proposed models. These models included linear, polynomial, and exponential models. The consideration of variables was tested using both forward and backward stepwise regression. In the forward regression the model starts with no coefficients and then adds explanatory variables as needed. In the backward regression approach the model starts with a full model and then removes variables as needed. Different variables and the

various variable interactions was taken into consideration. Response and explanatory variables were transformed as needed to satisfy the normality error assumption. Considering all possible transformations, variables, and variable interactions resulted in the over 100 models. Failure of a large number of models is attributed to their inability to describe the variance in the data. A list of sample candidate models with their statistic metrics is presented in Table 1.

Several criteria were used for model selection. These included the sum of residuals, Mallows (C_p), the coefficient of determination (R^2), the adjusted coefficient of multiple determination (R^2_{adj}), the residual mean square error (MS_{res}), the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the PRESS residuals.

Comparing the candidate models, the proposed model clearly outperforms the other models based on most of the selection criteria. The proposed model shows the lowest mean residual error, with the highest coefficient of multiple determination, which reflects the model's capability to model approximately 88.9% of the data compared with the other models. The model also has the best trade-off between the estimated parameters and the model goodness of fit using the BIC and AIC values. With this in mind, testing the model precision and avoiding model overfitting was done by comparing the Mallows C_p value for each model and choosing the lowest C_p value.

Assuming that vehicle BA time is continuous and the demand data (number of passengers loading, alighting, and onboard the transit vehicle) are integer values, a general linear model was fit to the data considering a normal distribution of the form in Eq. (6).

$$\ln(T_{BA}) = \beta_1\sqrt{X_1} + \beta_2\sqrt{X_2} + \beta_3 X_3 + \varepsilon \tag{6}$$

Here T_{BA} is the duration of time the transit vehicle door is open (BA time), βi's are the model coefficients, X_1 is the of the ratio of the number of boarding passengers to the bus capacity, X_2 is the ratio of the number of passengers on board the bus to the bus capacity, X_3 is the ratio of the alighting passengers to the bus capacity, and ε is the error term.

Transforming Eq. (6), the final equation for estimating the bus BA time is computed using Eq. (7). This equation puts a lower and upper bound to the estimated BA times to ensure that the values are realistic. Here Min_{Stop} and Max_{Stop} are the minimum and maximum stop durations, taken to be 0 and 210 s in this study.

$$T_{BA} = \begin{cases} Min_{Stop} & X_1 + X_3 = 0 \\ \min\left(e^{\beta_1\sqrt{X_1}+\beta_2\sqrt{X_2}+\beta_3 X_3}, Max_{Stop}\right) & X_1 + X_3 > 0 \end{cases} \tag{7}$$

The model coefficient values along with their p values are summarized in Table 2. Furthermore, the calibrated model was validated using the validation data subset, with the model showing the ability to explain 88.4% of the total error of the randomly selected data. The table presents the dimensionless coefficients (normalized by the bus capacity of 82 passengers) and the non-dimensionless coefficients based on the actual number of boarding, on-board and alighting passengers.

Table 2
Statistical model calibration results.

Coefficients	Coefficient values (dimensionless)	Coefficient values	SError	p-Value
β_1	8.57127	0.9465	0.00490	<0.0001
β_2	1.94351	0.2146	0.00650	<0.0001
β_3	3.84286	0.0469	0.00524	<0.0001

The model demonstrates that the expected BA time is set to Min_{Stop} (in our case zero) when there are no boarding and alighting passengers regardless of the number of passengers onboard the bus. Accordingly, the expected BA time increases in the range of 1.5 s to 3.7 s per unit increase in boarding passengers in the range between 1 and 10, which is similar to the recommended serving time for boarding passengers according to the TCQSM. Alternatively, the BA time only increases by 0.1 s for the addition of an alighting passenger in the range of 1 to 10 alighting passengers. Finally the BA times increases by 0.2 s per unit increase in the number of passengers on-board the bus, without any change to the other variables included in Eq. (7). Fig. 2 demonstrates that the BA times are more sensitive to the number of boarding passengers than to the number of alighting passengers.

The adequacy and validity of the models shown in Table 2 was done using the general linear model assumptions. Choosing the best model among a set of models means that the selected model should have a normal residual error, as shown in Fig. 3.

Fig. 3 shows that the fitted values approximately follow the normal distribution, which is indicated by the solid blue line. The figure also shows that most of these values are bounded by the prediction interval, which is denoted by the dashed blue lines. However, it also shows that model results in large values at the tails of the distribution compared with standardized model assumption. This issue was resolved by setting a maximum and minimum stop duration, as demonstrated in Eq. (7). The maximum value is set at 210 s, which is the maximum recommended value by TCQSM. The minimum value is used when the sum of boarding and alighting passengers equals zero. In our case we assumed that the bus would not stop and thus set a value of zero.

6.2. Stochastic model development

This research deploys a nonparametric bootstrap technique to estimate the sampling distribution without the need to assume the population distribution. For this study, the case resampling technique was conducted using the model in Eq. (7) to capture the model coefficient distribution with the number of generated samples established as 100,000. The generated sample was checked using a subset of the large sample by a factor of 10, which results in sub-setting the sample by 10,000 observations after

Table 1
Candidate models with their summary statistics.

ID	Model variable	MS_{res}	PRESS	R^2	R^2_{adj}	C_p	AIC	BIC
1	Intercept	22.60	4,251,346	0.0000	0.0000	4242.70	75,491.32	75,505.37
2	Intercept, X_1	20.43	3,473,767	0.1835	0.1834	3449.91	73,807.11	73,828.19
3	Intercept, X_2	22.57	4,239,822	0.0029	0.0028	4208.72	75,468.87	75,489.95
4	Intercept, X_3	22.17	4,090,529	0.0383	0.0381	4000.17	75,168.45	75,189.53
5	Intercept, X_1, X_2	20.35	4,239,822	0.1898	0.1896	4208.72	73,744.89	73,773.00
6	Intercept, X_1, X_3	20.33	3,440,498	0.1917	0.1915	3370.05	73,724.84	73,752.95
7	Intercept, X_2, X_3	22.17	4,090,206	0.0386	0.0383	3900.48	75,168.10	75,196.20
8	Intercept, X_1, X_2 X_3	20.17	3,338,031	0.2042	0.2040	3238.44	73,596.79	73,631.92
9	Intercept, X_1^2	20.43	3,473,767	0.1835	0.1834	3449.91	73,807.11	73,828.19
10	Intercept, $\ln(X_1)$	20.44	3,476,882	0.1827	0.1826	3450.32	73,814.68	73,835.76
11	Intercept, X_1^2, X_2	20.35	3,448,031	0.1898	0.1896	33,789.29	73,744.89	73,773.00
12	Proposed model	0.85	5971	0.8895	0.8895	5.00	20,851.25	20,879.36

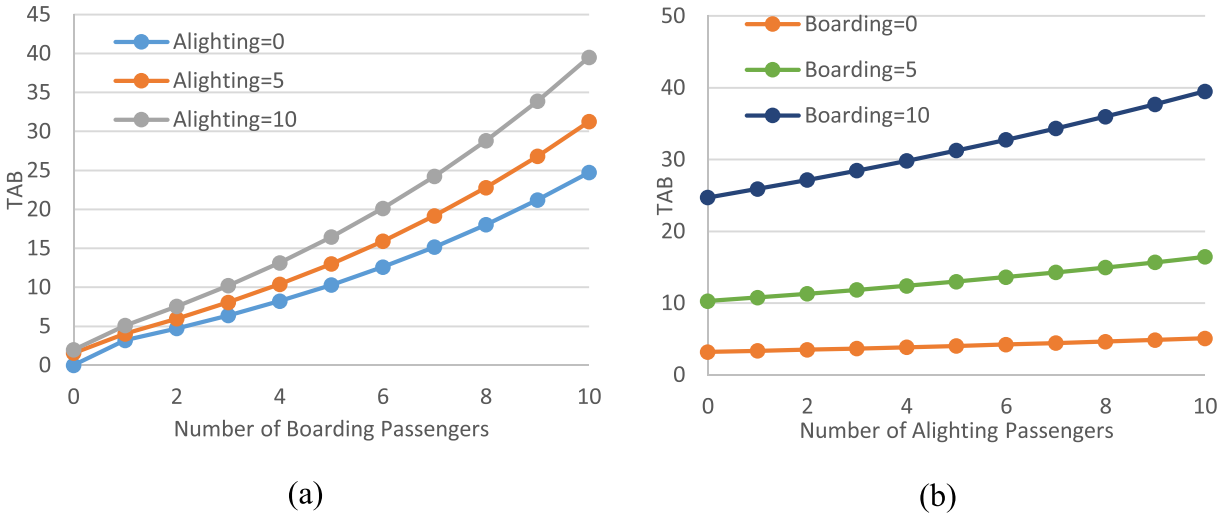


Fig. 2. Variation in BA times as a function of the number of (a) boarding passengers and (b) the number of alighting passengers.

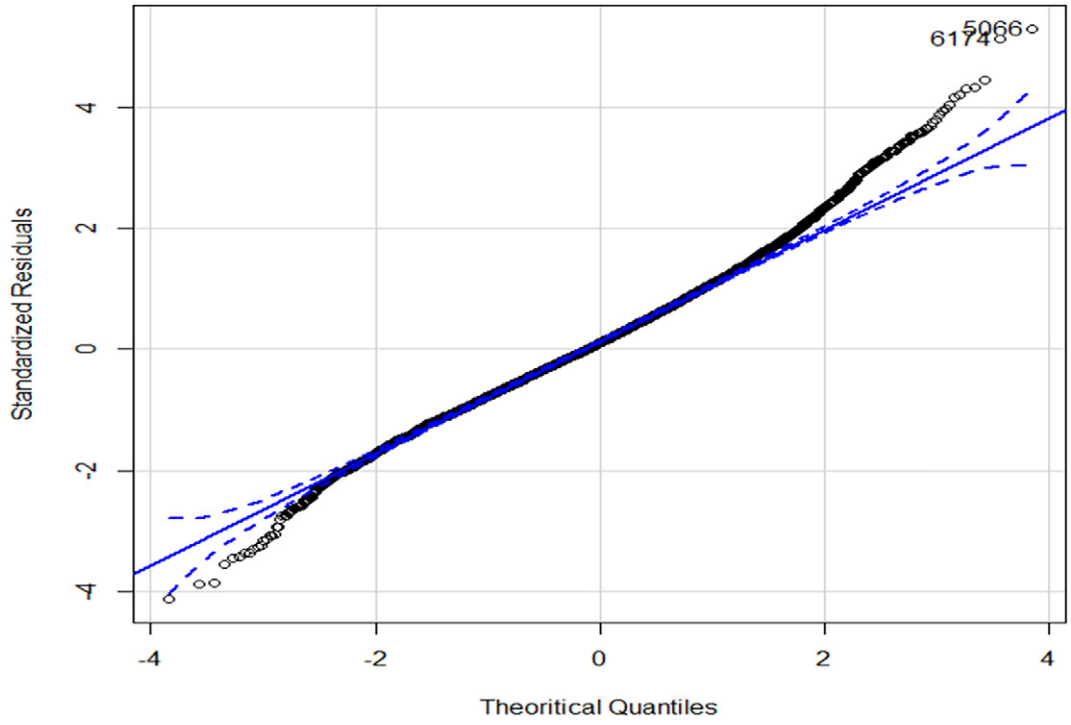


Fig. 3. Normal quantile plot of BA times.

considering every 10th observation (Singh and Xie, 2008). The bootstrapping approach that was implemented within R is an algorithm designed to sample from a distribution with an arbitrary density function, known only up to a constant of proportionality. This is needed for sampling from a complicated posterior distribution whose normalization constant is unknown.

A comparison between the original population parameters and the bootstrapped population parameters is presented in Table 3. This comparison was done to ensure that the resampling technique that was done was sufficient enough to replicate the original population.

The results summarized in Table 3 demonstrate that there is no significant difference between the population parameters of the original dataset and the bootstrapped population.

The algorithm generates random samples and fits Eq. (7). Also, the algorithm is able to generate more parameters, including the mean, standard deviation, and median for the model coefficients. The results for the generated bootstrapped samples are shown in Table 4.

Table 3
Population parameter summary statistics.

	Mean		Skewness	
	Original	Bootstrapped	Original	Bootstrapped
Boarding	0.0243	0.0288	1.8548	1.9096
Riders	0.2176	0.2176	3.4473	3.4639
Alighting	0.0230	0.0230	0.7987	0.8027
Loading time	17.3409	17.2697	5.3786	5.4349

Table 4
Bootstrap coefficient results.

	R	Original	BootBias	BootSE	BootMed	bootSkew	bootKurtosis
Boarding	10,000	8.5373	$-3.07e^{-4}$	0.0396	8.5374	−0.0285	−0.0092
Riders	10,000	1.9542	$-1.49e^{-4}$	0.0493	1.9544	0.0027	−0.0608
Alighting	10,000	3.7863	$1.65e^{-3}$	0.4124	3.7842	0.0201	−0.0316

Table 5
Confidence intervals for the coefficients.

	2.5% confidence limit	97.5% confidence limit
Boarding	8.298	8.842
Riders	1.848	2.042
Alighting	3.060	4.662

Table 5 shows that bootstrapping gives the original coefficient values, while also computing the average statistics of the generated samples using the formula in Eq. (8).

$$\overline{T}^* = \hat{E}^*(T^*) = \sum_{b=1}^{R} \frac{T_b^*}{R} \tag{8}$$

The bias of the bootstrapping was estimated by using $\hat{B}^* = \overline{T}^* - T$. The standard error was calculated for the bias using Eq. (9).

$$\widehat{SE}^*(T^*) = \sqrt{\frac{\sum_{b=1}^{R}\left(T_b^* - \overline{T}^*\right)^2}{R-1}} \tag{9}$$

To compare the standard error values for the model coefficient using conventional modeling and the bootstrapping model, the values were used to construct the regression coefficient confidence interval based on the normality assumption as both of Bootstrap skewness and Kurtosis indicated that the distribution of these coefficients follow the normal distribution as these values tend to be close to zero.

A histogram of the model's coefficient distribution, kernel density, bootstrapping, and confidence intervals of the values was individually generated for each explanatory variable coefficient. Fig. 4 shows the histograms for the bootstrapping case. The vertical dashed lines express the coefficient point-estimate in the conventional regression model (presented in Eq. (7)). Conversely, the horizontal lines express the bootstrap confidence interval for the coefficients, which is presented in Table 5. Interestingly, the figure shows that both of the distribution functions follow a normal distribution and this distribution is very strong for all coefficients.

Fig. 4 subplots show the model parameter distribution using the bootstrap technique. The X-axis shows the parameter values in second/number of subplot name (i.e., boarding, Rider, Alighting) divided by the total capacity, while the y-axis shows the density for these values. The solid black line along the x-axis in each subplot indicates the confidence interval for each parameter, while the dashed black line is the estimated value for each parameter. In each subplot, the black solid line and the dashed line values are not different from the values in Tables 4 and 5. Checking the

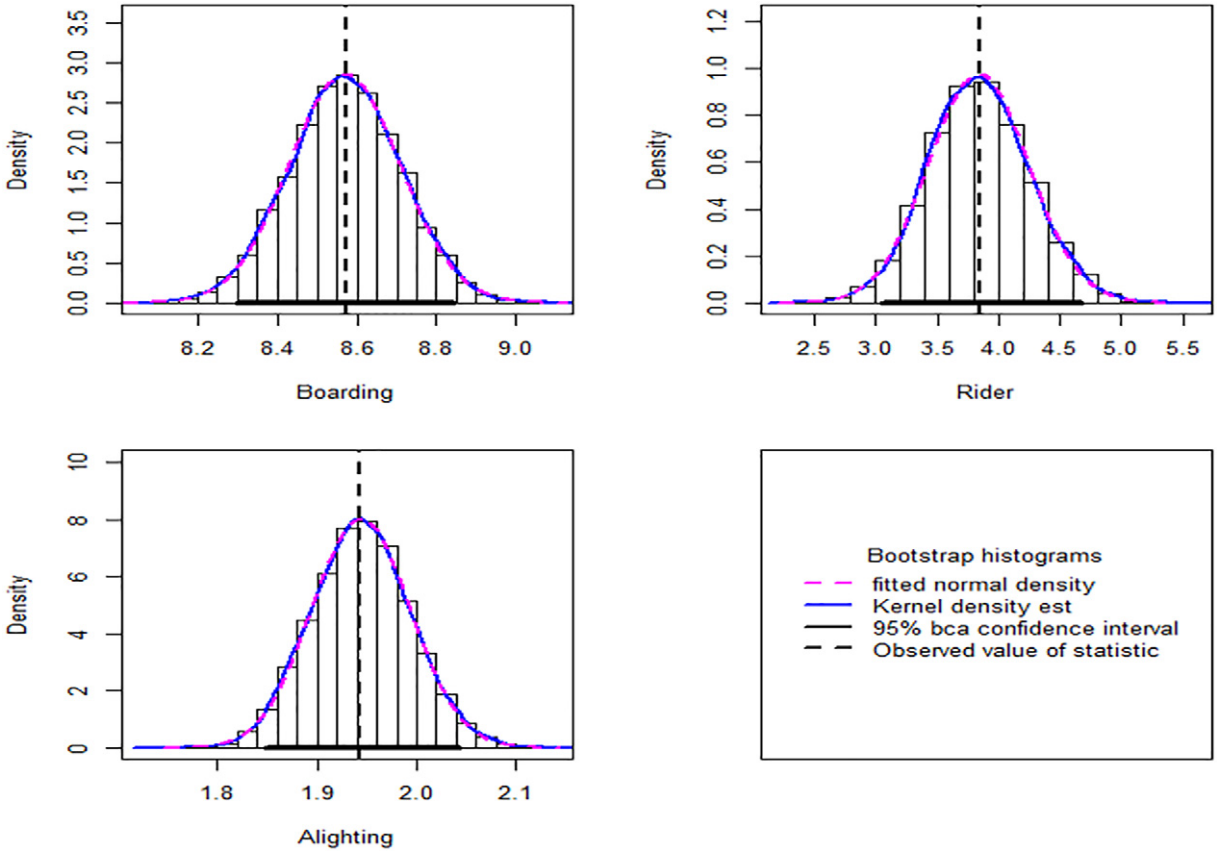


Fig. 4. Bootstrap coefficients.

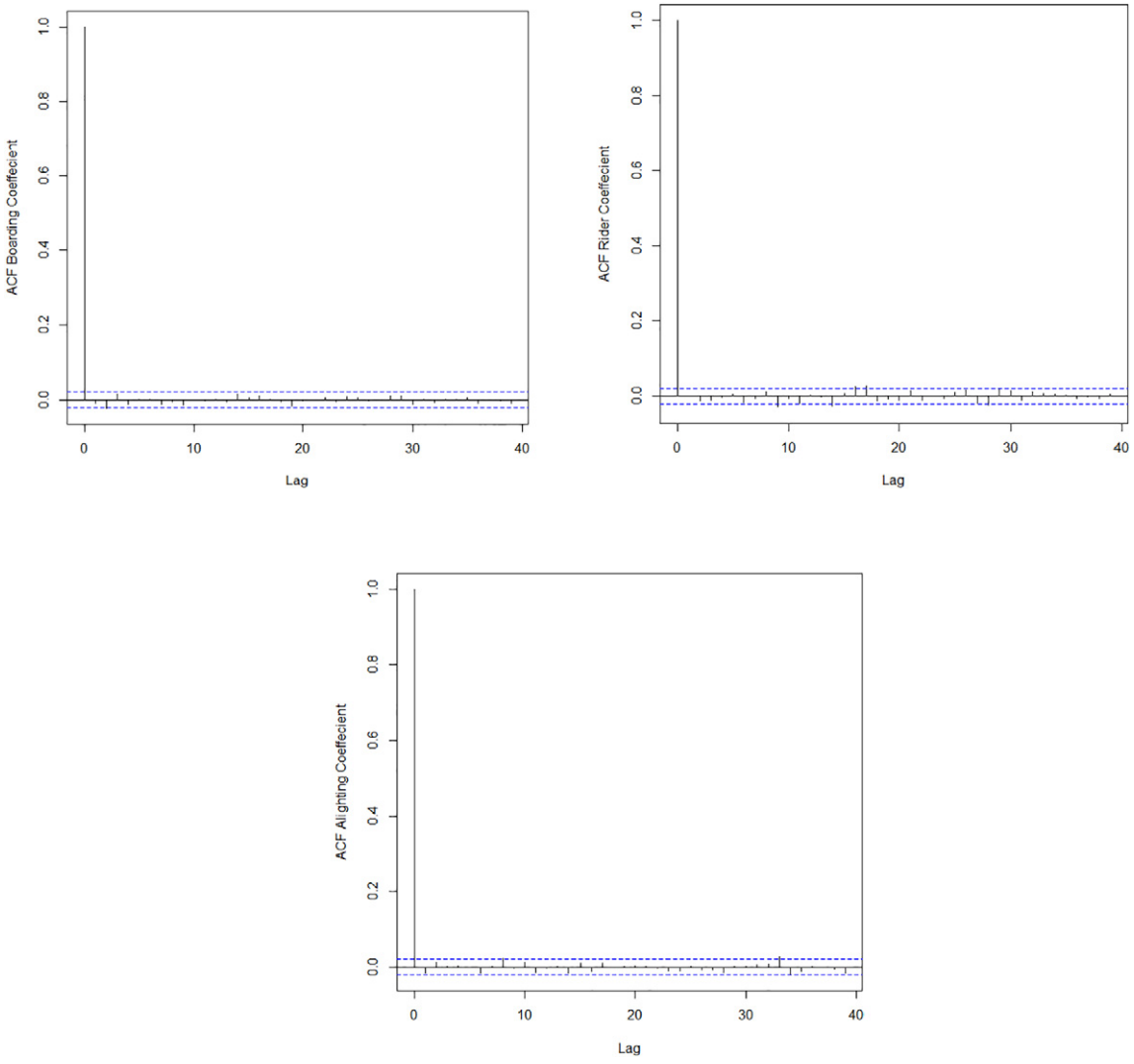


Fig. 5. Variations in the autocorrelation function of the model parameters.

distribution of variables in the proposed model as stated in Eq. (7), the pink dashed line shows the fitted parameter density, while the blue dashed line shows the normal distribution for the parameters. In order to verify the independency in the resampling process, the coefficients' convergence was checked by testing whether the generated sample size was sufficient using the autocorrelation function (ACF), a tool for checking the randomness of the data points, considering the between-the-coefficients sample points, as shown in Fig. 5.

If the data points are random, the ACF value should be near zero along all-time lag separations between them. The ACF predicts the value of a selected data point at time (t) given the prior value in time ($t - \Delta t$). Fig. 4 also shows that the coefficients' values are independent of each other. The dashed blue lines represent the 95% confidence limits, while the stem lines represent the ACF.

As expected based on bootstrapping a random sample point selection, an autocorrelation of 1.0 is observed for a lag of 0, which means that the probability of choosing the second sample would not be different from the previously selected sample. The values are within the confidence limits for larger lag times and close to zero. Fig. 6 illustrates that there is interaction between the model parameters.

Fig. 6 shows the interactions between the model coefficients (β_1, β_2, and β_3). Each subplot shows the relation between two coefficients, illustrating that the parameters are highly correlated.

The plots also show that the data points are intensively spread between the parameter confidence intervals for each subplot. Also, there is a clear linear relationship between β_1 versus β_2, and β_2 versus β_3. The relationship between β_1 versus β_3 can be modeled linearly, as the relationship between β_1 and β_2 is linear. The parameters' interaction models can be deduced and validated using another statistical technique (which is presented in the next section).

6.3. Model application using Cholesky decomposition

The Cholesky decomposition, described earlier, was used to capture the stochastic nature of the model coefficients while capturing the model coefficient correlations. The correlation coefficients were computed as $\rho_{12} = -0.79$; $\rho_{13} = -0.12$ and $\rho_{23} = -0.34$ with probabilities < 0.001, demonstrating that all coefficients are correlated. Table 6 provides the standard deviation and various parameters for the model coefficients as derived from the Cholesky decomposition.

Comparing these statistics to those of the bootstrap results in Fig. 6 reveals that the Cholesky approach is valid and can be used for BA data. Again, the inter-dependence of the generated 10,000 replications was plotted over those of the Cholesky model (shown in Fig. 7), and both show similar results.

7. Conclusions and recommendations for further research

The contribution of this research is threefold. First, it proposes and provides standards to validate APC and AVL data. Second, it develops two

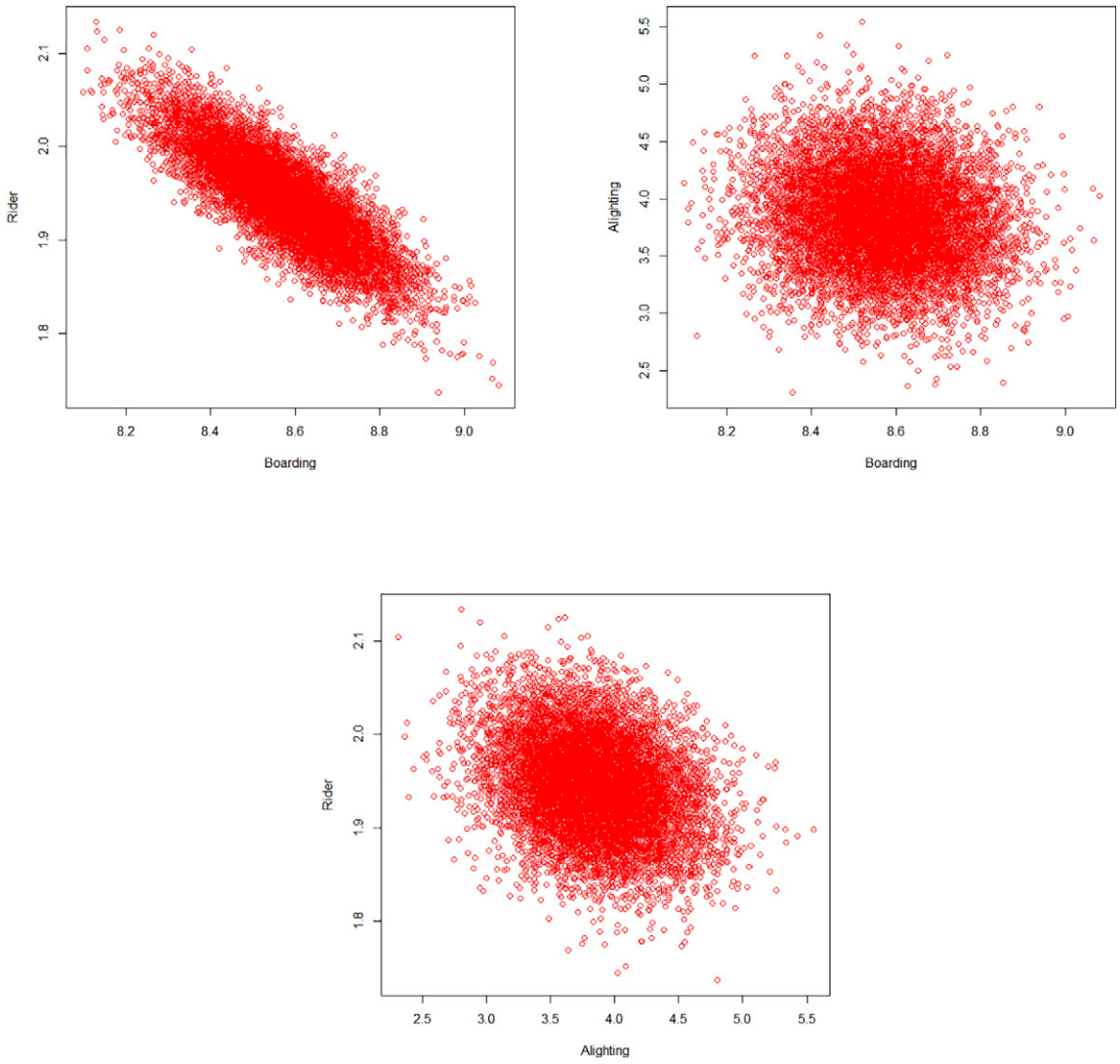


Fig. 6. Scatterplots of model coefficient interdependencies.

statistical boarding/alighting time models using a total of 8341 AVL and APC empirical observations from King County Metro in Seattle, Washington. Third, it identifies the significant factors to estimate the boarding/alighting times and develops an approach to capture the stochasticity in the boarding/alighting times. The primary objective of this research was to develop two models that can predict boarding/alighting times as a function of boarding, on-board and alighting passengers.

The first model is a frequentist linear model and the second model is a stochastic model. Several models were tested and compared based on different statistical metrics including residual errors, goodness-of-fit, AIC and BIC. The best model was selected based on its ability to express the variation in the data with the lowest residual errors.

Table 6
Summary statistics of the Cholesky decomposition model parameters.

Parameter	Mean (?tf=□)	Quantiles		St. Dev. (σ)	Skewness (γ_3)	Kurtosis (γ_4)
		$Q_{0.025}$	$Q_{0.975}$			
β_1	8.570	8.072	8.665	0.141	−0.015	−0.079
β_2	1.947	0.392	2.213	0.041	−0.011	0.054
β_3	3.843	3.766	3.856	0.019	0.028	−0.015

The stochastic bootstrapping technique showed consistency with the frequentist model with the addition of stochastic model coefficients, which can be utilized to conduct stochastic analyses. A key drawback of using the bootstrapping technique is that it requires large computational resources to store the thousands of model parameter realizations. As this study demonstrates, Cholesky decomposition is capable of capturing the desired stochasticity with the use of only $3 \times n$ parameters (mean, variance, and correlations), where n is the number of model coefficients. The Cholesky decomposition is demonstrated to yield results similar to that of the bootstrapping technique, with the use of minimum computational resources.

Inclusion of the boarding/alighting time models are important because they can be used to better estimate bus dwell times and bus travel times to enhance transit system operations. The stochastic boarding/alighting time can be used in further operational applications such as stochastic estimation of transit travel times and enhancing the estimation of bus arrival times at intersections for use in Transit Signal Priority (TSP) systems to better estimate green truncation and extension times. Finally, the results of this work can be useful for transit agencies to enhance bus operations and provide transit users with better estimates of bus arrival times.

Although the study achieved its objective of modeling bus passenger boarding/alighting times, further research is needed in the following two areas: (1) validating the model on other datasets, and (2) extending the

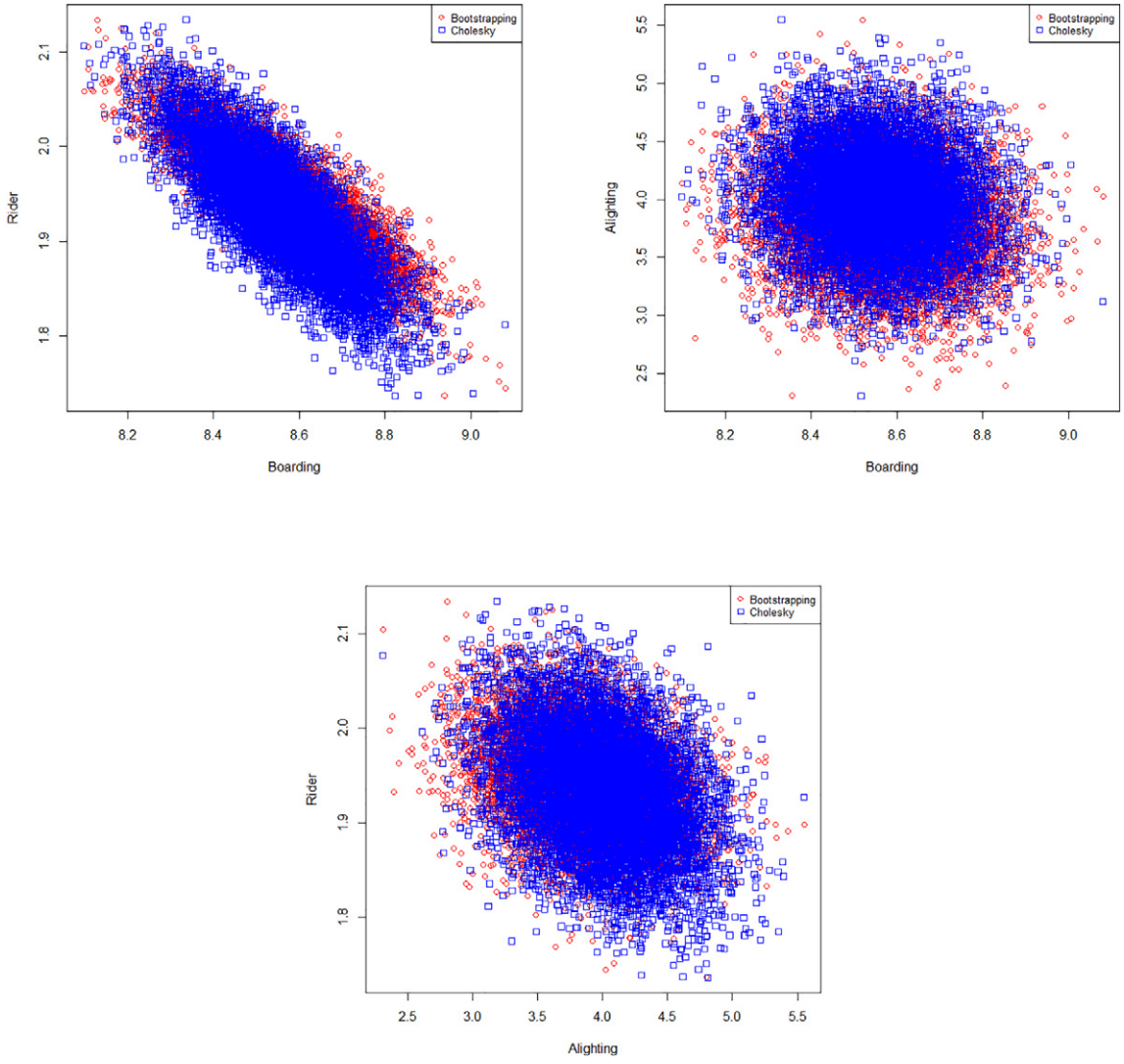


Fig. 7. Interdependence of the Cholesky decomposition model parameters.

model to different bus configurations. In our study the buses had a maximum capacity of 82 passengers (sitting and standing) and had two doors: a front door for boarding passengers and a rear door for alighting passengers. The model can be strengthened if the passenger characteristics (i.e. senior passengers and cyclists) are available and considered. A. Additional research is also needed to develop an evaluation tool for transit agency personnel that would allow for the evaluation BA and dwell times to assist in route restructuring and timing.

Acknowledgements

This effort was funded by the US Department of Transportation through the University Mobility and Equity Center (Award 69A3551747123). Alhadidi is a Jordanian Fellow supported by Al-Ahliyya Amman University (AAU).

Author contribution statement

The authors confirm contribution to the paper as follows: study conception and design: Alhadidi and Rakha; data reduction: Alhadidi; analysis and interpretation of results: Alhadidi and Rakha; draft manuscript preparation: Alhadidi and Rakha. All authors reviewed the results and approved the final version of the manuscript.

References

Bian, B., Zhu, N., Ling, S., Ma, S., 2015. Bus service time estimation model for a curbside bus stop. Transportation Research Part C: Emerging Technologies 57, 103–121.
Chen, M., Yaw, J., Chien, S.I., Liu, X., 2007. Using automatic passenger counter data in bus arrival time prediction. J. Adv. Transp. 41 (3), 267–283.
Fox, J., 2002. An R and S-PLUS Companion to Applied Regression. Bootstrapping Regression Models. Sage, Thousand Oaks, CA.
Fricker, J.D., 2011. Bus dwell time analysis using on-board video. TRB 2011 Annual Meeting.
Furth, P.G., Hemily, B., Muller, T.H., Strathman, J.G., 2006. Using Archived AVL-APC Data to Improve Transit Performance and Management. TCRP Report 113. Washington, D.C.
Glick, T.B., Figliozzi, M.A., 2017. Measuring the determinants of bus dwell time new insights and potential biases. Transp. Res. Rec. (1), 109–117.
Guenthner, R.P., Hamat, K., 1988. Transit dwell time under complex fare structure. J. Transp. Eng. 114 (3), 367–379.
Guenthner, R.P., Sinha, K.C., 1983. Modeling bus delays due to passenger boardings and alightings. Transp. Res. Rec. (915), 7–13.
Gurmu, Z.K., Fan, W.D., 2014. Artificial neural network travel time prediction model for buses using only GPS data. J. Public Transp. 17 (2), 3.
Kathuria, A., Parida, M., Sekhar, C., Pathak, M., 2016. Examining bus lost time dynamics for a bus rapid transit station. J. Public Transp. 19 (2), 168–182.
Kay, J.L., 1990. Advanced Traffic Management Systems-An Element of Intelligent Vehicle-Highway Systems. Vehicle Electronics in the 90's: Proceedings of the International Congress on Transportation Electronics. IEEE.
Kraft, W.H., 1975. An Analysis of the Passenger Vehicle Interface of Street Transit Systems With Applications to Design Optimization. New Jersey Institute of Technology (Ph.D).
Levine, J.C., Torng, G.-W., 1994. Dwell-time effects of low-floor bus design. J. Transp. Eng. 120 (6), 914–929.

Levinson, H.S., 1983. Analyzing transit travel time performance. Transp. Res. Rec. 915, 1–6.
Mazloumi, E., Rose, G., Currie, G., Sarvi, M., 2011. An integrated framework to predict bus travel time and its variability using traffic flow data. J. Intell. Transp. Syst. 15 (2), 75–90.
Meng, Q., Qu, X., 2013. Bus dwell time estimation at bus bays: a probabilistic approach. Transportation Research Part C: Emerging Technologies 36, 61–71.
Rajbhandari, R., Chien, S.I., Daniel, J.R., 2003. Estimation of bus dwell times with automatic passenger counter information. Transp. Res. Rec. 1841 (1), 120–127.
Rashidi, S., Ranjitkar, P., 2015. Bus dwell time modeling using gene expression programming. Computer-Aided Civil and Infrastructure Engineering 478–489.
Singh, K., Xie, M., 2008. Bootstrap: A Statistical Method. Rutgers University, USA.
Strathman, J.G., Hopper, J., 1989. An evaluation of automatic passenger counters: validation, sampling, and statistical inference. Transp. Res. Rec. 1308, 69–77.
TRB, 2013a. Transit Capacity and Quality of Service Manual.
TRB, 2013b. Transit Capacity and Quality of Service Manual, Third Edition. The National Academies Press, Washington, DC.
Wang, C., Ye, Z., Wang, Y., Xu, Y., Wang, W., 2016. Modeling bus dwell time and time lost serving stop in China. J. Public Transp. 19 (3), 55–77.

Update

Transportation Research Interdisciplinary Perspectives
Volume 9, Issue , March 2021, Page

DOI: https://doi.org/10.1016/j.trip.2020.100256

Transportation Research Interdisciplinary Perspectives 9 (2021) 100256

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

journal homepage: www.elsevier.com/locate/trip

Erratum regarding missing declaration of competing interest statements in previously published articles

A **Declaration of Competing Interest** statements were not included in the published version of the following articles that appeared in previous issues of Transportation Research Interdisciplinary Perspectives.

The appropriate Declaration/Competing Interest statements, provided by the Authors, are included below.

1. "Evaluation on the coordinated development of air logistics in Beijing-Tianjin-Hebei" [Transportation Research Interdisciplinary Perspectives, 2019; 1: 100034]
2. "Progress or regress on gender equality: The case study of selected transport STEM careers and their vocational education and training in Japan" [Transportation Research Interdisciplinary Perspectives, 2019; 1: 100009]
3. "What passengers really want: Assessing the value of rail innovation to improve experiences" [Transportation Research Interdisciplinary Perspectives, 2019; 1: 100014]
4. "Audio on the go: The effect of audio cues on memory in driving" [Transportation Research Interdisciplinary Perspectives, 2019; 1: 100004]
5. "Michigan's public transportation: An application of statewide performance assessment and management" [Transportation Research Interdisciplinary Perspectives, 2019; 1: 100013]
6. "Planning for walking and cycling in an autonomous-vehicle future" [Transportation Research Interdisciplinary Perspectives, 2019; 1: 100012]
7. "Are you going to get a ticket or a warning for speeding? An autologistic regression analysis in Burlington, VT" [Transportation Research Interdisciplinary Perspectives, 2019; 1: 100001]
8. "Physical activity of electric bicycle users compared to conventional bicycle users and non-cyclists: Insights based on health and transport data from an online survey in seven European cities" [Transportation Research Interdisciplinary Perspectives, 2019; 1: 100017]
9. "An informal transportation as a feeder of the rapid transit system. Spatial analysis of the e-bike taxi service in Shenzhen, China" [Transportation Research Interdisciplinary Perspectives, 2019; 1: 100002]
10. "Modeling bus passenger boarding/alighting times: A stochastic approach" [Transportation Research Interdisciplinary Perspectives, 2019; 2: 100027]
11. "Disaggregation of aggregate GPS-based cycling data – How to enrich commercial cycling data sets for detailed cycling behaviour analysis" [Transportation Research Interdisciplinary Perspectives, 2019; 2: 100041]
12. "Reducing global warming by airline contrail avoidance: A case study of annual benefits for the contiguous United States" [Transportation Research Interdisciplinary Perspectives, 2019; 2: 100033]
13. "Effect of social capital on the life satisfaction of paratransit drivers in Sri Lanka" [Transportation Research Interdisciplinary Perspectives, 2019; 2: 100050]
14. "To drive or not to drive? A qualitative comparison of car ownership and transport experiences in London and Singapore" [Transportation Research Interdisciplinary Perspectives, 2019; 2: 100030]
15. "Incorporating systems thinking approach in a multilevel framework for human-centered crash analysis" [Transportation Research Interdisciplinary Perspectives, 2019; 2: 100031]

DOI of original article: https://doi.org/10.1016/j.trip.2019.100027; https://doi.org/10.1016/j.trip.2019.100041; https://doi.org/10.1016/j.trip.2019.100050; https://doi.org/10.1016/j.trip.2019.100030; https://doi.org/10.1016/j.trip.2019.100033; https://doi.org/10.1016/j.trip.2019.100031; https://doi.org/10.1016/j.trip.2019.100009; https://doi.org/10.1016/j.trip.2019.100014; https://doi.org/10.1016/j.trip.2019.100012; https://doi.org/10.1016/j.trip.2019.100017; https://doi.org/10.1016/j.trip.2019.100013; https://doi.org/10.1016/j.trip.2019.100034; https://doi.org/10.1016/j.trip.2019.100001; https://doi.org/10.1016/j.trip.2019.100002; https://doi.org/10.1016/j.trip.2019.100004;

https://doi.org/10.1016/j.trip.2020.100256
Available online 2 January 2021
2590-1982/