Approximation and Short-Term Prediction of Bus Dwell Time using AVL Data

Soroush RASHIDI[a], Prakash RANJITKAR[b]

[a,b] *Department of Civil and Environmental Engineering, University of Auckland, Auckland 1142, NEW ZEALAND*
[a]*Email: sras029@aucklanduni.ac.nz*
[b]*Email: p.ranjitkar@auckland.ac.nz*

Abstract: A significant proportion of bus travel time is contributed by bus dwell time for passenger boarding and alighting. This paper reports an investigation conducted on approximation and short-term prediction of bus dwell time based on historical AVL data collected from selected bus routes in Auckland, New Zealand. Three distribution functions including normal, lognormal and Wakeby distribution functions were assessed to approximate the distribution of bus dwell time. Autoregressive Integrated Moving Average (ARIMA) is assessed for the first time to make a short-term prediction of bus dwell time. Wakeby distribution outperformed the most commonly used distribution function namely lognormal distribution to approximate the dwell time for both peak and off-peak periods while ARIMA performed reasonably well for a short-term prediction of the dwell time.

Keywords: bus dwell time, AVL data, Wakeby distribution, ARIMA Model

1. INTRODUCTION

A great proportion of bus travel time is contributed by bus dwell time for passengers boarding and alighting. Highway Capacity Manual (2000) defines bus dwell time as the amount of time a bus spends whilst stopped to serve passengers. Levinson (1983) conducted a cross section study of U.S. cities and stated that nearly 26% of total bus travel time is contributed exclusively by bus dwell time. The effect of bus dwell time on vehicle bunching is generally proven, which may cause variation in headway; leading to an increase in the number of passengers waiting at bus stops; affecting capacity of the transit system [TRB, 2003].

The majority of research conducted on distribution of bus dwell time proposed lognormal distribution function as the best fit approximation [Rajbhandari et al., 2003; Guenther and Hamat, 1988]. However, Koshy and Arasan (2005) recommended that bus dwell time might be normally distributed under heterogeneous traffic conditions. Normal distribution with only two parameters i.e. mean and standard deviation is a quite useful function to demonstrate the distribution of error and residual in regression analysis, however it might not be appropriate for the approximation of bus dwell time due to its negative infinity margin [Neter et al., 1990]. The non-negative feature of lognormal distribution function is quite useful to approximate some non-negative human behaviour related parameters, such as drivers' reaction time and sensitivity parameters [Ranjitkar et al., 2005; Ranjitkar et al., 2010]. However, the interpretation of data is more difficult as its parameters are not in the scale of the original data due to the logarithmic transformation.

A great deal of research works conducted in the past on bus dwell time modelling is based on regression analysis [Levinson, 1983; Rajbhandari et al., 2003; Guenther and Sinha, 1983]. Due to technical as well as monetary constraints, it might not be always feasible to

collect all the data required for regression analysis to produce reasonable results. Besides, the factors affecting bus dwell time may vary between different bus routes and locations of bus stops.

In this paper, we assessed Wakeby distribution for the first time to approximate the distribution of bus dwell time based on historical AVL data collected from selected bus routes in Auckland, New Zealand. With relatively larger number of parameters used in Wakeby distribution function it is more flexible and hence has potential to improve the accuracy of bus dwell time approximation. We assessed Autoregressive Integrated Moving Average (ARIMA) a time series based method also for the first time to make a short-term prediction of bus dwell time. The main advantage of a time series based methods is that it requires only historical time series data to make predictions, which can be collected using AVL system. A brief description of Wakeby distribution function and ARIMA is presented in the following section followed by a description of test bed and AVL data used in this study in the next section. Then results are presented in section four under two subheadings: distribution analysis and short-term prediction. Finally, the outcomes of this paper are summarized in the last section.

2. LITERATURE REVIEW

2.1 Wakeby Distribution

Wakeby distribution is named after Wakeby Pond in Cape Cod, Massachusetts; invented by H.A. Thomas in 1976 and introduced by Houghton in 1978 to model flood flows. The model formulation can be expressed as follows:

$$X = \xi + \frac{\alpha(1-(1-U)^{\beta})}{\beta} - \frac{\gamma(1-(1-U)^{-\delta})}{\delta} \tag{1}$$

where,

X is a quintile function for variable U
U is a standard uniform random variable (0, 1).
β, δ are shape parameters;
α, γ are scale parameters and
ξ is a location parameter.

Here β, δ, α and γ are always positive real numbers while ξ could be any real number. Some special features of Wakeby distribution as mentioned by Hosking and Wallis (1997) are as follows:

- Wakeby distribution has more suitable parameters to mimic shapes of many skewed distributions such as lognormal or log-gamma.
- With a considerable increase in the number of parameters to five, Wakeby distribution can approximate a wider range of distribution shapes than other traditional distribution functions such as normal and lognormal distributions.
- Wakeby distribution with heavy upper tail will direct more attention to the data set containing outliers. In bus dwell time data rare events such as lift operation can make the dwell time distribution heavily right-skewed.

A method of L-moments is used in this paper to estimate the parameters of Wakeby distribution function. This method, when compared with the conventional moment and maximum likelihood method, has less error in the estimation of model parameters, is less

sensitive to outliers and, can describe a wider range of distribution coverage. Moreover, it is generally more accurate than maximum likelihood method for small sample sizes [12]. Among non-parametric tests, Kolmogorov-Smirnov (K-S) test, Chi-Square (C-S) test and Anderson Darling (A-D) test are the most widely used Goodness-Of-Fit (GOF) test. We used K-S GOF test as recommended by Frank and Masssey (1951) to assess how well the three distribution functions perform including normal, lognormal and Wakeby distribution functions.

2.2 ARIMA

ARIMA model also known as Box-Jenkins model consists of three parts autoregressive, moving average and differencing. ARIMA models are generally represented as ARIMA (p, d, q); where the first digit (p) represents the number of time steps that the model looks back to find the best correlation. The second digit (d) represents the degree of differencing and the third digit (q) represents the number of points used for moving average calculation. It is expressed as follows:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \alpha_t - \theta_1 \alpha_{t-1} - \cdots - \theta_q \alpha_{t-q} \quad (2)$$

The correlation between the data point at time t (y_t) and previous time steps (y_{t-1}, y_{t-2},…, y_{t-n}) is an important consideration in this model. That is any given value of y_t is a linear function of its previous value y_{t-1} and random error in the past time α_{t-1}, plus error α_t which is assumed to follow a normal distribution. More detailed information on the ARIMA method can be found in [Pankratz, 1983].

Model Estimation and Diagnosis: Unlike regression analysis which assumes that any individual observations are independent of each other and errors in their measurements are not related, time series based methods considers correlation between individual observations made at different times. This hypothesis in time series is termed as autocorrelation. Partial autocorrelation is a partial correlation between time series variables y_t and y_{t-l}, which takes into account any lags less than L to remove its effect [McCleary and Hay, 1980].

The importance of autocorrelation and partial autocorrelation functions in time series analysis is in the estimation of model parameters. If ACF or PACF is significantly different from zero in some lags, then they should be taken into account in the model parameter estimation. In general, the shape of ACF and PACF and their cut off from their boundary will give the broad understanding of the correlation in data points. Pankratz (1983) recommended that the absolute value of *t* for autocorrelations for the first three lags should be less than 1.25 and the rest should be less than 1.60. The portmanteau test can be used to find out whether there are any significant autocorrelation in the model, which can be expressed as follows:

$$Q_{LB} = N(N+2)\sum_{L=1}^{L}\left(\frac{1}{N-L}\right)r_L^2 \quad (3)$$

where,

N is the sample size;

r_L^2 is the sample autocorrelation at lag L for the residual of estimated model; and

L is number of lags.

Model Accuracy: A number of measures are proposed in literatures to evaluate accuracy of a

model, which can be categorized in two groups: scale-dependent and non-scale-dependent [Hyndman and Koehler, 2006]. A scale dependent error has the same scale that of the data. Mean Square Error (MSE) is one of the most commonly used scale-dependent approaches. Mean Absolute Percentage Error (MAPE) as a non-scale-dependent performance measure is used in this study due to its ease of interpretation of error measure. MAPE can be expressed as follows:

$$MAPE = \frac{\sum_{t=1}^{n} \frac{|A_t - F_t|}{A_t} \times 100}{n} \quad (4)$$

where,

A is the actual value; and
F is the forecasted value.

3. TEST BED

Bus stops along a major arterial road in Auckland CBD are selected as test bed for this study. The AVL data was provided by Auckland Regional Transport Authority then, which is now merged into Auckland Transport. It is a signpost type of AVL system where each signpost has its own unique code, which detects the buses within its range of influence; records location, time and identifier number of buses; and sends this information to the dispatcher through a transmitter.

The data was separated in two categories for analysis: one for distribution analysis and the other for short-term prediction. Three months data starting from 2 March 2010 to 28 May 2010 was used for distribution analysis, which was split into three parts: early morning off peak from 6:00AM to 7:30AM), morning peak from 7:30AM to 9:30AM) and late morning off peak period from 9:30AM to 12:00AM. In the second category, the data from 2 March 2010 to 26 March 2010, which include four days a week data from Tuesday to Friday during the morning peak periods starting from 7:30 AM to 9:30 AM was used. Nearly 20% of the second category data was used for validation purpose while the rest (80%) was used for model estimation purpose. The model performance was evaluated based on validation results rather than calibration results as suggested in the literatures.

The data for four bus stops near Auckland CBD were investigated for the distribution analysis, which include stop number 159, 160, 5118 and 5117 as seen in figure 1a. While for short-term prediction only a single stop numbered 159 was investigated. It shall be noted that these stops generally have high number of boarding and alighting passengers as they are located in Auckland CBD and also close to city campus of two large universities namely the University of Auckland and Auckland University of Technology. Each bus stop serves several bus routes along Symonds Street. For distribution analysis, we investigated the data from a single bus route number 2742 for bus stops numbering 159, 160, and 5118 while for bus stop number 5117 the data from bus route number 2242 was investigated. For short-term prediction, the data from all bus routes using bus stop number 159 was combined.

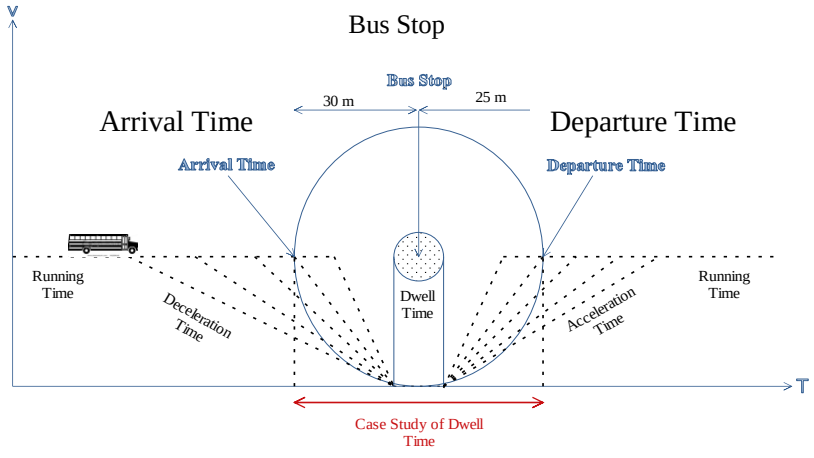
All public transport buses in Auckland have lowered floors and passengers can board or alight from the front door, whereas, the rear door is used only for alighting. Bus tickets can be bought on-board from the operator or alternatively relatively faster “Go and Ride” card, an electronic payment method can be used. We computed bus dwell time from the arrival and departure times recorded by the AVL system, which can be expressed as follows:

$$DT = ART - DET \tag{5}$$

where,

DT is dwell time;
ART is arrival time; and
DET is departure time.

(a)



(b)

Figure 1. (a) Location map of Symonds Street bus stops; (b) A conceptual diagram of dwell time computed from AVL data

The arrival time is recorded when a bus enters the circle and departure time is logged as the bus leaves the circle. It shall be noted that the dwell time used in this study represents the time spent by buses while they are within a predefined circle of influence of the AVL system, which can be approximately 55 m in diameter as shown in figure 1b. Hence, it may include the time taken for decelerating to come to stop at the bus stop, accelerating when departing from the bus stop as well as the dwell time itself for boarding and alighting passengers while stopping at the bus stop.

4. RESULTS AND DISCUSSION

The AVL data used in this study contains the dwell time with and without stopping at the bus stops. The first step of data processing is to separate the dwell time data with stopping from those without stopping as the latter one does not include bus dwell time itself. For this purpose, we used frequency polygon method to demonstrate visually the number of observations in a set of intervals plotted for each stop. Figure 2 presents the results from this analysis where at 15 seconds time duration a clear shift can be easily observed; shown by dotted circles for each stop, which confirms that the data points lower than this value is significantly different from those over this value.

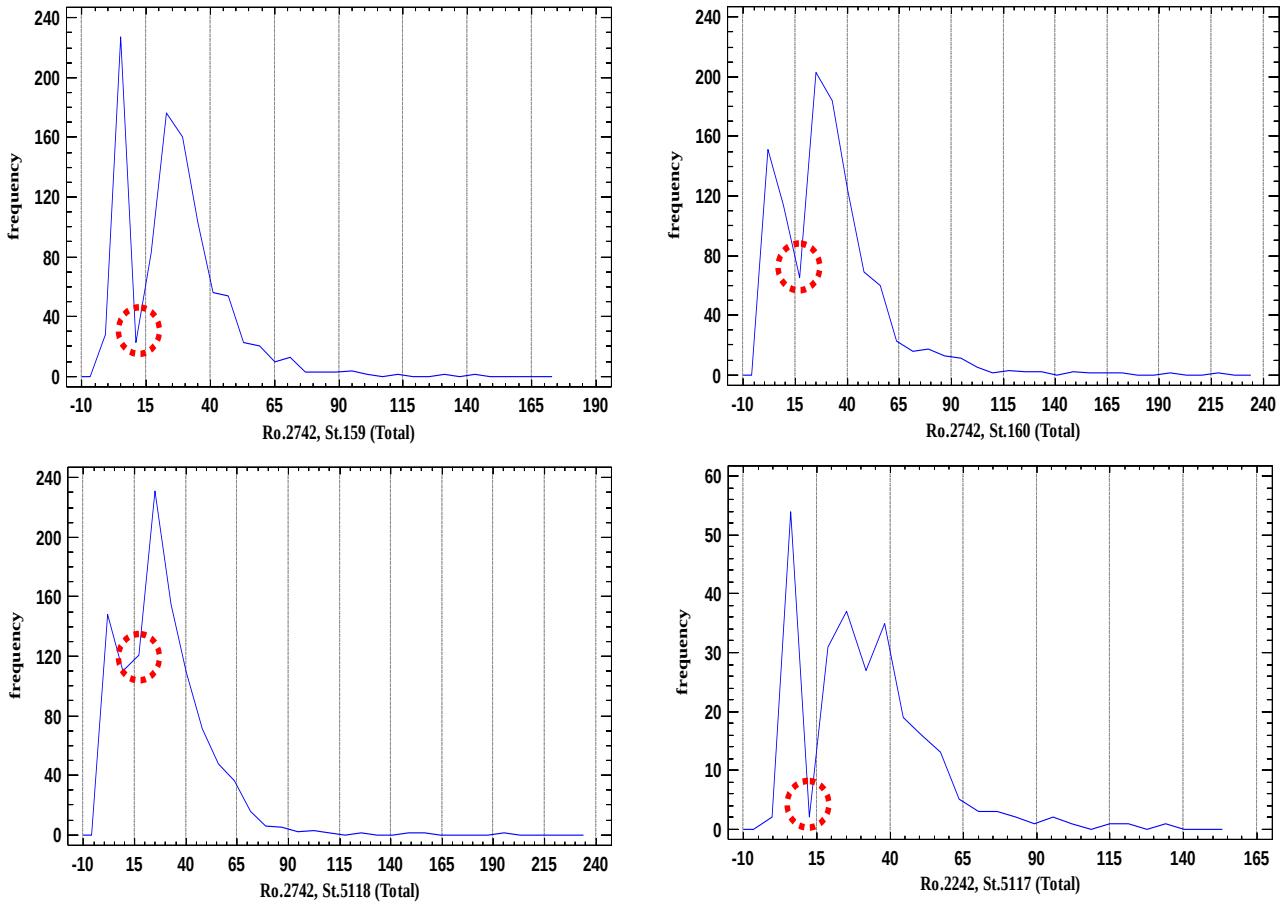


Figure 2. Frequency polygon results to separate data points with and without stopping at the bus stops

4.1 Distribution Analysis

Figure 3 presents distribution of the bus dwell time computed from the AVL data along with the respective probability distribution functions for normal, lognormal and Wakeby distributions for the bus stop number 159. It can be observed in figure 3(a) that Wakeby distribution function gives the closest approximation to the data particularly close to the peak points demonstrating a better distribution fit than normal and lognormal distribution functions. Lognormal distribution function also performed well compared to normal distribution. Similar trends can be observed in figures 3(b) and (c) for Wakeby distribution. While in figure 3(c) a big difference can be observed between the approximated values by

lognormal distribution and the field data near the peak point, demonstrating its failure to approximate the dwell time data.

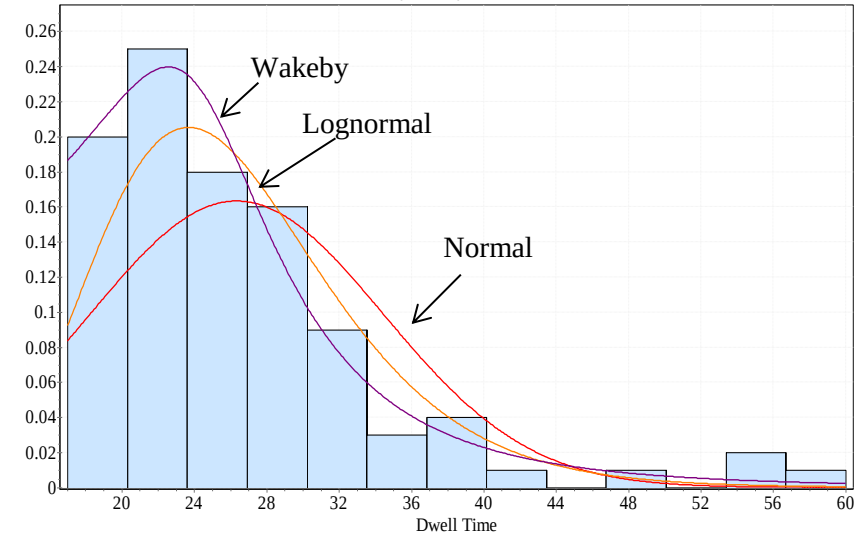
Table 1 presents results from K-S GOF test. The un-shaded cells represent the rejected hypothesis while shaded cells represent the accepted hypothesis. All 12 out of 12 cases accepted the hypothesis for Wakeby distribution as the respective K-S statistic values remained lower than their critical values representing a close fit in all cases. For lognormal distribution, 1 case was rejected at bus stop 159 for late morning peak period out of 12 cases tested. In that particular case, the K-S statistic value of 0.086 is higher than its critical value of 0.074. For normal distribution, the hypothesis was rejected for 10 out of 12 cases tested. Two cases where the hypothesis was accepted include bus stop 5117 for early morning and late off peak period. This could be due to low stopover in this stop in the off peak period. Based on these results, it can be concluded that Wakeby distribution outperforms the other two distribution functions for this test bed. Lognormal distribution also performed satisfactorily while normal distribution is not suitable to approximate the dwell time for this test bed.

4.2 Short-Term Prediction

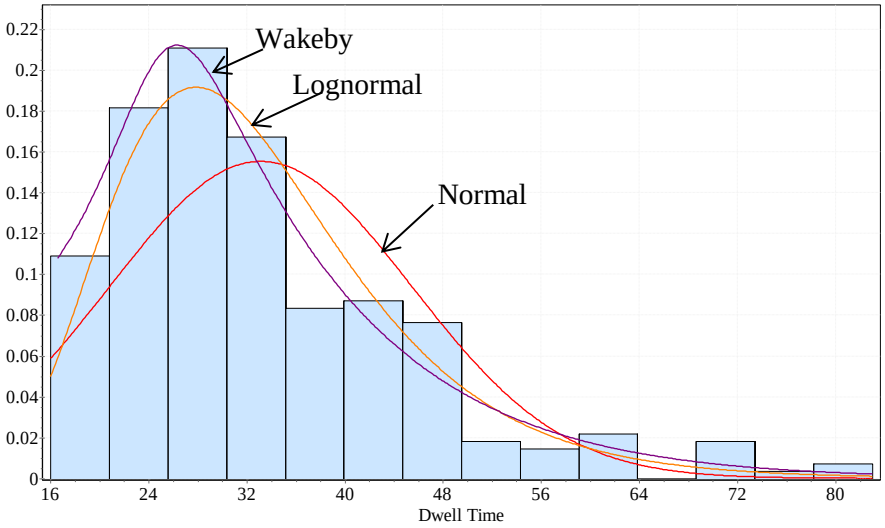
Time series analysis requires having equally spaced time interval. To make the bus dwell time equally spaced we groped the bus dwell time data into several equal 15 minutes time intervals. In this study, we used short term as we have chosen rather small time window to model and predict the bus dwell time. Less demanding bus stops should have bigger time interval. For example bus dwell time at stops which are located at routes with 30 minute bus headway should be grouped into several equal 30 minutes or more time intervals. Therefore, the choice of time interval is purely depends on several factors like the quality of the data, frequency and headway of the bus route, time of the day and day of the week.

As a first step to build an ARIMA model, the dwell time data was transformed to stationary series by differencing. A spike in lag one in ACF and a decreasing pattern in PACF may suggest suitability a model of no autoregressive with one moving average order written as (0, 1). In this case if a single order differencing is used to transform the series into the stationary one then the model can be termed as ARIMA (0, 1, 1). We used three different approaches to test the adequacy of the model and choose the best ARIMA model, which include portmanteau test, normality of residual and ACF residual examination as recommended by Pankratz (1983). After several rounds of testing, it was concluded that ARIMA (5, 1, 0) is the best ARIMA model for these data sets. The parameters of the ARIMA model according to problem can be any number, however it is recommended to keep the parameter as small as possible because increasing the parameter of the model will increase its complexity. As it is the first time we apply time series analysis for estimation and prediction of bus dwell time there are no recommendations regarding the value of the parameters. Due to different underling pattern at different bus stop we strongly suggest optimization the parameter of the ARIMA model for other bus stops.

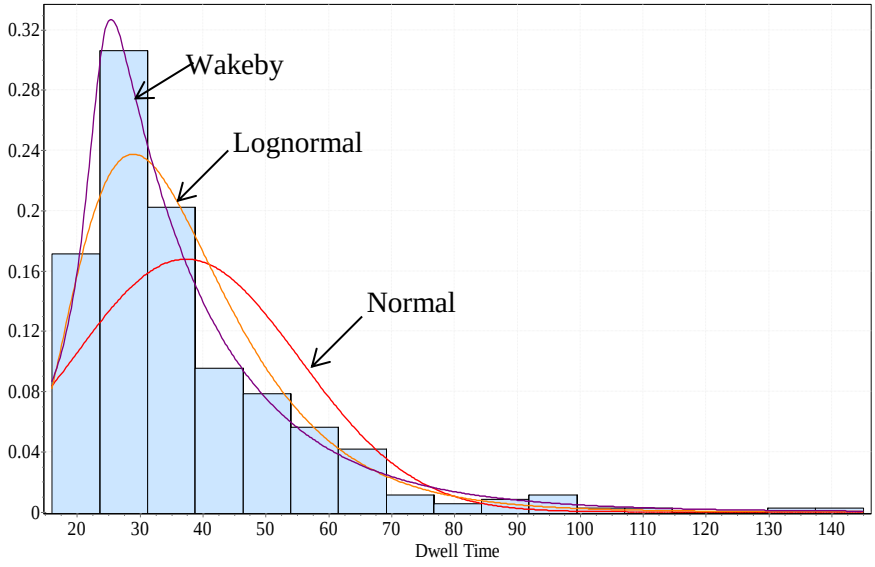
As can be seen in Figure 4 the ARIMA (5, 1, 0) can reasonably well mimic the pattern of data in both estimation and validation periods. Based on the model in estimation and validation period the forecast was made for Tuesday 30 march. A MAPE value of 18.68% was obtained for the ARIMA (5, 1, 0) model, which can be considered as a reasonably acceptable for the bus dwell time prediction. A MAPE value of 10% or less is generally considered excellent; those in a range from 10% to 20% considered as good; while there are a number of cases presented in literatures where MAPE values are in a range of 20%-30% or even higher (IPREDICT IT website). Generally there are two main goals for using time series



(a) 6-7:30AM



(b). 7:30-930 AM



(c) 9:30-12AM

Figure 3. Probability distribution function for Stop 159 at different time of day

Table 1. Comparison of Distribution Fitting

Distribution Functions	*Route*	*Bus Stop*	*K-S (6:00-7:30)*	*Critical Value*	*K-S (7:30-9:30)*	*Critical Value*	*K-S (9:30-12:00)*	*Critical Value*
Wakeby	2742	159	0.06	0.13	0.04	0.08	0.04	0.07
		160	0.09	0.15	0.04	0.08	0.03	0.07
		5118	0.06	0.12	0.04	0.08	0.04	0.07
	2242	5117	0.11	0.28	0.05	0.15	0.08	0.18
Lognormal	2742	159	0.11	0.13	0.07	0.08	0.09	0.07
		160	0.15	0.15	0.07	0.08	0.05	0.07
		5118	0.11	0.12	0.07	0.08	0.07	0.07
	2242	5117	0.16	0.28	0.08	0.15	0.08	0.18
Normal	2742	159	0.16	0.13	0.14	0.08	0.17	0.07
		160	0.19	0.15	0.14	0.08	0.14	0.07
		5118	0.17	0.12	0.10	0.08	0.14	0.07
	2242	5117	0.20	0.28	0.17	0.15	0.11	0.18

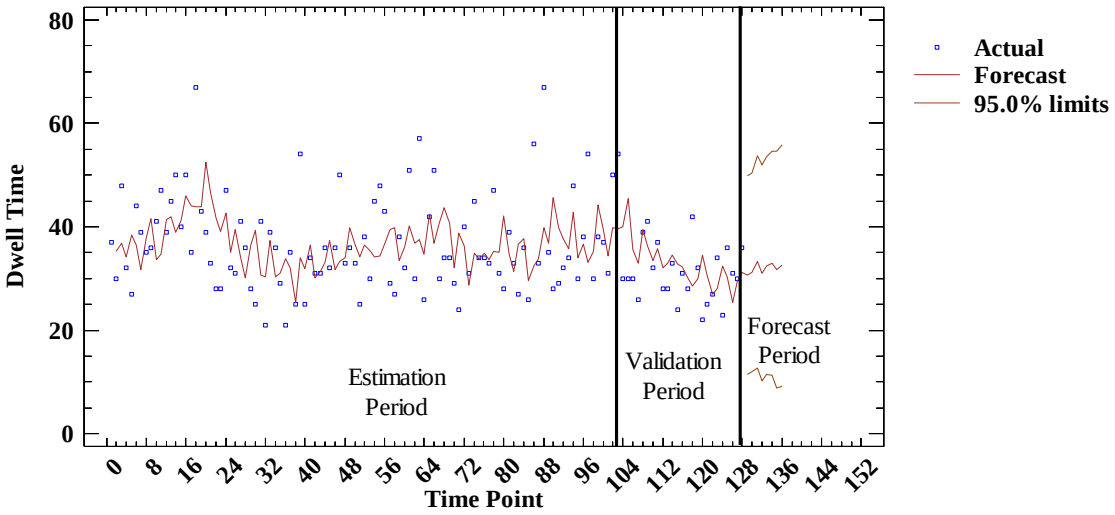


Figure 4. Dwell time modelling and prediction using ARIMA (5, 1, 0)

analysis. The first, understanding the underlying pattern of the data and the second, forecasting based on the model estimation. Proposed ARIMA model can be applied for other bus stop, however due to different underling pattern at different bus stop we strongly suggest try to optimize the parameter of the ARIMA model for other bus stops.

5. CONCLUDING REMARKS

This paper has investigated bus dwell time data collected from selected bus routes in Auckland, New Zealand using AVL system. Three distribution functions are assessed to approximate the distribution of bus dwell time including normal, lognormal and Wakeby distribution functions. Linear regression models have received considerable attention in bus dwell time analysis, mainly for reasons of its well-known theoretical concepts and its availability in almost any statistical packages. Despite all these advantages, most often the linear regression approach requires several different independent variables to explain the variations of dependent variable. In contrast, time series models only require historical time sequence of data which makes these models free of any independent variable. Therefore, in

the absence of number of boarding and alighting people's record, time series analysis can be used to model bus dwell time. In this study, the potential applicability of time series analysis to model and predict bus dwell time investigated. The following outcomes can be drawn from the time series analysis presented in this paper:

- Wakeby distribution outperformed the other two distribution functions for this test bed.
- Lognormal distribution also performed satisfactorily while normal distribution is not suitable to approximate the dwell time for this test bed.
- Time series based prediction models can be implemented to predict the bus dwell time.
- ARIMA (5, 1, 0) performed the best among ARIMA models for this test bed.

REFERENCES

Frank J., Massey, J. (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, Vol. 46, No. 253, p.68-78.

Guenther, R.P., Hamat, K. (1988) Transit Dwell Time under Complex Fare Structure. *Journal of Transportation Engineering*, 1141, p.367-379.

Guenther R.P., Sinha, K.C. (1983) Modelling Bus Delays Due to Passenger Boarding and Alighting. *Transportation Research Record*, 915, p.7-13.

Highway Capacity Manual (2000) Transportation Research Board, National Highway Research Council. Washington, D.C.

Hosking, J.R.M. (1990) L-moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society*, Vol. 52, No.1, p.105-124.

Hosking, J.R.M., Wallis, J.R. (1997) *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, Cambridge, U.K.

Houghton, J.C. (1978) Birth of a Parent: the Wakeby Distribution for Modelling Flood Flows, *Water Resource Research*, 14, p.1105–1110.

Hyndman, R., Koehler, A. (2006) Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, Vol. 22, 2006, p.679-688.

IPREDICT It, Time Series Forecasting, Error Statistics. www.ipredict.it/ErrorStatistics.aspx/. Accessed on Feb. 5, 2013.

Koshy, R.Z., Arasan, V.T. (2005) Influence of Bus Stops on Flow Characteristics of Mixed Traffic. *Journal of Transportation Engineering*, ASCE, p.640-643.

Levinson, H.S. (1983) Analysing Transit Travel Time Performance. *Transportation Research Record*, 915, p.120-127.

McCleary, R., Hay, R.A. (1980) *Applied Time Series Analysis for the Social Sciences*. Beverly Hills, Sage.

Neter, J., Wasserman, W., Kutner, M.H. (1990) *Applied Linear Statistical Models.* 3rd ed. Burr Ridge, IL; Richard D. Irwin, Inc.

Pankratz, A. (1983) *Forecasting with Univariate Box-Jenkins Model*. John Wiley & Sons, Inc., New York.

Rajbhandari, R., Chien, S., Daniel, J. (2003) Estimation of Bus Dwell Times with Automatic Passenger Counter Information. *Transportation Research Record*, 1841, p.120-127.

Ranjitkar, P., Nakatsuji, T., Azuta, Y., Asano. M., Kawamura, A. (2005) A Contemporary Reassessment of GM Car-Following Model Using RTK GPS Data. *Proceedings of Japan Society of Civil Engineers*, Vol. 793, p. 121.

Ranjitkar, P., Nakatsuji, T., (2010) Analysing Transit Travel Time Performance. *Proceedings of 89th Annual Meeting of Transportation Research Board*, Washington, D.C.

TRB and Kittelson & Associates, Inc. (2003) TCRP Report 100: *Transit Capacity and Quality of Service Manual, Second Edition.* TRB.