INTRODUCTION

In recent years the traditional objective of improving road network efficiency is being supplemented by greater emphasis on safety, incident detection management, driver information, better car park utilisation, environmental issues and the provision of priority to public transport and pedestrians. In line with this increasing interest in *Intelligent* or *Smart* transport systems (ITS), the current and future trend is moving towards proactive systems. The essential component of a proactive system is short term prediction of traffic flow. The predicted traffic flow then becomes the input to many integrated applications such as proactive traffic control system, travel information system, dynamic route guidance system and incident management system. Without a predictive capability, ITS can only provide services in a reactive manner.

The objective of this paper is to review techniques that are used to predict travel flow using field data from Melbourne's freeway to evaluate their robustness and accuracy of the techniques. The techniques used in the evaluation are:

- Regression,
- Historical average,
- ARIMA (Auto Regressive Integrated Moving Average), and
- SARIMA (Seasonal Auto Regressive Integrated Moving Average).

REVIEW OF SHORT-TERM TRAFFIC FLOW PREDICTION TECHNIQUES

Regression

Regression analysis is a statistical technique that is often applied when some relationship is presumed to exist between a single dependent variable and one or more independent variables. The objective of regression analysis is to determine (ie. predict) the expected value of a dependent variable in response to changes in one or more independent variables. The typical form of a multi-variate regression equation with *n* number of independent variables is as follows:

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

where:

y is the predicted value of the dependent variable;

a is the y-intercept; and

b_i is the coefficient assigned to the independent variable, x_i.

Historical Average

The historical average model uses an average of past traffic flows to forecast the future traffic flow.

$$q(t+1) = q_h(t+1)$$

The basic premise behind the historical data based algorithm is that traffic patterns are seasonal. In other words, a knowledge of *typical* traffic conditions on Tuesday at 5:30pm will allow one to predict the conditions on any particular Tuesday at 5:30pm.

Using historical data helps capture the shape of the traffic flow pattern, including the degree of peaking and its starting and finishing times.

Various refinements of historical average have been made to enhance the predictive accuracy of this technique. One example is the use of linear scaling shown in the equation below.

$$q(t+1) = q_h(t+1) + k\,[q_h(t) - q(t)]$$

where:

$q_h(t+1)$ is the historical average flow at time t+1

$q(t+1)$ is the predicted flow at time t+1, and

k is a constant.

Although, these algorithms perform reasonably well during normal operating conditions, they do not respond well to external changes in the system such as weather, special events, or modified traffic control strategies

ARIMA (Auto Regressive Integrated Moving Average)

ARIMA (Auto Regressive Integrated Moving Average) is a statistical based method of the time-series analysis popularised by Box and Jenkins (1970) in the early 1970s. The ARIMA model is based on the premise that the knowledge of past values in a time series is the best predictor of the variable in question. In other words, the ARIMA model can produce accurate short term forecasts based on a synthesis of historical patterns in data and does not assume any pattern in the historical data of the time series.

A non-seasonal ARIMA model *ARIMA(p,d,q)* refers to the *p* degree of the AR process, *d* degree of the I component and *q* degree of the MA process. The number of p, d and q terms start from 0.

The autoregressive (AR) term is the self deterministic part of the series and is simply the time-lagged values of the forecast variable, expressed in the form:

$$Y_t = c + (\phi_1 B + \phi_2 B^2 + \ldots + \phi_p B^p)\; Y_t \; + e_t$$

$$Y_t = c + \phi_1\, Y_{t-1} \; + \phi_2\, Y_{t-2} \; + \ldots + \phi_p\, Y_{t-p} + e_t$$

where

B is the backward shift operator, $BY_t = Y_{t-1}$

e_t is the error term that represents random event not explained by the model,

p is the number of AR terms.

$\phi_1 \ldots \phi_p$ are the autoregressive coefficients, and c is a constant.

The integrated (I) term refers to the differencing of the data series to make the series stationary. A stationary series means that the data fluctuate around a constant mean, independent of time, and the variance of the fluctuations remains essentially constant over time. By allowing differencing of the data series, the ARMA model can be extended to non-stationary series and is said to be an "integrated" version of a stationary series. The d degree of differencing I(d) can be expressed in the form:

$$(1 - B)^d \, Y_t = c + e_t$$

where

d is the number of non-seasonal differences.

For first difference, I(1):

$$(1 - B) \, Y_t = c + e_t$$

$$Y_t - Y_{t-1} = c + e_t$$

It is often convenient to redefine the first difference series $Y_t^{'}$ as $Y_t - Y_{t-1}$. If the first difference does not convert the series to a stationary form, then the first difference of the first difference (*second order differencing*) can be created.

Second order difference: $(1 - B)^2 \, Y_t = c + e_t$

$$Y_t = 2Y_{t-1} - Y_{t-2} + c + e_t$$

$$Y_t^{''} = Y_t - 2Y_{t-1} + Y_{t-2}$$

Note a distinction between second order difference, $Y_t^{''}$, defined above and a second difference $(Y_t - Y_{t-2})$.

The general equation for the ARIMA(p,d,q) model can be written as:

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) \, (1-B)^d \, Y_t = c + (1-\theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) \, e_t$$

The moving average (MA) term is the disturbance component of the series and is a moving average of the successive error terms, expressed in the form:

$$Y_t = c + (1 \; \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) \, e_t$$

$$Y_t = c + e_t - \theta_1 \, e_{t-1} - \theta_2 \, e_{t-2} - \ldots - \theta_q \, e_{t-q}$$

where

q is the number of MA terms,

$\theta_1 \ldots \theta_q$ are the moving average coefficients, and

$e_{t-1} \ldots e_{t-q}$ are previous values of residuals.

Seasonal ARIMA (SARIMA)

The ARIMA models that have been discussed so far are linear functions of the most recent few observations. If the time series is seasonal (ie. a series with a pattern that repeats itself over fixed intervals of time), a seasonal ARIMA can be applied to handle the seasonal aspects of the time series.

The general notation for seasonal ARIMA model is: ARIMA(p,d,q)(P,D,Q)$_s$

where:

(p,d,q) is the non-seasonal part of the model explained above,

(P,D,Q)s is the seasonal part of the model,

P is the seasonal autoregressive (SAR) terms,

D is the seasonal differences,

Q is the seasonal moving average (SMA) terms, and

s is the number of periods per season, for example for a monthly series with a pattern that repeats itself year after year, s=12.

The equation for the ARIMA(p,d,q)(P,D,Q)s model is

$$(1-\phi_1 B\ldots-\phi_p B^p)(1-\Phi_1 B^s\ldots-\Phi_P B^{s+P-1})(1-B)^d(1-B^s)^D Y_t = c+(1-\theta_1 B\ldots-\theta_q B^q)(1-\Theta_1 B^s\ldots-\Theta_Q B^{s+Q-1})\, e_t$$

where

$\Phi_1\ldots\Phi_P$ are the seasonal autoregressive coefficients, and

$\Theta_1\ldots\Theta_Q$ are the seasonal moving average coefficients.

CASE STUDY

Site Description

This study focuses on the inbound section of Melbourne's Eastern Freeway between Doncaster Road and Hoddle St (see *Figure 1*). The data used in this study was obtained from VicRoads raw traffic data collected on freeways in Melbourne. Inductive loop detectors, placed every 500m along each lane of the freeway, were used to measure speed, flow and lane occupancy data every 20 seconds. Data from a select number of detectors were used in the study.

Software was developed to extract the raw data collected on Eastern Freeway and aggregate the data into specified time intervals (eg. 15 minute volume). The software has a compliance factor feature that allows "incomplete" data to be scaled up. For example aggregating 20 seconds data into 15 minute intervals (ie. 60 data points) with a compliance factor of 95% would allow data sets with greater or equal to 54 data points, to be scaled up to a full 15 minutes data. This feature increases the amount of useable data and also gives the flexibility to specify the level of tolerance required for the analysis.

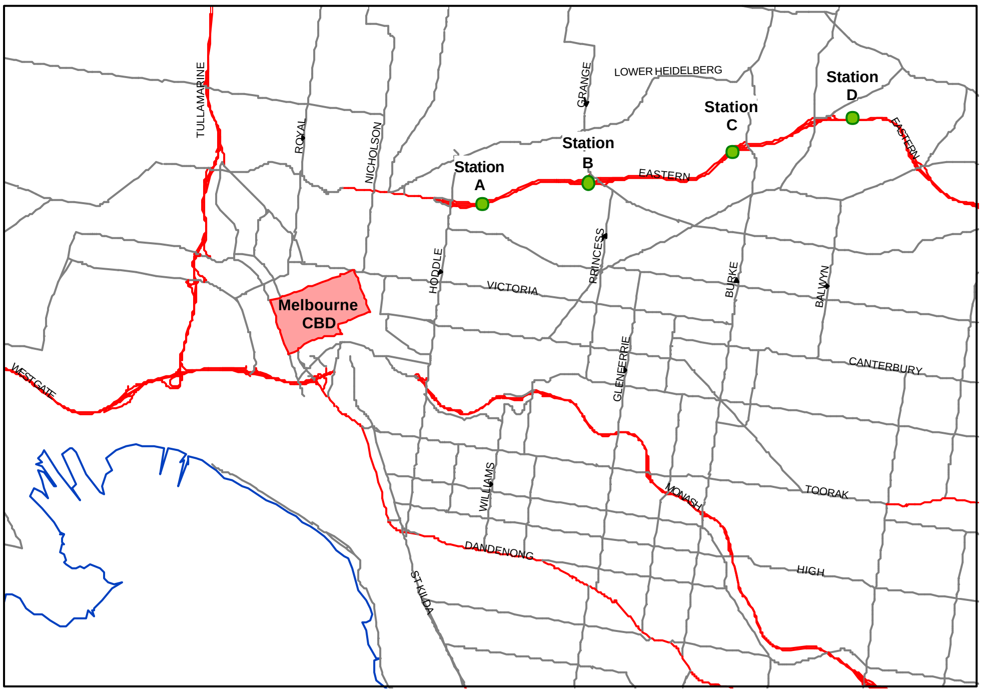

Figure 1 – Eastern Freeway Site

Data Analysis

Becoming familiar with a data set is advisable before applying any prediction techniques. Without a thorough examination of the data the performance of any candidate technique can be undermined. By studying the original data set one can minimise the impact of any outliers or missing data and the potential for violating assumptions (eg. normality), which is important as these effects can be compounded across several variables to produce significant levels of error.

To gain some general understanding of the data's qualitative properties the simplest exercise to perform is to graph the data. Figure 2 presents the flow measured at detector station D between 5 am and 10 pm for one week from Monday 13th July to Sunday 19th July, 1998. This graph indicates the presence of some patterning in the flow by time of day and the repetitive or "seasonal" quality of the data.

A more analytical approach of determining the qualitative characteristics (ie. categorical subgroups) of the data is to undertake a cluster analysis. The objective of this technique is to identify relatively mutually exclusive, homogeneous groups within a sample of entities abased on the similarities between the individuals.

In this study the agglomerative hierarchal clustering technique was applied to the data set. The average linkage between groups clustering method and the squared Euclidean distance measurement were adopted. The cluster analysis was conducted using data measured at the

D detector site during the six weeks between the 19th June to 26th June and 13th July to 11th August, 1998. It was intended that the cluster analysis be conducted on data from typically normal days.

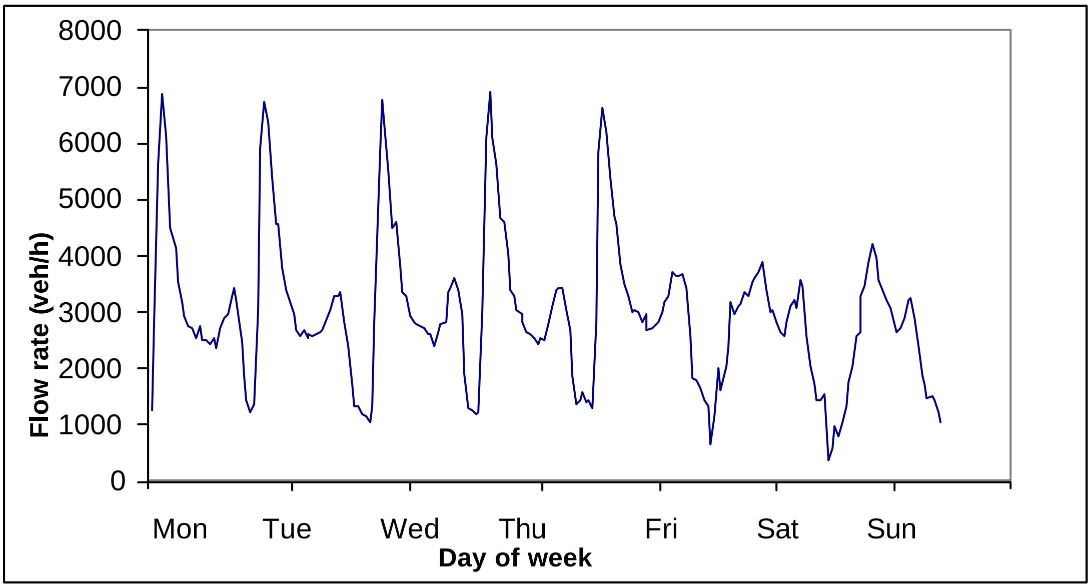

Figure 2 – Flow at detector station D between 13/7/98 – 19/7/98 (from 5am to 10pm)

Data during the school holiday period (between the 27th of June to the 12th of July) were excluded from the analysis. The data were aggregated into 30 minute intervals between 5.30 am and 10pm. The final data set consisted of 1216 samples for each variable: flow, day of week and time of day. Cases were eliminated if any of the three variables contained missing data.

Cluster analysis was undertaken for the weekdays and weekend using the flow and day of week as variables. The cluster solutions show fairly consistent behaviour over the five weekdays, with the time of day (in particular the morning peak) as the main distinguishing characteristic of each cluster solution (see *Figure 4*). Although the weekend was not singled out as a different cluster, Saturday and Sunday clearly exhibit different daily patterns (ie. the absence of extremely high traffic flows during the morning peak period) compared to the other days of the week.

Holiday Period Versus Normal Traffic Flow

Data during the holiday period was also examined. As part of the first step of the data analysis, the traffic flow on selected days during the holiday period was plotted against the same day of the week during the normal period (see *Figure 3*). The traffic flow pattern over the holiday period does not appear to be very different from that of the normal traffic flow pattern. A paired samples t-test was also conducted to confirm this matter. The little difference between the traffic flow patterns indicates that the majority of traffic using the freeway is work-based travel so is largely unaffected by school holidays.

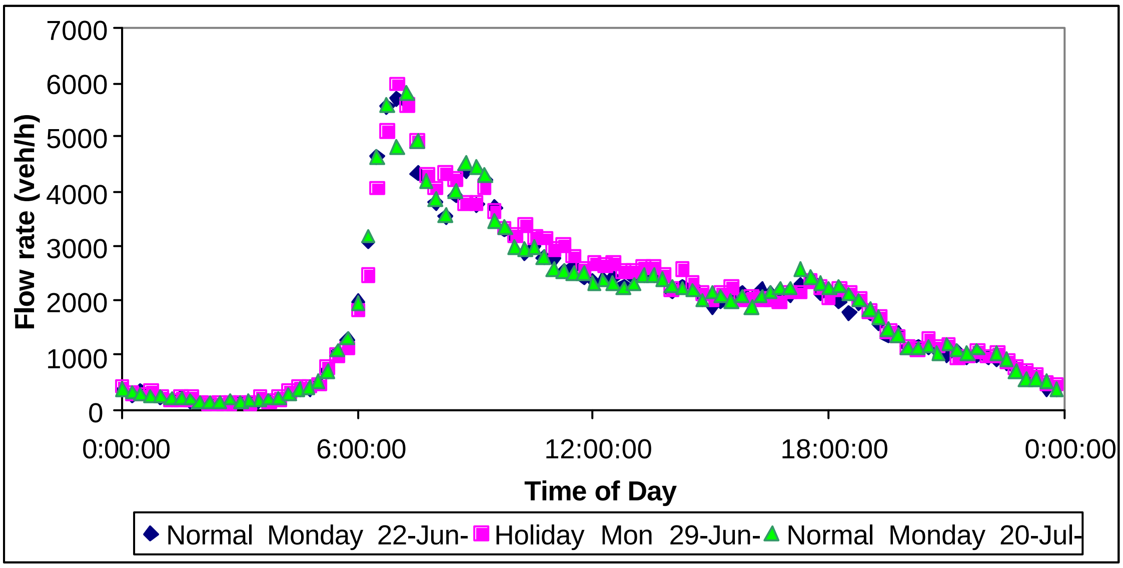

Figure 3– Traffic flow on a Monday during the holiday period and under normal conditions

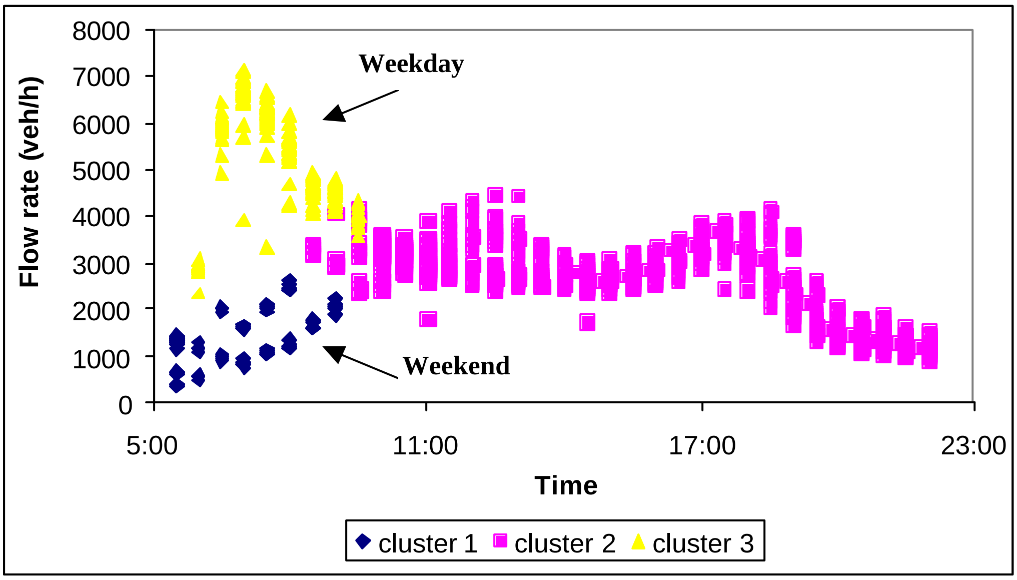

Figure 4 – Flow versus time of day depicting three cluster solution for flow, time of day and day of week variables

MODEL SELECTION AND CALIBRATION

This section provides an outline of the calibration of the prediction techniques (historical, regression, seasonal and non-seasonal ARIMA) that are subsequently evaluated in next section.

Data from detector station A were used for calibrating and validating the techniques evaluated. The calibration data consists of only weekdays (19, 22-26 June 1998, 13-17 July 1998). The techniques were validated using weekday (20-24 July 1998) and weekend data (25-26 July 1998). Validation against weekend data would highlight the robustness of the techniques.

A 15 minute prediction horizon was used for this study. Prediction horizons between 5 and 30 minutes represent a worthwhile forecasting window. Shorter periods than 15 minutes tend to produce unstable results. On the other hand, longer time intervals remove high frequency variations and smooth the trend and pattern within the data.

Two statistical error measures: mean absolute error (MAE) and root mean square error (RMSE) were used for model identification, parameter estimation and forecasting accuracy estimation. AIC (Akaike's Information Criterion) indicator, was also applied to the time series models for model selection. AIC is not applicable to other techniques. The mean absolute percentage error (MAPE) is also presented for each of the techniques investigated.

Regression

The various regression models that were developed to determine the flow at time t at station A (ie. the dependent variable), used different combinations of the following independent variables:

- the flow, q, at times t-1, t-2, t-3 at station A, and
- the flow, q_u, at times t-1, t-2, t-3 at upstream station B, and
- the flow, q_{u+1}, at time t-1 at station C further upstream.

Regression models are commonly developed using the least squares method. The objective of the method is to minimise the sums of the squared residuals or error (ie. the difference between the observed value and the estimate), as a criterion to obtain the best fit.

In this study, various regression models were developed using the confirmatory method based on the findings of a correlation matrix of all candidate variables and the results of several stepwise model estimations.

Table 1 presents the calibration results of only some of the multiple linear regression models that were developed during the analysis.

Table 1 – Calibration Results of Multiple Linear Regression Models

Model No.	Regression Model	MAE	RMSE	MAPE (%)
1	q(t)=Aq(t-1)+B	201	341	13.3

2	$q(t)=Aq(t-1)+Bq_u(t-1)+C$	200	327	13.6
3	$q(t)=Aq(t-1)+Bq_{u+1}(t-1)+C$	225	377	15.2
4	$q(t)=Aq(t-1)+ Bq(t-2)+Cq_u(t-1)+ Dq_u(t-2)+E$	189	280	14.4
5	$q(t)=Aq(t-1)+ Bq(t-2)+ Cq(t-3)+Dq_u(t-1)+ Eq_u(t-2)+F$	187	280	14.5

The results presented in Table 1 show that including a lot of variables does not greatly improve the model's performance. For example, there is only less than 7% difference between Models 1 and 2 and Models 4 and 5. The results also indicate that using traffic flow values measured too far upstream do not enhance the model's performance.

The greatest difference between these models is shown when the predicted and observed traffic flows are plotted against the time of day (refer to *Figure 5*).

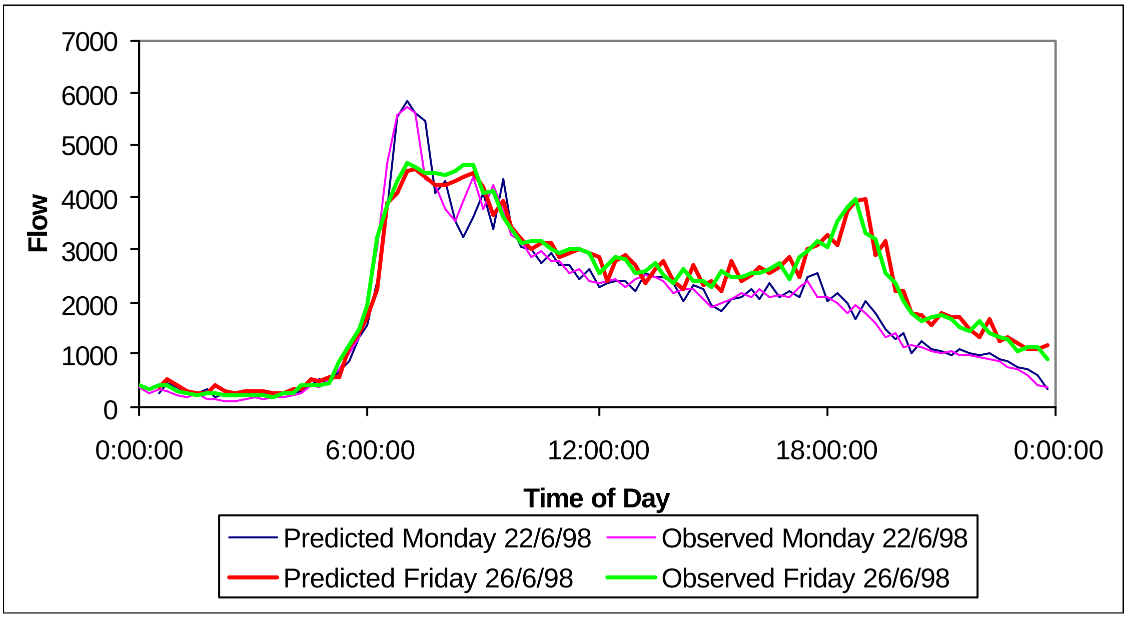

Figure 5 – Traffic flow versus time of day predicted using Model 4

All of the regression models demonstrate some degree of "lagging" between the predicted and the observed traffic conditions. That is, the regression models cannot quickly adapt to changes in relatively current traffic conditions without a point of reference. Increasing the number of variables in the model (ie. the inclusion of flow values measured at times further into the past) decreases the "lag" because there is more information about the previous traffic flow values and patterns.

The presence of lagging is not sufficient justification for including a large number of independent variables into the regression model. A ratio of past and current traffic flow variables was included in the model (eg. $q(t-1)/q(t-2)$) to introduce some information about the past. A variety of different variable combinations were calibrated. Table 2 presents the results.

The inclusion of the variable ratio slightly improves the performance of the models in relation to the indicators. The ability to reduce the presence of "lagging" between the regression model and the observed traffic conditions is fairly similar (especially in Model 7) to that of Models 4 and 5.

Table 2- Calibration Results of Multiple Regression Models Using Ratios

Model No.	Regression Model	MAE	RMSE	MAPE (%)
6	q(t)=Aq(t-1)+Bqu(t-1)+C(qu(t-1)/qu(t-2))+D	204	302	20.5
7	q(t)=Aq(t-1)+Bqu(t-1)+C(q(t-1)/q(t-3))+D	190	281	17.5
8	q(t)=Aq(t-1)+B(q(t-2)/q(t-3))+C	201	313	18.6

Models 1 and 7 were subsequently validated using the validation data set. The results of which are presented in the next section.

Historical Average

Two historical average models, simple historical average, $q(t+1) = q_h(t+1)$ and a variant of historical average $q(t+1) = q_h(t+1) + k\ [q_h(t) – q(t)]$, were evaluated. Both models (*simple* and *enhanced*) rely on the calculation of a historical average to be used as a reference. Weekday data from 19th to 26th June were used to form the historical reference. Estimation of parameter k in the *enhanced* model was carried out using the performance indicator MAE and RMSE.

The two models performance are presented in Table 3. The calibration results showed that using an adjustment factor k with the historical average to reflect the measured traffic flow condition performs better than using historical average alone.

Time Series

The Box-Jenkins methodology that involves a three-stage cycle was used. For a detailed description of the Box-Jenkins methodology refer to Makridakis et al. (1998) and Newbold and Bos (1994). The first step requires the selection of the appropriate degree, d, of differencing, and the autoregressive and moving average orders, p and q of the ARMA model.

Once the potential models from the general ARIMA class have been selected for fuller analysis, the second step is to estimate the unknown coefficients of all the potential models.

The best model is selected according to the closeness of fit to data using AIC (Akaike's Information Criterion). MAE and RMSE were also used to assist in the selection.

ARIMA(1,1,0) and ARIMA(2,1,0) were found to be the most suitable for this study (see Table 3). The lower autoregressive, order ie. ARIMA(1,1,0) was selected for validation because the higher autoregressive order did not significantly increase the predictive accuracy.

Seasonal ARIMA (*SARIMA*) was applied to the calibration data using a season of 1 day (s=96) and 1 week (s=672) to represent traffic flow patterns that repeats itself every weekday and every specific day of the week respectively. SARIMA(2,1,0)(1,0,0)$_{672}$ is the model closest fit to the calibration data and performs better than the same model with a season of 1 day, SARIMA(2,1,0)(1,0,0)$_{96}$ (see Table 3). SARIMA(2,1,0)(1,0,0)$_{672}$ is selected for validation. Also, the SARIMA model does not have the “lagging” effect of the ARIMA model (see *Figure 7* and *Figure 8*).

Table 3 – Calibration Results of Historical Average and Time Series Models

Model Type	Model	MAE	RMSE	MAPE (%)
Historical Average	$q(t+1) = q_h(t+1)$	169.6	248.7	10.1
Historical Average	$q(t+1) = q_h(t+1) + k\,[q_h(t) – q(t)]$	119.0	176.0	7.4
Time series	ARIMA(1,1,0)	196.3	308.2	11.6
Time series	ARIMA(2,1,0)	193.7	307.1	11.3
Time series	SARIMA(2,0,1)(1,0,0)$_{96}$	136.6	229.0	8.9
Time series	SARIMA(2,0,1)(1,0,0)$_{672}$	83.6	184.6	4.9

MODEL RESULTS AND EVALUATION

A total of six models were calibrated and the validation results are presented in Table 4. Based on MAE and RMSE, the *enhanced* historical average model (see *Figure 6*) has the best performance, and the regression and ARIMA(1,1,0) models have the worst performance (see *Figure 7*). Although the SARIMA(2,0,1)(1,0,0)$_{672}$ has the best performance in the calibration (see *Figure 8*), this model did not perform as well as the enhanced historical average model.

A practical measure of the predictive accuracy of the models is to evaluate the percentage of predicted flow within say 5 percent error. For the enhanced historical model, 74.6% and 91.9% of the predicted flows are within 5 and 10 percent errors respectively. Compared to enhanced historical average, SARIMA(2,0,1)(1,0,0)$_{672}$, has 69.7% and 85.4% of the predicted flows within 5 and 10 percent errors respectively.

The strength of historical average models is that the uses of historical data helps capture the shape of the traffic flow pattern, including the degree of peaking and its starting and finishing times. This characteristic is an advantage over some other techniques, such as the regression model, that has difficulty predicting the shape of the traffic flow pattern and produce "lagging" behind the actual traffic flow pattern.

To test the robustness of the enhanced historical average model and the SARIMA model, weekend data (25th -26th July 1998) were used. The results (see Table 5 and *Figure 9*) clearly highlight the weakness of the historical average model, that is its inability to predict traffic pattern when current traffic pattern is very different from the historical pattern. Hence it is important that historical data are categorised into homogeneous groups.

Although the historical average models perform well during normal operating conditions, they do not respond well to external changes in system such as weather, special events, or modified traffic control strategies

Table 4 – Validation Results

Model Type	Model	MAE	RMSE	MAPE (%)
Regression	$q(t)=Aq(t-1)+B$	194.0	345.0	14.3
Regression	$q(t)=Aq(t-1)+Bqu(t-1)+C(q(t-1)/q(t-3))+D$	181.0	281.0	18.2
Historical Average	$q(t+1) = q_h(t+1)$	158.2	231.0	10.5
Historical Average	$q(t+1) = q_h(t+1) + k\,[q_h(t) – q(t)]$	110.2	172.9	7.3
Time series	ARIMA(1,1,0)	193.2	311.8	12.0
Time series	$\text{SARIMA}(2,0,1)(1,0,0)_{672}$	154.7	270.3	8.8

Table 5 – Validation Using Weekend Data Results

Model Type	Model	MAE	RMSE	MAPE (%)
Historical Average	$q(t+1) = q_h(t+1) + k\,[q_h(t) – q(t)]$	252.9	434.3	20.3
Time series	$\text{SARIMA}(2,0,1)(1,0,0)_{672}$	150.3	200.1	10.3

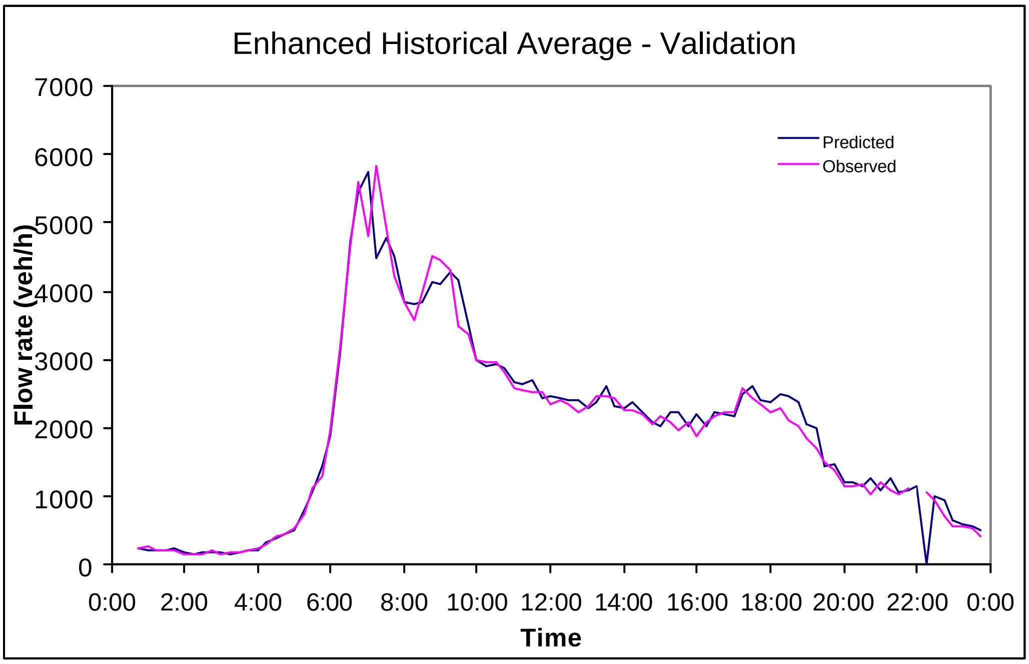

Figure 6 – Validation of Enhanced Historical Average Model

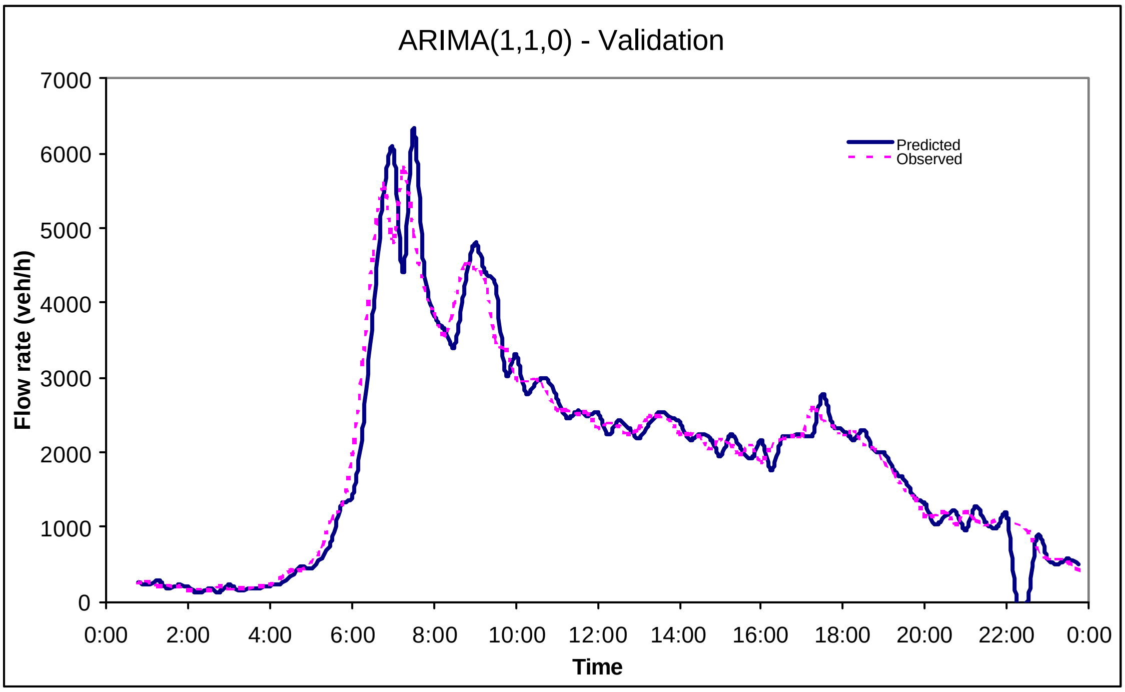

Figure 7 – Validation of ARIMA(1,1,0)

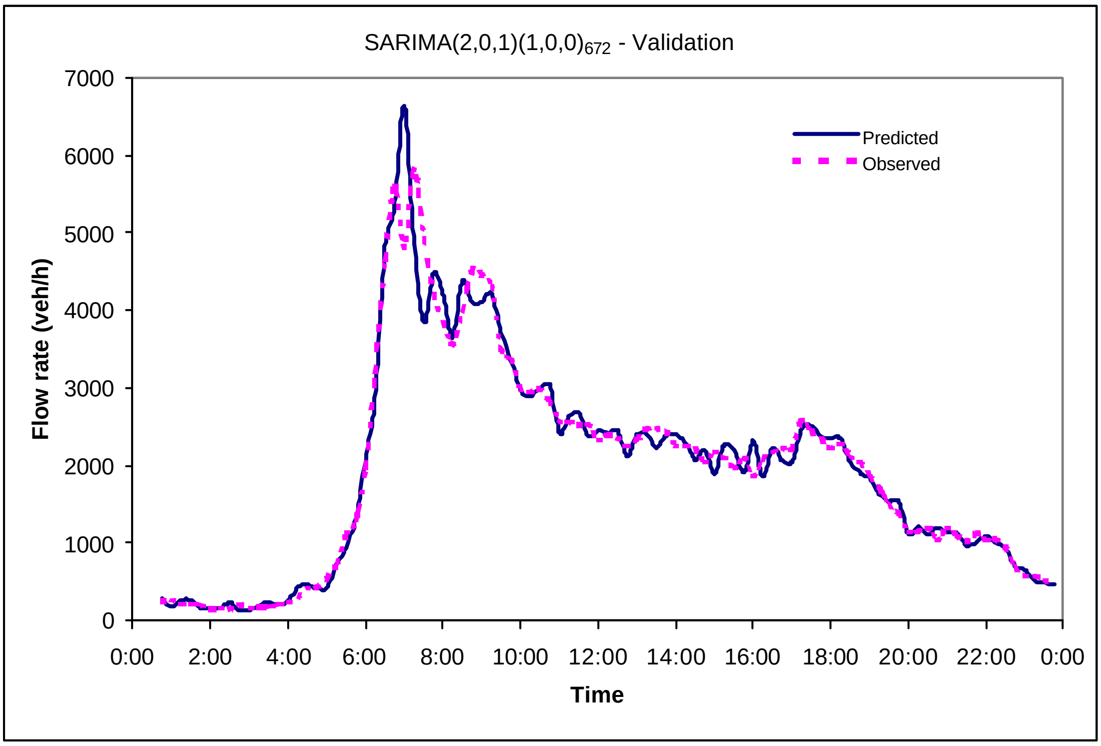

Figure 8 – Validation of SARIMA(2,0,1)(1,0,0)$_{672}$

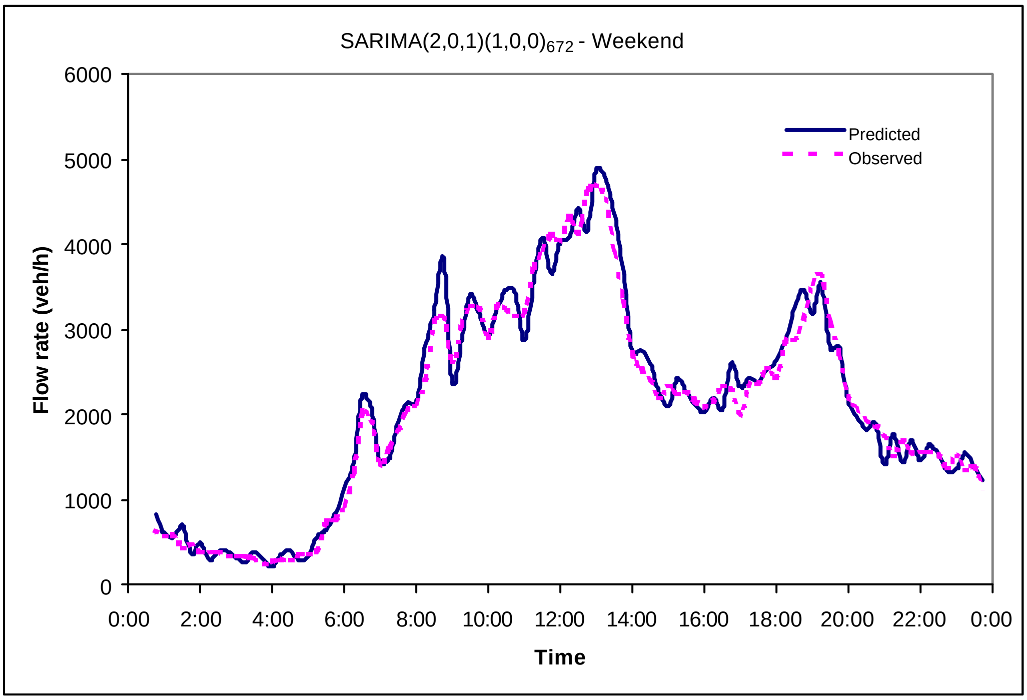

Figure 9 – Validation of SARIMA(2,0,1)(1,0,0)$_{672}$ using Weekend data

CONCLUSION

The enhanced historical average and SARIMA hold considerable promise for application to traffic flow prediction. The strengths of these models have been demonstrated using real life freeway data.

One common weakness of the regression and ARIMA models is their inability to forecast the traffic flow pattern and the production of a "lagging" effect. The SARIMA model has a seasonal component that captures the seasonal aspect of the traffic flow pattern and eliminates the "lagging" effect. The enhanced historical average can also capture the shape of the traffic flow pattern using historical data.

Advantages of the historical average include the ease of implementing this model and its high execution speed.

Although the enhanced historical average performs well during normal operating conditions, it does not respond well to external changes in the system such as weather or special events. This study demonstrates the weakness of using the historical average technique based on a weekday historical pattern for predicting weekend traffic flow.

The limitation of the historical average could be overcome by categorising data into different homogeneous groups that can be applied to the appropriate condition.

Time series models rely on past data for prediction and there is an issue with the handling of missing data when implementing a time series model.

Further study should be carried to test SARIMA and the historical average technique on other sites and to test the predictive accuracy of SARIMA by feeding the measured data back to the model. It would also be beneficial to test the performance of SARIMA on shorter-term prediction horizons of 2-10 minutes, as some traffic management and information systems need to predict traffic conditions in few minutes time for effective traffic management.

Dia (2000) demonstrated that neural network (time-lag recurrent network) has the capability to reduce or eliminate the lagging effect and to produce prediction accuracies of up to 95 percent. It would be valuable to compare the performance of SARIMA and neural network based on the same data set in future work.

REFERENCES

BOX, G.E.P. and JENKINS, G.M. (1970). *Time series analysis forecasting and control*. (Holden-Day, San Francisco, California)

DIA, H. (2000). An object-oriented neural network approach to short-term traffic forecasting. *Special Issue of the European Journal of Operation Research*, December 2000.

MAKRIDAKIS, S.G., WHEELWRIGHT, S.C. and HYNDMAN, R.J. (1998). *Forecasting : methods and applications.* 3rd Edition. (Wiley: New York)

NEWBOLD, P. and BOS, T. (1994). *Introductory business & economic forecasting.* 2nd Edition. (South-Western Publishing Co.: Cincinnati, Ohio)