A study of travel time predictability in Auckland

David Siddle, Jacobs

1 Introduction

1.1 Overview

In 2012, the NZTA appointed Sinclair Knight Merz (SKM) to carry out research into Travel Time Predictability. This research intends to clarify how historical base line data combined with near real-time data including environmental conditions, incidents and traffic flow can contribute to the calculation of reliable and timely-delivered travel time predictions.

1.2 Background

Reliable journey time is a key parameter in travellers' route choice and has important applications in transport planning and modelling. For transport users, it affects their choice of mode, journey route and also their activity patterns. For transport planners and policy makers, journey time estimates are used to provide key indicators for performance monitoring, congestion management, travel demand modelling and forecasting, traffic simulation, air quality analysis, evaluation of travel demand and traffic operations strategies.

Real time travel time information is becoming increasingly important for a variety of transport applications - including Advanced Traveller Information Systems (ATIS), Advanced Traffic Management Systems (ATMS), Route Guidance Systems (RGS), etc., which all form part of the collective Intelligent Transportation System (ITS).

As Intelligent Transportation Systems (ITS) are deployed more widely throughout the world, managers of transport systems have increasing access to large amounts of historical and "real-time" status data. Traffic flows, speeds and densities on transport network are being continuously measured by different monitoring systems such as loop detectors, Automatic Number Plate Recognition systems (ANPR), Closed Circuit Television (CCTV) monitoring and, more recently, probe vehicles and mobile phone data.

The collected information can be used to guide the use of dynamic traffic management measures such as variable message signs, information provision by radio and internet, etc. in order to reduce congestion and improve network efficiency.

The availability of real-time traffic information, developments in Information Technology and the need for predicting short-term traffic conditions have raised the question: *Can we predict travel time?* It has been recognized by researchers and practitioners alike that the benefits of ITS capabilities cannot be fully realised without an ability to anticipate traffic conditions in the short-term (i.e. less than one hour into the future).

Predictive modelling capability can theoretically provide the ability to forecast the performance of a transport network in the short term and may also allow the impact of planned and unplanned events and incidents on the network to be assessed in near real-time.

1.3 Research objectives

The key objectives of the research were to:

- Understand the ***current state of research*** locally and internationally on real-time prediction methodology – specifically identify from current research the most suitable methodology for predicting journey times from real-time data sources for strategic highway networks.
- Understand ***what data sources*** currently exist and identify the gaps in the current information sources which can lead to accurate journey time predictions.
- Develop ***summary statistics*** of the observed data to identify the impact of key parameters on journey times in '*normal'* traffic conditions and also during ***'****abnormal'* conditions resulting from planned and unplanned events and incidents
- Develop an appropriate ***forecasting model*** to predict travel times with a reasonable degree of confidence and develop a testing methodology for verifying the results
- Summarise the findings in a comprehensive ***research report*** which also provides a programme for delivery of a Travel Time Predictive Capability for the strategic highway network.

The research needs to deliver a modelling framework that can forecast journey times based on near real-time traffic information and historic data sources. The modelling approach should attempt to take into account all explanatory variables which will affect the reliability of forecast results. The model will also need to incorporate ***Spatio-Temporal*** relationships which have a direct impact on traffic on the highway sections where journey time is being forecast.

One of the most import aspects of the research is to understand existing data sources and gaps in traffic information which can affect the reliability of modelling results. In addition, the research will also focus on the impact of planned and unplanned disruptions in the highway network capacity.

1.4 Research stages

The study involves three main stages of work:

- A literature review, which examines the various prediction methods and case studies which have been developed. The objective of this stage is to gain an insight into the current state of the art methodologies and research into developing techniques. This literature review will inform the development of a travel time prediction methodology for this specific case study in New Zealand.
- A review and analysis of available data sources. The study will require a large quantity of existing, available real time data sources in New Zealand and for this purpose, the strategic motorway network of Auckland has been taken up as the core study network due to extensive coverage of real time continuous traffic monitoring stations and availability of over two year worth of historic data.
- Development of a working "demonstrator" model which will attempt to apply the preferred methodology to the historic data and predict future journey times. These predictions will be validated against the actual observations.

2 Literature Review

2.1 Overview

The first stage of the study is a comprehensive Literature Review, which forms a descriptive review of the existing literature in travel time prediction. Various prediction methods and case studies have been reviewed to develop an overview of the current methodologies and research into developing techniques. The key objective of this literature review is to provide a survey of existing literature and the applied methods for travel time and other real time traffic prediction studies. This chapter provides a brief overview of the Literature Review and a summary of its findings.

There are several papers detailing research into travel time prediction spanning the last two decades (Peeta and Mahmassani, 1995; Ben-Akiva et al., 1997; Al-Deek et al, 1998; Kamarianakis et al., 2003; Zhang and Rice, 2003; Meschini and Gentile, 2010; Gentile and Meschini, 2011). Many of these papers provide evidence showing the successful application of short-term travel time prediction around the world. By and large, the literature points to evidence that the reliability of the journey time predictions depends on the quality, level of accuracy and availability of real-time data, the extent of the transport network and the chosen prediction methodology (D'Angelo et al. 1999; Van Lint et al. 2002, 2003, 2005).

The literature also highlights that the "Forecasting Platform" needs to be sufficiently fast (very close to real time) in order to allow network managers to detect and mitigate, minimise or remove the impact of events and incidents within the network (Gentile, 2010).

2.2 Key variables and data requirements

Travel time predictions require an extensive understanding of general traffic behaviour and underlying trends along with real time traffic state measures from the field. Historical data is required to develop relationships between traffic speeds and other inputs, and to also develop an understanding of traffic patterns during different time periods and seasons.

Review of available literature (Vanajakshi 2004; Kamarianakis et al. 2005; Min et al. 2007; Meschini and Gentile 2009) reveals that a multi-variable travel-time prediction framework benefits from the following primary data:

- Average speed
- Average vehicle length
- Traffic volume (unclassified or classified)
- Average occupancy (of detector loops)
- Highway geometric data (e.g. turning radii, section lengths, etc.)
- Network configuration (e.g. banned turns, one way streets, etc.)
- Incident data (location and duration)
- Route choice (e.g. turning proportions at junctions).

Of this data, traffic speed, volume and occupancy data are the key variables in models that have been successfully developed. Highway geometry and network configuration are other inputs into the model which can be introduced. Route choices and interaction between network links can be imported from a developed and calibrated strategic transport model.

In addition to the above, other supplementary data sources can be used to provide further segmentation of the model, thus allowing separate relationships to be developed for varying network conditions. These include:

- Weather conditions
- Day-type classification (e.g. weekday, weekend, holiday period, etc)
- Planned and unplanned incidents record.

2.3 Data Sources

There are two main types of data sources required to develop accurate travel predictions:

- "Real-time" data: live traffic data streams used as inputs to the predictive model tool; and,
- "Historic" data: primarily, this is the same data as above, but collected over a period of time to allow periodic calibration of the model. In addition, this data is supplemented by secondary data (e.g. event data) used to develop an understanding of how traffic conditions vary seasonally, or in response to events and other external influences.

The data requirements for each model will depend on the required outputs and the specifics of the study area. For example, traffic state prediction for an urban network will require detailed traffic flow patterns including OD information, routes taken by vehicles, vehicle classification, key traffic generators and attractors. For statistical travel time predictions on a strategic highway corridor, the above information may not be required as route-choice and the number of attractors and generators will be limited.

2.4 Modelling approaches

Existing research and linkages

Real time traffic modelling and predictive analysis has been under research for a number of years. There is a large interest in understanding the progression of traffic and predicting traffic conditions in the near future. The research papers variously propose a wide-range of applications, predominantly online information systems, vehicle and probe guidance systems and input to other traffic management systems used by road traffic authorities.

Various studies have approached the problem from different perspectives largely owing to different project objectives and the availability of data. The review of literature reveals that two main categories of modelling approaches have been successfully used in the development of predictive and real-time modelling tools:

- ***Dynamic Simulation Modelling***: This uses extensive network representation and assignment of historic travel patterns using real-time traffic data inputs (Ben-Akiva et al. 1997; Yang et al.2000; Kaufman et al. 2001; Gentile 2010). These models use first principles to describe drivers' route choice behaviour based on dynamic traffic assignment and progression of traffic based on network supply constraints.
- ***Statistical Modelling***: These are data-driven methods and predict travel times based on current and past real-world detector data, without explicitly considering the physical processes considered by the simulation based methods (Smith et al., 2002; Ishak, 2003; Yang, 2005; Min et al., 2007). Auto Regressive (AR) time series analysis and other stochastic methods are used based on historic and near real-time data (Kamarianakis and Prastacos, 2003; Milkovits, 2010; Hu and Ho, 2010).

Dynamic simulation modelling approach uses first principles of traffic flow theory in determining traffic flows and speeds through the network and thereby attempts to predict the progress of vehicles (either individually or in aggregate) through the network. Secondary variables such as journey time, saturation, delay and queues can then be calculated following the simulation.

The statistical modelling approach uses a calibrated function to predict the travel time on each link based on current link speeds, historical traffic data and spatio-temporal transport network correlations (i.e. speed and flows on other parts of the network). Statistical modelling requires separately calibrated models for each model segment.

Dynamic Simulation Models

The Simulation modelling approach is based on an explicit and physical interpretation of the network and traffic conditions. This is achieved through detailed simulation of the interaction between travel demand and road network (supply) through assigned routes based on a previously developed equilibrium assignment model.

The dynamic modelling techniques use a traditional transport modelling approach where the travel demand is assigned on highway network as a continuous load. The simulation-based model predicts traffic conditions based on real-time data inputs and an existing calibrated traffic model. It uses established modelling theory to assign traffic for a number of small, discrete time periods in the near-future. The output from each time period is used to inform the next. Travel time and other network statistics are available as standard outputs of the model. This approach requires the modeller to estimate traffic conditions at the boundaries of the model (this includes traffic generators within the study area).

Examples of the successful application of this approach are evidenced by DYNAMIT (Ben-Akiva et al. 2003), METANET (Smulders et al. 1999), DYNASMART (Hu, 2001), and CORSIM (Liu, et al. 2006b).

This forecasting methodology has benefits in simultaneously and holistically forecasting network performance and travel metrics for the entire network. It also has an added advantage in cases of unplanned incidents where new forecasts can be derived from “re-assignment” of expected demand. On the other hand, simulation modelling can be extremely data intensive and slower than numerical models using a statistical approach. Simulation modelling is most beneficial when used for urban traffic management or in deriving efficiency gains at traffic signal control centres. The approach could theoretically be used for simpler networks (for example, motorways), but in these situations the statistical approach may be able to provide more accurate results, and do so quicker and much more cost-effectively than by using simulation modelling.

Statistical Models

The Statistical modelling approach uses interpolation, interference, data mining, and mathematical models to derive near-future segment speeds based on current observations of traffic speeds and volumes. This approach uses historic data to determine and synthesize a mathematical relationship so that near-future conditions can be determined from trends in real-time traffic volume and speeds. Statistical methods such as time series analysis and the Kalman filtering method are used to calibrate the model and to provide an estimation of future travel time from real-time data.

In the simplest example of a statistical model, the current speed on a particular network segment will provide a good indication of the speed on downstream links in the near future, the time difference depending on the distance between the two segments. The problem becomes more complex when two traffic streams merge or separate at on- and off-ramps, or where there is a change in link capacity.

These models are extremely useful in predicting traffic speeds in simple networks with low traffic volatility. The challenge this methodology faces is when there are unprecedented traffic flow patterns due to significant spikes in travel demand or highway network disruption (e.g. roadworks) which change the network configuration or capacity.

2.5 Comparison of modelling approaches

Simulation Method

The Simulation method uses a traffic model to calculate the routing of vehicles from first principles based on established modelling algorithms. It is potentially a very powerful tool, provided that the underlying model is robust and there is sufficient data to allow accurate calibration. The advantages and disadvantages of this approach are listed below.

Advantages

Unlike the Statistical method which is suited to simple networks, there is theoretically no mathematical limit to the size or complexity of the network which can be covered by a Simulation-based predictive model. This is because the method is based on an underlying model which contains information about vehicle routing, which means that the number of calculations in the model grows roughly linearly as the number of modelled links increases (as opposed to the exponential increase required in Statistical models). Model size will be limited by the speed at which assignments can be undertaken so that predictions can be delivered timely, but computer processing power is now sufficient to allow even large cities to be covered in detail.

The models upon which Simulation-based systems are based also provide the ability to forestall the impact of events. Whereas the Statistical method relies on mathematical relationships to determine future trends, strategic models contain much more fundamental data on the road network and traffic demand and can forecast future behaviour from first principles. This information can be used to predict how travel behaviour will change regardless of whether a particular event or incident has been observed before. An example could be the temporary closure of a road which could be simulated in the model before the event takes place, ensuring that the remainder of the network is optimised in anticipation of the event. Furthermore, a Simulation-based system could be used to test a range of traffic management scenarios to determine the optimum solution in response to a particular planned or unplanned event. The Simulation method can also handle signalised junctions much more easily in terms of incorporating the impact of current plans on delays and routing, but additionally, a Simulation-based system could potentially assist in the selection of appropriate plans to minimise delays on the network.

Another advantage that the Simulation method has is data redundancy and its ability to still produce relatively accurate predictions despite gaps in the incoming data stream. This is because the model can be updated and recalibrated based on whatever data is available, yet still maintain a good representation of network conditions across the network.

However, the biggest advantage that the Simulation method has over the Statistical method is that it already quite well established with at least two well-developed commercial packages available (albeit at significant cost).

Disadvantages

It is no surprise that the complexity of Simulation-based tools is also a big disadvantage, not least because of the large quantities of time and money that must be spent in both implementing and maintaining them. The cost of the commercial software aside, there is usually great expense required to develop and calibrate the traffic model which will form the basis of the system. For the system to work accurately, the model must be robust and based on high-quality and spatially detailed data. This will usually require a very comprehensive programme of surveys to determine the underlying travel behaviours. Construction of the model itself will require specialist expertise to ensure that it is built to the highest standard. In addition to the up-front costs, the model will require periodic re-calibration every few years requiring further expense.

Statistical Method

The Statistical method is based on mathematical relationships developed through statistical regression analysis. It is the simpler of the two methodologies which means it should be quicker and cheaper to implement, but this also limits its capabilities. The advantages and disadvantages of this approach are described below.

Advantages

The main advantage of the statistical method is its simplicity. Although the relationships and functions used to undertake the predictions can be become quite complex, the framework in which they sit is very basic and will require little in the way of manual input or monitoring. Provided the inputs and outputs to the system are automated, a prediction tool based on the statistical method should be self-maintained. The system can be configured to periodically recalibrate itself and because there is no detailed network, any minor changes to the transport network will be incorporated into the model automatically over time. On a simple network, such as a trunk road or motorway network, this method could provide accurate results at a fraction of the cost compared to the Simulation method.

Disadvantages

The Statistical method's simplicity is also its key weakness. As transport networks become more complex, the number of connections between links in the network quickly multiplies to the point where regression can no longer identify which connections have significant mathematical relationships. Furthermore, networks with anything more than nominal route choice will most likely render the method ineffective, as will any network with a large number of traffic signals; both of these situations lead to interactions which are too complex for the model to handle with any degree of accuracy.

Another weakness of the Statistical method is that it is "reactive", i.e. the functions are re-calibrated based on recent observations. As noted above, this is a desirable feature as the model can update itself and react to changing conditions. However, this does mean that planned events cannot be pre-empted unless they have occurred and observed before; examples include sudden event-related peaks of traffic, or temporary lane closures. In these cases the model will offer predictions based on the best information available, but with varying degrees of accuracy.

Finally, the Statistical method suffers because it has been largely overlooked in the past by transport software firms. This is probably because the method is best suited to sparse, strategic networks, whereas the focus of predictive modelling so far has been in urban settings. This means that, although a "Statistical" system would be simple to operate, it will require highly-specialised skills to set up. This is due to the need to undertake very complex statistical analysis and to develop a working system from first principles.

Method comparison

Table 2.1 shows a high level comparison of Statistics-based and Simulation-based approaches.

Table 2.1: Comparison of Modelling Approaches

Features	Simulation-based approach	Statistics-based approach
Basis of system	Traffic model calculating vehicle routes from first principles	Mathematical relationships between different road segments and time periods
Network size/complexity	Can be highly complex, e.g. urban road network. Can include signalised junctions and route choice.	Small or simple networks (eg. Single corridor or a motorway network). Limited signals and route choice
Data Requirements	Complex: Additionally requires data on travel behaviour and signal information.	Basic: “Live feed” of traffic flows, speeds and external events, plus historic log of same data.
Ability to predict during events	Good. Plus, can be used to pre-empt planned events and to test a range of traffic management responses	Good if event has occurred before, otherwise limited.
Cost to implement	High	Modest
Cost to monitor and maintain	High. Will generally need constant monitoring and require periodic recalibration	Low. Will be largely self-maintaining and will require low levels of monitoring.
Software available?	Yes.	No

Survey of travel time prediction studies

The review of the studies shows that the Statistical approach was mostly used when the study network was a corridor with a limited number of links and junctions. This modelling approach does not require information regarding the origin and destination of trips, zones or route choices. The forecast run-time is quick and in most of the examples provided this approach has been capable of accurate predictions.

The Simulation method has been used to model larger or more complex areas. However, the number of examples of this type of model is lower than for the Statistical method.

2.6 Literature Review conclusions

The primary objective of the literature review was to highlight the state of research and the applications of predictive modelling. This included travel time forecasts as well as general forecasts of traffic speeds and volumes in varying conditions.

Two main prediction techniques were discussed:

- The Simulation method.
- The Statistical method.

The Simulation method requires the development of a transport model where traffic demand is simulated and calibrated in real-time based on incoming traffic data. The second method relies on the estimation of mathematical functions based on historical data which calculate future traffic speeds based on current network conditions.

It is clear from the review that there are significant limitations with the Statistical method, but that it is much simpler and quicker to implement and easier to maintain. For study areas which are suited to the Statistical method then this would offer the best option. This would include simple strategic trunk or motorway networks with limited route choice and signalised junctions. For more complex networks such as city centres or other urban areas, the Simulation method would be recommended.

The study network adopted for this research is a motorway corridor with fixed access and egress points, which is perfectly suited to the Statistical method. The Simulation modelling approach is unlikely to add much value without significantly increasing the resource requirement to build a complete traffic demand model. As such, it was recommended that the Statistical method be used for this study

3 Data Review and Methodology

3.1 Overview

The Literature Review examined various research papers and identified two general approaches. Of these the statistical approach was recommended as this was considered to be most suited to the client's requirements. The second stage of the project reviewed the data which had been received and undertook some initial analysis of the data based on some of the statistical methods identified. The outcome of this analysis would allow a more detailed methodology to be developed for the final stage of the study

3.2 Statistical methods

The Literature Review identified two main groups of statistical modelling which have been successfully implemented in the past and these are described below.

Auto-regressive moving average

Later methods, most specifically that proposed by Min, Wyntwer and Amemiya in the 2007 IBM Research Paper *"Road Traffic Prediction with Spatio-Temporal Correlations"*, suggested the use of Auto-Regressive Moving Average (ARMA) models. ARMA models look at how a variable is influenced by the past values of independent variables and possibly by its own past values. In the specific case of travel time prediction, this would mean determining the relationship between the current travel time on a particular link and recent travel times on upstream links; this relationship would then be applied to the current travel times on the upstream links to (theoretically) get the future travel times on the link under scrutiny.

We understand that the ARMA methodology proposed in the 2007 paper is similar to the one which is used in the IBM Traffic Prediction Tool which has been recently piloted for the NZTA. This tool has reported reasonably small errors of less than 6% for a prediction horizon of 30 minutes. However, the results for individual links varied and some links had 10-min prediction errors as high as 7.33%. The segment with the highest accuracy in the IBM study (2.7% average error) has more than twice the average error of the non-linear method.

Non-linear time-series analysis

Up until recently, the most common statistical approach had been that proposed by D'Angelo, al-Deek and Wang in their 1999 paper, *"Travel-Time Prediction for Freeway Corridors"* and implemented on the I-4 freeway in Florida. The methodology used non-linear time-series analysis to predict the travel times on individual freeway sections, based on historic data from the same freeway section. A prediction horizon of just 5 minutes was recommended in the paper; the horizon is constrained by the data sample interval and increasing this interval reduces the accuracy of the model. The paper reported that the method achieved an average error of 1.3%, although it was less accurate during congested periods. In total, 98% of predictions fell within 10% of the actual times. A subsequent evaluation of the method by Ishak and al-Deek found that errors increased significantly for longer prediction horizons.

3.3 General approach

Our approach

Our approach to selecting the most appropriate methodology has been to undertake initial analysis of the available data using each of the available statistical processes. This analysis is done at a basic level, involving simple statistical tests including regression and time-series analysis.

The purpose of the analysis is to identify which of the methodologies is most likely to provide the best statistical fit for the data. At this stage, no attempt was made to undertake predictions; the methodology selected as part of this process would be developed in the next stage of work and this process is described in Chapter 4.

Despite reporting slightly less accurate predictions at 5 minutes, the ARMA methodology is theoretically capable of greater accuracy over longer prediction horizons (up to 30 minutes). The evidence from the literature review suggested that this methodology was the most likely to produce significant advances over the current predictive capability. There are a number of different, specialised ARMA methodologies and by combining these with other factors (e.g. weather and incident data) it is possible that accuracy could be suitably improved. Our efforts therefore concentrated on this technique. In addition, we also examined ways in which the time-series analysis method could be extended. The tests undertaken are described in more detail in Section 3.4.

Data received

One of the main issues in undertaking any analysis has been obtaining quality data. The traffic data (flows, segment speeds) is generally of good quality, although there are large gaps in the data where detectors were clearly out of action for long periods. Only one link (out of nearly 500) contained a full 13-month period of data with no missing observations. In total, 5.2% of all observations across all links were missing from the data, although a small number of links had detectors which were malfunctioning or switched off for several months and accounted for most of the missing data. The majority of links had only a few days' data missing at some point over the period. Figure 3.1 shows how the missing data points were distributed across the links.

Figure 3.1 ranks the links according to the number of missing observations and it can be seen that the links range from no missing data to nearly 50,000 missing data points (or about half of the total number of possible observations). Most links have between 100 and 1,000 missing observations. It should be noted that data is recorded in 5 minute intervals, so a single day of missing data would result in 288 missing observations. A week's loss of data would represent 2,016 missing records, and 30 days would be the equivalent of 8,640 missing records. Table 3.1 shows the percentage of links below these thresholds. This table shows that around 83% of links had less than one month of missing data. In total there were 485 links of which 82 had to be initially discarded due to the volume of missing data. However, some of the discarded links were reinstated for later tests where only a subset of data was required and this did not coincide with the missing period.

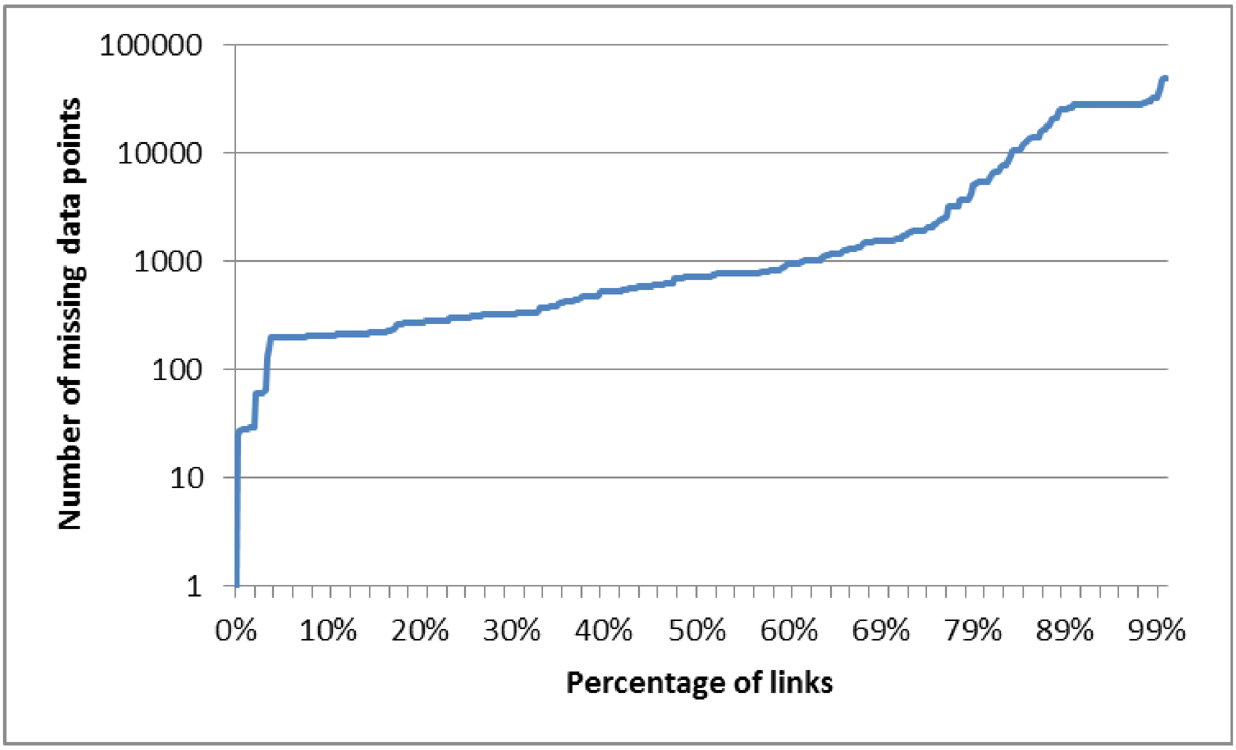

Figure 3.1: Distribution of missing data points over all links

Table 3.1: Percentage of links versus number of missing observations

Duration of missing period	Percentage of links with fewer missing observations than threshold
One day	22.8%
One week	74.3%
One month (30 days)	83.1%

By comparison, the weather data was of very poor standard. The data is incomplete, with rainfall and solar radiation provided for just one of the five sites. Furthermore, the data covers the wrong period (June 2012 to June 2013 rather than January 2012 to January 2013).

There are also some issues with the incident data which came as no surprise. It was always expected that unplanned events would not be very well recorded, and examination of the data shows this to be the case with just 401 incidents over a 13-month period. More than half of the recorded incidents (208) were listed as “Cautions” and were usually associated with maintenance issues such as lack of street lighting or mud on the road which did not necessarily equate to delays on the network. With regard to planned events, the dataset does contain a sufficiently large number, but many of these are for full road closures and are not very specific with regard to time or location, or if they even went ahead as planned.

3.4 Data Analysis

ARMA methodology

With the Auto-Regressive Moving Average (ARMA) model, the observed traffic conditions on each link in the road network can potentially have an influence on the vehicle speeds on other links in later time steps. The theory is that, if a relationship between two links can be demonstrated, then the current observations on one link can be used to predict near-future observations on downstream links.

To determine if the data is suitable for this type of model, a correlation function is used to search the data for similar patterns and trends within the data. This function looks at the time series data of each link and compares it to data on other links (cross-correlation) and to itself (auto-correlation) for a number of different time-offsets. For demonstration, a simplified explanation of this process is described below.

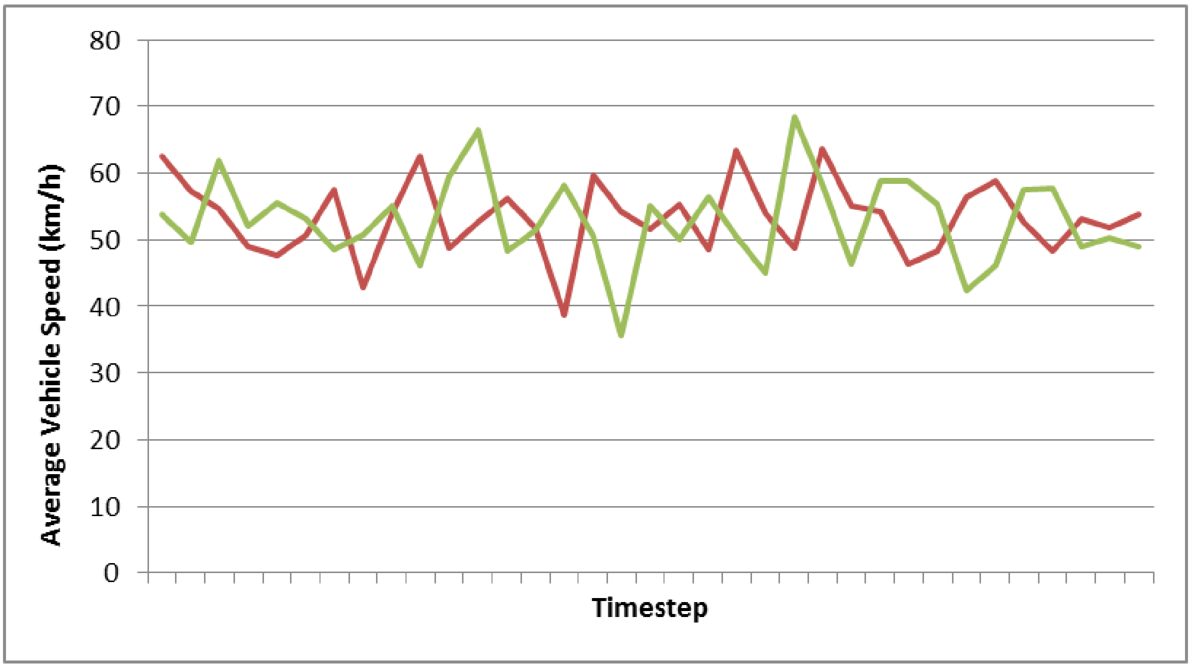

Figure 3.2: Sample data demonstrating correlation function

Figure 3.2 above shows some sample data which will be used to demonstrate the principle behind the correlation function. The two lines represent the speed observations over a number of time periods for two separate links. The correlation function, firstly subtracts the mean value of each link from each data point, this transforms the data series so that “peaks” in the data generally have values greater than 0 and “troughs” generally have data less than 0. When analysing the full dataset a moving average is used instead of the series mean value to avoid seasonality effects. The data is also normalised so that data from links with different speed limits can also be compared.

The images in Figure 3.3 below show the transformed data as dotted lines. The correlation function then multiplies the corresponding data points from the two links; this is shown by the solid blue line in the Figure 3.3 below. Where a peak in the data for one link corresponds to a peak in the data for the second link, the product is a large positive value. The same is true where there are matching troughs in the data (-ve x –ve = +ve). Where a peak in one dataset coincides with a trough in the second dataset, the product is a large negative value. The correlation between the two time series is therefore the area under the blue line.

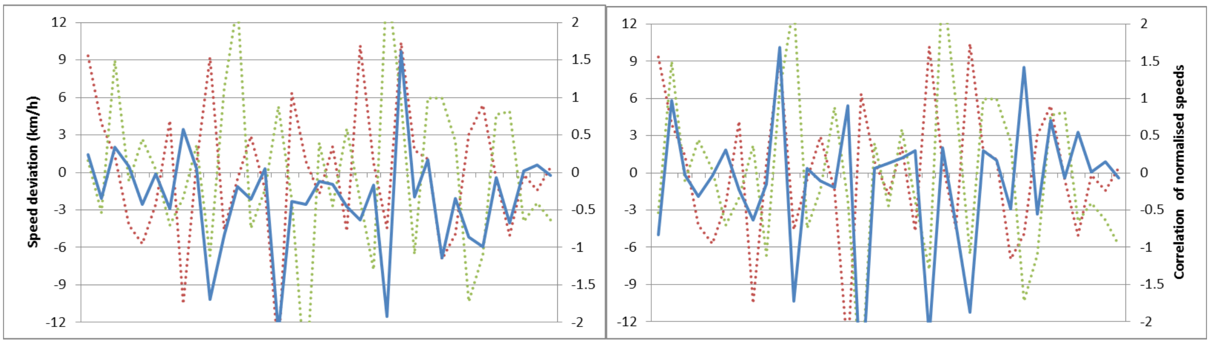

Figure 3.3: Correlation of sample data at T=0 (left image) and T=-1 (right image)

The left image in Figure 3.3 shows what happens when the data from the two sample links is compared. The correlation of the two links is then checked for a variety of lags by offsetting the data from one link. The right image in Figure 3.3 above shows how the correlation changes as the "green" link is offset to the left by one time-step. The correlation improves, but not significantly. However, Figure 3.4 below shows what happens when the offset is increased to two time-steps. This shows a marked increase in correlation with very few negative values. This would indicate that there is a high correlation between the conditions on the "red" link and those on the "green" link, two time steps later (10 minutes in this case).

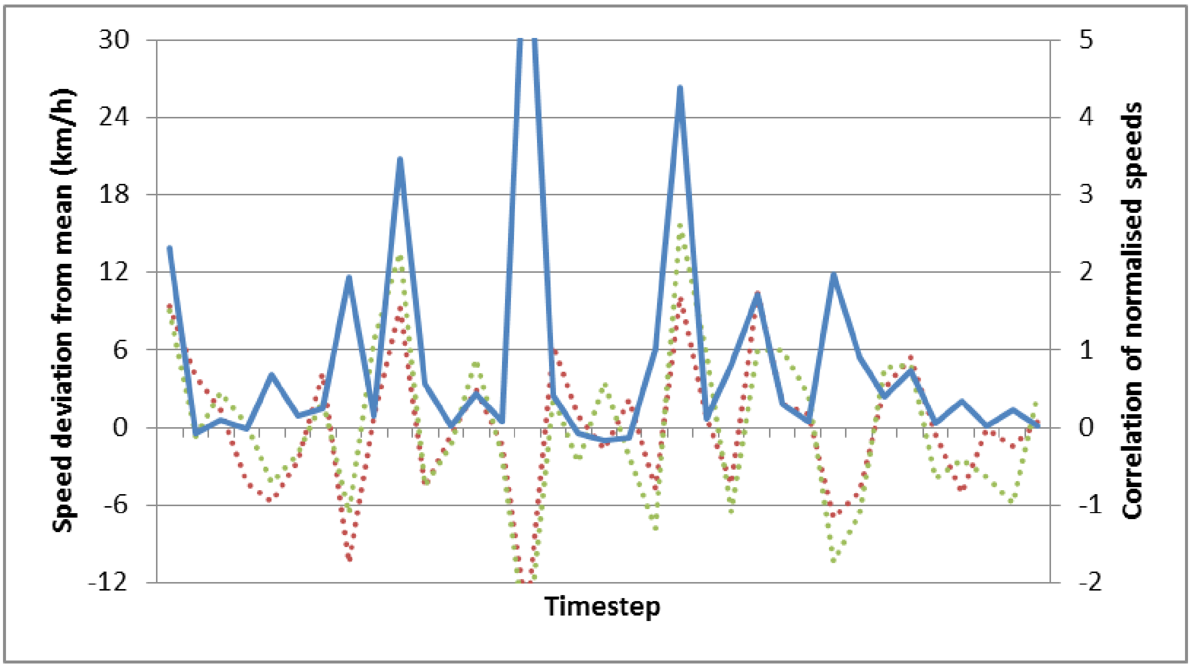

Figure 3.4: Correlation of sample data at T=-2

This process is repeated for a number of lag values (from 5 minutes to 1 hour in our analysis) and the correlation is calculated at each time step. Figure 3.5 below shows the correlation results for the sample data for a variety of lag values. This confirms that a lag of two time-steps provides the best correlation.

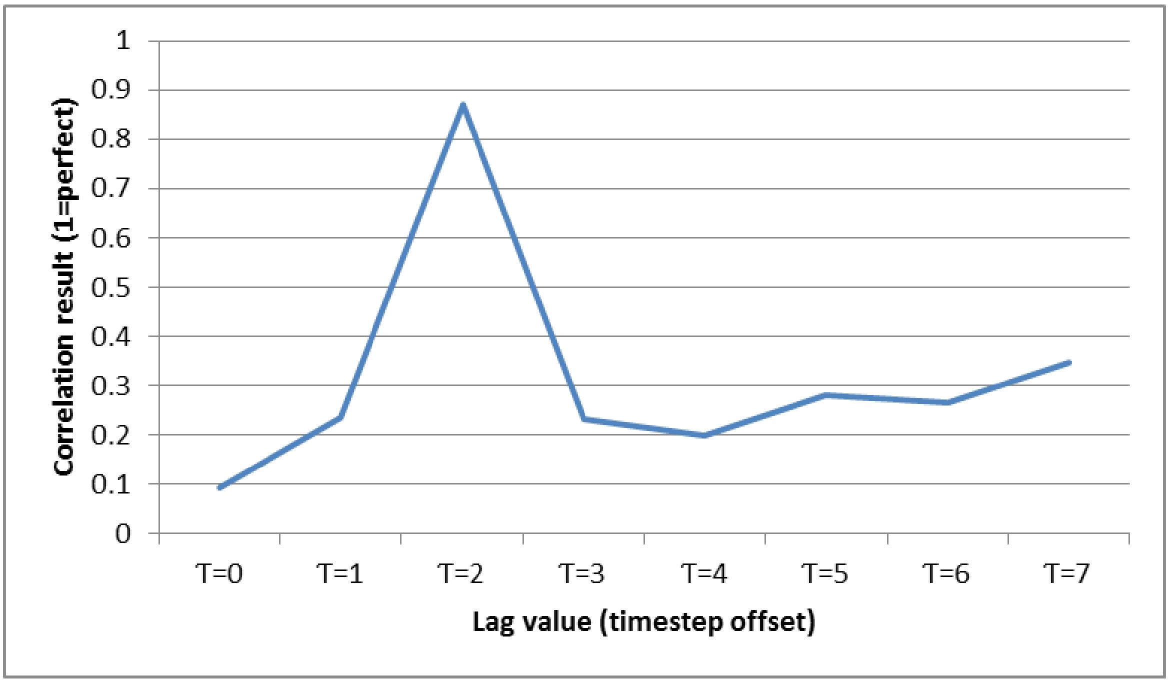

Figure 3.5: Correlation result vs lag for sample data

In addition to comparing data from two different links, it is possible to compare data from the same link against itself. The purpose of this is to determine if there are any cyclic patterns within the data which could then be predicted by the ARMA function. Once the best lag value has been determined, this can be applied to the appropriate dataset and the least squares method can be undertaken to assess the goodness of fit. This analysis is shown for the sample data in Figure 3.6 below.

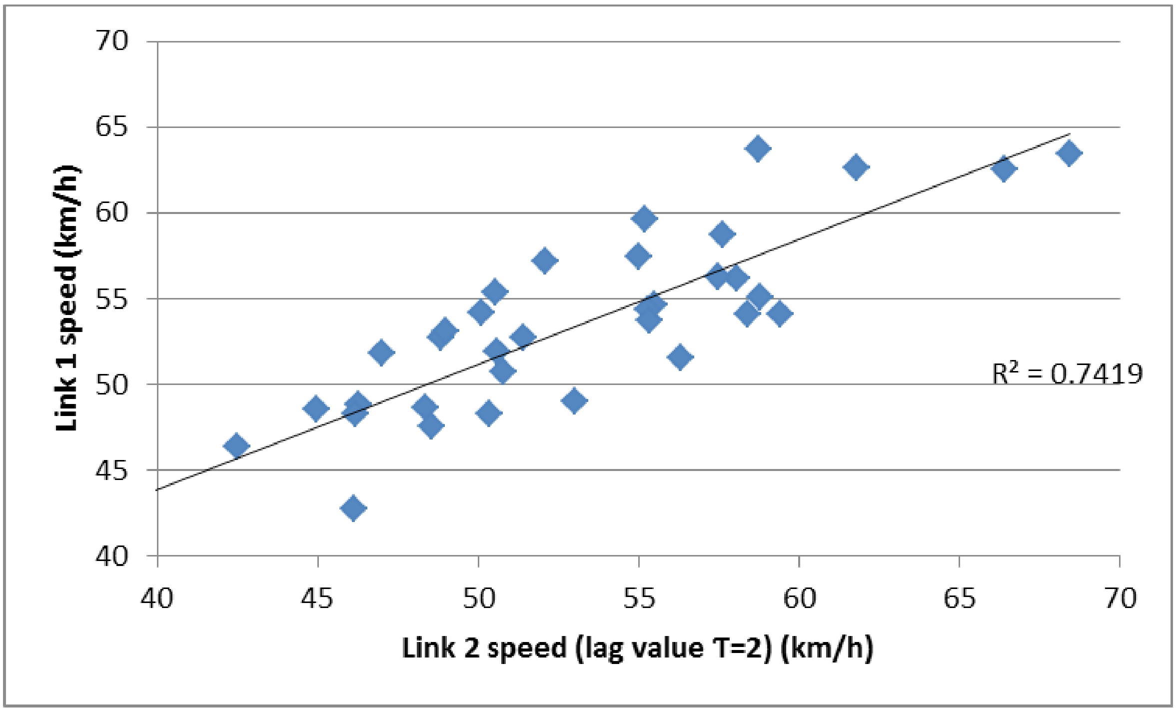

Figure 3.6: Regression of sample data with Ƭ=-2

Figure 3.6 clearly shows that there is a good fit between the two data series, once the lag has been corrected. The coefficient of determination (R^2) is shown on the graph to be 0.74 which indicates a reasonable closeness of fit (a value of 1 would be a perfect fit). Furthermore, the F statistic for the sample data is 94.9 which is substantially higher than the critical-value of 4.14, indicating that it is highly unlikely that this relationship occurred by chance (in fact, the probability of this match occurring by chance is calculated as 0.00001%).

The above calculation can take a long time when done for large datasets such as that used in this study. If every link was compared against every other link, this would require nearly 185,000 correlation tests to be undertaken. Clearly it is impractical to undertake that number of tests, particularly when a range of time offsets between each link pair is also considered.

A small number of links were therefore chosen at random to be the focus of the initial analysis and the correlation function was used in an attempt to identify the "lag" and the strength of the relationship between those links and every other link. Some typical results of are provided below in Table 3.2.

Table 3.2: Indicative results of preliminary correlation analysis

Link A	Link B	Lag value with max correlation	Correlation result	R-squared value
428	10580	Ƭ=-1	0.45	0.09
428	109	Ƭ=-1	0.68	0.36
10080	652	Ƭ=-1	0.34	0.02
10080	38	Ƭ=-1	0.50	0.13
677	212	Ƭ=-1	0.38	0.03
677	223	Ƭ=-1	0.56	0.25

It can be seen from the table above that the correlation between links was generally very poor with typical R^2 values of 0.1 or less. The maximum R^2 value returned by these initial results was around 0.4 which still indicates a poor goodness of fit. Most link pairs examined in this analysis indicated that the optimum lag value should be one time step (or 5 minutes) despite the fact that some of the pairs were at opposite ends of the road network.

On the basis of these results, several further attempts were used to identify relationships between pairs of links, this time by selectively grouping links on the basis of their proximity. Unfortunately, none of these tests provided any significant improvement on the results from the initial analysis.

The results from the second, more selective analysis were surprising since the link pairs used in these tests were mostly adjacent to each other, and it was expected that the correlation should have improved, if only slightly. After further examination of the analysis, it was concluded that the reason for the equally poor results was because the cumulative journey time along the link pairs was shorter than the 5 minute observation interval in nearly all cases. The analysis did not compare data with a zero lag value, as this

relationship (whilst potentially significant) is not useful for the purposes of predicting speeds on the downstream link. This is because by the time data has been collected for a particular link and the predictive calculations undertaken, the vehicles to which the prediction is relevant will already have passed through the link.
Further pairs of links were therefore selected, this time focussing on longer links were journey times were likely to be in excess of 5 minutes. Two link pairs were identified which met these requirements and the results of the correlation tests on these link pairs are shown below in Table 3.2.

Table 3.2: Results of correlation analysis on long links

Link A	Link B	Lag value with max correlation	Correlation result	R-squared value
243	247	Ꚍ=-1	0.64	0.30
262	266	Ꚍ=-1	0.624	0.35

The table above shows that the R^2 values are still low for these links indicating that there isn't a good correlation between observations on these link pairs. Despite this, it was decided to attempt to fit an ARMA model to this data.

Numerous different types of ARMA model were considered although most were not appropriate for this problem. In the end, three different ARMA models were investigated, all of which had been used in previous research: the general Auto-Regressive Integrated Moving Average (ARIMA), Auto-Regressive Distributed Lag (ARDL) and Vector Auto-Regressive Moving-Average (VARMA). In fitting an ARMA model to the observed data, there are a number of parameters which must be estimated. This is done using an optimisation routine which attempts to find those parameters which minimise the residual errors between the model and the observed data. This is a complex process and is undertaken using specialist statistical software. The output from the analysis is a measure of the statistical significance of the model and a summary of the residual errors. As expected, all three ARMA methodologies produced similar results to those already reported by the simpler correlation test. Furthermore, the results indicated that there were a large number of unexplained residual errors which indicated that the models were not a good fit for the observed data.

A further test was undertaken which included weather data in an attempt to reduce the residual errors. This test proved difficult as the analysis then had to be undertaken simultaneously on multiple variables, and also necessitated limiting the analysis to the period where the two data sets overlapped (i.e. June 2012 to Jan 2013). In the end, the complexities of incorporating the weather data meant that only the ARIMA method could be used and the output from this analysis were worse than those achieved previously.

Further analysis of the data revealed that there were significant fluctuations in average vehicle speeds on some of the links used in the analysis. For example, the speed on link 262 varied between 20kph and 100kph; given the distance between this link and the downstream link, 266, the "lag" between the two links could vary from 2 minutes to over 10 minutes. This would mean that the time offset required between the two offsets would vary from 0 to 2 time steps. This could explain the difficulties in fitting the ARMA models to the observed data. Additional tests were undertaken which attempted to calculate the time offset between links more accurately based on vehicle speeds. Initial spreadsheet testing of this hypothesis seemed to indicate a potential improvement in correlation but unfortunately, it ultimately proved impossible to manipulate the data in such a way that it could be input to the ARMA model.

Generally speaking the analysis into the use of ARMA models proved to be inconclusive, with very weak statistical relationships best described as "tenuous". It can only be concluded that this method is not best suited to the topography of the Auckland road network and the data available. Previous studies which have examined these techniques appear to have mostly concentrated their efforts on simpler networks, and it may be that the method may work better if analysis is confined to a long, single stretch of road with detectors spaced at regular intervals.

During the analysis into the ARMA methodology, the only significant results were obtained when data from a particular link was compared against itself (auto-correlation). For example, the autocorrelation of link 296 produced an R^2 value of 0.88 for a lag of 5 minutes. Several other links produced similar results. This may simply be because the speeds on these links did not fluctuate significantly. Nevertheless, this may indicate that there are cyclical patterns on these links which could be utilised using a simpler methodology such as the non-linear time series analysis which is described below.

Expansion of the time-series methodology

Evidence gathered during the Literature Review indicates that non-linear time-series analysis can give a more accurate prediction than the ARMA methodology explored above. However, this method for predicting travel times is severely limited in terms of the achievable prediction horizon. The method only allows travel times to be predicted for one "time-step" in the future. The time-step itself is dictated by the sample intervals of recent historic data. So if data is read in 5 minute intervals, the model can only predict 5 minutes into the future. To predict 10 minutes into the future the data would need to be aggregated into 10 minute intervals, but the accuracy of the predictions then suffers due to the reduction in the detail of the historic data.

There have been a number of papers which have examined ways in which the short-term prediction horizon could be extended slightly (for example to 10, 15 or 30 mins); ultimately, these have all been limited in their success. However, it may be possible to make larger increases to the prediction horizon (i.e. to a day or even a week). Whilst this may seem counter-intuitive given that accuracies decrease significantly as the horizon increases, once it is extended to 24 hours more patterns start to emerge which could be predicted. Certainly, some of the research into ARMA models has concluded that the best "lag" value is 24 hours (in other words, the travel time on a particular link can be better estimated from yesterday's data than from data 5 or 10 minutes ago). Using the non-linear method in this way would mean that the travel times at 9am next Monday morning are predicted based on a time series analysis of travel times at 9am for the last *n* Mondays.

To explore this issue, time series were created by taking the speed on each of a number of links at a specific time each day (e.g. 0830, 1300, 1900), or a specific time on a specific day of each week (e.g. Monday 0900, Saturday 1230). The analysis into the ARMA methodology had already demonstrated that there may be a cyclical pattern to the data over short periods, and further analysis showed that this may also be true for daily or weekly time offsets.

The non-linear method works by predicting the change in the Hölder exponent (which describes the concavity or convexity of the time-series curve at a given point). Without building a fully working predictive model to test the method, it stands to reason that any time series in which the Hölder exponent changes only gradually should lend itself to this method. In other words, if the Hölder exponent has a similar value at the same time each day this would be indicative of a similar profile of vehicle speeds (even if the absolute speed values varied). Examination of the weekly time series shows this to be the case, with seasonal changes resulting in small week-on-week changes, bar a small number of exceptions. However, the daily time series tests were less conclusive, largely due to the large differences between weekday flows and speeds compared to those at the weekend.

Overall, the examination of the daily and weekly time series indicate that the non-linear time-series analysis method may be able to provide reasonably accurate medium-term predictions.

Regression against weather data

In addition to the two methods above, the weather data has also been compared against journey times to determine if there is any statistical relationship between the two datasets. Initial regression analysis indicates that there may be some correlation between journey times and the limited weather data provided. The strength of the relationship is not sufficient that the weather could be used as the primary means to predict journey times, but combined with another method (for example the time-series analysis) this could increase the accuracy of predictions.

However, there is substantial difficulty in combining the two techniques, i.e. to adjust the time series output to take account of weather, there is a need to similarly “un-adjust” the historical data that is input to the time series analysis. As a result of this complexity and the poor quality of the weather data, this has not been attempted as part of this study.

3.5 Data review and modelling conclusions

We have examined the data that has been made available, investigated two main methods for predicting journey times and also looked into the influence the weather has on journey times.

The Auto-Regressive Moving Average (ARMA) method could potentially be used to increase the accuracy of predictions in the short term (5 to 30 minutes) thereby improving traffic management. However the findings from our analysis are generally inconclusive.

The non-linear time-series analysis method proposed by D’angelo et al is more accurate at very short term predictions (5 minutes), but is limited in its prediction horizon. We have looked into the possibility of extending this method to have prediction horizons of 1 day or 1 week. Initial analysis comparing some of the data to the model function indicates that it could be possible to build a model based on this principle, thereby improving the accuracy of journey times for the purposes of travel planning.

Based on the analysis to date, there is little confidence that the ARMA methodology would provide any improvement over existing prediction techniques, and hence there would be no benefit in undertaking any additional analysis on this methodology. The conclusion of this stage of work was that the non-linear time-series methodology should be taken forward to the next stage which is to build a test model around the methodology. This process is described in Chapter 4.

4 Model development

4.1 Overview

In the final stage of the project, a model will be constructed around the preferred methodology. This model will take inputs from the available data as if they were “real-time”, these inputs will be processed by the model algorithm and a prediction of journey times will be made. These journey times will then be compared against the actual observations so that the accuracy of the model can be determined.

4.2 Predictive function

Following the findings of the Literature Review and the Data Review, the chosen methodology is based on a non-linear time series analysis technique proposed by D’Angelo, Al-Deek and Wang in their 1999 paper “Travel-Time prediction for Freeway Corridors”. Rather than attempting to predict values in the time-series directly, the study proposed that the curvature of the time series could be predicted more accurately and used to derive the next value in the series

For each value in the time series, the following equation describes the curvature of the data:

$$\propto_n = \log_3 \frac{\left(S_{n-1} + S_n + S_{n+1}\right)}{S_n}$$

Where S is the average vehicle speed on the segment, and the subscript denotes the time step (n is the current time step, n-1 is the last time step and n+1 is the next time step). In the equation, α is known as the Holder exponent. If the Holder exponent is 1 then the values in the time series are linear (i.e. they are increasing or decreasing by the same amount in each time step). If the Holder exponent is less than 1, then time series is “curved” downwards at that point; similarly values greater than 1 represent points where the time series is “curved” upwards. Figure 4.1 below shows some example data curves with Holder exponents equal to, less than or greater than 1. It can be seen that the Holder exponent has no bearing on whether the data series is increasing or decreasing, but simply the rate of change in the data’s slope.

As the equation used to calculate the Holder exponent requires data from both the preceding and following time steps in the series, it can be seen that the Holder exponent can only be calculated for historic observations. However if the value of the Holder exponent could be predicted for the current time step, then the next value in the time series can be derived.

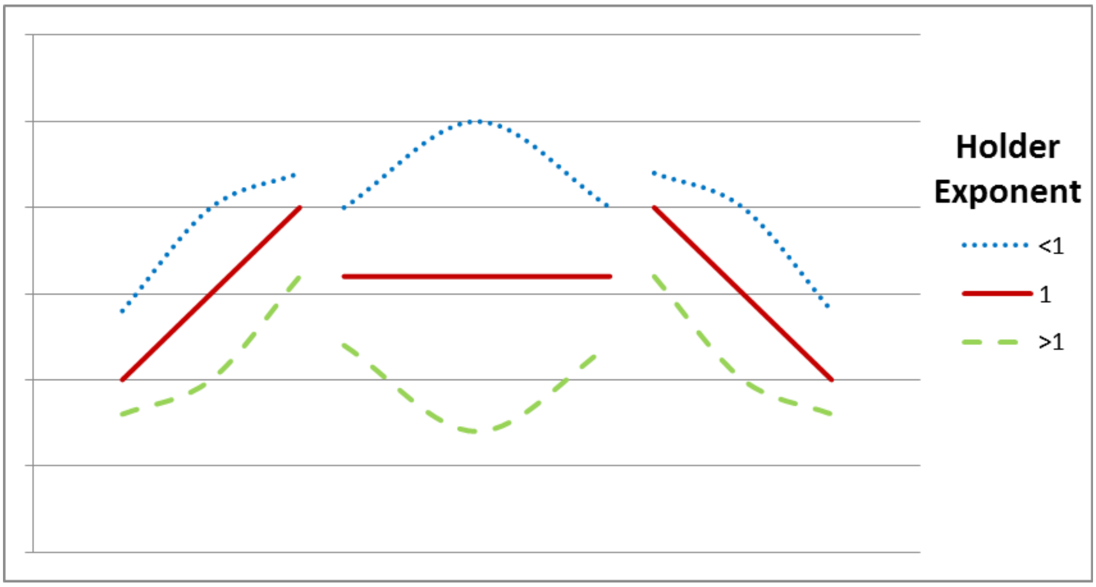

Figure 4.1: Illustration of different curves and their equivalent Holder exponent.

The Holder exponent itself can be predicted by examining recent historic data and how this changes over time. The assumption is that the change in the speed of vehicles over time is similar to a wave function. Although the amplitude of the wave may vary, the basic shape of the wave (i.e. the rate at which the average vehicle speed changes) should always be similar for each road segment as this is a function of both the road geometry and driver behaviour which can be considered constant. The model examines the recent time series data and attempts to identify where similar wave patterns have appeared in the past and use this as the basis of the prediction. Figure 4.2 below shows how this process works.

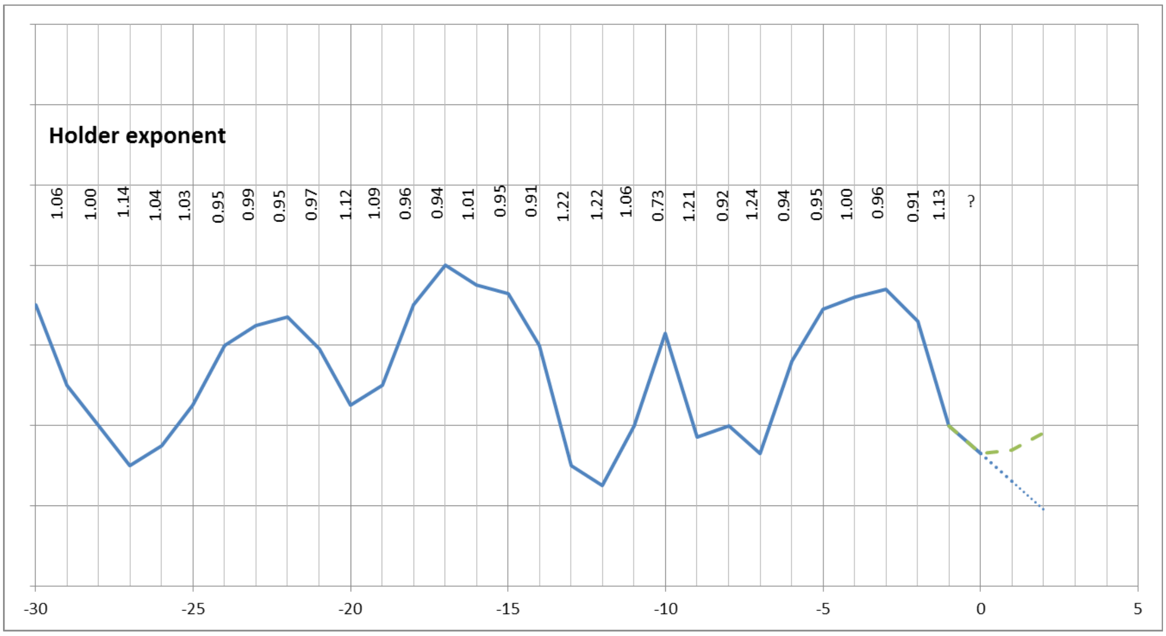

Figure 4.2: Sample data to illustrate model behaviour.

Figure 4.2 provides some sample data which can be used to demonstrate, in simple terms, how the model works. The process begins by firstly calculating the Holder exponent for each of the 30 most recent observations. These are recorded above the data series in Figure 4.2. The model then takes the Holder exponent for the previous observation (which is 1.13 in this case) and compares it to the exponents of the earlier data. In the sample data, there are two occurrences, a Holder exponent of 1.14 at S_{n-27}, and 1.12 at S_{n-20}. The model then examines the values of the Holder exponents that immediately follow these occurrences (i.e. at time steps S_{n-26} and S_{n-19}). In our example, these have the values 1.04 and 1.09 respectively. On the basis of these new values, the model can predict that the Holder exponent for the current time step is going to be similar. In this case, we can assume that this value is likely to be around 1.065.

Once the most likely value of the Holder exponent has been predicted for the current time step, the following equation can then be used to calculate the corresponding speed value for the following time step.

$$S_{n+1} = (3^{\alpha_n} - 1) S_n - S_{n-1}$$

It can be seen from Figure 4.2 that if a simple linear prediction was made, the average vehicle speed would be predicted to decrease in the next time step (as shown by the blue, dotted line to the right of the figure). However the prediction returned by the model (indicated by the green, dashed line) shows a slight increase in vehicle speeds which is much more likely.

4.3 Model construction

The model has been constructed within an Excel spreadsheet for the purposes of demonstration. Excel has been used so that each of the steps within the process can be seen clearly, allowing for easier debugging. A more efficient process would need to be created if this process was to be used in a production environment.

The model undertakes predictions for individual links, but over the entire date range for which data has been provided. This model has been created in this way to reduce the volume of data which needs to be stored within the Excel spreadsheet. Again this differs from how the model would work in production, where only the most recent data would be imported but for all links; the prediction would then be undertaken for a single time step for every link.

Within the spreadsheet model there is a drop-down list containing all of the links within the study area. Once a link has been selected, the data for that link is read into the spreadsheet from the main database, using a VBA macro. Any gaps in the data are infilled, interpolating speed data and journey time data where necessary.

A second VBA macro is then used to step through each of the values within the time series and undertakes the following process for each time step:

- A table within the spreadsheet reads the vehicle speed and journey time data from the 30 observations prior to the current time step and the Holder exponent is calculated for each.
- A second table is used to identify those Holder values which are similar to the last time step and then to determine the most likely value of the current time step.
- The speed value for the next time step is then calculated.

A comparison is made between the prediction and the actual value so that the accuracy of the model can be measured.

These comparisons are stored in a separate results table for later analysis. The model then moves on to the next time step and the process is repeated for the entire data set.

The predicted value of the Holder exponent at each time step can actually be determined in a number of ways and we have created several versions of the model which use different methods in an attempt to find the best one. In the original study by D'Angelo et al, a method was used which grouped the earlier values of the Holder exponent into a number of equal intervals, counting how many of the time steps fell into each interval to determine the most frequently observed value (effectively creating a Markov model at each time step). We have found through experimentation that simply taking an average of the appropriate values provides similarly accurate results.

The model is set up to predict speeds for the next value in the time series only. This means that the prediction horizon of the model (how far into the future the vehicle speeds can be predicted) is limited to the time interval of the observed data. As the data is recorded every 5 minutes, this is the default prediction horizon used in the model (and was also the prediction horizon used in the original study). However, we have set up the model so that, for example, every second observation is used to provide a 10-minute time step and therefore a 10 minute prediction horizon. The model allows any time step interval / prediction horizon between 5 minutes and 1 week to be used, although the accuracy decreases as the prediction horizon increases as explained in Section 5.3 below.

5 Results

5.1 Testing process

The model has been tested for all of the links for which data was available. For each link, predictions have been made for every five minute interval for a one week period. The reasons for making predictions for a single week were due to limitations in processing power; running predictions for the whole network for a single prediction horizon for one weeks' worth of data took more than 24 hours of computing time.

This is good news for the model, as it means that real-time data can be processed in slightly less time than the observation interval. However when attempting to test the model using a large dataset, the runtimes were unacceptably large. If the entire dataset had been used, it would have taken nearly two months to extract the results for a single prediction horizon, or more than a year for all of the tests we were proposing.

The model has been tested using seven different prediction horizons, ranging from five minutes to one week. To calculate the accuracy of predictions a comparison has been made between the predicted value and the actual observed value for each time-step and each link. This difference is expressed as a percentage error. The results for each link are summarised for each link in two ways:

The average percentage error on each link over all time steps (Measure A).

The percentage of time steps on each link where the percentage error is less than 10% (Measure B).

The results of these tests are summarised below

5.2 Initial findings

The model was initially tested using a five-minute prediction horizon, so that its accuracy could be compared against the original study. The results of these initial tests show that, for around one sixth of the links, the model is able to accurately predict future journey times (more than 85% of time steps have errors less than 10%). The results for a sample of links are provided below in Table 5.1.

Table 5.1: Model accuracy (5-min prediction horizon): most and least accurate links.

Link Ranking (by Criterion B)	Carriageway Segment ID	A) Total Average Error	B) Percentage of predictions with <10% error
1 (most accurate)	400	0%	100%
43 (90th percentile)	120	6%	85.3%
107 (75th percentile)	78	6%	76.4%
215 (median)	235	9%	64.5%
322 (75th percentile)	10090	14%	51.6%
430 (least accurate)	10099	2971%	17.5%

The above table is deceptive in that it appears to show that the total average error (Measure A) is low for the majority of links, with the 75th percentile link having a total average error of just 14%. However, this is because the table ranks each link in terms of their performance on Measure B. On close inspection of the underlying results, it is apparent that links with similar Measure B performance can have significantly different results for Measure A. For example, the median link has an average error of 9%, but the link ranked immediately above this has an average error of 93% despite having an almost identical result on Measure B (64.5%). In total, only around 43% of links have less than a 15% total average error for Measure A. Figures 5.1 and 5.2 show this more clearly by illustrating the distribution of results for both measures.

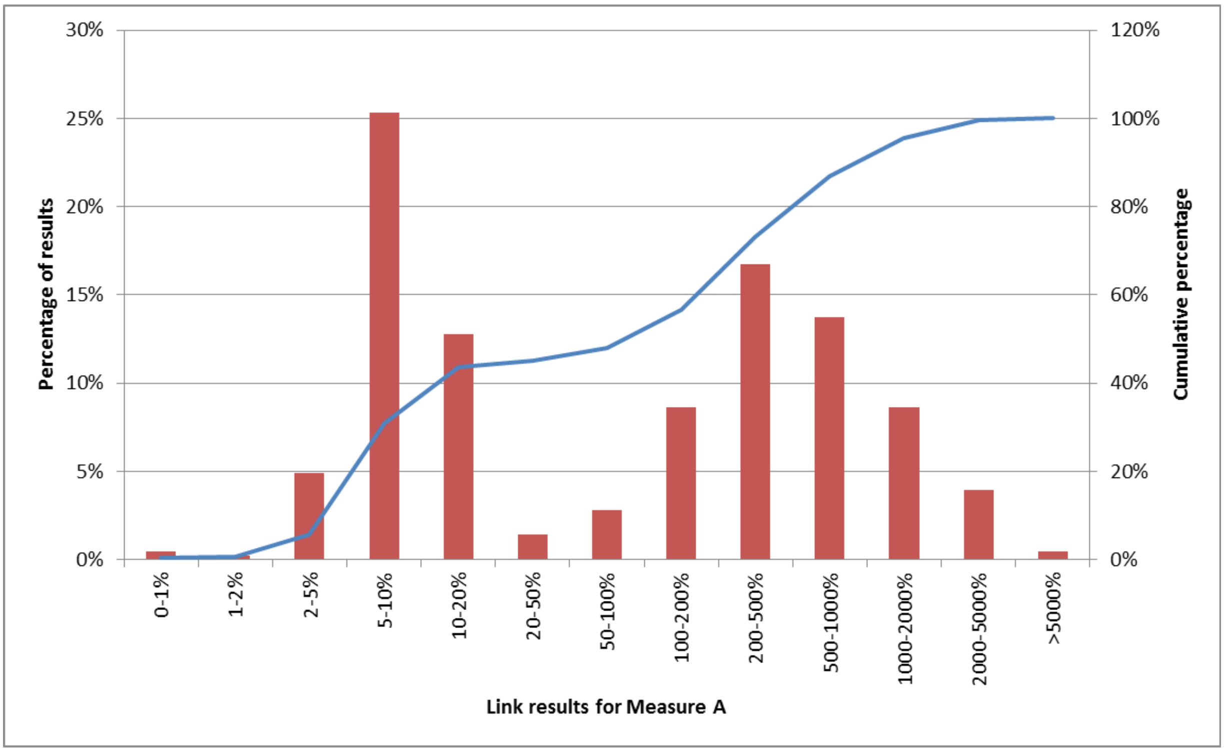

Figure 5.1: Distribution of results for Measure A

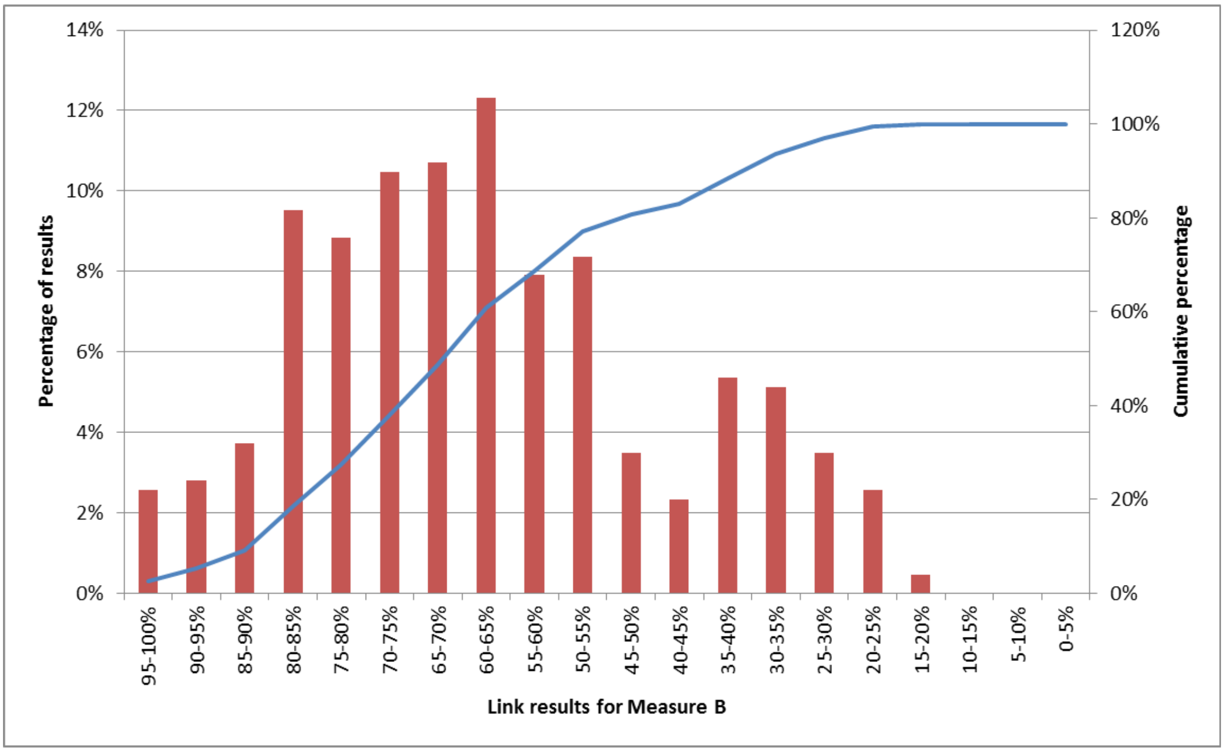

Figure 5.2: Distribution of results for Measure B

It can be seen from the figures above that whilst there is a normal distribution of results for Measure B, there are two distinct groups of links when they are compared on Measure A. Nearly half of the links (188) have an average error of less than 20%, whilst a similar number (224 links) have an average error of more than 100%. There are only 18 links which have an average error between 20% and 100%, suggesting that there may be some characteristic which prevents the prediction algorithm from working on certain links.

Some analysis has been undertaken to establish the reason for the poor results, by undertaking a comparison of results against known characteristics of the links (e.g. link length, traffic volume). The results of this analysis indicate that there is no single characteristic which explains why some links should give much better results than others. A summary of some of this analysis is presented below.

Figure 5.3 demonstrates how the predictive ability of the model (using Measure A) compares to the link length. It is logical to assume that percentage errors would be higher for shorter links (as absolute errors would be larger relative to the observed journey times), and that this may be one reason for the contrasting results. However, the evidence does not support this and Figure 5.3 clearly shows that accurate predictions are possible for short links and that, equally, some longer links are producing large percentage errors.

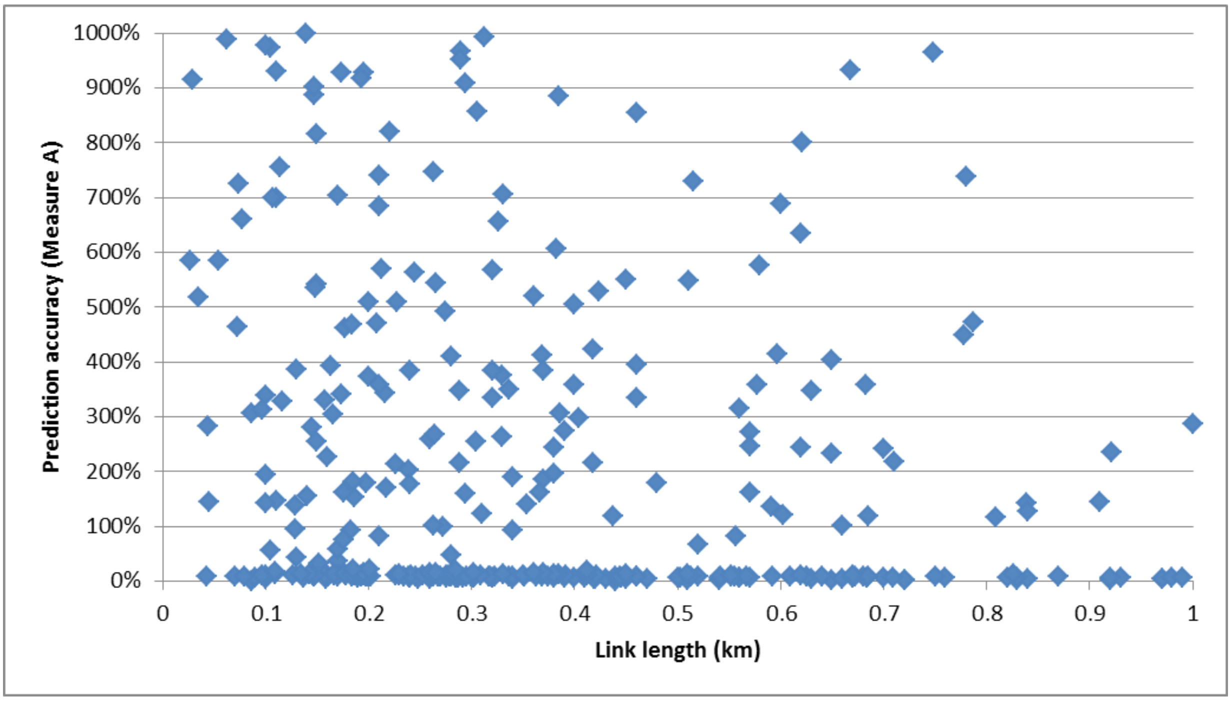

Figure 5.3: Comparison of model accuracy versus link length

There are many link characteristics which are not quantified in the data provided, for example the proximity of the link to major intersections. In an attempt to identify some of these characteristics, and to determine if they are a determinant factor in the model performance, the model results were plotted according to the link location. Figure 5.4 below shows graphically how prediction accuracy differs by location across the network. It appears from this figure that the links with accurate predictions tend to be located between major intersections, indicating that the reason for poor prediction results on some links may be related to the intersections. This may suggest that some links are simply more prone to congestion and hence the journey time on these links is much more varied and difficult to predict.

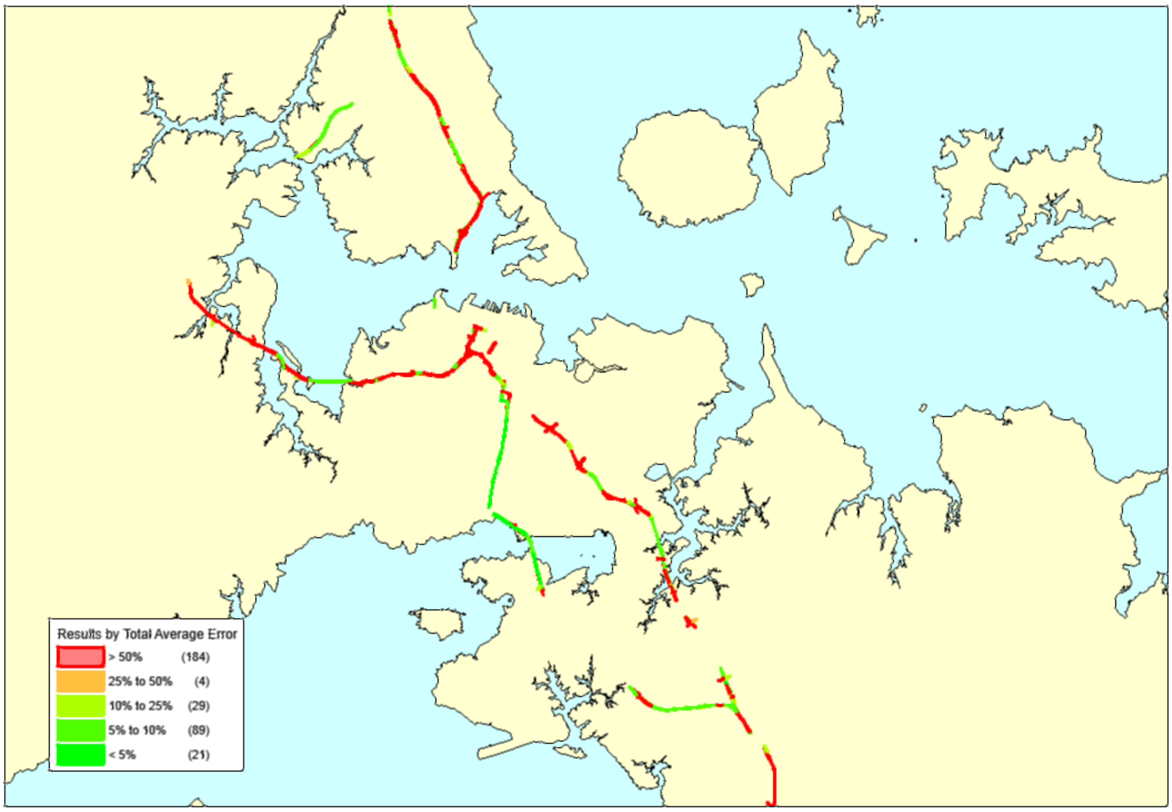

Figure 5.4: Model accuracy (5 min time step): prediction error by location

Another analysis tool that was used was to look at how prediction accuracy varies over a single day, and compare this for a link with good predictive results against another link with poor results. Figure 5.5 shows data for a segment with an average of error of 1%, compared with Figure 5.6 which shows data for a segment with an average error of 20%.

It can be seen from these two figures that the variation in journey times is broadly similar for both links, with constant fluctuations throughout the day. Figure 5.6 shows that there are three noticeably large prediction errors which could account for the increased average error. However, this doesn't explain why only 38% of the predicted values are within 10% of the observed values for this link segment, compared with 98% for the link segment shown in Figure 5.5.

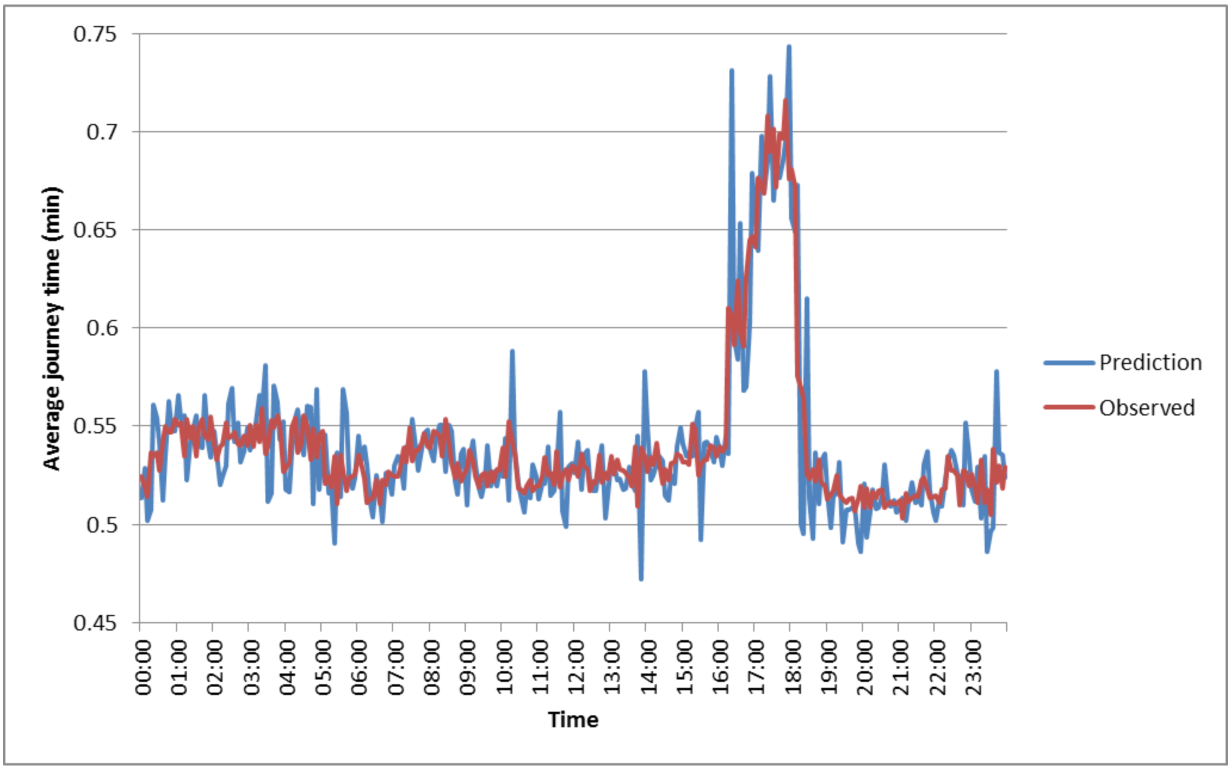

Figure 5.5: Model accuracy (5 min time step, 1 day period): predicted times vs observed times, segment 709.

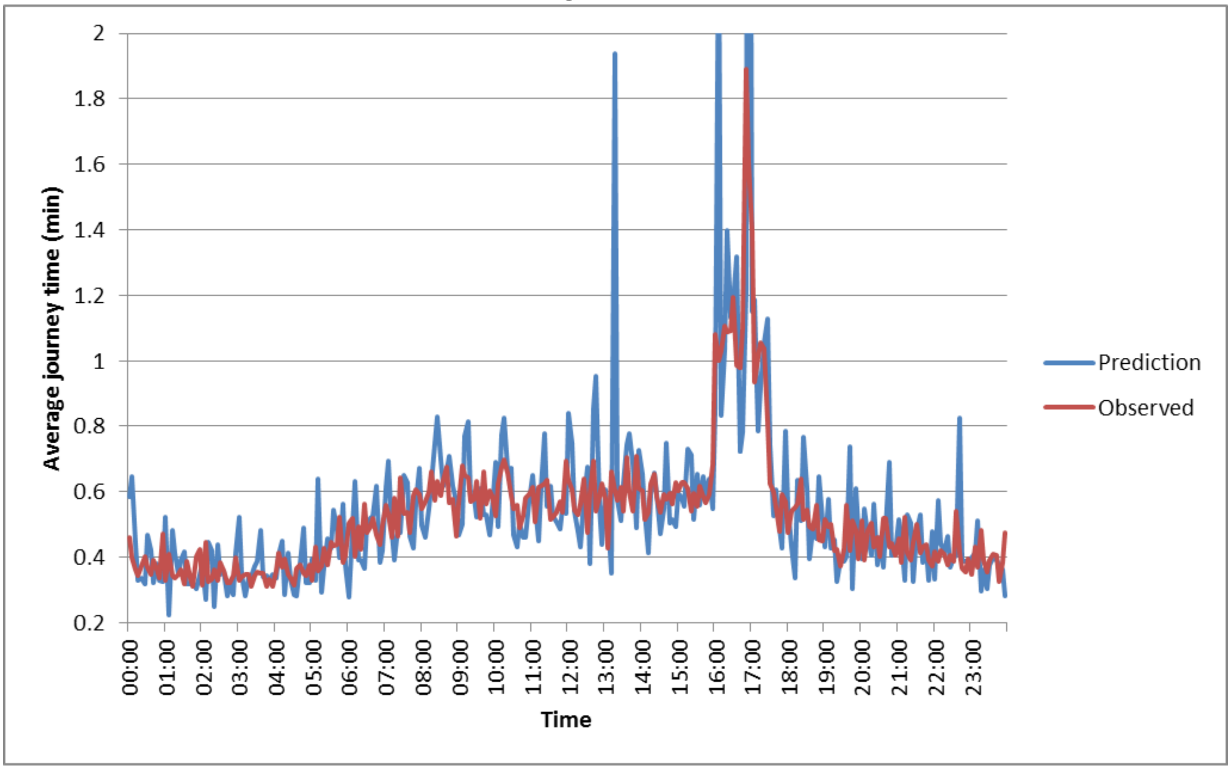

Figure 5.6: Model accuracy (5 min time step, 1 day period): predicted times vs observed times, segment 229.

Closer examination of the two plots above highlights two interesting aspects which may give clues to the model's incongruent results. Firstly, in both figures, the prediction algorithm tends to exaggerate every peak and trough, suggesting that it is oversensitive to sudden changes in journey times. There is potential for some improvements to be made to the model algorithm to dampen this effect, increasing model accuracy for all links.

Secondly, although the average journey time across the day is similar for both links, the journey time on the "poor" segment (Figure 5.6) drops between the hours of 11pm and 5am. Examination of the underlying data shows that this corresponds to a significant reduction in vehicles to just one or two vehicles every 5 minutes. The same analysis of the "good" segment shows that it carries a much higher volume of traffic all day and that the overnight volume remains relatively high. This indicates another potential reason for the contrasting results. Comparison of the prediction results for individual time steps on the "poor" link against vehicle flow for the same time step has been undertaken and this is shown in Figure 5.7. Unfortunately, there is no apparent correlation, but the relationship between link flow and prediction accuracy could be explored further.

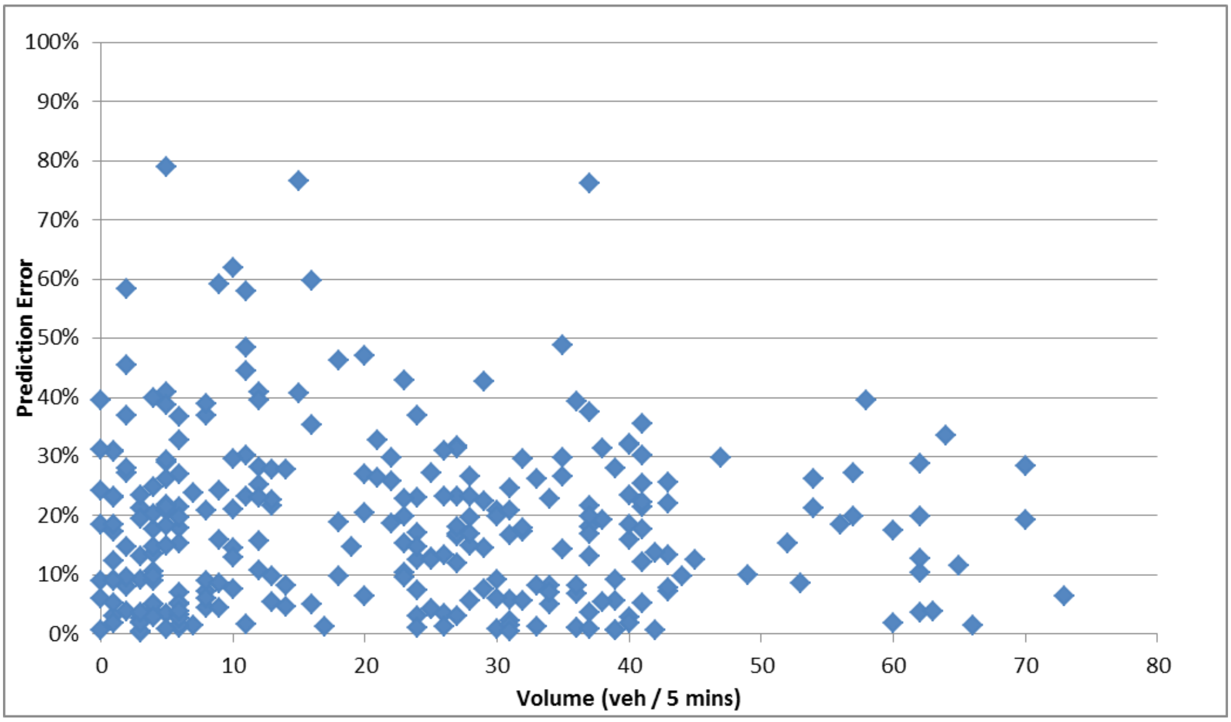

Figure 5.7: Model accuracy (5 min time step, 1 day period): prediction error vs segment flow, segment 229.

In summary, the model is able to make accurate predictions for many of the links, but there is some characteristic of the other links which results in poor prediction results. Analysis of the links indicates that there is no single characteristic causing this, but that the proximity of the link to intersections and low volumes may both be contributing factors.

5.3 Testing for longer prediction horizons

All of the links have also been tested for a variety of prediction horizons and the results of this analysis indicate that for those links where journey times can be accurately predicted over a 5-minute prediction horizon, reasonable accuracy can also be achieved over a longer prediction horizon. Figure 5.8 below shows how five minute accuracies compare against 1 week accuracies for all links.

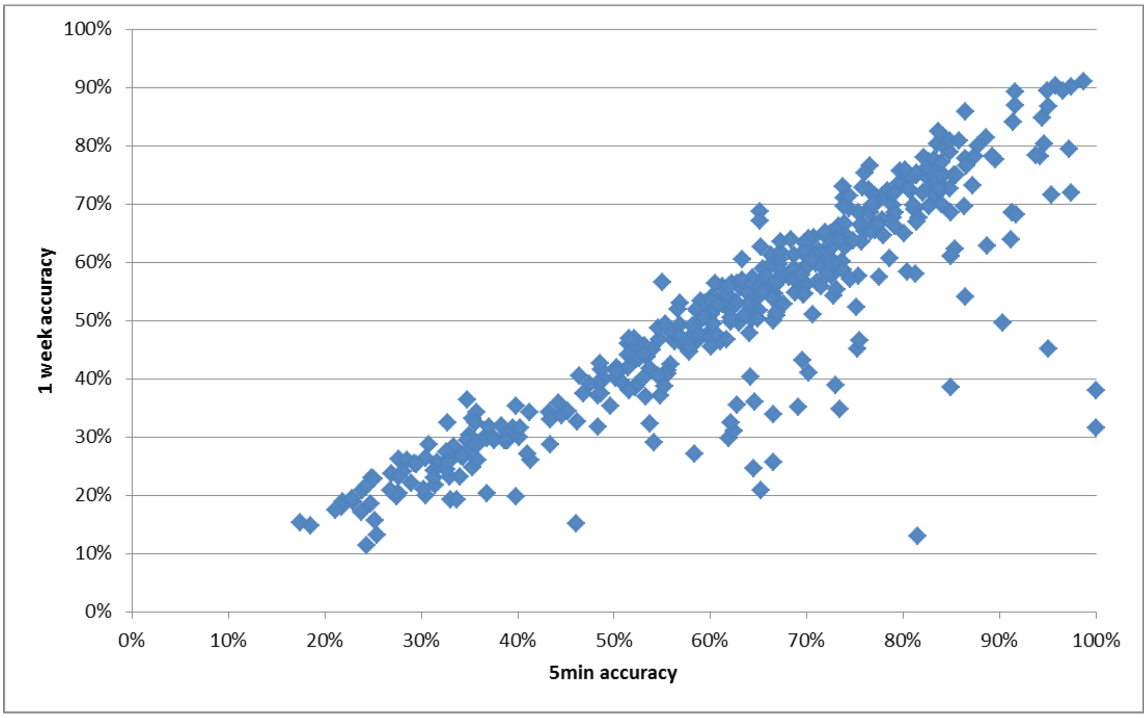

Figure 5.8: Impact of prediction horizon on model accuracy (Measure B)

The figure above clearly shows a good correlation between 5-minute accuracy and 1-week accuracy, with the latter being slightly lower in nearly all cases.

To examine how the accuracy of the model changes as prediction horizon increases, the results of an individual link are presented below in Table 5.2. This table clearly shows that the accuracy of the modelling deteriorates slightly as the time step interval increases.

One interesting point to note is that when the time step interval is set to one day, there is a significant reduction in model accuracy. This is because the prediction algorithm is taking consideration of journey times at the same time every day for the previous 30 days; this will include a mixture of weekdays and weekends. If the model could be adapted to only consider the previous 30 weekdays, this result is likely to improve. We have not been able to test this hypothesis at the present time.

The table does show, however, that even with a time step of 1 week, the model can still be reasonably accurate with an average error of just 8%. If the issue preventing accurate predictions for certain links can be overcome, this result indicates that it could be possible to predict journey times up to a week in advance for all links.

Table 5.2: Impact of prediction horizon on model accuracy (Link 715)

Time step interval	Average error	Percentage predictions <10% error
5 mins	2.29%	96.86%
10 mins	3.98%	90.77%
15 mins	4.94%	87.21%
30 mins	6.15%	83.18%
1 hour	7.88%	77.67%
1 day	24.68%	67.26%
1 week	8.00%	73.49%

6 Lessons Learnt

This section outlines the key lessons which have been learnt from this research. Is intended to provide NZTA and other transport decision makers with guidance on the further development of a travel time predictability system.

6.1 Data

Data Quality

The accuracy of the data available provided some issues and limitations with developing a successful model.

The traffic data (flows, segment speeds) was generally of good quality, although there are large gaps in the data where detectors were clearly out of action for long periods. Only one link (out of nearly 500) contained a full 13-month period of data with no missing observations. In total, 5.2% of all observations across all links were missing from the data, although a small number of links had detectors which were malfunctioning or switched off for several months and accounted for most of the missing data. The majority of links had only a few days' data missing at some point over the period. Most links have between 100 and 1,000 missing observations. Although, these seem minor, when undertaking a complex modelling process and looking for patterns between data sets, it is critical to have accurate and reliable data.

The weather data was of very poor standard. The data is incomplete, with rainfall and solar radiation provided for just one of the five sites. Furthermore, the data covered the wrong period (June 2012 to June 2013 rather than January 2012 to January 2013). This made it very difficult to make any use of the data provided.

There are also issues with the incident data. It was always expected that unplanned events would not be very well recorded, and examination of the data shows this to be the case with just 401 incidents over a 13-month period. More than half of the recorded incidents (208) were listed as "Cautions" and were usually associated with maintenance issues such as lack of street lighting or mud on the road which did not necessarily equate to delays on the network. With regard to planned events, the dataset does contain a sufficiently large number, but many of these are for full road closures and are not very specific with regard to time or location, or if they even went ahead as planned.

Any future work undertaken would need to ensure that accurate and reliable data could be provided across the network of interest.

Rationalisation of Base Data

The investigation of different forms of relationship between modelled and observed journey time has, in the first instance, involved the use of the entire data set. It is acknowledged that due to gaps in the data and some doubts over the accuracy and reliability of the data, that this could have influenced the strength of any resulting relationship. If additional research does eventuate on this topic, it is recommended that appropriate time be spent on sorting the data to produce sub sets defined by:

- Weekday and weekend
- Highway links free from the influential effects of ramps or intersections on neighbouring roads
- Links categorised by speed limit
- Links categorised by different ADT range bands
- Incident and non-incident data sets
- Inclement weather

It is felt that if such a rationalisation of data could be undertaken at the front end of the project, stronger modelling outcomes may well be achieved and it would be possible to better understand the key parameters influencing the accuracy of travel time predictions and their effects.

6.2 Statistical Modelling Approach

The literature review which was undertaken identified two main prediction models which could be used to undertake travel time predictability analysis. These included:

A Simulation method.

A Statistical method.

The Simulation method requires the development of a transport model where traffic demand is simulated and calibrated in real-time based on incoming traffic data. The second method relies on the estimation of mathematical functions based on historical data which calculate future traffic speeds based on current network conditions.

It was clear from the review that there are significant limitations with the Statistical method, but that it is much simpler and quicker to implement and easier to maintain. For study areas which are suited to the Statistical method then this would offer the best option. This would include simple strategic trunk or motorway networks with limited route choice and signalised junctions. For more complex networks such as city centres or other urban areas, the Simulation method would be recommended.

The study network adopted for this research is a motorway corridor with fixed access and egress points, which is perfectly suited to the Statistical method. The Simulation modelling approach is unlikely to add much value without significantly increasing the resource requirement to build a complete traffic demand model. As such, it was recommended that the Statistical method be used for this study.

Although the motorway network could be classified as simple in comparison to a city centre, the Auckland motorway network does include some complicating factors; these include closely spaced intersections, ramp signalling, variable message signs and speed limits in some locations, and frequentchanges to the motorway structure (lane drops, gains etc). Although the use of a simulation model would require significant data collection and calibration, given the complexities of the network mentioned above and the outcomes of this research, future studies may be better focusing on the simulation model approach.

6.3 Motorway Network

For the purposes of this study the full Auckland Motorway was investigated. This allowed full assessment of all the data and all the links, and improved the likelihood of identifying a location where travel time predictability could be assessed and analysed further. Given the complexity of the network and significant data requirements, any future studies could focus on a smaller section of the motorway network. Although this may limit the likelihood of finding correlation between links and data, it will allow more intimate assessment of the factors which are impacting the ability to predict travel times along the network. Again, the success of this approach is very much dependent on the quality and reliability of the data available.

There were a wide range of factors which impacted on the ability of the model to predict travel times. As part of this research, investigations have been undertaken on the features which limited the success of the model, it is believed the following features of a motorway network would be best suited to a statistical modelling approach:

Simple network – minimal on and off ramps and other motorway features which change traffic patterns

Full, Accurate and reliable data – Correlation of data sets is critical to the success of the model and travel time predictability.
Link Length – the length of the link needs to exceed the time interval being assessed to allow improved possibility of correlation, for example, in a 100kph zone, and an interval of five minutes, the link lengths should exceed 8.33km

7 Conclusion

7.1 Current state of research

A comprehensive Literature Review has been undertaken which examined a wide range of previous studies and research. The outcome of this review indicated that of the two potential solutions the Statistical approach was the most suitable, based on the availability of the data and the client's requirements.

7.2 Data sources

Detailed data analysis has then been undertaken of the available data, using two different statistical methodologies. This research indicated that no conclusive results emerged using the auto-regressive method, but that non-linear time series analysis techniques may provide the basis for a suitable prediction tool.

7.3 Forecasting model

A demonstration model has been built based on the non-linear time series analysis technique. The model is able to accurately predict journey times over a 5-minute prediction horizon for certain links, but is inaccurate for other links. For those links where the model is able to produce an accurate 5-minute prediction, it is also capable of predicting journey times with reasonable accuracy up to 1 week into the future.

7.4 Summary statistics

Analysis has been undertaken in an attempt to identify what factors may influence the model's ability to accurately predict times for certain links and not others. There appears to be no single contributory factor and that there is no clear correlation between model accuracy and traffic volumes or link length. Spatial analysis of the results indicates that links in the vicinity of major intersections may tend to demonstrate poorer prediction accuracies.

7.5 Research report

The overall conclusion of the study is that it is currently very difficult to predict journey times using the statistical method and current data availability with sufficient accuracy across the Auckland network. However, this research indicates that accurate predictions are certainly possible in some situations and that further study may result in improved accuracy across the network.

8 Recommendations

Whilst it is disappointing that the model accuracy was not better for all links, there were still a number of positive outcomes from this research which could be taken forward. It is encouraging that the model was able to predict travel times for certain links with such a high level of accuracy, and it would be hoped that this could be extended to other parts of the network as well.

There are a number of areas where further research could be undertaken, in an attempt to improve the model further:

- There are a number of issues with the underlying data, namely that there are a number of large gaps in the data for some of the links. Sourcing of improved or alternative data sources, or improved interpolation may result in more accurate predictions.
- Further analysis needs to be undertaken on the results of the current model to better understand the reasons for contrasting results. This may need to look at characteristics of the links which are not quantified in the current data set (e.g. proximity to major intersections, number of lanes and/or road geometry).
- Further analysis of the effect of traffic volumes on the prediction accuracy. Clearly, low volumes will be more difficult to predict and at these times it may be appropriate to substitute the model output with an average observed journey time from historical data. It may also be necessary to remove some links from the dataset where journey times are erratic or link lengths are short. A process of combining data from adjacent links to form longer segments may also improve accuracy.
- There are a number of potential improvements which could be made to the model algorithm, by focussing on those links which currently give good results. Additional study of alternative methods for determining the most likely Holder exponent could improve predictions significantly. This is evidenced by Figures 3.5 and 3.6 which clearly show that the prediction algorithm is over-sensitive to sudden changes in journey times.
- There should also be consideration of additional explanatory variables, such as the weather, which have not been considered within this study due to data issues.

If any further study is successful in improving overall prediction accuracy, then it may be prudent to revisit some of the earlier data analysis, particularly the tests which examined the potential of ARMA prediction methods. As discussed earlier, the results of these analyses indicated that the ARMA methodology was not suitable. At the time there was an assumption that a constant "lag" value could not be found due to the varying vehicle speeds and / or segment lengths. In retrospect, the inability to find any significant results may have been due to other link-specific issues explored in this report. It may be the case that, if the dataset is limited to those links with good predictive results from the time series analysis, then more favourable results may also be obtained using the ARMA method.

9 References

. (2008*) Online learning solutions for freeway travel time prediction.* Intelligent Transportation Systems 9, 38-47.

Al-Deek, H.M.; D'Angelo, M.P.; Wang, M.C. (1998) *Travel time prediction with non-linear time series. In Applications of advanced technologies in transportation.* Proceedings of the 5th international conference, ed.s C. Hendrickson, S.G. Ritchie American Society of Civil Engineers, Reston, VA, 317-324.

Ashok, K. (1997). *Estimation and prediction of time-dependent origin-destination flows*. PhD Thesis, Massachusetts Institute of Technology.

Antoniou, C. (1997). *Demand simulation for Dynamic Traffic Assignment*. Master Thesis, Massachusetts Institute of Technology.

Bajwa, S.; Chung, E.; Kuwahara, M. (2005). *Performance evaluation of an adaptive travel time prediction model.* In Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, ed. E. Smid, IEEE, Vienna, Austria, 1000-1005

Ben-Akiva, M.; Bierlaire, M.; Bottom, J.; Koutsopoulos, H. (1997). *Development of a route guidance generation system for real-time application*. Poceedings of the IFAC Transportation Systems 97 Conference, Chania.

Chang, G.L.; Mahmassani, H.S. (1988). *Travel time prediction and departure time adjustment behavior dynamics in a congested traffic system*. Transportation Research B 22, 217-232.

Chen, M.; Chien, S. (2001). *Dynamic freeway travel-time prediction with probe vehicle data: Link based versus path based*. Transportation Research Record 1768, 157-161.

Chien, S.; Kuchipudi, C. (2003). *Dynamic travel time prediction with real-time and historic data*. Journal of Transportation Engineering 129, 608-616.

Chiu, Y.C.; Mahmassani, H.S. (2002). *Hybrid real-time Dynamic Traffic Assignment approach for robust network performance*. Transportation Research Record 1783, 89-97.

D'Angelo, M.P.; Al-Deek, H.M.; Wang M.C. (1999). *Travel-time prediction for freeway corridors.* Transportation Research Record 1676, 184-191.

Florian, M.; Mahut, M.; Tremblay, N. (2001). *A hybrid optimization-mesoscopic simulation dynamic traffic assignment model*. Proceedings of IEEE Intelligent Transportation Systems 2001, 118- 121.

Gentile, G. (2010). *The General Link Transmission Model for Dynamic Network Loading and a comparison with the DUE algorithm*. In New developments in transport planning: advances in Dynamic Traffic Assignment, ed.s L.G.H. Immers, C.M.J. Tampere, F. Viti, Transport Economics, Management and Policy Series, Edward Elgar Publishing, MA, USA, 153-178.

Gentile, G.; Meschini, L. (2011). *Using dynamic assignment models for real-time traffic forecast on large urban networks*. Proceedings of the 2nd International Conference on Models and Technologies for Intelligent Transportation Systems, Leuven, Belgium.

Giacomini, R.; and C. W. J. Granger (2004). *Aggregation of Space-Time Processes*. Journal of Econometrics.

Harvey, A. C. *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge University Press, Inc., New York, 1992.

Hoogendoorn, S.P.; Van Zuylen, H.J. (2002). *Freeway travel time prediction with state-space neural networks: modeling state-space dynamics with recurrent neural network*s. Transportation Research Record 1811, 30-39.

Hu, T.Y.; Chen, L.W. (2008). *Real-time implementation of simulation-based Dynamic Traffic Assignment model*. Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems, 645-650.

Hu, T.; and Ho, W. (2010). *Travel Time Prediction for Urban Networks: the Comparisons of Simulation-based and Time-Series Models.*

Idè, T.; Kato, S. (2000). *Travel-time prediction using Gaussian process regression: a trajectory based approach*. SIAM, 1185-1196.

Kamarianakis, Y.; Prastacos, P. (2003). *Forecasting Traffic Flow Conditions in an Urban Network: Comparison of Multivariate and Univariate Approaches*. In Transportation Research Record: Journal of the Transportation Research Board, No. 1857, Transportation Research Board of the National Academies, Washington, D.C.

Kamarianakis, Y.; Kanas, A.; Prastacos, P. (2005). *Modelling Traffic Volatility Dynamics in an Urban Network*. In Transportation Research Record, Journal of Transportation Research Board.

Kamarianakis, Y.; Prastacos, P. (2005). *Space-Time Modelling of Traffic Flow*. Computers and Geosciences.

Kaufman, D.E.; Smith, R.L.; Wunderlich, K.E. (1991). *An iterative routing/assignment method for anticipatory real-time route guidance*. Proceedings of the Vehicle Navigation and Information Systems Conference, 693-700.

Kwon, J.; Coifman, B.; Bickel, P. (2000). *Day-to-Day travel time trends and travel time prediction from loop detector*. Transportation Research Record 1717, 120-129.

Lin, W.H.; Kulkarni, A.; Mirchandani, P. (2004). *Short-term arterial travel time prediction for advanced traveler information systems*. Intelligent Transportation Systems 8, 143-154.

Lin, H.E.; Zito, R.; Taylor, M.P.A. (2005). *A review of travel-time prediction in transport and logistics*. Proceedings of the Eastern Asia Society for Transportation Studies 5, 1433-1448.

Mahmassani, H.S. (2001) *Dynamic network traffic assignment and simulation methodology*. Network Spatial and Economics 1, 267-292.

Milkovits, M.; Huang, E.; Antoniou, C.; Ben-Akiva, M.; Lopes, J.A. (2010). *DynaMIT 2.0: The next generation real-time Dynamic Traffic Assignment system*. Proceeding s of the Second International Conference on Advances in System Simulation, 45-51.

Min, W.; Wyntwer, L.; Amemiya, Y. (2007). *Road Traffic Prediction with Spatio-Temporal Correlations.*

Nakata, T.; Takeuchi, J.I. (2004). *Mining traffic data from probe-car system for travel time prediction.* Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, 817-822.
Nanthawichit, C.; Nakatsuji, T.; Suzuki, H. (2003) *Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway*. Transportation Research Record 1855, 49-59.

Park, D.; Rilett, L.R.; Han, G. (1999). *Spectral basis neural networks for real-time travel time forecasting*. Journal of Transportation Engineering 125, 515-523.

Rice, J.; Van Zwet, E. (2004). *A simple and effective method for predicting travel times on freeways.* IEEE Transactions on Intelligent Transportation Systems 5, 200-207.

Meschini, L.; Gentile, G. (2010). *Real-time traffic monitoring and forecast through OPTIMA – Optimal Path Travel Information for Mobility Actions*. Proceedings of Models and Technologies for Intelligent Transportation Systems, International Conference Rome 2009, ed. G. Fusco, Aracne, Rome, Italy, 113-121.

Meschini, L.; Gentile, G. (2009) *Real-time traffic monitoring and forecast through OPTIMA – Optimal Path Travel Information for Mobility Actions.*

Peeta, S.; Mahmassani, H.S. (1995) *Multiple user classes real-time traffic assignment for online operations: a rolling horizon solution framework.* Transportation Research C 3, 83-98.

Papageorgiou, M. (1990) *Dynamic modeling, assignment, and route guidance in traffic networks.* Transportation Research B 24, 471-495.

Rice, J.;Zwet, E. V. (2004). *A Simple and Effective Method for Predicting Travel Times on Freeways.*

Smith, B.L.; Williams, B.M.; Oswalsd, R.K. (2002). *Comparison of parametric and nonparametric models for traffic flow forecasting.* Transportation Research Part C.

Sundaram, S. (2000). *Development of a Dynamic Traffic Assignment system for short-term planning applications.* Master Thesis, Massachusetts Institute of Technology.

Van Lint, J.W.C. (2004). *Reliable travel time prediction for freeways.* PhD Thesis, Delft University.

Van Lint, J.W.C.; Hoogendoorn, S.P.; Van Zuylen, H.J. (2005). *Accurate freeway travel time prediction with state-space neural networks under missing data*. Transportation Research Part C 13, 347–369.

Van Lint, J.W.C; Wunderlich, K.E.; Kaufman, D.E. (2000). *Link travel time prediction for decentralized route guidance architectures*. IEEE Transactions on Intelligent Transportation Systems 1, 4-14.

Vanajakshi, L.D. (2004). *Estimation and prediction of travel time from loop detector data for Intelligent Transportation Systems.* PhD Thesis, Texas A&M University.

Wu, C.H.; Ho, J.M.; Lee, D.T. (2004). *Travel-time prediction with support vector regression*. IEEE Transactions on Intelligent Transportation Systems 5, 276-281

Xia, J. (2006). *Dynamic freeway travel time prediction using single loop detector and incident data.* PhD Thesis, University of Kentucky.

Yang, J.S. (2005). *Travel time prediction using the GPS test vehicle and Kalman filtering techniques.* Proceedings of American Control Conference, 2128-2133.

Yang, Q.; Koutsopoulos, H.N.; Ben-Akiva, M.E. (2000). *Simulation laboratory for evaluating Dynamic Traffic Management systems.* Transportation Research Record 1710, 122-130.

Yu, J.; Chang, G.L.; Ho, H.W.; Liu, Y. (2008). *Variation based online travel time prediction using clustered neural networks.* Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems, 85-90.

Zhang, X.; Rice, J.A. (2003). *Short-term travel time prediction*. Transportation Research Part C 11, 187–210.

Zhou, X. (2004). *Dynamic origin-destination demand estimation and prediction for off-line and online dynamic traffic assignment operation.* PhD Thesis, University of Maryland.

Acknowledgements

We wish to express our thanks to the Project Steering Group and Peer Reviewers for their assistance and support throughout this project:

Steering Group
Henry Pretorius (NZ Transport Agency)
Ken Lee-Jones (Auckland Transport)
Iain McGlinchy (Ministry of Transport)

Peer Reviewers
Deryk Whyte
Leon Wee