SHORT TERM TRAFFIC PREDICTION ON THE UK MOTORWAY NETWORK USING NEURAL NETWORKS

Carl Goves
Transport Systems Catapult

1. INTRODUCTION

The ability to predict traffic conditions over the short-term has the potential to improve traffic management by allowing decisions to be proactive to changing traffic conditions rather than reactive. This paper explores the application of a machine learning technique, specifically neural networks, to predict traffic conditions on a section of the UK motorway network 15 minutes ahead of time.

This paper is structured as follows:

- Section 2 discusses the potential applications of the research, an overview of the machine learning technique used and a literature review of previous relevant research;
- Section 3 outlines the technical approach undertaken for this research;
- Section 4 details the data collected which has been used as an input to the analysis;
- Section 5 reports on and discusses the results of the research;
- Section 6 concludes the findings; and
- Section 7 identifies a number of future research opportunities which have arisen from this study.

2. BACKGROUND

2.1 Potential application

Short term traffic prediction has a number of real world applications and could be used to better manage congestion on the transport system. For example, some intelligent transport systems (ITS) systems react to current traffic conditions and introduce measures to try and mitigate the impact of congestion. Amongst others, these include variable speed limits on busy motorways and urban traffic control systems which help control traffic within complex urban environments. ITS systems used to manage traffic would benefit from being able to anticipate the onset of congestion through use of better short term predictions (i.e. if the system knows congestion is expected in the near future it could take proactive measures to help mitigate the impact of the expected congested future state). The need for such systems is not waning either with the ITS global market expected to grow to be worth over $33bn by 2020 (Intelligent Transportation System Market by Component, 2015). In the UK, local authorities such as Transport for London recognises the value of accurate short term predictions and their integration with ITS as its to procure a predictive signalling system that adjusts traffic signal timings in response to short-term forecasts of traffic conditions (Short-term traffic forecasts to help TfL combat capital's jams, 2015). Also in the UK, £15.2bn was outlined in the Roads Investment Strategy and featured, amongst other schemes, the continued rollout of smart motorways which uses ITS to help manage traffic flows on the UK motorway network (DfT, 2014).

As well as the potential to refine automatic traffic management through ITS, better short term predictions could be used by traffic controllers to make proactive decisions on managing the network. This could be through warnings of expected congestion which would then allow more time for controllers to evaluate different mitigation strategies rather than making decision in reaction to the congestion materialising. A further extension could be that the predictions are made visible to the public in the form of a “traffic-cast”. This could benefit the transport system as it could allow users to optimise their travel arrangements by either re-routing or re-timing their trip (depending on how far into the future the “traffic-cast” predicts).

2.2 Neural networks

The prediction model developed as part of this research takes the form of an artificial neural network. Artificial neural networks are a type of learning model inspired by biological neural networks. They can be used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. Further information on neural networks is readily available in a host of publications. One such publication includes Stergiou and Siganos (Stergiou and Siganos, n.d.).

2.3 Previous research

Attempting to predict traffic conditions over the short term is not new. Numerous studies have used statistical techniques using empirical data to conduct such prediction. Zheng et al (2006) reports that previous research using statistical models includes those using multivariate time-series (Ahmed and Cook, 1979, cited in Zheng, 2006; Hamed et al, 1995, cited in Zheng , 2006; Williams et al, 1998, cited in Zheng , 2006), the Kalman filtering method (Okutani and Stephanedes, 1984, cited in Zheng , 2006), and the nonparametric regression model (Davis and Nihan, 1991, cited in Zheng, 2006; Smith et al, 2002, cited in Zheng , 2006). Indeed, neural networks have also been used to predict traffic conditions (Smith and Demetsky, 1994, cited in Zheng, 2006; Zhang et al, 1997, cited in Zheng, 2006; Dougherty and Kirby, 1998, cited in Zheng, 2006; Park et al, 1998, cited in Zheng, 2006). More recently Hodge et al used binary neural networks to predict traffic conditions at three urban locations in the UK (Hodge et al, 2011). The results of this study showed promise and the system developed was to be incorporated into an intelligent decision support (IDS) system to test against real world data in London, Kent and York in the UK.

Kumar et al used neural networks for traffic conditions on non-urban highways (Kumar et al, 2013). This study developed a neural network with a single output which predicted the traffic volume at a single site from inputs describing 19 different characteristics of that site from the previous 45 minutes. Features included, day of week, time of day, traffic flow by vehicle class, average speed by vehicle class and overall traffic density. Such an architecture yielded promising results with traffic flow 15 minutes into the future matching well with observed values.

For inter-urban, freeway applications, Zheng et al combined several single neural networks to form a Bayesian combined neural network (BCNN) to predict traffic flows on a freeway in Singapore (Zheng et al, 2006). Rather than use data from a single site to predict traffic conditions at that same site, Zheng et al also used data from upstream detectors to inform the traffic conditions. This was limited however to up to two upstream detectors.

3. APPROACH

3.1 Originality

From the literature research undertaken, the use of neural networks to predict traffic conditions is not in itself novel. However, previous studies appear to be dominated by using a single neural network to predict traffic conditions at either single geographical sites or a limited number of sites along a corridor. Further to this, the influence which traffic conditions at sites outside the immediate vicinity of a particular traffic detector is limited. Some of these existing models also use temporal information such as time of day to inform future conditions which may mean that predictions are driven by typical daily profiles of recurrent congestion and may struggle to adapt to predicting traffic conditions when a daily profile is atypical. Therefore, the approach undertaken for this research concentrates on the following principal aspects:

- Short term prediction across a wide geographical area using traffic information from all detectors within that study area; and
- Predictions are time independent, i.e. the predictions will be driven by the "pattern" of traffic conditions from preceding time intervals from all detectors within the chosen study area rather than including the specific time of day as an explanatory variable.

The hypothesis here is that traffic conditions at a single site are influenced by traffic conditions elsewhere in the geographical vicinity. Also by predicting conditions which are independent of temporal information it is surmised that such a prediction model will be more adaptive to non-recurrent traffic conditions than one which uses temporal information such as time of day.

3.2 Methodology

Ideally, the architecture of the neural network which would be built for this research would have multiple input nodes. Essentially, each input node would represent a metric informing a traffic condition during a preceding time interval at a detector within the chosen geographical area. The number of traffic condition metrics could include traffic volume and speed by traffic lane if available. The number of outputs would be the predicted traffic condition for each lane of each traffic site.

Across a wide geographical area this can result in many input and output nodes. For example if a chosen area had 50 sites reporting volume and speed for three lanes of traffic and a neural network which explains predictions using traffic conditions from three previous time periods is needed, this would result in 900 input nodes and 300 output nodes (a third of the input nodes since data from three previous time intervals is being used in this model). There are a number of issues with such a large network:

- The neural network could spend time learning the dependency between input neurons rather than the relationship between inputs and outputs, i.e. correlated inputs creates redundancy in the model (May et al, 2011);
- Large neural networks are computationally expensive to run. For example for every hidden node added to the hidden layer of a fully connected single layer

neural network, the number of connection vectors increases by $I_n + O_n$ where I_n is the number of input nodes and O_n is the number of output nodes; and

- In real world applications, the completeness and reliability of datasets of disaggregate information such as traffic flow and speed by lane is not always available.

To overcome some of the issues identified above, a three stage compression process of the input and output data was used:

- Aggregation of data by lane (averaging of speed and summing of traffic volume);
- Use of traffic density (traffic volume divided by traffic speed) to indicate traffic conditions which removes any potential redundancy present between traffic volume and traffic speed metrics; and
- Dimensionality reduction using an autoencoder.

The first two phases of the compression are trivial. The third phase is less so and uses an autoencoder to reduce the dimensions of the input set. An autoencoder is a form of neural network which attempts to reconstruct its input using one or more hidden layers which have less nodes than the input set. Once an autoencoder architecture is found which allows for the input data to be replicated to an acceptable level of accuracy, the hidden node values from the hidden layer with the fewest nodes can be used as inputs to the neural network model. Once trained, the outputs from the neural network can be decompressed using the front end of the autoencoder.

4. DATA COLLECTION

4.1 Overview

The data used in this study has been extracted from Highways England’s Motorway Incident Detection and Automatic Signalling (MIDAS) system. This system is installed across the UK motorway network and comprises traffic sensors, mainly inductive loops for each lane and located at around 500m intervals. Traffic flow and speed data are recorded by the loops which is used to monitor the network to assist with either manual or automated traffic management through the use of variable message signs (i.e. to set variable speed limits in some locations). This dataset has been chosen given the extensive coverage of the MIDAS system which provides a detailed picture of traffic conditions over a wide geographical area.

4.2 Study area

For this study, traffic data collected along approximately 20km of the M60, M62 and M602 motorways near Manchester was used. Specifically, the motorway network bounded by Junctions 8 and 15 of the M60, Junction 3 of the M602 and Junction 11 of the M62 was used. Figure 1 shows the study area and the location of traffic detectors within this area.

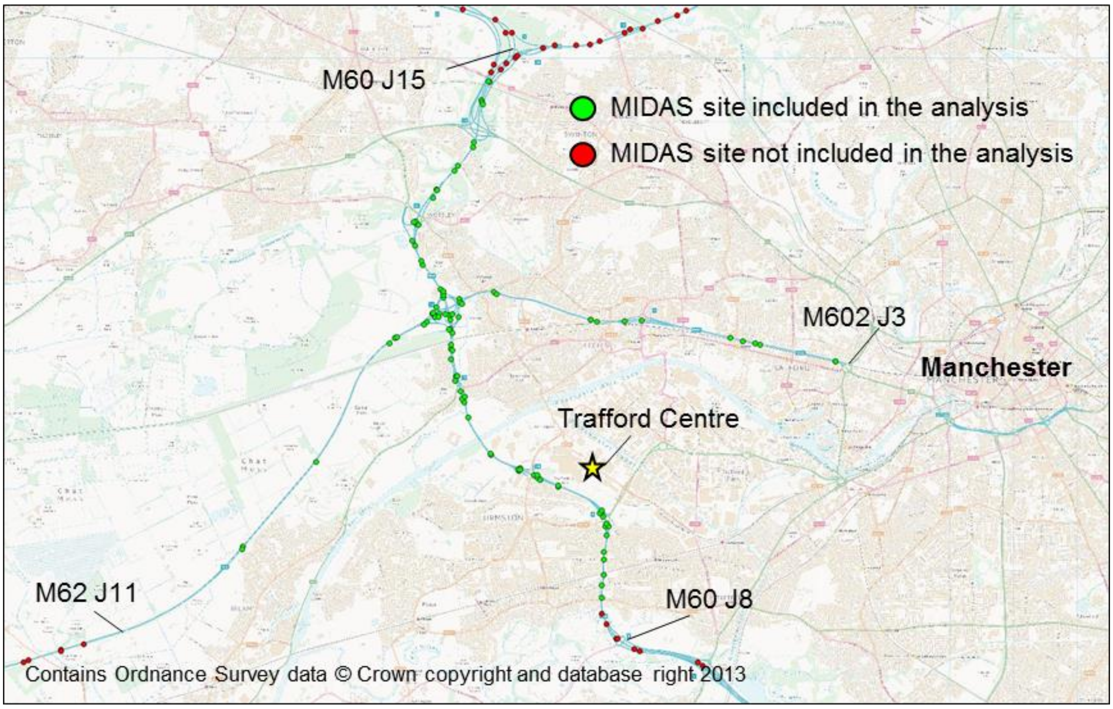
M60 J15
🟢 MIDAS site included in the analysis
🔴 MIDAS site not included in the analysis
M602 J3
Manchester
Trafford Centre
M62 J11
M60 J8
Contains Ordnance Survey data © Crown copyright and database right 2013

Figure 1 - Study area

The area chosen comprises 102 traffic detectors which record traffic volume and speed by lane every minute. This area was chosen as it is an extremely busy part of the network given that it is located adjacent to Manchester and near one of the UK's largest retail outlets, the Trafford Centre.

4.3 Data processing

Traffic data for every day in 2014 was extracted from the MIDAS database for the 102 detectors within the study area. The database was however not complete and contained some missing data due to detector faults which occurred at intermittent times. Therefore, to process the data, a phased approach was developed which is summarised in Table 1.

Table 1 - Data processing approach

Stage	Name	Tasks completed
0	Error detection	Data records missing or thought to be in error were flagged as missing and set to zero
1	Aggregation	Data aggregated over 15 minute intervals and over all lanes Volumetric data was summed whilst speed data was averaged Partially observed 15 minute records were expanded according to the proportion of data flagged as missing within each 15 minute time period
2	Sensor exclusion	If a sensor contained 10% or more missing data it was excluded from the database
3	Infill - interpolation	For missing data positioned between adjacent observed data, interpolation was used to estimate missing values
4	Infill - adjacent week data	For blocks of missing data, observed data from the same time period one week before or one week after was used to infill. If the adjacent week also contained missing data then data from two weeks away was used to infill. If missing data still existed data from 3, 4 etc. weeks away was used to infill.
5	Traffic density	Traffic density was calculated by dividing traffic volumes by average traffic speed

Once the sensors with more than 10% of missing data was excluded from the database, just 2.9% of the remaining records had been expanded due to partial observations within a 15 minute interval. The distribution of those expansion factors is shown in Table 2 below. 73% and 99% of the records expanded were done so with a value of 1.5 or less and 2.0 or less respectively.

Table 2 - Distribution of expansion factors

Metric	1.01 to 1.5	1.51 to 2.0	2.01+
% total records	2.09%	0.76%	0.03%
% expanded records	72.6%	26.4%	1.0%

For stages 1 to 4 in the data processing task, summary statistics of the total number of records in the database and the percentage which were missing, expanded and infilled are provided in Table 3.

Table 3 - Database summary statistics

Stage	Task	No sensors	Total records	% missing	% expanded	% infill
1	Aggregation	102	3,574,080	5.6%	2.6%	0.0%
2	Sensor exclusion	92	3,223,680	0.5%	2.9%	0.0%
3	Interpolation	92	3,223,680	0.5%	2.9%	0.0%*
4a	1 week infill	92	3,223,680	0.0%	2.9%	0.5%
4b	2 week infill	92	3,223,680	0.0%	2.9%	0.5%

* rounds to zero due to absolute number of records interpolated being only 369

On completion of the data processing task, just 2.9% had been expanded and 0.5% infilled.

4.4 Standardisation

Although not necessary, training of neural networks often perform better with input data that has been standardised. For this study the data has been standardised by first zero-centreing the data by applying a Gaussian transformation (i.e. subtracting the mean and dividing by the standard deviation). Secondly the data is transformed to an interval of [-0.9, 0.9] using min-max scaling. An example standardised density profile for six of the sensors in the database is shown in Figure 2.

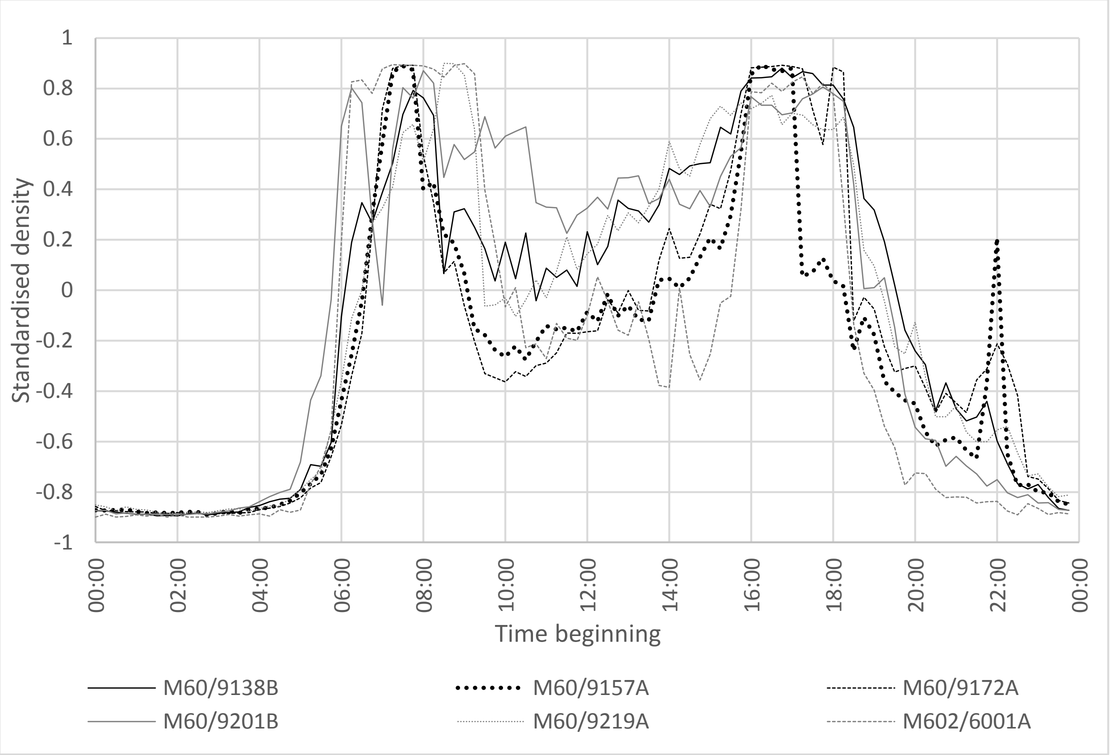

Figure 2 - Example standardised traffic density profiles at multiple detector sites

4.5 Database segmentation

In order to build generalised neural networks, the input database needs to be split into three subsets, namely:

- Training subset – this is the majority of the database, 70% in this case, which is used in training the neural networks;
- Validation subset – here 15% of the databased has been reserved for validation. At each epoch of the training regime, the validation dataset is fed forward through the neural network to ascertain whether the fit of the model continues to improve. It is essentially used to prevent overfitting of the model in order that it provides a generalised solution; and
- Test subset – the remaining 15% of the databased is reserved for testing. Like the validation dataset this is used as another independent test of the generality of the model. Unlike the validation dataset, the test dataset is only passed through the neural network on completion of the training regime and the fitness statistics associated with it are used for comparison with the fitness of alternative neural network structures.

The division of the database into these three subsets was undertaken randomly using a stratified approach in that each 15 minute interval in the day had an equal number records within the training, validation and test datasets. This ensured that when the neural networks were trained, they were not biased by data representative of one or more particular times of the day.

4.6 Sample size

Although a complete years' worth of data was processed, due to the amount of exploratory work that was required in training a representative neural network (for example the setting of hyper parameters such as learning rate and the number of hidden nodes is done by trial and error), the first 10% of the database was used in the analysis. This equates to over 36 days in the year and incorporates over 320,000 data points.

5. RESULTS AND DISCUSSION

Results for this research are presented in two sections. The first reports on the results of the dimensionality reduction achieved through application of an autoencoder. The second section reports on the results of the short term prediction conducted using the reduced dimensioned database.

5.1 Dimensionality reduction

The autoencoder trained to reduce the dimensions of the input data took the form of a neural network with a single hidden layer with less nodes than the input layer and trained using a batched backpropagation algorithm. Hyper-parameters such as learning rate and batch size were determined via trial and error. Fifteen experiments were then conducted, each trained for 500 epochs, to ascertain the performance of the autoencoder assuming different numbers of hidden nodes: 1, 2, 3, 4, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80 and 90. The author has found no clear guidance on what performance metrics should be monitored to help decide which structure is most appropriate. Therefore, the following metrics calculated using the test dataset were monitored:

- Root mean square error (RMSE);
- The percentage of reconstructed inputs which lie within 0.01, 0.05 and 0.1 of the input values; and
- The mean, median, interquartile range and range of the absolute error between reconstructed and input values.

Figure 3 shows the results of all these metrics graphically. Note that the results have been plotted on a continuous scale with the underlying assumption that intermediate hidden node values not tested would yield interpolated performance between adjacent tested values. On all charts, the number of hidden nodes is plotted on the horizontal axis, with the vertical axis showing the metric(s) indicated within the associated chart legend.

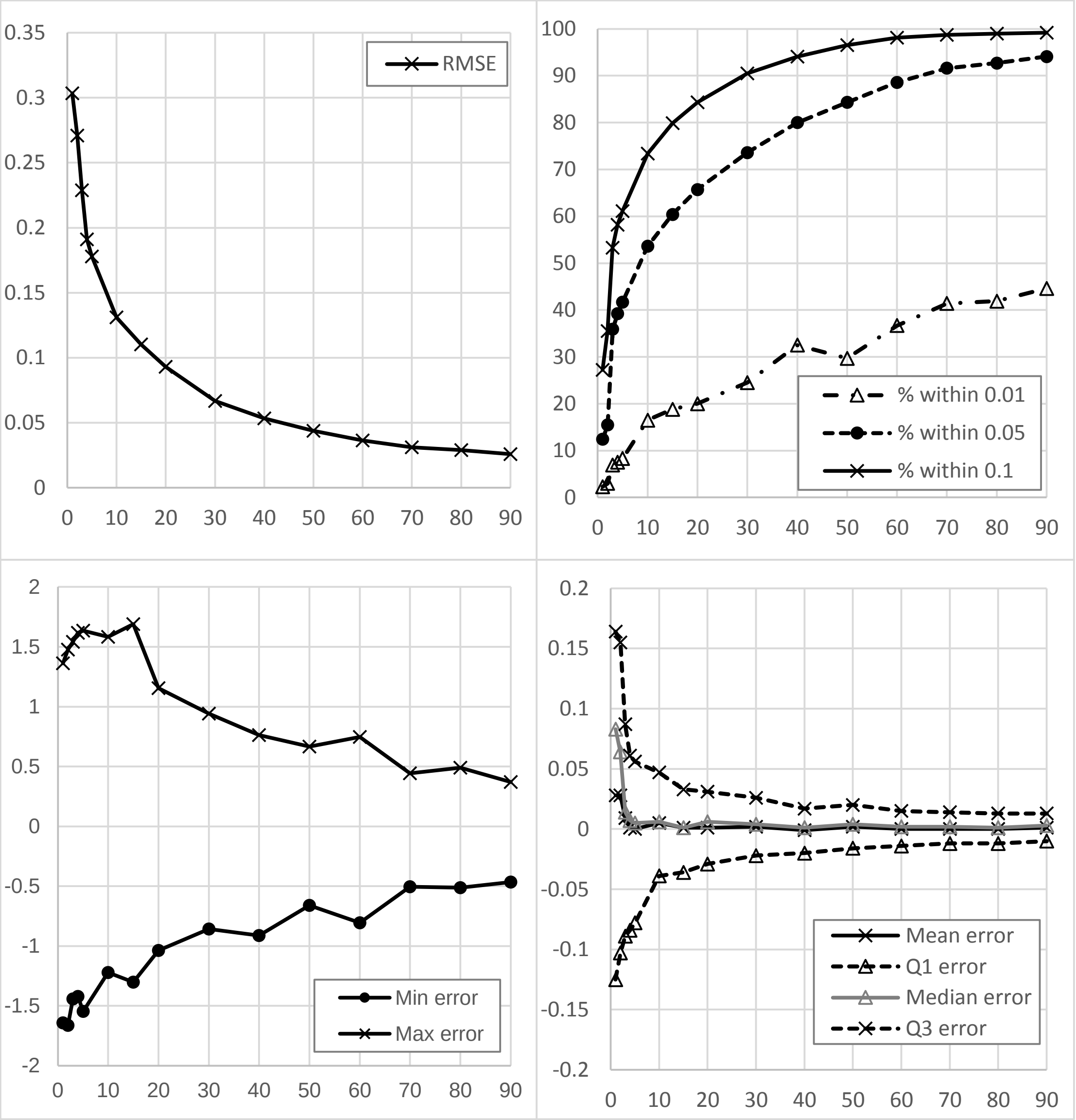

Figure 3 - Autoencoder accuracy results (test dataset)

From the results above, the error between the reconstructed input and the input values reduces as the number of hidden nodes increases. This is as expected as increasing the number of hidden nodes decreases the difference to the original number of dimensions (input nodes). Deciding which autoencoder to select is a trade-off between accuracy and reducing the number of dimensions sufficiently and further investigations beyond this study should be undertaken to determine the sensitivity of different levels of dimensionality reduction upon the performance of a prediction model. For the purposes of this study, an autoencoder with 30 hidden nodes would appear reasonable and was selected for training the prediction model. Such an autoencoder achieves an accuracy of more than 90% when replicating input values to within 0.1 of standardised density values. To put this into context, the figure below shows a selection of encoded density inputs next to the standardised observed values and shows a very good fit.

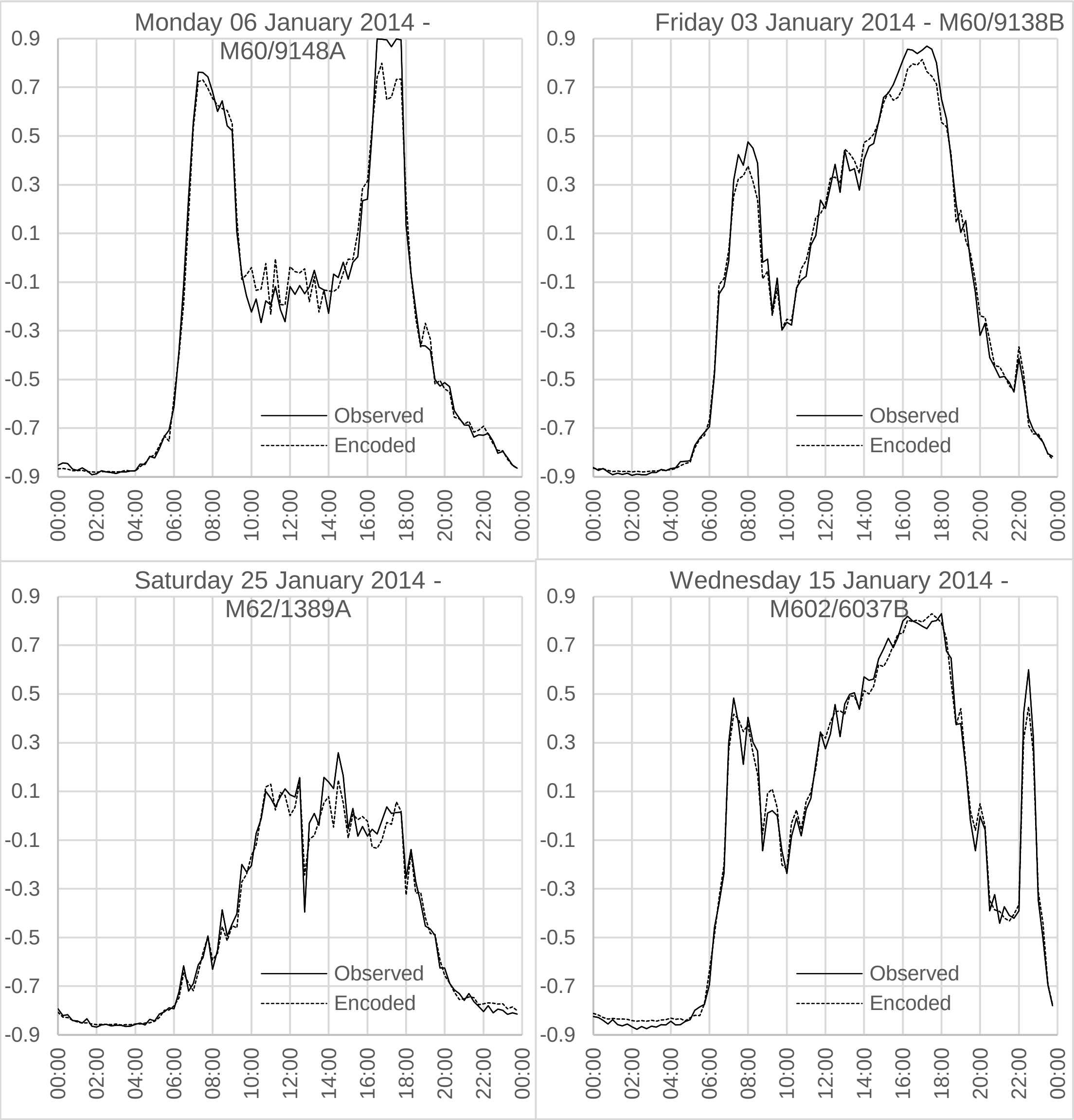

Figure 4 - Example density profile comparisons: observed vs encoded

Investigations were briefly made into reducing the dimensions of the inputs further through use of a stacked autoencoder (essentially the reduced dimension from this auto encoder is then passed through another autoencoder to reduce dimensions further and this process is repeated until the error propagation is not acceptable). However, these initial results showed the error propagation beyond a single layer autoencoder was unacceptable. This issue was only briefly visited and so further investigation is warranted.

5.2 Short term traffic prediction model

The encoded database reported on in section 5.1 was taken forward as the input dataset for the short term traffic prediction model which was developed. The objective of this exercise was to be able to predict accurately traffic conditions in the next time interval (in this case the next 15 minutes). As with the autoencoder, the training regime employed to train the neural network was batch backpropagation. Trial and error was used to ascertain adequate hyper-parameter values such as the learning rate and batch size. All neural networks were trained with up to 5,000 epochs although early stopping was introduced once the model appeared to be overfitting (i.e. once the validation cost function begins increasing).

A series of experiments were undertaken in which the number of hidden layers, hidden nodes and the number of previous time intervals used for the prediction were adjusted. Traffic density from up to three previous time periods were tested, (i.e. if the traffic density (*d*) in the current time interval (*t*) is d_t then for an experiment which explored using three time intervals, values of d_t, d_{t-1} and d_{t-2} were used to try to predict d_{t+1}). Up to three hidden layers were tested each with a varying number of hidden nodes. The RMSE for each of the experiments conducted is shown in Table 4 and includes the lowest RMSE emphasised in red.

Table 4 - Prediction model RMSE by different neural network structures

Hidden nodes by hidden layer			RMSE by the number of previous time intervals included in the model		
Layer 1	Layer 2	Layer 3	t	t, t-1	t, t-1, t-2
50			0.1592	0.1517	0.1527
100			0.1547	***0.1507***	0.1521
250			0.1544	0.1518	0.1515
500			0.1532	0.1510	0.1516
1000			0.1542	0.1524	0.1524
50	50		0.1548	0.1508	0.1539
50	100		0.1526	0.1520	0.1527
50	250		0.1541	0.1513	0.1522
50	500		0.1539	0.1524	0.1522
50	1000		0.1567	0.1537	0.1552
50	50	50	0.1541	0.1530	0.1533
50	50	100	0.1548	0.1515	0.1531
50	50	250	0.1551	0.1521	0.1523
50	50	500	0.1546	0.1532	0.1546
50	50	1000	0.1572	0.1519	0.1538

Table 4 shows that the RMSE of many of the experiments are similar. In some cases the differences could be due to stochastic fluctuations brought about by the initial weights in the networks being randomly set rather than representing relative differences in the predictive power of the models. However, Table 4 consistently shows that the predictive power of the model improves when traffic density information from two previous time intervals are included in the model as opposed to one. It is less conclusive whether increasing the number of hidden nodes or hidden layers adds to the predictive power of the model. Based on the results shown in Table 4, a neural network which incorporated two previous time intervals of traffic density and a single hidden layer consisting of 100 hidden nodes was selected for further investigation.

Using the selected predictive model, the predictions were decompressed using the front end of the autoencoder reported on in section 5.1. Next the predictions were transformed back to traffic density values by reversing the standardisation described in section 4.4. These predicted values were then compared against observed values. Figure 5 shows two comparisons of daily density profiles chosen at random.

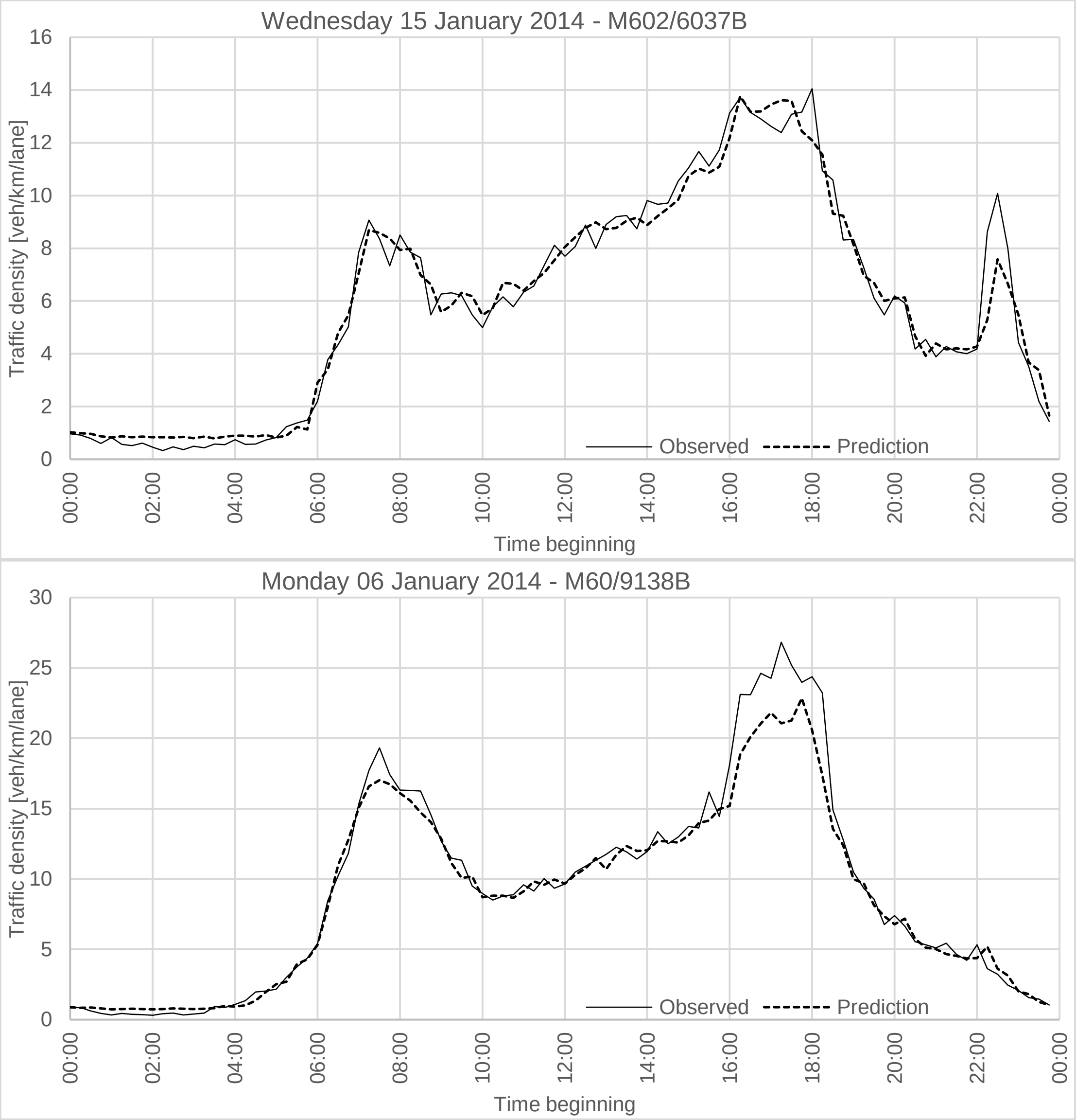

Figure 5 - Example density profile comparisons: observed vs predicted

From observation of profiles such as those shown above, it would appear the accuracy of the prediction model is very good. Predictions generally follow the rise and fall of traffic congestion and there appears to be limited lag from the observed values. The errors that are present could be partially due to the prediction model using encoded inputs. Also the fit of predictions for some sensors is better than others. This could be due to the proximity of the detectors to the edge of the study area boundary. For example, if there are few or no upstream detectors for a site on the edge of the study area then there would be limited dependent information to drive the prediction. Further research into this effect would be needed to confirm whether this is true.

To understand more fully the accuracy of the prediction model, the percentage of predictions within an absolute difference to observed traffic density was calculated. As well as calculating the percentage accuracy across all observations, the calculation was also disaggregated by congestion bands. Four congestion bands were derived: low, medium, high and very high which translate to the first, second, third and fourth quartiles respectively of observed traffic densities. The results of this analysis is shown in Figure 6 below.

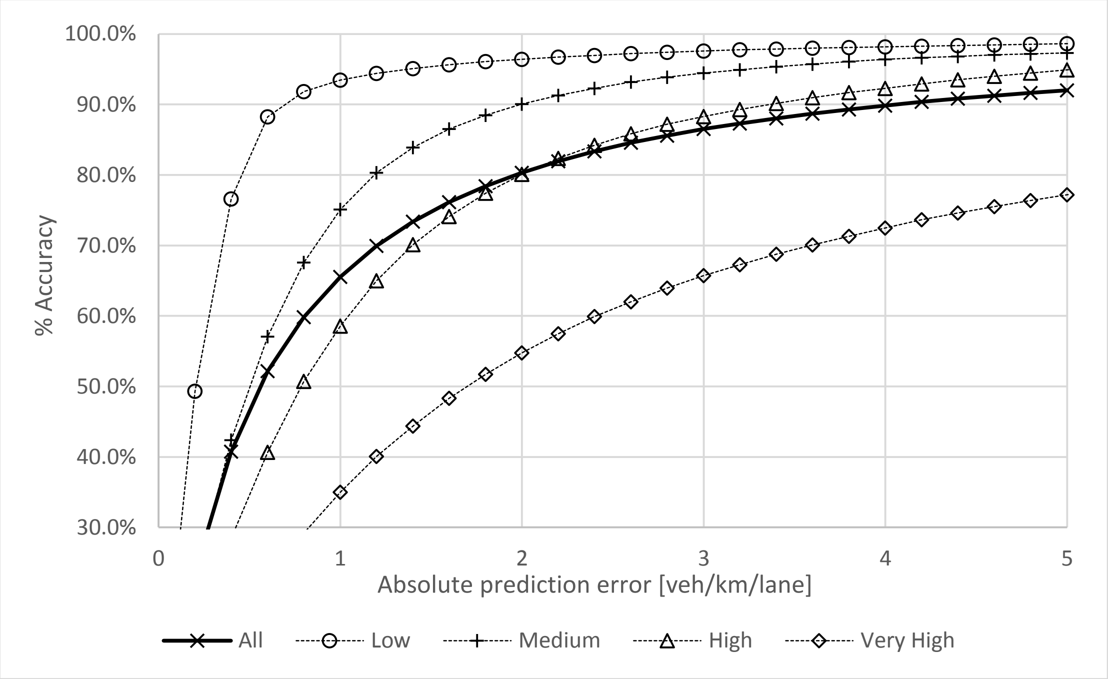

Figure 6 - Prediction model accuracy

Figure 6 shows that overall 80% of all the predictions are within 2 veh/km/lane of observed traffic density values and 90% within 4 veh/km/lane. Predictions for low congestion situations are the most accurate with 90% of predictions within 0.8 veh/km/lane of observed values. For medium and high levels of congestion, 90% accuracy is achieved within a tolerance of 2 veh/km/lane and 3.4 veh/km/lane respectively. Accuracy is the lowest for very high levels of congestion with 80% of predictions within 5.8 veh/km/lane. The significance of the prediction model accuracy is something that needs to be investigated further and will be dependent on the application of the model.

6. CONCLUSIONS

The research reported on in this paper has successfully developed a model using a neural network that 90% of the time predicts future traffic density fifteen minutes into the future within 4 veh/km/lane of accuracy. This pilot model has been developed over a wide geographical area covering approximately 20 km of the UK motorway network, estimating traffic conditions across 92 sensors using over 320,000 data points to train, validate and test the model. This research has multiple real world applications including the refinement of prediction engines in ITS systems, informing proactive decisions to be taken by traffic controllers and opens up the opportunity, with the integration of real time data, to produce a “traffic-cast” of future traffic conditions.

7. FUTURE RESEARCH OPPORTUNITIES

The research presented in this paper represents an initial attempt at using neural networks to perform short term traffic prediction over a wide geographical area. Future research opportunities related to the investigations undertaken here include:

- The use of stacked autoencoders to perform reduce dimensions of traffic data further;
- Investigations into the accuracy of predictions in relation to the proximity of detectors to the boundary of a selected geographical area;
- The sensitivity of predictions to the level of data encoding, size of study area and size of training dataset;
- Testing of the prediction model using real time data; and
- Benchmarking against other prediction models.

BIBLIOGRAPHY

Department for Transport. (2014) *Road Investment Strategy: Investment Plan.* Available from: <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/382813/dft-ris-road-investment-strategy.pdf>. [24 August 2015]

Hodge, V.J., Krishnan, R., Jackson, T., Austin, J., and Polak, J (2011), Short term traffic prediction using a binary neural network, paper presented to Universities Transport Study Group 2011, Milton Keynes, 2011. Available from: <http://www-users.cs.york.ac.uk/vicky/myPapers/Hodge-etal-UTSG2011.pdf>. [24 August 2015]

Intelligent Transportation System Market by Component. (2015). Available from: <http://www.marketsandmarkets.com/Market-Reports/intelligent-transport-systems-its-market-764.html. [24 August 2015]

Kumar, K., Parida, M. and Katiyar, V.K. (2013). Short term traffic flow prediction for a non urban highway using Artificial Neural Network. *Procedia – Social and Behavioural Studies,* 104, 755-764

May R., Dandy G. and Maier, H. (2011). Review of Input Variable Selection Methods for Artificial Neural Networks, *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, Prof. Kenji Suzuki (Ed.), ISBN: 978-953-307-243-2, InTech, Available from: <http://cdn.intechopen.com/pdfs-wm/14882.pdf>. [25 August 2015]

Short-term traffic forecasts to help TfL combat capital's jams. (2015) *Local Transport Today*, (678), p.1.

Stergiou C. and Siganos D. n.d. *Neural Networks*. Available from: <http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html>. [27 August 2015]

Zheng W., Lee D, Shi, Q (2006). Short-Term Freeway Traffic Prediction: Bayesian Combined Neural Network Approach. *Journal of Transportation Engineering,* 132(2), 114-121.