Introduction and Data Description

This investigation serves as a review of spatio-temporal methodologies for forecasting traffic flow data across a section of London's road network. There is currently a lack of research focused on the performance of forecasting traffic flows in heterogeneous road networks (HRN). Research emphasis has instead centred on forecasting single road segments i.e. motorways (Asif *et al*., 2013). In reality, HRNs have inherent spatiotemporal interdependencies - traffic in one location is innately tied to traffic both upstream or downstream in the network (Yue & Yeh, 2008). Therefore, there is a need to include spatio-temporal metrics to account for space-time structures and to forecast traffic flows more effectively (Cheng *et al.* 2012).

This study aims to provide exploratory spatiotemporal analysis to help uncover such space-time structures/interdependencies, whilst evaluating the performance of three machine-learning models: (i) random forest regression (RF); (ii) support vector regression (SVR); and, (iii) a Long-Short Term Memory neural network (LSTM).

This study's data derives from Transport for London's London Congestion Analysis Project (LCAP). The study uses estimates for travel times between a pair of cameras based across 256 road links recorded in 5 minute intervals between 6am–9pm between 1st–31st January 2011. This report selects 18 road links based around Elephant and Castle (Figure 1). This area has been chosen as it displays qualities of an interdependent HRN comprising a core section of the south London transport flow. We aim to create conditions for 'effective' space-time forecasting (Yue & Yeh, 2008).

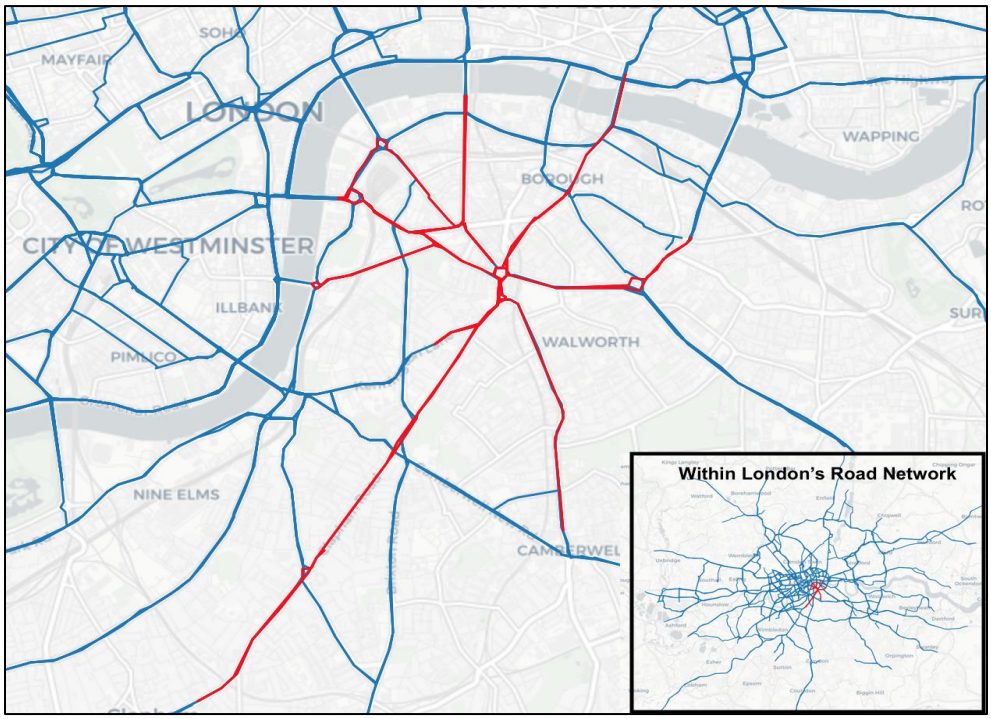
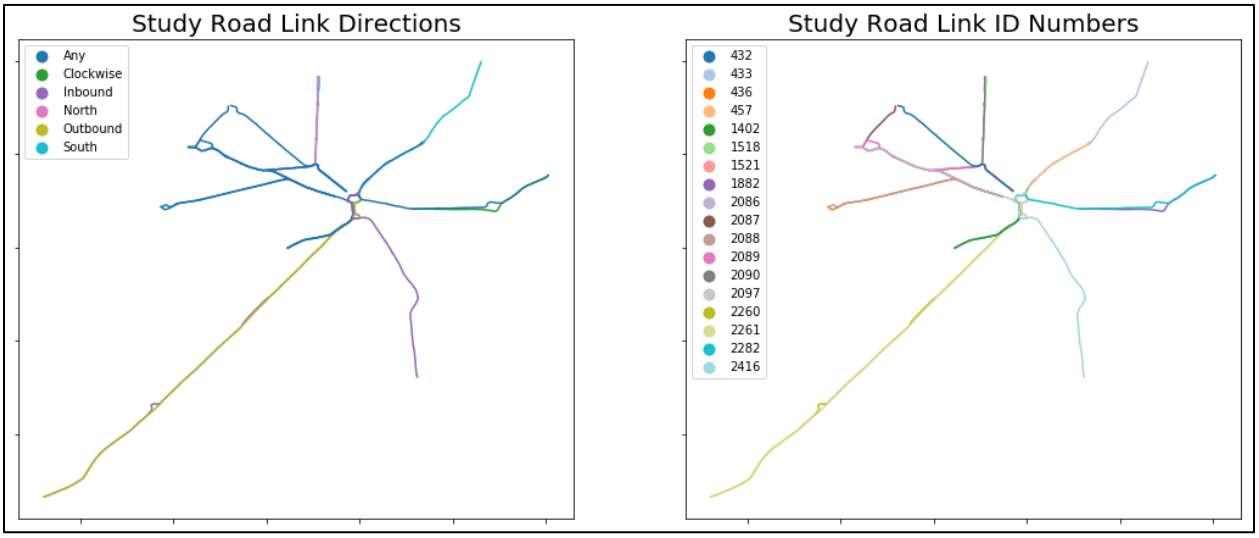


Figure 1: Selected road links

The chosen area incorporates six different directionality classes (Figure 2). A directed second order neighbour adjacency matrix was calculated based on where edges of the network connected upstream and downstream (Figure 3).

Figure 2: Road link directions (Left) and ID numbers (Right)



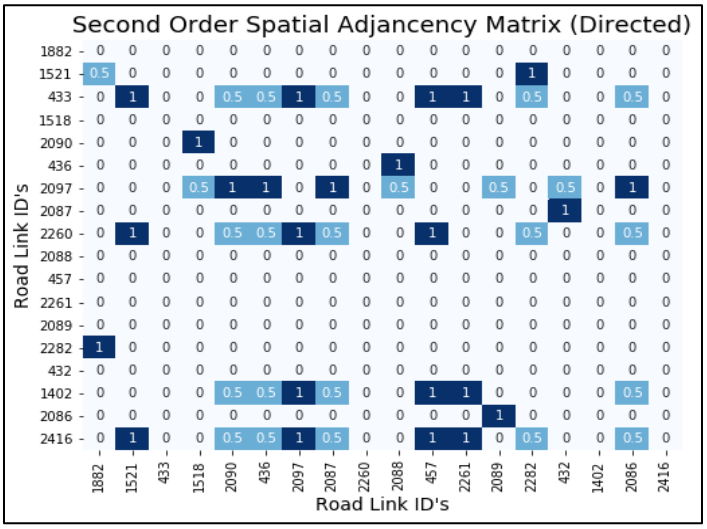
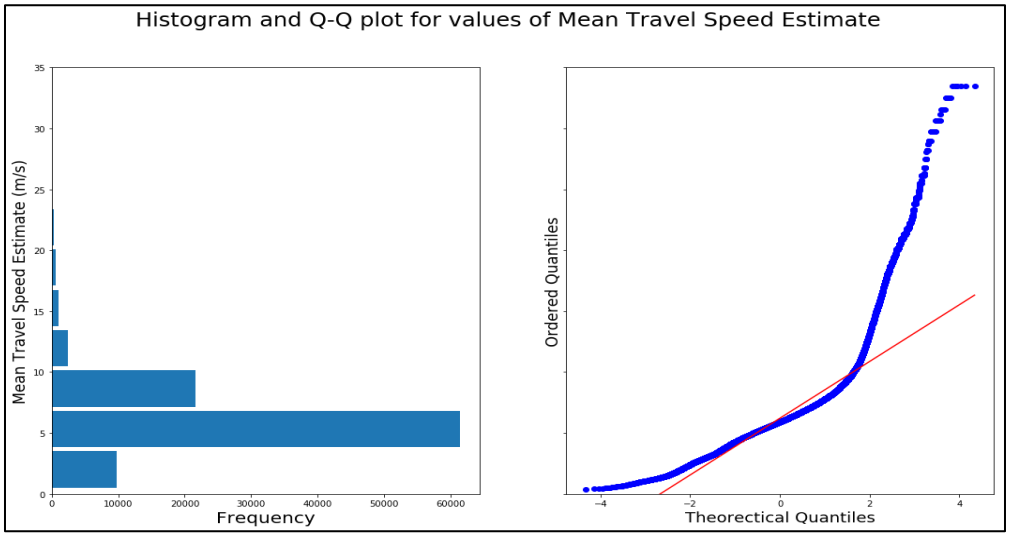


Figure 3: Spatial weight matrix used in the study

We elected to transform the travel speed estimate (TSE) from seconds per metre to metre per second for interpretability. The distribution of TSE across the study area is shown in Figures 4 and 5. TSE has a mean of 6.21 (m/s) and a positive skew as indicated by the Q-Q plot (Figure 4). Figure 5 indicates that the chosen road links exhibit a heterogeneous range of TSE values, with greater interquartile ranges in roads 1882, 1402 and 2416.

Figure 2: Histogram (Left) and Quantile-Quantile plot (Right) of mean TSE



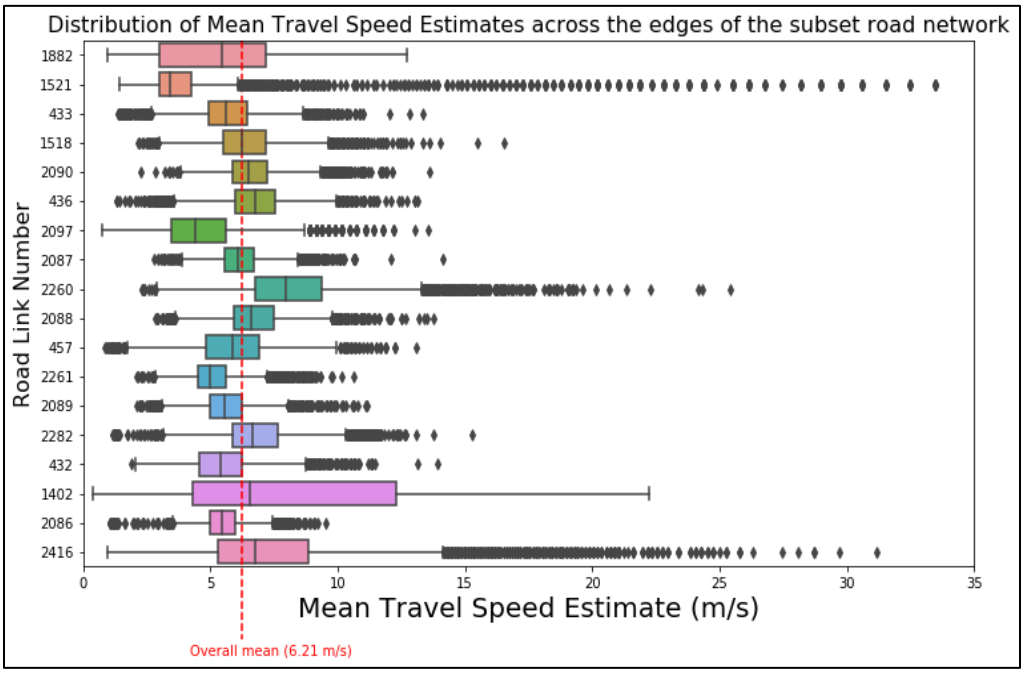


Figure 5: TSE distribution across all road links of study area

Exploratory Spatio-Temporal Data Analysis (ESTDA)

For this report, we split the data into three defined sections of day (after Cheng *et al.* 2012) detailed in Table 1 to help inform spatio-temporal information passed to the models.

Table 1: Selections of day defined for report

Section	Times	Days	n (each day)
Morning Peak	7:00 – 9:00	Monday – Friday	24
Evening Peak	17:00 – 19:00	Monday – Friday	24
Non-Peak	Times outside morning and evening peaks	Monday – Sunday	132–180

Temporal Analysis:

TSE values remain fairly consistent on average across the road links from 7:00–19:00 with a more pronounced drop in values during the evening than morning peak (Figure 6). In Figure 7, the overall probability distribution of TSE between each distinct section of the day is found to be statistically different as determined by 2-sample Kolmogorov–Smirnov tests (Table 2).

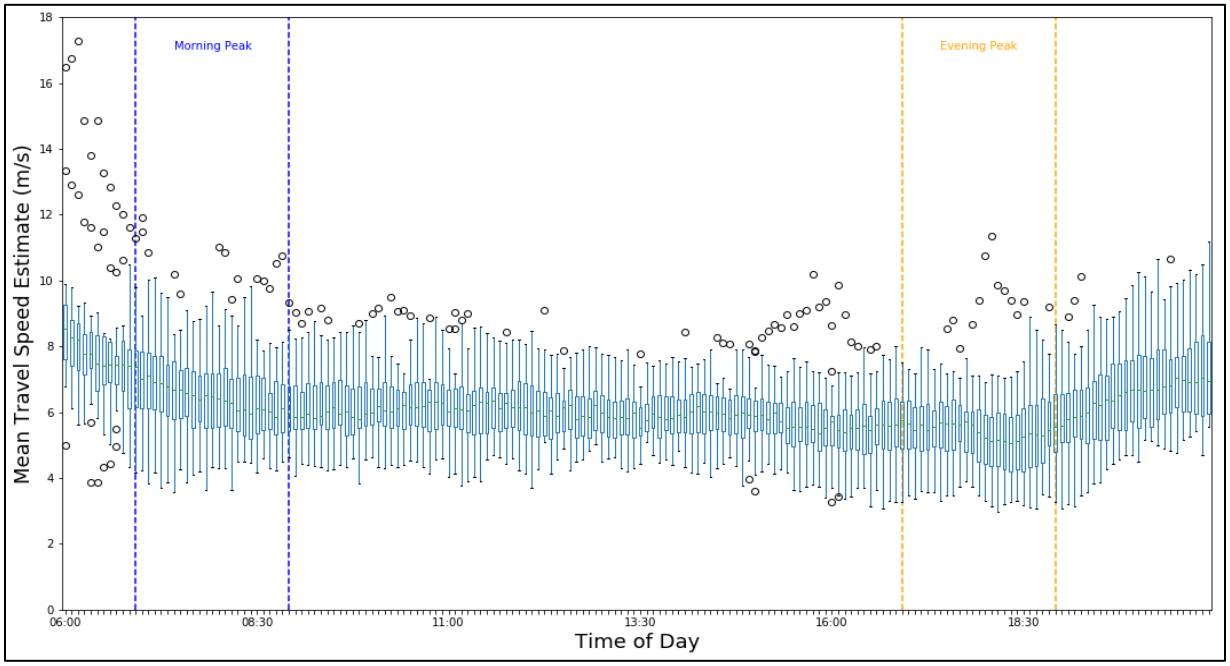


Figure 6: Average TSE box-plots of each five minute interval across all road links for the days Monday–Friday.

Figure 7: Kernel Density Estimation of mean travel speed

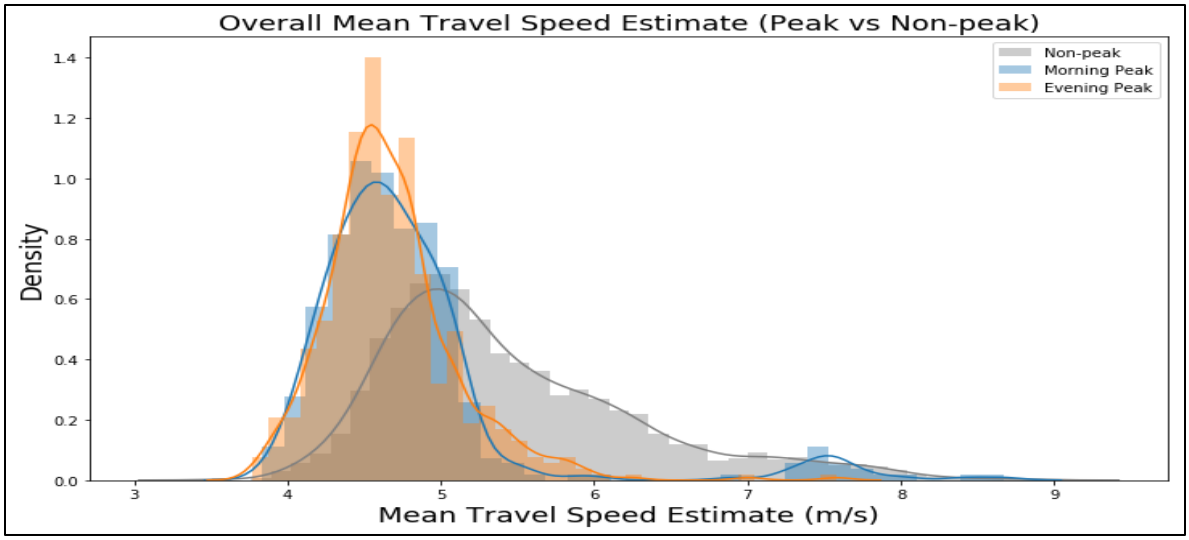
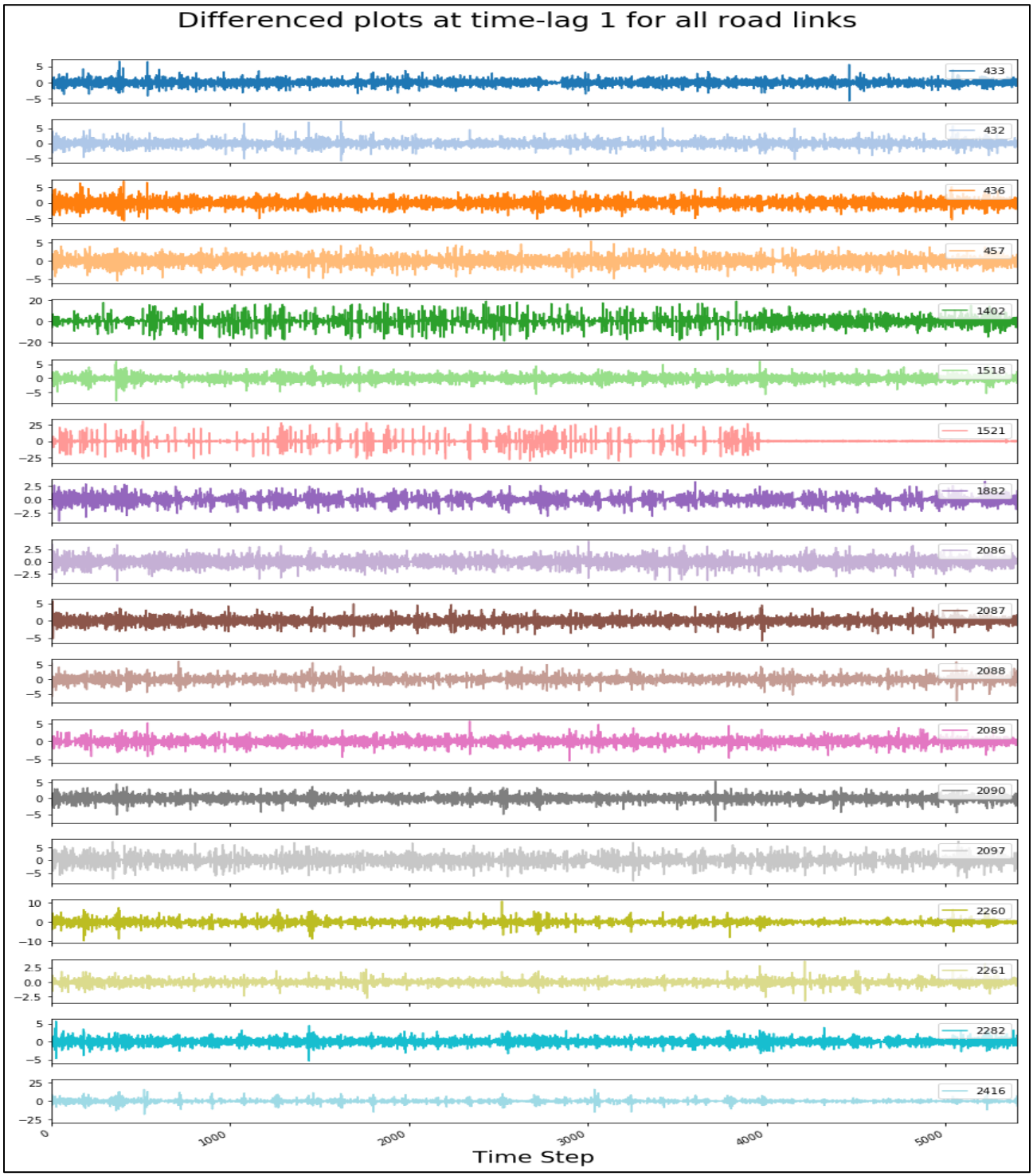


Table 2: Kolmogorov–Smirnov 2-sample tests for the distribution of the TSE at each section of the day (significant values above 95% confidence interval shown in bold)

Comparison	K-S Statistic	p-value
Morning Peak vs Non-Peak	0.344	**0.00**
Evening Peak vs Non-Peak	0.510	**0.00**
Morning Peak vs Evening Peak	0.203	**0.00**

Differencing (at a time-lag of one) of the TSE in applied across each road link and is shown in Figure 8. Following this, we decided to remove link '1521' from our models as the temporally-differenced trend is shown to level out around 4100 time steps, and it is thought, related to road works at this location.

Figure 8: Differencing results of individual road links



Temporal autocorrelation and partial autocorrelation plots across all the road links are shown in Figure 9a and 9b. A significant linear relationship exists across time for up to 34 lags (170 minutes). However, Figure 9b suggests that the importance of additional temporal information is diminishes after six temporal

lags (30 minutes) across the road network. An Augmented Dickey-Fuller (ADF) unit root test found each individual road link to exhibit statistically significant (p-value > 0.05) stationarity of TSE across the 30 days of data.

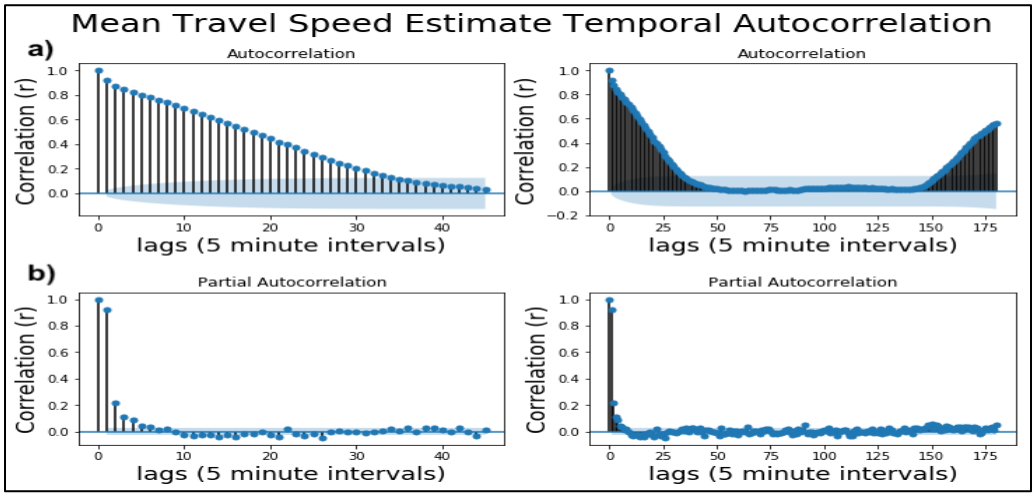


Figure 9: Temporal autocorrelation (a) and partial temporal autocorrelation (b) of TSE values across all road links

Spatial Analysis:

Figure 10 shows a correlation matrix of average speeds of all 18 roads in the study area. This suggests that the roads are more correlated in the morning peak period than evening peak. However, Moran's I results (Table 3) indicate the spatial autocorrelation is relatively insignificant.

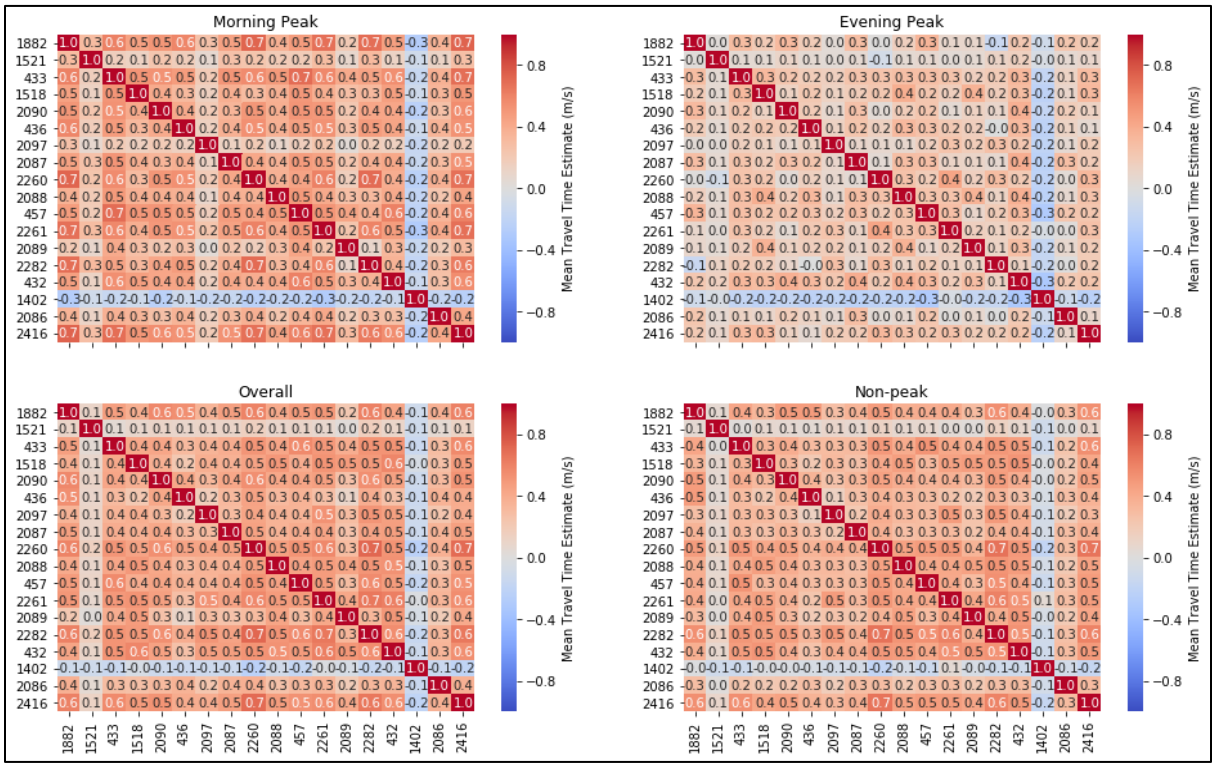


Figure 3: Correlation matrix between roads by time of day

Table 3: Moran's I statistics across road links

Time Period	Moran's *I*	*p-value*	*n*
Overall	–0.186	0.54	18
Morning (AM) Peak	0.510	0.76	18
Evening (PM) Peak	0.203	0.44	18
Non-peak	–0.182	0.55	18

Spatio-Temporal Analysis:

Using the second-order adjacency matrix, the spatial-temporal autocorrelation functions (ST-ACF and ST-PACF) were calculated across the 30 days of TSE data (Figures 11 and 12). The ST-ACF shows a strong correlation which persists until about 36 lags (three hours). Across the study period, there is a clear periodic cycle (repeated daily) indicating the seasonality of the data. Both of these trends are likely due to the effects of weekday rush-hour period. Despite this, the ST-PACF indicates no additional information is added above a temporal lag of one and spatial lag of one, although this may be a result of averaging all the road links (Figure 13).

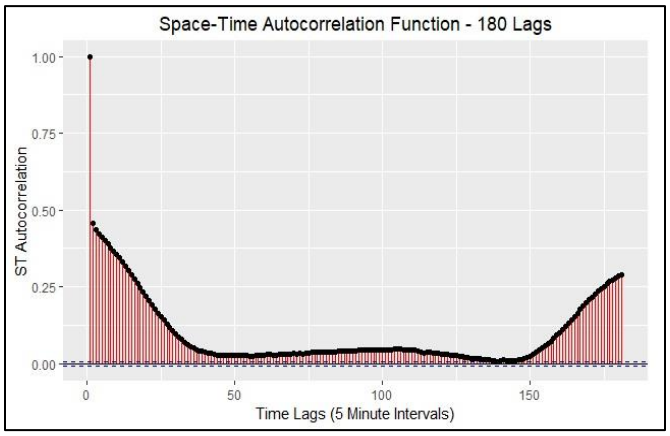


Figure 11: 180 lag ST-ACF

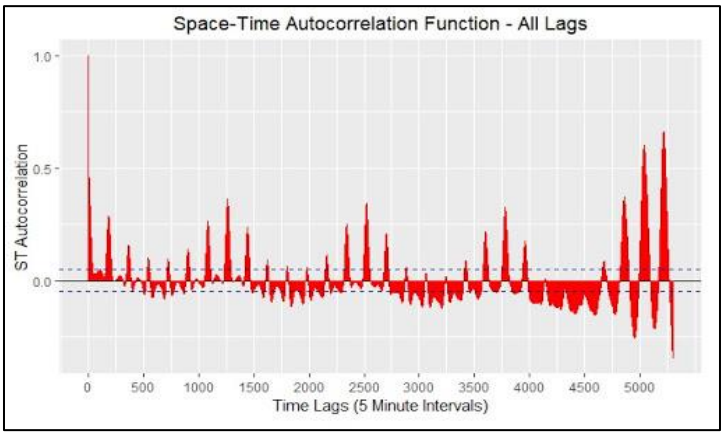


Figure 12: 5400 lag (30 days) ST-ACF

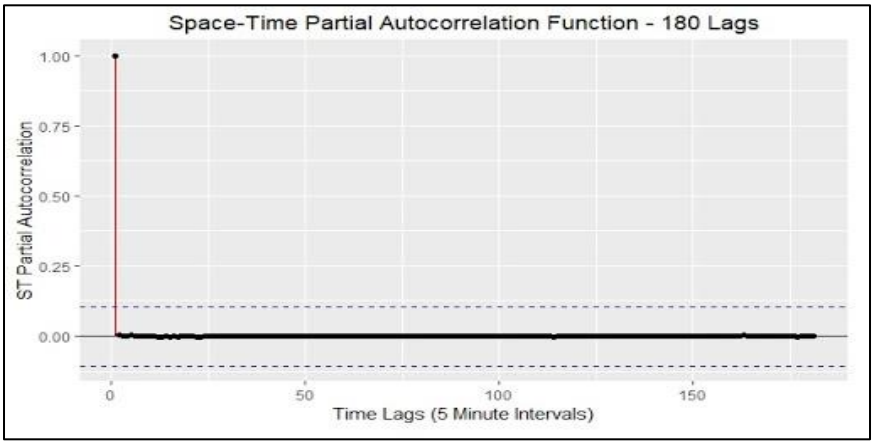
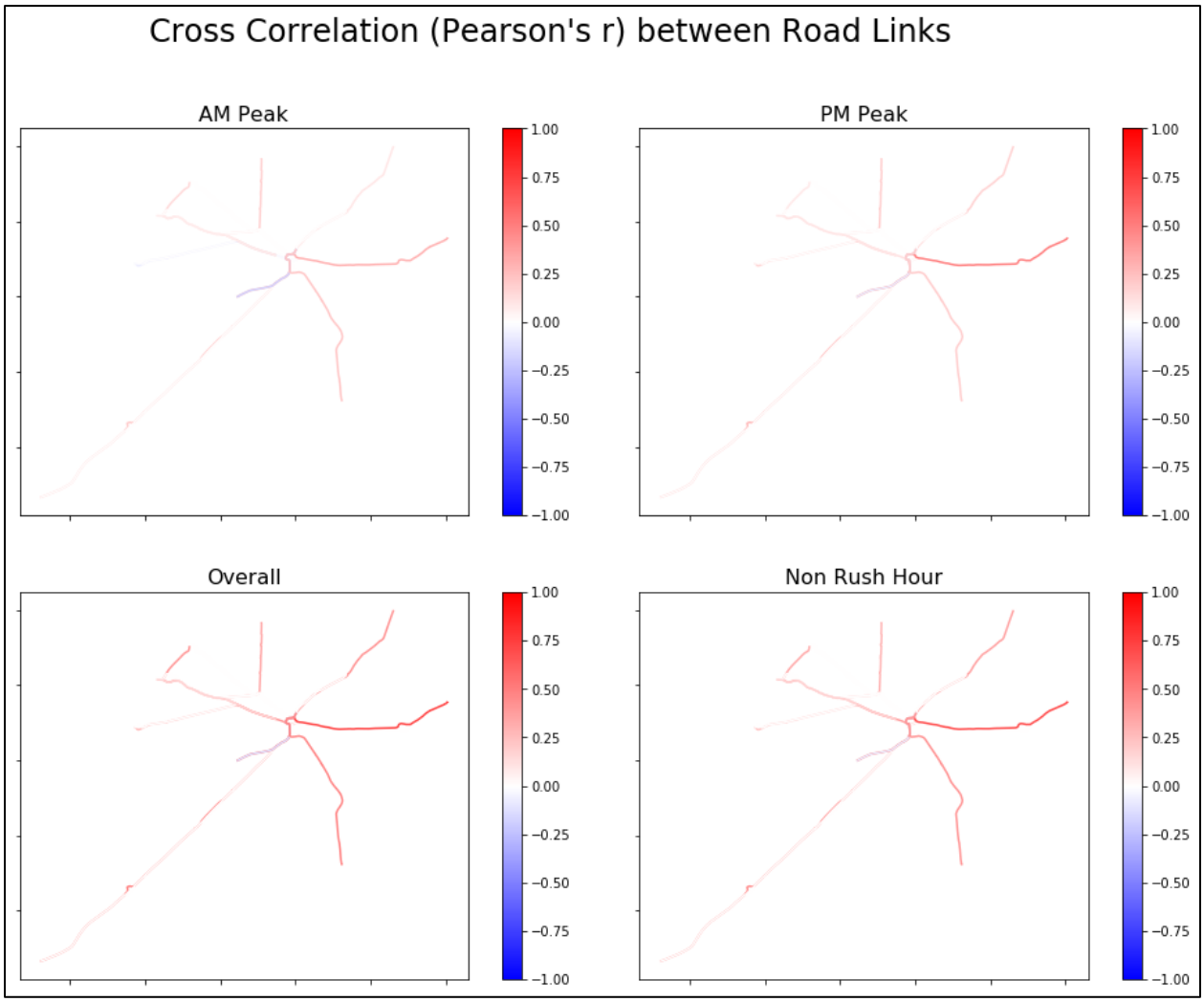


Figure 13: 180 lag (one day) ST-PACF

The cross-correlation function (CCF) uses Pearson's r coefficient to measure the spatio-temporal interdependence of an individual link of a network to all the other links within the network at a given time lag (Yue & Yeh, 2008). CCF is shown spatially at temporal lag one across the study area in Figure 14. There is evidence that spatio-temporal interdependence varies in time (throughout the day) and space (across the road network), favouring more interdependence in the morning and towards the west of the network.

Figure 14: Cross correlation between road links (temporal lag one)



The coefficient of determination (CCF^2 or CoD), is used to measure the shared variance between each road link and all the other links. This was calculated for each road link pair between temporal lags of -90 and 90 and then averaged for each lag (Figure 15; Cheng *et al.* 2012). Links with greater values of shared variance contain more useful spatio-temporal information at a given number of lags - that is, they help explain more of the spatio-temporal interdependence of the network and thus are more useful in the forecasting of all road links (Asif *et al.* 2014). Roads 2282, 2260, and 2416 are found to display this dynamic at up to +-20 lags. Opposingly, 1402 shows a low shared variance with the other links, suggesting any space-time forecast for this link is known as a 'blind forecast' (Yue & Yeh, 2008).

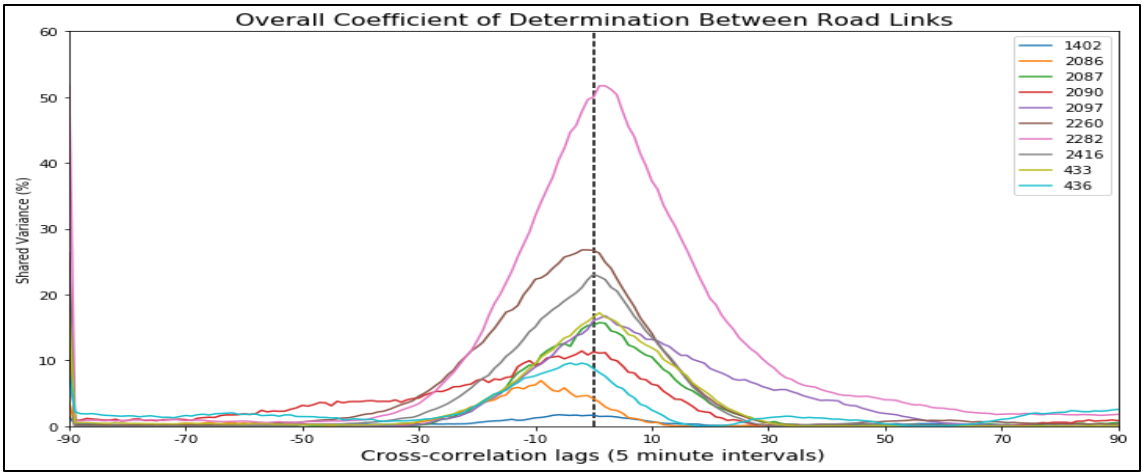


Figure 15: CoD between each road link and the others across the dataset

Figure 16 shows the CoD for the morning and evening peaks. Here we see that the the shared variance between the links is much more suppressed with no values above 5% in the morning and only road 2282 above 10% in the evening. This suggests that these would be ‘‘blind’ forecasts if we isolated these time periods.

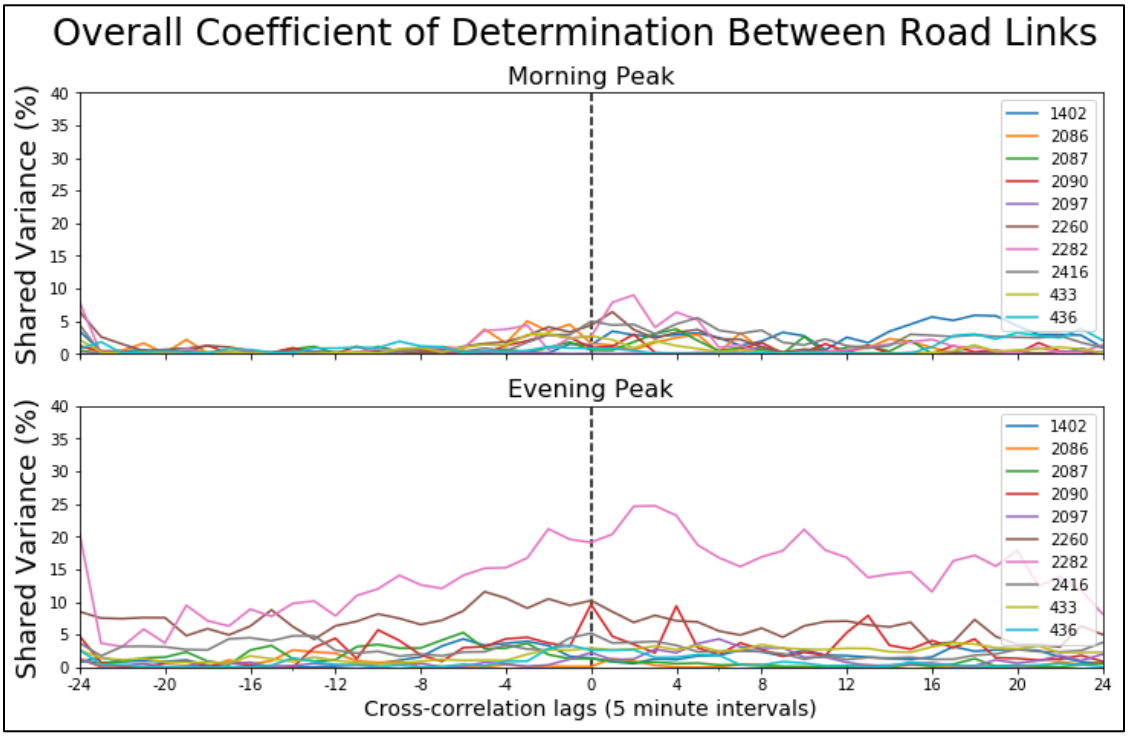
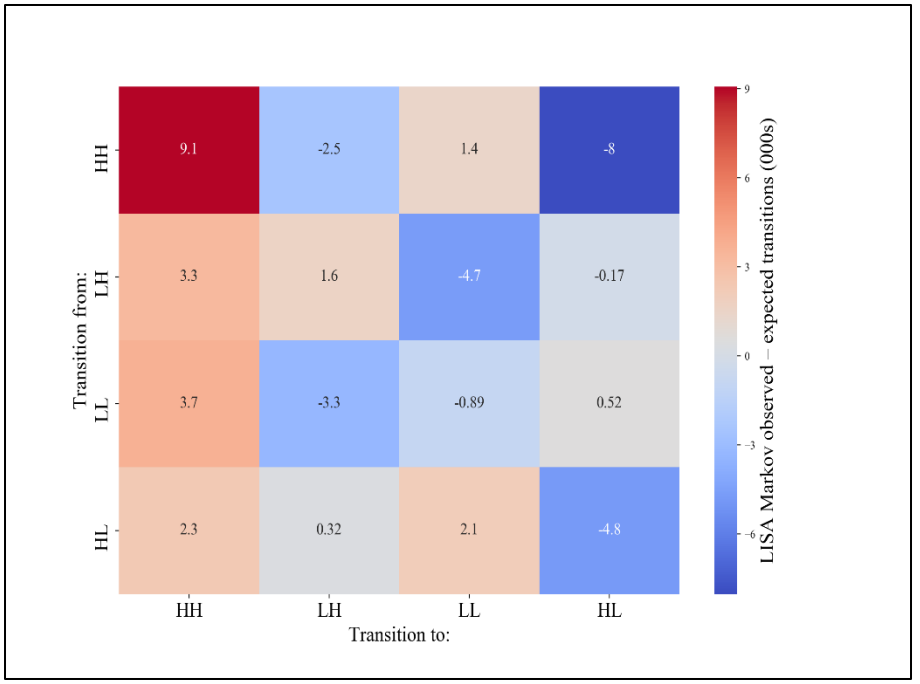


Figure 16: CoD between all road links at morning (Top) and evening (Bottom) peaks

Discrete Markov-chains are a methodology for modelling the probabilities of state transitions based on previous temporal observations. The methodology is limited in a spatio-temporal context as the framework assumes that transitions are independent to space (Clark and Rey, 2017). However, Markov-chains have been extended to accommodate spatial interdependencies in a method known as a spatial Markov that incorporates spatial lag values into the framework (Clark and Rey, 2017).

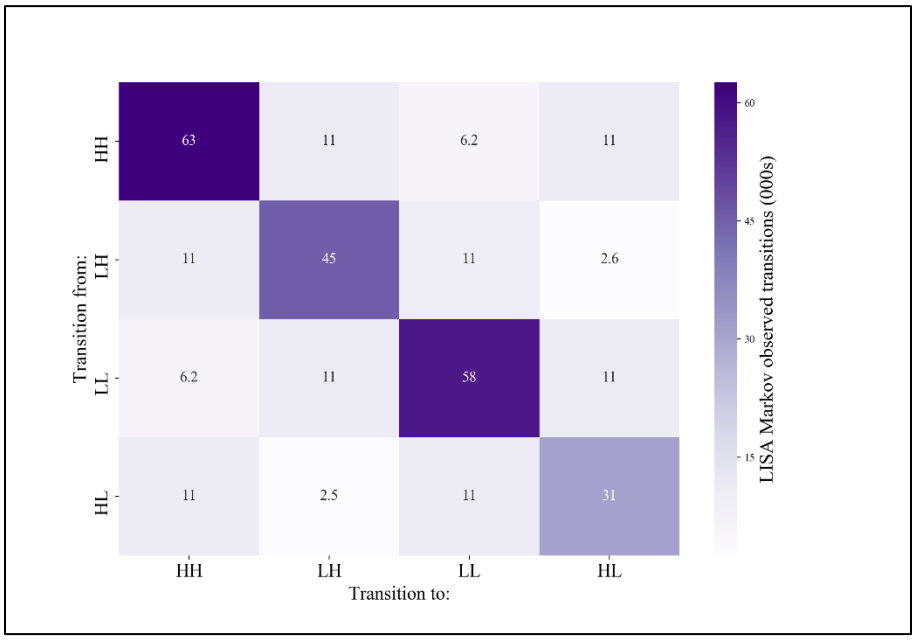
Figure 17: LISA Markov observed transitions of traffic speeds.



The spatial Markov framework is applied to local indicators of spatial association (LISA) that examine (dis)similar values of an individual spatial unit and its neighbouring values (Anselin, 1995). These are expressed as four High or Low permutations (HH, HL, LH, LL) based on the value of the spatial unit in relation to neighbouring spatial lag values. The LISA Markov framework exploits these four states as the possible transition classes for each time step. In our study, it investigates how TSE transitions co-evolve with spatial lag (Rey, 2015).

The complete LISA Markov TSE transitions are shown in Figure 17. This is compared to the expected number of transitions (if transitions were independent of spatial lag) in Figure 18. Most transitions were same-state which may be expected given the high temporal resolution. The difference between observed and expected transitions show that there were 9.1K more HH-HH, and 4.8K less LL-LL transitions than would be expected if transitions were independent of space.

Figure 18: LISA Markov observed - expected transitions of traffic speeds.



Methodology A: Random Forest Regression

Name: Robert Howd

Random Forest Regression

Random Forest (RF) was one method used to forecast the final seven days of our time-series data. RF is an extension of basic decision trees where each node represents a feature, each branch represents a decision, and each leaf represents an output. RF is a combination of multiple Classification and Regression Trees (CARTs). CARTs function by dividing data into increasingly similar groupings. While single decision trees are weak predictors, RF improves them by creating an ensemble of randomized decision trees, as shown in Figure A1 (Mennitt et al, 2014).

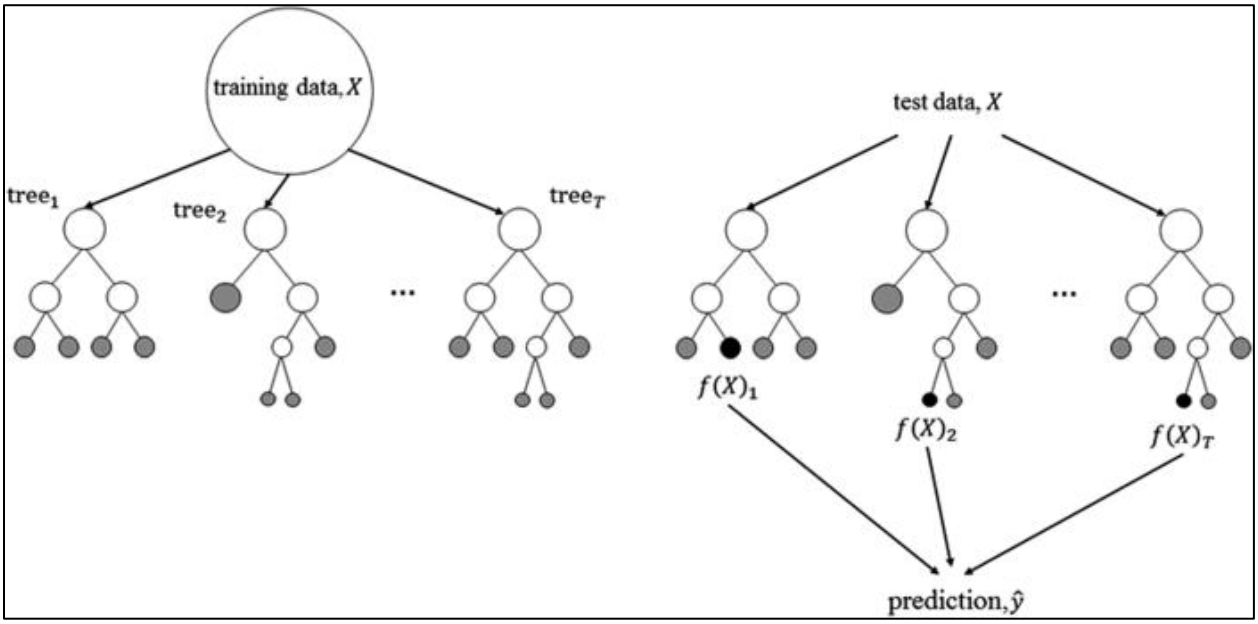


Figure A1: Example diagram of random forest process

In our study, we use RF regression to predict the TSE of our traffic flows. Forests for regression are formed by growing randomised trees so that the tree predictor takes inputs and outputs numeric values (Breiman, 2001). At each node, the observations of the *y variable* are averaged to make the prediction. In this way, RF algorithms are a type of "crowd sourcing." Splits are determined by measuring the reduction in variance using metrics such as sum squared errors.

RF is a non-linear algorithm that can handle large datasets with complex interactions between variables (Taalab et al., 2018). Some studies have shown RF to be a stronger model than ARIMA for time-series forecasting (Kane et al, 2014). A unique attribute of RF is the ability to assess feature importance. The algorithm is user-friendly as the two main parameters are (1) the number of trees in the forest and (2) the number of variables in the random subset at each node (Liaw and Wiener, 2002).

Experiment Setup

Data Transformation

We elected to use all of the training data (23 days) and maintain the five minute time intervals for forecasting. While this may increase model training time, RF models are encouraged to use all available data (Breiman, 2001). We combined the converted TSE data matrix of our study area with the time steps, seen in Figure A2.

DateTime	1882	1402	…	2416
1/1/2011 6:00	9.217436	8.357382537	…	17.66198923
1/1/2011 6:05	9.015595	7.796065772	…	11.37686701
…	…	…	…	…
1/30/2011 20:55	7.817319	8.098238894	…	8.988129826

Figure A2: Sample of initial data transformation

The dataset combines the dates and time stamp (5400 rows) to our road links (17 columns).

Inclusion of Spatio-Temporal Data

One downside of RF is the inability to explicitly account for spatio-temporal details or autocorrelation. We attempted to transform the data so that spatio-temporal details are present in the modeling. A spatio-temporal model, such as ARIMA, would be practical for forecasting the initial, non-adapted data as it is single variable (travel times). By incorporating spatio-temporal details into our data we create a multi-variable dataset for forecasting.

Temporal Data:

We incorporated temporal data into our analysis by two methods. First, shown in Figure A3, is the inclusion of binary variables to indicate time of the week (weekday vs weekend) and period of the day (AM/PM rush hour) defined in the ESTDA.

Is_weekend	morning_peak	evening_peak
0	0	0
1	0	1
…	…	…
1	1	0
1	1	0

Figure A3: Dummy variables to indicate time of day and week (road link 1402)

Second, we included the TSE of the six previous time steps for each road (Figure A4). We chose the six previous speeds based on the temporal PACF results shown in the ESTDA (Figure 13).

1402_prev_1	1402_prev_2	1402_prev_3	1402_prev_4	1402_prev_5	1402_prev_6
7.854682776	8.424780748	8.098238894	8.633659645	7.796065772	8.357382537
1.479706535	7.854682776	8.424780748	8.098238894	8.633659645	7.796065772
…	…	…	…	…	…
1.461080858	1.302584555	1.479706535	7.854682776	8.424780748	8.098238894

Figure A4: Previous time step TSE (road link 1402)

Spatial Data:

Additionally, we attempted to include spatial information. We did this by adding the TSE of the current time step of all linked/neighbouring roads for the road being predicted. No road had more than four neighbours. This structure is seen in Figure A5.

neighbour1	neighbour2	neighbour3	neighbour4
7.616423043	7.11563258	4.664979853	0
3.84829797	7.324915887	5.537524974	0
…	…	…	…
5.712317282	8.490243433	5.124089853	0

Figure A5: Spatial details included (road link 1402)

The final, combined dataset that we use for each road in our analysis is below in Figure A6.

DateTime	Speed	Is_weekend	morning_peak	evening_peak	neighbour1	neighbour2	neighbour3	neighbour4	1402_prev_1	1402_prev_2	1402_prev_3	1402_prev_4	1402_prev_5	1402_prev_6
1/1/2011 6:30	1.479706535	1	0	0	7.616423	7.11563258	4.66497985	0	7.854682776	8.424780748	8.098238894	8.63365964	7.796065772	8.35738254
1/1/2011 6:35	1.302584555	1	0	0	3.848298	7.32491589	5.53752497	0	1.479706535	7.854682776	8.424780748	8.09823889	8.633659645	7.79606577
1/1/2011 6:40	1.461080858	1	0	0	6.7701538	6.61187096	6.73060241	0	1.302584555	1.479706535	7.854682776	8.42478075	8.098238894	8.63365964
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1/30/2011 20:55	18.01160037	1	0	0	3.3540212	6.27849932	6.05574257	0	16.07188943	14.50934453	10.76982284	14.3105865	11.11354053	13.3932411

Figure A6: Complete, transformed data with additional temporal and spatial variables (road link 1402)

Parameter Tuning

With our data transformed and spatio-temporal aspects accounted for, we conducted a grid search to identify the optimal parameter. Although an advantage of RF is using the "out of box" parameter settings, a study by Mark Segal suggests that performance gains can be attained by additional tuning related to tree and node size (2004). The parameter range we tested was from 100-700 trees, and the node size as the number of features or the square root (sqrt) number of features.

The data was divided into *X* and *y* sets for training. The *y* data consisted of the TSE and the *X* set was the variables. These subsets are seen in Figures A7 and A8. The models were trained on subset one, while the

X dataset of subset two was provided to the trained models to predict the *y* of subset two. Subset two's *y* values were used to measure accuracy as they are the actual travel speeds recorded for the last seven days.

Subset 1 – Training Data

DateTime	Is_weekend	morning_peak	evening_peak	neighbour1	neighbour2	neighbour3	neighbour4	433_prev_1	433_prev_2	433_prev_3	433_prev_4	433_prev_5	433_prev_6
1/1/2011 6:00	1	0	0	5.523225417	7.616423043	7.11563258	4.664979853	7.128078091	7.38886142	7.270639621	8.114553147	7.017991928	7.734723009
1/1/2011 6:05	1	0	0	7.653612382	3.84829797	7.324915887	5.537524974	7.767777386	7.128078091	7.38886142	7.270639621	8.114553147	7.017991928
…	…	…	…	…	…	…	…	…	…	…	…	…	…
1/23/2011 20:55	0	0	0	3.307116454	7.460985852	8.894540678	6.063849301	6.18251671	6.757100045	7.669451076	7.801115491	8.114553147	6.885075396

DateTime	Travel Speed
1/1/2011 6:00	7.767777386
1/1/2011 6:05	7.902869186
…	…
1/23/2011 20:55	6.445602524

Figure A7: Model training data divided into X (Top) and y (Bottom) sets. The first 23 days of road 433

Subset 2 – Testing Data

DateTime	Is_weekend	morning_peak	evening_peak	neighbour1	neighbour2	neighbour3	neighbour4	433_prev_1	433_prev_2	433_prev_3	433_prev_4	433_prev_5	433_prev_6
1/24/2011 6:00	0	0	0	3.74652354	4.404678401	5.88300329	5.807301832	6.445602524	6.18251671	6.757100045	7.669451076	7.801115491	8.114553147
1/24/2011 6:05	0	0	0	3.369514881	4.569853826	6.278499325	5.859890602	7.605271561	6.445602524	6.18251671	6.757100045	7.669451076	7.801115491
…	…	…	…	…	…	…	…	…	…	…	…	…	…
1/30/2011 20:55	1	0	0	3.799665711	3.354021159	6.278499325	6.055742565	5.680187226	5.178518257	5.592799718	5.592799718	5.491419649	5.959540666

DateTime	Travel Speed
1/24/2011 6:00	7.605271561
1/24/2011 6:05	8.573867469
…	…
1/30/2011 20:55	5.863419064

Figure A8: Model testing data divided into X (Top) and y (Bottom) sets. The final seven days of road 433.

We used the Python tool GridSearchCV for our parameter search. GridSearchCV includes cross-validation in selecting the best parameter settings. The five-fold cross-validation executes training and validation on

multiple subsets of the data to evaluate the ability to predict new data. This helps avoid problems of overfitting.

An additional aspect of the tuning process was deciding which scoring measure GridSearchCV uses to identify the strongest parameters. Before attempting to model all our road links, we tested multiple metrics on two road links to inform our decision. The road links selected – 432 and 433 – represent roads with zero linked neighbours (432) and four linked neighbours (433). The metrics tested were Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2.

The structure of RF models ensures very few assumptions are needed in our data. Unlike an ARIMA model, we do not require time-series differencing for stationarity. However, we did test if standardised data increased the performance. Results of these parameter tests are shown in Figure A9.

Road	R2 Scoring Measure			MAE Scoring Measure			MSE Scoring Measure			Standardised (MSE Scoring Measure)		
	MAE	MSE	R2	MAE	MSE	R2	MAE	MSE	R2	MAE	MSE	R2
432	0.645	0.753	0.609	0.646	0.754	0.609	0.646	0.754	0.609	0.457	0.377	0.608
433	0.556	0.520	0.547	0.556	0.520	0.547	0.556	0.520	0.547	0.444	0.332	0.547

Figure A9: Scoring measure parameter test results.

Results show that choice of scoring metric made little difference and using standardised data did not increase performance. We used non-standardised data (for increased interpretability) and the MSE scoring metric for our parameter tuning.

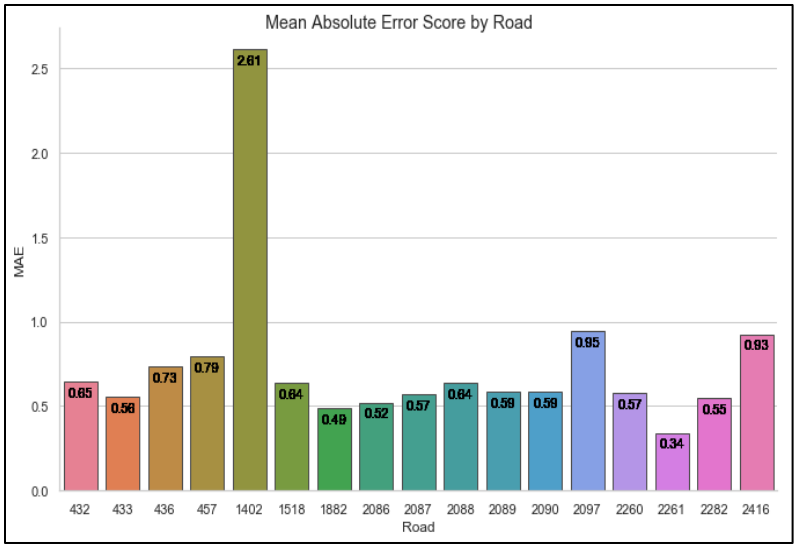
Random Forest Results

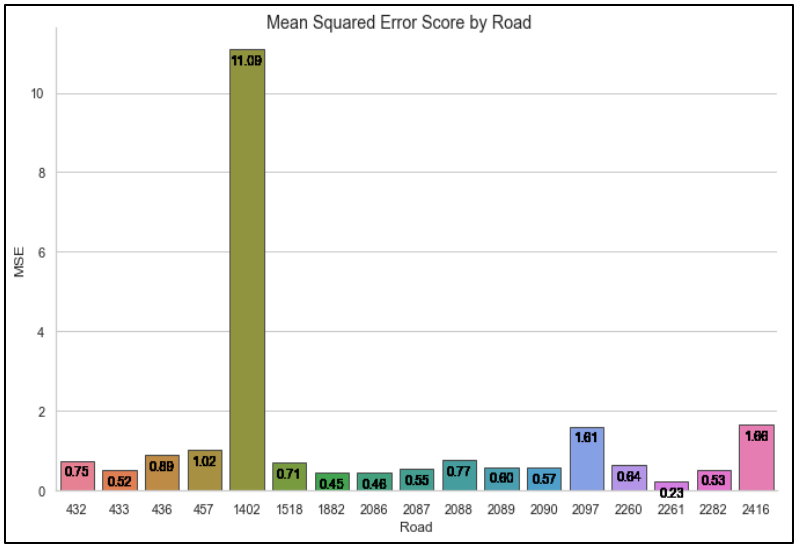
We created 17 different models customized for each road. Using the first 23 days to train each model we predicted the TSE of each five minute time step for the remaining seven days. The metrics recorded for each model were MAE, MSE, and R^2. Full results are presented in Table A1, while Figure A10 graph them individually.

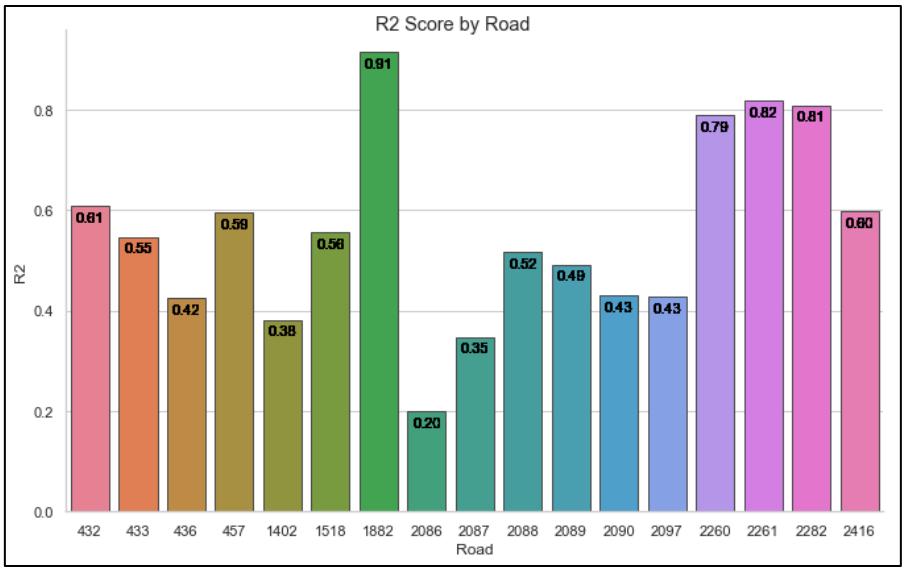
Table A1: Full random forest regression results predicting every time step of the final seven days for each road. Colour coding indicates strong (green) and weak (red) model performance.

Road	Metrics			Best Parameters	
	MAE	MSE	R2	Trees	Max Features
432	0.646	0.754	0.609	300	sqrt
433	0.556	0.520	0.547	500	N Features
436	0.734	0.894	0.425	700	sqrt
457	0.793	1.016	0.595	300	sqrt
1402	2.611	11.091	0.382	700	N Features
1518	0.640	0.708	0.556	700	sqrt
1882	0.486	0.450	0.915	500	N Features
2086	0.522	0.456	0.199	300	N Features
2087	0.571	0.547	0.348	500	sqrt
2088	0.640	0.774	0.516	700	sqrt
2089	0.588	0.597	0.492	400	sqrt
2090	0.587	0.574	0.430	700	sqrt
2097	0.946	1.610	0.428	700	sqrt
2260	0.574	0.643	0.789	500	sqrt
2261	0.343	0.231	0.819	700	N Features
2282	0.548	0.532	0.809	600	sqrt
2416	0.926	1.664	0.598	700	sqrt
AVERAGE	**0.748**	**1.357**	**0.556**		

Figure A10: Prediction scoring metrics by road: MAE (Top Left); MSE (Top Right); R2 (Bottom).



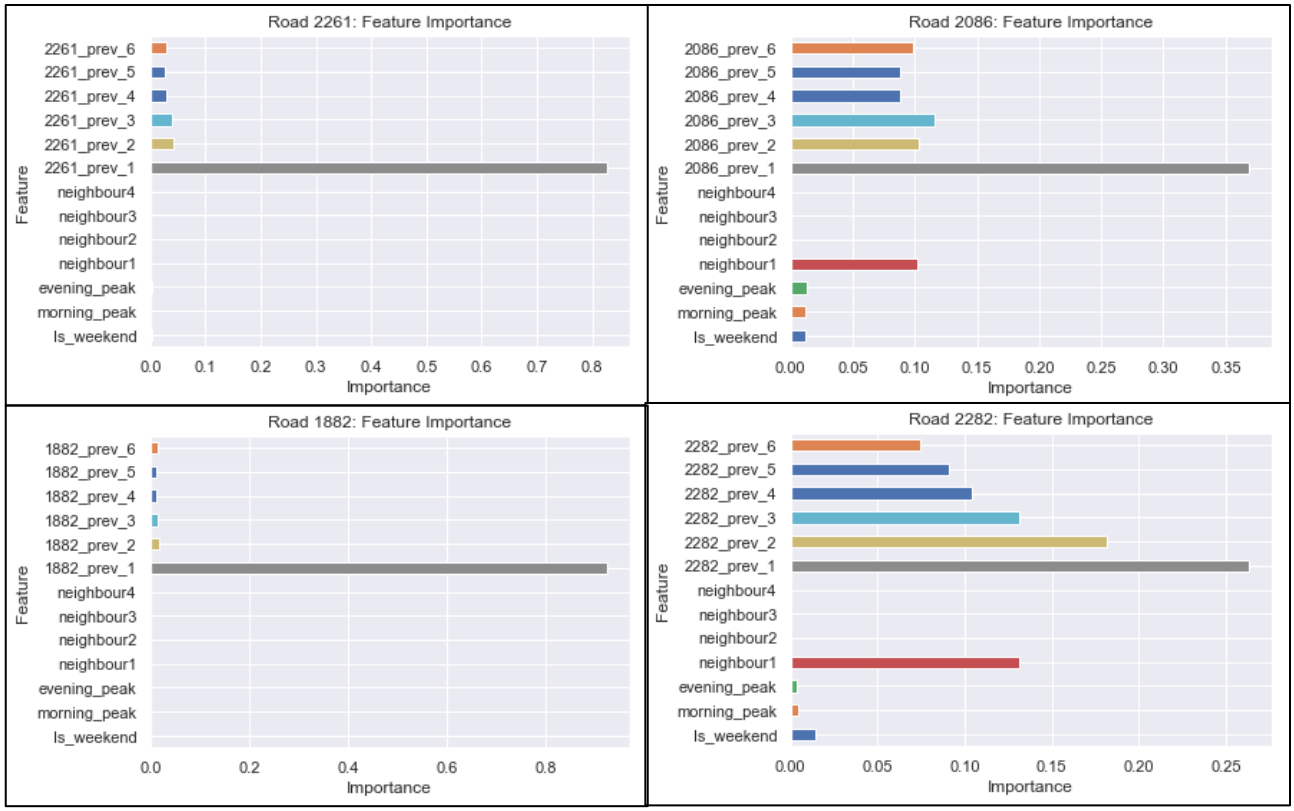




Road 1402 is the main outlier with the poorest MAE and MSE scores by a substantial margin. Perhaps the road work that affected road 1521 had additional impacts on 1402.

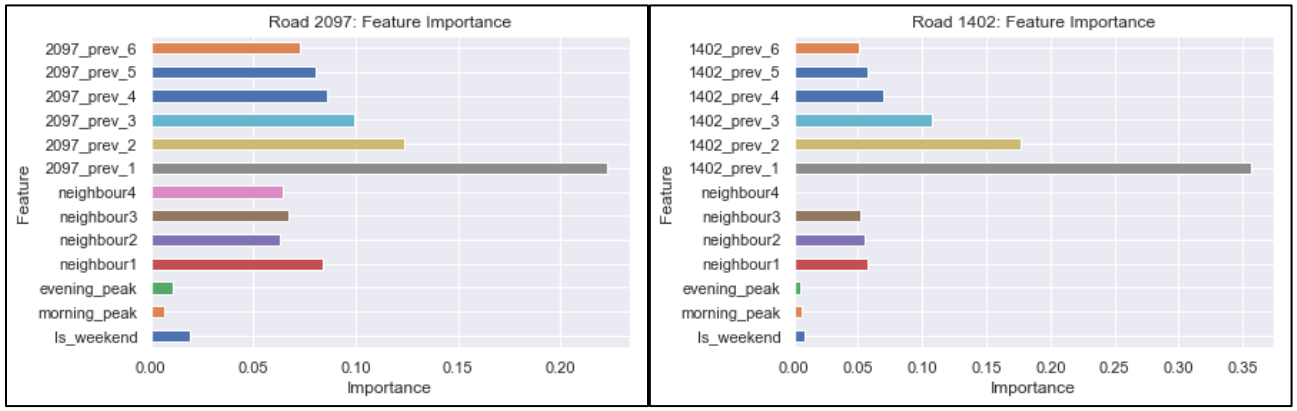
We also extracted the feature importance of each model. Figure A11 shows the feature importance of the top four performing models based on MSE score – 2261, 1882, 2086, and 2282.

Figure A11: Feature importance charts for the top four performing models – roads 2261, 1882, 2086, and 2282



The feature importance charts show that the previous time step data was the most important predictor of all models, and particularly the best models. Interestingly, the roads with high amounts of neighbours performed the worst (Figure A12). The temporal data related to time of the week and rush hours proved, comparatively, unimportant in all models.

Figure A12: Feature importance charts of poor performing models with high numbers of neighbours – roads 2097 and 1402



Finally, we can see a R^2 comparison of the predicted versus actual travel speeds of the best model (road 1882) and worst model (road 2086) in Figure A13.

Figure A13: Actual vs Predicted TSE for worst (Top) and best (Bottom) models

The Predicted speed is overlaid the Actual speed, showing the poor performance of the 2086 model in predicting TSE. It underestimated most TSE throughout the forecasted period (including most peaks) as compared to the 1882 model which was a much closer forecaster.

Methodology B: Support Vector Machine

Name: George Pyne

Overview:

The simplified theory behind the support vector machine (SVM) methodology is shown in Figure B1 where a separating line has drawn to separate and classify the blue and red data points. Many algorithms exist to execute this task to draw what is known as the 'separating hyperplane' (Kirk, 2017). Usually the maximum margin hyperplane methodology is used, this places the hyperplane at the furthest distance from the edge data points known as the 'support vectors' (James *et al.*, 2013; Figure B1).

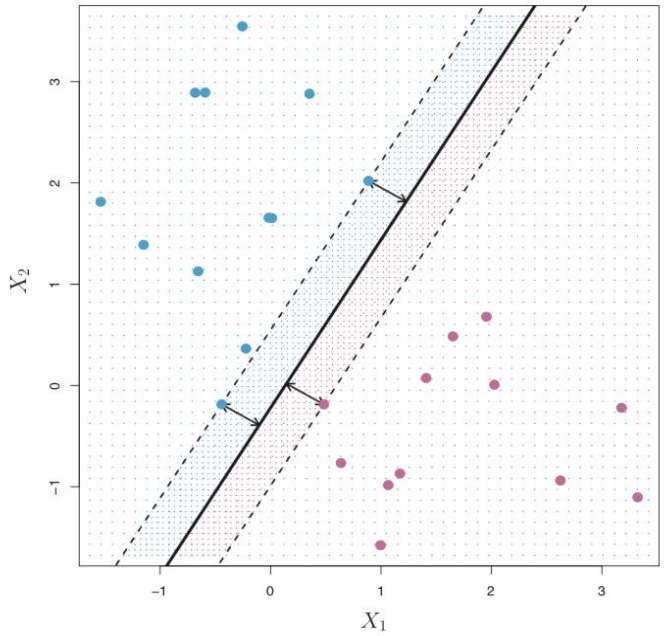


Figure B1: Maximum margin hyperplane methodology (James et al., 2013, pp. 340)

This becomes problematic when a boundary between the support vectors are non-linear (see Figure B2). The SVM algorithm can overcome this by using a 'kernel trick' to map the feature space into a higher dimension, this can convert the established linear model to a non-linear boundary. Once the SVM has mapped the input data into a higher dimensional feature space it can then identify the optimal separating hyperplane (Asif *et al.*, 2014). Figure B2 demonstrates this, whereby the kernel trick transforms the two-dimensional feature space into three dimensions in order to draw a separating hyperplane.

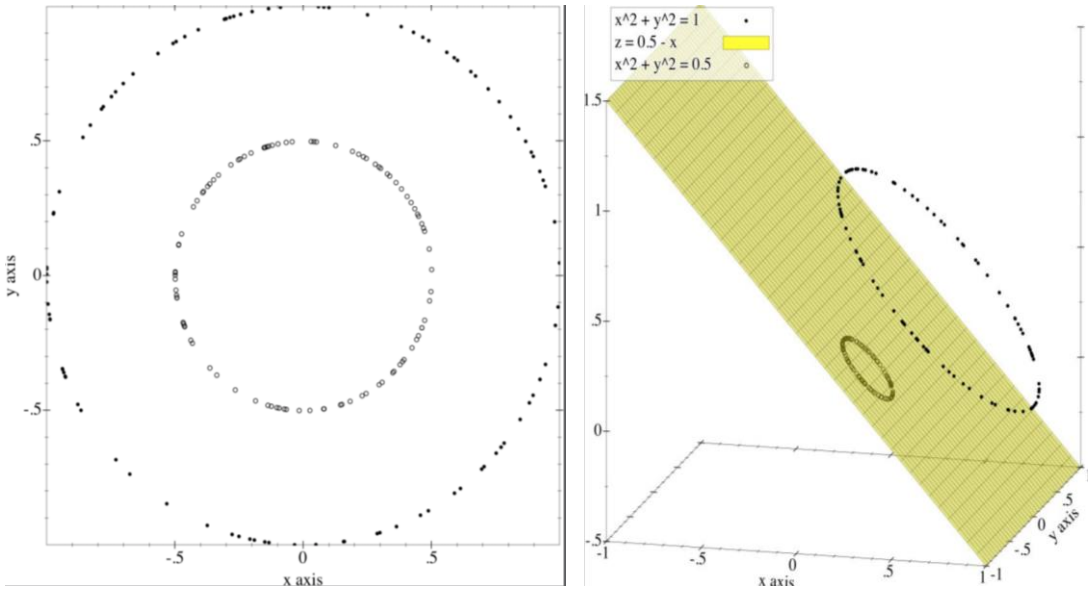


Figure B2: Non-linear class boundary (left) mapped into higher dimensional feature space (right) (Kirk, 2017, pp. 246)

The SVR extension to the SVM formulae can predict quantities. Where SVM is parameterised with a cost function (C) to minimise training errors, the SVR alternatively attempts to minimise a generalised error bound. The aim of traditional linear regression is to fit a model minimising the deviation between observed date and predicted responses. In SVR, the aim is instead to minimise this generalised error bound through a combination of regularisation functions (responsible for altering the complexity of feature space) and from the C function of errors (Basak *et al.*, 2007). The generalised error bound means that the SVR model isn't concerned about errors, provided they are generalised below a certain threshold (Smola and Schölkope, 2004).

Setup:

The temporally static (auto)correlation analysis of the road links (see Table 3; Figure 10) didn't necessarily show spatial dependence within TSEs. However, the literature argued this is vital due to upstream or downstream flow dynamics (Yue and Yeh 2008). Further analysis with the LISA Markov (see Figure 28) hinted at this dynamic as TSE transitions co-evolved with spatial lag in a way that seemed interdependent with space. Therefore TSEs were included for each of the individual road link's neighbours; each road had between 0-4 neighbours. It was hoped where temporally static (auto)correlation analysis failed to pick up these upstream and downstream dynamics, SVR may approximate this non-linear pattern.

Figure 13 established a degree of temporal autocorrelation, and partial autocorrelation suggested that additional information on average was provided by 6 temporal lags (30 mins). Therefore, for each of the 18 roads in the network the previous 6 time steps were added as variables. Lastly, the temporal dynamics of HRNs change in spatial locations over time, their cross correlations can change between the AM peak, inter-peak and PM peak periods (Cheng *et al*., 2014). This model attempted to include this by introducing time as a continuous variable throughout the day, and with dummy variables of whether it was the AM peak, PM peak or the weekend. The example training data is shown in Table B1.

Table B1: Example of the training data for individual and HRN SVR models

Weekend	AM peak	PM peak	Min	N1	N 2	N3	N4	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6
0	1	0	450	3.70	4.87	5.37	4.16	4.58	4.84	5.14	5.86	7.17	7.96
0	1	0	455	3.54	5.71	4.91	4.20	4.51	4.58	4.84	5.14	5.8	7.17
0	1	0	460	3.19	6.89	8.12	4.66	4.44	4.51	4.58	4.84	5.14	5.86
0	1	0	465	3.57	4.94	3.95	4.86	4.34	4.44	4.51	4.58	4.84	5.14
0	1	0	470	2.95	5.37	3.73	4.75	4.10	4.34	4.44	4.51	4.58	4.84

Multiple methodologies exist for optimising SVR hyperparameters such as: bootstrapping (Chatzimichali and Bessant, 2015), genetic algorithms (Cheng *et al*., 2017) or brute-force methods such as grid-search (sDong et al., 2005). Grid-search was selected as the benefits from heuristic approaches can still be achieved by brute-force when a smaller dimension of parameters exist (Jain *et al*., 2014). Overfitting is the phenomenon whereby the results returned from a model are excessively tailored to the training data (Xiao, 2017). Therefore, *k*-fold cross validation was used to help prevent this. This method randomly divides the data into *k* folds which are equally sized - the first fold is used as a validation fold and the SVR is fitted to the remaining $k - 1$ folds (James *et al*., 2013).

The grid-search process was fit to the first 23 days of TSE data, then the optimal parameters were recorded as an average of the cross-validation. Results were recorded by cross-validated average MSE an MAE scores of the SVR predictions for the unseen data of the last seven days. Two SVR models were deployed, one was fitted to each road individually (k=5) and another (the HRN SVR) was fitted to the entire road network (k=3). This was to measure differences in performances, which could help uncover the spatiotemporal complexities of each individual road, and the HRN as a whole.

Table B2: Best parameters and model evaluations of each road segment (cross-validation k = 5)

Road ID	Neighbours	C	gamma	kernel	MAE	MSE
1882	0	150	0.001	rbf	0.477269348	0.459030677
1521	1	20	0.001	rbf	0.429806549	21.76595842
433	4	10	0.001	rbf	0.534612111	0.659329267
1518	0	20	0.001	rbf	0.645906549	0.84571132
2090	1	10	0.001	rbf	0.569341067	0.603661534
436	1	10	0.001	rbf	0.717271785	1.013768359
2097	4	10	0.001	rbf	0.935295798	1.888613296
2087	1	10	0.001	rbf	0.554925706	0.6275362
2260	3	15	0.001	rbf	0.583772028	1.478270722
2088	0	15	0.01	rbf	0.633001134	0.7754417
457	0	150	0.001	rbf	0.765003731	1.062882127
2261	0	115	0.001	rbf	0.327305972	0.145990209
2089	0	75	0.001	rbf	0.577662214	0.572471309
2282	1	20	0.001	rbf	0.565850535	0.5425806
432	0	50	0.001	rbf	0.631427348	0.735899621
1402	3	10	0.001	rbf	2.686127993	12.43422126
2086	1	10	0.001	rbf	0.495992673	0.476998339
2416	4	10	0.001	rbf	0.865790568	4.172054071

Results:

Table B2 shows the results from each SVR model. This shows that the RBF kernel was universally the most effective for predicting TSEs, which supports much of the literature (Asif et al., 2014; Cheng et al., 2017; Hong et al., 2011; Ahn et al., 2015). There was a wide range of optimised C values which indicates the spatiotemporal heterogeneity within the HRN.

The individual SVR models had a wide range of accuracies (see MSE and MAE; Table B2), MSE for example had a range between 0.45 and 21.7. Individual SVR models performed modestly well across the study area

though, but performed poorly on roads 1521 and 1402. Table B2 suggested that the number of neighbours (and thus spatial dependence) may not be the most important feature for SVR accuracy – this was because 'blind forecasts' (where neighbours = 0) were surprisingly accurate (see Figure B4).

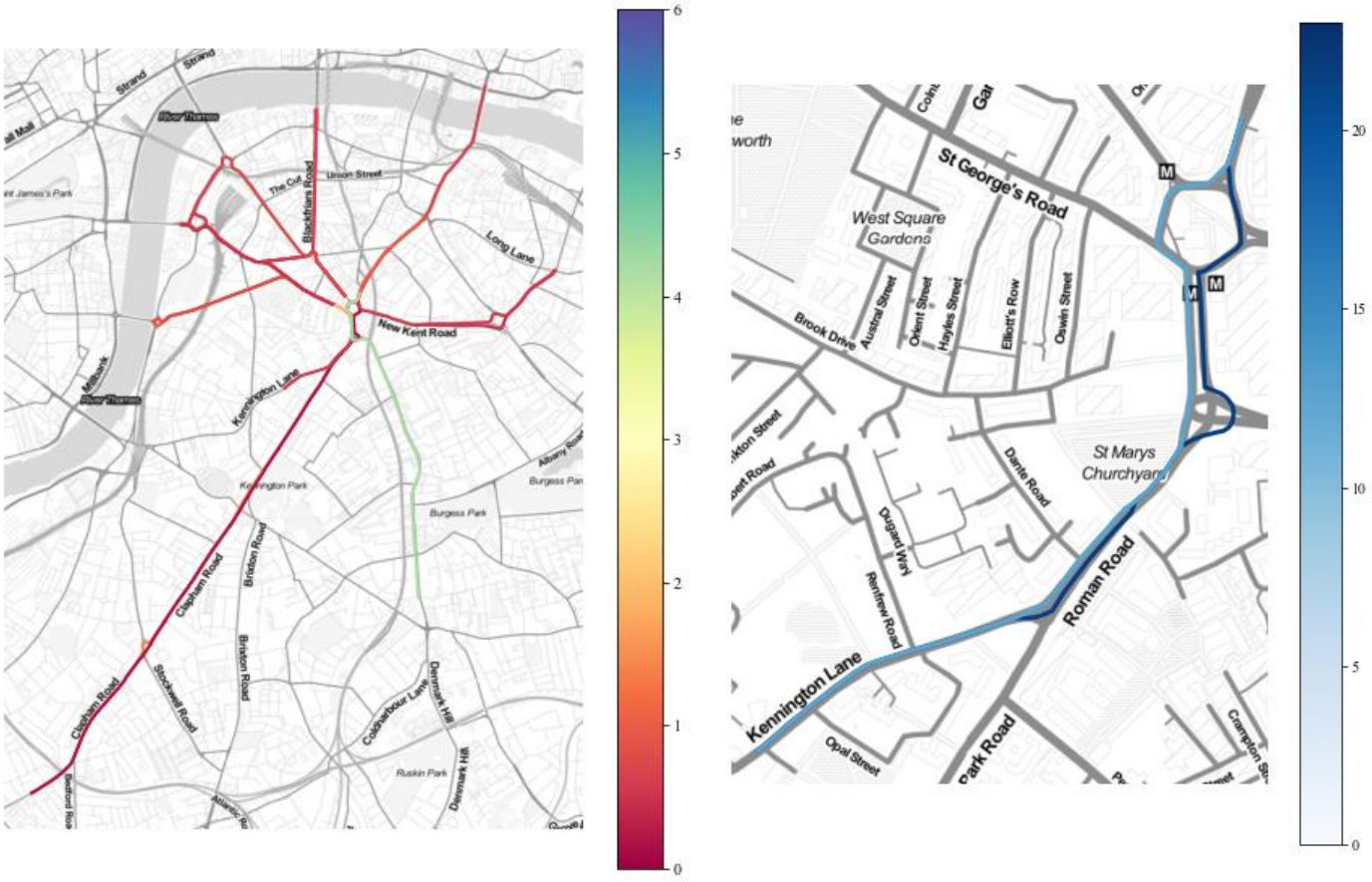


Figure B3: Map of multiple SVR MSE scores < 6.0 (left) and > 10.0 (right)

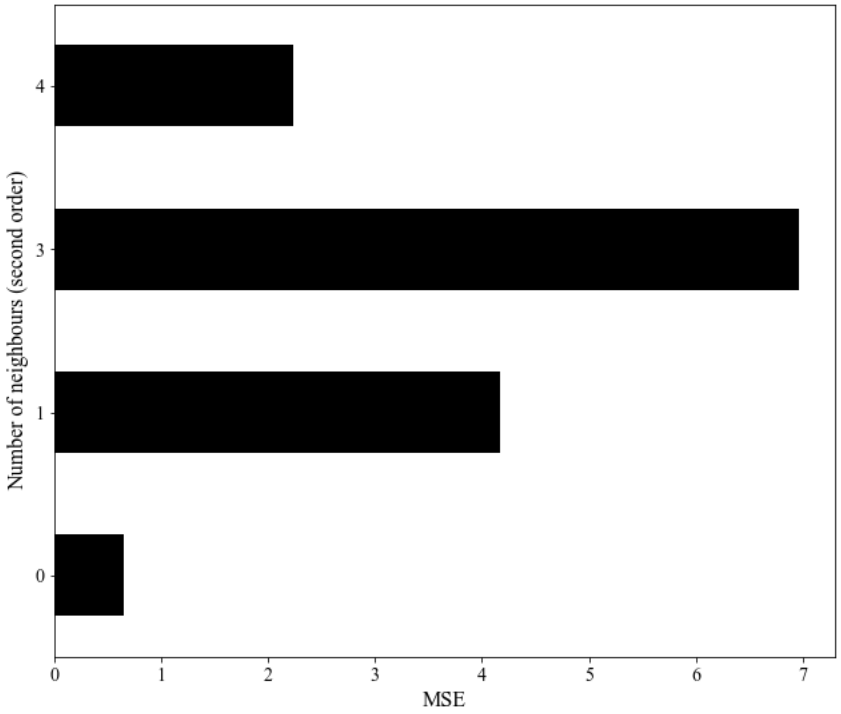


Figure B4: MSE scores averaged per neighbour count.

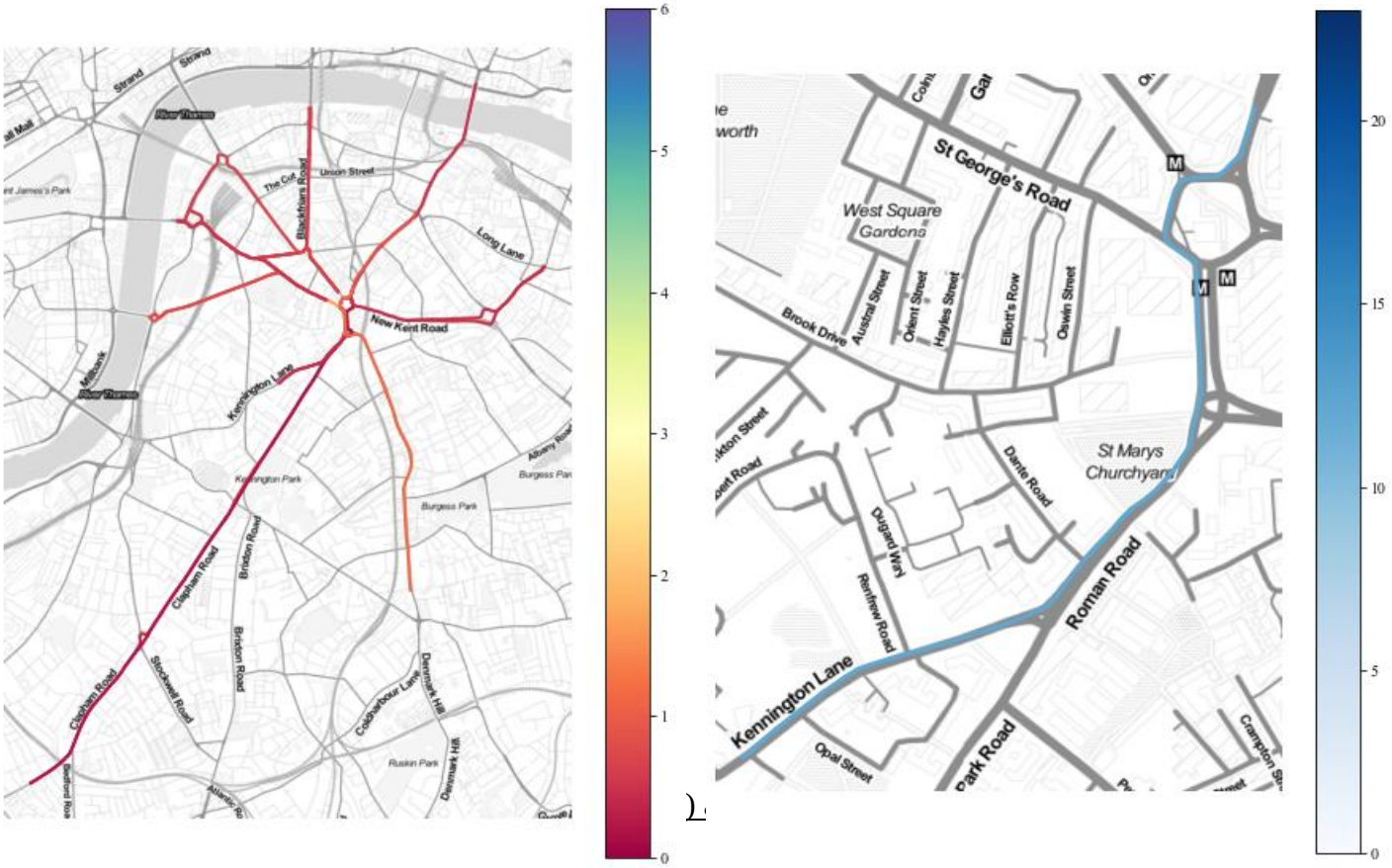
When individual SVR results are projected spatially (Figure B3) it shows individual SVR performance modestly for larger roads across the study area, but with higher errors in the South-East. When roads with the highest prediction errors (1521 and 1402; Table B2) were removed and mapped individually they appeared largely contiguous - so the cause of these error may have been shared.

Figure 5: TSE prediction cross-validation (k=3) scores for individual SVR and HRN SVR

Road ID	**Neighbours**	**MAE** (individual)	**MAE** (HRN SVR)	**MSE** (individual)	**MSE** (HRN SVR)
1882	0	0.477269348	0.51207801	0.459030677	0.47951708
1521	1	0.429806549	0.36525089	21.76595842	0.21675984
433	4	0.534612111	0.51896679	0.659329267	0.47728876
1518	0	0.645906549	0.6465593	0.84571132	0.71642713
2090	1	0.569341067	0.58215663	0.603661534	0.58226992
436	1	0.717271785	0.74021902	1.013768359	0.92152983
2097	4	0.935295798	0.94601044	1.888613296	1.74451837
2087	1	0.554925706	0.58906092	0.6275362	0.60664625
2260	3	0.583772028	0.51268359	1.478270722	0.50797861

2088	0	0.633001134	0.62492073	0.7754417	0.7544221
457	0	0.765003731	0.75622269	1.062882127	0.92422691
2261	0	0.327305972	0.32201677	0.145990209	0.19522396
2089	0	0.577662214	0.57610318	0.572471309	0.57682432
2282	1	0.565850535	0.52009531	0.5425806	0.45900071
432	0	0.631427348	0.62377485	0.735899621	0.70215156
1402	3	2.686127993	2.62897101	12.43422126	12.2986748
2086	1	0.495992673	0.54743523	0.476998339	0.49764379
2416	4	0.865790568	0.74665714	4.172054071	1.15459341

Table B3 shows the HRN SVR compared to individual SVR, the most notable result was a severe drop in MSE with road 1521 from 21.7 to 0.21. The HRN SVR outperformed individual SVR in 17 roads when scored by MSE and 11 roads with MAE (Table B3). When these were mapped (Figure B5) the generalised MSE performance improvements are shown across the study area with improvements in the South-East and with only one roads with an MSE > 10.



Evaluation:

Overall the results support the literature that SVR is a capable methodology for TSE forecasting (Cheng *et al*., 2017; Wu et al., 2004; Su and Zhang 2007; Lam and Toan 2008; Ahn et al., Hong et al., 2011; Asif *et al*., 2014). The HRN SVR model was more accurate than individual SVR, despite individual differences in optimal hyperparameters (see C values; Table B2). The most notable example of this improved performance was the reduction of errors for Road 1521 when the information of the entire HRN was available for SVR. This suggests the HRN SVR might have better approximated the non-linear spatiotemporal dynamics which were hinted in the CCF and LISA Markov analysis (see Figure 15-18).

It is hard to infer what information reduced the error of Road 1521 however. Only five other instances of single-neighbour road links became available to the HRN SVR, whereas 17 further instances of temporal training data became available. Crudely, this might suggest that temporal information was more useful for TSE forecasting, especially given the high degree of accuracy of 'blind forecasts' where neighbours = 0 (see Figure B4).

These interpretations of space–time feature importance should be taken with caution, this a limitation of SVR as it can't assess feature importance; it becomes hard to adjudge the ergodicity of the HRN. Further limitations were that no explicit spatial information was included in the SVR, instead only neighbouring road TSE records were included. Although the training data wasn't geographically referenced it still may be influenced by the modifiable areal unit problem (MAUP) where a modification of the aggregation criterion (of road start and end points) or areal scale of analysis could alter the outcome from the SVR (Arbia and Petrarca, 2011).

Six temporal lags were inputted for each road, this was the average of the whole network. This was a limitation as partial autocorrelation of each individual road may yield improve results. The handling of the temporal dimensions of TSE data such as the aggregation, segmentation (5 minute intervals) or boundaries (between 6AM-9PM) may have also influenced the SVR results due to the modifiable temporal unit problem (MTUP; Cheng and Adepeju, 2014).

The autoregressive nature of the data may have missed exogenous information such as influences from London's entire road network, or temporal influences (such as one-off events or cyclical seasonal changes). Overall the results from SVR, whilst moderately accurate, tell us little about the actual spatiotemporal dynamics of traffic flow. Cheng *et al*. (2017) for example, used SVR to dynamically predict further characteristics such as traffic occupancy (the concentration of traffic over different lanes; Arasan and Dhivya, 2009) and flow (how concentrated traffic was in relation to TSE; Arasan and Dhivya, 2009) to better understand spatiotemporal dynamics.

Methodology C: Long-Short Term Memory Neural Network

Name: Thomas Keel

Overview:

A specialised recurrent neural network (RNN) called a Long-Short Term Memory network will be used to forecast the TSE data in this section (LSTM). In its simplest form, an artificial neural network (ANN) is a non-linear 'black-box' model which comprises connections between an input node layer, hidden node layer(s) and output layer (see Figure C1). In this framework, data are input and sent to a hidden layer via weighted synapses. The weighting implies how useful an input is to the output of the model, relative to other inputs (TDSB, 2016). Weights are initially randomised, before the network 'learns' and adjusts these weights to minimise the prediction error in the output layer (a process called back-propogation). In the hidden layer, the nodes apply bias to the sum of the weighted inputs it receives from the input layer and a non-linear activation function determines whether these weighted sums are 'fed-forward' to make a prediction in the output layer or 'not'. The output layer sums the activated information from the hidden layer and makes a prediction, it then back-propogates the error to adjusts the values of bias and weights throughout the model (TDSB, 2016).

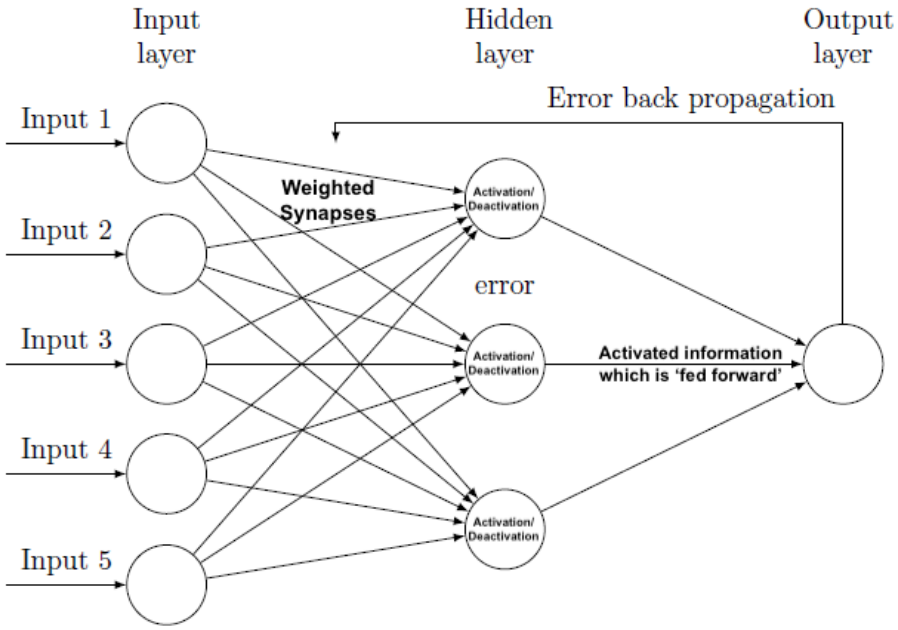


Figure C1: Basic Structure of a "Feed-Forward" Artificial Neural Network

A RNN, of which the LSTM is a type, is a form of ANN in which the number of hidden neurons in the hidden layer is determine by the number of time-steps in the input (Figure C2; Olah, 2015). In this model the useful parts of the output of one neuron in the hidden layer is passed into the next neuron in the hidden layer along with new input data for each time step, creating a loop (Olah, 2015). This allows information important for prediction to persist across time-steps.

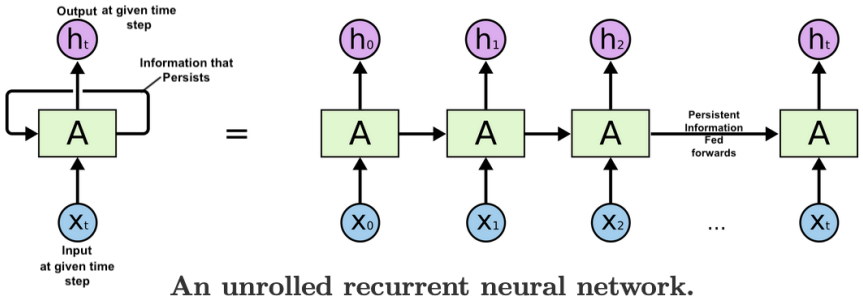


Figure C2: Structure of a RNN (adapted from: Olah, 2015)

The LSTM specifically uses a 'memory cell' to allow information from the previous time-step to be 'remembered' or forgotten based on how useful the input was to the prediction at previous time-steps (Ma et al., 2015). In the literature, there is a wide use of specialised forms of RNNs for traffic-flow forecasting (Vlahogianni et al. 2014), as they can handle highly dimensional data and can keep important long-term temporal information (Ma et al. 2015). Despite this, the LSTM, inherently have no mechanism to capture spatio-temporal structures/trends in the data (Cheng et al., 2011; Han et al., 2019).

Setup:
The LSTM has been applied in two distinct ways to forecast the traffic flow across the study area to:

F1. Model the individual road links and their neighbours at 6 time-lags (see Figure 3).

F2. Model the entire network at 6 individual time-lags.

Note, the use of 6 time steps have been informed by the PACF (Figure 9b), where on average across the road network, additional temporal information above 6 lags drops below the 95-percentile confidence interval. Of course, the temporal, and spatio-temporal, dependency of each road link to the others inherently varies throughout the links (see CoD in Figure 15), 6 time steps are preferred to produce are comparable models.

The data has been split into training and testing sets containing the first 23 days (4140 5-minute intervals) and last 7 (1260 intervals). Note that for models containing time-lag (1-6), the training set has thus been reduced (i.e. 4134–4140). For the forecast of the entire network (F2), the data has been differenced at a temporal lag of 1 (see Figure 8) and will be used in conjunction with un-differenced data to the compare effectiveness of the LSTM for forecasting (relatively) stationary and non-stationary inputs (henceforth, F2D and F2U).

The LSTM has been built using the Keras library. The model contains 4 neurons in its hidden layer and weights are adjusted via back-propogation using the 'Adam' optimisation algorithm (Figure C1; Kingma & Ba, 2014). This iterative algorithm is chosen as it can better deal with long-term temporal information held by the 'memory cells' (Kingma & Ba, 2014; Brownlee, 2017b). The model is explicitly told how many temporal lags to include in its prediction (here known as amount of 'look-back').

A grid-search has been employed for hyper-parameter tuning (as opposed to heuristic tuning) for the number of epoch and batch size used in each LSTM model. Specifically, the epoch is the number of runs that the model is trained on, and the batch size is the number of inputs processes before estimating error and adjusting the weights via back-propagation in each run (Brownlee, 2017a). This grid-search is limited though, to 3 epoch, 2 batch sizes and 6 look-back values, due to limitation in GPU.

As LSTMs, inherently automate feature selection no changes to the input features, other than differencing (Pavluk, 2019). The Mean-Squared Error (MSE) and Mean-Absolute Error (MAE) will be used to evaluate each model's performance.

Results:

Figure C3, shows the hyper-parameter tuning carried out using look-back values between 1-6. Here, there is a large gap between the performance of 50 and 100 epochs, which indicates that the model is optimised between this range. Roughly, batch sizes of 64 perform better across all values of look-back. When more temporal information is included i.e. using more than 1 look-back step the models tend to perform better. This relative improvement with more time-lag highlights that some degree of lagged temporal information is needed to forecast the network, something which is expected in HRNs (Asif et al., 2014).

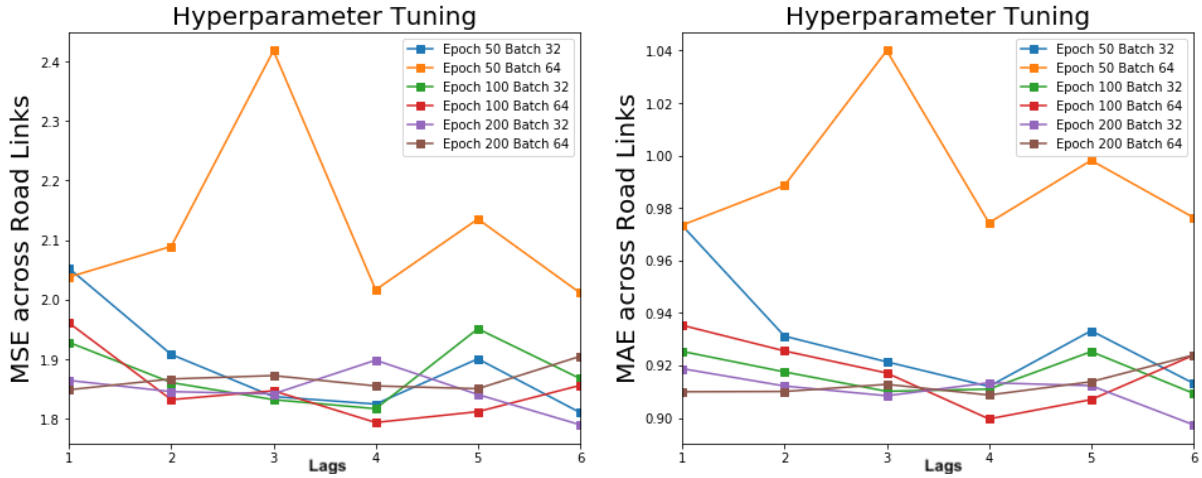


Figure C3: Results from hyper-parameter tuning of the differenced data (F2D) across the entire road network which is evaluated by mean-squared error (left) and mean-absolute error (right).

As demonstrated in Table C1, the LSTM models aiming to forecast the entire network perform (F2) better than the individual forecast (F1) for all links except 2090, 2088 and 432. From figures 2-3, we know that these specific links only have no direct neighbours and are towards the north-west of the central Elephant Castle roundabout. On average, the differenced data (F2D) performs better than both the other models and a lot better on links 2282 and 2260, which were found to share high CoD with the rest of the network (Figure 15). Across the time lags and over all road links, both the F2D and F2U share similar scores (Table C2).

Table C1: Comparison of results from two distinct version of the LSTM model at a look-back (time-lag) of 1 across each individual road link.

Forecast Type	*F1. Individual*		*F2U. Un-differenced*		*F2D. Differenced*	
Road Links	***MAE***	***MSE***	***MAE***	***MSE***	***MAE***	***MSE***
1882	0.66	0.79	0.57	0.57	**0.51**	**0.48**
433	2.51	23.49	0.66	0.70	**0.59**	**0.64**
1518	0.72	0.92	**0.70**	**0.83**	0.77	1.02
2090	**0.61**	**0.64**	0.62	0.66	0.68	0.83
436	0.81	1.12	**0.74**	**0.92**	0.86	1.25
2097	1.07	1.97	1.08	**1.94**	**1.05**	2.17
2087	0.60	0.65	**0.58**	**0.60**	0.69	0.85
2260	0.98	1.84	1.24	2.41	**0.58**	**0.70**
2088	**0.71**	**0.88**	0.73	0.91	0.71	1.04
457	0.87	1.25	0.94	1.42	**0.86**	**1.20**
2261	0.87	1.25	0.64	0.64	**0.33**	**0.23**
2089	0.88	1.26	0.70	0.80	**0.65**	**0.77**
2282	0.61	0.62	0.87	1.27	**0.57**	**0.57**
432	**0.62**	**0.63**	0.70	0.83	0.72	1.01
1402	**2.52**	14.75	2.75	12.81	2.56	**12.19**
2086	2.47	14.76	**0.55**	**0.52**	0.63	0.66
2416	1.55	5.03	1.74	5.21	**0.85**	**1.58**

Table C2: Overall results of LSTM model for Forecast F2 comparing the impact of differencing of the model's results.

Time Lags	*1*		*2*		*3*		*4*		*5*		*6*	
Forecast	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
F2U	0.81	1.59	0.80	1.59	0.80	1.58	0.80	1.57	0.79	1.57	0.79	1.57
F2D	0.80	1.58	0.79	1.58	0.80	1.58	0.79	1.58	0.80	1.60	0.80	1.56

The model results are shown visually in Figure C4+C5 and mapped onto the road network in Figure C6 for models using up to 6 look-back steps. Note that, in most of the models the addition of more temporal information past 1 lag in the model has a relatively negligible impact. The F1 models are shown to perform a lot worse for a selection of road links.

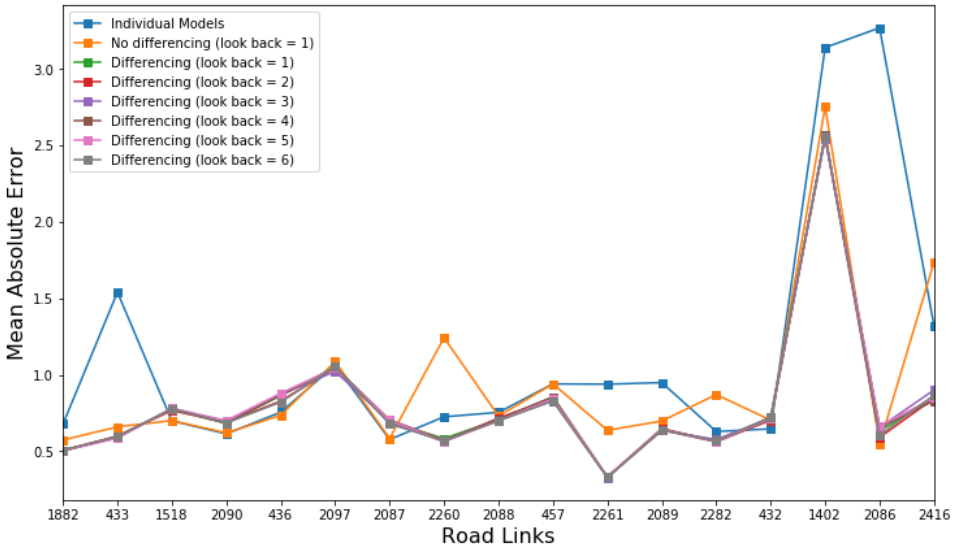


Figure C4: Comparison between the LSTM models using Mean-Absolute Error across the road links

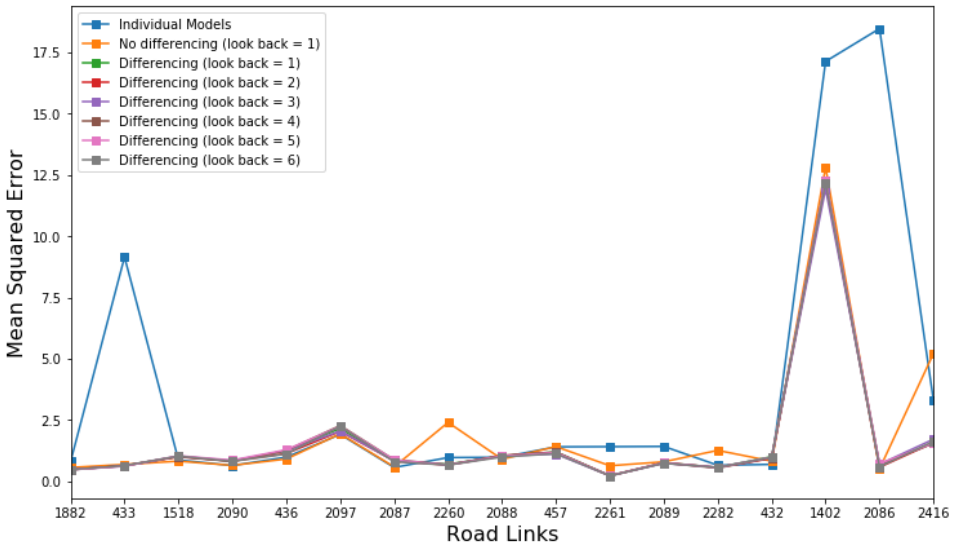


Figure C5: Comparison between the LSTM models using Mean-Squared Error across the road links.

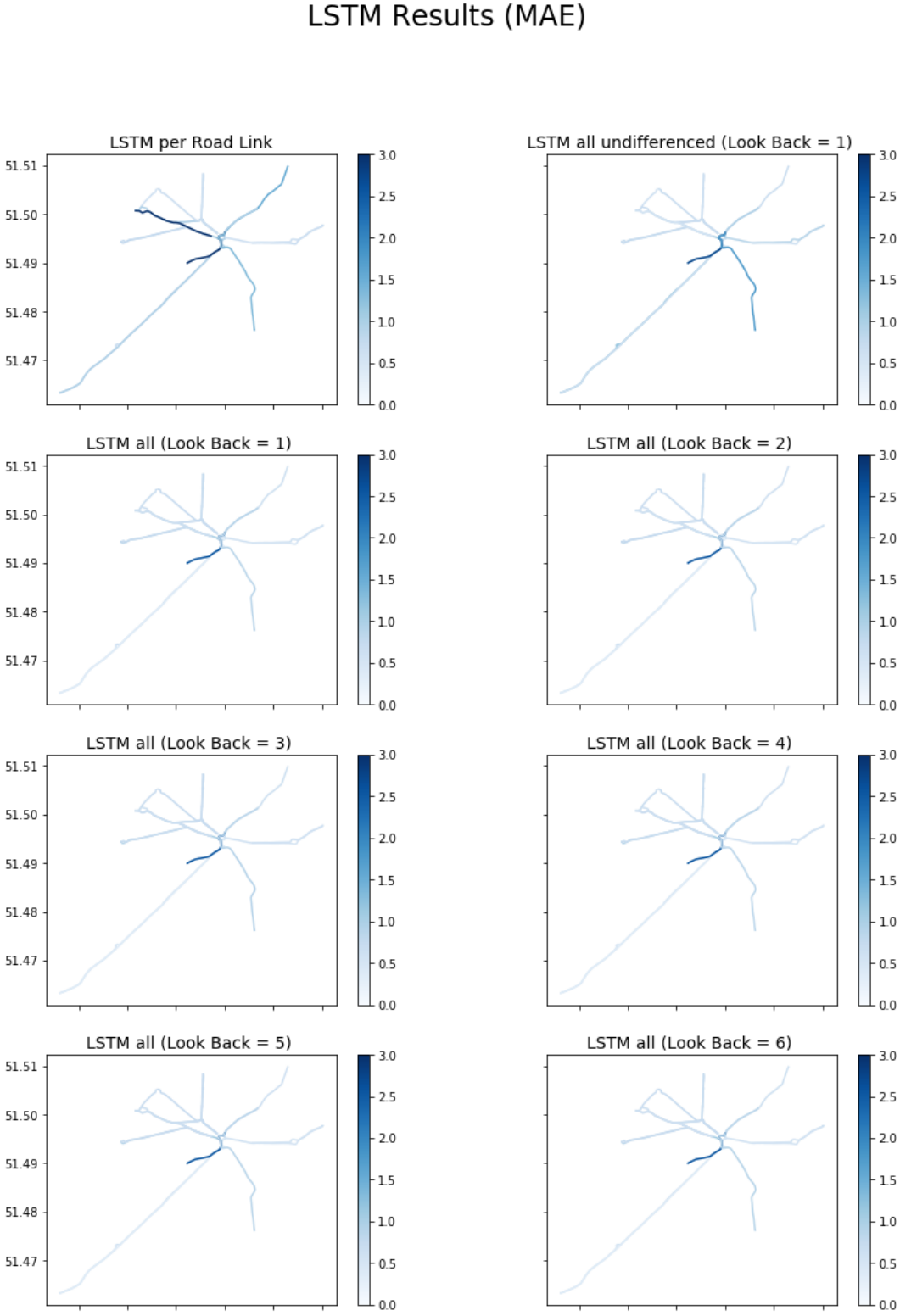


Figure C6: Comparison of Mean-Absolute Error between the LSTM models mapped onto the study area.

Finally, the prediction for 31st January 2011 is shown for two road links with contrasting performance scores in Figures C7-C8. Note, no models of '1402' come close to the accounting for the variance of the differenced TSE. In '2086', where the LSTM follows the general trend effectively.

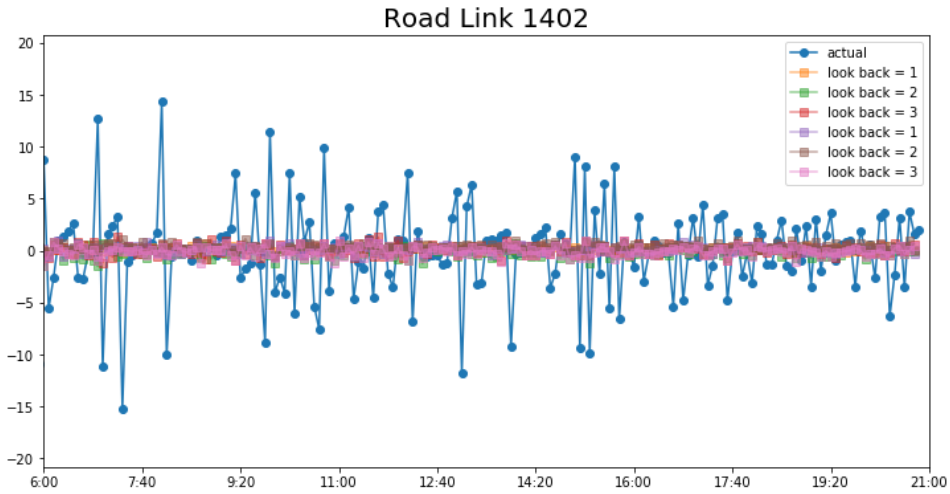


Figure C7: Forecast made the LSTM models for the 31st January for Road Link 1402 (data has been differenced).

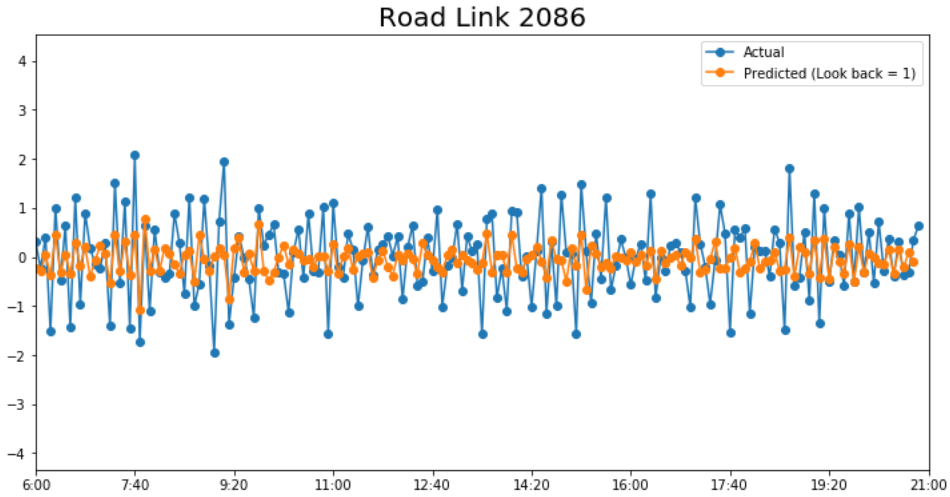


Figure C8: Forecast made the LSTM models for the 31st January for Road Link 2086 (data has been differenced).

Model Evaluation:

The LSTM model was moderately effective in forecasting the traffic flow across most of the section of road network chosen for this study. Although, the model's ability to forecast the TSE in each road link was improved when it had been trained on the overall the network (F2), as opposed to on individual roads neighbour pairings (F1). This may indicate that there is certain degree of spatial and temporal dependency which is not accounted in most of the individual models as opposed to modelling the overall network. This is not un-expected, however, as the forecasted road links rely purely on a second-degree adjacency matrix, in reality, individual roads have dependence to upstream and downstream traffic flow at much higher spatio-temporal orders (Cheng et al., 2017).

A major strength of ANN in general they do not require any prior assumptions to be made about the data thus they can deal with highly-dimensional and varied inputs (Ma et al., 2015). Also, LSTM inherently prevent over-fitting of the data by using a drop-out function to forget information passed a threshold (Brownlee, 2018).

A major drawback to LSTM models is that they are a 'black boxes', which mean that we have been given no direct indication of 'cause and effect' between the road link at given time-lags i.e. we do not know which link the model finds to be important for the forecasting of others (Pavluk, 2019). Moreover, this form of LSTM does not explicitly include spatio-temporal information, which may have improved the forecast of road links with a greater CCF with the others (i.e. 2282; see Figure 15). One such way to make this improvement could be integrating Convolution neural networks (CNN) into the LSTM framework, which can allow the treatment of spatial units as the 'memory cells' (Zhao et al., 2017; Han et al., 2019).

Discussion, and Conclusion

Summary of Models

The combined performance results of the RF, (individual) SVR, and LSTM methods are shown in Table 4. Here, each method was trained on each individual road link for cross-comparison and the results from the LSTM model used differenced input data. The overall lowest individual MSE and MAE scores were recorded by the SVR, but RF scored the lowest on average, lastly the LSTM had the highest interquartile range.

Table 4: Overall results evaluated by Mean-Squared Error and Mean-Absolute Error. Note: LSTM uses differenced TSE data at temporal lag of 1.

Road	RF		SVR		LSTM	
	MAE	MSE	MAE	MSE	MAE	MSE
432	0.646	0.754	0.631	0.736	0.723	1.011
433	0.556	0.520	0.535	0.659	0.590	0.639
436	0.734	0.894	0.717	1.014	0.863	1.251
457	0.793	1.016	0.765	1.063	0.856	1.206
1402	2.611	11.091	2.686	12.434	2.566	12.193
1518	0.640	0.708	0.646	0.846	0.778	1.023
1882	0.486	0.450	0.477	0.459	0.507	0.483
2086	0.522	0.456	0.496	0.477	0.634	0.656
2087	0.571	0.547	0.555	0.628	0.691	0.848
2088	0.640	0.774	0.633	0.775	0.708	1.038
2089	0.588	0.597	0.578	0.572	0.647	0.770
2090	0.587	0.574	0.569	0.604	0.684	0.834
2097	0.946	1.610	0.935	1.889	1.051	2.166
2260	0.574	0.643	0.584	1.478	0.583	0.704
2261	0.343	0.231	0.327	0.146	0.331	0.234
2282	0.548	0.532	0.566	0.543	0.565	0.571
2416	0.926	1.664	0.866	4.172	0.846	1.584
AVERAGE	**0.748**	**1.357**	**0.739**	**1.676**	**0.801**	**1.601**

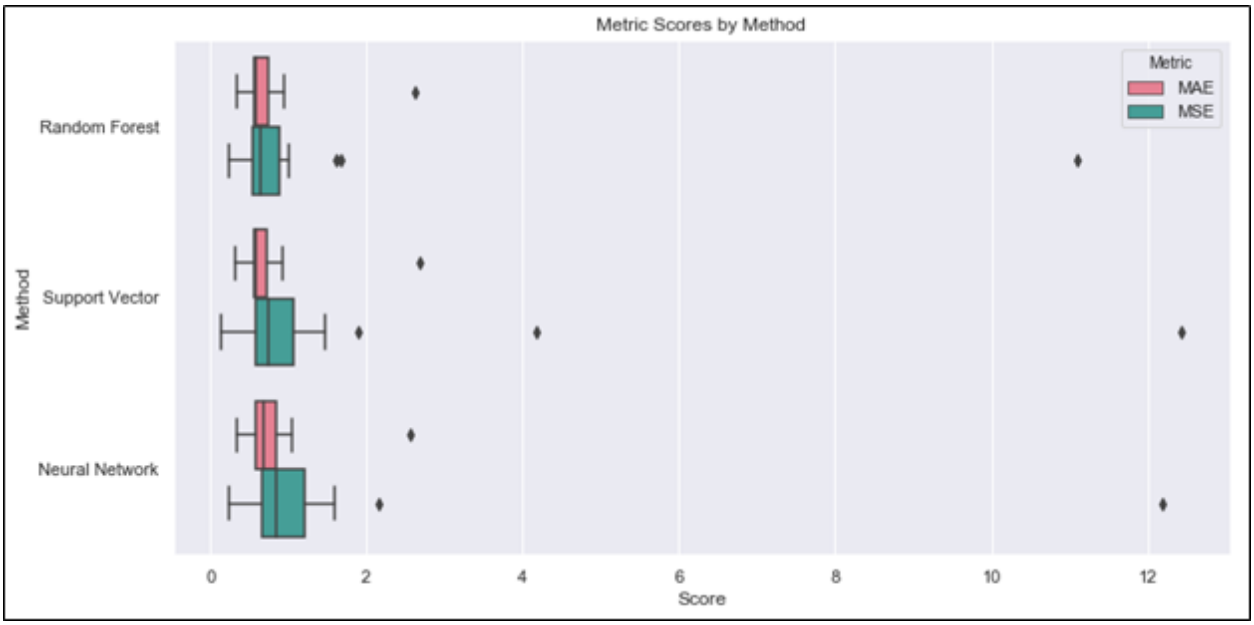


Figure 19: Box-plots comparing the performance metrics for each method used in this report.

Discussion

The models performed relatively well in forecasting the TSE with each method having the most success with the road links 2261, 1882, and 2086 and struggled the most with 457,1402, and 2097. The particular values that the model performed well on tend to be towards the South and West of Elephant and Castle versus the North for the road links the models struggled with. This may be indicative of a spatio-temporal 'edge-effect' towards the north of this area (i.e. heading towards the Strand and City of London), indicating that these are governed by different dynamics not accounted for by the model. As such, to forecast more effectively in this area the models need more information about the spatio-temporal structure outside the current study area (Cheng *et al.* 2012*)*. Then again we would expect to see a spatial-temporal edge effect regardless of how much information we included (see MAUP, Arbia and Petrarca, 2011; MTUP, Cheng & Adepuju, 2014).

We identified two main reasons for the moderate results. Firstly, machine learning is the unable to directly handle spatio-temporal structure (Cheng *et al.*, 2011). As seen in the ESTDA, this autocorrelation is present in our data, although this has not been specifically dealt withj. However, there are uses of machine learning methods that have been extended to account for the limitations of working with spatio-temporal data (such as the integration of convolutional neural networks and LSTM methods (Shi *et al*., 2015; Yu *et al.*, 2017; Han *et al.*, 2019)) which may improve the ability to forecast TSE.

Limitations:

Interpretation of model results varies between the methods. The LSTM and SVR prove to be the more difficult to interpret as they are "black-box model". RF provides a stronger interpretation through the visualisation of decision trees and feature importance. In this study, it is not reasonable to visualise the full random forest, but the feature importance report provides information highlighting the importance of temporal data over spatial data in the prediction process.

A significant advantage that all the methods maintain is the ability to handle and make generalisation about non-linear data. Additionally, all models avoided overfitting by using a five-fold cross-validation

process. That said, RF maintains a comparative implementation advantage. RF can easily handle a combination of non-standardised binary, numeric, and categorical data and requires no data assumptions prior to training. It takes few parameters to create a successful model and the structure of decision tree ensembles avoid overfitting instinctively.

SVR, meanwhile, required the one-hot encoding of temporal features (weekday vs weekend, rush hours) as it can only handle numeric data. SVR benefits from being a stochastic gradient descent model, which can enable scalability for high temporal datasets (Jain *et al*., 2014). This was shown as the HRN SVR forecasted the entire study area effectively with only one model fitted to the entire road network.

As suggested above, both the SVR and LSTM models can be computationally intensive and thus slower to compute. The SVR process is slowed from the incorporation of kernels while LSTM requires a high amount of hyperparameters. RF is quicker in computational cost, but still took some time to construct due to the high number of time lags included in our data.

Future Research:
Moving forward, we identified three main possible forecasting improvements. The first is to account for the non-stationarity in the data. The LSTM method used differenced data to account for this and perhaps a similar approach could improve the RF and SVR results. Second, we could add additional spatial-related information i.e. spatio-temporal weigh (Cheng *et al.,* 2014). We included the current TSE of each road's neighbours, but including the previous time steps for each neighbour may prove fruitful.

Finally, further research should be directed to comparisons of ARIMA or STARIMA models as they inherently include spatial and temporal variables being purposely built for time-series analysis. For data with clear seasonal patterns STARIMA may be most suitable to maximize accuracy for long-term TSE forecasting (Kane *et al.,* 2014).

Conclusion:
Overall, the RF method proved to be the strongest predictive model based on the MSE metric. That said, the results for all methods were moderate and further model refinement or data transformation could improve results. While this study provides an insightful comparison of the RF, SVR, and LSTM methods in forecasting this specific selection of roads, the broader merits of these methods related to traffic forecasting require further analysis.

An extension of this study would include a STARIMA model for comparison. Due to the special nature of the spatio-temporal data, a STARIMA model that explicitly accounts for these significant space-time correlations may prove to be the best model in forecasting travel speed.

References

Ahn, J., Ko, E. and Kim, E. (2015). Predicting Spatiotemporal Traffic Flow Based on Support Vector Regression and Bayesian Classifier. *2015 IEEE Fifth International Conference on Big Data and Cloud Computing*.

Arasan, V. and Dhivya, G. (2009). Measurement of Occupancy of Heterogeneous Traffic Using Simulation Technique. *IFAC Proceedings Volumes*, 42(15), pp.19-24.

Arbia, G. and Petrarca, F. (2011). Effects of MAUP on spatial econometric models. *Letters in Spatial and Resource Sciences*, 4(3), pp.173-185.

Arora, A., 2018. Why Random Forests can't predict trends and how to overcome this problem? Available at: https://medium.com/datadriveninvestor/why-wont-time-series-data-and-random-forests-work-very-well-together-3c9f7b271631.

Asif, M. T., Member, S., Dauwels, J., Member, S., Goh, C. Y., Oran, A., … Jaillet, P. (2014). Spatiotemporal Patterns in Large-Scale Traffic Speed Prediction. *IEEE Transactions on Intelligent Transportation Systems*, *15*(2), 794–804.

Basak, D., Pal, S. and Patranabis, D. (2007). Support vector regression. *Neural Information Pro- cessing-Letters and Reviews*, *11*(10), pp.203-224.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. Available at: https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2017a) How to Use Features in LSTM Networks for Time Series Forecasting [Online]. Available at: https://machinelearningmastery.com/use-features-lstm-networks-time-series-forecasting/ [Accessed 23rd April 2019].

Brownlee, J. (2017b) ADAM optimisation algorithm for deep learning [Online]. Available at: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/ [Accessed 23rd April 2019].

Brownlee, J. (2018) How to Develop LSTM Models for Multi-Step Time Series Forecasting of Household Power Consumption [Online]. Available at: https://machinelearningmastery.com/how-to-develop-lstm-models-for-multi-step-time-series-forecasting-of-household-power-consumption/ [Accessed 23rd April 2019].

Brownlee, J. (2017a) How to Use Features in LSTM Networks for Time Series Forecasting [Online]. Available at: https://machinelearningmastery.com/use-features-lstm-networks-time-series-forecasting/ [Accessed 23rd April 2019].

Bezuglov, A., & Comert, G. (2016). Short-term freeway traffic parameter prediction: Application of grey system theory models. *Expert Systems with Applications*, *62*, 284–292.

Chatzimichali, E. and Bessant, C. (2015). Novel application of heuristic optimisation enables the creation and thorough evaluation of robust support vector machine ensembles for machine learning applications. *Metabolomics*, 12(1).

Cheng, A., Jiang, X., Li, Y., Zhang, C. and Zhu, H. (2017). Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method. *Physica A: Statistical Mechanics and its Applications*, 466, pp.422-434.

Cheng, T., Wang, J., & Li, X. (2011). A Hybrid Framework for Space – Time Modeling of Environmental Data, *43*, 188–210.

Cheng, T., Haworth, J., & Wang, J. (2012). Spatio-temporal autocorrelation of road network data. *Journal of Geographical Systems*, 389–413.

Cheng, T., Wang, J., Haworth, J., Heydecker, B., & Chow, A. (2014). A Dynamic Spatial Weight Matrix and Localized Space-Time Autoregressive Integrated Moving Average for Network Modeling. *Geographical Analysis*, *46*(1), 75–97.

Cheng, T. and Adepeju, M. (2014). Modifiable Temporal Unit Problem (MTUP) and Its Effect on Space-Time Cluster Detection. *PLoS ONE*, 9(6), p.e100465.

Clark, S. and Rey, S. (2017). Temporal dynamics in local vehicle ownership for Great Britain. *Journal of Transport Geography*, 62, pp.30-37.

Dong, B., Cao, C. and Lee, S. (2005). Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, 37(5), pp.545-553.

Han, D., Chen, J., & Sun, J. (2019). A parallel spatiotemporal deep learning network for highway traffic flow forecasting. International Journal of Distributed Sensor Networks, 15(2).

Haworth, J., Shawe-Taylor, J., Cheng, T., Wang, J., 2014. Local online kernel ridge regression for forecasting of urban travel times. Transportation Research Part C: Emerging Technologies 46, 151–178.

Hochreiter, S., & Urgen Schmidhuber, J. (1997). Long Short Term Memory. *Neural Computation*, *9*(8), 1735–1780.

Hong, W.-C., Dong, Y., Zhang, W.Y., Chen, L.-Y., K. Panigrahi, B., 2013. Cyclic electric load forecasting by seasonal SVR with chaotic genetic algorithm. International Journal of Electrical Power & Energy Systems 44, 604–614.

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 20th April 2019.

Jain, R., Smith, K., Culligan, P. and Taylor, J. (2014). Forecasting energy consumption of multi- family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy*, 123, pp.168-178.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*. London: Springer, 102, pp.303-368.

Hong, W., Dong, Y., Zheng, F. and Wei, S. (2011). Hybrid evolutionary algorithms in a SVR traffic flow forecasting model.*Applied Mathematics and Computation*, 217(15), pp.6733-6747.

Kane, M.J., Price, N., Scotch, M., Rabinowitz, P., 2014. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics 15, 276. Available at: https://doi.org/10.1186/1471-2105-15-276.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *3rd International Conference for Learning Representations* (pp. 1–15).

Kirk, M. (2017). *Thoughtful machine learning with Python*. 1st ed. London: 9781491924136, pp. 243-256.

Lam, S.H.M. and Toan, T.D., 2008. *Short-term travel time prediction using support vector regression*(No. 08-0670).

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest 2, 6.

Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies, 54, 187–197.

Mennitt, D., Sherrill, K., Fristrup, K., 2014. A geospatial model of ambient sound pressure levels in the contiguous United States. The Journal of the Acoustical Society of America 135, 2746–2764. Available at: https://doi.org/10.1121/1.4870481

Olah, C. (2015) Understanding LSTM Networks [Online]. Available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ [Accessed 23rd April 2019].

Qu, L., Li, W., Li, W., Ma, D., & Wang, Y. (2019). Daily long-term traffic flow forecasting based on a deep neural network. *Expert Systems with Applications*, *121*, 304–312.

Pavlyuk, D. (2019). Feature selection and extraction in spatiotemporal traffic forecasting: a systematic literature review. *European Transport Research Review*, *11*(1).

Taalab, K., Cheng, T., Zhang, Y., 2018. Mapping landslide susceptibility and types using Random Forest. Big Earth Data 2, 159–178. Available at: https://doi.org/10.1080/20964471.2018.1472392

The Data Science Blog (TDSB) (2016) A Quick Introduction to Neural Networks [Online]. Available at: https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/ [Accessed 23rd April 2019].

Rey, S. (2015). *Python Spatial Analysis Library (PySAL): An Update and Illustration*. In *Geocomputation*. Los Angeles: SAGE Publications Ltd, pp.233-254.

Segal, M.R., 2004. Machine Learning Benchmarks and Random Forest Regression.

Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong & W.-c. Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems, 802-810.

Smola, A. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Comput- ing*, 14(3), pp.199-222.

Su, H., Zhang, L. and Yu, S., 2007, August. Short-term traffic flow prediction based on incremental support vector regression. In *Third International Conference on Natural Computation (ICNC 2007)*(Vol. 1, pp. 640-645). IEEE.

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, *43*, 3–19.

Wu, C.H., Ho, J.M. and Lee, D.T., 2004. Travel-time prediction with support vector regression. *IEEE transactions on intelligent transportation systems*, *5*(4), pp.276-281.

Xiao, N. (2017) Machine Learning. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*. pp.1-9.

Yue, Y., & Yeh, A. G. O. (2008). Spatiotemporal traffic-flow dependency and short-term traffic forecasting. *Environment and Planning B: Planning and Design*, *35*(5), 762–771.
Yu, H., Z. Wu, S. Wang, Y. Wang & X. Ma (2017) Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks. Sensors, 17, 1501.

Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y., & Liu, J. (2017). LSTM network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, *11*(2), 68–75.