

Modelling Urban Link Travel Time with Inductive Loop Detector data using the k-NN method.

Steve Robinson
Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London
Exhibition Road, London SW7 2AZ
United Kingdom
Tel: +44 – 20 7594 6153
Fax: +44 – 20 7594 6102
E-mail: steve.robinson@imperial.ac.uk

John W. Polak
Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London
Exhibition Road, London SW7 2AZ
United Kingdom
Tel: +44 – 20 7594 6089
Fax: +44 – 20 7594 6102
E-mail: j.polak@imperial.ac.uk

Text = 5742 words
Tables and Figures = 7, i.e. 1750 words
Total Words = 7492 words

Submission date = 31st July 2004
Revised date = 2nd November 2004

Submitted for presentation at the 84th Transportation Research Board Annual Meeting,
January 9 - 13, 2005, Washington, D.C. and publication in the Transportation Research
Record

ABSTRACT

The need to measure urban link travel time (ULTT) is becoming increasingly important for the purposes both of network management and traveller information provision. This paper proposes the use of the k nearest neighbors (k-NN) technique to estimate ULTT using single loop inductive loop detector (ILD) data. This paper explores the sensitivity of travel time estimates to various k-NN design parameters. It finds that the k-NN method is not particularly sensitive to the distance metric, although care must be taken in selecting the right combination of local estimation method (LEM) and value of k . A robust LEM should be used. The optimised k-NN model is found to provide more accurate estimates than other ULTT methods. A potential application of this approach could be to aggregate GPS probe vehicle ULTT records from different times but the same underlying travel time distribution together, to obtain a more accurate estimate of ULTT.

Keywords

ILD, inductive loop detector, Automatic Number Plate Recognition, ANPR, GPS, arterial link, urban link, travel time, k-NN, k nearest neighbors, data fusion.

1 INTRODUCTION

The need to measure urban link travel time (ULTT) is becoming increasingly important for the purposes both of network management and traveller information provision. Although there has already been much research and success in estimating link travel times on highways (1), there has been less success in the urban environment, due to reasons such as; greater heterogeneity of the vehicle population, greater prevalence of devices such as traffic lights, and more complex network topologies. This paper proposes the use of the k Nearest Neighbor (k-NN) method to model ULTT using single loop inductive loop detectors (ILDs). Section 2 provides a literature review on measuring ULTT and modelling ULTT using ILD data. Section 3 introduces the k-NN method, providing a theoretical basis for its use, and highlighting its key design parameters. The k-NN method is optimised for use as a ULTT model in section 4, and in section 5 its performance is compared to existing ULTT models. Finally, section 6 then proposes an application of the k-NN method allowing for high-quality low-quantity travel time data from GPS probe vehicles to be merged with low-quality high-quantity data from inductive loop detectors (ILDs) to enable better estimations of ULTT.

2 MODELLING ULTT USING SINGLE LOOP ILD DATA

2.1 Problems with measuring ULTT directly

Various technologies exist to allow the direct measurement of the travel time of a vehicle over an urban link. Such technologies include; the matching of Automatic Number Plate Recognition (ANPR) data, moving vehicle observers (MVO), and GPS probe vehicles. According to Robinson and Polak (2) there are three main criteria for any technology measuring ULTT.

- To be able to measure with known precision and accuracy the speed of individual vehicles.
- To be able to identify whether a vehicle has or has not travelled ‘in a reasonable fashion’ between the start and end of the link – i.e. the driver did not stop en-route to fill up with petrol, pick up a passenger etc. A vehicle travelling in a reasonable fashion is termed a valid vehicle.
- To be able to capture a representative sample of the total population of valid vehicles.

The last criterion that is often overlooked is whether a representative sample of the total population of valid vehicles is obtained. ANPR systems do tend to capture a large proportion of the traffic stream. Research has suggested a recognition rate between 50-90% of all vehicles in one study (3) and 86% in another (4). Probe vehicle methods tend to capture a far smaller proportion of the vehicle population. In addition probe vehicle data may also not be representative of the vehicle population whose travel time is desired. For example a study by Gühnemann et al (5) used taxis as sources of travel time data; yet taxis sometimes have access to priority lanes and frequently stop mid-link to pick up customers and are thus atypical in their behaviour of the general population of vehicles using a link.

It is important to capture a representative sample of vehicles since ULTT varies significantly in the short term. For example even at night-time on a 700-metre link in central London containing several sets of traffic lights, the travel time was seen to vary between 50 and 140 seconds within the same 15 minute period (6). This short-term variability is caused by various factors such as traffic lights, stopping buses, etc which delay some vehicles but not

others. Due to this variability the ULTT can be described by a random variable with a distribution. The travel time record from each probe vehicle or matched ANPR record is simply one realisation of this distribution. It is thus necessary to obtain a sufficient number of vehicle travel time records in order to characterise the link travel time distribution. Srinivasan and Jovanis (7) showed that the number of vehicle travel time records needed in any time period increased as the underlying variability of travel time increased.

Although it is not possible to measure all vehicle travel times, it is possible to use ILDs to measure other traffic characteristics of the link continuously. These other traffic characteristics can then be used as input in a model to estimate ULTT. This approach is studied in the next section. Alternatively, these other traffic characteristics could be used to aggregate vehicle travel time records from different time periods, but the same underlying travel distribution, to enable a more accurate estimate of ULTT. This approach will be proposed by use of the k-NN method in section 3.

2.2 Modelling ULTT using single loop ILD data

This section discusses the current literature in measuring and modelling urban link travel time (ULTT) using single loop inductive loop detectors (ILDs).

Single loop ILDs are wires, commonly of size 2m x 2m, laid several centimetres below the surface of the road. They detect the presence of metallic objects located above them by measuring a change of inductance in the loop. Their primary data output is the flow (the number of vehicles passing over the detector per unit time) and occupancy (the proportion of time that a vehicle is located directly above the detector). Their primary use in transport has been to measure demand on a link, allowing the operation of traffic lights to be optimised. Research has been undertaken trying to relate ULTT with data from single loop ILDs. This research can be divided into 4 main methods; spot-speed, regression, artificial neural networks (ANNs), and signature matching.

On highways spot-speed methods are prevalent and much work has been undertaken with this method (8, 9, 10, 11). This method relies on the assumption that the space-mean speed (inversely proportional to link travel time) is a simple function of the spot-speed measured by the single loop ILD. Where little variation between vehicles and between locations exists, this method works well. However, in urban environments the above two conditions tend not to be achieved (12).

The second class of ULTT models are based on regression, using variables such as flow, occupancy, signal settings, and road characteristics as the independent variables. Examples of such models include that by Wardrop (13), Gault and Taylor (14), Sisiopiku and Rouphail (15), and Zhang and He (16). The latter model was actually a hybrid model recognising the fact that links may need to be modelled differently depending on whether there is light or heavy demand on the link.

A third category of ULTT models are based on Artificial Neural Networks (ANNs) (17, 18, 19, 20). A common criticism of ANNs is that they lack transparency and thus make it difficult to gain useful insights into the factors giving rise to variability in ULTT. In addition, the literature suggests that more research is needed to optimise the architecture of the ANN for use in estimating ULTT.

Another category of link travel time model using ILD data involves signature matching. In these methods an attempt is made to match the signal at the downstream detector with the signal recorded at the upstream detector. Systems have been proposed that do this with individual vehicles (8) or platoons of vehicles (21, 22). However, many ILDs do not output an analogue signal so it is not possible to match individual vehicles. Moreover the latter method of platoon matching depends on the microstructure of the underlying traffic stream (i.e. the patterns of platoons) remaining intact. However, in urban areas there are many factors (i.e. traffic signals, stopping buses) which substantially alter the microstructure, diminishing the usefulness of this method in estimating ULTT.

It is apparent from the literature that there is much scope for the development of new ULTT models. A novel approach of estimating ULTT based on the k-NN method is thus proposed in the next section.

3 K-NEAREST NEIGHBOR METHOD

3.1 Previous Application Of The k-NN Method In Traffic Characterisation

This section proposes the use of the k-NN method as a model to estimate ULTT. The k-NN method has already been used in traffic characterisation, notably in the estimation of traffic flow. A short review of its use is given in this section.

The first explicit use of the k-NN method was by Davis and Nihan (23) to forecast flow and occupancy at a time interval t , given various flow and occupancy readings at the time interval $t-1$. They found that the k-NN method did not outperform traditional time-series analysis using the Box-Jenkins approach.

Smith and Demetsky (24) used the k-NN method to forecast traffic flow. They compared the results of the k-NN model with those using historical averages, ARIMA models, and a back-propagation neural network. They found that the nearest neighbor approach gave the smallest error of all the methods and also gave an error distribution which tended not to be skewed. They also found that it was the most successful when transferred to a different site. However in later research, Smith et al (25) found the k-NN unable to outperform a SARIMA model when undertaking short-term forecasting of flow.

Clarke (26) also had success with applying the k-NN method to forecast traffic states from measures of flow, speed, and occupancy. He used an adaptation of the city-block method but failed to consider other distance metrics.

There are three instances in the literature where the k-NN method has been applied to estimate link travel time on highways. Handley et al (27) used the k-NN method to estimate travel time on freeways in the San Diego area of the USA. Their research indicated that on freeways the most important input variables were flow and occupancy rather than other features such as time-of-day. However, the full details of the research are unknown, since much of this work was proprietary.

You and Kim (28) used a k-NN method to estimate spot speeds on a highway in Korea and travel time in an urban area in Seoul. They used travel times from probe vehicles obtained over the previous 15 to 60 minutes as input to the model rather than ILD data. They report a MAPE in the region of 8 to 10%. However, they do not compare this accuracy with that achieved by other methods which makes these figures rather meaningless.

More recently Bajwa et al (29, 30) have used the k-NN method to estimate travel time on expressways in Tokyo using data from ultrasonic sensors. Of the literature identified in this section, this is the only paper that explicitly attempts to optimise the k-NN parameters – this is attempted using a genetic algorithm.

It has been shown that although research into the use of the k-NN method in traffic characterisation has been undertaken, there has tended to be little effort in optimising the key parameters, and other than work by Oswald et al (31), no guidelines towards its implementation. In addition no reference to the underlying mathematical theory has been made. The next sections address these two issues. In addition a fair comparison between the k-NN method and existing ULTT models, using identical datasets, will be undertaken.

3.2 Basic Concepts

The basic idea behind the k-Nearest Neighbor (k-NN) approach is to match the current input variables with historical observations that have similar input variables. The set of input variables is collectively called the *feature vector*. The current output is then defined to be some function of the known outputs of those records with similar *feature vectors*. This is explained with the aid of the example below.

[FIGURE 1]

FIGURE 1 shows a scatterplot of the independent variable, x , against a dependent variable y . In this example the feature vector is comprised of only one attribute; i.e. x . Suppose it is desired to use the k-NN method to estimate the value of y when $x = x_j$. If k is chosen to be 5, then the 5 historical observations with a value of x closest to x_j are identified. These observations are shown by the shaded points in the diagram. The y values of these 5 closest observations are then used to estimate the value of y when $x = x_j$. It is common to use the arithmetic mean of these five y values.

3.3 Key Assumptions Of The k-NN Method

The k-NN method relies on several assumptions.

- *Similar a posteriori Distributions* - The k-NN method relies on the assumption ‘that observations which are close together in feature-space are likely to belong to the same class or to have the same *a posteriori* distributions of their respective classes’ (32). It is thus important that the feature vector contains sufficient attributes to allow observations to be classified into their correct class of distribution.
- *Large Historical Database* - It can be shown that the distance from the input feature-vector to the k-nearest neighbors will tend to zero as the number of historical observations, N , tends to infinity (32, 33). In addition as long as k/N tends to zero as both k and N tend to infinity, the risk (i.e. the probability of misclassification) of the k-NN classification rule approaches the minimal possible (commonly referred to as the Bayes risk, occurring when complete knowledge of the underlying system is available) (34). It is thus desirable to have a historical database of a sufficient size (32). However increasing the size of the database does increase the computation time of the method, although techniques do exist to reduce this problem (35).
- *Random Distribution Of The k-Nearest Neighbors* - A further assumption made in the k-NN method is that the k nearest neighbor historical observations are distributed randomly in all n dimensions around the input feature vector. However, if the input feature-vector is located at the boundary of the feature-space of existing observations, then this assumption will not hold since most selected neighbors will tend to be on one side of the input feature vector. Thus input feature vectors located at the boundary will be subjected to larger bias error. This is explored by Robinson (36) where a way of quantifying the likely biasedness of the selected k-nearest neighbors is given. Techniques exist for overcoming these boundary problems; for example, instead of calculating the arithmetic mean of travel time for the k-nearest neighbors, these k-nearest neighbors can be used to define a suitable regression line through the local area (37).

3.4 Theoretical Basis Of The k-NN Method

Although the k-NN method has been applied in the existing traffic characterisation literature, the underlying theoretical properties of the k-NN method have not been widely discussed. These properties are important in the context of devising optimal strategies for the implementation of k-NN methods.

The first reference to the k-NN method was by Fix and Hodges (38, 39) in the 1950s. However due to limitations of computation power it was not until the late 1960's and early 1970's that interest in the method grew. Cover and Hart (33) showed that given an infinite sample set, then when using the 1-NN model the probability of error of the k-NN classification rule will not exceed twice the Bayes risk (which is the best performance possible). An important contribution to the study of the k-NN method has been made by Fukunaga (40, 41). He identified that the optimum value of k to use was a function of; sample size, inherent dimensionality of the feature space, and the underlying distribution of the dependent variable against the independent variables (40). He formulated an expression for the Mean Square Error (MSE) of the density (i.e. the difference between the actual and estimated density) of the k-NN method as a function of the number of nearest neighbors, k (41).

$$MSE \cong p^2 \left[\frac{1}{k} + \frac{1}{4} \alpha^2 (cp)^{-4/n} \left(\frac{k}{N} \right)^{4/n} \right] \quad (1)$$

k = number of nearest neighbors

α = represents the rate of change of the underlying distribution and is thus a measure of the non-linearity of the system. i.e. in linear systems, $\alpha = 0$.

c = constant relating the volume that encapsulates the k nearest neighbors to the distance of the k-th nearest neighbor. The value of c is dependent on the distance metric chosen.

p = Mixture density function

n = number of dimensions in the feature space.

N = number of historical observations in the database.

The important thing to note is that there is a trade-off between the measure of imprecision (variance error) and the measure of bias (bias error). Variance error, represented by the first term in equation (1), takes into account that travel time is a random variable, and several realisations of it are needed to reduce the effects of the random error component. Bias error, represented by the second term in equation (1), is caused by the non-linearity of the underlying system, and represents the fact that a record on one side of the input feature vector may influence the estimate more than a record at the same distance but on the other side of the input feature vector. It can be seen that as k increases, the variance error decreases, but the bias error increases. The optimal value of k, k^* , is given by differentiating equation (1) with respect to the MSE.

$$k^* = \left[\frac{n(cp)^{4/n}}{\alpha^2} \right]^{\frac{n}{n+4}} \cdot N^{\frac{4}{n+4}} \quad (2)$$

For a comprehensive introduction to the k-NN method the reader is referred to Fukunaga (40), and Dasarthy (35).

3.5 Key Parameters Of The k-NN Method

There are four key design parameters that need to be optimised when using the k-NN method.

The first is the choice of which attributes to include in the feature vector. If too few attributes are used, then the model will not have enough features with which to differentiate

the input observation from the historical observations, and inappropriate historical observations will be chosen. That is historical observations coming from a different underlying distribution of travel time will be chosen. However, if too many attributes are included in the feature vector, then the model will be prone to the *curse of dimensionality*. This describes the situation when irrelevant attributes in the feature vector dominate the distance metric, reducing the influence of the relevant attributes on the distance metric.

Various methods of determining which attributes to include have been proposed. These feature selection methods can be divided into two categories, filter and wrapper methods (42). A filter method chooses the attributes independently of the final model, whereas a wrapper model uses the final model to evaluate the performance of the current feature vector. The latter method is prone to overfitting. A good summary of feature selection methods is presented by Kittler (43) and Kudo and Sklansky (44).

The second key parameter is the distance metric. This metric defines the ‘closeness’ between two points in multivariate space. A selection of possible measures of distance are given in TABLE 1 along with the appropriate equation for calculating the distance between two n-variate observations, x_a and x_b (45). Qi and Smith (46) also formulated an alternative approach of measuring the distance between two points. Theoretical work undertaken by Fukunaga and Hostetler (40) suggests that the optimal distance metric is the Mahalanobis metric.

[TABLE 1]

The third key parameter is the number of nearest neighbors to use in determining the output value. The trade-off between bias and variance error has already been noted in equation (1).

The estimated travel time for the current feature vector will be calculated using a function of the travel times of the k-nearest neighbor records. This function is called the local estimation method, and is the final key design parameter of the k-NN method. Most applications of the k-NN, estimate the output value as the arithmetic mean or median of the output values of each of the k-nearest neighbors identified. Another approach is to fit a curve through these k points using regression analysis (37). The latter approach is useful for reducing the effects of boundary bias. A robust version of regression analysis called LOWESS (LOcally WEighted Scatter plot Smoothing) has also been proposed (47).

Before using the k-NN method to estimate ULTT, it will be necessary to determine the optimal levels of these parameters, and identify which parameters the model is sensitive to. This will be undertaken later in this paper.

4 OPTIMISATION OF THE KEY DESIGN PARAMETERS OF THE K-NN MODEL

4.1 Methodology

This section presents an experiment to determine the optimal values of the k-NN model to be used in estimating ULTT. Four key design parameters in the k-NN method were identified in section 3.5. These were:

- Attributes to include in the feature vector
- Distance Metric, (DM)
- Value of k
- Local Estimation Method (LEM)

Theory suggests that these parameters may not be independent of one another. For example the value of k and the LEM are in some instances related (37). Thus it is desirable to undertake full-factorial analysis where all combinations of parameters are tested against one another to determine the sensitivity of the parameters and the optimal levels of the parameters. This exhaustive approach is not well suited to the k -NN method since it is itself computationally intensive. Thus before the full factorial analysis was undertaken it was necessary to reduce the number of distinct levels associated with each parameter, and hence the number of total combinations, to be considered. Therefore the first analysis considered each of these parameters independently of one another, enabling the identification of those levels that are obviously much poorer than the rest. To further reduce the computational demands of the full factorial analysis, it was decided to fix the feature vector to include input from only 3 ILDs on the link.

4.2 Data For Experiments

Real world data was provided by Transport for London (TfL), for a south-bound link around Russell Square in central London, from 37 separate days in March and April 2003 – see FIGURE 2. Travel time data for each individual vehicle was obtained from Automatic Number Plate Recognition (ANPR) cameras used to monitor London’s congestion charging scheme, taking full account of data protection principles. The cameras were sited at either end of the link. Erroneous travel time records were filtered out using the overtaking rule (48). The overtaking rule has been shown to outperform other filtering mechanisms when used to filter London’s ANPR data (49).

For each individual vehicle link travel time record the 15-minute flow and occupancy measured at three detectors on the Russell Square link were also recorded. These 6 values constituted the feature vector. Flow and occupancy data were ‘verified’ using the Daily Statistic Algorithm (DSA) (Chen et al, 2003). Records whose ILD data failed the DSA test were removed from the data set. The resulting data set contained 24,718 records.

One of the 37 days was chosen at random to provide the validation data. From this single day, 200 travel time records were chosen at random to act as the validation data set. To ensure independence between the training and validation data sets, all records from the same day as the validation data were removed from the remaining data set. A randomly selected subset of 15,000 records was chosen from this remaining data set to create the training data set. The k -NN model was then used to estimate the travel times of the validation data set. Comparing the estimated and actual travel times, it was then possible to calculate the Mean Absolute Percentage Error (MAPE), and the Root Mean Square Error, (RMSE). Due to the random nature of selecting both the validation and historical data sets, this whole process was repeated 25 times for each setting using different validation and training sets each time.

4.3 Parameters Considered Independently

As a first approach to identifying the optimal settings of the k -NN model, a series of univariate analyses were undertaken to gain a preliminary understanding of each parameter. This analysis suggested a strong correlation between the performance of the ULTT model and the combination of the value of k and the local estimation method (LEM) used. Other univariate analysis suggested that the Qi and Smith (46) distance metric (DM) was less accurate and more computationally expensive than other DMs given in TABLE 1. Thus it was possible to eliminate this DM from further analysis.

A final analysis indicated that the optimum value of k to use increased as the size of the historical database used increased. This can be seen in FIGURE 3.

[FIGURE 3]

4.4 Parameters Considered Simultaneously

A full factorial analysis was undertaken to determine the optimal setting of the k-NN model, and the sensitivity of the estimate of ULTT to these parameters. Following the work of section 4.3, the Qi-Smith DMs were omitted from the study. The size of the historical database was fixed at 15,000. The three parameters and the various levels of these parameters are given below:

- **Distance metric, DM** – 4 levels: SE-VAR, SE-STD, Mahalanobis, UnitMap (refer to TABLE 1).
- **Local Estimation Measure, LEM** – 4 levels: mean, median, regression, Lowess.
- **Value of k** – 14 levels: 20, 50, 100, 200, 300, 400, 500, 750, 1000, 1500, 2000, 3000, 4000, 5000.

Thus there were a total of 224 possible settings. Each setting was run 25 times to take account of the random selection of the historical database. In each run 200 validation records were used to determine the accuracy of the prediction resulting from each set of parameters.

The optimal MAPE setting was found to use 400 nearest neighbors with the UnitMap distance metric and the median as the LEM. The optimal RMSE setting was found to use 1500 nearest neighbors with the UnitMap distance metric and the regression as the LEM. It was noticeable that those settings with a good MAPE performance tended not to have such a good RMSE performance, and vice-versa. Ideally the optimal settings should perform well when measured by both RMSE and by MAPE. They should also be computationally quick. To determine the overall optimal settings, an ad-hoc performance index, PI, was defined in equation (3).

$$PI = \text{Rank MAPE} + \text{Rank RMSE} + \frac{\text{rank Time}}{5} \quad (3)$$

Using this performance index the best settings were seen to use approximately 2000 nearest neighbors with the SE-VAR distance metric and the lowess as the LEM. Further refinement using the Fibonacci line search technique suggested the use of 2160 nearest neighbors.

In order to identify which levels of the ULTT model the MAPE was sensitive to, a linear regression model was then constructed in which every level of each given design parameter was characterised by a regression parameter. This regression parameter measured the effect of each level relative to a nominal base. Several interaction terms were introduced to investigate the possibility of interactions between the various design parameters. In addition various terms were dropped from the model to ensure that there was no collinearity between variables. In particular adding interactive terms involving the distance metric and values of k improved the model performance. A total of 2400 records were used in this model. The results are shown in TABLE 2.

[TABLE 2]

From the results the following conclusions can be drawn about the k-NN ULTT model:

- The k-NN is not particularly sensitive to the DM chosen, with only the SE-STD distance metric significantly worse (β_{DM1}).
- The robust LEMs, i.e. the median and the Lowess method (β_{LEM2} , β_{LEM4}) significantly outperform the non-robust methods (β_{LEM3}).
- There are significant interaction terms involving the LEM and value of k. The Lowess and Regression LEM perform significantly better using high values of k ($\beta_{LEM3, k1000}$; $\beta_{LEM3, k2000}$) and significantly worse with low values of k. ($\beta_{LEM3, k100}$;

$\beta_{LEM3, k200}$; $\beta_{LEM2, k200}$). Likewise the median performs well with low values of k ($\beta_{LEM2, k200}$) and worse with high values of k ($\beta_{LEM2, k5000}$).

Thus from this study it was concluded that the optimum settings were as follows:

- Distance metric: SE-VAR
- Local estimation method: Lowess
- Value of k : 2160

In the next section these settings were used when comparing the k -NN model against other ULTT models.

5 COMPARISON OF THE OVERALL ACCURACY OF THE K-NN METHOD AGAINST OTHER ULTT METHODS

One criticism of the literature on ULTT models is the failure to compare the performance of different models. The overall accuracy of the k -NN method was thus tested against various other models. The optimal settings derived in section 4.4 were used (k -NN lowess). For all ILD based methods the same data set and methodology was used as described in section 4.2. In addition three ULTT models which used solely time information and two naïve ULTT estimators were included:

- **Gault and Taylor** – adapted from the model used by Gault and Taylor (14).
- **Basic Regression** – uses the occupancy, and ratio of occupancy to speed, from each of the 3 ILDs.
- **Artificial Neural Network** - The ANN was a back-propagation network with 12 log-sigmoid neurons in the hidden layer. Levenberg-Marquardt training was used (45). It was given a 120 second training time on a Pentium 4, 2.8GHz machine.
- **K-NN median** – k -NN method using the median LEM, SE-VAR DM, and 225 nearest neighbours.
- **Naïve Mean** – this estimate is the mean travel time of all the training records.
- **Naïve Median** – this estimate is the median travel time of all the training records.
- **Day-of-Week** – All the historical records belonging to the same day-of-week as the validation data are first selected. The estimate is the median of these selected travel times.
- **Period 15** – All the historical records belonging to the same 15 minute period of the day as the validation data are first selected. The estimate is the median of these selected travel times.
- **Day-of-Week & Period 15** - All the historical records belonging to the same day-of-week and 15 minute period of the day as the validation data are first selected. The estimate is the median of these selected travel times.

[TABLE 3]

The results in TABLE 3 show that when the MAPE is considered, the k -NN method clearly outperforms all other methods (column 1). Although the MAPE for all models look quite high, it must be remembered that due to the short-term variability of travel times (some vehicles get stopped by the traffic lights and others do not), there is an inherent limit to the performance of any ULTT model (50). Since the ANN and regression methods aim to minimise the total squared error, it is also useful to consider the RMSE (column 2). Using this metric, the k -NN method is still the best ULTT model.

The percentage of times that the k-NN Lowess gave a more accurate estimate than the other ULTT models was calculated. This is shown in column 3.

A t-test was then undertaken to see whether the distribution of the travel time estimates differed significantly from the actual TT distribution. All of the ULTT models were found to be significantly different from the actual travel time distribution (column 4). This difference is borne out by the high negative correlation between the actual travel time and the residuals (column 5). That is all the models tend to underestimate ULTT at high actual travel time, and overestimate ULTT at low actual travel time. This effect was seen to be reduced with the k-NN lowess ULTT model.

It was hypothesised that due to its inherent nature of using only local data, the k-NN model may perform well at the extremes of the operation of the model – i.e. at low and high travel times. FIGURE 4 shows the percentage of times that the estimate given by the k-NN lowess method is better than the other ULTT model, but disaggregated by the actual travel time. This figure clearly shows the k-NN method to be superior to other methods at low and very high actual travel times. At other times the k-NN does no worse than other methods, except for the naïve average, which by definition performs very well around the average actual travel time. The figure also shows the k-NN median to be better than the k-NN lowess model at low travel times, but worse at high travel times. This may explain why in TABLE 3 the k-NN median has a lower overall MAPE but a higher overall RMSE than the k-NN lowess model.

[FIGURE 4]

6 POTENTIAL APPLICATION OF K-NN METHOD TO GPS DATA

It is realised that ULTT can be measured directly on a link with ANPR cameras. The primary purpose of using such data was to demonstrate that the performance of the k-NN method was better than other ULTT methods. In effect the k-NN allowed travel time records from different times but the same underlying travel time distribution to be aggregated together to allow for an accurate estimate of the ULTT given the current flows and occupancies measured on the link. It could be possible to use travel time records from GPS probe vehicles rather than travel time records from ANPR systems. This would help overcome one of the limitations of measuring ULTT using data from GPS probe vehicles, namely an insufficient number of probe vehicles. Travel time records from GPS probe vehicles would not be used explicitly to estimate the travel time in the period but would be added, along with the appropriate flow and occupancy readings, to the historical database. Thus if it were desired to estimate ULTT the only input to the k-NN model would need to be the current flow and occupancy measurements at the time of interest. Unlike the model proposed by Xie et al (51), probe vehicle records from the current period would not be used directly to estimate the ULTT of the current period.

The research in this paper suggests that this approach would be successful. Certainly the results in TABLE 3 suggest that using travel time records with similar ILD data rather than from a similar day-of-week or time-of-day would result in more accurate estimates of ULTT.

7 CONCLUSIONS

This paper has researched the use of the k-NN method to model ULTT using high quality low quantity travel time data, and low quality, high quantity ILD data. The k-NN method has a

solid theoretical foundation, although the practitioner of the method must ensure that the key parameters are set properly. Research into these key parameters concluded the following:

- The k-NN is not particularly sensitive to the distance metric chosen.
- A robust local estimation method such as the median or Lowess should be used.
- The optimal value of k depends on the size of the historical database.

The performance of the k-NN method was then compared with various ULTT models. The k-NN method was seen to have the best accuracy measured by both MAPE and RMSE. In particular the model performed well at low and very high levels of actual travel time.

This paper has shown that the k-NN method provides an attractive framework by which travel time records from different times but with similar underlying traffic conditions can be aggregated together to provide accurate estimates of travel time. An application of this approach would be to aggregate high quality low quantity GPS data, and low quality, high quantity ILD data.

ACKNOWLEDGEMENTS

The authors would like to thank Charles Buckingham and Dr John Tough of Transport for London, for providing the ANPR and ILD data respectively. The work reported in this paper was partially supported by the UK Engineering and Physical Sciences Research Council (EPSRC).

REFERENCES

- 1 van Lint, H. (2004). "Reliable travel time prediction for freeways." Trail research school, Delft, Netherlands. ISBN: 90-5584-054-8.
- 2 Robinson, S. and Polak, J. W. (2004). "Modelling urban link travel-time using data from inductive loop detectors." WCTR 2004 - 10th world conference on transport research, 4th - 8th July 2004, Istanbul, CD-ROM of un-revised papers.
- 3 Van der Zijpp, N. (1997). "Dynamic OD-Matrix estimation from Traffic Counts and automated vehicle identification data." Transportation Research Record, 1607, 87-94.
- 4 Wiggins, A. E. (1999). "Helsinki Journey Time Monitoring System." IEE Seminar on CCTV and Road Surveillance (Ref. No. 1999/126).
- 5 Gühnemann, A., Schäfer, R.-P., Thiessenhusen Kai-Uwe, and Wagner, P. (2004). "New approaches to traffic monitoring and management by floating car data." WCTR 2004 - 10th world conference on transport research, 4th - 8th July 2004, Istanbul, CD-ROM of un-revised papers.
- 6 Papagianni, S. (2003). "Using Automatic Number Recognition Data to estimate vehicle speeds in central London." MSc Dissertation, Centre for Transport Studies, Imperial College London.
- 7 Srinivasan, K. and Jovanis, P. (1996). "Determination of number of probe vehicles required for reliable travel time measurement in urban network." Transportation Research Record, 1537, 15-22.
- 8 Turner, S., Eisele, W. L., Benz, R. J., and Holdener, D. J. (1998). "Travel Time Data Collection Handbook." Federal Highway Administration, FHWA-PL-98-035; Research Rept 07470-1F; Final Report.
- 9 Ishimaru, J. M. and Hallenbeck, M. E. (1999). "Flow Evaluation Design - Technical Report." Washington State Transportation Center.

- 10 Dailey, D. J. (1999). "A statistical algorithm for estimating speed from single loop volume and occupancy measurements." *Transportation Research B*,(33), 313-322.
- 11 Coifman, B. (2001). "Improved velocity estimation using single loop detectors." *Transportation Research A*, 35, 863-880.
- 12 Barbosa, H., Tight, M., and May, A. D. (2000). "A model of speed profiles for traffic calmed roads." *Transportation Research A*, 34, 103-123.
- 13 Wardrop, J. G. (1968). "Journey Speed and Flow in Central Urban Areas." *Traffic Engineering + Control*, 9, 528-539.
- 14 Gault, H. E. and Taylor, I. G. (1981). "The use of the output from vehicle detectors to assess delay in computer-controlled area traffic control systems." University of Newcastle, Transport Operations Research Group, Research Report no.37.
- 15 Sisiopiku, V. and Roupail, N. (1994). "Travel Time Estimation from Loop Detector Data for Advanced Traveler Information Systems Applications." (A Technical Report in Support of the ADVANCE Project, Urban Transportation Center, University of Illinois).
- 16 Zhang, M. and He, J. C. (1998). "Estimating Arterial Travel Time Using Loop Data." Public Policy Center, University of Iowa.
- 17 Palacharla, P. and Nelson, P. (1995). "On-line travel time estimation using fuzzy neural network." *Proceedings 2nd World Congress on Intelligent Transport Systems*. In VERTIS (Ed). Yokohama. Vol.1, 112-116.
- 18 Anderson, J. and Bell, M. G. H. (1997). "Travel Time Estimation in Urban Road Networks." *IEEE Conference on Intelligent Transportation System, 1997. ITSC 97*.
- 19 Park, D. J., Rilett, L. R., and Han, G. H. (1999). "Spectral basis neural networks for real-time travel time forecasting." *Journal of Transportation Engineering*, 125(6), 515-523.
- 20 Cherrett, T., Bell, H. A., and McDonald, M. (2001). "Estimating Vehicle Speed using single Inductive Loop Detectors." *Proc of the Institution of Civil Engineers, Transport*(147), 23-32.
- 21 Dailey, D. J. (1993). "Travel Time Estimation using cross correlation techniques." *Transportation Research B*, 27b(2), 97-107.
- 22 Petty, K. F., Bickel, P., Ostland, M., Rice, J., Schoenberg, F., and Jiang, J. (1998). "Accurate Estimation of travel times from single loop detectors." *Transportation Research A*, Vol 32, no.1, pp.1-17.
- 23 Davis, G. A. and Nihan, N. L. (1991). "Nonparametric regression and short-term freeway traffic forecasting." *Journal of Transportation Engineering*, 117(2), 178-188.
- 24 Smith, B. L. and Demetsky, M. J. (1997). "Traffic Flow Forecasting: Comparison of Modeling Approaches." *Journal of Transportation Engineering*, 123(4), 261-266.
- 25 Smith, B. L., Williams, B. M., and Oswald, K. (2002). "Comparison of parametric and nonparametric models for traffic flow forecasting." *Transportation Research C*, 10, 303-321.
- 26 Clark, S. (2003). "Traffic Prediction using multivariate nonparametric regression." *Journal of Transportation Engineering*, 129(2), 161-168.
- 27 Handley, S., Langley, P., and Rauscher, F. A. (1998). "Learning to predict the duration of an automobile trip." *Proceedings of the International Conference on Knowledge Discovery and Data Mining (4th : 1998 : New York, N.Y.)*, 219-223.
- 28 You, j. and Kim, T. J. (2000). "Development and Evaluation of a hybrid travel time estimation model." *Transportation Research C*,(8), 231-256.
- 29 Bajwa, S. u. I., Chung, E., and Kuwahara, M. (2003). "A travel time prediction method based on pattern matching technique." 21st ARRB and 11th REAAA Conference, Cairns, Australia, 2003.
- 30 Bajwa, S. u. I., Chung, E., and Kuwahara, M. (2003). "Sensitivity Analysis of Short Term Travel Time Prediction Model's Parameters." *World Congress on Intelligent Transport Systems (10th : 2003 : Madrid, Spain)*.
- 31 Oswald, K., Scherer, W. T., and Smith, B. L. (2000). "Traffic flow forecasting using approximate nearest neighbor nonparametric regression."
- 32 Devijver, P. A. and Kittler, J. (1982). "Pattern Recognition: A statistical approach." Prentice Hall International, ISBN 0-13-654236-0.
- 33 Cover, T. M. and Hart, P. E. (1967). "Nearest Neighbor Patter Classification." *IEEE Transactions on Information Theory*, 13(1), 21-27.

- 34 Cover, T. M. (1968). "Estimation by the Nearest Neighbor rule." *IEEE Transactions on Information Theory*, 14(1), 50-55.
- 35 Dasarathy, B. V. (1991). "Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques." *IEEE Computer Society Press*, ISBN 0-8186-8930-7.
- 36 Robinson, S. "Estimating Travel Time using the k-Nearest Neighbor Method." Working Paper.
- 37 Simonoff, J. S. (1996). "Smoothing Methods in Statistics." *Springer-Verlag*, New York, ISBN 0-387-94716-7.
- 38 Fix, E. and Hodges, J. L. (1951). "Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties." Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolph Field, Texas, 261-279.
- 39 Fix, E. and Hodges, J. L. (1952). "Discriminatory Analysis: Nonparametric Discrimination: Small Sample Performance." Project 21-49-004, Report Number 11, USAF School of Aviation Medicine, Randolph Field, Texas, 280-322.
- 40 Fukunaga, K. and Hostetler, L. (1973). "Optimization of k nearest neighbor density estimates." *IEEE Transactions on Information Theory*, 19(3), 320-326.
- 41 Fukunaga, K. (1990). "Introduction to Statistical Pattern Recognition." *Academic Press Limited*, ISBN 0-12-269851-7.
- 42 Liu, H. and Motoda, H. (1998). "Feature selection for knowledge discovery and data mining." *Kluwer Academic Publishers*, ISBN 0-7923-8198-X.
- 43 Kittler, J. (1978). "Feature set search algorithms." in: C.H.Chen (Ed.), *Pattern Recognition and Signal Processing*, Sijthoff & Noordhoff, Netherlands. ISBN 90 286 0978 4, 41-60.
- 44 Kudo, M. and Sklansky, J. (2000). "Comparison of algorithms that select features for pattern classifiers." *Pattern Recognition*, 33, 25-41.
- 45 MathWorks (2000). "MATLAB release 12 user documentation." *The Mathwork Limited*.
- 46 Qi, Y. and Smith, B. L. (2004). "Identifying Nearest Neighbors in a Large Scale Incident Data Archive." *Transportation Research Board. Meeting (83rd : 2004 : Washington, D.C.). Compendium of papers CD-ROM*.
- 47 Härdle, W. (1990). "Applied nonparametric regression." *Cambridge University Press*. ISBN 0-521-38248-3.
- 48 Robinson, S. and Polak, J. "Overtaking Rule approach for the cleaning of matched license plate data." Submitted to *Journal of Transportation Engineering*.
- 49 Begon, C. (2004). "Using automatic number plate recognition data to estimate vehicle speeds in central London." MSc dissertation, *Imperial College London*.
- 50 Robinson, S. and Polak, J. W. (2004). "Some new perspectives on Urban Link Travel Time Models: Is the k-nearest neighbours approach the solution?" *36th Annual Conference of the Universities Transport Studies Group, 5th - 7th January 2004, Newcastle, UK*.
- 51 Xie, C., Cheu, R. L., and Lee, D.-H. (2004). "Improving Arterial Link Travel Time Estimation by data fusion." *Transportation Research Board. Meeting (83rd : 2004 : Washington, D.C.). Compendium of papers CD-ROM*.

LIST OF TABLES AND FIGURES

TABLE 1 – Distance Metrics For Use With Multivariate Data. Equation gives the distance between two points a and b located at x_a and x_b in a p-dimensional feature space.

TABLE 2 – Linear Regression Model Showing Which Levels The k-NN Based ULTT Model Is Sensitive To.

TABLE 3 – Overall Accuracy Of Five ULTT Models. * - significant at 90% level.

FIGURE 1 – Example Of The Use Of The k-NN Method When $k = 5$.

FIGURE 2 – Data collected from a link near Russell Square in central London

FIGURE 3 – Relationship Between MAPE And The Value Of k For Various Sizes Of Database (size of database shown on the right hand axis). Data has been smoothed.

FIGURE 4 – Performance of the k-NN Lowess based ULTT model against other ULTT models over the whole range of operation.

TABLE 1

TABLE 1 – Distance Metrics For Use With Multivariate Data. Equation gives the distance between two points a and b located at x_a and x_b in a p-dimensional feature space.

Name	Description	Equation
Euclidean	Distance that the ‘crow flies’ between two points. Unscaled.	$d_{ab}^2 = (x_a - x_b)(x_a - x_b)'$
Unit Map	All values are mapped onto the unit hyper-cube – i.e. each value is normalised such that the minimum value of the attribute is mapped to 0 and the maximum value of the attribute is mapped to 1.	$d_{ab}^2 = (f(x_a) - f(x_b))(f(x_a) - f(x_b))'$
Standardized Euclidean (SE-STD and SE-VAR)	This metric weights each value by another value constant to each attribute. Two common attribute-constants are: <ul style="list-style-type: none"> • Standard deviation (S.E-STD) i.e. $V = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_p)$ • Variance (S.E-VAR) i.e. $V = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$ 	$d_{ab}^2 = (x_a - x_b)V^{-1}(x_a - x_b)'$
Mahalanobis	This metric weights each variable by its variance and covariance with other variables. NB. $\Sigma = p \times p$ dimension Variance-Covariance matrix.	$d_{ab}^2 = (x_a - x_b)\Sigma^{-1}(x_a - x_b)'$
City Block	The sum of the difference of each of the variables – i.e. the distance a taxi would take in a street-grid system getting from one point to another.	$d_{ab} = \sum_{j=1}^p x_{aj} - x_{bj} $
Minkowski	A generalised distance metric: $c = 1$ gives the city block metric $c = 2$ gives the Euclidean metric.	$d_{ab} = \left[\sum_{j=1}^p x_{aj} - x_{bj} ^c \right]^{1/c}$

TABLE 2

TABLE 2 – Linear Regression Model Showing Which Levels The k-NN Based ULTT Model Is Sensitive To.

Variable Name	Description	Coefficient	t	p
Constant		21.642	134.634*	0.000
β_{DM1}	DM = SE-STD	0.377	2.147*	0.032
β_{DM3}	DM = Mahalanobis	0.150	0.853	0.394
β_{DM4}	DM = Unit Map	-0.192	-1.090	0.276
β_{LEM2}	LEM = median	-1.995	-10.081*	0.000
β_{LEM3}	LEM = regression	1.554	7.508*	0.000
β_{LEM4}	LEM = Lowess	-2.143	-9.730*	0.000
$\beta_{LEM2, k100}$	LEM = median, k = 100	-1.043	-2.202*	0.028
$\beta_{LEM2, k200}$	LEM = median, k = 200	-1.105	-2.332*	0.020
$\beta_{LEM2, k500}$	LEM = median, k = 500	-0.911	-1.924	0.054
$\beta_{LEM2, k1000}$	LEM = median, k = 1000	-0.534	-1.126	0.260
$\beta_{LEM2, k1500}$	LEM = median, k = 1500	-0.149	-0.315	0.753
$\beta_{LEM2, k5000}$	LEM = median, k = 5000	3.001	6.335*	0.000
$\beta_{LEM3, k100}$	LEM = regression, k = 100	4.047	8.476*	0.000
$\beta_{LEM3, k200}$	LEM = regression, k = 200	1.574	3.297*	0.001
$\beta_{LEM3, k500}$	LEM = regression, k = 500	-1.652	-3.459*	0.001
$\beta_{LEM3, k1000}$	LEM = regression, k = 1000	-2.413	-5.054*	0.000
$\beta_{LEM3, k1500}$	LEM = regression, k = 1500	-2.464	-5.159*	0.000
$\beta_{LEM3, k5000}$	LEM = regression, k = 5000	-2.412	-5.052*	0.000
$\beta_{LEM4, k100}$	LEM = Lowess, k = 100	9.326	8.957*	0.000
$\beta_{LEM4, k200}$	LEM = Lowess, k = 200	3.557	6.453*	0.000
$\beta_{LEM4, k500}$	LEM = Lowess, k = 500	0.198	0.407	0.684
$\beta_{LEM4, k1000}$	LEM = Lowess, k = 1000	-0.474	-0.981	0.326
$\beta_{LEM4, k1500}$	LEM = Lowess, k = 1500	-0.515	-1.066	0.287
$\beta_{LEM4, k5000}$	LEM = Lowess, k = 5000	-0.222	-0.459	0.646

NB. DM = distance metric – see table 1, LEM = local estimation method. (Adjusted $R^2 = 0.142$). NB. * denotes significant at the 95% confidence level.

TABLE 3

TABLE 3 – Overall Accuracy Of Five ULTT Models. * - significant at 90% level.

ULTT model	Overall			estimated vs observed	
	(1) MAPE, %	(2) RMSE, secs	(3) % worse than k-NN Lowess	(4) t-test	(5) corr p
ANN	21.33	122.13	58.1	-2.077*	-0.846
Basic Regression	21.16	121.51	59.1	-2.330*	-0.854
Gault & Taylor	20.84	121.05	57.6	-2.084*	-0.848
k-NN lowess	18.24	120.84	NaN	-7.049*	-0.817
k-NN median	17.83	123.56	51.6	-17.364*	-0.877
Naïve Mean	30.04	140.67	63.8	-4.638*	-1.000
Naïve Median	27.47	142.74	63.1	-20.231*	-1.000
Time: day of week	27.82	143.40	63.7	-17.573*	-0.989
Time: day of week & period 15	29.77	168.79	59.7	-0.407	-0.709
Time: period 15	20.44	135.83	54.9	-19.457*	-0.930

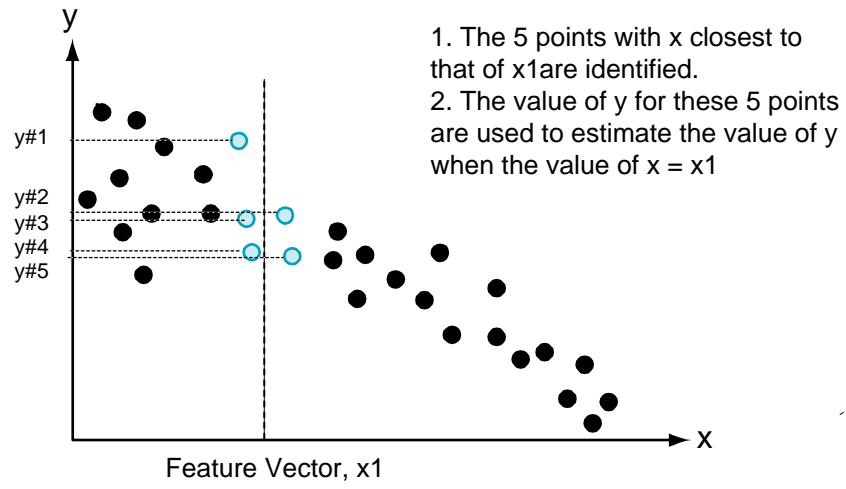
FIGURE 1**FIGURE 1 – Example Of The Use Of The k-NN Method When $k = 5$.**

FIGURE 2

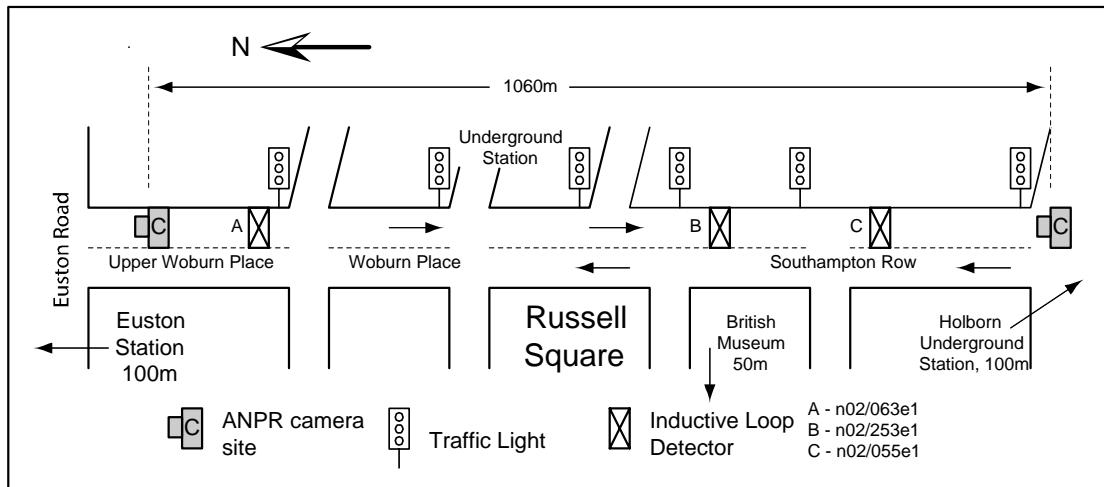


FIGURE 2 – Data collected from a link near Russell Square in central London

FIGURE 3

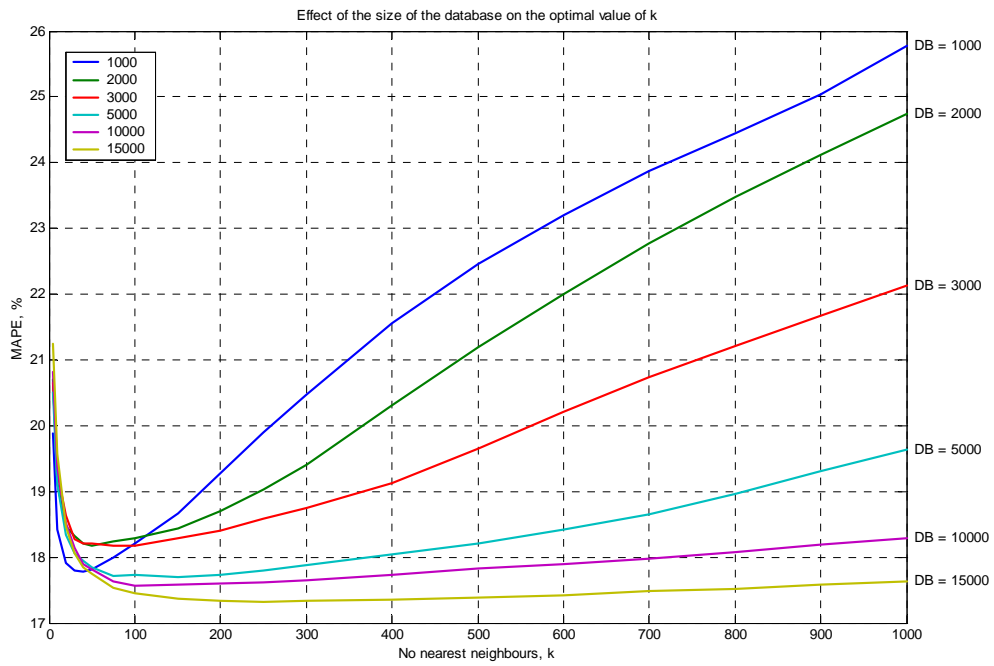


FIGURE 3 – Relationship Between MAPE And The Value Of k For Various Sizes Of Database (size of database shown on the right hand axis). Data has been smoothed.

FIGURE 4

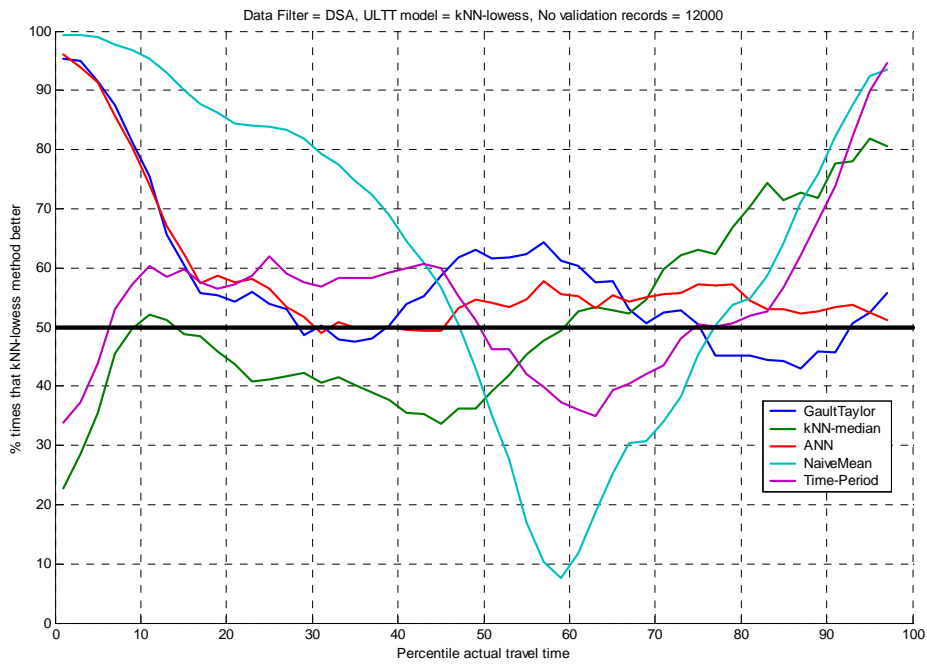


FIGURE 4 – Performance of the k-NN Lowess based ULTT model against other ULTT models over the whole range of operation.