Computers, Environment and Urban Systems 36 (2012) 538–550

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Non-parametric regression for space–time forecasting under missing data

James Haworth *, Tao Cheng [1]

Civil, Environmental and Geomatic Engineering, University College London, United Kingdom

ARTICLE INFO

Article history:
Available online 30 September 2012

Keywords:
Kernel regression
Missing data
Spatio-temporal
Transport
Journey time
Imputation

ABSTRACT

As more and more real time spatio-temporal datasets become available at increasing spatial and temporal resolutions, the provision of high quality, predictive information about spatio-temporal processes becomes an increasingly feasible goal. However, many sensor networks that collect spatio-temporal information are prone to failure, resulting in missing data. To complicate matters, the missing data is often not missing at random, and is characterised by long periods where no data is observed. The performance of traditional univariate forecasting methods such as ARIMA models decreases with the length of the missing data period because they do not have access to local temporal information. However, if spatio-temporal autocorrelation is present in a space–time series then spatio-temporal approaches have the potential to offer better forecasts. In this paper, a non-parametric spatio-temporal kernel regression model is developed to forecast the future unit journey time values of road links in central London, UK, under the assumption of sensor malfunction. Only the current traffic patterns of the upstream and downstream neighbouring links are used to inform the forecasts. The model performance is compared with another form of non-parametric regression, *K*-nearest neighbours, which is also effective in forecasting under missing data. The methods show promising forecasting performance, particularly in periods of high congestion.

1. Introduction

In recent decades, the availability of (near) real time spatio-temporal datasets has increased massively, with data routinely being collected on transportation networks, communications networks, networks of environmental monitoring stations, the internet and the world wide web amongst others. As a result, the traditional problems of data scarcity and lack of computational power that constrained research in the latter half of the 20th century have been replaced with fresh challenges of data mining, knowledge discovery and forecasting. Modelling of spatio-temporal data presents a unique set of problems as they often exhibit spatio-temporal dependence, nonlinearity and heterogeneity, which violate the normality assumption of classical statistics and render standard statistical models such as ordinary least squares (OLSs) ineffective. As a result, there has been great interest in developing models to deal with such data.

The state of the art in statistical modelling of spatio-temporal processes represents the outcome of several decades of cross-pollination of research between the fields of time series analysis, spatial statistics and econometrics. Some of the methods described in the literature include space–time autoregressive integrated moving average (STARIMA) models (Pfeifer & Deutsch, 1980) and variants, multiple ARIMA models, space–time geostatistical models (Heuvelink & Griffith, 2010; Kyriakidis & Journel, 1999), spatial panel data models (Baltagi, 2005; Elhorst, 2003; Wooldridge, 2002), geographically and temporally weighted regression (Huang, Wu, & Barry, 2010) and eigenvector spatial filtering (Griffith, 2010). In parallel to the development of statistical space–time models, there was a multidisciplinary explosion of interest in non-parametric machine learning methods, and many of these have been successfully adapted to work with spatio-temporal data due to their innate ability to model complex nonlinear relationships. Examples include artificial neural networks, support vector machines and non-parametric regression techniques, and the texts of Kanevski and Maignan (2004), Kanevski (2008) and Kanevski, Timonin, and Pozdnukhov (2009) provide a good introduction to their application to spatial problems. Generally, space–time forecasting models have been developed to be applied to near complete space–time series where there are few missing data and their performance suffers when large amounts of data are missing.

Unfortunately, missing data is one of the major problems that impacts on the application of spatio-temporal models to real life problems. Missing data complicates the application of many spatio-temporal models, and has been recognised as a problem in various application areas including environmental monitoring (Glasbey, 1995; Smith et al., 1996, 2003), video image reconstruction (Kokaram & et al., 1995) and hydrology (Amisigo & Van De

* Corresponding author. Tel.: +44 781 607 6958.
E-mail addresses: j.haworth@ucl.ac.uk (J. Haworth), tao.cheng@ucl.ac.uk (T. Cheng).
[1] Tel.: +44 781 560 1230.

0198-9715/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2012.08.005

Giesen, 2005). Many statistical space–time models require a complete series to ensure the proper ordering of the temporal dimension and calibration of parameters. Additionally, multivariate statistical analyses rely on the calculation of means and covariance matrices. Under missing data, their calculation is usually not admissible. Covariance matrices calculated from incomplete data may not be positive semidefinite and may produce negative eigenvalues, thus affecting the application of methods such as principal components analysis that rely on eigenvector decomposition (Schneider, 2001).

Missing data is usually a result of failure in the data collection process, e.g. that caused by a faulty sensor. There are three types: (1) missing completely at random, whereby the missing points are independent of each other; (2) missing at random, whereby missing points are related to the neighbouring points; and (3) not missing at random, whereby the missing data has some pattern, possibly linked to a long term sensor malfunction (Qu et al. (2009)). These failures result in incomplete space–time series and make accurate forecasting difficult or even impossible as no data can mean no forecast. The alternative is to calculate estimates of the missing values. The first two types of missing data can often be modelled effectively using univariate methods due to their short length. However, the latter type is more difficult to handle due to the extended period of missing data where no local temporal information is available. This will be the focus of this study.

Furthermore, dealing with missing data can be implemented in two settings, which is known as imputation in an offline setting and forecasting in an online (real time) setting. Various offline imputation methods have been used in the spatio-temporal modelling literature, many of them based on principal components analysis (Smith & et al., 1996) or expectation maximisation (Schneider, 2001; Smith, Kolenikov, & Cox, 2003). However, there has not been a great deal of research that has focussed specifically on dealing with missing data in a real time setting, where its presence necessitates long range forecasting. One application domain where researchers have begun to tackle this issue is transport (van Lint, Hoogendoorn, & van Zuylen, 2005; Whitlock & Queen, 2000). This has been motivated by the frequently high levels of missing data that are present on traffic monitoring networks.

In the next section, methods that are currently used for online and offline imputation of traffic data are reviewed. This leads to the development of a spatio-temporal approach to traffic forecasting under missing data based on kernel-regression, which is introduced in Section 3. Kernel regression is a non-parametric regression technique that is used to estimate the conditional expectation of a random variable. It is a memory based pattern matching method that produces forecasts as a combination of historical data points, weighted by a kernel function. The case study is presented in Section 4, using unit journey time data collected on London's road network. The London Congestion Analysis Project (LCAP) network is described, and a missing data analysis is carried out to highlight the missing data issue. The experimental design is then outlined. The results are presented and discussed in Section 5. Finally, some conclusions and directions for future research are given in Section 6.

2. Existing approaches in forecasting under missing data

Providing accurate, up to date information to road users is one of the primary goals of intelligent transportation systems. Forecasting future traffic conditions typically involves the application of algorithms that forecast based on the current conditions on the network. Often, this is accomplished by univariate methods, including statistical time series models such as ARIMA and seasonal variants (Williams et al., 1998), neural networks (Dougherty, 1995; Dougherty & Cobbett, 1997) and non-parametric regression (Smith, Williams, & Keith Oswald, 2002). Vlahogianni, Golias, and Karlaftis (2004) provide a good review and Karlaftis and Vlahogianni (2011) compare the relative merits of neural and statistical approaches. More recently, there has been an increase in the application of space–time models to transport network data. This is a natural progression as traffic networks are systems of flows. Flows observed at one spatial location at one time will be observed at another location downstream at a later time. Furthermore, Lighthill Witham Richards (LWRs) theory states that, in congested conditions, shockwaves propagate in the opposite direction to the flow of traffic (Lighthill & Whitham, 1955; Richards, 1956). These phenomena can be clearly observed on highways, although they are disrupted by exogenous factors in the urban environment.

Some of the spatio-temporal models that have been successfully applied to the forecasting of traffic variables to date include space–time autoregressive integrated moving average (STARIMA) (Kamarianakis & Prastacos, 2005), neural networks (van Lint et al., 2005), Bayesian networks (Queen & Albers, 2008) and state space models (Stathopoulos & Karlaftis, 2003). For forecasting to be as successful as possible, it is desirable to have complete data. However, traffic sensors are notoriously prone to failure, and some sensor networks can experience capture rates of as low as 10–30% (Sharma, Lingras, & Zhong, 2004). The missing data issue has led to considerable interest in the development of offline imputation methods. More recently, as real time forecasting has become a feasible goal, methods are required to simultaneously deal with missing data and produce forecasts. In the following subsections, the current methods for offline imputation and forecasting under missing data are reviewed.

2.1. Univariate imputation

Missing data is typically dealt with using offline imputation techniques. In practice, simple time series or factor based approaches are often used which do not account for the dynamics of the real traffic situation (Zhong, Lingras, & Sharma, 2004b). However, in recent years, a number of more sophisticated approaches have been developed that attempt to improve estimates. The studies of Zhong et al. (2004a, 2004b) and Sharma et al. (2004) compared the performance of factor approaches and ARIMA with genetically trained time delay neural networks (TDNNs) and locally weighted regression. The models were applied to hourly flows from permanent traffic counters (PTCs) on the highway network in Alberta, Canada. The best performing model was a seasonal local regression model trained by a GA that made use of data from either side of the failure.

A subsequent study by Zhong, Sharma, and Lingras (2006) proposed a simple imputation method based on pattern matching using data from before and after the failure. The method extracts a candidate set of normalised traffic patterns from the historical dataset and selects the best fitting pattern based on a minimum squared error (MSE) between the candidate pattern and the pattern under study. This best fitting pattern is then used to impute the data. The method was tested on traffic count data collected using permanent traffic counters (PTCs) on the ATR C002181 in Alberta, Canada and was found to outperform factor approaches, ARIMA and exponential moving average. The use of a single traffic pattern for imputation may be appropriate for the smooth, hourly data used in Zhong et al. (2006), but is unlikely to be sufficient when applied to data of higher temporal resolution that is corrupted by noise. This was recognised by Liu, Sharma, and Datla (2008) who proposed a k-NN method for imputation of missing traffic data during holiday periods. To determine the k neighbours, a state vector is defined, augmented with historical averages (after Smith et al., 2002). To take into account the ranking of the neighbour set, the weights vary inversely with distance from the

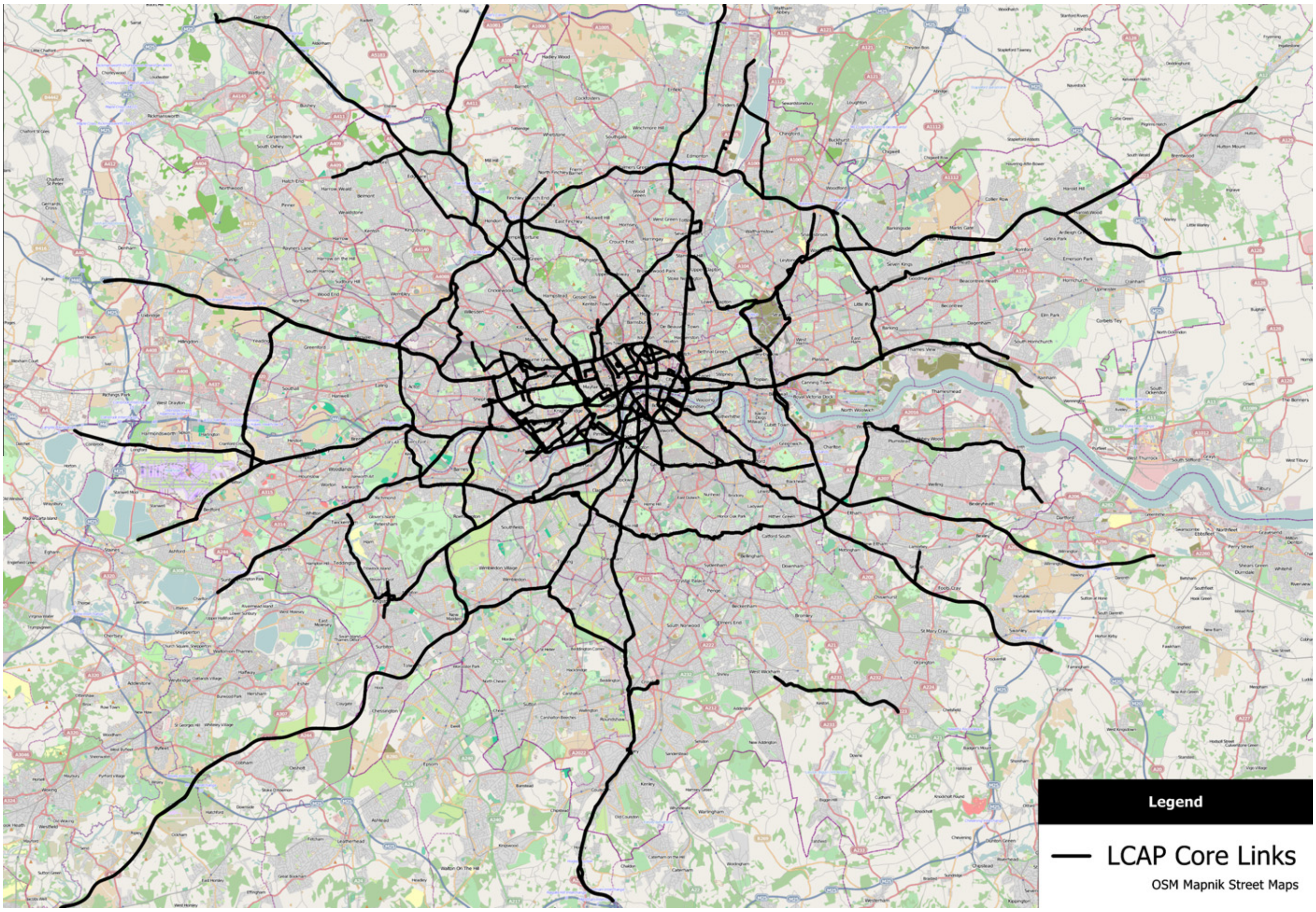

Fig. 1. Spatial extent of the LCAP core links network (originally published in Cheng et al, 2011).

state vector. The results of Zhong et al. (2004b, 2006) highlight the effectiveness of simple pattern based approaches for imputation.

Qu et al. (2009) proposed a probabilistic principal components analysis (PPCA) method for urban traffic flow imputation. The authors note two desirable properties of PPCA. Firstly, it uses only the major information and avoids overfitting faced by spline methods and regression methods. Secondly, the model parameters can be computed directly from the data via efficient eigenvector decomposition. The method can also be used to identify outliers or unusual patterns in the data. Components that cause a notable increase in variance and covariance in the original variables represent extreme values. The PPCA method is tested on 5 min flow data from loop detectors in Beijing, China and is found to outperform spline and historical imputation methods, particularly when the ratio of missing data is high. It also produces imputed values that are statistically consistent with the distribution of flow data.

2.2. Multiple imputation

The methods described thus far all make a single imputation for each missing point. Their effectiveness is assessed by comparison with validation data. However, in real situations where the data is actually missing, this validation data is not available. Therefore, having some indication of the reliability of the estimated data can be important. In light of this, multiple imputation is a common approach. Ni et al. (2005) developed a multiple imputation method for imputing traffic flow. The idea behind multiple imputation is to simulate multiple random draws from a population in order to estimate an unknown parameter. The expectation maximisation (EM) algorithm is used to generate maximum likelihood estimates of the missing values and these are used as inputs to a data augmentation (DA) procedure which is run k times. The whole process is repeated n times to produce n sets of imputed data. The consistency of the missing data estimates can then be examined by analysing the distribution of the n sets.

Wang, Zou, and Chang (2008) proposed an alternative approach that combines non-parametric regression techniques with multiple imputation. k-NN is used as the imputation algorithm, with the historical dataset first being categorised into free-flowing, moderately congested and highly congested patterns. The current traffic pattern is then assigned to one of these categories prior to the k-NNs being computed. A second method further classifies the data into different subsegments based on road characteristics. The procedures are carried out multiple times to impute travel times on 10 roadside detectors on the I-70 in Maryland, USA and are found to outperform mean substitution and Bayesian forecasting methods. In this method, spatial information is included through the grouping of sensors into subsegments; however, there is no discrimination between patterns collected on the sensor in question and the adjacent sensors.

2.3. Spatio-temporal approaches

Univariate techniques can be very effective but their applicability decreases with the length of the period of missing data. Many imputation methods assume that missing data are sparsely located within the series so that neighbouring temporal information can be used for univariate forecasts. This is an unrealistic assumption as

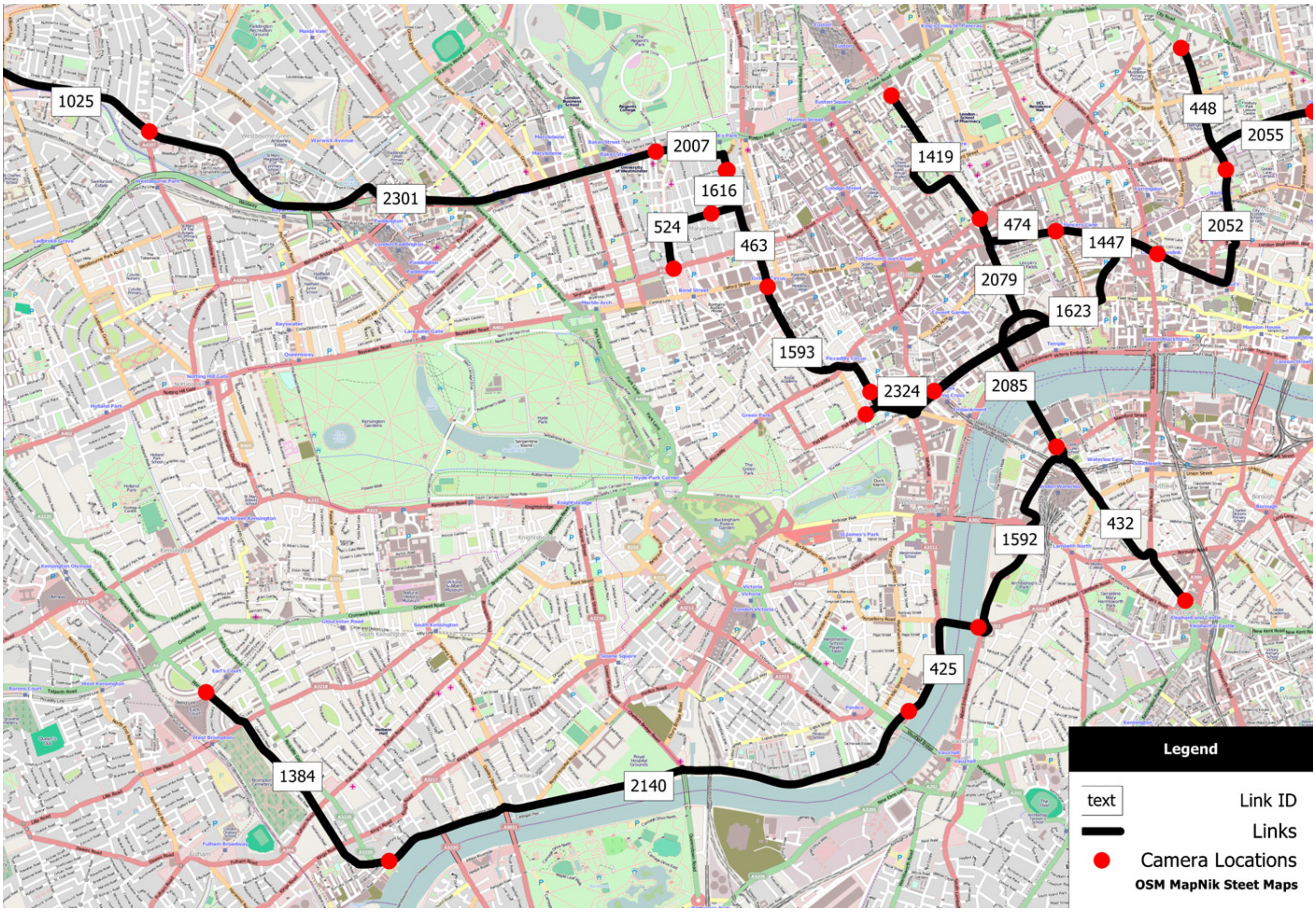

Fig. 2. Location of the test network in Central London. (originally published in Cheng et al, 2011).

sensors often malfunction and stop collecting data for several consecutive time periods, i.e. when data is missing not at random. Surprisingly, the use of spatial information in imputation and forecasting of traffic under missing data has not received much attention in the literature, although some notable exceptions exist. Whitlock and Queen (2000) proposed a dynamic graphical model for forecasting traffic flows under missing data. In their approach, the traffic flow at a measurement site is considered to be independent of all other sites given its parents, which are its upstream neighbours. Methods were developed for dealing with situations where the data from one or more parents and/or children are missing. It was found that strong apriori knowledge of the missing series, which can be assumed in most cases, resulted in good performance. However, this is dependent on the regression parameters being fairly constant over time, which may not be the case in an urban setting.

van Lint et al. (2005) developed a state space neural network (SSNN) that is robust to missing data. It was discovered that training the SSNN with perfect data reduces its capacity to deal with missing data. By incorporating simple imputation schemes such as spatial interpolation and exponential smoothing, the algorithm becomes less sensitive to missing data, despite slight changes in the statistical properties of the input data. The method was shown to outperform SSNN with no data imputation on simulated (FOSIM) data with up to 40% missing data. Additionally, the method was tested on real highway data taken from the MONICA system in the Netherlands and was shown to outperform the online estimator being used at the time. This method highlights the importance of effective imputation, and it is likely that its accuracy would be improved further if more accurate techniques were initially used to impute the missing data in the training set.

Table 1
Description of patch types used to replace missing data.

Patch type	No. of missing points	Interpolation method
1	None	None
2	1	Average of adjacent observations
3	2–6	Interpolated from adjacent observations
4	>6	Replaced with historical profile data (average of each time point)

2.4. Summary

A range of imputation models have been described in the preceding subsections with varying degrees of sophistication, ranging from simple univariate averaging methods to complicated spatio-temporal neural network architectures which are capable of operating in a real time forecasting setting. It can be concluded that univariate methods will usually perform well when the length of missing data is short but that spatio-temporal models will be more effective when extended periods of missing data are evident. Additionally, multiple imputation methods have benefits in terms of interpreting uncertainty. However, in a real time forecasting setting, multiple imputation methods are not appropriate as they are computationally intensive. Computationally efficient models are needed that are suitable for real time use. In the following

Table 2
Breakdown of data on the test network by patch type, 2009.

Link	Count 1	Count 2	Count 3	Count 4	Perc. 1	Perc. 2	Perc. 3	Perc. 4
R1025	6676	884	1673	5743	44.58	5.90	11.17	38.35
R2301	3383	698	2406	8489	22.59	4.66	16.07	56.68
R2007	5291	532	745	8408	35.33	3.55	4.97	56.14
R1616	7875	805	1397	4899	52.58	5.38	9.33	32.71
R524	9689	1418	1970	1899	64.70	9.47	13.15	12.68
R463	10593	768	771	2844	70.73	5.13	5.15	18.99
R1593	12063	899	636	1378	80.55	6.00	4.25	9.20
R2324	10911	1474	1214	1377	72.86	9.84	8.11	9.19
R2085	2233	348	2542	9853	14.91	2.32	16.97	65.79
R432	13901	697	291	87	92.82	4.65	1.94	0.58
R1592	6327	1342	3306	4001	42.25	8.96		26.72
R425	13614	336	306	720	90.91	2.24	2.04	4.81
R2140	10501	584	1362	2529	70.12	3.90	9.09	16.89
R1384	7951	491	992	5542	53.09	3.28	6.62	37.01
R2079	4852	1087	4505	4532	32.40	7.26	30.08	30.26
R1419	14537	160	124	155	97.07	1.07	0.83	1.03
R474	10680	1606	1764	926	71.31	10.72	11.78	6.18
R1447	13634	367	208	767	91.04	2.45	1.39	5.12
R1623	4206	887	4336	5547	28.08	5.92	28.95	37.04
R2052	3084	547	3387	7958	20.59	3.65	22.62	53.14
R448	10261	1593	2377	745	68.52	10.64	15.87	4.97
R2055	12330	1450	1000	196	82.33	9.68	6.68	1.31
Avg.	8845.09	862.41	1696.00	3572.50	59.06	5.76	11.32	23.85
S.D.	3853.96	445.55	1290.33	3100.50	25.73	2.98	8.62	20.70

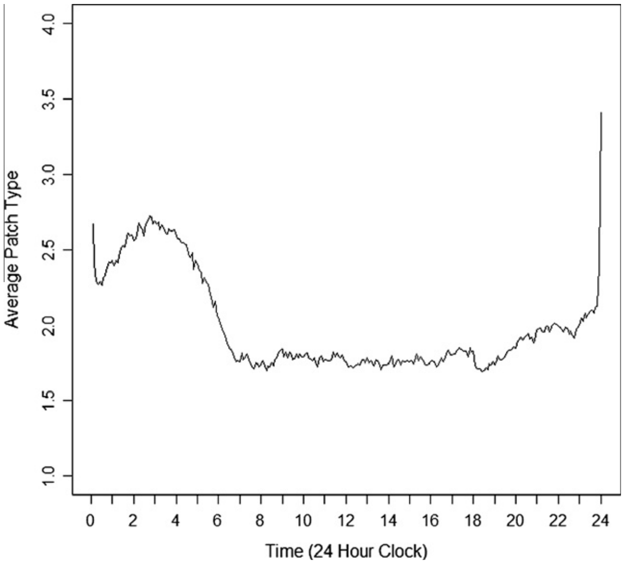

Fig. 3. Average patch type on the test network over 24 h.

section, kernel regression is introduced as a method for forecasting of space–time series under missing data that is both easy to interpret and simple to implement. In addition, *K*-nearest neighbours, which has been shown to be effective in univariate imputation, is presented as an alternative method since it has not yet been used in this setting.

3. Kernel regression for space–time forecasting

In the absence of local temporal information, it is necessary to use local spatial information in order to obtain forecasts. The literature review in Section 2 revealed that relatively simple pattern matching techniques such as locally weighted regression and *k*-NN can be very successful in univariate imputation of traffic data. However, there has been little research into the efficacy of such approaches when temporal information is not available. In this study, kernel regression is applied to the problem of spatio-temporal traffic forecasting under complete missing data. It is compared with three additional models; two forms of *K*-Nearest Neighbours (KNN) regression and a combined KR and KNN model. These are introduced in turn in the following subsections. In order to define these models, it is necessary to firstly define space–time series.

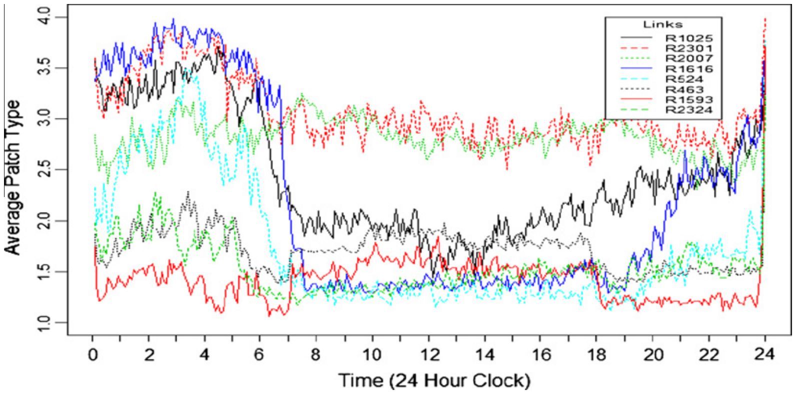

Fig. 4. Average patch types for eight links of the test network.

Table 3
Percentage and count of forecastable points for each link in the upstream and downstream directions. "–" Indicates that there is no neighbour. Cells shaded in grey denote links that are removed from the analysis.

Link	% Upstream	Count upstream	% Downstream	Count downstream
R425	98.8	1775	97.2	1747
R432	1.0	18	–	–
R448	–	–	1.1	20
R463	4.1	73	89.4	1607
R474	61.1	1098	15.4	277
R524	53.0	952	–	–
R1025	–	–	0.4	7
R1384	92.0	1653	–	–
R1419	1.4	26	–	–
R1447	1.0	18	43.7	785
R1592	1.0	18	99.4	1787
R1593	82.5	1482	60.5	1087
R1616	3.4	61	49.2	885
R1623	0.5	9	–	–
R2007	0.2	4	7.3	131
R2052	9.0	162	0.6	11
R2055	–	–	0.9	17
R2079	23.9	429	7.3	131
R2085	11.7	211	18.3	329
R2140	98.3	1767	84.9	1525
R2301	6.6	118	1.3	23
R2324	81.5	1464	0.0	0

3.1. Space–time series

A space–time series consists of $N * S$ observations $z_1^{(1)}, \ldots, z_N^{(S)}$ of a spatio-temporal process collected at $t = 1, 2, \ldots, N$ times which are usually equally spaced; and $s = 1, 2, \ldots, S$ often irregularly spaced locations. Under the missing data assumption, the value of the series $z^{(s)}(t + \tau)$ at location s at time $t + \tau$ is modelled as a function of the value of the series at its spatial neighbours at $s + l$ at m previous times, where l is a spatial lag determining the separation distance, τ is the temporal lag determining the forecast horizon and m is the embedding dimension. This leads to an embedded series with the following form:

$$z_{t+\tau}^{(s)} = f\left(z_t^{(s+l)}, z_{t-\tau}^{(s+l)}, \ldots, z_{t-(m-1)\tau}^{(s+l)}\right) \tag{1}$$

For ease of presentation, the LHS of Eq. (1) will be denoted $Y^{(s)}$ and the RHS will be denoted $X^{(s+l)}$, leading to following equation:

$$Y_{t+\tau}^{(s)} = f(X_t^{(s+l)}) \tag{2}$$

The spatial lag 1 is incorporated into the models using $S * S$ spatial weight matrices $W_1, W_2, \ldots, W_L$ that determine the spatial relationship between neighbouring locations. W_1 is a binary [0,1] matrix that contains all first order spatial relations between road links, where 1 indicates that two road links are adjacent, zero otherwise. Spatial weight matrices $W_2, \ldots, W_L$ contain the spatial relations between road links up to spatial order L. For a given location s and weight matrix W_1, the space time series $z_{t+\tau}^{(s)}$ is constructed by concatenating the series for which $w_{sj} = 1$.

3.2. Kernel regression

Kernel methods are a class of algorithms that are used for pattern analysis tasks, including classification and regression. The best known example is the support vector machine (SVM), which has been shown to provide generalisation performance that either matches or is significantly better than competing methods in a wide range of applications (Burges, 1998). The performance gain of kernel methods in general is provided by the use of kernels. A kernel is a function k that for all $x, x' \in X$ satisfies:

$$k(x, x') = \langle \phi(x), \phi(x') \mathring{a} ngle \tag{3}$$

where ϕ is a mapping from X to a feature space F (Shawe-Taylor and Cristianini, 2004):

$$\phi : x \to \phi(x) \in F \tag{4}$$

Kernels define similarity functions between data patterns in high, possibly infinite dimensional spaces, where linear algorithms are applied to find nonlinear relations. This allows established and theoretically well founded linear algorithms to be applied to nonlinear problems. Furthermore, explicit evaluation of the coordinates of points in the feature space is not required, making kernel methods computationally efficient.

In this study, kernel regression (KR) is used for forecasting space–time series under missing data. Kernel regression (KR) is a non-parametric regression technique simultaneously developed by Nadaraya (1964) and Watson (1964) that is used to estimate the conditional expectation of a random variable. It has been alternatively described as a general regression neural network (GRNN, Specht, 1991). Given a set of n pairs of variables $(X_1, Y_1), \ldots, (X_n, Y_n)$, the goal of KR is to estimate a regression function of Y on X according to following equation:

$$m(x) = E(Y|X = x) \tag{5}$$

Continuing with the notation from Section 3.1, a KR estimator for the space–time series can be defined as follows:

$$Y_{t+\tau}^{(s)} = \frac{\sum_{i=1}^{D} k(X_t^{(s+l)},\ X_i^{(s+l)}) Y_i^{(s)}}{\sum_{i=1}^{D} k(X_t^{(s+l)},\ X_i^{(s+l)})} \tag{6}$$

where $i = 1, 2, \ldots, D$ are the training data and k is a radial basis function (RBF) kernel with bandwidth σ:

$$k(x, x') = \exp\left(-\frac{||x - x'^{2}||}{2\sigma^2}\right) \tag{7}$$

The KR estimator gives a weighted average of the observed independent variables $Y_i^{(s)}$ and the denominator ensures that the weights sum to 1. The RBF kernel in Eq. (7) is a Gaussian (bell-shaped) function. It assigns higher weight to those training patterns that are closer in Euclidean distance to the testing pattern.

(a)

Link	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
425	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
432	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
448	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
463	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
474	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
524	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1384	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1419	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1447	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1592	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1593	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1616	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1623	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2007	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2052	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
2055	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2079	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2085	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
2140	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2301	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2324	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0

(b)

Link	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
425	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
432	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
448	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
463	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
474	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
524	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1025	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1384	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1419	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1447	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1592	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1593	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1616	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1623	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2052	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
2055	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2079	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2085	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2140	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2301	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2324	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Fig. 5. First order spatial weight matrices for the (a) upstream and (b) downstream directions.

The tunable kernel parameter σ determines the bandwidth of the function, which allows the rate of distance decay to be controlled. Selecting a large value for σ results in a smooth function, while a small value results in a more complex function. The bell shape of the Gaussian function ensures that the weights tend to zero as the distance between the test pattern and the training data increases. Kernel functions of this form have been shown to perform well in most applications, although many types of kernel have been defined in different domains (see, for example, Cristianini & Shawe-Taylor, 2004).

3.3. K nearest neighbours

To provide a benchmark for model performance in the missing data setting, K Nearest Neighbours (KNN) is used. KNN is a machine learning technique that has been widely used for classification, regression and imputation (see e.g. Mitchell, 1997). The basic idea is to compute the distance (usually Euclidean) between a test sample and each of the samples in a training set:

$$d_i = \|X_t^{(s)} - X_i^{(s+l)}\| \tag{8}$$

where $i = 1, 2, \ldots, D$ are the training data. A forecast is made as a function of the K nearest samples in the training set. This function can take many forms and we consider two popular ones here. The first and simplest is to produce the output as the average of the neighbours. Using the notation from Section 3.1 and assuming the K neighbours have been found, this can be written as equation:

$$Y_{t+\tau}^{(s)} = 1 \Big/ K \sum_{i=1}^{K} Y_i^{(s)} \tag{9}$$

Table 4
Fitted model parameters.

Link	Downstream					Upstream				
	KR	KNNavg	KNNdist	KNNkernel		KR	KNNavg	KNNdist	KNNkernel	
	Sigma	*K*	*K*	Sigma	*K*	Sigma	*K*	*K*	Sigma	*K*
R425	0.075	99	92	0.25	100	0.025	84	100	0.25	84
R463	0.05	33	33	0.1	98					
R474	0.05	97	97	0.25	97	0.05	100	100	0.25	100
R524						0.025	93	93	0.1	93
R1384						0.25	100	100	1	100
R1447	0.1	92	92	0.1	92					
R1592	0.25	96	100	1	96					
R1593	0.075	100	100	1	100	0.05	100	100	1	100
R1616	0.05	100	100	0.1	100					
R2007	0.075	100	100	0.25	100					
R2052						0.075	100	100	0.25	100
R2079	0.05	100	100	1	100	0.025	87	96	0.25	87
R2085	0.25	98	100	1	98	0.075	100	100	0.1	100
R2140	0.05	99	100	0.5	99	0.1	50	84	1	50
R2301						0.025	82	99	1	82
R2324						0.05	100	100	1	100

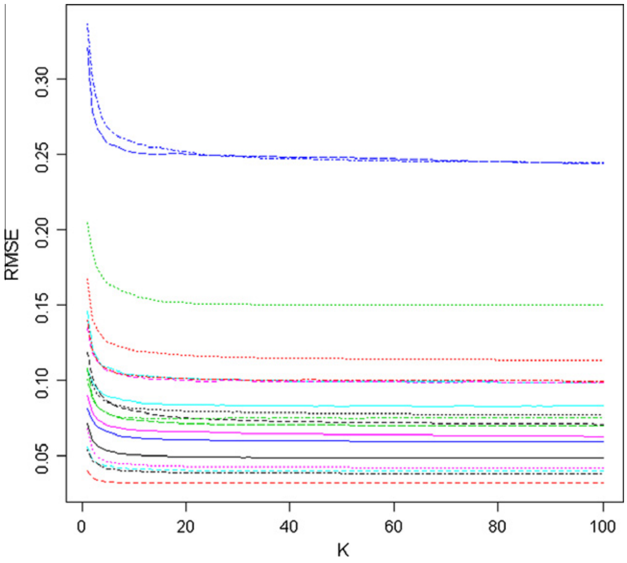

Fig. 6. Relationship between *K* and RMSE for the KNNavg model.

where *K* is the number of neighbours. This model will be referred to as KNNavg. The drawback of this approach is that it ignores the ranking of the neighbours and assigns them equal weight regardless of their similarity to the test pattern. This means that the distance information contained in the ranking is lost. Therefore, a second approach that may perform better in practice is to apply an inverse distance weighting function to the neighbours:

$$Y_{t+\tau}^{(s)} = \frac{\sum_{i=1}^{K}\left(1/d_i^2\right)Y_i^{(s)}}{\sum_{i=1}^{K}1/d_i^2} \qquad (10)$$

This model will be referred to as KNNdist and is similar to the model used in Liu et al. (2008). It assumes that the usefulness of a point $Y_i^{(s)}$ for forecasting varies inversely with the square of the distance between the corresponding training pattern $X_i^{(s+l)}$ and the test pattern $X_t^{(s)}$. This gives higher weight to those neighbours that are closer to the test sample and avoids the issue of giving high weight to samples that are high ranked neighbours but dissimilar in terms of distance. The distance function in KNNdist works in a similar way to the RBF kernel, but has two differences. Firstly, it does not have a tunable bandwidth parameter and so has less flexibility. Secondly, the shape of the function is different.

3.4. Combined model

In addition to the KNN and KR models, a third model is tested that combines the ideas of both, which is referred to as KNNkernel. It is basically KR on the set of *K* nearest neighbours and is given in equation:

$$Y_{t+\tau}^{(s)} = \frac{\sum_{i=1}^{K}k\left(X_t^{(s+l)},X_i^{(s+l)}\right)Y_i^{(s)}}{\sum_{i=1}^{K}k\left(X_t^{(s+l)},X_i^{(s+l)}\right)} \qquad (11)$$

Note the difference from Eq. (6) that the summation is over *K* rather than *D*, meaning that this equation has two parameters to tune; σ and *K*.

4. Case study

In this section, the case study data of travel times collected on London's road network is introduced and a data quality assessment is carried out. The aim of this is to identify the types of missing data the network faces in order to motivate the use of a spatio-temporal forecasting approach. Following this, the experimental design is outlined.

4.1. The LCAP data and test network

London, the capital city of the United Kingdom, is an urban conurbation with a population of around 8 million citizens. Being a city with ancient origins, its ageing transport infrastructure struggles to cope with the demands of modern commuting behaviour and there has been a long term trend of increasing congestion. During the period from 1980/1992 to 2006/2009, average weekday travel speeds in London fell by 18% in the morning peak period, 14% in the inter peak period and 12% in the evening peak period. Overall, London's roads account for around 20% of the congestion in the UK, 75% of which is concentrated on just 0.5% of the nation's roads (Transport for London, 2010).

The London Congestion Analysis Project (LCAP) network is a system of automatic number plate recognition (ANPR) cameras maintained by Transport for London (TfL) that monitor journey times on London's road network (Fig. 1). The system is link based, and cameras placed at each end of a link read number (license) plates as vehicles pass them. The individual journey time observations are averaged over a 5 min period to give the journey time data (in seconds) used in this study. Although data is available for the whole

Table 5
Root mean squared error on the training set.

Link	Downstream				Upstream				Best model
	KR	KNNkernel	KNNdist	KNNavg	KR	KNNkernel	KNNdist	KNNavg	
R425	0.041	0.041	0.040	0.040	0.048	0.048	0.049	0.048	KNNavg Down
R463	0.134	0.136	0.134	0.132					KNNavg Down
R474	0.276	0.271	0.272	0.271	0.250	0.245	0.254	0.245	KNNavg Up
R524					0.098	0.099	0.091	0.099	KNNdist Up
R1384					0.065	0.063	0.064	0.063	KNNkernel Up
R1447	0.100	0.100	0.102	0.100					KNNavg Down
R1592	0.074	0.073	0.073	0.073					KNNkernel Down
R1593	0.250	0.246	0.250	0.246	0.250	0.244	0.247	0.244	KNNkernel Up
R1616	0.080	0.080	0.081	0.080					KR Down
R2007	0.080	0.082	0.089	0.082					KR Down
R2052					0.100	0.100	0.100	0.100	KNNdist Up
R2079	0.071	0.070	0.070	0.070	0.071	0.070	0.071	0.070	KNNavg Up
R2085	0.059	0.059	0.059	0.059	0.059	0.059	0.061	0.059	KR down
R2140	0.043	0.043	0.043	0.043	0.041	0.040	0.040	0.040	KNNkernel Up
R2301					0.043	0.042	0.040	0.042	KNNdist Up
R2324					0.038	0.038	0.038	0.038	KNNkernel Up

Table 6
Root mean squared error on the testing set.

Link	HA	Downstream				Upstream				Best model
		KR	KNNkernel	KNNdist	KNNavg	KR	KNNkernel	KNNdist	KNNavg	
R425	0.027	0.043	0.043	0.042	0.043	0.043	0.043	0.043	0.043	HA
R463	0.426	0.192	0.191	0.197	0.200					KNNkern Down
R474	0.576	0.359	0.364	0.364	0.364	0.339	0.346	0.346	0.346	KR Up
R524	0.076					0.082	0.082	0.082	0.082	HA
R1384	0.073					0.051	0.050	0.050	0.050	KNNavg Up
R1447	0.102	0.086	0.089	0.087	0.089					KR Down
R1592	0.055	0.064	0.062	0.062	0.062					HA
R1593	0.186	0.182	0.184	0.183	0.184	0.273	0.248	0.257	0.248	KR Down
R1616	0.292	0.096	0.097	0.097	0.097					KR Down
R2007	0.880	0.087	0.087	0.089	0.087					KR Down
R2052	0.152					0.101	0.100	0.100	0.100	KNNdist Up
R2079	0.130	0.075	0.073	0.073	0.073	0.075	0.075	0.075	0.075	KNNdist Down
R2085	0.129	0.100	0.100	0.100	0.100	0.100	0.102	0.102	0.102	KNNavg Down
R2140	0.056	0.052	0.052	0.052	0.052	0.059	0.056	0.056	0.056	KR Down
R2301	0.250					0.178	0.177	0.177	0.177	KNNavg Up
R2324	0.281					0.199	0.198	0.198	0.198	KNNavg Up

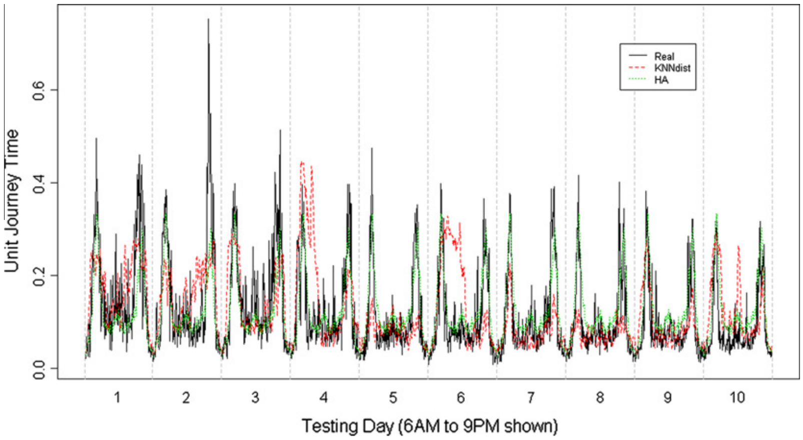

Fig. 7. Testing performance of link 425.

ANPR network, a subsection is chosen in order to investigate the feasibility of the methods (Fig. 2). A centrally located network was chosen because the average link lengths are lower in the central area, meaning that neighbouring locations are more likely to provide meaningful information. The test network comprises 22 links which are those used in Cheng, Haworth, and Wang (2011). Even on this small network, the link lengths vary considerably, ranging from 473.4 m to 3.85 km with an average length of 1.4 km.

Thirty-five consecutive Tuesdays beginning on January 6th 2009 were selected for case study. Tuesday data only is chosen as the behaviour of traffic is known to be different on different days of the week and Tuesday is presumed to be close to an 'average'

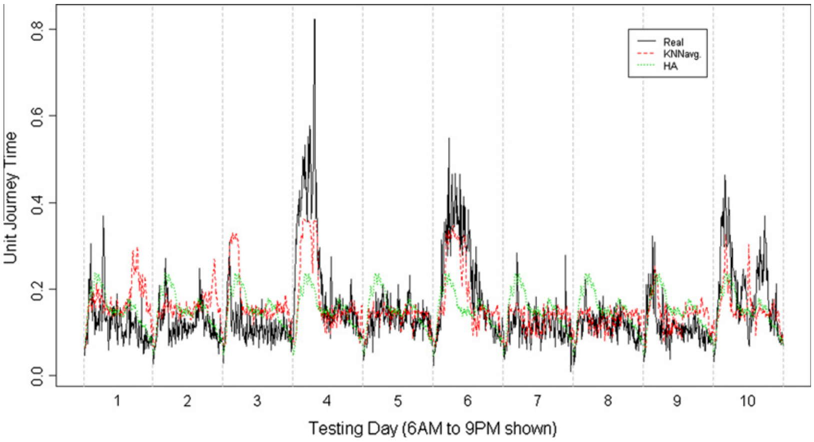

Fig. 8. Testing results for link 1384.

weekday, separate from the influences of weekend traffic patterns. The ANPR camera network was originally installed for operational purposes so collecting journey time data is a secondary use. It is prone to missing data as the technology is ageing and fails often. The cameras also have difficulty identifying number plates in highly congested conditions where there is insufficient headway between vehicles. In the following subsection, the frequency and types of missing data that are present on the network are examined in detail.

4.2. Missing data analysis

The ANPR data are subject to high levels of missing data, which are imputed offline by TfL using a process called patching. Patching involves replacing the missing values with estimates, which vary according to the number of points that are missing in succession. In total, there are four patch types, which are coded here from 1 to 4. A description of each patch type is given in Table 1.

The proportion of each patch type has a significant effect on the usability of the data for forecasting purposes. Patch type 1 is the desired patch type for obvious reasons as it indicates the presence of real data. The effect of patch type 2 is likely to be small as it represents the absence of a single 5 min observation and the replacement value is calculated from the true observed traffic pattern. Patch type 3 is of more concern, particularly as the size of the gap increases. Six points missing corresponds to 30 min at the 5 min aggregation, and the dynamics of the traffic pattern can be lost in this time window. With patch type 4, all local information about the current conditions on the link is lost. Roughly, the patch types can be categorised according to the definitions of Qu et al. (2009), with patch types 2–4 being missing completely at random, missing at random and missing not at random respectively. Table 2 shows the percentage of each patch type recorded on each of the links in the test network over the course of 2009.

On examining Table 2, it is evident that missing data is a serious issue on the LCAP network, the average percentage of patch type 1 is just 59%. Worryingly, the standard deviation of this figure is very high at 25.7%, indicating that some links have very high data quality and some have very low data quality. This is reflected in the average percentage of patch type 4; 23.9% with a standard deviation of 20.7%. However, missing data is less problematic if it is present at times where accurate forecasts are less important such as during the night.

It can be seen from Fig. 3 that the lowest capture rates are between the hours of around 8 pm and 7 am. The most likely reason for this is fewer cars being on the road at night time and, possibly, low light conditions affecting the cameras. Between 7 am and 8 pm, the average patch type is between 1.5 and 2. However, there is significant spatial variation in the level of patching; Fig. 4 shows the average patch types for eight links in the test network.

It can be seen that while some links such as link R524 and R2324 have average patch types during the daytime of less than 1.5, other links such as R2301 and R2007 have average patch types approaching 3. These abnormally low capture rates are caused by periods where sensors collected no data for a number of consecutive days and can be attributed to data that is “not missing at random” (Qu et al., 2009). Times where there are long term sensor failures are the most difficult to forecast because there is a complete lack of local temporal information. They almost certainly require a spatio-temporal approach to ensure an acceptable level of accuracy. It is these situations that we attempt to forecast in this study.

4.3. Experimental design

For the empirical example, only data between 6 am and 9 pm are used as the night time data was shown in the previous subsection to exhibit very high levels of patching. The data are converted to unit journey times (UJT) to provide a relative measure of link performance and are normalised to fall within the range [0, 1]. The 35 Tuesdays are separated into three sets; 25 days for training and validation and 10 days for testing. K-fold cross validation is used to train each of the models. The training data is divided into $k = 5$ equal sets and each one is used in turn as the validation set. Although this means that future information is used to forecast the past in the training phase, it is assumed that the traffic situation on consecutive Tuesdays is independent so this does not bias the model. The values of the parameters that were tested were $K = 1, 2, \ldots, 100$ and $\sigma = 0.01$, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1. The range of σ was determined using a rough initial search for parameter values. In each case, the embedding dimension of the series is set to 3. Once the models are trained, they are used to forecast the 10 days of testing data.

The high level of missing data complicates the experimental design because, in order to get results that can be validated, it is necessary to have unpatched data available for both the target link and its neighbouring links. Including patched data in the testing patterns would bias the results as they are often interpolated from future observations. To get the most comprehensive set of results possible, the models were tested on all situations where no patched data were present. Table 3 shows the number and percentage of forecasts that are possible for each link.

It can be seen that many of the links have a very low count of forecastable points. It is decided to remove those links and directions from the analysis where the percentage of forecastable points is below 5% (shaded grey in Table 3). This leads to links 432, 448, 1025, 1419, 1623 and 2055 being removed from the analysis completely. A consequence of the low capture rates is that few of the links have good data quality data in both the upstream and downstream directions, which makes it difficult to create combined forecasts. Therefore, it is decided to use upstream and downstream separately in order to create two forecasts for each link. Two separate embedded space–time series are created using separate weight matrices for the upstream and downstream directions, which are shown in Fig. 5.

4.4. Benchmark model

The performance of the spatio-temporal models is tested against the best available univariate method. As mentioned in Section 2, there are various univariate methods that have been used to model traffic data but these naturally rely on at least some local temporal information being available. In the absence of such information, the best estimate available is an average over the historical dataset. Therefore, the benchmark univariate model is the historical average (HA):

$$Y^{(s)}_{(t+\tau)} = \frac{1}{P}\sum_{i=1}^{P} Y^{(s)}_{(t-iq)+\tau} \tag{12}$$

where $i = 1, 2, \ldots, P$ is the number of days in the training data and q is the day length. The HA model is one of the simplest forecasting models available. Therefore, any more sophisticated model must outperform it as a minimum requirement. As HA does not take into account current spatial and temporal information, it will only perform well when conditions are close to average.

5. Experimental results

Table 4 shows the selected model parameters for each of the models in the upstream and downstream directions. The KNN based models tend to have values of k around 100. Fig. 6 shows the relationship between training error and k for the KNNavg model. A similar relationship can be observed on each of the other KNN based models.

Training error decreases rapidly as k increases to a value of around 20, after which it levels off. Although the minimum errors are generally found at $k \sim 100$, the performance increase from exceeding this number is minimal. In terms of the KR model, the optimal values of σ vary from 0.05 to 0.25. This reflects the differences in variability on each of the test links.

Table 5 shows the training performance of each of the models (the HA model is not shown here as it does not require training). It can be seen that the performance of each of the models on each link is generally quite similar, which is due to the fact that they are all based on the Euclidean distance metric and hence all use a similar measure of similarity. The relative training performance of upstream and downstream neighbours is also fairly similar in most cases, indicating that upstream and downstream are equally suitable for forecasting under missing data.

Table 6 shows the testing results for each of the models in the upstream and downstream directions. As was the case in the training process it can be seen that, for the most part, there is very little difference in performance between each of the pattern based models on each of the links. However, the pattern based methods outperform the HA method on all but three of the test links. The main reason for this is that they harness the local spatial information from the spatial neighbourhood in order to produce forecasts. This means they are able to capture the current conditions on the road network. We will now examine two situations in detail; (1) where the non-parametric models underperform; (2) where the non-parametric models outperform the benchmark model.

Fig. 7 shows the performance of the models on link 425. For presentation purposes, only the best performing pattern based method is shown, which is KNNdist downstream. On this link, the historical average model performed best. The reason for this is clear; the pattern of traffic changed little from week to week during the test period, with sharp morning and evening peaks observed with an intervening period of lower travel times. The pattern based models are able to model the general trend; however, they have a tendency to over forecast some of the peaks, for example on the AM peaks of the 4th and 6th days of the test period. The implication of this is that congestion occurred on link 2140 downstream of link 425 at these times but that this congestion remained spatially isolated.

This is contrary to traffic flow theory as congestion is expected to propagate from downstream to upstream. It is a significant finding because it provides insights into the dynamics of spatio-temporal autocorrelation on urban road networks. Urban roads are different to highways because there are various factors that disrupt the flow of traffic such as signals, pedestrian activity and stationary vehicles (loading and unloading, etc.). In addition to this many sensor networks, including the one used in this study, are spatially sparse meaning that there is a lot of uncertainty as to what is happening between two sensor locations (Qu et al., 2009), particularly if there are many entries and exits on a particular link. These factors mean that the strength of the correlation between locations is dynamic in time and heterogeneous in space, which was a finding of our previous study (Cheng et al., 2011). Sometimes congestion on a section of road will have a spatio-temporal effect and sometimes it will remain spatially isolated. Although it is beyond the scope of this study, consideration of which congestion events are likely to have a spatio-temporal effect and which are likely to remain isolated will be an interesting topic for future research.

Fig. 8 shows an example where the pattern matching models outperformed the HA model. On this link (link 1384), there was considerable variability in UJTs from week to week. On some weeks, UJT remained low throughout the day. However, there were 2 days in the test period that exhibited unusually high UJT levels during the AM peak period. These peaks can be seen as non-recurrent congestion, i.e. congestion caused by incidents, roadworks, special events and other unusual events. It is during these types of events that the availability of accurate current and predictive information is critical because they have the greatest impact on road users' journey times. The historical average model, by definition, is unable to capture any of this variability and systematically under or over forecasts the AM peak times. On the other hand, the pattern matching method successfully captures both of the large peaks. The implication of this result is that congestion was observed concurrently on link 1384 and its upstream neighbour, link 2140. It is likely that the direction of influence was from downstream to upstream in this case in accordance with traffic flow theory. Essentially, the influence of link 1384 on link 2140 is implicitly captured and used to forecast in the opposite direction. A consequence of this is that there is likely to be a temporal lag in the forecast values, which is evident on the 4th and 6th days. However, when downstream information is not available, as is the case here, this is an acceptable compromise.

For the most part, the non-parametric regression algorithms are also able to model the days where UJT is lower than average, although they over forecast UJT in the PM peak on some days, notably the 1st, 2nd and 3rd days. Whilst over estimation of UJTs may be less of a concern than underestimation in the context of real time forecasts, this type of error would affect the calculation of diagnostic statistics in the data that may be required offline for operational purposes and must be dealt with in future work. However, it should be noted that the HA model also overestimates the first 3 days of the testing set, indicating that link 1384 was experiencing a sustained period of lower than average UJTs over this time. It is possible that this was not reflected in the space of training patterns, which exposes one of the weaknesses of memory based approaches.

6. Conclusions

In this study, kernel regression was employed to forecast the future values of a space–time series using spatial neighbourhood information under the assumption of data that is missing not at random due to sensor failure. Furthermore, three types of KNN model were built for comparison purposes. Each of the methods significantly outperforms the historical average benchmark model in the majority of cases. They have the additional benefit of being able to forecast situations where the level of the series is higher than average, indicating strong performance in extreme conditions when accurate forecasts are more important. The methods all display similar performance, however, the KR, KNNdist and KNNavg models are desirable as they have only a single parameter to train.

The methods show promise for application to spatio-temporal datasets that exhibit high levels of missing data. One of the main advantages of pattern matching approaches is their ease of interpretation and implementation, which is a benefit over complicated statistical models and ‘‘black box’’ methods such as neural networks. However, they do require a database of training patterns to be stored and accessed from memory each time a forecast is made. It is essential that the methods presented here are scalable for implementation in real time on a network scale as the intention is to apply them to the entire LCAP network, which comprises around 1000 road links. They are well suited to this type of application for the following reasons: Firstly, the models are trained locally (one model per location), and as such applying them to a large network can be seen as an embarrassingly parallel problem (Harris & et al., 2010). Emerging parallel computing technologies such as general purpose graphics processing units (GPGPUs) and distributed computing resources can be used to achieve this goal. Secondly, the methods described here do not require a fixed model to be trained. Once the kernel bandwidth and/or number of neighbours have been determined, they respond to changes in the traffic state in the spatio-temporal neighbourhood and access patterns from the historical database accordingly. Therefore, the historical database is not required to remain fixed and the algorithms are all well suited to online application.

There are many ways in which this could be achieved, the simplest of which is to add each new test pattern to the historical dataset as it is encountered. However, this would lead to a continual growth in the size of the dataset and an unacceptable increase in computation time. To counter this, the size of the dataset could be fixed and the oldest pattern removed each time a new pattern is added in a sliding window approach. However, this has the drawback that some of the more useful patterns may be removed, while less interesting patterns remain. A more appropriate option may be to pass each new pattern through a filter, and retain it if it is novel and discard it otherwise. This process could also be used to minimise the database size in the model training stage.

A complementary approach that could be taken would be to examine where the K nearest neighbours are drawn from for each forecast point. It is likely that they will tend to be drawn from the training data at similar times of day, which would motivate the development of seasonal non-parametric regression models. Conversely, when dealing with non-recurrent events, the methods would benefit from access to a wider set of patterns that are not necessarily local in space but are representative of extreme conditions.

One of the main difficulties when trying to forecast missing data by any method is uncertainty about the results. The case study outlined here considers the simple case of using single upstream and downstream neighbours to forecast UJT on a road link. One deficiency of this approach is that it sometimes cannot distinguish between spatially isolated, short term events and permanent changes in the traffic state. This highlights the complicated spatio-temporal correlation structures that exist on the road network. A use of nonlinear correlation measures such as mutual information could help to untangle some of these relationships in order to improve forecasts.

Although the results are promising, there is scope for considering a larger spatial neighbourhood of upstream and downstream links. By taking into account conditions on the wider network, it may be possible to better estimate which events are localised and which will lead to measurable spatio-temporal effects. Additionally, the use of more spatial information may provide some insights into the uncertainty of forecasts. For instance, if one were to make forecasts using a number of spatial neighbours, the forecasts could be assessed for their individual consistency and combined in a multiple imputation setting. This could be further enhanced through the use of anisotropic kernels, whereby different kernel bandwidths are trained for each spatial neighbour.

Acknowledgements

This work was completed as part of the STANDARD project (EP/G023212/1), funded by EPSRC and in partnership with Transport for London. The authors would like to acknowledge the support of Andy Emmonds, Jonathan Turner and Alex Santacreu at TfL, as well as Benjamin Heydecker, Andy Chow, John Shawe-Taylor and the rest of the STANDARD team at UCL.

References

Amisigo, B. A. & Van De Giesen, N. C. (2005). Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series. <http://hal.archives-ouvertes.fr/hal-00304820/> Accessed 26.09.11.

Baltagi, P. B. (2005). *Econometric analysis of panel data* (3rd ed.). Wiley.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Cheng, T., Haworth, J., & Wang, J. (2011). Spatio-temporal autocorrelation of road network data. *Journal of Geographical Systems*. http://dx.doi.org/10.1007/s10109-011-0149-5.

Cristianini, N., & Shawe-Taylor, J. (2004). *Kernel methods for pattern analysis*. Cambridge, UK: Cambridge University Press.

Dougherty, M. (1995). A review of neural networks applied to transport. *Transportation Research Part C: Emerging Technologies, 3*(4), 247–260.

Dougherty, M. S., & Cobbett, M. R. (1997). Short-term inter-urban traffic forecasts using neural networks. *International Journal of Forecasting, 13*(1), 21–31.

Elhorst, J. P. (2003). Specification and estimation of spatial panel data models. *International Regional Science Review, 26*(3), 244–268.

Glasbey, C. A. (1995). Imputation of missing values in spatio-temporal solar radiation data. *Environmetrics, 6*(4), 363–371.

Griffith, D. A. (2010). Modeling spatio-temporal relationships: Retrospect and prospect. *Journal of Geographical Systems, 12*(2), 111–123.

Harris, R. et al. (2010). Grid-enabling geographically weighted regression: A case study of participation in higher education in england. *Transactions in GIS, 14*(1), 43–61.

Heuvelink, G. B. M., & Griffith, D. A. (2010). Space–time geostatistics for geography: A case study of radiation monitoring across parts of germany. *Geographical Analysis, 42*(2), 161–179.

Huang, B., Wu, B., & Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science, 24*(3), 383–401.

Kamarianakis, Y., & Prastacos, P. (2005). Space–time modeling of traffic flow. *Computers & Geosciences, 31*(2), 119–133.

Kanevski, M. (2008). *Advanced mapping of environmental data*. Wiley-ISTE.

Kanevski, M., Timonin, V., & Pozdnukhov, A. (2009). *Machine learning for spatial environmental data: Theory, applications, and software Har/Cdr*. EFPL Press.

Kanevski, M., & Maignan, M. (2004). *Analysis and modelling of spatial environmental data*. EPFL Press.

Karlaftis, M. G., & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies, 19*(3), 387–399.

Kokaram, A. C. et al. (1995). Interpolation of missing data in image sequences. *IEEE Transactions on Image Processing, 4*(11), 1509–1519.

Kyriakidis, P. C., & Journel, A. G. (1999). Geostatistical space–time models: A review. *Mathematical Geology, 31*(6), 651–684.

Lighthill, M. J., & Whitham, G. B. (1955). On kinematic waves. II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 229*(1178), 317–345.

Liu, Z., Sharma, S., & Datla, S. (2008). Imputation of missing traffic data during holiday periods. *Transportation Planning and Technology, 31*, 525–544.

Mitchell, T. (1997). *Machine learning*. Singapore: McGraw-Hill.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications, 9*(1), 141.

Ni, D. et al. (2005). A multiple imputation scheme for overcoming the missing values and variability issues in ITS data. *ASCE Journal of Transportation Engineering, 131*(12), 931–938.

Pfeifer, P. E., & Deutsch, S. J. (1980). A three-stage iterative procedure for space–time modelling. *Technometrics, 22*(1), 35–47.

Qu, Li. et al. (2009). PPCA-based missing data imputation for traffic flow volume: A systematical approach. *IEEE Transactions on Intelligent Transportation Systems, 10*, 512–522.

Queen, C. M., & Albers, C. J. (2008). Forecasting traffic flows in road networks: A graphical dynamic model approach. In *Proceedings of the 28th international symposium of forecasting*. International Institute of Forecasters.

Richards, P. I. (1956). Shock waves on the highway. *Operations Research, 4*(1), 42–51.

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate, 14*, 853–871.

Sharma, S., Lingras, P., & Zhong, M. (2004). Effect of missing values estimations on traffic parameters. *Transportation Planning and Technology, 27*, 119–144.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge University Press.

Smith, T. M. et al. (1996). Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *Journal of Climate, 9*, 1403–1420.

Smith, R. L., Kolenikov, S., & Cox, L. H. (2003). Spatio-temporal modeling of PM2.5 data with missing values. *Journal of Geophysical Research – Atmospheres, 128*, 10–1029.

Smith, B. L., Williams, B. M., & Keith Oswald, R. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies, 10*(4), 303–321.

Specht, D. F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks, 2*(6), 568–576.

Stathopoulos, A., & Karlaftis, M. G. (2003). A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C: Emerging Technologies, 11*(2), 121–135.

Transport for London (2010). *Travel in London: Report 3*. <http://www.tfl.gov.uk/assets/downloads/corporate/travel-in-london-report-3.pdf> Accessed 30.08.11.

van Lint, J. W. C., Hoogendoorn, S. P., & van Zuylen, H. J. (2005). Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C: Emerging Technologies, 13*(5–6), 347–369.

Vlahogianni, E. I., Golias, J. C., & Karlaftis, M. G. (2004). Short-term traffic forecasting: Overview of objectives and methods. *Transport Reviews: A Transnational Transdisciplinary Journal, 24*(5), 533.

Wang, J., Zou, N., & Chang, G.-L. (2008). Travel time prediction: Empirical analysis of missing data issues for advanced traveler information system applications. *Transportation Research Record: Journal of the Transportation Research Board, 2049*, 81–91.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A, 26*(4), 359–372.

Whitlock, M. E., & Queen, C. M. (2000). Modelling a traffic network with missing data. *Journal of Forecasting, 19*(7), 561–574.

Williams, B., Durvasula, P., & Brown, D. (1998). Urban freeway traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing models. *Transportation Research Record, 1644*(1), 132–141.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.

Zhong, M., Lingras, P., & Sharma, S. (2004b). Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. *Transportation Research Part C: Emerging Technologies, 12*(2), 139–166.

Zhong, M., Sharma, S., & Lingras, P. (2004a). Genetically designed models for accurate imputation of missing traffic counts. *Transportation Research Record: Journal of the Transportation Research Board, 1879*(1), 71–79.

Zhong, M., Sharma, S., & Lingras, P. (2006). Matching patterns for updating missing values of traffic counts. *Transportation Planning and Technology, 29*, 141–156.