Improving forecasting under missing data on sparse spatial networks

J. Haworth[1], T. Cheng[1], E. J. Manley[1]

[1]SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, Chadwick Building, Gower Street, London WC1E 6BT
Telephone: +44
Email: j.haworth@ucl.ac.uk, t.cheng@ucl.ac.uk, edward.manley.09@ucl.ac.uk

1. Introduction

Missing data is a major issue in many real world sensor networks. It can complicate the calculation of diagnostic statistics in an offline setting, as well as making prediction of processes difficult in a real time setting. The longer the period of missing data, the more difficult it is to mitigate its effects. In spatial sensor networks, data from neighbouring locations can be used to impute or forecast missing values. Efforts have been made to deal with missing spatio-temporal data in environmental monitoring (Glasbey, 1995; Smith et al, 1996, 2003) and traffic forecasting (Whitlock and Queen, 2001; van Lint et al, 2005; Haworth and Cheng, 2012) amongst others.

Crucial to the use of spatio-temporal approaches for dealing with missing data is correctly capturing the dependency structure between locations. In networks of flows this can be implicit in the network structure: for instance, on road networks traffic flows from upstream to downstream in free flowing conditions, and queues build up in the opposite direction in congested conditions. Although this relationship is complicated by traffic signals in the urban environment, it can still be captured using physical models if flow data are available with sufficient spatial and temporal resolution. However, on real world sensor networks, the data collection system often does not capture a sufficient level of spatial and/or temporal granularity, or indeed the right type of data, to model the physical process in detail. In this study, we extend our previous work (Haworth and Cheng, 2012) to improve forecasting of road link travel times under missing data. We combine two networks; the data collection network and the underlying road network. The structure of the road network is used to improve spatio-temporal forecasting of missing data on the sensor network, which is spatially sparse.

2. Data and Methods

2.1 The sensor network and the road network

The sensor network is a system of Automatic number plate recognition (ANPR) cameras that record travel times, as part of Transport for London's (TfL) London Congestion Analysis Project (LCAP). Automatic number plate recognition (ANPR) technology is used to record the elapsed time between vehicles entering and exiting a road link, and the individual observations are averaged at 5 minute intervals. Although the LCAP network is a fully connected network, it is spatially sparse, consisting only of major roads that are

of strategic importance to TfL. Therefore, many of the spatial connections between LCAP links are ignored in its network structure.

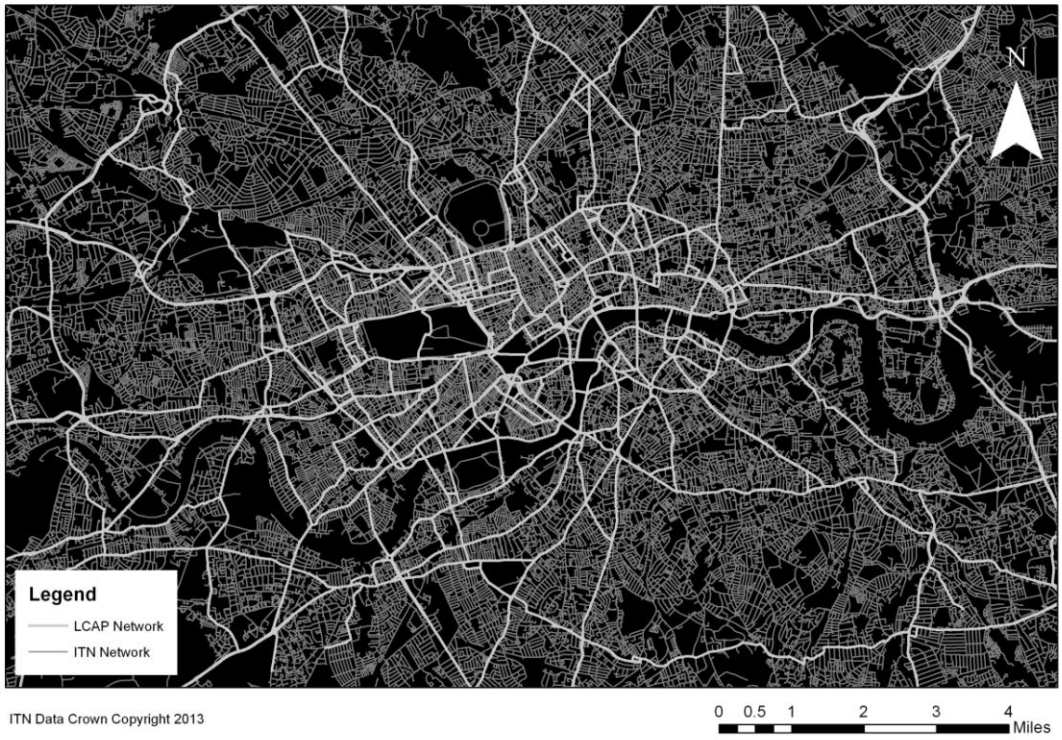

Figure 1. Connecting the LCAP and ITN Networks

A more complete representation of London's road network is provided by Ordnance Survey's Integrated Transport Network (ITN), upon which the LCAP network lies. The two networks are shown in Fig. 1, illustrating the sparse coverage of the LCAP network. A comparison of their characteristics is provided in Table 1. In this study, we use the ITN to build up a more complete network structure that captures the possible flows between LCAP links in order to improve spatio-temporal forecasting. This is described in the following section.

Variable	LCAP	ITN
No. Links	~1500	500,000+
Avg. Link length (m)	2700	97
Total Length (km)	~3800	60,000+

Table 1. Comparison of LCAP and ITN in the London area

2.2 Methodology

First, the ITN links that intersect with each of the LCAP links are identified. Following this, the shortest paths are calculated between each of the LCAP links on the ITN network, to provide a measure of network proximity. This is shown in Fig 2, and is described below.

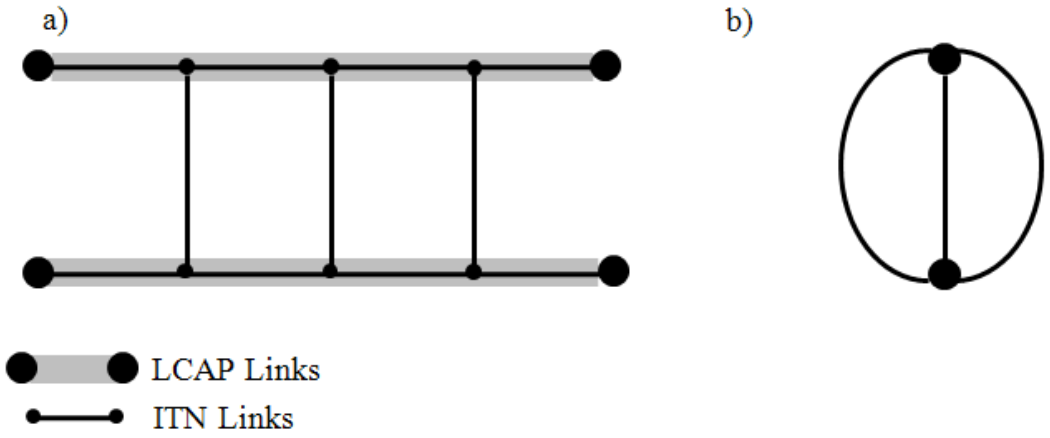

Figure 2. Connecting the LCAP and ITN Networks

Fig. 2 a) shows an example of the typical relationship between LCAP and ITN links. Each LCAP link consists of several ITN links, and there are various paths between LCAP links on the ITN network. Fig. 2 b) depicts the approach we take here graphically. The ITN links associated with an LCAP link are viewed as a single node, with the ITN network forming edges between them. The arcs in the diagram represent the ITN links in Fig. 2 a). Shortest paths are calculated along the ITN network between all LCAP links using the igraph library in R statistical package. An N*N spatial weight matrix W is then created from these paths, where N is the number of LCAP links. The i,j elements of W contain the average shortest path distance between the ITN links on LCAP links i and j.

3. Experimental setup

To train the models we use the kernel regression model from Haworth and Cheng, 2012 (Eq. 6, p. 543). For brevity, the equations are not repeated here. For each LCAP link L_i, a neighbour set is constructed consisting of those LCAP links L_j that are within a specified network distance $d_{i,j} \leq s$. The spatial weight matrix W is converted into a binary adjacency matrix, where $i,j = 1\ if\ d_{i,j} \leq s$, 0 otherwise. When making the forecasts, we assume *no data is being collected at the current sensor location*, therefore $ii = 0$. In total, the travel times of 108 LCAP links are forecast. The size of the training set is 25 days, and k-fold cross validation is used as the training method, with k=5. The selected model is the one which minimises the root mean squared error criterion. A further 10

days, immediately following the training period, are used as a validation set to evaluate the performance of the fitted models.

3.1 Comparison Model

For comparison purposes, a further model is created based on the connectivity structure of the LCAP network. For each LCAP link, its directly adjacent upstream and downstream neighbours are used to forecast its values. The experimental setup is identical to that outlined above.

4. Initial Results

For the new model, the average RMSE of training across the 108 test links was 94.7 seconds. This rises to 101.2 seconds for the validation data. The RMSE of training and testing for the comparison model was 96.04 and 98.5 seconds respectively. The new model performs slightly better than the comparison model in terms of training, but slightly worse in terms of forecasting, indicating that the LCAP connectivity structure alone is a good predictor of network conditions overall. However, in terms of the performance of the models on individual links, the new model performed better than the comparison model on 46 (42%) of the 108 links, with 4 having exactly equal performance. This indicates that, despite the comparison model having better forecasting performance overall, performance can be improved on a significant number of the LCAP links by incorporating the ITN network structure. An extreme example of this is given in Fig. 3.

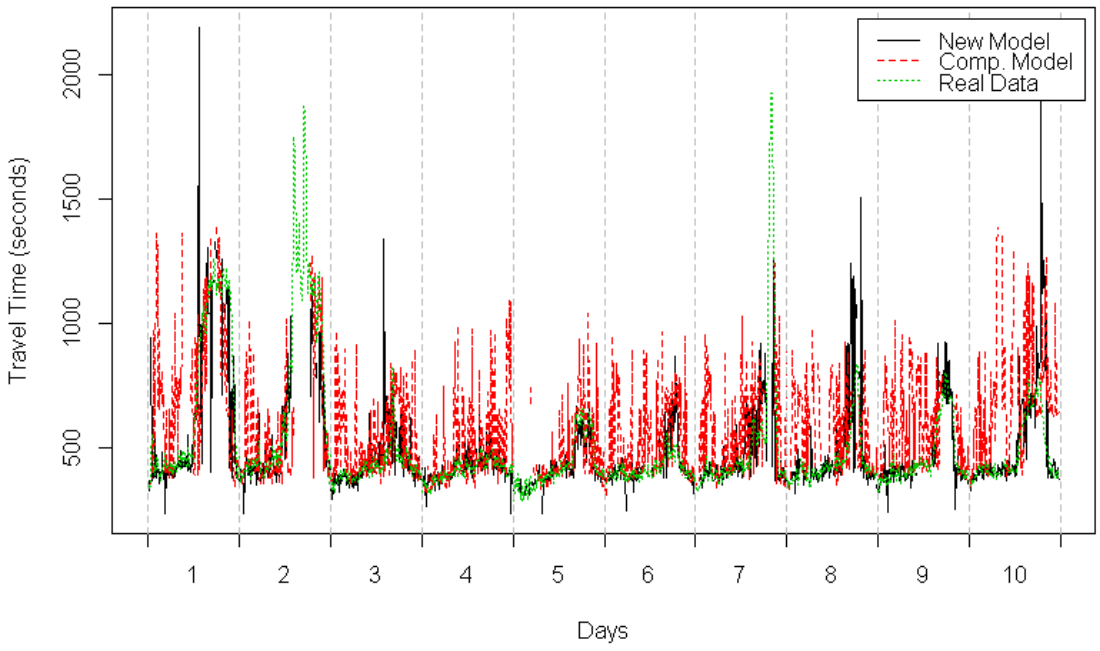

Figure 3. Example of model performance

5. Conclusions

The approach used here demonstrates that spatio-temporal forecasts on sparse sensor networks can be improved by incorporating knowledge of the underlying network structure. In this example, network shortest paths were used to build up spatial neighbourhoods of road links, which were used for forecasting under the assumption of missing data. This study opens up a number of avenues for continuing research. Firstly, it is necessary to examine the characteristics of LCAP links that lead to one model being favoured over the other. It can be speculated that centrally located links, where the network is denser, will see a greater performance increase from the inclusion of the ITN network. Secondly, although shortest path lengths were used to define the connectivity structure in this study, other measures could be used. For instance, the paths could be weighted using other traffic variables such as historical travel times. Furthermore, the number of; or strength of, connections between links on the sensor network could be used a proxy for their degree of connectedness. A further consideration is the weighting of each of the members of the neighbour set. In this study they are weighted equally, but an inverse distance weighting could improve results further.

6. References

Haworth, J & Cheng, T, 2012. Non-parametric regression for space–time forecasting under missing data. *Computers, Environment and Urban Systems*, 36(6): 538–550.

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

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.

Smith, T M, Reynolds, R W, Livezey, R E & Stokes, D C. 1996. Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *Journal of Climat,e* 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.

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.