Urban Network Travel Time Prediction Based on a Probabilistic Principal Component Analysis Model of Probe Data

Preprint of paper in
IEEE Transactions on Intelligent Transportation Systems
DOI: 10.1109/TITS.2017.2703652

Erik Jenelius and Haris N. Koutsopoulos

Department of Transport Science,
KTH Royal Institute of Technology,
SE-100 44 Stockholm, Sweden
(e-mail: jenelius@kth.se)

Department of Civil and Environmental Engineering,
Northeastern University,
Boston, MA 02115
(e-mail: h.koutsopoulos@neu.edu)

June 14, 2017

Abstract

The paper proposes a network travel time prediction methodology based on probe data. The model is intended as a tool for traffic management, trip planning and online vehicle routing, and is designed to be efficient and scalable in calibration and real-time prediction, flexible to changes in network, data and model extensions, and robust against noisy and missing data. A multivariate probabilistic principal component analysis (PPCA) model is proposed. Spatio-temporal correlations are inferred from historical data based on MLE and an efficient EM algorithm for handling missing data. Prediction is performed in real-time by computing the expected distribution of link travel times in future time intervals, conditional on recent current-day observations. A generalization of the methodology partitions the network and applies a distinct PPCA model to each subnetwork. The methodology is applied to the network of downtown Shenzhen, China, using taxi probe data. The model captures variability over months and weekdays as well as other factors. Prediction

with PPCA outperforms k-nearest neighbors prediction for horizons from 15 to 45 minutes, and a hybrid method of PPCA and local smoothing provides the highest accuracy.

1 Introduction

Short-term traffic prediction, referring to horizons from a few minutes up to around 60 minutes depending on the network, is a crucial component of intelligent transportation systems (ITS), particularly for traffic control, information provision and trip planning. The spatio-temporal scope of the prediction varies depending on the application; for traffic control, focus may be on motorways or major signalized arterials. For user-centered information provision and trip planning, prediction at the network level gives the highest flexibility in terms of routes and origin-destination pairs.

Until now, most research effort has focused on motorways and freeways, employing univariate statisticial models to predict traffic volume based on data from single location sources [1, 2]. Recently, the emergence of new technologies like Automatic Vehicle Identification (AVI) systems and GPS have facilitated the collection of travel time data for network segments and routes, as well as wider range of applications [3, 4]. These developments have led to more research on travel time prediction, although still focusing mainly on motorways. A number of remaining challenges in traffic prediction have been identified [2], including

1. moving from motorways and single arterials to the network level
2. capturing the spatial and temporal patterns of traffic
3. incorporating new data sources such as probe data
4. properly handling missing data

Meeting the first challenge requires the prediction method to be efficient and scalable. Several univariate prediction methods for motorway or arterial travel times based on probe or AVI data have been proposed, some using statistical approaches [5, 6, 7], others using artificial neural networks [8, 9]. However, such an approach tends to be demanding in terms of calibration, maintenance and computations for large networks [10]. In contrast, multivariate models have the potential to be scalable and feasible even for large-scale networks.

Several papers have proposed network-level methods for travel time estimation on probe data, e.g., [11, 12], but fewer papers address the challenges of short-term prediction in this context. Proposed methods include spatio-temporal autoregressive models [13] and Bayesian network models [14], [15], which specify rigid model structures for the correlations between links and time intervals, and spatio-temporal k-nearest neighbors models [16], which use data-driven approaches based on similarities with historical data. In general, however,

the ability of the models to handle few or missing data in a consistent way have not been much discussed.

Methods utilizing mesoscopic traffic simulation models for traffic prediction in general networks, known as Dynamic Traffic Assignment (DTA) models, have also been developed [17, 18]. These models are behaviorally rich and can capture the interrupted flows and route choices of arterial and mixed networks, including driver response to information under incident conditions. On the other hand, they are also complex and present challenges from a calibration and application point of view. Another line of research uses data fusion methods such as Kalman filtering to combine real-time traffic data with simulation models. For motorways, macroscopic flow models are well suited for fusion with probe vehicle and fixed sensor data [19, 20].

This paper aims to address the challenges raised by [2] by proposing a multivariate network-level travel time prediction methodology based on probe data, which handles missing data in a consistent way. The model is intended as a tool for trip planning, traffic management and online vehicle routing, and to be an efficient complement and baseline to DTA-based prediction approaches. The paper takes elements from both model-based and data-driven approaches and proposes a method in which neither network structure nor temporal progression are explicitly modeled, but the relevant spatio-temporal patterns are inferred from a long series of historical data. In addition to prediction accuracy, the model is designed with several important properties for large-scale applications in mind:

- *Efficiency:* The calibration of parameters and the real-time application of the prediction model is computationally efficient and scalable.
- *Flexibility:* The model is flexible to changes in the network and data to which it is applied. It is also flexible to extensions of the model itself.
- *Robustness:* The model is able to handle missing data and measurement noise in both historical and real-time data and still produce robust predictions.

Link travel times over multiple time intervals are modeled probabilistically as projections from latent random variables in a lower dimensional space, and each day represents a realization of the latent variables. Each loading represents a different dimension of variability and together create a spatio-temporal correlation structure among the observed variables. Specifically, link travel times are assumed to be generated by a probabilistic principal component analysis (PPCA) model [21]. The model parameters are estimated on historical probe data with missing data handled consistently using an efficient expectation-maximization (EM) algorithm [22].

PPCA has been used for imputing missing values in traffic volume data from single fixed detectors [23], but the application to network-wide travel time prediction is novel. Prediction is performed by computing the distribution of link travel times in future time intervals conditional on observed values in past

time intervals the current day. The use of the PPCA model for prediction is equivalent to the supervised PPCA (SPPCA) model [24] and the probabilistic partial least squares regression (PPLSR) model [25].

The prediction methodology is applied to the network of downtown Shenzhen, China, using four months of link travel time observations from taxis probes. Prediction accuracy is compared with k-nearest neighbors prediction and a univariate smoothing method. A hybrid of PPCA and local smoothing is also considered, as well as separate PPCA models for different functional classes. The robustness of the prediction methods against missing values is assessed, and the spatio-temporal correlation structure identified by the PPCA model is studied.

The remainder of the paper is organized as follows. The prediction methodology is described in Section 2, Section 3 describes the application to Shenzhen, China, with results presented in Section 4. Section 5 concludes the paper.

2 Methodology

2.1 Definitions

Let t_{ikn} be the average travel time of link k in time-of-day interval i on day n based on observations from probe data, and let d_k be the length of link k. Then $v_{ikn} = d_k/t_{ikn}$ is the mean speed of link k according to Edie's definition [26], assuming time intervals are sufficiently long in relation to the probe sampling frequency. The methodology allows a one-to-one transformation $u_{ikn} = f(v_{ikn})$ of the mean speeds to be used as variables. Outputs $\hat{u}_{ikn}$ of the prediction method are transformed to predicted mean speeds by using the inverse transformation $\hat{v}_{ikn} = f^{-1}(\hat{u}_{ikn})$, and to predicted travel times by the transformation $\hat{t}_{ikn} = d_k/f^{-1}(\hat{u}_{ikn})$. The transformed variables for all links for a given time interval i and day n are collected in the $K \times 1$ column vector $\mathbf{u}_{in}$.

In particular, the case study in Section 3 shows that the logarithm of the mean speeds leads to lower prediction errors than using the original mean speeds. Hence, the logarithmic transformation $u_{ikn} = \log(v_{ikn})$ is used, and predicted speeds are recovered as $\hat{v}_{ikn} = \exp(\hat{u}_{ikn})$.

The prediction horizon is specified by a parameter $F \geq 1$. Prediction is based on current-day observations from the $P \geq 1$ most recent intervals. Assuming that the current time interval is h, link speeds are thus predicted for time intervals $\{h+1, \ldots, h+F\}$, and predictions are based on observations from interval $h-P+1$ to interval h. Here, $h-P+1$ and $h+F$ should be interpreted modulo I, i.e., the intervals may extend past midnight to the previous and the next day, respectively. The number of future intervals F is an input parameter to the analysis, while P is a model parameter that should be calibrated to minimize prediction error.

Historical observations from all time intervals from $h-P+1$ to $h+F$ are used to build a model of link travel times. Let $D = K(P+F)$ denote the total number of link/time-of-day interval combinations. The observations for day n

are stacked in the $D \times 1$ vector $\mathbf{u}_n$,

$$\mathbf{u}_n = \begin{pmatrix} \mathbf{u}_{h-P+1,n} \\ \vdots \\ \mathbf{u}_{h+F,n} \end{pmatrix} \tag{1}$$

2.2 Network Travel Time Model

The observations $\mathbf{u}_n$ are assumed to be generated from the PPCA model [21],

$$\mathbf{u}_n = W\mathbf{x}_n + \boldsymbol{\mu} + \boldsymbol{\epsilon}_n, \quad n \in \{1, \ldots, N\} \tag{2}$$

$\mathbf{x}_n$ is a $Q \times 1$ column vector of latent random variables, assumed to be i.i.d. Normal with mean zero and variance one,

$$\mathbf{x}_n \sim \mathcal{N}(0, I), \quad n \in \{1, \ldots, N\} \tag{3}$$

Q is a model parameter that represents the dimensionality of the latent feature space. W is a $D \times Q$ parameter matrix that represents the linear mapping between latent space and the observed variables. $\boldsymbol{\mu}$ is a $D \times 1$ parameter column vector that represents the mean value for each variable, i.e., for each link and time-of-day interval.

$\boldsymbol{\epsilon}_n$ is a $D \times 1$ column vector of random errors, assumed to be i.i.d. Normal with mean zero and variance σ^2,

$$\boldsymbol{\epsilon}_n \sim \mathcal{N}(0, \sigma^2 I), \quad n \in \{1, \ldots, N\} \tag{4}$$

The errors represent random variability in the observations that is not generated from the latent structure, but arises from sources such as variability between vehicles, measurement noise etc.

The larger the number of dimensions Q, the more of the observed variability in the historical data is preserved. However, for moderate-sized to large networks and a limited historical data set, there is a risk of over-fitting the model at the expense of prediction accuracy. Thus, Q is calibrated by cross-validation.

The model asserts that the variables $\mathbf{u}_n$ are Normal. If the logarithm of link speed is used as model variable, as discussed in Section 2.1, it implies that link speeds and travel times are distributed log-Normal, which has been found to be a good approximation in many empirical studies [27].

Note that there are no explicit temporal and spatial components in (2). Instead, the model assumes a general correlation structure for the observations, and extracting the spatial and temporal components is part of the estimation of the model.

ML estimators of $\boldsymbol{\mu}$, W and σ^2 with complete data for a fixed Q have been derived [21]. The ML estimation of W corresponds essentially to a PCA of the observations. Because of the connection with PCA and factor analysis, the columns of W are often called factor loadings. However, while PCA is a computational method to analyze and reduce the dimensions of variance in a

given set of data, PPCA is a generative model that describes how observations are generated from a lower-dimensional space of random variables.

An efficient EM algorithm for estimating $\boldsymbol{\mu}$, W and σ^2 in the presence of missing values has been presented [22]. The algorithm is shown in Table 1, where $u_{jn,o}$ equals u_{jn} if the value is observed and 0 otherwise, $N_{j,o}$ denotes the number of observed values among $(u_{j,1}, \ldots, u_{j,N})$, N_h is the total number of missing values in the observation set, $\mathbf{e}_{n,h}$ denotes the subvector of missing values of $\mathbf{e}_n = \mathbf{u}_n - \boldsymbol{\mu}$, and δ is a threshold for iteration convergence.

2.3 Network-Wide Prediction

Let $\boldsymbol{\omega}_i$ be the $K \times 1$ vector of variables for all links in time interval i from the current day. All time intervals are stacked in a $D \times 1$ vector $\boldsymbol{\omega}$. It is assumed that the variables $\boldsymbol{\omega}$ from the current day are generated from the same model (2) as the historical data. The marginal distribution of $\boldsymbol{\omega}$ is thus

$$\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu}, WW^T + \sigma^2 I) \tag{5}$$

At time interval h only the variables from time intervals $\{h - P + 1, \ldots, h\}$ are observed, while the variables in time intervals $\{h + 1, \ldots, h + F\}$ are to be predicted. The vector $\boldsymbol{\omega}$ can thus be split into the observed values $\boldsymbol{\omega}_{\mathrm{P}}$ and the future values $\boldsymbol{\omega}_{\mathrm{F}}$,

$$\boldsymbol{\omega}_{\mathrm{P}} = \begin{pmatrix} \boldsymbol{\omega}_{h-P+1} \\ \vdots \\ \boldsymbol{\omega}_h \end{pmatrix} \quad \boldsymbol{\omega}_{\mathrm{F}} = \begin{pmatrix} \boldsymbol{\omega}_{h+1} \\ \vdots \\ \boldsymbol{\omega}_{h+F} \end{pmatrix} \tag{6}$$

The loading matrix W can be similarly split into the $KP \times Q$ matrix W_{P} and the $KF \times Q$ matrix W_{F}, and $\boldsymbol{\mu}$ into $\boldsymbol{\mu}_{\mathrm{P}}$ and $\boldsymbol{\mu}_{\mathrm{F}}$.

The distribution of the future variables $\boldsymbol{\omega}_{\mathrm{F}}$ conditional on the past observed values $\boldsymbol{\omega}_{\mathrm{P}}$ can now be derived. From (5)–(6), the properties of multivariate normal distributed variables and the matrix inversion lemma, it follows that $\boldsymbol{\omega}_{\mathrm{F}} \,|\, \boldsymbol{\omega}_{\mathrm{P}} \sim \mathcal{N}(\hat{\boldsymbol{\omega}}_{\mathrm{F|P}}, \Sigma_{\mathrm{F|P}})$, where

$$\hat{\boldsymbol{\omega}}_{\mathrm{F|P}} = \boldsymbol{\mu}_{\mathrm{F}} + W_{\mathrm{F}}(W_{\mathrm{P}}^T W_{\mathrm{P}} + \sigma^2 I)^{-1} W_{\mathrm{P}}^T (\boldsymbol{\omega}_{\mathrm{P}} - \boldsymbol{\mu}_{\mathrm{P}}) \tag{7}$$

$$\Sigma_{\mathrm{F|P}} = \sigma^2 W_{\mathrm{F}} \left(W_{\mathrm{P}}^T W_{\mathrm{P}} + \sigma^2 I \right)^{-1} W_{\mathrm{F}}^T + \sigma^2 I \tag{8}$$

The mean vector $\hat{\boldsymbol{\omega}}_{\mathrm{F|P}}$ represents the point predictor of the future variables, while the covariance matrix $\Sigma_{\mathrm{F|P}}$ provides information about the variability around the point predictions. Note that the historical means $\boldsymbol{\mu}_{\mathrm{F}}$ serve as baseline for the prediction, and the factor loadings are applied to the residuals between observations and historical means in past intervals, $\boldsymbol{\omega}_{\mathrm{P}} - \boldsymbol{\mu}_{\mathrm{P}}$. From the way input data are ordered, the predictors for time intervals $h + 1, \ldots, h + F$ are stacked as

$$\hat{\boldsymbol{\omega}}_{\mathrm{F|P}} = \begin{pmatrix} \hat{\boldsymbol{\omega}}_{h+1|P} \\ \vdots \\ \hat{\boldsymbol{\omega}}_{h+F|P} \end{pmatrix} \tag{9}$$

Table 1: EM algorithm for estimating PPCA model parameters with missing values (adapted from [22]).

Estimation of $\boldsymbol{\mu}$:

$$\mu_j \leftarrow \frac{1}{N_{j,o}} \sum_{n=1}^{N} u_{jn,o}, \quad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)^T$$

Initialization:

$$\Psi_{\text{old}} \leftarrow \infty$$

$$\mathbf{e}_n \leftarrow \mathbf{u}_n - \boldsymbol{\mu}, \quad \mathbf{e}_{n,h} \leftarrow 0, \quad E = (\mathbf{e}_1, \ldots, \mathbf{e}_N)$$

$$W_{jq} \leftarrow \mathcal{N}(0,1), \quad W = (W_{jq})_{j=1,\ldots,D}^{q=1,\ldots,Q}$$

$$X \leftarrow (W^T W)^{-1} W^T E, \quad X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$$

$$\hat{\mathbf{e}}_n \leftarrow W\mathbf{x}_n, \quad \hat{\mathbf{e}}_{n,h} \leftarrow 0$$

$$\sigma^2 \leftarrow \frac{1}{ND - N_h} \sum_{n=1}^{N} ||\mathbf{e}_n - \hat{\mathbf{e}}_n||^2$$

E step:

$$M \leftarrow \sigma^2 (W^T W + \sigma^2 I)^{-1}$$

$$\sigma^2_{\text{old}} \leftarrow \sigma^2$$

$$\mathbf{e}_{n,h} \leftarrow W\mathbf{x}_n, \quad E = (\mathbf{e}_1, \ldots, \mathbf{e}_N)$$

$$X \leftarrow MW^T E/\sigma^2, \quad X = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$$

M step:

$$W \leftarrow EX^T (X^T X + NM)^{-1}$$

$$\sigma^2 \leftarrow \frac{1}{ND} \left(\sum_{n=1}^{N} ||\mathrm{e}_n - W\mathrm{x}_n||^2 + N\mathrm{Tr}(WMW^T) + N_h \sigma^2_{\text{old}} \right)$$

Stopping criterion:

$$\Psi \leftarrow ND + N\left(D\log(\sigma^2) + \mathrm{Tr}(M) - \log|M|\right) + \mathrm{Tr}(X^T X) - N_h \log(\sigma^2_{\text{old}})$$

If $1 - \frac{\Psi}{\Psi_{\text{old}}} < \delta$ stop, otherwise $\Psi_{\text{old}} \leftarrow \Psi$, and repeat E and M steps.

If the model variables are a transformation of link travel times (e.g., the logarithm of link speed), the inverse transformation needs to be applied to the point and interval predictions obtained from (7)–(8) as discussed in Section 2.1. If the observations $\boldsymbol{\omega}_{\mathrm{P}}$ contain missing values, the prediction is carried out in a straightforward way by removing the corresponding rows of $\boldsymbol{\omega}_{\mathrm{P}}$, W_{P} and $\boldsymbol{\mu}_{\mathrm{P}}$ before applying equations (7)–(8). Travel times are thus consistently predicted also for links missing some or all current-day observations based on the historical means and the latent variable correlation structure.

2.4 Subnetwork Prediction

In the basic form of the prediction methodology, a single PPCA model is used for the full network. A natural generalization of the methodology is to partition the network and associate a distinct PPCA model with each subnetwork. The partitioning can be based on functional classes, spatial regions, network topology, etc., in order to capture relevant correlation patterns for each particular subnetwork and largely independent variation between subnetworks. There is a trade-off regarding the size of the subnetworks, however, since correlation between subnetworks is neglected and the number of available observations for each subnetwork is reduced.

2.5 Hybrid PPCA and Local Prediction

The PPCA-based prediction extracts (sub)network-wide correlation patterns from historical data to predict future travel times for all links. A different paradigm is to consider each link in isolation and predict travel times using only local information from past intervals. A hypothesis is that a combination of PPCA-based network-wide prediction and local prediction may give lower prediction error than each method individually. For the local prediction a simple scheme is considered based on a smoothed average of deviations between observed link speeds and historical means; for $i \in \{h+1, \ldots, h+F\}$,

$$\begin{aligned} \hat{\boldsymbol{\omega}}_i^{\mathrm{local}} = \boldsymbol{\mu}_i &+ \alpha^{i-h}(\boldsymbol{\omega}_h - \boldsymbol{\mu}_h) + \ldots \\ &+ \alpha^{i-h+P-1}(\boldsymbol{\omega}_{h-P+1} - \boldsymbol{\mu}_{h-P+1}) \end{aligned} \tag{10}$$

where $\alpha \in [0, 1)$ is a weight parameter. The weights decay geometrically according to the time difference between the prediction interval and the past interval.

The PPCA model can be used not only to predict future time intervals but also to estimate the mean values in past intervals conditional on the current-day observations. Thus, the conditional mean values $\hat{\boldsymbol{\omega}}_{\mathrm{P|P}}$ are estimated as

$$\hat{\boldsymbol{\omega}}_{\mathrm{P|P}} = \boldsymbol{\mu}_{\mathrm{P}} + W_{\mathrm{P}}(W_{\mathrm{P}}^T W_{\mathrm{P}} + \sigma^2 I)^{-1} W_{\mathrm{P}}^T (\boldsymbol{\omega}_{\mathrm{P}} - \boldsymbol{\mu}_{\mathrm{P}}) \tag{11}$$

where the intervals $\{h-P+1, \ldots, h\}$ are stacked as

$$\hat{\boldsymbol{\omega}}_{\mathrm{P|P}} = \begin{pmatrix} \hat{\boldsymbol{\omega}}_{h-P+1|P} \\ \vdots \\ \hat{\boldsymbol{\omega}}_{h|P} \end{pmatrix} \tag{12}$$

The network-wide and local prediction models are combined by replacing the historical mean $\boldsymbol{\mu}_i$ in (10) for every time interval i with the PPCA conditional mean, $\hat{\boldsymbol{\omega}}_{i|\mathrm{P}}$. Thus, the local smoothing is performed over the deviations from the PPCA means conditional on current-day observations rather than the historical means. The hybrid network-wide and local predictor for future time interval $i \in \{h+1, \ldots, h+F\}$ is

$$\begin{aligned} \hat{\boldsymbol{\omega}}_i^{\mathrm{hybrid}} = \hat{\boldsymbol{\omega}}_{i|\mathrm{P}} &+ \alpha^{i-h}(\boldsymbol{\omega}_h - \hat{\boldsymbol{\omega}}_{h|\mathrm{P}}) + \ldots \\ &+ \alpha^{i-h+P-1}(\boldsymbol{\omega}_{h-P+1} - \hat{\boldsymbol{\omega}}_{h-P+1|\mathrm{P}}) \end{aligned} \tag{13}$$

Pure PPCA prediction is recovered by setting $\alpha = 0$.

2.6 Computational Aspects

In practical applications, model parameters P, Q, W, $\boldsymbol{\mu}$, σ^2 and α would be updated relatively infrequently over a sliding time window based on historical data, where older observations may be assigned lower weight; the appropriate window length and updating frequency would depend on the dynamics of the particular network. First, the desired prediction horizon F is set. For each time-of-day interval, parameters P, Q and α are then calibrated using cross-validation on the historical data. When the best parameter values have been selected, the historical mean $\boldsymbol{\mu}$, observation noise σ^2 and factor loadings W are estimated on the historical data. The parameter values for each time-of-day interval are stored in a database.

Prediction is carried out using current-day observations from the P most recent time intervals. Prediction is a computationally very efficient operation given that all model parameters are retrieved from the database. Computing the matrix $(W_{\mathrm{P}}^T W_{\mathrm{P}} + \sigma^2 I)^{-1}$ in (7), (8) and (12), which has dimensions $Q \times Q$ independently of the number of links K or time intervals P, is the most expensive operation computationally. Since Q is typically chosen much smaller than KP, the inversion of the matrix is computationally affordable.

3 Case Study

The prediction methodology is applied to the road network of downtown Shenzhen, China. The data contain many link attributes including length, speed limit and functional class. In this study, road segments of functional class 2 (urban expressways), 3 (main arterial roads) and 4 (minor arterial roads) are used. The network consists of 966 links covering an area of approximately 25 km^2. Figure 1 shows the network and the functional class of each link.

Link speeds based on probe data from ca. 40,000 taxis are available in 15-minute intervals. The taxis record current time, GPS coordinates and occupancy status at random intervals. The average speed of a link during a particular time interval is estimated as the ratio of the total distance travelled and the total time spent on the link by occupied taxis during the time interval according to

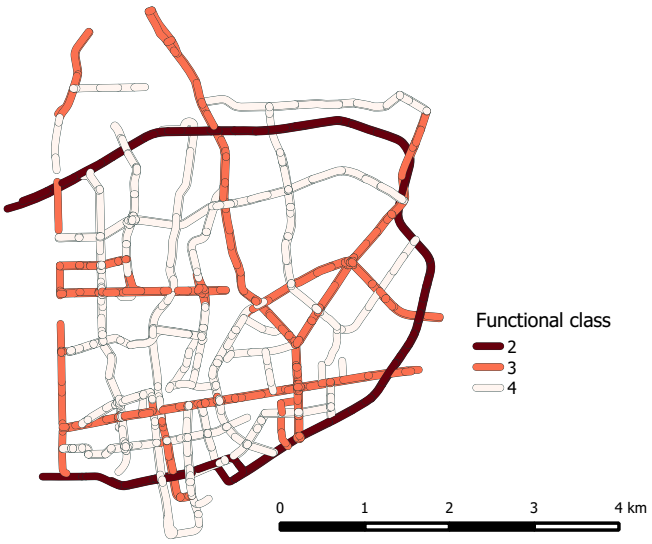

Figure 1: Shenzhen network.

Edie's definition [26]. Links and intervals without probe records are treated as missing values. The case study makes use of four months of data, from 1 March to 30 June 2014. Saturdays, Sundays and three holidays, as well as five days that were identified as anomalous (possibly due to incidents, construction work or special events), were removed from the data set to avoid the historical mean having an unfair disadvantage compared to the other prediction methods. In the end, data from 69 days were used. Of these, 10 days were randomly selected as the test set while the remaining 59 days were used as training set.

Figure 3 shows the variation of link speeds across the day. There is evidence of severe congestion. The average speed is the highest, more than 40 km/h, around 6 a.m., then rapidly falls below 30 km/h around 8 a.m. Congestion appears to be worst in the afternoon rush hour with the lowest speed, ca. 27 km/h, occurring around 6 p.m.

The dotted lines show the variability across days, in terms of the mean plus/minus the standard deviation, of the network-wide average speed. The variation is low, around 1 km/h, which suggests that at an aggregate level the network follows a highly predictable speed trajectory every day. Meanwhile, the dashed lines show the mean speed plus/minus the network-wide average standard deviation of link speed across days. Thus, for individual links the average standard deviation is about 7 km/h, which is considerably larger than for the network as a whole. One reason for this is likely variability arising from traffic signals, map-matching, link allocation etc. Another plausible explanation is that the distribution of traffic in the network varies between days, which leads to large variations at the local level but small variations at the aggregate level.

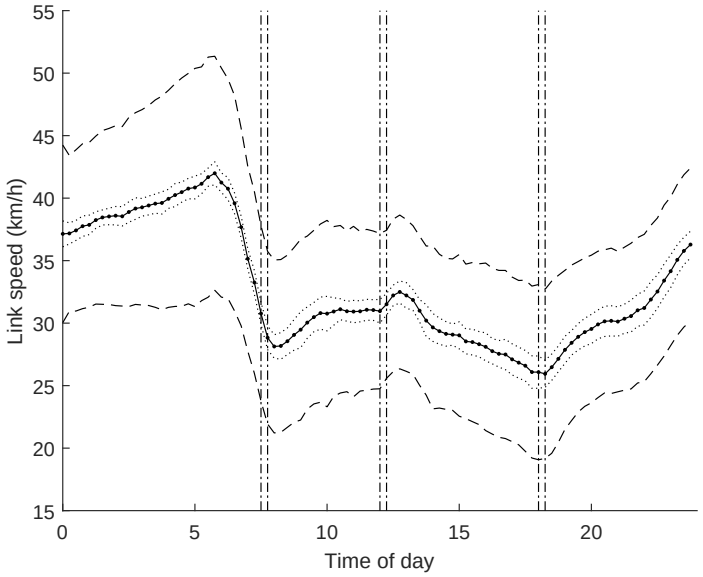

Table 2: Missing values across days in the Shenzhen data set. IQR: inter-quartile range.

Time period	Missing values (%)			
	Average	Min	Max	IQR
07:00–08:30	6.9	4.7	53.2	0.64
11:30–13:00	5.8	3.3	62.1	0.82
17:30–19:00	4.7	3.3	20.9	0.69

If this hypothesis is accurate, it is expected that the PPCA model be able to find some of these variation patterns.

The case study focuses on prediction from three different time-of-day intervals h: 7:30–7:45, 12.00–12:15, and 18:00–18:15 (indicated by the vertical lines in Figure 3). Prediction is performed up to 45 minutes into the future, i.e., $F = 3$.

Many links lack observations during some intervals on certain days; hence, some values in the historical and current-day observation vectors are missing. The variation of the fraction of missing values across days is shown in Table 2.

Five different prediction methods are evaluated:

1. Historical mean
2. k-nearest neighbors (k-NN) clustering
3. Local smoothing (Section 2.5)
4. PPCA model prediction (Section 2.3)

Table 3: Model parameter values calibrated through cross-validation on training set.

	7:30–7:45	12:00–12:15	18:00–18:15
k-NN	$P = 3$, $\kappa = 30$	$P = 3$, $\kappa = 44$	$P = 3$, $\kappa = 26$
Local	$P = 3$, $\alpha = 0.20$	$P = 3$, $\alpha = 0.15$	$P = 3$, $\alpha = 0.40$
PPCA	$P = 3$, $Q = 4$	$P = 3$, $Q = 2$	$P = 3$, $Q = 6$
Hybrid	$P = 3$, $Q = 4$, $\alpha = 0.15$	$P = 3$, $Q = 2$, $\alpha = 0.13$	$P = 3$, $Q = 6$, $\alpha = 0.33$

5. Hybrid (PPCA + local smoothing) (Section 2.5)

In all prediction methods, the logarithm of link speed is used as internal variable. The model output is transformed back to link speeds before the prediction accuracy is assessed (Section 2). Experiments show that this gives higher prediction accuracy compared to using link speed directly as internal model variable.

k-NN-based prediction is a common machine learning technique. The method is applied at the network-level, using the Euclidean distance between the variables across all links and past time intervals as distance metric. Each day in the historical database is weighted inversely proportionally to the distance to the current day during the P most recent time intervals. Prediction is performed by calculating the weighted average values in the future time intervals across the κ most similar historical days. Missing values are imputed with the historical mean for the corresponding time-of-day interval. The local smoothing method (Section 2.5) uses a weighted average of the deviations between observations and historical means in the P most recent time intervals for each link to predict future values. The weights decay geometrically according to the gap between the prediction interval and the past interval, where the same baseline weight parameter α is used for all links.

The parameters P and κ in the k-NN method, P and α in the local smoothing and hybrid methods, and P and Q in the PPCA and hybrid methods, are calibrated by cross-validation. A leave-one-out calibration scheme is used, in which the ability of the model to predict a day that is kept out of the training set is repeatedly assessed for each day in the training set. The prediction accuracy for a particular day held outside the training set is evaluated in terms of the average root mean squared error (RMSE) across all links and prediction intervals. The overall prediction accuracy is evaluated as the average across every day held out of the training set. For each prediction method, the set of parameters that maximizes overall prediction accuracy is found by a grid search approach.

The calibrated parameter values are shown in Table 3. For all time-of-day intervals and prediction methods, more than three past intervals lead to insignificant improvements or even reductions in prediction accuracy. For PPCA, relatively few latent dimensions are found significant for predictions outside of the training data. The remaining variability is attributed to random noise in the observations.

4 Results

4.1 Prediction Performance

Table 4 summarizes the prediction errors for each method and time-of-day interval in terms of average RMSE and mean absolute percentage error (MAPE) across the 10 test days. Due to the noisy nature of the probe data, residuals when compared to observations will be positive even for the most accurate prediction methods. Thus, relative differences between the methods are more

Table 4: Prediction accuracy on the test set. RMSE in km/h, MAPE in per cent, relative improvement from historical mean in parentheses.

Current time interval	Prediction interval					
7:30–7:45	7:45–8:00		8:00–8:15		8:15–8:30	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Historical mean	7.37	22.2	7.54	23.3	7.60	25.0
k-NN	7.26	21.8	7.47	22.9	7.57	25.5
	(1.46%)	(1.64%)	(0.87%)	(1.82%)	(0.42%)	(1.83%)
Local	7.33	21.8	7.52	23.2	7.60	25.0
	(0.64%)	(1.48%)	(0.22%)	(0.48%)	(0.04%)	(0.07%)
PPCA	7.25	21.7	7.47	22.9	7.54	24.4
	(1.62%)	(2.24%)	(0.82%)	(2.00%)	(0.83%)	(2.17%)
Hybrid	7.23	21.5	7.47	22.8	7.54	24.4
	(1.93%)	(3.07%)	(0.90%)	(2.21%)	(0.84%)	(2.19%)
12:00–12:15	12:15–12:30		12:30–12:45		12:45–13:00	
Historical mean	6.52	17.8	6.74	18.2	6.54	17.1
k-NN	6.50	18.0	6.74	18.3	6.58	17.2
	(0.33%)	(−0.76%)	(0.00%)	(−0.12%)	(−0.47%)	(−0.63%)
Local	6.41	17.6	6.73	18.2	6.54	17.1
	(1.60%)	(1.13%)	(0.25%)	(0.24%)	(0.04%)	(0.03%)
PPCA	6.50	17.8	6.73	18.1	6.55	17.0
	(0.40%)	(0.23%)	(0.11%)	(0.70%)	(−0.03%)	(0.19%)
Hybrid	6.42	17.6	6.72	18.1	6.55	17.0
	(1.57%)	(1.23%)	(0.28%)	(0.84%)	(−0.01%)	(0.21%)
18:00–18:15	18:15–18:30		18:30–18:45		18:45–19:00	
Historical mean	7.69	25.7	7.67	25.9	7.48	24.5
k-NN	7.45	25.3	7.55	25.9	7.44	24.6
	(3.09%)	(1.49%)	(1.58%)	(−0.01%)	(0.50%)	(−0.20%)
Local	6.91	22.9	7.34	24.9	7.33	24.1
	(10.1%)	(10.7%)	(4.34%)	(4.00%)	(1.94%)	(1.77%)
PPCA	7.30	24.6	7.46	25.3	7.36	24.0
	(5.01%)	(4.35%)	(2.74%)	(2.43%)	(1.50%)	(1.92%)
PPCA + Local	6.87	22.8	7.32	24.8	7.31	23.9
	(10.7%)	(11.3%)	(4.57%)	(4.43%)	(2.16%)	(2.56%)

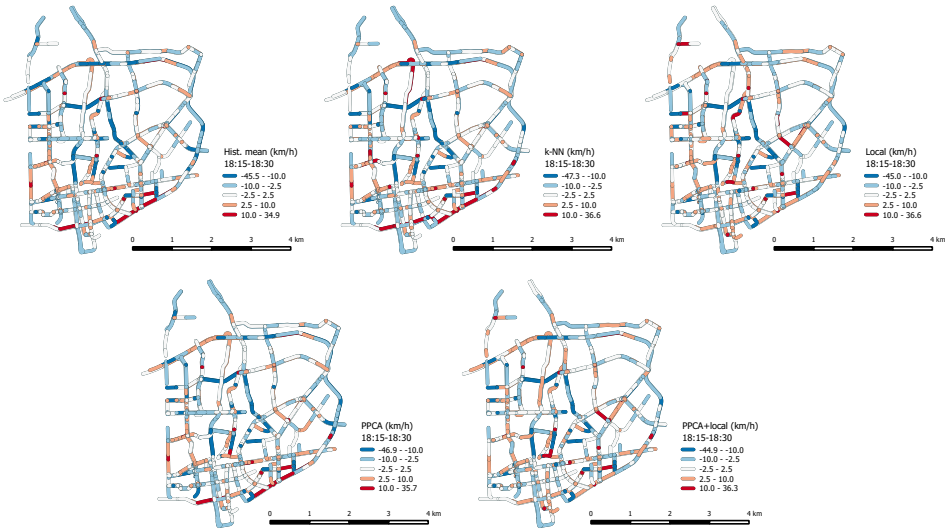

Figure 2: Prediction residuals for time interval 18:15–18:30, March 20, 2014. From top left to bottom right: Historical mean (RMSE 7.16), k-NN (RMSE 6.89), local smoothing (RMSE 6.60), PPCA (RMSE 6.82), and hybrid (RMSE 6.51).

relevant than absolute values of errors.

For all intervals and prediction horizons, current-day observations help improve prediction accuracy compared to the historical mean. In absolute values, the prediction accuracy is not always decreasing with prediction horizon, since the baseline variability of traffic conditions varies between time intervals. As

Table 5: Prediction accuracy for 15-minute prediction for separate functional classes. Current time h = 18:00–18:15. RMSE in km/h, relative improvement from historical mean in parentheses.

	Single network model				Separate subnetwork models			
	FC2	FC3	FC4	Total	FC2	FC3	FC4	Total
Historical mean	12.8	6.98	7.06	7.69	12.8	6.98	7.06	7.69
k-NN	11.1	6.90	7.05	7.45	10.2	6.78	7.07	7.30
	(13.5%)	(1.15%)	(0.03%)	(0.33%)	(20.7%)	(2.87%)	(-0.15%)	(5.10%)
Local	8.05	6.44	7.04	6.91	8.05	6.44	7.04	6.91
	(37.3%)	(7.82%)	(0.15%)	(10.1%)	(37.3%)	(7.82%)	(0.15%)	(10.1%)
PPCA	10.8	6.69	7.03	7.30	8.82	6.62	7.03	7.06
	(15.7%)	(4.13%)	(0.39%)	(5.01%)	(31.4%)	(5.16%)	(0.35%)	(8.26%)
Hybrid	8.09	6.38	6.99	6.87	7.58	6.37	6.94	6.80
	(37.0%)	(8.56%)	(0.95%)	(10.7%)	(41.0%)	(8.75%)	(1.57%)	(11.6%)

expected, however, the performance of the other methods relative to the historical mean is the highest for the shortest prediction horizon and then gradually decreases.

The relative predictive power of all methods compared to the historical mean is the highest for the 18:00–18:15 time interval, where the contribution of current-day observations is significant even up to the 45 minutes horizon. PPCA reduces RMSE by 5% compared to the historical mean, and the hybrid PPCA and local method reduces RMSE by almost 11%. At 07:30–07:45 the PPCA and hybrid methods improve over the historical mean for all prediction horizons, while at 12:00–12:15 the relative improvement is much smaller, and insignificant for horizons beyond 15 minutes. This suggests that current-day observations are particularly useful during intervals with significant congestion. Meanwhile, absolute prediction errors are lower at 12:00–12:15 than at the other time intervals even when using the historical mean for prediction.

Of the two network-wide methods, PPCA outperforms k-NN for all intervals and prediction horizons. PPCA also outperforms local smoothing at 7:30–7:45, but is outperformed at 12:00–12:45 and 18:00–18:15. The combination of PPCA and local smoothing outperforms all other methods for all time intervals and prediction horizons. The results indicate that PPCA picks up wider spatial correlation patterns while local smoothing picks up trends at the link level, and the hybrid approach is able to explain a larger share of the observed travel time variability.

Fig. 2 shows the residuals between predicted and observed link speeds for a particular day in the test set, March 20 2014, for the one step ahead prediction interval from 18:00–18:15. In general, the spatial patterns of the historical mean, k-NN and PPCA residuals are similar, although PPCA yields the smallest overall prediction error and the historical mean the largest. On this particular test day, the prediction methods tend to underestimate link speeds in the north and the centre of the network but overestimate them in the south. The local smoothing method and the hybrid method are similar to each other but slightly different from the other three methods. In particular, the local methods reduce prediction errors along the major semi-ring road highlighted in Fig. 1.

4.2 Functional Class Subnetwork Prediction

The prediction performance is analyzed separately for each functional class in Table 5. It is seen that prediction errors are significantly larger for FC2 than for FC3 and FC4. On the other hand, PPCA and the hybrid method achieve the greatest reduction in errors for FC2 (37% for the hybrid method). This is partly due to fewer missing values (0.73% for FC2, 3.65% for FC3, 6.23% for FC4).

Following Section 2.4, the network is partitioned into subnetworks based on functional classes, and prediction is performed separately for each subnetwork. Thus, a separate PPCA model is estimated for each subnetwork. The subnetwork approach reduces prediction errors for all functional classes and for the overall network (Table 5), even though the same, likely suboptimal, parame-

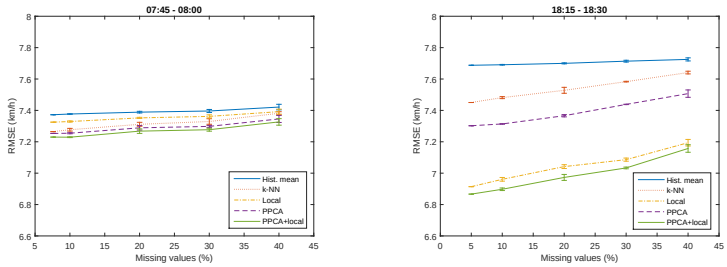

Figure 3: Prediction accuracy for 15-minute prediction with different amounts of missing values. Left: Current-time h = 07:30–07:45. Right: Current-time h = 18:00–18:15. Averages across three random draws, error bars show standard errors.

ter values of P, Q, κ and α calibrated for the full network are used for every subnetwork. The hybrid method achieves 41% reduction in prediction error for FC2. The results highlight the potential of the subnetwork prediction approach; refinement of the partitioning method may increase prediction accuracy further.

4.3 Impact of Missing Values

The data contain between 4.7% and 6.9% missing values depending on the time interval (Section 3). The robustness of the prediction methods against higher shares of missing values is investigated by removing additional observations from the historical and current-day data completely at random. The prediction accuracy is evaluated for 10%, 20%, 30%, and 40% missing values across multiple random draws. The model parameters given in Table 3 are used in all scenarios.

Fig. 3 shows the RMSE for the one step ahead (15 minutes) prediction interval (for current time interval 8:00–8:15 to the left and 18:00–18:15 to the right). All methods utilize in some way the historical mean as baseline predictor, and the benefit of the other methods relative to the historical mean decreases with increasing missing values, but prediction errors increase quite slowly. All methods thus display robustness against missing values. The hybrid model is the most accurate method regardless of the amount of missing values.

4.4 Principal Components Analysis

This section looks at the inferred spatio-temporal correlation structures of the PPCA model in more detail. The analysis focuses on the three intervals from 17:30 to 18:15.

By design, the columns of W are sorted in decreasing order according to the magnitude of variability in the observations explained by the dimensions of the latent space. Fig. 4, top row, shows the first column of the loadings matrix W. Large positive loading coefficients are represented by dark red colors while large negative coefficients are represented by dark blue colors. Links with

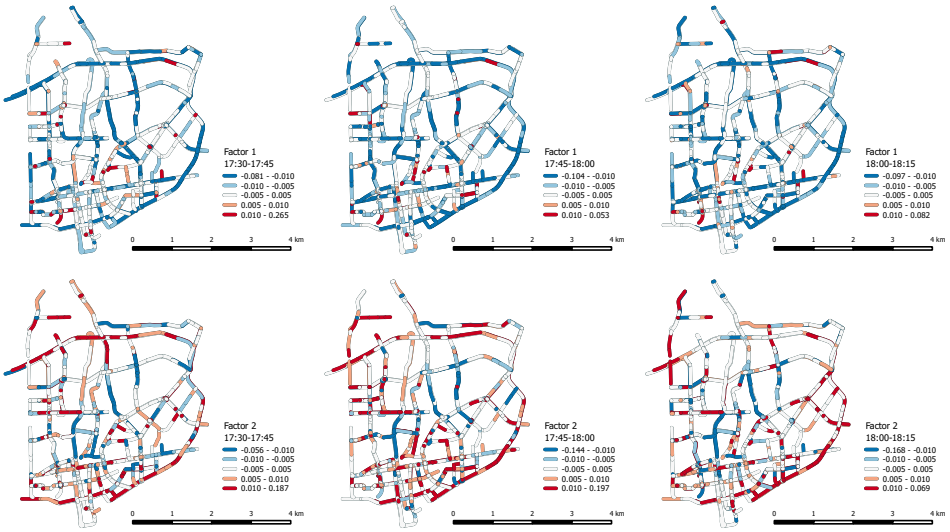

Figure 4: Factor loadings from 17:30–17:45 to 18:00–18:15. Top row: principal factor 1. Bottom row: principal factor 2.

dark colors are positively correlated with links of the same color and negatively correlated with links of the opposite color, while links of intermediate color are more weakly correlated with other links.

The first principal factor captures positive correlation between many of the major urban roads (Fig. 1). This suggests that there are strong dependencies between links in the main urban network, while dependencies with minor links are often weaker. The factor loadings are quite stable over time; e.g., the correlation across all links between the first time interval and the two following intervals is 0.44 and 0.52, respectively. This is not an assumption in the model but purely a result of the estimation.

Fig. 4, bottom row, shows the second column of the loadings matrix W.

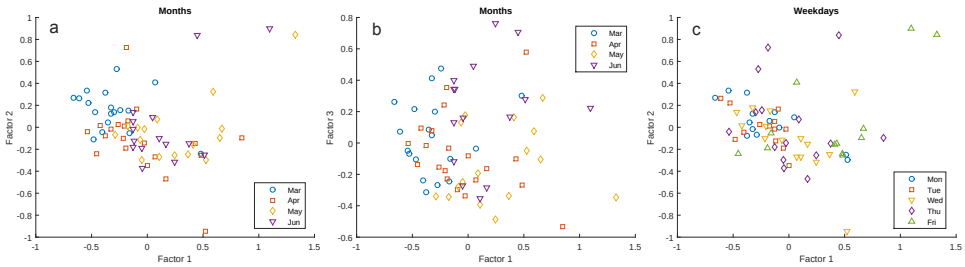

Figure 5: Expected values for principal factor 2 (a, c) and factor 3 (b) against principal factor 1, distinguished by month (a, b) and weekday (c).

The second principal factor displays a different spatio-temporal pattern than the first, and the overall correlation across links and time intervals between the first and the second factor loadings is only 0.04. The factor captures strong positive correlation among the links in the major semi-ring road (Fig. 1), but also strong negative correlation between the ring road and the network center.

According to the model, each day in the historical data is generated from a point in the latent space. The point for day n can be projected from the observations as the expected value of the latent variables conditional on the observations, $\hat{\mathbf{x}}_n = E[\mathbf{x}_n|\mathbf{u}_n] = (W^T W)^{-1} W^T(\mathbf{u}_n - \boldsymbol{\mu})$, $n = 1, \ldots, N$. Here $\hat{\mathbf{x}}_n$ is a $Q \times 1$ vector where element q is the expected value in the qth principal latent dimension, $q = 1, \ldots, Q$.

Fig. 5 plots the first few principal factors against each other for all historical days. There are no anomalous days in terms of extreme outliers, which is a verification of model validity. To investigate whether there are any systematic patterns across days in the principal factors, days are distinguished by month and by weekdays, respectively. There is evidence of clustering of the points based on the month. In Fig. 5a in particular, days are roughly ordered along a negatively sloping line in the temporal order of the months, with days from March in the upper left corner and days from May and June in the bottom right corner. This indicates that the first two principal factors partly capture long-term variations between months of the year. Figure 5b shows more separation between days from May and June, suggesting that the third principal dimension picks up some variation between the two months.

There is less clear evidence of clustering based on weekdays (Fig. 5c). In general, Mondays to Thursdays are clustered relatively closely together, while Fridays stand out as a somewhat distinct cluster. The fact that traffic on Fridays is different from other weekdays is a common phenomenon in urban networks, and the results are thus in line with expectations.

Student's t-tests confirm that both the first and second principal factors are significantly different (at the 5% level) for days in March compared to the other months; for the first principal factor, days in April are also significantly different from days in May on average. For the third principal factor, June is significantly different from April and May, while March is significantly different from May. Regarding weekdays, the first principal factor captures significant differences between Fridays and the other days, as well as between Tuesdays and Thursdays. Thus, the first principal factor represents a combination of month and weekday effects, while principal factors two and three capture additional month effects.

5 Conclusion

The paper proposes a probabilistic principal components analysis model of network travel times. The model assumes that observed link travel times are projections from a lower-dimensional latent space, which determine the spatio-temporal correlations among links and time intervals. Because of the dimen-

sionality reduction of PPCA, prediction conditional on observed current-day link travel times is performed efficiently, with consistent handling of missing values, in real-time even for large-scale networks. The methodology is flexible to changes in network, data and model extensions, and robust against noisy and missing data through the consideration of both temporal and spatial correlations. In contrast to model-free approaches like k-NN, historical data do not have to be reconsidered for every new prediction. Furthermore, it does not suffer from the time lag problems of many autoregressive time series approaches.

The model captures the wider spatio-temporal correlation patterns between days that are lost with methods that only consider local correlations between nearby links. The case study of Shenzhen, China, shows that some of the captured variation constitutes long-term variations over months, while some of the variation constitutes differences between weekdays. However, much of the variation comes from other, unidentified sources, which highlights the value of not imposing prior labeling and clustering of observations for the prediction.

The case study shows that a combination of the PPCA model and a local model, in this case a simple smoothing method, can give the highest prediction accuracy. Partitioning the network into subnetworks based on functional class further increases accuracy. Furthermore, the prediction methods are robust against missing values.

Further work should be dedicated to developing the integration between local and network-wide prediction, including the methods for partitioning the network into subnetworks. Under anomalous conditions such as incidents, data-driven approaches require that similar events are captured in the historical data set. Relaxing the spatio-temporal constraints of the model is an interesting direction of development. Further, the PPCA framework allows exceptional conditions to be identified based on the likelihood of observed speeds, and offers a signal about the validity of prediction based on historical patterns compared to, e.g., simulation-based approaches.

Acknowledgment

The authors would like to thank Ding Luo, Shenzhen Urban Transport Planning Center, for kindly providing the data.

References

[1] H. van Lint and C. van Hinsbergen, "Short-term traffic and travel time prediction models," in *Artif. Intell. Appl. Crit. Transp. Issues*. Transp. Res. Board, 2012, pp. 22–41.

[2] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where were going," *Transp. Res. Part C*, vol. 43, pt. 1, pp. 3–19, 2014.

[3] M. Rahmani and H. N. Koutsopoulos, "Path inference from sparse floating car data for urban networks," *Transp. Res. Part C*, vol. 30, pp. 41–54, 2013.

[4] C. Antoniou, R. Balakrishnan, and H. N. Koutsopoulos, "A synthesis of emerging data collection technologies and their impact on traffic management applications," *Eur. Transp. Res. Rev.*, vol. 3, pp. 139–148, 2011.

[5] F. Yang, Z. Yin, H. X. Liu, and B. Ran, "The reliability of travel time forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 162–171, 2010.

[6] A. Simroth and H. Zähle, "Travel time prediction using floating car data applied to logistics planning," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 243–253, 2011.

[7] J. Haworth and T. Cheng, "Non-parametric regression for space-time forecasting under missing data," *Comput. Environ. Urb. Syst.*, vol. 36, no. 6, pp. 538–550, 2012.

[8] Q. Ye, W. Y. Szeto, and S. C. Wong, "Short-term traffic speed forecasting based on data recorded at irregular intervals," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1727–1737, 2012.

[9] F. Zheng and H. van Zuylen, "Urban link travel time estimation based on sparse probe vehicle data," *Transp. Res. Part C*, vol. 31, pp. 145–157, 2013.

[10] G. Fusco, C. Colombaroni, L. Comelli, and N. Isaenko, "Short-term traffic predictions on large urban traffic networks: Applications of network-based machine learning models and dynamic traffic assignment models," in *2015 Int. Conf. Models Techn. Intell. Transp. Syst. (MT-ITS)*, 2015, pp. 93–101.

[11] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data," *Transp. Res. Part B*, vol. 53, pp. 64–81, 2013.

[12] B. S. Westgate, D. B. Woodard, D. S. Matteson, and S. G. Henderson, "Large-network travel time distribution estimation for ambulances," *Eur. J. Oper. Res.*, vol. 252, no. 1, pp. 322-333, 2016.

[13] W. Min, and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transp. Res. Part C*, vol. 19, pp. 606–616, 2011.

[14] A. Hofleitner, R. Herring, P. Abbeel, and A. Bayen, "Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, pp. 1679–1693, 2012.

[15] X. Zhan, S. V. Ukkusuri, and C. Yang, "A Bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data," *Autom. Construct.*, vol. 72, pt. 3, pp. 237–246, 2016.

[16] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun, "A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting," *Transp. Res. Part C*, vol. 62, pp. 21–34, 2016.

[17] W. Burghout, H. N. Koutsopoulos, and I. Andreasson, "Incident management and traffic information: Tools and methods for simulation-based traffic prediction," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 2161, pp. 20–28, 2010.

[18] M. Ben-Akiva, M. Bierlaire, D. Burton, H. N. Koutsopoulos, and R. Mishalani, "Network state estimation and prediction for real-time transportation management applications," *Netw. Spat. Econ.*, vol. 151, no. 1, pp. 13–22, 2002.

[19] Y. Wang and M. Papageorgiou, "Real-time freeway traffic state estimation based on extended Kalman filter: a general approach," *Transp. Res. Part B*, vol. 39, no. 2, pp. 141–167, 2005.

[20] D. B. Work, O.-P. Tossavainen, S. Blandin, A. M. Bayen, T. Iwuchukwu, and K. Tracton, "An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices," in *Proc. 47th IEEE Conf. Decision Control*, 2008, pp. 5062–5068.

[21] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. Royal Stat. Soc. Series B (Stat. Meth.)*, vol. 61, no. 3, pp. 611–622, 1999.

[22] J. Porta, J. Verbeek, and B. Krose, "Active appearance-based robot localization using stereo vision," *Auton. Robots*, vol. 18, no. 1, pp. 59–80, 2005.

[23] L. Qu, L. Li, Y. Zhang, and J. Hu, "PPCA-based missing data imputation for traffic flow volume: A systematical approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 512–522, 2009.

[24] S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu, "Supervised probabilistic principal component analysis," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*. ACM, 2006, pp. 464–473.

[25] S. Li, J. O. Nyagilo, D. P. Dave, W. Wang, B. Zhang, and J. Gao, "Probabilistic partial least squares regression for quantitative analysis of Raman spectra," *Int. J. Data Mining Bioinf.*, vol. 11, no. 2, pp. 223–243, 2015.

[26] L. Edie, "Discussion on traffic stream measurements and definitions," *2nd Int. Symp. Theory Traffic Flow*, pp. 139–154, 1963.

[27] N. Uno, F. Kurauchi, H. Tamura, and Y. Iida, "Using bus probe data for analysis of travel time variability," *J. Intell. Transp. Syst: Techn. Plan. Oper.*, vol. 13, no. 1, pp. 2–15, 2009.