REAL-TIME URBAN ROAD NETWORK TRAVEL TIME PREDICTION BASED ON PROBE DATA: FRAMEWORK AND STOCKHOLM CASE STUDY

Matej Cebecauer (corresponding author)
KTH Royal Institute of Technology
Department of Transport Science
Teknikringen 10, SE-100 44 Stockholm, Sweden
Email: matejc@kth.se

Erik Jenelius
KTH Royal Institute of Technology
Department of Transport Science
Teknikringen 10, SE-100 44 Stockholm, Sweden
Email: erik.jenelius@abe.kth.se

Wilco Burghout
KTH Royal Institute of Technology
Department of Transport Science
Teknikringen 10, SE-100 44 Stockholm, Sweden
Email: wilco@kth.se

Word count: 5,587 words in text with references + 7 figures (250 words each)
= 7,337 words in total

August 4, 2016
ABSTRACT
The paper presents the architecture and methodology for an urban road network travel time prediction framework based on low frequency probe vehicle data. The prediction framework is intended to be used in a real-time traffic management and information provision tool. The design of the framework has to satisfy three crucial aspects: computational efficiency of real-time prediction; flexibility to changes in the network; and robustness against noisy and missing data. Correlation patterns between links and time intervals are computed, updated and stored at infrequent intervals based on historical data and a Probabilistic Principal Component Analysis (PPCA) model. Prediction for future time intervals is performed in real-time using stored model parameters of correlation patterns and recent current-day observations. A novel multivariate hybrid method of PPCA and local smoothing, which considers local correlations among neighboring links, is proposed. The methodology is applied to the road network of Stockholm, Sweden and probe data from taxis. Computational experiments show that the proposed method provides high accuracy for both the peak hours as well as off-peak traffic conditions. The effects of missing data and link speed variation on the prediction accuracy are also analyzed. For links with large link speed variations in particular, the PPCA based methods significantly outperform the historical mean, whereas for links with high proportions of missing data the hybrid PPCA method is more reliable than pure PPCA due to its ability to use local neighboring link correlations.

Keywords: travel time prediction, PPCA, probe data
INTRODUCTION

Real-time travel time estimation and prediction is a topic that has sparked considerable research interest in recent years. Prediction is crucial for traffic control, information provision, online vehicle routing 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 information provision, vehicle routing and trip planning, prediction at the network level is generally desirable.

GPS devices in vehicles or smart phones provide information that recently has proven useful for traffic management applications (1). These opportunistic traffic sensors allow the collection of data during any time and day in urban road networks at low marginal costs (1, 2). However, there are a number of limitations that make the use of probe vehicle data challenging, and advanced and sophisticated methods are needed to generate useful information to support traffic management decisions (3). Regarding traffic prediction, the main identified challenges are: moving from motorways and single arterials to the network level; properly handling missing data; and making use of new data sources such as probe data (4).

With the increasing availability of probe data, the literature on real-time travel time estimation has been growing significantly in the past years (1, 2, 5, 6, 7, 8). Meanwhile, most research on short-term prediction has focused on motorways and major arterials (4, 9). Relatively few papers consider short-term prediction in the context of large urban networks. One proposed framework considers a dynamic Bayesian network model of the spatio-temporal dependencies (10). The state of the prediction for the next time interval is only affected by the current state of the neighboring links. Another method, which is an extension of k-nearest neighboring, does not inherently handle missing data (11). A further line of research is utilizing simulation models for traffic prediction, known as Dynamic Traffic Assignment (DTA) models (12, 13). However, while these models are behaviorally rich and can capture driver response to traffic information, they are complex and their calibration and application is challenging. Finally, a data-driven methodology for network-wide travel time prediction is proposed in (14). The method combines advantages of the Probabilistic Principal Component Analysis (PPCA) and local smoothing. One advantage of the method is that no explicitly modeled network structure or temporal progression is needed. Furthermore, the approach produces multivariate network-wide predictions, which has so far received relatively little attention in the literature.

This paper presents a real-time framework for traffic prediction using probe vehicle data. Part of this framework is a novel hybrid PPCA method. The hybrid PPCA methodology proposed in (14) for network-wide short-term prediction is adopted in order to capture network-level spatio-temporal correlation patterns and handle missing data. The local smoothing is extended to neighboring links in order to capture local spatio-temporal correlation patterns. Furthermore, the paper investigates and quantifies the effects of probe data properties, e.g., the fraction of missing observations and link speed variability, on prediction performance.

Computational experiments show that the designed framework produces predictions for the complete road network of the case study (1,880 links) under 1 second using the hybrid PPCA method. The proposed method is the most accurate among all considered methods, especially during peak hours when the link speed variability is high, and thus the accuracy of the prediction is very sensitive. The PPCA-based methods are able to better adapt to different variation patterns and generate significantly better prediction than the historical mean. Furthermore, based on the computational experiments, the effects of the fraction of missing observations and link speed variability
on the prediction accuracy is investigated.

The paper is organized as follows. The prediction framework and methods are presented in Section 3. Section 4 introduces the data and case study. Results of numerical experiments are reported in Section 5. Section 6 concludes the paper.

METHODOLOGY
In this section the real-time travel time prediction framework and PPCA methodology are introduced. Prediction based on PPCA can be considered as a two-stage approach. In the first stage, historical observations are used to estimate the model parameters. In the second stage, current-day observations are combined with the output of the first stage to produce predictions in real-time. Finally, a novel smoothing approach is proposed based on the smoothed average value across past current-day time intervals and neighboring links.

Notation
The paper uses the following notation:

- $K$: number of links, indexed $k, l = 1, \ldots, K$
- $A_k$: set of the adjacent links to the link $k$
- $I$: number of time-of-day intervals (intervals), indexed $i = 1, \ldots, I$
- $h$: current interval
- $F$: number of future intervals in prediction horizon, indexed $i = h + 1, \ldots, h + F$
- $P$: number of past intervals to base prediction on, indexed $i = h - P + 1, \ldots, h$
- $D = K(P + F)$: number of observed model variables
- $N$: number of historical days used in estimation, indexed $n = 1, \ldots, N$
- $Q$: dimensionality of latent space, indexed $q = 1, \ldots, Q$
- $W, \mu, \sigma^2$: parameters of the PPCA method
- $\alpha, \beta$: parameters of the local smoothing method
- $v_{ikn}$: speed of link $k$ in interval $i$ on day $n$
- $\omega_{ik}$: speed of link $k$ in interval $i$ on current day

Framework
The real-time traffic prediction framework for urban road networks can be divided into three main components: (i) stream processing infrastructure; (ii) background processes; and (iii) real-time services. Figure 1 shows the framework architecture.

Stream Processing Infrastructure
This component receives and handles the incoming stream of probe data. The map-matcher module projects the probes to the road network considering the most likely vehicle paths. The projected pairs are then stored in the database. The map-matching and path inference framework is described in more detail in (2).

Background Processes
Two modules of this component are running in the background at specified time intervals. A travel time estimation module is running in our case each 15 minutes. The module takes the past 15 minutes interval to compute the travel time estimation based on the projected probe data for this
FIGURE 1 Architecture of the real-time traffic prediction framework.
particular interval. The estimated link travel times for this interval are then stored in the historical
database as well as in the real-time database. The real-time database is only keeping data for a
certain number of past intervals within the current day.

The module for the calibration of parameters is the most computationally intensive com-
ponent of the framework. The task of the module is to calibrate and update of the parameters $P$,
$Q$, $W$, $\mu$, $\sigma^2$, $\alpha$ and $\beta$ for the real-time prediction. The module can run relatively infrequently, e.g.,
onece per month, to update prediction parameters based on a long time series of historical data.
Because of this decomposition of model calibration and prediction, current-day link travel times
can be estimated or predicted in real-time.

**Real-Time Services**
The real-time modules of this component receive lists of links for which the travel times are pre-
dicted or estimated. To predict for the future intervals, the historical correlations captured by the
calibrated parameters of the PPCA model and residuals from current-day observations are consid-
ered. Finally, service modules adapt the results in an appropriate way for client applications.

**Prediction Methodology**

**PPCA Method and Network-wide Prediction**

PPCA is a probabilistic model that assumes that observed variables are generated from latent (un-
observed) variables of lower dimension (15). The PPCA model and its use for travel time predic-
tion is described in more detail in (14). Here, the logarithm of the space-mean speed on each link
is used as the observed variable. Let $\tilde{v}_{ikn} = \log v_{ikn}$ be the logarithm of link speed of link $k$ in
interval $i$ on day $n$. The variables for all links in interval $i$ and day $n$ are stored in the $K \times 1$
vector $\tilde{v}_n$, and the variables for all links and intervals from $h - P + 1$ to $h + F$ are stacked in the $D \times 1$
vector $\tilde{v}_n$. Thus,

$$\tilde{v}_n = \tilde{\mu} + W x_n + \epsilon_n$$

where $x_n$ is a $Q \times 1$ column vector of latent random variables, assumed to be i.i.d. Normal with
mean zero and variance one. $W$ is a $D \times Q$ parameter matrix, often called factor loadings, that can
be interpreted as the linear mapping between the latent variables and the observed variables. $\epsilon_n$
is a $D \times 1$ column vector of measurement noise, assumed to be i.i.d. Normal with mean zero and
variance $\sigma^2$. $Q$ is a model parameter; it is typically assumed that $Q < D$, i.e., that the observed
variables are generated from a lower-dimensional latent space. $\tilde{\mu}$ is a $D \times 1$ parameter column
vector that represents the mean logarithm of speed for each link and interval.

Maximum likelihood estimators of factor loadings $W$, historical mean $\tilde{\mu}$ and observation
noise $\sigma^2$ for fixed $Q$ and $P$ are derived in (15) for complete data an in (16) for incomplete data.
The EM algorithm of (16) is used here to handle missing values in the historical data.

In order to predict network-wide link travel times in future intervals, let $\tilde{\omega}_{ik}$ be the logarit-
mic speed variable for link $k$ in interval $i$ of the current day. The variables for all links in interval $i$
are stacked in the the $K \times 1$ vector $\tilde{\omega}_i$, and the variables for all links and intervals from $h - P + 1$
to $h + F$ are stacked in the $D \times 1$ vector $\tilde{\omega}$. The variables $\tilde{\omega}$ are assumed to be generated from
the PPCA model in Eq. (1). The $\tilde{\omega}$ vector can be split into the observed values $\tilde{\omega}_P$ and the future
values $\tilde{\omega}_F$. 
The loading matrix $W$ can be similarly split into $W_p$ and $W_f$, and $\mu$ into $\mu_p$ and $\mu_f$.

The future variables $\omega_i$, $i = h + 1, \ldots, h + F$, are predicted conditional on the past observed values $\omega_i$, $i = h - P + 1, \ldots, h$. The point predictor for interval $i$ is

$$
\hat{\omega}_{i|h} = \mu_i + W_i (W_p^T W_p + \sigma^2 I)^{-1} W_p^T (\hat{\omega}_p - \mu_p)
$$

The historical mean $\mu_i$ is the base value for the prediction, and the factor loadings are applied to the residuals between observations $\omega_p$ and $\mu_p$. The PPCA model can also estimate the mean values in the past intervals conditional on the current-day observations, by applying Eq. (3) to past interval $i \in \{h - P + 1, \ldots, h\}$.

It is important to note that if the observations $\omega_p$ contain missing values, the prediction is performed after removing the corresponding rows of $\omega_p$, $W_p$ and $\mu_p$. This allows travel times to be predicted even for links entirely missing current-day observations.

**Hybrid Network-wide and Local Prediction Considering Neighboring Links**

The PPCA method captures network-wide correlation patterns across days, but is less well suited to capture local correlations between consecutive time intervals for single links or small neighborhoods of links. For this purpose, a hybrid PPCA and local smoothing methodology introduced in (14) is adopted here. The methodology is extended by taking correlations between neighboring links into consideration. The local smoothing updates the predicted speed of the pure PPCA method $\omega_{ik|h} = \exp(\hat{\omega}_{ik|h})$ for future interval $i$ and link $k$ with the sum of smoothed residuals between observed link speeds $\omega_{jk}$ and PPCA estimated mean values $\omega_{jk|h} = \exp(\hat{\omega}_{jk|h})$ for the past intervals $j = h - P + 1, \ldots, h$. Thus, the network-wide and local predictor for future interval $i$ with considering the neighboring links is:

$$
\hat{\omega}_{ik} = \omega_{ik|h} + \alpha \sum_{j=h-P+1}^{h} I_{jk} \cdot (\omega_{jk} - \omega_{jk|h}) + \beta \sum_{j=h-P+1}^{h} \sum_{l \in A_k} O_{il} \cdot I_{jl} \cdot (\omega_{jl} - \omega_{jl|h})
$$

$I_{jk}$ is equal to 1 if $\omega_{jk}$ is observed and 0 if the value is missing. $O_{k} = I_{h-P+1,k} + \cdots + I_{h,k}$ is the number of observed $\omega_{jk}$ values across the $j = h - P + 1, \ldots, h$ past intervals on link $k$. Note that the pure PPCA prediction is recovered by setting parameters $\alpha = 0$ and $\beta = 0$ in Eq. (4). For the remainder of this paper this method is referred to as the hybrid PPCA method.

**DATA AND CASE STUDY**

The travel time prediction framework is applied to the road network of the northern inner city of Stockholm, Sweden. In the digital road network each link has a number of attributes, including length, presence of a traffic signal, functional road class, and speed limit. The functional class categorizes roads from class 1 (major roads) to class 5 (minor service or side streets). All functional classes are included except the highest one (class 5: minor service or side streets), given that for that class more than 80% of links observations across all days and intervals are missing. The case
study road network consists of 1,886 links (functional class 1: 248 links, class 2: 79 links, class 3: 1,118 and class 4: 441). Figure 2 shows the speed limit and functional classes of all links included in the case study.

Link travel times are estimated for 15 minutes intervals using probe data from ca. 1,500 taxis operating in the area. The 15 minutes interval represents the overall trade-off between the temporal resolution and the number of observations per link that guarantees reasonable travel time estimation quality. For details on the travel time estimation method see (1, 2). One year of data (year 2014) is used. Saturdays, Sundays and all holidays as well as school holidays are removed from the data set. After this filtering, data from 179 days are used. Of these, 30 days are randomly selected as a calibration set, used by the calibration module of the framework to calibrate parameters $P$, $Q$, $\alpha$ and $\beta$. Further, 30 randomly selected days are used for the prediction performance evaluation in Section 5. The remaining 119 days are used as the training set for the PPCA model to estimate parameters $W$, $\mu$ and $\sigma^2$ based on the historical observations.

Figure 3 shows network speed and speed variations for different functional classes across the day. There is presence of congestion for a number of links and intervals. The highest average speed across all links is around 35 km/h; during peak hours the average link speed falls below the 25 km/h. The congestion is especially severe for lowest functional classes 3(b-c) that normally have the highest travel speed and represent natural bottlenecks because they are the part of the commuting backbone, which is congested during the peak hours in the morning (7:00-9:00) and in the evening (15:00-19:00).

Functional class 4 has the highest Fraction of Missing Observations (FMO) (see Figure 3(e)). $\text{FMO}(i, k)$ denotes fraction of missing observation across $P$ past intervals for link $k$ in interval $i$ and let $\text{FMO}(i)$ refer to the mean $\text{FMO}(i, k)$ across all links in interval $i$. For all functional classes, the largest $\text{FMO}(i)$ is observed in the early morning hours (1:30-8:00).

To describe the link speed variability is not trivial, therefore, we use three different statistics. First, the standard deviation of the network-wide average speed (dotted lines in Figure 3) shows that the variability across days is usually small, around 1-2 km/h (see Figure 3(d-e)). The highest variability is for links with lowest functional classes, as shown in Figure 3(b-c)). The low network-wide variability suggests that network speed at an aggregate level is similar across days.

Second, the dashed lines in Figure 3 show the standard deviation of individual link speed

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Case study road network of the Stockholm, Sweden. (a) Speed limits. (b) Functional classes of links.}
\end{figure}
across days, which is larger than the standard deviation for the network as a whole. This variability may be caused by traffic signals, crossroads, or because of variations in the traffic distribution across days. In addition, the map-matching and link allocation algorithms may affect variability as well (1).

Third, the relative mean absolute difference (RMD) is used to evaluate link speed variability across different links and functional classes. RMD considers all differences between values in the sample set independently from the mean contrary to the standard deviation and coefficient of variations. The mean speed of link \( k \) in interval \( i \) across \( N \) evaluation days is denoted \( \mu_{ik} \). For the set \( \{ v_{ikn}, n = 1, 2, \ldots, N \} \), the relative mean absolute difference RMD\((i, k)\) for interval \( i \) and link \( k \) is defined as:

\[
RMD(i, k) = \frac{\sum_{n=1}^{N} \sum_{m=1}^{N} |v_{ikn} - v_{ikm}|}{N^2 \mu_{ik}}.
\] (5)

RMD\((i, k) = 0 \) when the link \( k \) has the same speed for all observations. Let RMD\((i)\) refer to the mean RMD\((i, k)\) across all links in interval \( i \). The RMD values show larger link speed variability, especially during the peak hours (see light green line in Figure 3).

It is expected that the accuracy of all prediction methods decreases with the growth of the RMD. On the one hand we expect the PPCA methods to outperform the historical mean when the RMD is larger. This should be possible thanks to the methods considering the historical correlation patterns that can detect such variations and consider current-day observations. On the other hand, RMD values close to zero indicate low link speed variability where the predicted value is with large probability close to the historical mean.

**Computational Experiments**

The case study focuses on network-wide short-term travel time prediction 15 minutes into the future (i.e., parameter \( F = 1 \)). All 96 15-min intervals from 0:00–0:14 to 23:45–24:59 are investigated. Three prediction methods are evaluated:

- Historical mean
- PPCA
- Hybrid PPCA

**Calibration**

The PPCA model uses the logarithm of link speeds as the observed variable (14). Before the calibration of parameters \( P, Q, W, \sigma^2, \mu \), the outputs of the model are transformed back to link speeds with the exponential function. The local smoothing is applied after the transformation and is thus based on link speeds directly.

All parameters \( P, Q, W, \sigma^2, \mu, \alpha \) and \( \beta \) are calibrated across 30 days (the calibration set) separately for each 15-minute interval. The 119 days are used as the training set to estimate parameters \( W, \sigma^2 \) and \( \mu \). The parameters \( P \) and \( Q \) directly effect the dimensions of \( W \). The combination of the parameters \( P, Q, \alpha \) and \( \beta \) producing the best overall prediction accuracy across calibration set is selected.
FIGURE 3  Network speed (km/h), fraction of missing observations (FMO(i)), and relative mean absolute difference (RMD(i)) during the day. (a) All links of the study area. (b-e) Results per link functional class: (b) class 1 (c) class 2 (d) class 3 and (e) class 4.
Calibration experiments show that care should be taken to avoid overfitting of the model. Calibration of separate parameters for each link reduces the average error in prediction accuracy of hybrid PPCA to only 1-2 km/h on the calibration set and can be better than pure PPCA (about 5 km/h or more). This effectiveness achieved on the calibration set highlights the potential of hybrid PPCA. For the evaluation set, however, the accuracy is about 2-4 km/h worse in average compared to when joint parameters for the network as one whole are calibrated. This confirms the expectation that there is larger variability in the observed speeds across days for individual links (see dashed lines in Figure 3). Parameter $P$ is limited to a range of reasonable values $1, 2, \ldots, 5$, and $Q$ to the range $1, 2, \ldots, 10$. The $\alpha$ and $\beta$ parameters are restricted to only seven values $0, 0.05, \ldots, 0.25, 0.3$ with regular step $0.05$, in order to avoid overcalibration.

Figure 4 shows the values of calibrated parameters across all intervals. The parameter $Q$ has a tendency to be higher when the FMO is small. In other words, when there are few missing observations, the PPCA model is able to find more useful correlation patterns for prediction in the historical data. Meanwhile, the parameter $\beta$ is always less than or equal to 0.05, and it seems that $\beta$ is applied when the FMO is large. In this case, the calibration uses neighboring links ($\beta > 0$) to increase the number of current-day observations in the local smoothing. On the other hand, parameter $\alpha$ is lower when the FMO is large, but this is partially an effect of the presence of parameter $\beta > 0$ and considering the neighboring links in the prediction.

The hybrid PPCA needs significantly more computational time as the pure PPCA to calibrate $\alpha$ and $\beta$ (see Figure 4(c)), mainly considering $\beta > 0$ is taking large fraction of computational time. Thus, effect of FMO which is present for pure PPCA is not crucial for hybrid PPCA. The number of links affects the computational time in the following way: 900 links takes about 9-10 hours and 5,800 links (all links with functional class 1-4 of Stockholm inner city) 105 hours. Please note that calibration run relatively infrequently and computational time is not critical if it is in reasonable limits.

**RESULTS**

**Congestion Patterns and Prediction Output**

Figure 5 visualizes the congestion patterns and prediction residuals between predicted and observed link speeds for a selected afternoon peak interval. The difference between the speed limit and predicted values in Figure 5(a) shows that during the peak the difference is highest for links with functional class 3 and 4, which can be considered bottlenecks (see dark/red links in Figure 5(a)). Figure 5(b) shows the accuracy of the hybrid PPCA method, which has a tendency to be lower near to the crossroads or traffic lights.

**Prediction Performance**

Prediction accuracy is evaluated in terms of Mean Absolute Error (MAE). The MAE for link $k$ and interval $i$ is denoted as $\text{MAE}(i, k)$, which can be interpreted as the mean of the absolute difference between predicted $\hat{\omega}_{ikn}$ and observed $v_{ikn}$ values on the link $k$ and interval $i$ across $N$ days:

$$\text{MAE}(i, k) = \frac{\sum_{n=1}^{N} |\hat{\omega}_{ikn} - v_{ikn}|}{N}.$$  

Further, the mean absolute error for interval $i$ across all links and $N$ days is denoted as $\text{MAE}(i)$.

As shown in Figure 6(a), for all intervals on the aggregated level the hybrid PPCA method has the highest prediction accuracy compared to the historical mean and the pure PPCA method.
FIGURE 4 Values of calibrated parameters $Q, P, \alpha, \beta$ and calibration time across all time intervals.
Furthermore, the use of smoothing parameters $\alpha$ and $\beta$ improves the pure PPCA in all cases and allow to the hybrid PPCA to be better than the historical mean on aggregated level in all intervals. The pure PPCA performs better than the historical mean in 86% of intervals especially when FMO($i$) is large. However, the largest improvement of the hybrid PPCA with respect to the PPCA is in the intervals 14:00-19:00, when the mean RMD($i$) is largest. PPCA methods in comparison to the historical mean are capable to predict some of the variation patterns and produce up to 40% more accurate prediction with respect to the historical mean (see Figure 6(b)). For all methods the accuracy increases when the FMO($i$) and RMD($i$) decrease.

Time Efficiency

The two-stage framework and PPCA methodology allow computation and estimation of the prediction parameters on the historical observations independently. Thus, the travel time prediction takes only 0.6-1 second for all 1,886 links of the case study. Pure PPCA is usually under 0.2 second. The number of links effects the hybrid PPCA computational time in the following way: 900 links takes about 0.1 second and 5,800 links (all links with functional class 1-4 of Stockholm inner city) 4-5 seconds.

Effects of Missing Observations and Link Speed Variability

Figure 6 shows the average impact of FMO on prediction accuracy in $P$ past intervals, across the evaluation set. For all presented methods, accuracy is lower (large MAE($i$) values in Figure 6) when the FMO is larger, which is directly affecting the local smoothing in hybrid PPCA methods. In this case, due to the missing observation PPCA methods are not able adequately adapt to the bias. MAE($i,k$) and MAE($i$) tend to increase with larger fractions of missing observations Figure 6 , 7(c-d).

Results in Figure 6 show that the link speed variability RMD($i$) is affecting the accuracy of prediction methods. Accuracy advantage of PPCA methods is more evident in intervals 14:00-19:00 when the average RMD($i$) across all links is the highest. Figure 6(b) shows extracted results for links with functional class 2 and highlight that link speed variability RMD($i$) strongly affects the accuracy of all presented methods, more than FMO.
FIGURE 6 Prediction methods accuracy across all intervals for: (a) all case study links (b) functional class 2 with the largest RMD values.
FIGURE 7 Visualization of MAE\((i, k)\), RMD\((i, k)\) and fraction of missing observations effects on the individual links for interval 17:15-17:30 with large RMD\((i)\) and highest improvement in comparison to the historical mean. (a-b) show RMD\((i, k)\) effects on the MAE\((i, k)\). (c-d) illustrate the impact of the fraction of missing values on the MAE\((i, k)\).
Figure 7 shows patterns between link speed variability (RMD($i, k$)), fraction of missing observations (FMO($i, k$)) and prediction performance (MAE($i, k$)) for the interval with the highest RMD($i$). The hybrid PPCA method outperforms the historical mean significantly for functional classes 1 and 2 (Figures 7(a-b)), when RMD($i, k$) is larger than 0.30 and FMO($i, k$) is under 0.40.

CONCLUSIONS

The paper proposes a two-stage framework for real-time travel time prediction. In the first stage, correlation patterns are extracted from historical travel time data using a hybrid method of probabilistic principal component analysis (PPCA) and local smoothing, and calibrated model parameters are stored in the database. In the second stage, travel times are predicted in real-time by combining current-day data and the correlation patterns stored in the estimated parameters. Furthermore, a novel PPCA hybrid method that considers neighboring links in the local smoothing was proposed.

From the case study and numerical experiments several conclusions may be drawn. It was found that the proposed hybrid PPCA method produces the most precise and robust travel time predictions of all presented methods. The computational experiments also reveal that both PPCA methods (including pure PPCA) have significant advantages when the link speed variability represented by the relative mean absolute difference (RMD) for individual links is larger than 0.30, and the fraction of missing observations (FMO) is under 40%. Considering historical correlations by PPCA methods can improve the accuracy compared to the historical mean by up to 40%. This is especially useful since link speed variability tends to be larger during the peak hours, when prediction accuracy is the most important. The accuracy of PPCA methods and historical mean are very close if the link speed variability is small (large probability that value is close to the mean) or if the fraction of missing observations is large (when PPCA is unable to make use of historical correlations and residuals from current-day observations). In this case, the pure PPCA method may perform worse than the historical mean. However, the hybrid PPCA method performs on the aggregated interval level always better than historical mean, in our case study.

Calibration experiments reveal the potential of hybrid PPCA and smoothing parameters $\alpha$ and $\beta$ but it can be easily overcalibrated. If calibrated separately for each link and calibration day the $\alpha$ and $\beta$ can be set in that way that prediction error is under 1 km/h and can be better than pure PPCA (about 5 km/h or more). However, in order to avoid overfitting the model parameters are calibrated to the whole set, and because of large link speed variability across different days in our case study, calibration sets the parameters to small values to fit best to the different days. Therefore the prediction results for hybrid PPCA are very close to the pure PPCA for this data set. On the other hand, where there are large proportions of missing data, the hybrid PPC proves more robust than the pure PPCA. In future work, more sophisticated methods such as estimating parameter sets $\alpha$ and $\beta$ for clusters of link and days may help to use the full potential of hybrid PPCA method.

The proposed real-time travel time prediction framework is producing predictions for all 1,886 links of case study in time less than 1 second which makes it a great candidate for real-time urban network prediction for large-scale networks. Although calibration of the hybrid PPCA takes significantly more computational time than the pure PPCA, this run relatively infrequently and therefore not time-critical. The average prediction error is under 5 km/h which is less than uncertainties in the input data (e.g. link speed estimation accuracy).

The average and maximum prediction residuals grow as the link speed variability increases. In future work this dependence can be used to assess the size of residuals associated with the link
or time-of-day interval that can be compared with other input parameters and help designers to identify parts of the network where the historical mean or PPCA methods are insufficient. Such problematic parts might require more sophisticated methods, such as combinations of PPCA and simulation (Dynamic Traffic Assignment (DTA)) models for traffic prediction (12, 13), or incorporation of other data sources to provide more accurate real-time travel time prediction for large-scale networks.

ACKNOWLEDGMENT
The work was funded in part by the Swedish Transport Administration through the Mobile Millennium Stockholm project, and TRENoP Strategic Research Area. The authors would like to thank Tomas Julner and Per-Olof Svensk at the Swedish Transport Administration for their support and provision of the data to the iMobility Lab at KTH.

REFERENCES


