T-ITS-22-02-0533

Filtering Limited Automatic Vehicle Identification Data for Real-time Path Travel Time Estimation without Ground Truth

Ang LI, William H.K. LAM, Wei MA, Andy H.F. CHOW, S.C. WONG, and Mei Lam TAM

***Abstract*—Automatic Vehicle Identification (AVI) technology has been widely used for real-time path travel time estimation. For a study path equipped with AVI sensors at both ends, the difference between the timestamps of vehicles entering and leaving the path is AVI data. In urban areas, there can be several alternative routes and vehicle entry/exit points for the study path. Consequently, invalid AVI data occur that fall outside the scope of the travel time of the study path. Some AVI technologies based on identification information of vehicles can match vehicles precisely. However, for cities like Hong Kong with concerns of privacy issues, only commercial vehicle data can be collected. Under this scenario, the resultant AVI data are accurate but with few valid samples in a relatively short time interval due to the unavailability of private car data. The estimation accuracy of path travel times on a real-time basis will then be affected significantly by the existence of invalid AVI data. In this paper, a novel unsupervised algorithm is proposed to filter out real-time invalid AVI data efficiently although there is no ground truth available for training purposes. It is tested and compared with other benchmark algorithms on two selected paths in the Hong Kong urban road network. It is found that the proposed unsupervised algorithm can still filter limited but accurate AVI data with satisfactory performance. Sensitivity tests with ground truth are also conducted with different sampling rates. Some insightful findings are given for filtering AVI data under various scenarios.**

***Index Terms*—data filtering, functional principal component analysis, automatic vehicle identification, advanced traveler information systems**

I. Introduction

OVER the past two decades, automatic vehicle identification (AVI) data have been increasingly explored for use in advanced traffic management systems (ATMSs) and advanced traveler information systems (ATISs). These data are collected using various AVI sensors, such as Radio Frequency Identification (RFID) tag readers, automatic license plate recognition (ALPR) cameras, and Bluetooth MAC address readers. Thus, a vehicle passing an AVI sensor has its specific identifiers (e.g., RFID tags for RFID tag readers and license plate numbers for ALPR cameras) and the corresponding timestamp recorded. These data from successive AVI sensors are matched to the vehicle and used to calculate its travel time [1]–[4], which is denoted as AVI data.

According to the uniqueness of the identifier of each vehicle, there are two types of AVI data. On the one hand, Bluetooth sensors can collect numerous AVI data. However, the MAC address identified by Bluetooth sensors can be provided by either vehicles, or passengers within the same vehicle and even pedestrians on the roadside through their mobile devices. Therefore, the collected AVI data is inaccurate. On the other hand, for AVI technologies requiring identifier information (e.g., RFID and ALPR), AVI data are collected accurately but the sampling rate is very few in a relatively short time interval due to privacy issues.

The availability of identifier information in the database depends on the corresponding privacy issues concerned by different cities [5], [6]. In Hong Kong, only AVI data on commercial vehicles are available for collection. The sampling rate is relatively low without the collection of AVI data on private cars. As a path is defined as the corridor of interest between two AVI sensors and path travel time refers to the time required to transverse the path, these AVI data can be used for real-time estimation of path travel times, and the estimates can then be supplied to travelers and management authorities [7]. In this paper, real-time path travel time estimation is referred to as an estimated path travel time for the current time interval on the current day.

Some AVI records from RFID tag readers and ALPR cameras may be inappropriate for real-time path travel time estimation. Similar to the data cleaning process of GPS data that can accurately capture the trajectory of vehicles for travel time estimation [8]–[14], AVI data also need data preprocessing before real-time path travel time estimation.

As discussed by researchers such as [1] and [15], errors may arise from vehicles being misidentified, stopping en-route (e.g., see Fig. 1), or choosing unusually long routes (e.g., detours) between two locations that are equipped with AVI sensors. Thus, invalid data (or outliers) are most often obtained if AVI sensors (i) are far apart, implying that vehicle detours or stops are more frequent, or (ii) contain many short-spacing intersections and frequent frontage access (which explains why

This work was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU R5029-18, R7027-18, PolyU/25209221, and PolyU/15206322), and a grant from Dean's Reserve at the Hong Kong Polytechnic University (Project No. P0034271) *(Corresponding author: Wei MA)*.

Ang LI, William H.K. LAM, Wei MA, and Mei Lam TAM are with the Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China (email: ang-leon.li@connect.polyu.hk; william.lam@polyu.edu.hk; wei.w.ma@polyu.edu.hk; trptam@polyu.edu.hk).

Andy H.F. CHOW is with the Department of Systems Engineering, City University of Hong Kong, Hong Kong SAR, China (email: andychow@cityu.edu.hk).

S.C. WONG is with the Department of Civil Engineering, The University of Hong Kong, Hong Kong SAR, China (e-mail: hhecwsc@hku.hk).

it can be more difficult to obtain valid AVI data from urban roads than from freeways).

The path travel times derived from an AVI system under these circumstances can be regarded as invalid AVI data, which must be removed by novel filtering algorithms to extract valid AVI data for use in real-time path travel time estimation. Fig. 1 illustrates one scenario in which AVI data from an AVI system may be invalid. Vehicle B travels to a petrol station after being detected by AVI sensor 1, and hence the travel time of vehicle B is much longer than that of vehicle A; accordingly, the AVI data collected from vehicle B is invalid.

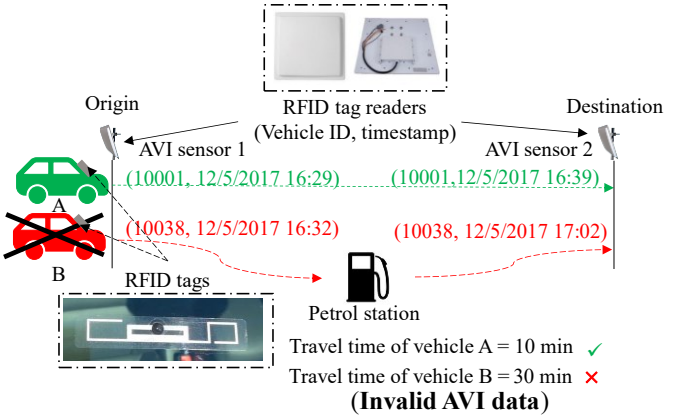


Fig. 1. Example of invalid AVI data in an AVI system.

AVI data is categorized as real-time AVI data or historical AVI data, depending on when it is collected. In this paper, real-time AVI data are collected in the current and previous time intervals on the current day, while historical AVI data are acquired on previous days. Both these two data categories contain valid and invalid AVI data. As real-time AVI data are generally used for real-time path travel time estimation [16]–[18], it is critically important to remove invalid real-time AVI data by novel filtering algorithms to enable real-time path travel time estimation.

Furthermore, for AVI technologies based on identifier information in cities with privacy issues, the collected AVI data are accurate with fewer samples. It is more challenging to distinguish invalid real-time AVI data from limited accurate real-time AVI data with a low sampling rate. Therefore, this paper focuses more on filtering accurate but limited real-time AVI data.

Data-filtering and outlier-detection algorithms have been developed for other traffic variables, including flow [19] and speed [20]. These algorithms assume that most data are valid and hence they remove only small portions of invalid data [21]. However, for AVI data with low sampling rates, its distribution can be more scattered and varied. Thus, a large proportion of AVI data may be invalid. Furthermore, the occurrence of longer travel times by path is more frequent when traffic is congested. It is a challenge to distinguish invalid data from comparatively long travel times by path under this scenario [22]. As existing filtering algorithms only make use of real-time AVI data, the resultant time windows lack rigorous mathematical guarantees, particularly for limited real-time AVI data with low sampling rates. Therefore, existing filtering algorithms may not be effective.

There are some existing offline algorithms for the filtering of historical AVI data. These algorithms are devoted to the data clustering or modeling of travel time distributions using a large amount of historical AVI data [23]–[27]. However, these algorithms lack sufficient computation time to provide time windows for filtering real-time AVI data. Consequently, various data-filtering algorithms have been developed to screen out invalid real-time AVI data in real-time applications.

A summary of the existing algorithms for filtering real-time AVI data is presented in Table I. It indicates that the performance of existing filtering algorithms depends largely on real-time AVI data, which means that their performance drastically decreases if the collected real-time AVI data are limited. Therefore, there is a need for a novel filtering algorithm capable of effectively extracting real-time AVI data, especially when they are limited. The use of historical AVI data including both valid and invalid AVI data collected in previous days can be very helpful in this regard.

The effect of complex network structure in urban areas has not been investigated in most previous studies on AVI data filtering. Attention has been mainly given to freeways [28], which have relatively simple network topologies and very few entries and exits between pairs of AVI sensors. Moreover, the numerous entries, exits, and bus stops along the urban study paths (as used in the case study) may indicate that valid real-time AVI data can be limited, which adversely affects the performance of existing filtering algorithms used for real-time path travel time estimation.

When valid real-time AVI data is limited, it is worthwhile to investigate temporal variance-covariance (var-cov) relationships between path travel times at different time intervals and on different days from historical AVI data. They are significantly beneficial for filtering out invalid real-time AVI data and for real-time path travel time estimation.

It is the first-time historical AVI data is used for filtering real-time AVI data, as shown in Table I. A novel filtering algorithm is proposed to filter out invalid real-time AVI data for real-time path travel time estimation, without ground truth for training purposes. As no ground truth is used for training, it is also referred as to the proposed unsupervised algorithm in this paper.

The proposed unsupervised algorithm is particularly useful when privacy policies prohibit the availability of many valid AVI data from privately-owned vehicles (e.g., Hong Kong only allows the collection of AVI data from commercial vehicles. These commercial vehicles include goods vehicles, non-franchised and franchised buses, and private cars owned by commercial companies, which account for approximately 19% of the total vehicle fleet in Hong Kong[1]) for utilization in the development of various intelligent transportation systems (ITS).

[1]https://www.td.gov.hk/en/transport_in_hong_kong/transport_figures/index.html

Furthermore, most existing filtering algorithms use simple first-order central tendency measures, such as observed mean or median values, of AVI data. In contrast, the proposed unsupervised algorithm considers both first- and second-order statistical properties of AVI data via a functional principal component analysis (FPCA). The mean and standard deviation of estimated path travel times by FPCA can help to construct a dynamic time window for filtering out invalid real-time AVI data for real-time path travel time estimation on urban arterials.

TABLE I
SUMMARY OF FILTERING ALGORITHMS FOR REAL-TIME AVI DATA

Literature	Input data	Time window	Path travel time estimates	Road type	Distance between two AVI sensors (km)	Type of AVI sensors	Interval of updating time window (min)
[29]	R	Distribution center	-	Highway (3 km)	0.9–3.7	Dedicated Short Range Communication sensors	5
[30]	R	Distribution center	Mean from data	Freeway	1.6	Bluetooth MAC address readers	5
[31]	R	Median and variance	Median from data	Urban arterials	-	ALPR cameras	2 and 5
[32]	R	Mean and variance, and transition identification	Mean from data	Urban arterials	6.2	RFID tag readers	2
[33]	R	Mean and variance, and transition identification	Mean from data	Freeway and urban arterial	4.0 (freeway) and 1.9 (urban arterial)	RFID tag readers	2
[34]–[36]	R	Mean	Mean from data	Freeway	-	RFID tag readers	0.5, 2, and 15
This paper	R + H (First time)	Statistical model	Conditional mean from model	Urban arterials	4.3, 4.5, and 9.2	RFID tag readers	2

Note: **R** = real-time AVI data; **H** = historical AVI data.

FPCA is a statistical tool for functional data analysis that uses advanced feature approximation techniques. It has received increasing attention in recent related studies, as it can be used for analyzing highly stochastic data. For example, [37] proposed an FPCA model to predict traffic flows, and [38] and [39] have used FPCA to identify and monitor traffic patterns. In addition, [40] applied FPCA to model the variability and reliability of freeway travel times. Furthermore, [41] performed FPCA of global positioning system data to predict vehicle speed distributions. [37] and [38] adopted FPCA to satisfactorily predict and estimate link and path travel time variations. Moreover, [43] further highlighted the merits of FPCA on path travel time predictions under abnormal traffic conditions.

The FPCA model regards the path travel time as a stochastic process [42]–[44]. In this paper, the FPCA model has been extended to generate temporal var-cov relationships between path travel times. These relationships are then used to develop the proposed unsupervised algorithm for filtering limited but accurate real-time AVI data, which enables the real-time estimation of path travel times without ground truth for training purposes.

The major contributions of this paper are summarized into the following three categories.

1) It is the first time that a novel unsupervised algorithm is proposed, with the usage of historical AVI data but without using historical ground truth for training purposes, for constructing dynamic time windows to filter out invalid real-time AVI data from limited real-time AVI data.

2) A FPCA-based model is adapted to consider both the historical and real-time AVI data for modeling their temporal var-cov relationships between path travel times at different time intervals and on different days. Both mean and standard deviation of the path travel times are estimated by the proposed FPCA model and used for the improvement of the real-time AVI data filtering performance.

3) Sensitivity tests are conducted to examine the effects of different sampling rates of the real-time AVI data or the valid real-time AVI data only in order to verify the generality and robustness of the proposed unsupervised algorithm without or with the use of the ground truth for training purposes.

The remainder of this paper is organized as follows. Section II presents the methodology of the proposed unsupervised algorithm for filtering real-time AVI data. Section III reports the numerical results obtained for the two case studies in Hong Kong by applying the proposed unsupervised algorithm in comparison with the other three corresponding existing algorithms for screening out invalid real-time AVI data, together with sensitivity tests on the sampling rates of real-time AVI data and valid real-time AVI data on estimation accuracy without and with ground truth. Finally, concluding remarks and suggestions for future research are given in Section IV.

II. METHODOLOGY

To illustrate the essential ideas of the methodology, any given path with two AVI sensors at both ends is considered. In this setting, the i^{th} AVI data measured on the day d is denoted as $x_{i,d}$. The corresponding timestamp of $x_{i,d}$ when it is collected is represented by $t_{i,d}$. The set of days with historical AVI data is defined as D, while the current day is denoted as d^*. The assignment of d from set D or $\{d^*\}$ depends on whether it is historical or real-time AVI data.

The proposed unsupervised algorithm aims to provide a dynamic time window for screening out invalid real-time AVI data. The dynamic time window consists of the upper bound $U(t^*)$ and the lower bound $L(t^*)$, where t^* is the timestamp when real-time filtering is performed. In the proposed

unsupervised algorithm, the available data are the real-time AVI data before time t on the current day $x_{d^*,i}^{AVI}$, for $t_{d^*,i}^{AVI} \le t$, and historical AVI data $x_{d,i}^{AVI}$, for $d \in D$.

A. Proposed Unsupervised Algorithm

Fig. 2 provides the framework of the proposed unsupervised algorithm. There are two stages and five steps in the framework with corresponding equation numbers shown at each step. Detailed descriptions of these two stages and five steps are given in the following paragraphs.

Stage 1 involves offline training, which uses historical AVI data for the development of the trained FPCA models. Stage 2 concerns the real-time filtering of real-time AVI data, in which the trained FPCA models are used to construct the dynamic time windows to screen out invalid real-time AVI data.

As the backbone of the methodology framework, FPCA models are trained to map the predictor to the response [43]. Thus, the eigenfunctions and principal components must be trained for the predictor and the response. Then, the conditional distributions of the response based on the predictor can be obtained and represented by the trained eigenfunctions and principal components [41], [45], [46]. In the proposed unsupervised algorithm, the historical AVI data is considered as the predictor, and the response is the offline path travel time estimated from sufficient historical AVI data using existing filtering algorithms, such as TransGuide algorithm [34]. They are preliminary travel time estimates shown in Step 1 without the use of ground truth. Therefore, the i^{th} preliminary travel time estimate is denoted as $x'_{i,d}$, with the corresponding timestamp denoted as $t'_{i,d}$.

Selecting appropriate training set is performed in Step 2. It is based on the temporal var-cov relationships between the path travel times on different days. Afterward, learning training sets is proceeded with the modeling of the temporal var-cov relationships between the path travel times at different time intervals in Step 3. Dynamic time windows are constructed based on the mean and standard deviation of estimated path travel times provided by the proposed FPCA model in Step 4. The resulting FPCA models are adopted for real-time filtering to determine the dynamic time windows in a rolling horizon scheme for the latest real-time AVI data in Step 5.

B. Selecting Appropriate Training Set

Historical AVI data may reflect different traffic patterns due to the changing traffic demand and network supply (e.g., incidents and sensor failures). If the traffic patterns are different from that of the current day, then those historical AVI data may provide little useful information for constructing the current day's dynamic time windows. Accordingly, historical AVI data that contain similar traffic patterns to the current day are selected for filtering real-time AVI data.

To this end, the temporal var-cov relationships between path travel times on different days are modeled by FPCA to reflect the similarities of traffic patterns across multiple days. At time t, historical AVI data at time $t_{i,d} \in [t-T, t]$ is considered, where the T is the length of the study horizon and it is the unit for the rolling horizon scheme presented later. The AVI data $x_{i,d}$ is the sum of travel time and measurement error $\varepsilon_{i,d}$, and are given by (1), as follows:

$$x_{i,d} = \mu_X(d) + \sum_{k=1}^{K^D} \xi_k^D \phi_k(d) + \varepsilon_{i,d} \quad (1)$$

where $\mu_X(d)$ is the mean function of travel times from AVI data on day d, which is $X(d)$; ξ_k^D is the score of the k^{th} functional principal component; $\phi_k(d)$ is the eigenfunction of the k^{th} functional principal component from AVI data on day d for $|D|$ days according to the Karhunen-Loève representation; and K^D is the number of functional principal components from AVI data for $|D|$ days, where $\varepsilon_{i,d}$ represents the measurement error of the i^{th} travel time from AVI data on day d.

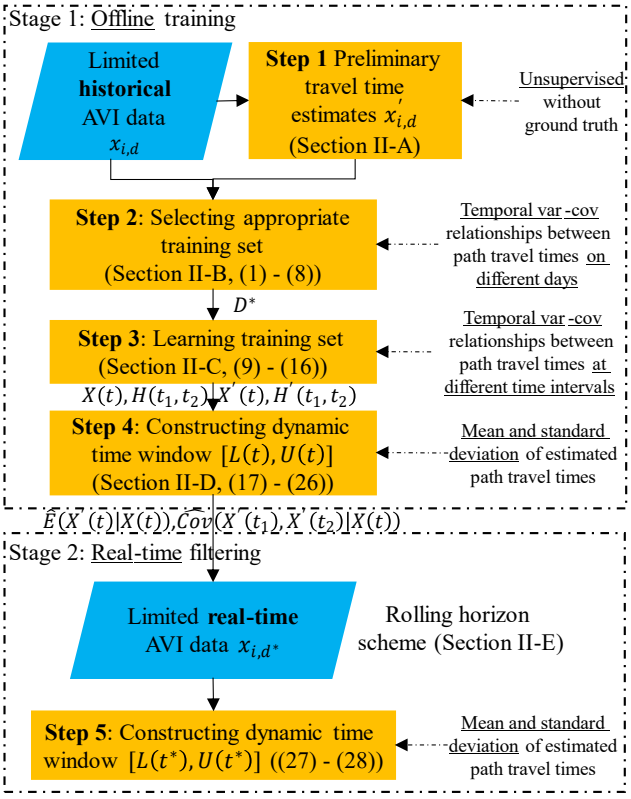


Fig. 2. Framework of the proposed unsupervised algorithm.

Equation (1) assumes that the path travel time data in $[t-T, t]$ is continuous in d, and the corresponding path travel time from the AVI data $X(d)$ is given by (2), as follows:

$$X(d) = \mu_X(d) + \sum_{k=1}^{K^D} \xi_k^D \phi_k(d) \quad (2)$$

where the function $\mu_X(d)$ is given by

$$\mu_X(d) = E\big(X(d)\big) \quad (3)$$

The covariance function of path travel times from AVI data between day d_1 and d_2 (temporal var-cov relationships between path travel times on different days) is denoted by $H(d_1, d_2)$ and is provided by (4), as below:

$$H(d_1, d_2) = \sum_{k=1}^{K} \lambda_k^D \phi_k(d_1)\, \phi_k(d_2) \quad (4)$$

where λ_k^D is the eigenvalue of the k^{th} functional principal component from AVI data.

It is assumed that the weighting or score of the functional principal component ξ_k^D has the statistical properties given by (5) and (6), as below:

$$E\big(\xi_k^D\big) = 0 \quad (5)$$

$$Var\left(\xi_k^D\right) = \lambda_k^D \tag{6}$$

The covariance $H(d_1, d_2)$ is derived by solving the following minimization (7) for the AVI data:

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{1\le d_3\le d_4}^{|D|} \sum_{i=1}^{N_d} \kappa_C\left(\frac{d_3-d_1}{h_C}\right)\kappa_C\left(\frac{d_4-d_2}{h_C}\right)\cdot \begin{pmatrix} \widehat{Cov}\left(x_{i,d_3}, x_{i,d_4}\right) - \beta_0 - \beta_1(d_3-d_1) \\ -\beta_2(d_4-d_2) \end{pmatrix}^2 \tag{7}$$

where $\widehat{Cov}\left(x_{i,d_3}, x_{i,d_4}\right)$ represents the estimated travel time covariance between day d_3 and d_4, the estimates of the model coefficients $\beta_0, \beta_1, \beta_2$ are dependent on days d_1 and d_2, and N_d is the number of samples within the study horizon on day d. The estimates of β_0 are denoted as $\hat{\beta}_0(d_1, d_2)$ and an estimate of $H(d_1, d_2)$ is obtained from $\widehat{H}(d_1, d_2) = \hat{\beta}_0(d_1, d_2)$. Moreover, κ_C is a kernel function in which h_C is the bandwidth that enables calibration of the covariance function.

Referring to the covariance function of path travel times for different days $\widehat{H}(d_1, d_2)$, the samples with larger covariance values are selected and used to calibrate the model. $D^*$is the set of days after sample selection, which is determined by (8):

$$D^* = \left\{d \,\middle|\, \left|\widehat{H}(d_1, d_2)\right| \ge H^*, d \in D\right\} \tag{8}$$

where H^* is the threshold of the path travel time covariance between different days.

C. Learning Training Set

In this section, two FPCA models are proposed to model temporal var-cov relationships between path travel times at different time intervals. The first FPCA model is based on the predictor (i.e., historical AVI data). The second FPCA model is based on the responses, which are preliminary travel time estimates in the proposed unsupervised algorithm.

The historical AVI data $x_{i,d}$ is modeled in (9), as follows:

$$x_{i,d} = \mu_X\left(t_{i,d}\right) + \sum_{k=1}^{K^T} \xi_k \phi_k\left(t_{i,d}\right) + \varepsilon_{i,d}, d \in D^* \tag{9}$$

where $\mu_X\left(t_{i,d}\right)$ is the mean function of the measured travel times at time $t_{i,d}$; ξ_k represents the score/weight of the k^{th} functional principal component; $\phi_k\left(t_{i,d}\right)$ is the eigenfunction of the k^{th} functional principal component from AVI data at time $t_{i,d}$; K^T is the number of functional principal components from AVI data during study horizon T.

Analogously, the path travel time based on AVI data $X(t)$ can be described as (10) below:

$$X(t) = \mu_X(t) + \sum_{k=1}^{K^T} \xi_k \phi_k(t) \tag{10}$$

where $\mu_X(t)$ is given by (11), as below:

$$\mu_X(t) = E\left(X(t)\right) \tag{11}$$

$H(t_1, t_2)$ is denoted as the covariance function of path travel times from AVI data between time t_1 and t_2 (temporal var-cov relationships between path travel times at different time intervals) in (12), as below:

$$H(t_1, t_2) = \sum_{k=1}^{K^T} \lambda_k \phi_k(t_1)\, \phi_k(t_2) \tag{12}$$

Again, the weighting/score of functional principal components has the same statistical properties as shown in (5) and (6).

If a response $x'_{i,d}$ is available at time $t'_{i,d}$ on day d, (9) can be expressed as (13):

$$x'_{i,d} = \mu_{X'}\left(t'_{i,d}\right) + \sum_{k=1}^{K'^T} \xi'_k \phi'_k\left(t'_{i,d}\right),\ d \in D^* \tag{13}$$

where $\mu_{X'}\left(t'_{i,d}\right)$ is the mean function of responses over study horizon T; ξ'_k is the score/weight of the k^{th} functional principal component of the responses; $\phi'_k\left(t'_{i,d}\right)$ is the eigenfunction of the k^{th} functional principal component of the responses at time $t'_{i,d}$; and K'^T is the number of functional principal components of the responses during study horizon T.

Correspondingly, the path travel time based on the responses can be expressed as (14), as below:

$$X'(t) = \mu_{X'}(t) + \sum_{k=1}^{K'^T} \xi'_k \phi'_k(t) \tag{14}$$

where $\mu_{X'}(t)$ given by (15), as follows:

$$\mu_{X'}(t) = E\left(X'(t)\right) \tag{15}$$

$H'(t_1, t_2)$ is denoted as the covariance function of path travel times for responses between time t_1 and t_2 during study horizon T (temporal var-cov relationships between path travel times at different time intervals), as below:

$$H'(t_1, t_2) = \sum_{k=1}^{K'^T} \lambda'_k \phi'_k(t_1)\, \phi'_k(t_2) \tag{16}$$

where λ'_k is the eigenvalue of the k^{th} functional principal component of responses.

The predictors $x_{i,d}$ and responses $x'_{i,d}$ can be used to calibrate the above-described FPCA-based models. The details of the procedure for calibrating mean functions, covariance functions, and functional principal components (including weighting/score and eigenfunctions) are available in the literature [41], [43], [45], [46]. The number of functional principal components is generally determined by applying one of the following three methods: the fraction of variance explained, the Akaike information criterion, or the Bayesian information criterion.

D. Constructing Dynamic Time Window

The principal analysis by conditional expectation (PACE) is now formulated for the FPCA models presented in the previous section, for use in data filtering. The objective is to relate the models derived from the predictors and the responses via the method of additive models [41], [43], [46]. Specifically, the conditional distributions of the responses derived from the AVI data are adopted. The advantage of this PACE approach is its superiority over other approaches under the Gaussian assumption [47].

Application of the functional additive model [46] provides the conditional model (17), as below:

$$E\left(X'(t)\middle|X(t)\right) = \mu_{X'}(t) + \sum_{q=1}^{K'^T}\left(\sum_{k=1}^{K^T} E\left(\xi'_q\middle|\xi_k\right)\right)\phi'_q(t) \tag{17}$$

Similar to the calibration procedure adopted in the general FPCA model, $f_{qk}(\xi_k) = E\left(\xi'_q\middle|\xi_k\right)$ on each day d, $f_{qk}(\xi)$ can be obtained by minimizing the following expression with respect to γ_0 and γ_1:

$$\min_{\gamma_0,\gamma_1} \sum_{d\in D^*} \kappa_f\left(\frac{\hat{\xi}_{k,d}-\xi}{h_f}\right)\left[\hat{\xi}'_{k,d} - \gamma_0 - \gamma_1\left(\xi - \hat{\xi}_{k,d}\right)\right]^2 \tag{18}$$

where $\hat{\xi}_{k,d}$ and $\hat{\xi}'_{k,d}$ are the estimated ξ_k and ξ'_k, respectively, on each day d. This leads to $\hat{f}_{qk}(\xi) = \hat{\gamma}_0(\xi)$. Moreover, the conditional covariance function is given by (19), as follows:

$$Cov\left(X'(t_1), X'(t_2)\middle|X(t)\right) = \sum_{q=1}^{K'^T} var\left(\xi_q'\middle|X(t)\right)\phi_q(t_1)\,\phi_q(t_2) \quad (19)$$

By using the property of variance, $var\left(\xi_q'\middle|X(t)\right)$ can be further expanded such that (19) can be rewritten as (20) below:

$$\begin{aligned} &Cov\left(X'(t_1), X'(t_2)\middle|X(t)\right) \\ &= \sum_{q=1}^{K'^T}\Big[var(\xi_q') + \sum_{k=1}^{K^T} E\left(\left(\xi_q'\right)^2 - \right. \\ &\left. var(\xi_q'|\,\xi_k)\right) - E^2(\xi_q'|\,\xi_k)\Big]\cdot\phi_q(t_1)\phi_q(t_2) \\ &= H'(t_1,t_2) + \\ &\sum_{q=1}^{K'^T}\sum_{k=1}^{K^T}[g_{qk}(\xi_k) - f_{qk}^2(\xi_k)]\phi_q(t_1)\phi_q(t_2) \end{aligned} \quad (20)$$

where $g_{qk}(\xi_k)$ is given by (21), as follows:

$$g_{qk}(\xi_k) = E\left[\left(\xi_q'\right)^2 - var(\xi_q'|\xi_k\right] \quad (21)$$

By setting $f_{qk}(\xi_k) = \hat{f}_{qk}(\xi_k)$, an estimate of $g_{qk}(\xi_k)$ can be further acquired by minimizing the following expression (22) with respect to η_0 and η_1:

$$\min_{\eta_0,\eta_1} \sum_{d\in D^*} \kappa_g\left(\frac{\hat{\xi}_{k,d}-\xi_k}{h_g}\right)\left[{\hat{\xi}_{q,d}'}^2 - var(\hat{\xi}_{q,d}') - \eta_0 - \eta_1(\xi_k - \hat{\xi}_{k,d})\right]^2 \quad (22)$$

which leads to $\hat{g}_{qk}(\xi_k) = \hat{\eta}_0(\xi_k)$.

The conditional mean of responses based on path travel times derived from AVI data and the conditional covariance of responses based on travel times from AVI data can be modeled as (23) and (24), respectively, as follows:

$$\hat{E}\left(X'(t)\middle|X(t)\right) = \hat{\mu}_{X'}(t) + \sum_{q=1}^{K'^T}\left(\sum_{k=1}^{K^T}\hat{f}_{qk}(\xi_k)\right)\hat{\phi}_q'(t) \quad (23)$$

$$\widehat{Cov}\left(X'(t_1), X'(t_2)\middle|X(t)\right) = \sum_{q=1}^{K'^T}\left(var(\hat{\xi}_q') + \sum_{k=1}^{K^T}\left(\hat{g}_{qk}(\xi_k) - \hat{f}_{qk}^2(\xi_k)\right)\right)\hat{\phi}_q'(t_1)\hat{\phi}_q'(t_2) \quad (24)$$

The conditional mean and covariance function of responses derived from path travel times determined from AVI data can be obtained from (23) and (24) by learning from historical information on the predictors and responses.

Proposition 1 presents the uniform convergence properties of the conditional model of path travel time estimation. The conditional mean and covariance of estimated path travel times are accurate when the number of path travel time data $|D^*| \to +\infty$. If more principal components are considered (i.e., K^T, K'^T is large), more data samples are required.

Proposition 1. (The uniform convergence of the conditional modeling of path travel time)

Suppose that the number of travel time data $|D^*| \to +\infty$ and the path travel time on each day in D^* are i.i.d., and that the mean $\hat{E}\left(X'(t)\middle|\varphi(t)\right)$ and the covariance $\widehat{Cov}\left(X'(t_1), X'(t_2)\middle|X(t)\right)$ in the calibrated conditional model of travel time in (23) and (24) approximate the actual conditional mean and covariance with the error rate $O_p\left(\frac{K^T K'^T}{\sqrt{|D^*|}}\right)$. Thus, mathematically (25) and (26) are presented:

$$\sup_{t\in T}\left|\hat{E}\left(X'(t)\middle|X(t)\right) - E\left(X'(t)\middle|X(t)\right)\right| = O_p\left(\frac{K^T K'^T}{\sqrt{|D^*|}}\right) \quad (25)$$

$$\sup_{t\in T}\left|\widehat{Cov}\left(X'(t_1), X'(t_2)\middle|\varphi(t)\right) - Cov\left(X'(t_1), X'(t_2)\middle|\varphi(t)\right)\right| = O_p\left(\frac{K^T K'^T}{\sqrt{|D^*|}}\right) \quad (26)$$

The corresponding proofs can be found in Appendix 1 in the online supplement[2].

Based on the conditional mean and covariance function of path travel times, $U(t^*)$ and $L(t^*)$ as upper and lower bounds the dynamic time windows can be obtained from (27) and (28):

$$U(t^*) = \hat{E}\left(X'(t^*)\middle|X(t^*)\right) + Z_{\alpha/2}\cdot\widehat{Cov}(X'(t_1^*), X'(t_2^*)|\varphi(t^*))^{\frac{1}{2}} \quad (27)$$

$$L(t^*) = \hat{E}\left(X'(t^*)\middle|X(t^*)\right) - Z_{\alpha/2}\cdot\widehat{Cov}(X'(t_1^*), X'(t_2^*)|\varphi(t^*))^{\frac{1}{2}} \quad (28)$$

The invalid real-time AVI data can then be filtered out if they are not falling within the dynamic time window $[L(t^*), U(t^*)]$. At each time interval t^*, the time window is updated based on the rolling horizon scheme, as detailed in the following section.

E. Rolling Horizon Scheme

The rolling horizon scheme is adopted following previous real-time applications [43], [48]. The dynamic time windows governed by $U(t^*)$ and $L(t^*)$ (which are determined from the proposed unsupervised algorithm) are updated in each time interval t^*, when these new real-time AVI data are streamed for filtering. The filtering framework generates the dynamic time windows for each rolling step (i.e., 2 min in this paper) using the flexible and adaptive rolling horizon (study horizon) T. In contrast, most existing data filtering algorithms have adopted a fixed rolling horizon for their applications [43], [48].

III. Numerical Experiments

In this section, the proposed unsupervised algorithm is examined in case studies of two selected paths using real-world data collected from the Hong Kong urban road network.

A. Traffic Data

The historical ground truth travel time data are obtained from the Hong Kong Journey Time Indication System (JTIS), the path travel time estimates of which have been independently validated using floating car survey data [32], [49]. The path travel time estimates provided by JTIS are instantaneous travel times. An example of the real-time information supplied by the JTIS [50] is depicted in Fig. 3. The numbers displayed in the digital signs are journey times (or path travel times) in minutes from the locations of these signs to the exits of the corresponding road tunnels crossing Victoria Harbor in Hong Kong. The colors of the digits in the display panel represent the congestion levels of each route: red digits indicate congested traffic (<25 km/h), yellow digits imply slow traffic (25–50 km/h), and green digits reveal free-flowing traffic (>50 km/h).

As there are a limited number of AVI sensors (RFID tag readers) in the JTIS, and the average distance between these sensors is relatively long, the rates at which AVI data are

[2]https://www.dropbox.com/s/07twpatvmr2rxrm/Online%20supplement.pdf?dl=0.

sampled in the JTIS are very low. Accordingly, some point sensors are also deployed in the JTIS to provide additional data at selected locations along major paths in urban areas. These point sensors collect the point speed data of vehicles traveling along the major paths. The combination of AVI and point sensor data enables the JTIS to generate updated real-time path travel time estimates along major routes in Hong Kong urban areas once every 2 min [49]. As reported, independent floating car surveys have confirmed the validity of JTIS path travel time estimates [32], [49]. Hence, the path travel time information provided by the JTIS is regarded as the ground truth for this study.

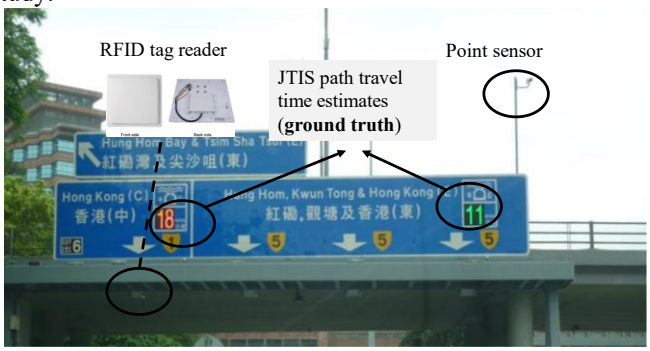


Fig. 3. Illustration of the Journey Time Indication System in Hong Kong.

B. Experimental Set-ups

Case studies on two selected paths in the Hong Kong urban road network are performed using real-world data. Fig. 4 and Table II show the locations and characteristics of these two selected paths, respectively.

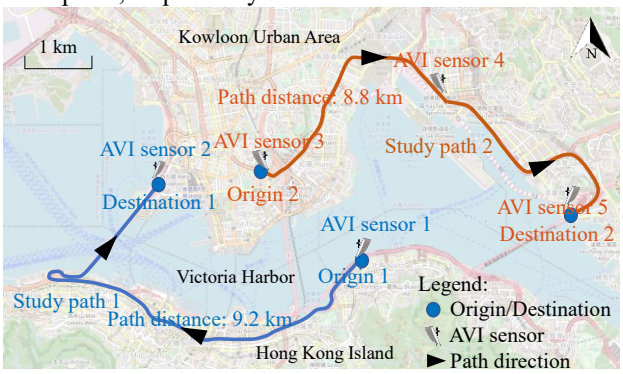


Fig. 4. Overview of the two study paths in Hong Kong.

Study path 1 is 9.2-km long and connects the Island Eastern Corridor on Hong Kong Island to the Western Harbor Crossing in Kowloon; its free-flow path travel time is 8.4 min. Study path 2 is 8.8-km long and connects Gascoigne Road and the entry of the Eastern Harbor Crossing; its free-flow travel time is 7.9 min. A pair of AVI sensors are installed at both ends of both paths. An additional AVI sensor 4 is installed in the middle of study path 2 to collect more AVI data in order to enable the real-time estimation of path travel times.

These two study paths differ primarily in the number of AVI sensors and bus stops in Table II. In addition, there is a signalized intersection on study path 1 but not on study path 2. The study paths contain several bus stops and frontage access with entries and exits. These site characteristics can lead to very few valid real-time AVI data available for real-time path travel time estimation. Fig. 5(a) shows the low sampling rates of valid real-time AVI data for both paths. Based on descriptions of [33], a low sampling rate refers to as representatively two or three AVI data per 2-min time interval. However, in the case study as shown in Fig. 5(a), there are only 12% and 30% of 2-min time intervals with no less than 2 valid real-time AVI data on study paths 1 and 2 respectively. It can be seen in Fig. 5(a) and Table II that the existence of signalized intersections and more bus stops on study path 1 further decreases the sampling rates of valid real-time AVI data. Moreover, there are more than 50% of 2-min time intervals without any valid real-time AVI data or real-time AVI data from Fig. 5(a) and Fig. 5(b). The latter consists of both valid and invalid AVI data.

The AVI data and JTIS ground truth collected on all weekdays in 2017 and January 2018 are used in these two case studies. Public holidays and days with adverse weather and incidents are excluded. Hence, data of 299 days in 2017 are employed for training. Data from January 8th to 12th in 2018 are adopted for testing and evaluation for the rest of the experiments unless other specifications. The rolling step chosen is 2 min, and the confidence level for the dynamic time window is 90%.

C. Results

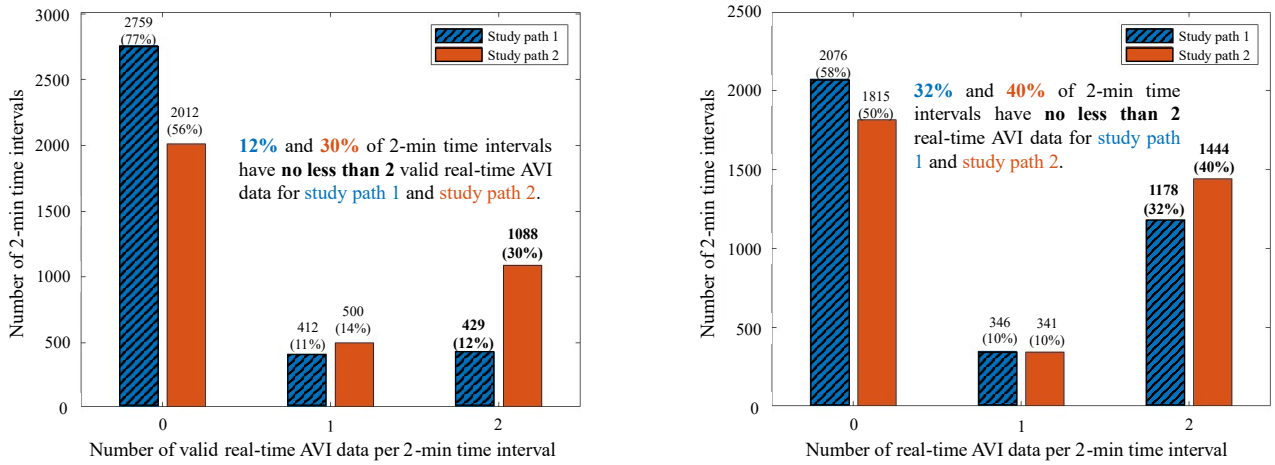
The performance of the proposed unsupervised filtering algorithm is compared with that of the three corresponding existing algorithms that are commonly used in practice for filtering real-time AVI data. The algorithm developed by [34] is used to generate preliminary travel time estimates for historical AVI data. The proposed unsupervised algorithm is denoted as U1, and the other three corresponding existing algorithms are the most up-to-date algorithms for filtering real-time AVI data, which have been used successfully for decades in various ITS projects [31], [33], [34]. The algorithm of [33] is denoted as U2; that of [31] as U3; and that of [34] as U4. These existing algorithms use real-time AVI data while U1 utilizes both real-time and historical AVI data.

The mean absolute error (MAE) and the mean absolute percentage error (MAPE), which are given by (29) and (30), respectively, are used to evaluate the filtering performance of algorithms with respect to the JTIS ground truth.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|X_t - \hat{X}_t\right| \tag{29}$$

$$MAPE = \frac{100}{T}\sum_{t=1}^{T}\frac{|X_t - \hat{X}_t|}{X_t} \tag{30}$$

where X_t are true values and $\hat{X}_t$ are estimated values.



(a) (b)

Fig. 5. Sampling rates of (a) valid real-time AVI data, and (b) real-time AVI data on both study paths.

TABLE II
SUMMARY OF TWO STUDY PATHS

	Study path 1	Study path 2
Number of AVI sensors along the study path	Two	Three
Road type	Urban arterials with bus stops and signalized junction	Urban arterials with bus stops only
Path length (km)	9.2	8.8
Number of bus stops	20	8
Number of entries along the study path (e.g., slip road and frontage access)	13	13
Number of exits along the study path (e.g., slip road and frontage access)	13	11
Free-flow travel time (min)	8.4	7.9
Speed limits (km/h)	70 (31%), 50 (18%), 60 (19%), 80 (32%)	70 (58%), 50 (20%), 80 (22%)
Number of point sensors	Seven	Five

Fig. 6 illustrates the contribution of using historical AVI data for filtering out invalid real-time AVI data. For limited real-time AVI data with low sampling rates, there is a chance that the transition between congestion and free-flow conditions can hardly be recognized properly by the existing filtering algorithms (e.g., U2, Dion's algorithm, which has already considered the transition recognition of the real-time traffic conditions by looking back real-time AVI data in consecutive preceding time intervals). As shown in Fig. 6, the black circles indicate that U2 fails to select valid real-time AVI data. In contrast, U1 with the use of historical AVI data performs well in filtering the limited real-time AVI data. The temporal var-cov relationships between path travel times on different days modeled in (4) can help to recognize traffic conditions by time of day. It is also observed that most of the relevant ground truth is captured within the dynamic time windows resulting from U1 throughout the day.

Table III compares the filtering performance of the proposed unsupervised algorithm with benchmarks with respect to the mean/standard deviation of estimated path travel times. U1 outperforms the other three existing unsupervised algorithms from both aspects. For the mean of path travel times, the MAPE of U1 is 19.3% for study path 1 and 16.1% for study path 2. For the standard deviation of path travel times, the MAE values of U1 are 0.61 min and 0.52 min for study paths 1 and 2, respectively. The comparison of results between U1 and the other three existing unsupervised algorithms provides evidence in supporting the contribution of making use of historical AVI data.

D. Sensitivity Analysis

In the real world, historical ground truth data on path travel times can be available (e.g., existing path travel time estimates from existing ATISs, and samples collected independently from floating car surveys). It is a special case of research problem in this paper when historical ground truth is ready for training purposes. Under this scenario, $x'_{i,d}$ denotes the i^{th} ground truth on path travel time, with the corresponding timestamp denoted as $t'_{i,d}$ in Step 1 of Fig. 2.

With the use of historical ground truth for training purposes, S1 represents the proposed unsupervised algorithm under this scenario. Three existing advanced supervised learning algorithms are selected for benchmark comparison. The long short-term memory neural network LSTM NN in [51] is denoted as S2. The LSTM encoder-decoder model in [52] is denoted as S3. The attention-based periodic-temporal neural network in [53] is denoted as S4. It should be noted historical ground truth in 2017 (299 weekdays) is used for training.

In contrast to other neural networks that have black-box procedures and non-explanatory performance, the FPCA model provides explicit descriptions of the temporal var-cov relationships between path travel times at different time intervals and on different days [43]. Moreover, as the FPCA model enables a better understanding of trends [54], it can be used to quantify the uncertainty of valid real-time AVI data with low sampling rates, particularly those data that are

scattered and time-varying. The input data used in the sensitivity analysis include historical AVI data, historical JTIS ground truth, and real-time AVI data.

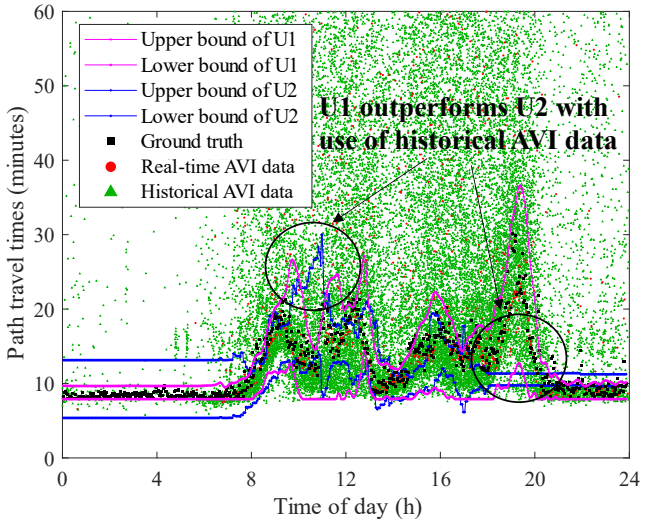


Fig. 6. Filtering performance of U1 and U2 on study path 2 by time of day.

TABLE III
COMPARISON OF FILTERING PERFORMANCE WITH RESPECT TO THE MEAN/STANDARD DEVIATION OF PATH TRAVEL TIMES

Algorithms	Study path 1		Study path 2	
	MAPE (%)	MAE (min)	MAPE (%)	MAE (min)
U1 (proposed unsupervised algorithm)	19.3/15.6	2.94/0.61	16.1/13.2	2.53/0.52
U2 (Dion's algorithm)	20.2/19.3	3.14/0.67	18.2/16.9	2.74/0.56
U3 (Median-based filter)	21.4/22.4	3.32/0.81	19.4/18.2	2.89/0.63
U4 (TransGuide)	22.1/27.1	3.49/0.94	20.1/20.4	3.01/0.68

Table IV gives the comparison results on the mean of estimated path travel times under this scenario. The MAPE of S1 is 11.4% for study path 1 and 5.1% for study path 2. It is found that S1 performs better than the other benchmarks. Besides, it should be noted that other benchmarks can only provide the mean of estimated travel times. But S1 can also produce the corresponding results and the standard deviation of estimated path travel times. In the case study, the corresponding MAPE and MAE are 12.7% and 0.5 min for study path 1, and 9.5% and 0.36 min for study path 2. These results can demonstrate the contribution of using the proposed FPCA model to capture the temporal var-cov relationships between path travel times at different time intervals and on different days for data filtering and path travel time estimation.

The average computational times of the proposed unsupervised and benchmarks without (U1-U4) and with (S1-S4) ground truth are provided. All experiments are conducted on a standard computer with an AMD Ryzen 5 5600X processor (3.7 GHz, 6 cores) in Table V. The average computational time required to obtain dynamic time windows and path travel time estimates varies from 0.07 to 0.63 min. It is found that the U1 is applicable for real-time ITS applications, that is, U1 can filter the real-time AVI data collected at about each 1.5-min time interval, and then rapidly (within 0.55 min) generate the real-time path travel time estimates.

TABLE IV
COMPARISON OF FILTERING PERFORMANCE ON THE MEAN OF PATH TRAVEL TIMES WHEN HISTORICAL GROUND TRUTH IS USED FOR TRAINING

Algorithms	Study path 1		Study path 2	
	MAPE (%)	MAE (min)	MAPE (%)	MAE (min)
S1 (proposed unsupervised algorithm with the use of historical ground truth)	11.4	1.79	5.1	0.81
S2 (LSTM NN)	15.3	2.38	7.5	0.95
S3 (encoder-decoder model)	14.1	2.21	6.6	0.88
S4 (periodic-temporal NN)	12.5	1.96	6.4	0.87

TABLE V
THE AVERAGE COMPUTATIONAL TIME OF THE ALGORITHMS

Algorithms		The average computational time required to obtain travel time estimates and filtering windows for each time interval (min)
Study path 1	U1/S1	0.52/0.55
	U2/S2	0.13/0.63
	U3/S3	0.07/0.5
	U4/S4	0.1/0.52
Study path 2	U1/S1	0.51/0.52
	U2/S2	0.2/0.57
	U3/S3	0.17/0.6
	U4/S4	0.15/0.55

Another sensitivity analysis is conducted to examine the effect of sampling rates of real-time AVI data on the performance of the proposed unsupervised algorithm. As study path 2 has more AVI sensors than study path 1, the range of sampling rate for study path 1 is greater. Moreover, it can be used to investigate the AVI data filtering problem under different sensor failure scenarios for further study. Therefore, study path 2 is used for the sensitivity analysis to examine the effects of sampling rates.

In Fig. 7, U1 performs much better than the other 3 benchmarks, in terms of the probabilities of absolute percentage errors of the estimates less than 20% (i.e., 83% against 56%-60%). However, S1 only performs slightly better than the other benchmarks (i.e., 93% against 88%-91%). It demonstrates that, with the use of historical AVI data only, the proposed unsupervised algorithm U1 can lead to promising results even when ground truth is not available for training purposes.

To further investigate the effects of sampling rates of valid real-time AVI data on the proposed unsupervised algorithm, another sensitivity test is carried out. 30 out of 299 weekdays in 2017 are randomly segregated from the original training set and used as the new validation set. The performance of the proposed unsupervised algorithm on both study paths is provided in Fig. 8. It is noted that when the sampling rate of valid real-time AVI data is no less than 2 valid AVI data per 2-min interval, the performance of U1 is similar on different datasets (95% of the absolute percentage errors less than 15.2% and 14.9% for study paths 1 and 2, respectively). It demonstrates the generality and robustness of the proposed unsupervised algorithm. In general, it is found in Fig. 8 that 95% of the absolute percentage errors of the estimates are less than 20%.

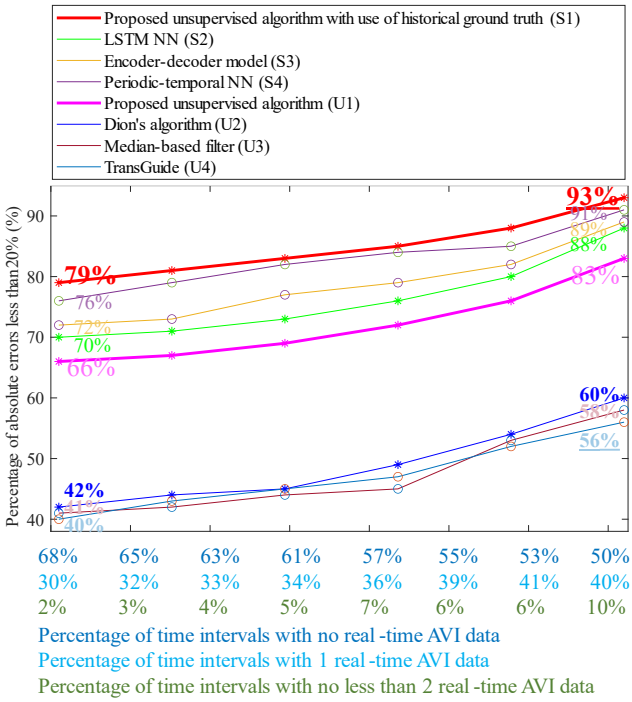


Fig. 7. Sensitivity test with various sampling rates of real-time AVI data on study path 2.

Additionally, as shown in Fig. 5(a), study path 2 with more AVI sensors would have a higher percentage of 2-min intervals with no less than 2 valid data than that of study path 1. Hence, the performance of both S1 and U1 for study path 2 is better than that of study path 1 as shown in Tables III and IV. The same finding can also be found in Fig. 8, even the validation dataset is different in Fig. 8 and Tables III and IV.

As the case study is performed using accurate but limited real-time AVI data, it is worthwhile to discuss the performance of U1 on inaccurate real-time AVI data with more samples (e.g., Bluetooth data). It is assumed that this type of AVI data has a much lower percentage of valid real-time AVI data. Therefore, the performance will deteriorate due to the extremely low sampling rate of valid real-time AVI data for U1. Further study should be carried out in the future if this type of AVI data is available.

To test the effect of historical ground truth data, a sensitivity test is performed by reducing the number of days with historical ground truth data. A percentage varying from 0% to 90% of historical ground truth data is removed to test the performance of S1. The result is given in Table VI. The percentage of absolute percentage errors less than 20% is reduced to 83% or lower if less than 50% of the historical ground truth is used for training purposes. It implies that U1 is better than S1 in practice particularly when less than half of the historical ground truth on path travel time is available for filtering of real-time AVI data and real-time path travel time estimation.

TABLE VI
SENSITIVITY TEST WITH DIFFERENT PERCENTAGES OF HISTORICAL GROUND TRUTH REMOVED IN 2017 DATA ON STUDY PATH 2

Percentage of historical ground truth removed in 2017 data (%)	Percentage of absolute percentage errors less than 20% (%)
0	93
10	90
30	86
50	83
70	78
90	76

With reference to the above (8), the threshold H^* of the covariance of path travel times on different days for the sampling of historical AVI data in S1 is examined in a sensitivity test, The relevant results are given in Table VII, in which the optimum thresholds of H^* for study paths 1 and 2 are 10.8 and 9.6 min^2, respectively. It also shows that the variation of H^* affects the results significantly. For study path 1, only 74% of absolute percentage errors are less than 20% when there is a 20% deviation from the optimum threshold. H^* can also be an annual average figure, as it is based on weekday data in 2017 (excluding public holidays, and days with adverse weather and incidents) to capture the seasonal variation of path travel times. Moreover, the optimum threshold is based on the current dataset, but H^* may deviate from the actual optimum threshold, as the latter will depend on the updated dataset.

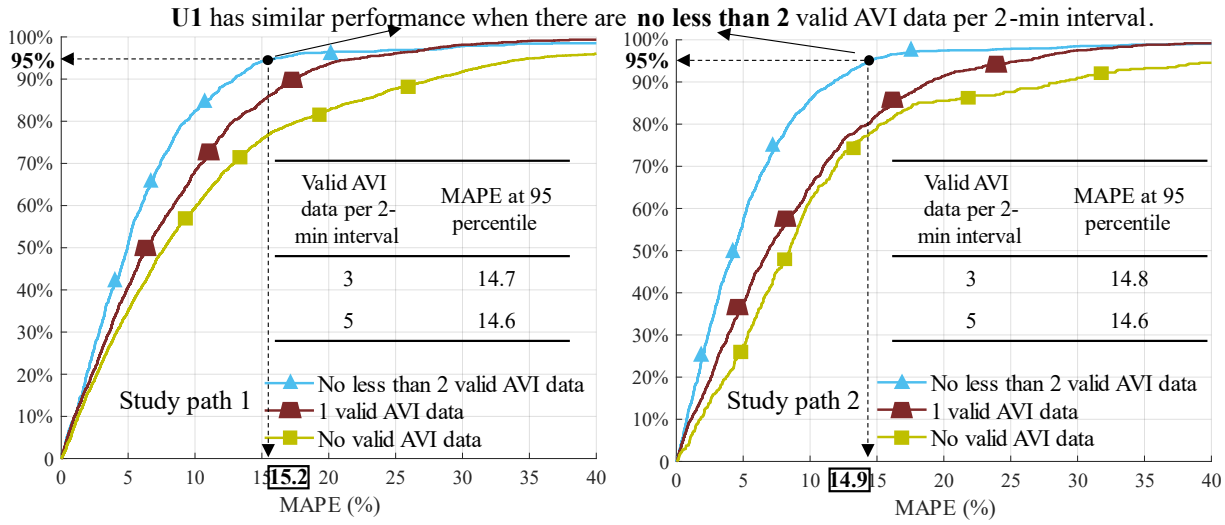


Fig. 8. Sensitivity test of sampling rates of valid real-time AVI data on U1 for study path 1 (left) and study path 2 (right).

TABLE VII
SENSITIVITY TEST OF THE PERCENTAGE DEVIATION OF THE RESULTS OF S1 FROM THE OPTIMUM THRESHOLD

		Percentage of deviation from optimum threshold (%)		
		0	10	20
Value of threshold (min^2)	Study path 1	10.8	11.9	13
	Study path 2	9.6	10.6	11.5
MAPE (%)	Study path 1	11.4	12.6	13.4
	Study path 2	5.1	7.8	9.5
Percentage of absolute percentage errors less than 20% (%)	Study path 1	83	78	74
	Study path 2	93	86	82

IV. CONCLUSION

This paper proposes a novel unsupervised algorithm (U1) for filtering limited but accurate real-time AVI data without ground truth for training. Instead, it makes use of real-time AVI data collected on the same day and historical AVI data collected on previous days. The temporal variance-covariance relationships between path travel times at different time intervals and on different days are explicitly considered in the proposed FPCA model. Both mean and standard deviation of the path travel times are estimated and used for the improvement of the real-time AVI data filtering performance. As the proposed FPCA model can effectively reduce the dimension of high-variability data, the corresponding PACE approach is used to construct the dynamic time windows in the proposed unsupervised algorithm. The real-time dynamic time windows are generated via a rolling horizon scheme. Furthermore, the asymptotic properties of the proposed unsupervised algorithm have been theoretically proven to demonstrate their ability to generate reliable dynamic time windows for real-time AVI data filtering and real-time path travel time estimation.

The performance of U1 is compared respectively with three existing data filtering algorithms in the case studies using real-world data collected from two selected paths in the Hong Kong urban road network. The comparison between the data filtering performance of U1 and U2 by time of day demonstrates the merit of using historical AVI data. It is also found that U1 outperforms the existing algorithms in terms of both mean and standard deviation of estimated path travel times.

A sensitivity analysis is conducted for a special case when ground truth is available for training. The proposed unsupervised algorithm with ground truth for training (namely S1) outperforms other benchmarks in terms of both mean and standard deviation of estimated path travel times. It illustrates the merit of using the proposed FPCA model for modeling temporal variance-covariance relationships between path travel times at different time intervals and on different days. Another sensitivity test is also performed to demonstrate the merits of using historical AVI data when real-time AVI data is sampled at a very low rate. U1 performs much better than the other 3 benchmarks, in terms of the probabilities of absolute percentage errors of the estimates less than 20% (i.e., 83% against 56%-60%).

The sensitivity test on sampling rates of valid real-time AVI data demonstrates the generality and robustness of the proposed unsupervised algorithm. When there are no less than 2 valid real-time AVI data per 2-min interval, 95% probability of generating absolute percentage errors of less than 20%. Moreover, the performance of U1 on different study paths is similar under this scenario, which implies that the proposed unsupervised algorithm is generalized with robust performance. The expected worsened performance on inaccurate real-time AVI data with more samples (e.g., Bluetooth) is also discussed with the assumed lower sampling rate of valid real-time AVI data. Filtering of this type of AVI data is suggested for further study if this dataset is available.

Moreover, an additional sensitivity test is carried out to show the advantage of U1. The percentage of absolute percentage errors less than 20% is reduced to 83% or lower if less than 50% of the historical ground truth is used for training purposes. It implies that U1 is better than S1 in reality when less than half of the historical ground truth on path travel time is available for filtering of real-time AVI data and real-time path travel time estimation.

In the future, other types of traffic-related data, such as weather, traffic accidents, vehicular flow data, bus frequencies, signal timing, and road types could be incorporated into the proposed unsupervised algorithm for improving the data filtering performance. As both AVI sensors and point sensors are deployed in the JTIS, it is interesting to explore the sensor-location problems and trade-offs of these two types of traffic sensors. Similarly, it is also planned to extend the proposed unsupervised algorithm to examine the effects of sensor failure on data from multiple AVI sensors at urban road corridors, by considering network topology and measurement errors.

REFERENCES

[1] A. H. F. Chow, A. Santacreu, I. Tsapakis, G. Tanasaranond, and T. Cheng, “Empirical assessment of urban traffic congestion,” *J. Adv. Transp.*, vol. 48, no. 8, pp. 1000–1016, Dec. 2014, doi: 10.1002/atr.1241.

[2] F. Soriguera and M. Martinez-DIaz, “Freeway Travel Time Information from Input- Output Vehicle Counts: A Drift Correction Method Based on AVI Data,” *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5749–5761, 2021, doi: 10.1109/TITS.2020.2992300.

[3] M. M. Ahmed and M. A. Abdel-Aty, “The viability of using automatic vehicle identification data for real-time crash prediction,” *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 459–468, 2012, doi: 10.1109/TITS.2011.2171052.

[4] X. Zhou and H. S. Mahmassani, “Dynamic origin-destination demand estimation using automatic vehicle identification data,” *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 105–114, 2006, doi: 10.1109/TITS.2006.869629.

[5] Y. Zhu, Z. He, and W. Sun, “Network-Wide Link Travel Time Inference Using Trip-Based Data from Automatic Vehicle Identification Detectors,” *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 6, pp. 2485–2495, 2020, doi: 10.1109/TITS.2019.2919595.

[6] D. Xia, L. Zheng, Y. Tang, X. Cai, L. Chen, W. Liu, and D. Sun, "Link-Based Traffic Estimation and Simulation for Road Networks Using Electronic Registration Identification Data," *IEEE Trans. Veh. Technol.*, vol. 71, no. 8, pp. 8075–8088, Aug. 2022, doi: 10.1109/TVT.2022.3171835.

[7] U. Mori, A. Mendiburu, M. Álvarez, and J. A. Lozano, "A review of travel time estimation and forecasting for Advanced Traveller Information Systems," *Transportmetrica A: Transp. Sci.*, vol. 11, no. 2, pp. 119–157, 2015, doi: 10.1080/23249935.2014.932469.

[8] R. Gao, F. Sun, W. Xing, D. Tao, J. Fang, and H. Chai, "CTTE: Customized Travel Time Estimation via Mobile Crowdsensing," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19335–19347, Oct. 2022, doi: 10.1109/TITS.2022.3160468.

[9] P. Wang, Z. Huang, J. Lai, Z. Zheng, Y. Liu, and T. Lin, "Traffic Speed Estimation Based on Multi-Source GPS Data and Mixture Model," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10708–10720, 2021, doi: 10.1109/TITS.2021.3095408.

[10] D. Correa and K. Ozbay, "Urban path travel time estimation using GPS trajectories from high-sampling-rate ridesourcing services," *J. Intell. Transp. Syst.*, pp. 1–16, Sep. 2022, doi: 10.1080/15472450.2022.2124867.

[11] C. Wang, F. Zhao, H. Zhang, H. Luo, Y. Qin, and Y. Fang, "Fine-Grained Trajectory-Based Travel Time Estimation for Multi-City Scenarios Based on Deep Meta-Learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15716–15728, Sep. 2022, doi: 10.1109/TITS.2022.3145382.

[12] Y. Zhu, Y. Ye, Y. Liu, and J. J. Q. Yu, "Cross-Area Travel Time Uncertainty Estimation From Trajectory Data: A Federated Learning Approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24966–24978, Dec. 2022, doi: 10.1109/TITS.2022.3203457.

[13] W. Zhou, X. Xiao, Y. Gong, J. Chen, J. Fang, N. Tan, N. Ma, Q. Li, C. Hua, S. Jeon, and J. Zhang, "Travel Time Distribution Estimation by Learning Representations Over Temporal Attributed Graphs," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 1–13, 2023, doi: 10.1109/tits.2023.3247884.

[14] Y. Ye, Y. Zhu, C. Markos, and J. J. Q. Yu, "CatETA: A Categorical Approximate Approach for Estimating Time of Arrival," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24389–24400, Dec. 2022, doi: 10.1109/TITS.2022.3207894.

[15] S. Robinson and J. Polak, "Overtaking rule method for the cleaning of matched license-plate data," *J. Transp. Eng.*, vol. 132, no. 8, pp. 609–617, 2006, doi: 10.1061/(ASCE)0733-947X(2006)132:8(609).

[16] X. Zhan, R. Li, and S. V. Ukkusuri, "Link-based traffic state estimation and prediction for arterial networks using license-plate recognition data," *Transp. Res. Part C: Emerg. Technol.*, vol. 117, p. 102660, 2020, doi: 10.1016/j.trc.2020.102660.

[17] M. Chen, G. Yu, P. Chen, and Y. Wang, "A copula-based approach for estimating the travel time reliability of urban arterial," *Transp. Res. Part C: Emerg. Technol.*, vol. 82, pp. 1–23, 2017, doi: 10.1016/j.trc.2017.06.007.

[18] K. Kwong, R. Kavaler, R. Rajagopal, and P. Varaiya, "Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors," *Transp. Res. Part C: Emerg. Technol.*, vol. 17, no. 6, pp. 586–606, 2009, doi: 10.1016/j.trc.2009.04.003.

[19] L. Li, X. Su, Y. Wang, Y. Lin, Z. Li, and Y. Li, "Robust causal dependence mining in big data network and its application to traffic flow predictions," *Transp. Res. Part C: Emerg. Technol.*, vol. 58, pp. 292–307, 2015, doi: 10.1016/j.trc.2015.03.003.

[20] P. Chakraborty, C. Hegde, and A. Sharma, "Data-driven parallelizable traffic incident detection using spatio-temporally denoised robust thresholds," *Transp. Res. Part C: Emerg. Technol.*, vol. 105, pp. 81–99, 2019, doi: 10.1016/j.trc.2019.05.034.

[21] S. Chen, W. Wang, and H. Van Zuylen, "A comparison of outlier detection algorithms for ITS data," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 1169–1178, 2010, doi: 10.1016/j.eswa.2009.06.008.

[22] Y. Shang, X. Li, B. Jia, Z. Yang, and Z. Liu, "Freeway Traffic State Estimation Method Based on Multisource Data," *J. Transp. Eng. A: Syst.*, vol. 148, no. 4, pp. 1–14, Apr. 2022, doi: 10.1061/JTEPBS.0000657.

[23] S. Washington, M. Karlaftis, F. Mannering, and P. Anastasopoulos, *Statistical and Econometric Methods for Transportation Data Analysis*., FL, USA: Chapman and Hall/CRC, Boca Raton, 2020.

[24] W. Qin, X. Ji, and F. Liang, "Estimation of urban arterial travel time distribution considering link correlations," *Transportmetrica A: Transp. Sci.*, vol. 16, no. 3, pp. 1429–1458, 2020, doi: 10.1080/23249935.2020.1751341.

[25] M. Yun, W. Qin, X. Yang, and F. Liang, "Estimation of urban route travel time distribution using Markov chains and pair-copula construction," *Transportmetrica B: Transp. Dyn.*, vol. 7, no. 1, pp. 1521–1552, 2019, doi: 10.1080/21680566.2019.1637798.

[26] E. Kazagli and H. Koutsopoulos, "Estimation of arterial travel time from automatic number plate recognition data," *Transp. Res. Rec.*, no. 2391, pp. 22–31, 2013, doi: 10.3141/2391-03.

[27] P. Duan, G. Mao, J. Kang, and B. Huang, "Estimation of Link Travel Time Distribution with Limited Traffic Detectors," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3730–3743, 2020, doi: 10.1109/TITS.2019.2932053.

[28] J. J. V. Diaz, A. B. Rodriguez Gonzalez, and M. R. Wilby, "Bluetooth Traffic Monitoring Systems for Travel Time Estimation on Freeways," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 123–132, Jan. 2016, doi: 10.1109/TITS.2015.2459013.

[29] H. Park and Y. Kim, "Model for Filtering the Outliers in DSRC Travel Time Data on Interrupted Traffic Flow Sections," *KSCE J. Civ. Eng.*, vol. 22, no. 9, pp. 3607–3619, 2018, doi: 10.1007/s12205-017-1333-z.

[30] A. Haghani, M. Hamedi, K. F. Sadabadi, S. Young, and P. Tarnoff, "Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors," *Transp. Res.*

Rec., vol. 2160, no. 1, pp. 60–68, Jan. 2010, doi: 10.3141/2160-07.

[31] X. Ma and H. Koutsopoulos, "Estimation of the automatic vehicle identification based spatial travel time information collected in Stockholm," *IET Intell. Transp. Syst.*, vol. 4, no. 4, pp. 298–306, 2010, doi: 10.1049/iet-its.2009.0149.

[32] M. L. Tam and W. H. K. Lam, "Using automatic vehicle identification data for travel time estimation in Hong Kong," *Transportmetrica*, vol. 4, no. 3, pp. 179–194, Jan. 2008, doi: 10.1080/18128600808685688.

[33] F. Dion and H. Rakha, "Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates," *Transp. Res. B: Methodol.*, vol. 40, no. 9, pp. 745–766, 2006, doi: 10.1016/j.trb.2005.10.002.

[34] "Automated Vehicle Identification Model Deployment Initiative System Design Document (A report prepared for TransGuide, Southwest Research Institute)", TxDOT, San Antonio, TX, 1998.

[35] H. TranStar, "TranStar Description," http://traffic.houstontranstar.org (accessed May. 19, 2023).

[36] K. C. Mouskos, E. Niver, and L. J. Pignataro, "Transmit System Evaluation," *Database*, pp. 1–170, 1998, [Online]. Available: http://ntl.bts.gov/lib/16000/16700/16703/PB20001045 37.pdf.

[37] J. M. Chiou, "Dynamical functional prediction and classification, with application to traffic flow prediction," *Ann. Appl. Stat.*, vol. 6, no. 4, pp. 1588–1614, 2012, doi: 10.1214/12-AOAS595.

[38] I. G. Guardiola, T. Leon, and F. Mallor, "A functional approach to monitor and recognize patterns of daily traffic profiles," *Transp. Res. B: Methodol.*, vol. 65, pp. 119–136, 2014, doi: 10.1016/j.trb.2014.04.006.

[39] I. M. Wagner-Muns, I. G. Guardiola, V. A. Samaranayke, and W. I. Kayani, "A Functional Data Analysis Approach to Traffic Volume Forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 878–888, Mar. 2018, doi: 10.1109/TITS.2017.2706143.

[40] J. M. Chiou, H. T. Liou, and W. H. Chen, "Modeling Time-Varying Variability and Reliability of Freeway Travel Time Using Functional Principal Component Analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 257–266, 2021, doi: 10.1109/TITS.2019.2956090.

[41] K. Chen and H. G. Müller, "Modeling conditional distributions for functional responses, with application to traffic monitoring via GPS-enabled mobile phones," *Technometrics*, vol. 56, no. 3, pp. 347–358, 2014, doi: 10.1080/00401706.2013.842933.

[42] R. X. Zhong, X. X. Xie, J. C. Luo, T. L. Pan, W. H. K. Lam, and A. Sumalee, "Modeling double time-scale travel time processes with application to assessing the resilience of transportation systems," *Transp. Res. B: Methodol.*, vol. 132, pp. 228–248, Feb. 2020, doi: 10.1016/j.trb.2019.05.005.

[43] R. X. Zhong, J. C. Luo, H. X. Cai, A. Sumalee, F. F. Yuan, and A. H. F. Chow, "Forecasting journey time distribution with consideration to abnormal traffic conditions," *Transp. Res. Part C: Emerg. Technol.*, vol. 85, pp. 292–311, 2017, doi: 10.1016/j.trc.2017.08.021.

[44] H. B. Celikoglu, "Flow-based freeway travel-time estimation: A comparative evaluation within dynamic path loading," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 772–781, 2013, doi: 10.1109/TITS.2012.2234455.

[45] F. Yao, H.-G. Müller, and J.-L. Wang, "Functional Data Analysis for Sparse Longitudinal Data," *J. Am. Stat. Assoc.*, vol. 100, no. 470, pp. 577–590, Jun. 2005, doi: 10.1198/016214504000001745.

[46] H. G. Müller and F. Yao, "Functional additive models," *J. Am. Stat. Assoc.*, vol. 103, no. 484, pp. 1534–1544, 2008, doi: 10.1198/016214508000000751.

[47] H. Ji and H. G. Müller, "Optimal designs for longitudinal and functional data," *J. R. Stat. Soc. Series B: Stat. Methodol.*, vol. 79, no. 3, pp. 859–876, 2017, doi: 10.1111/rssb.12192.

[48] T. L. Pan, A. Sumalee, R. X. Zhong, and N. Indra-Payoong, "Short-term traffic state prediction based on temporal-spatial correlation," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1242–1254, 2013, doi: 10.1109/TITS.2013.2258916.

[49] M. L. Tam and W. H. K. Lam, "Application of automatic vehicle identification technology for real-time journey time estimation," *Inf. Fusion*, vol. 12, no. 1, pp. 11–19, 2011, doi: 10.1016/j.inffus.2010.01.002.

[50] Hong Kong Transport Department, "Journey Time Indication System," https://www.td.gov.hk/en/transport_in_hong_kong/its/its_achievements/journey_time_indication_system_/index.html (accessed May. 19, 2023).

[51] Q. Ouyang, Y. Lv, J. Ma, and J. Li, "An LSTM-based method considering history and real-time data for passenger flow prediction," *Appl. Sci.*, vol. 10, no. 11, 2020, doi: 10.3390/app10113788.

[52] Z. Wang, X. Su, and Z. Ding, "Long-Term Traffic Prediction Based on LSTM Encoder-Decoder Architecture," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6561–6571, Oct. 2021, doi: 10.1109/TITS.2020.2995546.

[53] X. Shi, H. Qi, Y. Shen, G. Wu, and B. Yin, "A Spatial–Temporal Attention Approach for Traffic Prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 4909–4918, Aug. 2021, doi: 10.1109/TITS.2020.2983651.

[54] P. Z. Hadjipantelis and H.-G. Müller, "Functional Data Analysis for Big Data: A Case Study on California Temperature Trends," in *Handbook of Big Data Analytics*, New York, NY, USA: Springer, 2018, pp. 457–483.

Ang Li received B.Eng. degree in civil engineering from The Hong Kong Polytechnic University, Hong Kong and the M.Sc. degree in transportation engineering from The University of Hong Kong, Hong Kong. He is currently pursuing the Ph.D. degree in the Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University. His research interests include travel time estimation and prediction, and intelligent transportation systems.

William H. K. Lam received B.S. and M.S. degrees from the University of Calgary, Canada, and a Ph.D. degree in transportation engineering from the University of Newcastle upon Tyne, Newcastle, U.K. He is currently an Emeritus Professor of Civil and Transportation Engineering with the Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong. He has also been an Honorary Professor at the Institute for Transport and Logistics Studies, The University of Sydney, Australia since 2015.

Prof. Lam is currently a member of the International Scientific Committee of the International Symposium on Transportation Network Resilience (INSTR) and has been the convenor of the International Advisory Committee of the *International Symposium on Transportation and Traffic Theory (ISTTT)* from 2015-2022. He is also the Founding Editor-in-Chief of *Transportmetrica* and is now one of the Co-Editors-in-Chief of *Transportmetrica A: Transport Science*. His current research interests include transport planning and traffic forecasting, ITS technology and development, smart surveillance and traffic simulation, public transport, and pedestrian studies.

Wei Ma (Member, IEEE) received bachelor's degrees in Civil Engineering and Mathematics from Tsinghua University, China, master degrees in Machine Learning and Civil and Environmental Engineering, and PhD degree in Civil and Environmental Engineering from Carnegie Mellon University, USA. He is currently an assistant professor with the Department of Civil and Environmental Engineering at the Hong Kong Polytechnic University (PolyU). His research focuses on intersection of machine learning, data mining, and transportation network modeling, with applications for smart and sustainable mobility systems.

Andy H. F. Chow (Member, IEEE) is currently an Associate Professor in Systems Engineering at the City University of Hong Kong. His research lies in developing tools for analyzing and managing transport systems. His doctoral dissertation on optimal control of dynamic transport networks completed in London received a Gordon Newell Memorial Dissertation Prize in 2008. Dr. Chow is a Board Member of the Hong Kong Society for Transportation Studies (HKSTS), Chartered Member of the Chartered Institute of Logistics and Transport (CILTHK), Member of the Institute of Electrical and Electronics Engineers (IEEE).

S. C. Wong received the B.Sc. (Eng.) and M.Phil. degrees from The University of Hong Kong (HKU), and the Ph.D. degree in transport studies from University College London (UCL), U.K. He is currently a Chair Professor with the Department of Civil Engineering, HKU. His research interests include optimization of traffic signal settings, continuum modeling for traffic equilibrium problems, land use and transportation problems, dynamic highway and transit assignment problems, urban taxi services, and road safety.

Mei Lam Tam received her B.Sc. and Ph.D. degrees from The Hong Kong Polytechnic University, Hong Kong. She is currently a Senior Research Fellow with the Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong. Her research interests include transport planning and traffic forecasting, travel time estimation and prediction, and intelligent transportation systems.