Missing Data Estimation for Traffic Volume by Searching Optimum Closed Cut in Urban Networks

Shangbo Wang, *Student Member*, *IEEE*, and Guoqiang Mao, *Fellow, IEEE*

***Abstract*—Traffic data imputation has drawn significant attention from both academia and industry because traffic data often suffer from data missing problems, caused by temporary deployment of sensors, detector malfunction and lossy communication systems. To fully exploit the spatial-temporal correlation and road topological information in urban traffic network, we propose an Optimum Closed Cut (OCC) based spatio-temporal imputation technique, which is implemented in two stages: a) employing graph theory to search the OCC in the road network, for which the traffic on roads intersected by the closed cut has the maximum correlation with that on the target road while minimizing the number of intersected roads; b) estimating the missing data on the target road using OCC based Kriging estimator, incorporating both the road topological information and flow conservation law to improve the estimation accuracy. Experimental results using traffic data collected on real roads indicate that the OCC search algorithm can effectively capture the optimum set of neighboring sensors. OCC based estimator can provide more accurate imputation results compared to NHA (Nearest Historical Average) and correlative *k*-NN (*k*-Nearest Neighbors) methods. The road topological information and flow conservation law can be explored to further improve the estimation performance while reducing the number of sensors involved in the data imputation, hence improving the computational efficiency.**

***Index Terms*—Traffic data imputation, Optimum Closed Cut, NHA, *k*-NN**

I. Introduction

TRAFFIC flow refers to the number of vehicles passing through a certain fixed point within a unit time. Traffic flow information plays a vital role in Intelligent Transportation Systems (ITS). For example, the Advance Traveler Information Systems (ATIS), which acquire, analyze and present information to assist travelers navigating from the source to the destination, and the Advance Traffic Management Systems (ATMS), which integrate various technology to improve the road traffic flow and road safety, both rely heavily on reliable, accurate and consistent traffic flow information to provide users with up-to-date traffic information and guidance [**?**].

Missing data problem, where some subsets of traffic data become missing, has greatly hindered the collection and subsequent analysis, estimation and prediction of traffic flow data. Traffic data may become missing due to temporary deployment of sensors, detector malfunction or lossy communication systems. Specifically, due to high deployment costs, permanent traffic sensors may be installed on a subset of roads only [**?**] and some other roads may only be equipped with temporary sensors, which can provide traffic data within limited time periods. Furthermore, failures, caused by detector malfunction and lossy communication systems, may also result in incomplete traffic data [**?**], [**?**]. It was reported in [**?**] that at hundreds of detection points within PeMS (Performance Measurement System) traffic flow database, more than 5% of data are missing. The missing data has severe impact on many ITS applications, most of which rely on reliable, accurate and complete data [**?**], [**?**], [**?**]. For instance, traffic flow prediction relies on the complete historical data and the prediction performance will reduce sharply with incomplete data [**?**]. Therefore, developing methodologies to precisely estimate the missing data, i.e, traffic data imputation, is an important task.

A number of imputation methods have been proposed in the recent decade. Existing imputation techniques can be generally classified into three categories: interpolation based, prediction based and statistical learning based methods [**?**].

Some well known interpolation based methods include correlative *k*NN (*k* Nearest Neighbor) scheme [**?**], sectional *k*NN scheme [**?**] and LLS (Local Least Squares) scheme [**?**].

Prediction based methods only rely on known traffic prediction methods, e.g., Auto-regressive Integrated Moving Average (ARIMA) [**?**], [**?**], Seasonal ARIMA (SARIMA) [**?**], Space-Time ARIMA (ST-ARIMA) [**?**], [**?**].

The most frequently used statistical learning methods are Probabilistic Principal Component Analysis (PPCA) [**?**], Kernel Probabilistic Principal Component Analysis (KPPCA) [**?**] and tensor completion techniques [**?**], [**?**], [**?**].

Most existing traffic data imputation methods suffer from the following shortcomings: 1) there are few studies trying to find the optimum subset of detectors before imputation. It is well known that choosing all detector measurements may improve the accuracy of imputation but significantly increase the computational complexity, which consequently results in the imputation method becoming non-scalable; 2) the spatial correlation of the traffic data has not been fully utilized; 3) previous work mostly neglects the road topological information, which can be further exploited to improve the accuracy of the traffic data imputation.

Vehicles traveling through a specific road during a certain time interval can be classified into three portions: i) vehicles arriving from some neighboring roads equipped with detectors (termed *measured roads*), ii) vehicles coming from some neighboring roads without detectors (termed *unmeasured roads*), iii) vehicles coming from sources or traveling to

Shangbo Wang is with the School of Computing and Communications, The University of Technology Sydney, Sydney, Australia, e-mail: shangbo.wang@uts.edu.au.

Guoqiang Mao is with the School of Computing and Communications, The University of Technology Sydney, Sydney, Australia, e-mail: guoqiang.mao@uts.edu.au.

sinks within some specific sections (termed *flow generation or dissipation*). The first portion can be read from the detectors while sum of the second and the third one can be estimated from the empirical data. The road topology gives the sufficient information about on-ramp and off-ramp of each measured and unmeasured roads and thus can be exploited to alleviate the missing data problem.

Intuitively, when drawing a close circuit, or a closed cut, on a road map, the total amount of *long-term* traffic entering into the closed circuit must be equal to the total amount of *long-term* outgoing traffic. This implies that traffic on the roads intersected by the closed circuit must be correlated. Indeed, an equality can be established that relate traffic on those intersected roads. Motivated by the intuition and the aforementioned shortcomings of existing imputation techniques, in this paper, we propose an Optimum Closed Cut (OCC) based spatio-temporal imputation technique, where the OCC satisfies the following conditions: a) the close cut intersects the target road; *and* b) traffic on other intersected roads has the maximum correlation with that on the target road; *and* c) the number of intersected roads is minimized. The proposed technique then uses the spatio-temporal correlation of traffic on roads intersected by the OCC to estimate the missing data on the target road. The proposed technique utilizes both the road topological information and the spatio-temporal correlation among road traffic for imputation, while using a minimal number of sensor measurements. Therefore, it strikes a fine balance between imputation accuracy and computational complexity. Specifically, the main contributions of this paper are:

1) an optimum closed cut based spatio-temporal imputation technique is proposed that allows us to explicitly incorporate road topology information into imputation while using a small number of sensor measurements;
2) a graph-based technique is developed to select the optimum closed cut that achieves an optimum trade-off between the number of sensor measurements employed and the set of measured roads whose traffic has maximum correlation with that of the target road;
3) a spatial Kriging estimator is developed to explore the spatio-temporal correlation among road traffic for imputation.
4) experiments are conducted using real traffic data provided by Sydney Roads and Maritime Services (RMS), which validates the developed OCC based spatio-temporal imputation technique and demonstrates that the proposed technique can provide more accurate imputation compared with those in the literature.

This rest of the paper is organized as follows: Section II reviews the related work. Section III formulates the missing data imputation problem. Section IV presents the OCC based spatio-temproal imputation technique. Section V establishes the performance and validity of the proposed imputation strategy and compares its performance to those in the literature. Section VI concludes the study.

Notation: In this paper, bold capital characters stand for matrices, while bold and non-bold lowercase characters stand for vectors and scalars, with $I_N \in \mathbb{C}^{N\times N}$ being an identity matrix. The symbol "$^{\mathrm{T}}$" denotes matrix transpose operation. The symbols $E(\cdot)$ and $var(\cdot)$ represent expectation and variance, respectively, and $f(\cdot,\cdot)$ represents the flow between two vertices. The symbols "$\otimes$" and "$\otimes_V$" denote convolution operation and vector convolution operation, respectively.

II. Related Work

A number of studies have been carried out in exploiting spatial information to improve imputation performance. Tak et al. proposed sectional *k*-NN (Nearest Neighbor) method, which impute missing data based on road sections sharing the same traffic property [**?**]. Cai et al. introduced the correlative *k*-NN model which was superior than the original *k*-NN model because it replaces physical distances by the equivalent distances, which are determined by both the physical distance and the correlation coefficient between the historical traffic data of the two roads [**?**].

In [**?**] and [**?**], the authors explored the ability of tensor based method for multi-loop detector's missing data imputation, which completes the missing data by tensor decomposition. Qu et al. proposed the PPCA (Probabilistic Principal Component Analysis)-based method which integrated MLE (Maximum Likelihood Estimation) into traditional PCA (Principal Component Analysis) approach [**?**]. Li et al. compared PPCA method and KPPCA (Kernel Probabilistic Principal Component Analysis) method, which assumes a nonlinear relationship between observed samples and latent variables [**?**].

The aforementioned review reveals that most existing studies did not consider the problem of finding the optimum set of sensors for imputation. They either collected traffic data from all detectors or consider the detectors satisfying some given (often arbitrarily set) conditions.

Wang and Pagageorgiou utilized the macroscopic traffic flow model and the extended Kalman-filtering (EKF) method to estimate the freeway traffic state [**?**]. The considered freeway is subdivided into *N* segments. Traffic flow at boundary of each segment and some important parameters constitute the state vector. The key differences from our technique are that [**?**] mainly utilized the time evolution and measurements to estimate the state vector whereas our technique utilizes spatial-temporal correlation to estimate the missing data. Ng proposed a strategy which aims at determining the smallest subset of links in a traffic network for counting sensor installation in order to infer flows on all remaining links [**?**]. Ng presented the condition that all link flows can be inferred and proposed the inference method. Viti et al. studied the network sensor location problem (NSLP), which considered the case that the variables are partially observed [**?**]. Castillo et al. dealt with the over-specified network observability problem, which aims at determining link flow based on a subset of observed OD-pair and link flows [**?**]. The key differences between our technique and the three literature are that [**?**], [**?**], [**?**] dealt with network observability problem, which aims at optimizing the sensor location and determining link flow based on a subset of observed OD-pair and link flows, whereas our technique tries to utilize the spatial-temporal correlation between each

crossed link and the target link to estimate the missing data caused by temporary sensor failure based on the empirical measured data.

III. Problem Formulation

In this subsection, we will give a formal definition of the problem being considered in this paper. We consider an urban traffic network with a total of N_l links. Suppose there are N_{m} roads equipped with permanent or temporary detectors while the rest have no detector, Furthermore, each detector measures the data with the same sampling rate and delivers maximal M data points each day, and there are K observed days. Then, the traffic data can be viewed as a tensor $T \in \mathbb{R}^{N_{\mathrm{m}} \times M \times K}$. Denote the set of missing data in T by Q_{miss} and let N_{miss} be cardinality of Q_{miss}. Each element of Q_{miss} can be represented as q_{nmk} (true value of the missing data), where the subscripts n, m and k are the n-th road, m-th data point and k-th day, respectively. The missing data imputation aims at finding a function of the available measured data $f_{nmk}\left(T \backslash Q_{\mathrm{miss}}\right)$ to obtain the most likely estimates of Q_{miss} to minimize MAPE (Mean Absolute Percentage Error) and RMSE (Root Mean Squared Error) of the estimates, defined by

$$\mathrm{MAPE} = \frac{1}{N_{\mathrm{miss}}} \sum_{q_{nmk} \in Q_{\mathrm{miss}}} \frac{\left| f_{nmk}\left(T \backslash Q_{\mathrm{miss}}\right) - q_{nmk} \right|}{q_{nmk}} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{q_{nmk} \in Q_{\mathrm{miss}}} \left| f_{nmk}\left(T \backslash Q_{\mathrm{miss}}\right) - q_{nmk} \right|^2}{N_{\mathrm{miss}}}} \quad (2)$$

IV. OCC based Strategy

In this section, we introduce the OCC based spatio-temporal imputation technique. Prior to explaining the OCC based strategy, we will firstly review the SRE (Spatial Random Effects) model, which has been applied in [**?**] and [**?**]. Then, we will apply the SRE model to the OCC search algorithm. Cressie et al. defined the SRE model as a summation of the large-scale spatial variation, smooth small-scale spatial variation and the measurement error, where the unknown random variables are fixed in number, statistically independent, and coefficients of known but not-necessarily-orthogonal spatial basis function [**?**]. In a traffic network, the measured traffic flow $\boldsymbol{Z}(\boldsymbol{s})$ at a finite number of locations $\boldsymbol{s} = \{ s_1 \quad \cdots \quad s_N \}$ can be expressed by [**?**], [**?**]

$$\boldsymbol{Z}(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{B}(\boldsymbol{s})^{\mathrm{T}} \boldsymbol{\eta} + \boldsymbol{\xi}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}) \quad (3)$$

where the product of $\boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}}$ and $\boldsymbol{\beta}$ can be understood as the weighted sum of the average traffic flow from the L selected neighboring roads, the length of $\boldsymbol{\beta}$ represents the number of selected neighboring roads, $\boldsymbol{B}(\boldsymbol{s})^{\mathrm{T}} \boldsymbol{\eta}$ can be interpreted as the fluctuation caused by the varied traffic flow from the L selected neighboring roads, $\boldsymbol{\xi}(\boldsymbol{s})$ can be understood as a fine-scale variability on $\boldsymbol{s}$ due to the nugget effect and flow generation or dissipation within some specific sections, and $\boldsymbol{\epsilon}(\boldsymbol{s})$ denotes the measurement error. In geostatistics, nugget effect represents the discontinuity at the beginning of semivariogram graphs, which is generally caused by inadequate sampling size [**?**].

A. Traffic Flow Analysis with a Closed Cut

Consider a two-dimensional traffic network with N_l single lane bi-directional roads (Fig.1), which are composed of N_{m} roads with detectors and N_{un} roads without detectors, $N_l = N_{\mathrm{m}} + N_{\mathrm{un}}$.

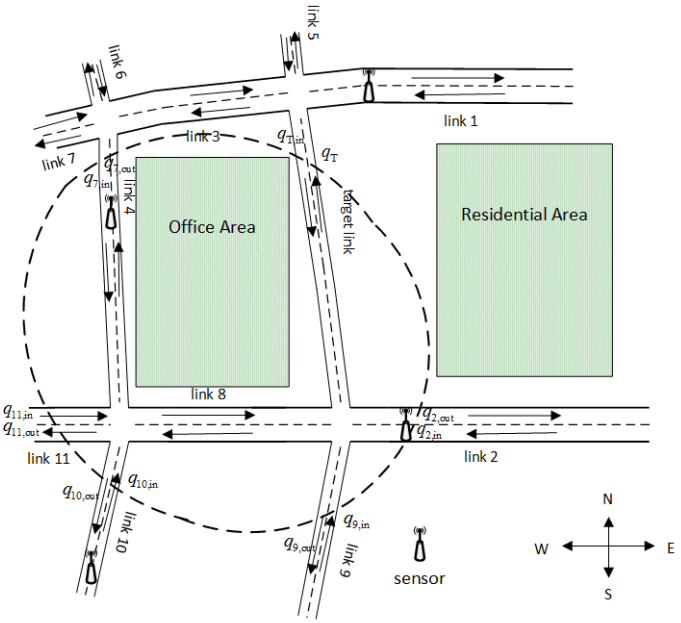

Fig. 1: Illustration of a two-dimensional urban network with sensors

The target road is defined as the road, on which the traffic flow should be estimated. The flow on the target road toward one direction is caused by the flow on its neighboring roads and the flow directly injected to sinks or dissipated from sources within some specific sections. For instance, the flow on the target road toward north consists of portion of eastbound flow from road 8, westbound flow from road 2, northbound flow from road 9 and the missing flow, where the flow from road 2 can be acquired from the detector while the road 8 and road 9 show a lack of data. To investigate the flow relation between the target road and its neighboring roads, we propose Theorem 1, which gives the mathematical expression for the continuous flow relation between two roads.

Theorem 1. *Consider a simple case of two directional roads and define $q_1(\tau)$, $q_2(\tau)$, $\mu_1(\tau)$ and $\mu_2(\tau)$ as traffic flow and traffic density on the two roads, respectively. Further assume that traffic flow on road 2 all comes from road 1, the case that traffic from road 1 may divert to other roads is allowed. Assume that $\mu_1(\tau)$ and $\mu_2(\tau)$ are smaller than μ_m defined by the limiting density [**?**]. Then the relation between $q_1(\tau)$ and $q_2(\tau)$ can be approximated by*

$$\begin{aligned} q_2(\tau) &\approx q_1(\tau) \otimes h_{12,m}(\tau, t) \\ &= \int_0^{\tau} q_1(t)\, h_{12,m}(\tau - t, t)\, dt \end{aligned} \quad (4)$$

where "$\otimes$" is two-dimensional convolution operation and $h_{12,m}(\tau,t)$ is a time-varied correlation coefficient function between two links.

Proof. Partition the input flow stream on road 1 into *N* cells, each of which is with a short distance δx. Thus, the length of input flow stream L_1 can be expressed by $L_1 = N\delta x$ and there are $\mu_1(\tau)\,\delta x$ vehicles traveling through road 1 at τ-th time slot. For an arbitrarily selected cell, its length may expand to $\delta x'$ when the cell arrives at the sensor 2 because of the speed variation and difference of the vehicles within the cell. Here we only consider the case that $\delta x' > \delta x$, because shrink of the partial stream length can be viewed as overlap of two adjacent expanded cells. The density within the short distance δx and $\delta x'$ are assumed to be constant because δx and $\delta x'$ are small enough. Thus, there are $\mu_1(\tau)\,\delta x$ and $\mu_2(\tau)\,\delta x'$ vehicles traveling through the road 1 and 2 at τ-th time slot, respectively. Then, the number of vehicles traveling through road 2 at τ-th time slot can be expressed by

$$\mu_2(\tau)\,\delta x' = \sum_{i=1\cdots N} \mu_1(\tau_i)\,\delta x p_{i,12,\mathrm{m}}(\tau) \tag{5}$$

where τ_i is the time slot in which the *i*-th cell travels through road 1, $p_{i,12,\mathrm{m}}(\tau)$ is a function that represents the fraction of vehicles expanding over time zone for the *i*-th cell arriving at road 2, and then for each cell, it has

$$\begin{cases} p_{i,12,\mathrm{m}}(\tau) = 0 & \text{if the } i\text{-th cell is diverted to other links} \\ \int_0^\infty p_{i,12,\mathrm{m}}(\tau) = 1 & \text{if the } i\text{-th cell arrives at the link 2} \end{cases}$$

When $\delta x, \delta x' \to 0$, (5) can be transformed as

$$\begin{aligned} \mu_2(\tau) &= lim_{\delta x,\delta x'\to 0}\left(\sum_{i=1\cdots N} \mu_1(\tau_i)\,\frac{\delta x}{\delta x'} p_{i,12,\mathrm{m}}(\tau)\right) \\ &= \int_0^\tau \mu_1(t)\,h_{12,\mathrm{m}}(\tau - t, t)\,\mathrm{d}t \end{aligned} \tag{6}$$

where $h_{12,\mathrm{m}}(\tau,t)$ is a time-varied correlation coefficient function between road 1 and 2. If $\mu_1(\tau)$ and $\mu_2(\tau)$ are smaller than the limiting density, the mean speed will be unaffected and flow-density curve is closed to linearity [**?**]. Thus, (4) can be obtained via multiplying both sides of (6) with the mean speed at the input and output streams, respectively. □

Remark 1. It is worth noting that Theorem 1 is also valid for the case that the two roads are not adjacent to each other. In such case, all road segments between the two roads traveled through by the input stream can be virtualized as an intersection with a delay function. Any congestion occurring between roads 1 and 2 has no impact on the validity of Theorem 1. In that case, the time delay caused by congestion occurring between the two roads is given by the time-varying correlation coefficient function between two roads $h_{12,\mathrm{m}}(\tau,t)$. In the case that there is congestion on road 1 or road 2, Theorem 2 can be slightly modified to express a relation in terms of traffic volume between each road, instead of flow:

$$\begin{aligned} \int q_2(\tau)\,\mathrm{d}\tau &= \int\int_0^\tau q_1(t)\,h_{12,\mathrm{m}}(\tau - t, t)\,\mathrm{d}t\mathrm{d}\tau \\ &= \int q_1(t)\int_0^\tau h_{12,\mathrm{m}}(\tau - t, t)\,\mathrm{d}\tau\mathrm{d}t \\ &= g_{12,\mathrm{m}}\int q_1(\tau)\,\mathrm{d}\tau \end{aligned}$$

where $g_{12,\mathrm{m}}$ is the correlation coefficient in terms of traffic volume between two roads, and can be expressed by

$$g_{12,\mathrm{m}} = \frac{\int q_1(t)\int_0^\tau h_{12,\mathrm{m}}(\tau - t, t)\,\mathrm{d}\tau\mathrm{d}t}{\int q_1(t)\,\mathrm{d}t}$$

In this paper, the assumption in Theorem 1 is fulfilled because our employed data shows that no congestion occurs on the inspected roads. Suppose the number of selected feeding sources of the target road is K, which consists of K_{m} measured roads and K_{un} unmeasured roads. From Theorem 1, the flow on the target road can be expressed by

$$\begin{aligned} &q_{\mathrm{T}}(\tau) \\ &\approx \bar{q}_{\mathrm{T}}(\tau) + \boldsymbol{C_{T,i}}(\boldsymbol{\tau},\boldsymbol{t})\,\boldsymbol{C_{i,i}^{-1}}(\boldsymbol{t}) \otimes_V (\boldsymbol{q_i}(\boldsymbol{\tau}) - \bar{\boldsymbol{q}}_{\boldsymbol{i}}(\boldsymbol{\tau})) + w_{\mathrm{T}} \end{aligned} \tag{7}$$

where $\bar{q}_{\mathrm{T}}(\tau)$ is the average flow on the target road at τ-th time slot, w_{T} is the missing flow and can be modeled as a stationary non-zero mean Gaussian variable, $\boldsymbol{q_i}(\boldsymbol{\tau})$ and $\bar{\boldsymbol{q}}_{\boldsymbol{i}}(\tau)$ are $K \times 1$ vectors containing the instantaneous and average flow on the K feeding sources, $\boldsymbol{C_{T,i}}(\boldsymbol{\tau},\boldsymbol{t})$ and $\boldsymbol{C_{i,i}}(\boldsymbol{t})$ are a time-varied $1 \times K$ Cross Correlation Matrix (CCM) between $q_{\mathrm{T}}(\tau)$ and $\boldsymbol{q_i}(\boldsymbol{\tau})$, and a time-varied $K \times K$ Auto Correlation Matrix (ACM) of $\boldsymbol{q_i}(\boldsymbol{\tau})$, respectively. By the definition, $\boldsymbol{C_{T,i}}(\boldsymbol{\tau},\boldsymbol{t})$ and $\boldsymbol{C_{i,i}}(\boldsymbol{t})$ can be expressed by

$$\begin{aligned} \boldsymbol{C_{T,i}}(\boldsymbol{\tau},\boldsymbol{t}) &= E\left((q_{\mathrm{T}}(\tau) - \bar{q}_{\mathrm{T}}(\tau))\,(\boldsymbol{q_i}(\boldsymbol{t}) - \bar{\boldsymbol{q}}_{\boldsymbol{i}}(\boldsymbol{t}))^{\mathrm{T}}\right) \\ \boldsymbol{C_{i,i}}(\boldsymbol{t}) &= E\left((\boldsymbol{q_i}(\boldsymbol{t}) - \bar{\boldsymbol{q}}_{\boldsymbol{i}}(\boldsymbol{t}))\,(\boldsymbol{q_i}(\boldsymbol{t}) - \bar{\boldsymbol{q}}_{\boldsymbol{i}}(\boldsymbol{t}))^{\mathrm{T}}\right) \end{aligned} \tag{8}$$

Note that "$\otimes_V$" is vector convolution operation. For example, given two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ with the elements $a_{i,i=1\cdots N}(t)$ and $b_{i,i=1\cdots N}(t)$, respectively. Then the convolution of two vectors can be expressed by

$$\begin{aligned} \boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} &= \begin{bmatrix} a_1(t) & a_2(t) & \cdots & a_N(t) \end{bmatrix} \otimes_V \begin{bmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_N(t) \end{bmatrix} \\ &= \sum_{i=1\cdots N} a_i(t) \otimes b_i(t) \end{aligned}$$

Remark 2. Note that (7) is valid no matter whether the corresponding road traffic is statistically independent or correlated. Equation (7) gives a general form to express the relation between the target flow and its neighboring flows for both statistically independent and correlated cases. In the case of correlated traffic, $\boldsymbol{C_{i,i}^{-1}}(\boldsymbol{t})$ is a $K \times K$ auto

correlation matrix, $\boldsymbol{C_{T,i}}(\boldsymbol{\tau},\boldsymbol{t})$ is a $1\times K$ vector. It follows that $\boldsymbol{C_{T,i}}(\boldsymbol{\tau},\boldsymbol{t})\,\boldsymbol{C_{i,i}^{-1}}(\boldsymbol{t})$ is a $1\times K$ vector, each element of which can be represented as $h_{\mathrm{T}l}(\tau,t)$. In the case of independent traffic, $\boldsymbol{C_{i,i}}(\boldsymbol{t})$ becomes a diagonal matrix and thus $h_{\mathrm{T}l}(\tau,t)$ becomes the correlation coefficient.

Then (7) can be rewritten as

$$q_{\mathrm{T}}(\tau)\approx\bar{q}_{\mathrm{T}}(\tau)+\sum_{l=1}^{K_{\mathrm{m}}}\tilde{q}_{l,\mathrm{m}}(\tau)\otimes h_{\mathrm{T}l,\mathrm{m}}(\tau,t)+\sum_{j=1}^{K_{\mathrm{un}}}\tilde{q}_{j,\mathrm{un}}(\tau)\otimes h_{\mathrm{T}j,\mathrm{un}}(\tau,t)+w_{\mathrm{T}}-\mu_w \quad (9)$$

where $\tilde{q}_{l,\mathrm{m}}(\tau)$, $\tilde{q}_{j,\mathrm{un}}(\tau)$, $h_{\mathrm{T}l,\mathrm{m}}(\tau,t)$ and $h_{\mathrm{T}j,\mathrm{un}}(\tau,t)$ are the flow variation on the *l*-th measured and *j*-th unmeasured roads, and the time-varied correlation coefficient function between the target and the *l*-th measured and *j*-th unmeasured roads, respectively. The flow variation $\tilde{q}_{l,\mathrm{m}}(\tau)$ and $\tilde{q}_{j,\mathrm{un}}(\tau)$ denote the gap between the instantaneous flow and the average flow on the *l*-th and *j*-th roads, w_{T} is a non-zero mean Gaussian variable and $w_{\mathrm{T}}\sim N(\mu_w,\sigma_w)$. Note that $q_{l,\mathrm{m}}(\tau)$ and $q_{j,\mathrm{un}}(\tau)$ are stochastic process which varies day to day at the same time slot. By the assumption of $q_{l,\mathrm{m}}(\tau)$ and $q_{j,\mathrm{un}}(\tau)$ keeping constant over a short time interval, (9) can be rewritten as

$$q_{\mathrm{T}}(\tau)\approx\bar{q}_{\mathrm{T}}(\tau)+\sum_{l=1}^{K_{\mathrm{m}}}\tilde{q}_{l,\mathrm{m}}\int_0^{\tau}h_{\mathrm{T}l,\mathrm{m}}(\tau-t,t)\,\mathrm{d}t+\sum_{j=1}^{K_{\mathrm{un}}}\tilde{q}_{j,\mathrm{un}}\int_0^{\tau}h_{\mathrm{T}j,\mathrm{un}}(\tau-t,t)\,\mathrm{d}t+w_{\mathrm{T}}-\mu_w \quad (10)$$

where $\bar{q}_{\mathrm{T}}(\tau)$ corresponds to the element of $\boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta}$ in (3), the sum of the second and the third terms can be transformed to the element of $\boldsymbol{B}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\eta}$ via Karhunen-Loeve expansion. Note that $\epsilon(\boldsymbol{s})$ in (3) is neglected because the measurement process is assumed to be identical to the hidden process. For simplicity, we define $g_{\mathrm{T}l,\mathrm{m}}(\tau)=\int_0^{\tau}h_{\mathrm{T}l,\mathrm{m}}(\tau-t,t)\,\mathrm{d}t$ and $g_{\mathrm{T}j,\mathrm{un}}(\tau)=\int_0^{\tau}h_{\mathrm{T}j,\mathrm{un}}(\tau-t,t)\,\mathrm{d}t$. Then (10) can be rewritten as

$$q_{\mathrm{T}}(\tau)\approx\bar{q}_{\mathrm{T}}(\tau)+\sum_{l=1}^{K_{\mathrm{m}}}\tilde{q}_{l,\mathrm{m}}g_{\mathrm{T}l,\mathrm{m}}(\tau)+\sum_{j=1}^{K_{\mathrm{un}}}\tilde{q}_{j,\mathrm{un}}g_{\mathrm{T}j,\mathrm{un}}(\tau)+w_{\mathrm{T}}-\mu_w \quad (11)$$

where the average target flow $\bar{q}_{\mathrm{T}}(\tau)$ and $g_{\mathrm{T}l,\mathrm{m}}(\tau)$ can be obtained from the empirical data, $\tilde{q}_{l,\mathrm{m}}$ can be acquired by sensor measurements, $\tilde{q}_{j,\mathrm{un}}$, $g_{\mathrm{T}j,\mathrm{un}}(\tau)$ and w_{T} are unknown parameters. By modeling sum of $\sum_{j=1}^{K_{\mathrm{un}}}\tilde{q}_{j,\mathrm{un}}g_{\mathrm{T}j,\mathrm{un}}(\tau)$ and $w_{\mathrm{T}}-\mu_w$ as a zero-mean Gaussian variable with variance of σ_η^2, the probability function of $q_{\mathrm{T}}(\tau)$ by given $\bar{q}_{\mathrm{T}}(\tau)$ and $\sum_{l=1}^{K_{\mathrm{m}}}\tilde{q}_{l,\mathrm{m}}g_{\mathrm{T}l,\mathrm{m}}(\tau)$ can be expressed by

$$\Pr\left(q_{\mathrm{T}}(\tau)\,|\bar{q}_{\mathrm{T}}(\tau),\sum_{l=1}^{K_{\mathrm{m}}}\tilde{q}_{l,\mathrm{m}}g_{\mathrm{T}l,\mathrm{m}}(\tau)\right)=\frac{1}{\sqrt{2\pi}\sigma_\eta}\exp\left(-\frac{\left(\begin{array}{c}q_{\mathrm{T}}(\tau)-\bar{q}_{\mathrm{T}}(\tau)\\-\sum_{l=1}^{K_{\mathrm{m}}}\tilde{q}_{l,\mathrm{m}}g_{\mathrm{T}l,\mathrm{m}}(\tau)\end{array}\right)^2}{2\sigma_\eta^2}\right) \quad (12)$$

To maximize the conditional probability, the neighboring measured roads should be selected to minimize $E\left(\left(q_{\mathrm{T}}(\tau)-\bar{q}_{\mathrm{T}}(\tau)-\sum_{l=1}^{K_{\mathrm{m}}}\tilde{q}_{l,\mathrm{m}}g_{\mathrm{T}l,\mathrm{m}}(\tau)\right)^2\right)$, i.e., ESEE (Expected Squared Estimation Error), and meanwhile the neighboring unmeasured roads should be selected to minimize σ_η. However, σ_η is an unknown parameter. To tackle the problem, flow conservation law is utilized to find optimum set of neighboring unmeasured roads. In graph theory, the sum of flows entering the vertex (or a closed curve) is equal to the sum of flows leaving the vertex (the closed curve) if the vertex is neither (the closed curve does not contain) a source nor a sink [**?**]. We next will apply the flow conservation law to the road network depicted in Fig.1.

We model the Fig. 1 as a directed graph $G(V,E)$ with a source set S and a sink set T, where V is the set of vertices and $E\in V\times V$ is the set of edges (road segments), respectively. A vertex $v_i\in V$ models a road intersection or an end of a road. An edge $e(v_i,v_j)$, which connects two vertices, represents a directed network segment. The size of source set and sink set are K_{S} and K_{T}, respectively. Then we can obtain Theorem 2.

Theorem 2. *Let us create an arbitrary closed cut in $G(V,E)$ and define the flow on each intersected edge (road segment) as $f_i, i=1\cdots K_C$, where K_C is the total number of edges intersected by the closed cut. Furthermore, let $K_{T,in}$ and $K_{S,in}$ be the number of sinks and sources located inside of the closed cut, respectively, and let $f(v,t_k)$ and $f(s_l,u)$ be a flow from the vertex v to a sink t_k and a flow from a source s_l to the vertex u, respectively. Then, the following equality should hold:*

$$\sum_{k=1}^{K_{C,in}}f_k-\sum_{l=1}^{K_{C,out}}f_l=\sum_{k=1}^{K_{T,in}}\sum_{v\in V}f(v,t_k)-\sum_{l=1}^{K_{S,in}}\sum_{u\in V}f(s_l,u) \quad (13)$$

where $K_C=K_{C,in}+K_{C,out}$ and $K_{C,in}$, $K_{C,out}$ are the number of flows entering the closed cut and leaving the closed cut, respectively.

Proof. The closed cut $\mathbb{C}$ partitions the graph into two disjoint vertex sets, denoted by V_1 and V_2. By defining a flow function between two sets of vertices X and Y as $f(X,Y)=\sum_{x\in X}\sum_{y\in Y}f(x,y)$, the left side of (13) can be expressed by $f(V_1,V_2)$. Reference [**?**] shows that for all $X\in V$, $f(X,X)=0$, and for all $X,Y,Z\in V$ with $X\bigcap Y=\phi$, $f(X\bigcup Y,Z)=f(X,Z)+f(Y,Z)$ and $f(Z,X\bigcup Y)=$

$f(Z,X)+f(Z,Y)$. Hence, we have

$$\begin{aligned} f(V_1,V_2) &= f(V_1,V)-f(V_1,V_1) \\ &= f(V_1,V) \\ &= f(S_{\text{in}},V)+f(T_{\text{in}},V)+f(V_1\backslash\{S_{\text{in}},T_{\text{in}}\},V) \end{aligned} \tag{14}$$

From the flow conservation law, $f(V_1\backslash\{S_{\text{in}},T_{\text{in}}\},V)=0$. Hence, (13) can be rewritten as

$$\begin{aligned} f(V_1,V_2) &= f(S_{\text{in}},V)+f(T_{\text{in}},V) \\ &= \sum_{k=1}^{K_{\text{T,in}}}\sum_{v\in V} f(v,t_k)-\sum_{l=1}^{K_{\text{S,in}}}\sum_{u\in V} f(s_l,u) \end{aligned} \tag{15}$$

Theorem 2 is proved. □

Remark 3. It is worth noting that Theorem 2 ignored the storage capacity of roads, i.e., vehicles stored in the road segments enclosed by the closed cut. Therefore, strictly speaking, the relationship depicted in Theorem 2 only applies to *long-term* traffic flows where the storage capacity is of negligible impact. When Theorem 2 is applied to short-term traffic flows, the equality no longer holds strictly. Moreover, the loop detectors may also create some uncertainties about the number of the passing vehicles. The mismatch between incoming and outgoing traffic flows caused by storage capacity and measurement uncertainties caused by loop detectors can be captured by an error term or can be modeled by a source/sink inside the closed cut.

A closed cut (dotted line) is shown in Fig. 1 as a dotted line, which intersects the target road, road 7, road 11, road 10, road 9 and road 2. Applying Theorem 2, we have

$$\begin{aligned} &q_{\text{T}}+q_{7,\text{out}}+q_{11,\text{out}}+q_{10,\text{out}}+q_{9,\text{out}}+q_{2,\text{out}}= \\ &=q_{\text{T,in}}+q_{7,\text{in}}+q_{11,\text{in}}+q_{10,\text{in}}+q_{9,\text{in}}+q_{2,\text{in}}+w_s+w_t \end{aligned} \tag{16}$$

where q_{T} is the missing target flow and w_s and w_t are the (total) source flow and sink flow within the closed cut, respectively. Note that each vehicle will spend a different amount of time traveling from an entrance to an exit of the closed cut. Then for each arbitrary closed cut crossing the target road, we can get the following relation:

$$\begin{aligned} q_{\text{T}}+\sum_{k=1}^{K_{\text{m}}} q_{k,\text{out}}+\sum_{l=1}^{K_{\text{un}}} q_{l,\text{out}} &= \sum_{k=1}^{K_{\text{m}}} q_{k,\text{in}}+\sum_{l=1}^{K_{\text{un}}} q_{l,\text{in}} \\ &+ w_s+w_t \end{aligned}$$

The above equation can be further rewritten in the following form:

$$\begin{aligned} q_{\text{T}}+\sum_{k=1}^{K_{\text{m}}} q_{k,\text{out}}-\sum_{k=1}^{K_{\text{m}}} q_{k,\text{in}} &= \sum_{l=1}^{K_{\text{un}}} q_{l,\text{in}}-\sum_{l=1}^{K_{\text{un}}} q_{l,\text{out}} \\ &+ w_s+w_t \end{aligned} \tag{17}$$

where the right side of (17) is unknown and it has strong impact on σ_η^2 given by (12). Combined with (12), the optimization objective function to find the OCC can be expressed by

$$\begin{aligned} &\text{C}_{\text{occ}}(\text{T})= \\ &=\underset{\text{C(T)}}{\text{argmin}}\left\{E\left[\begin{array}{c}\left(q_{\text{T}}-\bar{q}_{\text{T}}-\sum_{k=1}^{K_{\text{m}}}\tilde{q}_{k,\text{m}}g_{\text{T}k,\text{m}}\right)^2+ \\ +\left(q_{\text{T}}+\sum_{k=1}^{K_{\text{m}}} q_{k,\text{out}}-\sum_{k=1}^{K_{\text{m}}} q_{k,\text{in}}\right)^2\end{array}\right]\right\} \\ &=\underset{\text{C(T)}}{\text{argmin}}\left\{\begin{array}{c}-2\sum_{k=1}^{K_{\text{m}}} g_{\text{T}k,\text{m}}r_{\text{T}k,\text{m}}+\boldsymbol{g}_{\text{T,m}}^{\text{H}}\boldsymbol{R}_{\text{C(T)}}\boldsymbol{g}_{\text{T,m}} \\ +\sum_{k=1}^{K_{\text{m}}} r_{k,\text{out}}+\sum_{k=1}^{K_{\text{m}}}\sum_{l=1,l\neq k}^{K_{\text{m}}} r_{kl,\text{out}} \\ +2\sum_{k=1}^{K_{\text{m}}} r_{\text{T}k,\text{out}}-2\sum_{k=1}^{K_{\text{m}}} r_{\text{T}k,\text{in}}- \\ -2\sum_{k=1}^{K_{\text{m}}}\sum_{l=1}^{K_{\text{m}}} r_{kl,\text{in,out}}+\sum_{k=1}^{K_{\text{m}}} r_{k,\text{in}} \\ +\sum_{k=1}^{K_{\text{m}}}\sum_{l=1,l\neq k}^{K_{\text{m}}} r_{kl,\text{m}}\end{array}\right\} \end{aligned} \tag{18}$$

where the first part of (18) is to maximize the conditional probability and the second part is to minimize the expected squared unknown metrics, $r_{\text{T}k,\text{m}}$, $\boldsymbol{R}_{\text{C(T)}}$, $r_{k,\text{out}}$, $r_{kl,\text{out}}$, $r_{\text{T}k,\text{out}}$, $r_{\text{T}k,\text{in}}$, $r_{kl,\text{in,out}}$, $r_{k,\text{in}}$ and $r_{kl,\text{m}}$ are co-variance between the target flow and the *k*-th inflow, co-variance matrix of the K_{m} input flows, variance of the *k*-th outflow, co-variance between the *k*-th and *l*-th outflow, co-variance between the target flow and the *k*-th outflow, co-variance between the target flow and the *k*-th inflow, co-variance between the *k*-th inflow and the *l*-th outflow, variance of the *k*-th inflow and co-variance between the *k*-th inflow and the *l*-th inflow, respectively, and can be straightforwardly obtained by the empirical data. The equation (18) will be used in the next subsection to find the OCC.

B. Novel OCC Search Algorithm

It shows in (18) that the optimization procedure is a minimum cut finding problem. Stoer–Wagner algorithm is a classical recursive algorithm which can find the minimum cut in an un-directed graph [**?**]. Unfortunately, the algorithm cannot be applied in our scenario because we need find a minimum weighted closed cut in a directed graph. The closed cut should start and end at the target edge. Brute-force solution is to check all possible neighboring edges and select the one minimizing (18). However, the search complexity will increase exponentially with the number of edges. To tackle the problem, we propose an iterative searching strategy which is a Modified version of Viterbi Algorithm (MVA). VA is a recursive optimal solution to the problem of estimating the state sequence of a discrete time finite-state Markov process observed in memoryless noise [**?**]. The finite-states and transition probabilities in VA are deterministic while MVA has non-deterministic states and transition probabilities for each iteration. Our computer validation shows that the OCC can be efficiently captured for each target link.

To describe MVA more clearly, we firstly start from the scenario that all edges are equipped with detectors and the empirical data are available for all roads. Then we will extend MVA to the scenario of low density of detectors. For the former scenario, we only aim at minimizing ESEE given by

the first term in (18). MVA can be interpreted as an iterative searching solution initiating from the target edge and try to find the optimal detector at each iteration which can minimize ESEE. Let us define the finite-states at the *i*-th iteration as S_i, which contains N_i neighboring edges of l_{i-1} defined by the selected edge at the *i*-1-th iteration. The transition probability from the selected edge l_{i-1} to the s_i-th state is defined by $\pi_{l_{i-1}s_i}$, which is $1/N_i$. The ESEE at the *i*-th iteration for the n_i-th selected neighbor is represented by V_{i,n_i}. From (18), V_{i,n_i} can be determined by

$$\begin{aligned} V_{1,T} &= 0 \\ V_{i,n_i} &= \min_{\boldsymbol{l_{i-1}}} \begin{pmatrix} V_{i-1,l_{i-1}} + g_{\mathrm{T}n_i,\mathrm{m}} \boldsymbol{r}^{\mathrm{H}}_{\boldsymbol{l_{i-1}n_i}} \boldsymbol{g}_{\mathrm{T}\boldsymbol{l_{i-1}},\mathrm{m}} + \\ \boldsymbol{g}^{\mathrm{H}}_{\mathrm{T}\boldsymbol{l_{i-1}},\mathrm{m}} \boldsymbol{r}_{\boldsymbol{l_{i-1}n_i}} g_{\mathrm{T}n_i,\mathrm{m}} + \\ g_{\mathrm{T}n_i,\mathrm{m}} r_{n_i} g^{\mathrm{H}}_{\mathrm{T}n_i,\mathrm{m}} - 2 r_{\mathrm{T}n_i} g_{\mathrm{T}n_i,\mathrm{m}} \end{pmatrix} \\ R_{n_i} &= \begin{bmatrix} \boldsymbol{R}_{\boldsymbol{l_{i-1}}} & \boldsymbol{r}_{\boldsymbol{l_{i-1}n_i}} \\ \boldsymbol{r}^{\mathrm{H}}_{\boldsymbol{l_{i-1}n_i}} & r_{n_i} \end{bmatrix} \\ \boldsymbol{g}^{\mathrm{H}}_{\mathrm{T}\boldsymbol{n_i},,\mathrm{m}} &= \begin{bmatrix} \boldsymbol{g}^{\mathrm{H}}_{\mathrm{T}\boldsymbol{l_{i-1}},\mathrm{m}} & g_{\mathrm{T}n_i,\mathrm{m}} \end{bmatrix} \end{aligned} \tag{19}$$

where $\boldsymbol{l_{i-1}}$ contains all selected roads at the *i*-1 -th iteration, $\boldsymbol{r}_{\boldsymbol{l_{i-1}n_i}}$, r_{n_i} and $r_{\mathrm{T}n_i}$ are the co-variance vector between $\boldsymbol{l_{i-1}}$ and n_i, variance of n_i and co-variance between T and n_i. Then MVA can be described by Algorithm 1. In the line 3, a queue is created to store the ESEE and the selected road at the initial iteration, the "while" loop from the line 13 to line 23 selects each crossed edge by minimizing the ESEE at each iteration, and the OCC can finally be determined by tracing each edge back to its parent which is defined by the selected edge in the last iteration.

For the latter scenario, a number of the unmeasured roads appear in the network. The flow conservation given by (18) need to be utilized to improve the estimation performance because of the lack of detectors. However, algorithm 1 cannot be directly applied to the latter scenario because the second term in (18) cannot be determined unless a cut is pre-given. It makes impossible for algorithm 1 to iteratively incorporate the flow conservation during the searching procedure. To tackle the problem, an approximation is made for (18) aiming at minimizing the number of crossed unmeasured roads while minimizing the first term in (18). We try to select the edge at each iteration being able to provide the maximum average Variance of the Hypothetical Means (VHM), which is interpreted as the difference between the variance and the conditional variance, then divided by the number of crossed roads. From the law of total variance, the variance of the target flow can be expressed by

$$\begin{aligned} \mathrm{var}\left(q_{\mathrm{T}}\right) &= E_{\boldsymbol{q_m}}\left(var_{q_{\mathrm{T}}}\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)\right) + var_{\boldsymbol{q_m}}\left(E_{q_{\mathrm{T}}}\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)\right) \\ &= \textit{ESEE} + \textit{VHM} \end{aligned} \tag{20}$$

The dual problem of ESEE minimization is to maximize the VHM at each iteration. The modified algorithm can be found in algorithm 2.

Remark 4. Note that the number of cuts being found for the given area is determined by the empirical data and topological information. Therefore, the number of determined cuts varies with different target road segment. For each target road segment, there is only one optimal cut which is used to estimate the missing data. For an example, for the target roads "Snowy Mountains Highway" and "Monaro Highway", there are 4 and 18 closed cuts being determined via the proposed algorithm, respectively.

Algorithm 1 MVA for the former scenario

1: **Input:** target road *T*, graph *G*
2: **Output:** optimum closed cut *OCC*
3: Initialize an empty *OCC*, empty edge array *Ecell* and a new queue Q with the initial information about *T*,$V_{1,T} = 0$;
4: **While** Q is not empty **do**
5: Get the ESEE $\boldsymbol{V_{i-1}}$ and the crossed edges $\boldsymbol{l_{i-1}}$ at $i-1$-th iteration, pop Q
6: **For** $n_l = 1, 2, ..., \boldsymbol{L_{i-1}}$ **do**
7: Find the neighboring edges $\boldsymbol{S_i}$ for l_{i-1,n_l}
8: **For** $n_i = 1, 2, ...,$ *number of neighboring edges* **do**
9: **If** S_{i,n_i} approaches the target road
10: **Then** remove S_{i,n_i} and continue
11: **Else** update $V_{i-1,l_{i-1}}$ to V_{i,n_i} via (19)
12: Search the minimum ESEE $V_{i,n_i,\min}$ through Q
13: **If** $V_{i,n_i} = V_{i,n_i,\min}$
14: **Then** $\boldsymbol{V_i} = [\boldsymbol{V_{i-1}} \quad V_{i,n_i}], \boldsymbol{l_i} = [\boldsymbol{l_{i-1}} \quad S_{i,n_i}]$
15: **Else** remove S_{i,n_i} and continue
16: **If** S_{i,n_i} is a new edge
17: **Then** add the new edge to *Ecell*
18: **End For**
19: **If** $\boldsymbol{l_i}$ is not empty
20: **Then** construct a new element *info* with $\boldsymbol{V_i}$ and $\boldsymbol{l_i}$
21: push *info* into Q
22: **End For**
23: **End While**
24: *OCC* is the concatenation of the *parent* field of each *Ecell* element

Algorithm 2 MVA for the latter scenario

The following modifications should be made:
line 11: update $V_{i-1,l_{i-1}}$ to V_{i,n_i}
line 12: Search the maximum expected variance improvement $V_{i,n_i,\max}$ through Q
line 13: **if** $V_{i,n_i} = V_{i,n_i,\max}$
The rest lines are identical to those in algorithm 1.

C. OCC Based Novel Estimator

This subsection proposes the OCC based Kriging estimator and the OCC based novel estimator which incorporates the flow conservation law. After that OCC is determined by algorithm 2, the missing data at the target road can be estimated via a Kriging estimator [**?**].

$$\hat{q}_{\mathrm{T,kriging}} = \bar{q}_{\mathrm{T}} + \boldsymbol{g}^{\mathrm{H}}_{\mathrm{T,m}}\left(\boldsymbol{q}_{\mathrm{m}} - \bar{\boldsymbol{q}}_{\mathrm{m}}\right) \tag{21}$$

where $\boldsymbol{q}_{\mathrm{m}}$ and $\bar{\boldsymbol{q}}_{\mathrm{m}}$ contains the instantaneous flow and average flow for each crossed sensors, the vector $\boldsymbol{g}_{\mathrm{T,m}}$ consists of K_{m} scaling factors $g_{\mathrm{T}k,\mathrm{m};k=1\cdots K_{\mathrm{m}}}$ between the target flow and the

K_{m} measured flows and can be straightforwardly obtained by the empirical data. Then the conditional expectation $E\left(\hat{q}_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)$ and the conditional variance $\mathrm{var}\left(\hat{q}_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)$ can be expressed by

$$E\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)=\hat{q}_{\mathrm{T,kriging}} \tag{22}$$

and

$$\mathrm{var}\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)=\mathrm{var}\left(q_{\mathrm{T}}\right)-\boldsymbol{g}_{\mathrm{T,m}}^{\mathrm{H}}\boldsymbol{R}_{\boldsymbol{q}_{\mathrm{m}}}\boldsymbol{g}_{\mathrm{T,m}} \tag{23}$$

To our known, the conditional PDF (Probability Distribution Function) $\mathrm{P}\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)$ is a Gaussian function [**?**]. Thus, the OCC based Kriging estimation can be formulated as

$$q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}=\hat{q}_{\mathrm{T,kriging}}+\zeta \tag{24}$$

where $\zeta\sim N\left(0,\mathrm{var}\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)\right)$. To further reduce the uncertainty, we try to incorporate the flow conservation law given by (17). Let us define the OCC for the target road T as $C\left(\mathrm{T}\right)$. From (17), q_{T} can be written as

$$\begin{aligned}q_{\mathrm{T}} &= \sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{in}}-\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{out}}+\sum_{l=1}^{K_{\mathrm{un}}}q_{l,\mathrm{in}}-\sum_{l=1}^{K_{\mathrm{un}}}q_{l,\mathrm{out}}+w_s+w_t\\ &= \sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{in}}-\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{out}}+\gamma\end{aligned} \tag{25}$$

where $\gamma\sim N\left(\begin{array}{c}\sum_{l=1}^{K_{\mathrm{un}}}\bar{q}_{l,\mathrm{in}}-\sum_{l=1}^{K_{\mathrm{un}}}\bar{q}_{l,\mathrm{out}}\\ \sum_{l=1}^{K_{\mathrm{un}}}r_{l,\mathrm{in}}+\sum_{l=1}^{K_{\mathrm{un}}}r_{l,\mathrm{out}}+r_s+r_t\end{array}\right)$, $r_{l,\mathrm{in}}$, $r_{l,\mathrm{out}}$, r_s and r_t are the flow variance for the in and out direction at the l-th crossed edge, and the variance of generation flow and dissipation flow within the closed cut, respectively. Then the conditional expectation $E\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)$ and the conditional variance $\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)$ can be expressed by

$$E\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)=\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{in}}-\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{out}}+\sum_{l=1}^{K_{\mathrm{un}}}\bar{q}_{l,\mathrm{in}}-\sum_{l=1}^{K_{\mathrm{un}}}\bar{q}_{l,\mathrm{out}} \tag{26}$$

and

$$\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)=\sum_{l=1}^{K_{\mathrm{un}}}r_{l,\mathrm{in}}+\sum_{l=1}^{K_{\mathrm{un}}}r_{l,\mathrm{out}}+r_s+r_t \tag{27}$$

The better estimation can be obtained by maximizing the joint probability function $\Pr\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}},C\left(\mathrm{T}\right)\right)$. With the Bayesian theorem, the objective function can be expressed by

$$\begin{aligned}\hat{q}_{\mathrm{T,ML}} &= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\Pr\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}},C\left(\mathrm{T}\right)\right)\right)\\ &= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\frac{\Pr\left(\boldsymbol{q}_{\mathrm{m}},C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)\Pr\left(q_{\mathrm{T}}\right)}{\Pr\left(\boldsymbol{q}_{\mathrm{m}},C\left(\mathrm{T}\right)\right)}\right)\\ &= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\Pr\left(\boldsymbol{q}_{\mathrm{m}}|q_{\mathrm{T}}\right)\Pr\left(C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)\Pr\left(q_{\mathrm{T}}\right)\right)\\ &= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\Pr\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)\Pr\left(\boldsymbol{q}_{\mathrm{m}}\right)\Pr\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)/\Pr\left(q_{\mathrm{T}}\right)\right)\\ &= \underset{q_{\mathrm{T}}}{\mathrm{argmax}}\left(\begin{array}{c}-\frac{\left(q_{\mathrm{T}}-E\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)\right)^2}{2\mathrm{var}\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)}\\ +\frac{\left(q_{\mathrm{T}}-\bar{q}_{\mathrm{T}}\right)^2}{2\mathrm{var}\left(q_{\mathrm{T}}\right)}\\ -\frac{\left(q_{\mathrm{T}}-E\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)\right)^2}{2\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)}\end{array}\right)\end{aligned} \tag{28}$$

By setting the first derivative of (28) with regard to q_{T} to zero, $\hat{q}_{\mathrm{T,ML}}$ can be expressed by

$$\hat{q}_{\mathrm{T,ML}}=\frac{\frac{\bar{q}_{\mathrm{T}}}{\mathrm{var}\left(q_{\mathrm{T}}\right)}-\frac{E\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)}{\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)}-\frac{E\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)}{\mathrm{var}\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)}}{\frac{1}{\mathrm{var}\left(q_{\mathrm{T}}\right)}-\frac{1}{\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)}-\frac{1}{\mathrm{var}\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)}} \tag{29}$$

where $E\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)$, $E\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)$, $\mathrm{var}\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)$ and $\mathrm{var}\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right)\right)$ are given by (22), (26), (23) and (27), respectively, and they can be straightforwardly obtained by the empirical data.

Note that (28) and (29) are based on the assumption that the traffic flow on the crossed measured roads and the unmeasured roads are independent i.e. $\Pr\left(\boldsymbol{q}_{\mathrm{m}},C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)=\Pr\left(\boldsymbol{q}_{\mathrm{m}}|q_{\mathrm{T}}\right)\Pr\left(C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)$. In real scenario, however, it can depict the dependence between them. Thus, the conditional probability should be rewritten as

$$\Pr\left(\boldsymbol{q}_{\mathrm{m}},C\left(\mathrm{T}\right)|q_{\mathrm{T}}\right)=\Pr\left(\boldsymbol{q}_{\mathrm{m}}|q_{\mathrm{T}}\right)\Pr\left(C\left(\mathrm{T}\right)|\boldsymbol{q}_{\mathrm{m}},q_{\mathrm{T}}\right) \tag{30}$$

where the latter term represents the conditional probability of sum of the unmeasured flow and the missing flow based on input flow and the target flow, while the former term stands for the conditional probability of the input flow by giving the target flow. Because of the causal relationship between flows [**?**], the crossed flows on the OCC can be classified into causal flow and effect flow defined by the input and output of the OCC. Recall that γ is a set of unobserved data and can be modeled as a Gaussian variable, which is sum of the unmeasured input flow, the unmeasured output flow and the missing flow. The causal relation between $\boldsymbol{q}_{\mathrm{m}}$ and γ can be utilized to obtain a more accurate PDF. Then the second term of (30) can be transformed to

$$\begin{aligned}\Pr\left(C\left(\mathrm{T}\right)|\boldsymbol{q}_{\mathrm{m}},q_{\mathrm{T}}\right) &= \frac{\Pr\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right),\boldsymbol{q}_{\mathrm{m}}\right)\Pr\left(C\left(\mathrm{T}\right),\boldsymbol{q}_{\mathrm{m}}\right)}{\Pr\left(\boldsymbol{q}_{\mathrm{m}},q_{\mathrm{T}}\right)}\\ &= \frac{\Pr\left(C\left(\mathrm{T}\right),\boldsymbol{q}_{\mathrm{m}}\right)\exp\left(-\frac{\left(\begin{array}{c}q_{\mathrm{T}}-\sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{out}}+\\ \sum_{k=1}^{K_{\mathrm{m}}}q_{k,\mathrm{in}}-E\left(C\left(\mathrm{T}\right)|\boldsymbol{q}_{\mathrm{m}}\right)\end{array}\right)^2}{2\mathrm{var}\left(C\left(\mathrm{T}\right)|\boldsymbol{q}_{\mathrm{m}}\right)}\right)}{\Pr\left(\boldsymbol{q}_{\mathrm{m}}\right)\Pr\left(q_{\mathrm{T}}\right)\sqrt{2\pi\mathrm{var}\left(C\left(\mathrm{T}\right)|\boldsymbol{q}_{\mathrm{m}}\right)}}\end{aligned} \tag{31}$$

The reason that we write $\Pr\left(\boldsymbol{q}_{\mathrm{m}},q_{\mathrm{T}}\right)=\Pr\left(\boldsymbol{q}_{\mathrm{m}}\right)\Pr\left(q_{\mathrm{T}}\right)$ in (31) is the dependence between $\boldsymbol{q}_{\mathrm{m}}$ and q_{T} has been taken into

account in the first term of (30), and (31) only considers the causal relation between $\boldsymbol{q}_{\mathrm{m}}$ and γ, q_{T} and γ. Hence, (28) can be improved as

$$\begin{aligned}\hat{q}_{\mathrm{T,ML,improved}} &= \underset{q_{\mathrm{T}}}{\operatorname{argmax}} \left(\Pr\left(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}}\right)\Pr\left(q_{\mathrm{T}}|C\left(\mathrm{T}\right),\boldsymbol{q}_{\mathrm{m}}\right)/\Pr\left(q_{\mathrm{T}}\right)\right) \\ &= \underset{q_{\mathrm{T}}}{\operatorname{argmax}} \left(\begin{array}{c} \frac{\left(q_{\mathrm{T}}-\bar{q}_{\mathrm{T}}\right)^2}{2\mathrm{var}(q_{\mathrm{T}})} - \frac{\left(q_{\mathrm{T}}-E(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}})\right)^2}{2\mathrm{var}(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}})} - \\ \frac{\left(q_{\mathrm{T}}-\left(\sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{in}}-\sum_{k=1,k\neq \mathrm{T}}^{K_{\mathrm{m}}} q_{k,\mathrm{out}}\right)-E(C(\mathrm{T})|\boldsymbol{q}_{\mathrm{m}})\right)^2}{2\mathrm{var}(C(\mathrm{T})|\boldsymbol{q}_{\mathrm{m}})} \end{array} \right) \end{aligned} \tag{32}$$

Then $\hat{q}_{\mathrm{T,ML,improved}}$ can be expressed by

$$\begin{aligned}&\hat{q}_{\mathrm{T,ML,improved}} = \\ &= \frac{\frac{\bar{q}_{\mathrm{T}}}{\mathrm{var}(q_{\mathrm{T}})} - \frac{\left(\sum_{k=1}^{K_{\mathrm{m}}} q_{k,\mathrm{in}}-\sum_{k=1,k\neq \mathrm{T}}^{K_{\mathrm{m}}} q_{k,\mathrm{out}}\right)+E(C(\mathrm{T})|\boldsymbol{q}_{\mathrm{m}})}{\mathrm{var}(C(\mathrm{T})|\boldsymbol{q}_{\mathrm{m}})} - \frac{E(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}})}{\mathrm{var}(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}})}}{\frac{1}{\mathrm{var}(q_{\mathrm{T}})} - \frac{1}{\mathrm{var}(C(\mathrm{T})|\boldsymbol{q}_{\mathrm{m}})} - \frac{1}{\mathrm{var}(q_{\mathrm{T}}|\boldsymbol{q}_{\mathrm{m}})}}\end{aligned} \tag{33}$$

and

$$\begin{aligned} E\left(C\left(\mathrm{T}\right)|\boldsymbol{q}_{\mathrm{m}}\right) &= \bar{\gamma} + \boldsymbol{\sigma}_{\gamma,\boldsymbol{q}_{\mathrm{m}}}\boldsymbol{\sigma}_{\boldsymbol{q}_{\mathrm{m}},\boldsymbol{q}_{\mathrm{m}}}^{-1}\left(\boldsymbol{q}_{\mathrm{m}}-\bar{\boldsymbol{q}}_{\mathrm{m}}\right) \\ \mathrm{var}\left(C\left(\mathrm{T}\right)|\boldsymbol{q}_{\mathrm{m}}\right) &= \mathrm{var}\left(\gamma\right) - \boldsymbol{g}_{\boldsymbol{\gamma},\mathrm{m}}^{\mathrm{H}}\boldsymbol{R}_{\boldsymbol{q}_{\mathrm{m}}}\boldsymbol{g}_{\boldsymbol{\gamma},\mathrm{m}} \end{aligned} \tag{34}$$

where $\boldsymbol{\sigma}_{\gamma,\boldsymbol{q}_{\mathrm{m}}}$ and $\boldsymbol{g}_{\gamma,\mathrm{m}}$ are the co-variance matrix and conditional correlation coefficients between γ and $\boldsymbol{q}_{\mathrm{m}}$, respectively, and can be easily obtained by the empirical data.

V. Results and Discussion

In this section, we evaluate the proposed imputation strategy by comparing to other two imputation methods, NHA (Nearest Historical Average) and *k*NN. Various missing type and missing ratio are observed in performance comparison among the three imputation methods by MAPE and RMSE.

A. Data

Traffic flow data in this study was provided by Sydney RMS (Roads and Maritime Services). The selected data was collected by loop detectors on the arterial roads located in Sydney south area over 198 days (Fig. 2). All measured roads are single-lane bi-directional roads. Each detector provides the flow data of 1-hour interval from 00:00 - 23:00h on each day. Each green node in Fig. 2 represents a detector. The data can be represented as a tensor $T \in \mathbb{R}^{N_{\mathrm{m}}\times M\times K}$, where $N_{\mathrm{m}} = 10$, $M = 24$ and $K = 181$. We choose K to be 181 instead of 198 because the weekend and public days are removed due to the different traffic pattern.

B. Generation of Missing Data

To evaluate the imputation performance of each methods, the missing data are intentionally generated with different missing ratio that ranges from 20% to 50% at every 10% increment as usual in the research field ([**?**], [**?**]). To verify the robustness of the proposed methods, we consider three types of missing data in this research: 1) Missing Completely at Random manner (MCR), where the missing points are independently and uniformly distributed over the spatio-temporal domain. This may occur due to temporary from power or communication failures [**?**]. 2) Missing Group Randomly in the Temporal domain (MGRT), where the missing points appear as a group of fixed length sequential points lost at one road, and the group is independently and uniformly distributed over the temporal domain. This may occur due to a prolonged physical damage, malfunction of communication device or temporary detector deployment. 3) Not Missing Randomly (NMR), where the occurrences of missing data are scattered and simultaneous over different roads. NMR is often caused by a long time malfunction of the loop detectors [**?**].

Fig. 2: Selected map for experiment: Sydney South Area

C. Imputation Techniques for Comparison Analysis

NHA is the most common method in the data imputation because it shows a stable performance regardless of the missing data size with easy implementation [**?**]. NHA replaces the missing data by arithmetic average or weighted average of the nearest historical data [**?**]. NHA does not incorporate the information from neighboring roads at the same day and is based on the assumption that traffic pattern at the same detector at the same time is similar from day to day. In this study, we fill the missing data with the arithmetic average data of the same time over 10 historical days.

The second comparison method is *k*NN method which has been discussed in [**?**], [**?**]. The original *k*NN method is to fill the missing data with arithmetic or weighted average of data on k neighboring roads. The k neighboring roads are selected by searching for the data with close physical distances with the target road. In [**?**], the authors proposed the improved *k*NN which replaced the physical distance with the equivalent distance, which is related to the physical distance among roads h, connective grade of a road g and correlation coefficient between the historical time series of two roads r. The k neighboring roads are selected by a given suitable threshold of the equivalent distance. Then the missing data on the target road is estimated by the arithmetic average of the data on the k neighboring roads.

D. Results and Discussion

In this section, we examine the imputation performance of our novel approaches: OCC based Kriging (21), OCC based ML (29) and improved ML (33) and compare them to correlative *k*NN [**?**] and NHA [**?**] in terms of MAPE over different missing ratios and three missing patterns. The performance of the proposed approaches was evaluated with 198 days of the historical data. Missing data in testing were intentionally produced from the available data sheet and compared with the actual value for the performance evaluation.

Fig. 3 and Fig. 4 depict MAPE and RMSE of the three novel imputation methods for MCR pattern, respectively. The accuracy results of imputation represented by MAPE show that the three novel imputation methods dominantly outperform the correlative *k*NN and NHA over the missing ratio from 0.25 to 0.5. By incorporating the flow conservation law introduced in Theorem 2, the imputation performance can be further improved via OCC based ML and improved ML. Fig.5, 6 and Fig.7, 8 show the imputation performance for MGRT and NMR patterns, respectively. As shown in Fig. 5, OCC based ML and improved ML depict better estimation performance than correlative *k*NN and NHA over almost whole scale of missing ratios while OCC based Kriging is more appropriate for the missing ratio being lower than 0.35. Beyond 0.35, OCC based Kriging shows a worse performance than the comparing methods. For NMR pattern, the three novel approaches slightly outperforms the comparing methods over almost whole scale of missing ratio.

Comparing the three novel approaches, the improved ML shows the best imputation performance for three missing patterns because it incorporates the flow conservation law and takes into account the dependence between the measured and unmeasured roads.

Comparing the three missing patterns, three novel methods depict the best imputation performance for MCR pattern while the worst performance for NMR.

It is observed in Fig. 5 that the performance of OCC based Kriging for MGRT pattern becomes worse than NHA and correlative *k*NN when missing ratio is larger than 0.35. This is mainly because for group missing pattern, the number of the measured neighboring roads captured by OCC decreases quickly with the increase of the missing ratio and thus the correlation between the measured neighboring and the target roads plays less role compared to the flow conservation for the missing data imputation.

To summarize, our results show that the three novel methods perform better than the comparing methods in almost all missing pattern, with exception for MGRT for which the OCC based Kriging performs worse than the two comparing methods for the missing ratio being larger than 0.35. The improved ML outperforms all other methods for all missing patterns and missing ratios.

Remark 5. Although in this paper, we only considered the non-congested case. Theoretically, congestion and non-recurrent events pose no impact on the performance of our methods as long as the sampling period is much larger than the travel time. Because in this case, almost all traffic flow measured by

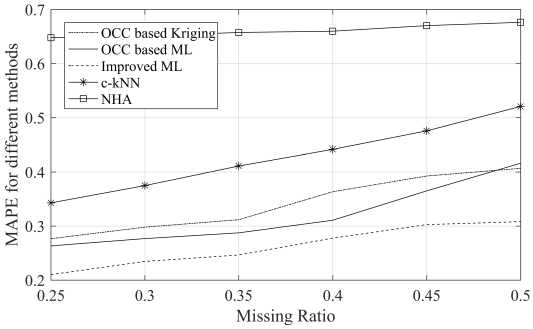

Fig. 3: MAPE of five different methods in MCR pattern

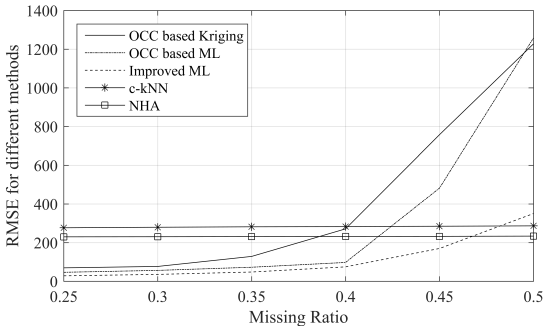

Fig. 4: RMSE of five different methods in MCR pattern

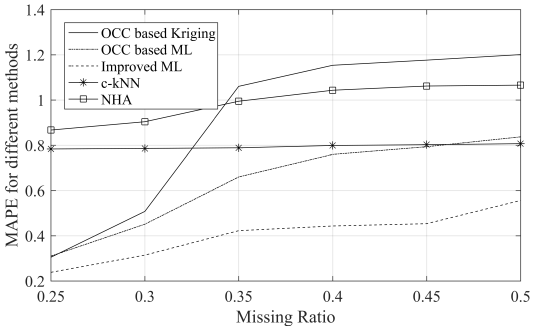

Fig. 5: MAPE of five different methods in MGRT pattern

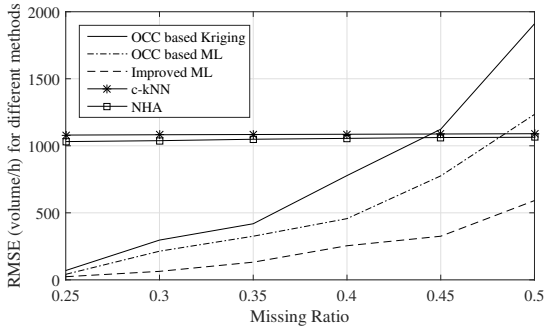

Fig. 6: RMSE of five different methods in MGRT pattern

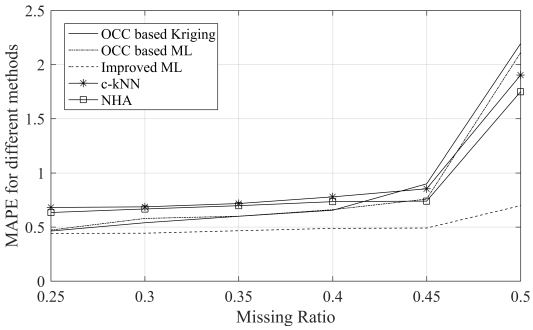

Fig. 7: MAPE of five different methods in NMR pattern

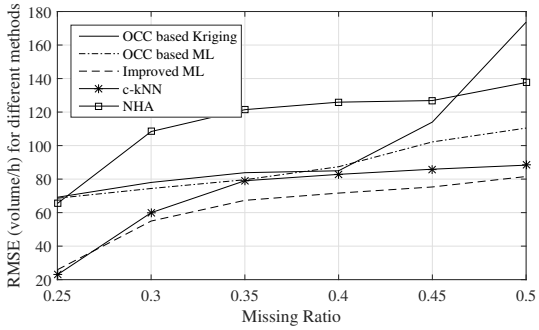

Fig. 8: RMSE of five different methods in NMR pattern

the target detector originates from the flow measured by its upstream detectors during the same time slot. In the case that the travel time is much larger than the sampling period due to congestion or non-recurrent events, the time lag should be considered to improve the performance.

VI. Conclusions

In this paper, an OCC based imputation strategy has been proposed for traffic flow incompleteness in urban network. Based on the determination of optimum sensors for imputation via novel OCC finding algorithm, we compare three different estimators: OCC based kriging, OCC based ML (Maximum Likelihood) estimator and improved ML estimator in terms of MAPE for three missing patterns: MCR, MGRT and NMR. In addition, our three novel methods are compared to NHA and correlative kNN. From our experimental results, we can conclude that

1) Our three novel methods outperforms NHA and kNN for three missing patterns over almost all missing ratios because the topological information was utilized and a sophisticated OCC finding algorithm was designed to determine the optimum sensors before imputation.
2) The two ML estimators can deliver a better estimation performance than OCC based Kriging because the flow conservation law has been incorporated.
3) By consideration of the dependence between the measured and unmeasured roads, the estimation accuracy can be further improved. Therefore, the improved ML estimator is the most appropriate imputation scheme for all missing patterns.

VII. Future Work

Our proposed methods are evaluated with 1-hour data due to unavailability of finer data. In future work, we will implement the proposed methods in finer data (15 mins, 5 mins, 30 seconds) and consider the effect of time-lag. The complexity of the proposed algorithm can further be reduced.

Acknowledgment

The authors thank Sydney RMS for their traffic data and kind assistance

Shangbo Wang received his master degree and Dr.-Ing degree in University of Duisburg-Essen, Germany in 2007 and 2014, respectively. Prior to joining in University of Technology Sydney in 2015 as research assistant, he has worked by Siemens AG in Munich, Germany and Continental AG in Lindau, Germany as research engineer and development engineer, respectively. His research interest includes statistical signal processing, intelligent transportation systems, digital signal processing, wireless localization and mobile communication.

Guoqiang Mao (S'98-M'02-SM'08) joined the University of Technology Sydney in February 2014 as Professor of Wireless Networking and Director of Center for Real-time Information Networks. Before that, he was with the School of Electrical and Information Engineering, the University of Sydney. The Center is among the largest university research centers in Australia in the field of wireless communications and networking. He has published about 200 papers in international conferences and journals, which have been cited more than 4500 times. He is an editor of the IEEE Transactions on Wireless Communications (since 2014), IEEE Transactions on Vehicular Technology (since 2010) and received "Top Editor" award for outstanding contributions to the IEEE Transactions on Vehicular Technology in 2011, 2014 and 2015. He is a co-chair of IEEE Intelligent Transport Systems Society Technical Committee on Communication Networks. He has served as a chair, co-chair and TPC member in a large number of international conferences. His research interest includes intelligent transport systems, applied graph theory and its applications in telecommunications, Internet of Things, wireless sensor networks, wireless localization techniques and network performance analysis.