Data-Driven Imputation Method for Traffic Data in Sectional Units of Road Links

Sehyun Tak, Soomin Woo, and Hwasoo Yeo, *Member, IEEE*

***Abstract*—Missing data imputation is a critical step in data processing for intelligent transportation systems. This paper proposes a data-driven imputation method for sections of road based on their spatial and temporal correlation using a modified k-nearest neighbor method. This computing-distributable imputation method is different from the conventional algorithms in the fact that it attempts to impute missing data of a section with multiple sensors that have correlation to each other, at once. This increases computational efficiency greatly compared with other methods, whose imputation subject is individual sensors. In addition, the geometrical property of each section is conserved; in other words, the continuation of traffic properties that each sensor captures is conserved, therefore increasing accuracy of imputation. This paper shows results and analysis of comparison of the proposed method to others such as nearest historical data and expectation maximization by varying missing data type, missing ratio, traffic state, and day type. The results show that the proposed algorithm achieves better performance in almost all of the missing types, missing ratios, day types, and traffic states. When the missing data type cannot be identified or various missing types are mixed, the proposed algorithm shows accurate and stable imputation performance.**

***Index Terms*—Imputation, kNN, big data, intelligent transportation system.**

I. Introduction & Motivation

As computing power grows ever more in this era of big data, the ITS community also meets a new opportunity to process its vast amount of data to extract precious information from it. Better knowledge from big data enables the transportation management systems to provide real-time services in a much larger scale, such as monitoring, travel time prediction, and traveler information, in order to improve mobility, safety and efficiency.

However, these transportation technologies have been limited in their accuracy due to the lack of complete datasets [1]–[10]. For example, Tan *et al.* [11] found more than 5% are missing from the PeMS traffic flow database, whereas Ni *et al.* [12] found a missing rate between 16 to 93% from the Texas Transportation Institute. Also Qu *et al.* [1] found missing ratio in Beijing usually around 10% but sometimes it amounts between 20 to 25% for various reasons.

Moreover, transportation data is unique due to the catastrophic changes of traffic state, sudden speed drops, and correlation with neighbors in time and space intervals. Because of these explicit characteristics of transportation data, the missing data imputation technique must be carefully chosen and needs to provide reliable and accurate data for ITS services such as travel time prediction. The existing data imputation can be categorized into two groups as shown in Fig. 1(a) and (b). Fig. 1(a) shows imputation methods that use temporal relations of individual sensors. Fig. 1(b) shows imputation methods that use both temporal and spatial relations of individual sensors.

First, many researchers have developed numerous methods that attempt to impute individual sensors with a temporal analysis as Fig. 1(a), for example using PPCA, neural network, k-Nearest Neighbors, ARIMA, multiple imputation, and Markov Chain Monte Carlo [1], [2], [5], [6], [9], [10], [12]–[16]. However these methods cannot perform well with large missing ratio often in practice with sometimes over a few days or months.

To overcome the problem with large missing ratio, many papers adopted spatial and temporal aspects of missing data imputation as shown in Fig. 1(b), including KPPCA, kernel regression, k-Nearest Neighbors, and mean matching multiple imputation method [3], [4], [7], [17], [18]. A few of these studies assumed spatial correlation between neighboring detectors in the arbitrary spatial window for imputation procedure, without assessing the actual relationship [3], [7]. However, this assumption may break when neighboring detectors are not correlated to each other, for example in a case where a junction or an on-ramp exists between the neighboring detectors, severing the continuity of traffic characteristics.

There are a few papers that have assessed correlation with neighboring detectors before imputation. Still, they lack of reasoning behind the threshold of correlation value in selection of neighboring detectors for imputation [17], [18]. For instance, Henrickson *et al.* used a clear cut on the correlation threshold, Pearson correlation coefficient great than 0.1 [18]. However, this number is arbitrary and a constant value for threshold may be inappropriate to apply for a wide range of detectors on the road because its scale may vary between different sites and detectors.

In order to spatially define the scope of imputation with scientific reasoning, it is possible to impute missing data based on sections of road, where a same traffic property will be shared within, as done by [19]. This study defines a section

Manuscript received October 15, 2014; revised April 21, 2015, October 13, 2015, and November 30, 2015; accepted January 30, 2016. Date of publication March 30, 2016; date of current version May 26, 2016. This work was supported by the Railroad Technology Research Program (RTRP) under Grant 14RTRP-B081103-01 funded by the Ministry of Land, Infrastructure and Transport of the Korean Government. The Associate Editor for this paper was W.-H. Lin.

The authors are with the Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: taksehyun@kaist.ac.kr; sssm731@kaist.ac.kr; hwasoo@kaist.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2530312

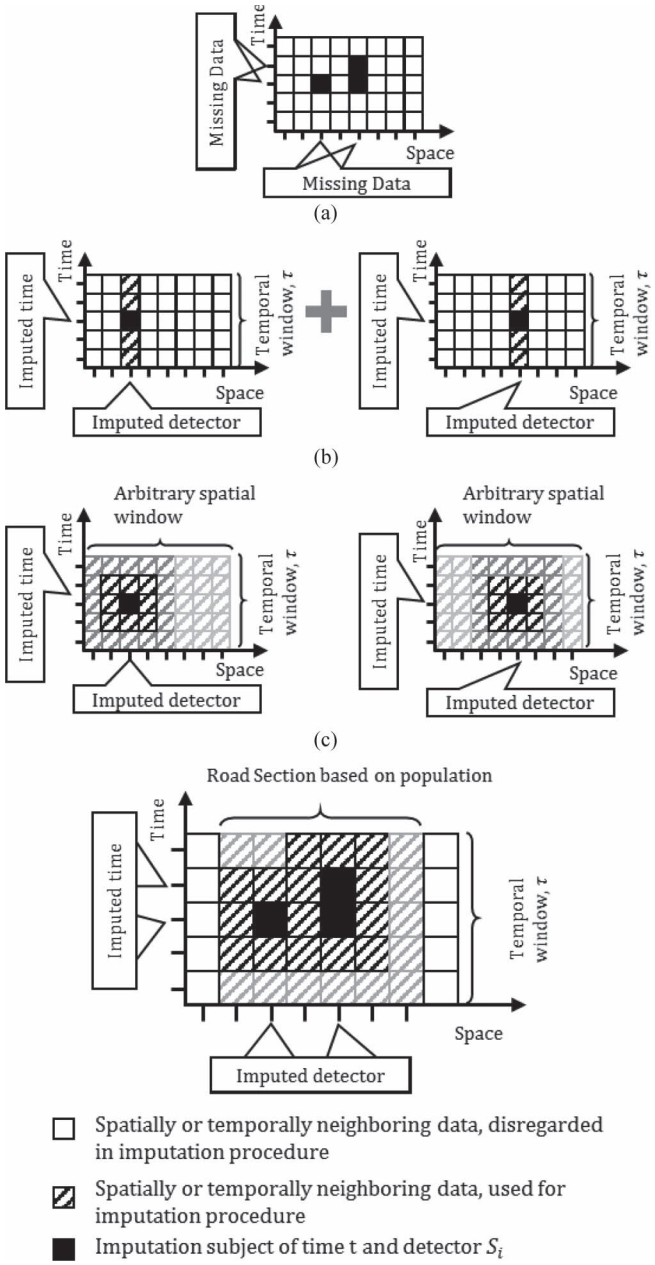

Fig. 1. Three categories of data imputation methods. (a) Temporal imputation method. (b) Temporal and spatial imputation method. (c) Temporal and spatial imputation method based on correlation between detectors.

of arterial road network as the spatial range for one single imputation process, however its Singular Value Decomposition method approximates and cannot be used to capture the detailed movement of traffic on highway.

One method to define the spatial scope to impute missing data based on sections of road is by grouping links with a shared traffic property. A group of links may share a traffic property if similar vehicle populations are contained in that group because this group would be severed from other link groups via geometry of road, i.e., junctions and on/off ramps, in a traffic engineering perspective. Because a group of links will spatially share a traffic property, missing data imputation can be done together at once. We will define this group unit in this paper as a road section.

Therefore in this paper, a road section is defined to capture its shared traffic property that severs them from other road links in a traffic engineering perspective, by observing the trend of spatial correlation along the detectors on the road. A sudden drop of spatial correlation of a moving window will represent a point where a road section finishes. This will define a road section that will be imputed as a unit, as shown in Fig. 1(c). This method analyzes both spatial and temporal relations of detectors but differs from previous studies in that it pre-defines the spatial range of imputation subject as a group of multiple sensors with correlation within themselves. Therefore, the imputation subject is no longer individual sensors, but multiple individual sensors in a road section.

For the imputation of highway traffic data, a modified kNN method is chosen in this paper for spatial and temporal missing data imputation of a section with multiples sensors, grouped with strong correlation within themselves. kNN method that once was deemed to be too expensive and inefficient to compute can now be adopted to a distributed-computing environment and can process a large amount of traffic data at a high speed with outstanding performance [20], [21]. Therefore, it can improve both computing efficiency and accuracy of imputation compared to the existing imputation methods.

II. Sectional Imputation Method

In this paper, we propose a sectional imputation method based on the kNN algorithm, which is a data-driven, heuristic approach to use non-parametric regression. The core idea of sectional kNN method is to search for the first k number of most similar historical data to the subject data for each divided road section, and then integrate them to impute for the missing values.

The key strength of kNN method in imputing missing data of road section is first because it can capture the details of traffic change without any approximation or smoothing because the input and output data are not aggregated nor represented by parametric values. This then gives a second strength of kNN in that it can capture the transition phase between congestion and free flow states and pictures the traffic fluctuation in more detail. Third, its process is convenient to manipulate the impact of neighboring detectors by mechanisms, for example weighting the similarity by correlation coefficient. Fourth, this non-parametric kNN method avoids over-fitting of historical data into one fit-it-all model and treasures the uniqueness of each individual historical data that can be used for imputing missing data of various transportation phenomena.

The main procedures of the proposed sectional kNN algorithm are based on four steps. At first, we divide road network into several sections based on a trend of correlation between detectors. Second, we locate the missing values in both historical and subject data in temporal and spatial span to generate a health matrix of each road section that visualizes the availability of the values. Third, we calculate the distance between the historical data and subject data to assess the similarity between them, by using weighting matrix from data availability and correlation coefficient between detectors. Fourth, algorithm generates the imputing values for the missing data by

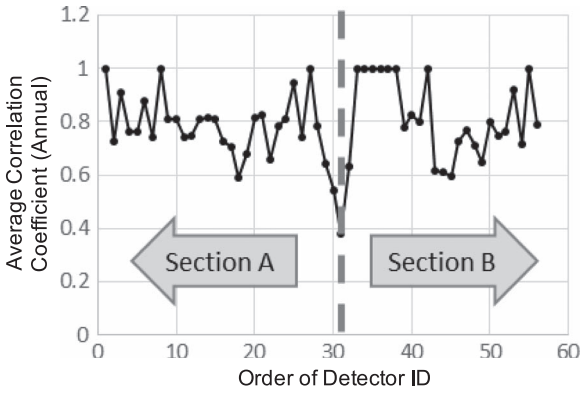

Fig. 2. Example of annual average correlation coefficient near junction points.

integrating the k nearest neighbors with highest weighted similarity to subject data. Following is the details of proposed sectional kNN algorithm.

A. Feature Matrix for Road Section

In the previous researches, kNN based algorithm finds the most similar data by matching each individual link with historical links for each loop detector. The feature matrix generally has contained the data of neighboring detectors as follows:

$$X_t^{(s)} = \begin{bmatrix} Z_{t-\tau}^{(s-l)} & \cdots & Z_{t-\tau}^{(s+l)} \\ \vdots & \ddots & \vdots \\ Z_{t+\tau}^{(s-l)} & \cdots & Z_{t+\tau}^{(s+l)} \end{bmatrix} \quad (1)$$

where Z_t^s is the observed value at $t = \{1, 2, \ldots, T\}$ time intervals and $s = \{1, 2, \ldots, S\}$ spatial locations, l represents the spatial range of the neighboring detectors in the kNN search, and τ represents the temporal range of neighboring time intervals.

This strategy of searching individual links requires a high computation power because the number of matching process increases as the number of missing detectors increase. For example, for imputing missing data in ten different detectors, the matching process is implemented by ten times. In Korean highway, there are historical data of approximately 1750 DSRC links and 7227 VDS (vehicle detection system) links over 472 observing days. With this large database, it is not wise to search the k nearest neighbors for each links by matching all the links respectively.

To improve the computation performance for matching within reasonable accuracy, we divide the road sections by considering the correlation coefficient with a moving window between detectors. Refer to Fig. 2, where x-axis represents the sequential order of detectors and y-axis represents annual average correlation coefficient between neighboring detectors. The correlation coefficient drops near a junction, which is located at 31th detector. After this detector, the correlation coefficient recovers again. The trend of correlation coefficient is repeated similarly near junction points. This drop of correlation coefficients with a moving window defines the length of a road

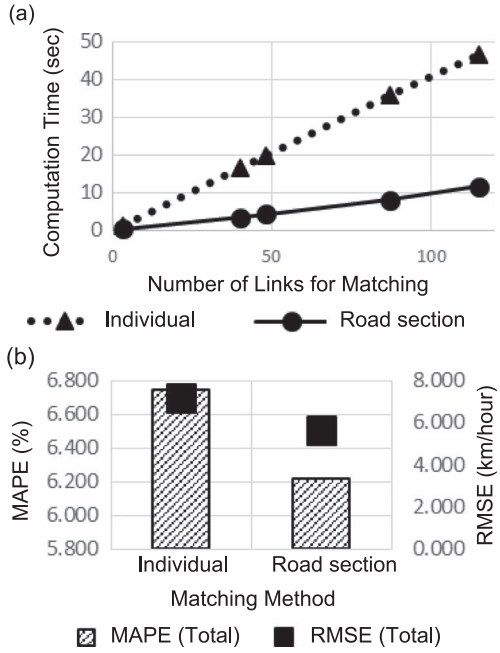

Fig. 3. Comparison analysis of two different matching strategy. (a) Computation time for matching. (b) Accuracy of matching.

section to be imputed. A feature vector of each section as a matrix is exemplified as follows:

$$X_t^{(r)} = \begin{bmatrix} Z_{t-\tau}^{(1)} & \cdots & Z_{t-\tau}^{(S^{(r)})} \\ \vdots & \ddots & \vdots \\ Z_{t+\tau}^{(1)} & \cdots & Z_{t+\tau}^{(S^{(r)})} \end{bmatrix} \quad (2)$$

where $Z_t^{(s)}$ is the observed value at $t = \{1, 2, \ldots, T\}$ time and $s = \{1, 2, \ldots, S^{(r)}\}$ detectors in $r = 1, 2, \ldots, R$ road sections. $S^{(r)}$ is the number of detectors in $r = 1, 2, \ldots, R$ road sections, and τ is represents the temporal range of neighboring time intervals in the kNN search.

While previous studies mostly impute missing data of each individual detector, this proposed feature matrix searches for the nearest patterns of each section with multiple detectors at once. As shown in Fig. 3, this method of imputing multiple missing data not only reduces the computation time, for instance by 75%, but also reduces error, for instance by 8% and 22% reduction of MAPE and RMSE, respectively. This increase in accuracy may be due to the removal of unrelated neighboring detectors. In other words, a possible bias from neighboring detectors with low correlation has been removed in the process of section definition. Since the historical data is confined to a spatially correlated set of detectors, the accuracy could naturally increase. Also, a section shares useful information such as congestion or free-flow state, and the imputation algorithm can efficiently capture this property from historical data to increase accuracy. Also, to impute n-number of detectors in a section, the kNN feature vector includes $n + 2$ number of detectors, whose detectors at two ends will be overlapping with other neighboring sections. These two detectors at the ends will not be imputed but only used for the historical data searching in the kNN process. This prevents the possible degradation

Fig. 4. Calculation method of Euclidean distance with incomplete historical data.

of imputation performance when imputing missing data of detectors at the end of original section, since it would have a removed data from its next detector.

B. Generation of Health Vector With Data Availability

A health vector represents data availability for both historical and subject data by considering missing and corrupted data as NA. This second step of health vector building based on data availability is a necessary procedure in imputation of multiple missing data in a section, because it is very difficult to find a complete historical data for the entire section. The usefulness of health vector is twofold. One, it is used to give weights to the historical data depending on their data availability. Two, it is used to efficiently calculate the Euclidean distance between historical and subject data, as shown in more detail later. A health vector is in the same dimension as a feature matrix, exemplified as follows:

$$H_t^{(r)} = \begin{bmatrix} h_{t-\tau}^{(1)} & \cdots & h_{t-\tau}^{(S^{(r)})} \\ \vdots & \ddots & \vdots \\ h_{t+\tau}^{(1)} & \cdots & h_{t+\tau}^{(S^{(r)})} \end{bmatrix} \tag{3}$$

where $h_\tau^{(s)}$ denotes the availability of detector data of link s at time interval t in the section r, τ is the temporal range of a section, and $S^{(r)}$ is the spatial range of a section r.

C. Distance Metric Between Historical and Subject Data

A distance metric is most commonly used for calculating similarity between historical data to a subject data in order to find the k nearest neighbors [3], [5], [22], [23]. There are various types of distance metrics, such as Euclidean, Unit Map, Quadratic Form, Standardized Euclidean, Mahalanobis, City block, and Qi and Smith distance metric. Among them, Euclidean distance is the most common to calculate the geometric distance in multi-dimensional space, formulated as follows [22]:

$$d(X_{\text{subject},r}, X_{\text{history},r}) = \sqrt{\sum_{t=1}^{T}\sum_{s=1}^{S}\left(V_{t,s}^{\text{subject}} - V_{t,s}^{\text{history}}\right)^2} \tag{4}$$

where, $V_{t,s}^{\text{subject}}$ is the speed of the subject data at time t in detector s and $V_{t,s}^{\text{history}}$ is the speed of the historical data at time t in detector s.

However, a conventional Euclidean distance can be used only with complete historical data, which does not happen in real life. Also, simply omitting the missing historical data could be dangerous in calculating the distance, since it can choose historical data with large missing data and only a few data similar with subject data as the best candidate for imputation because it will have the least distance to the subject data.

To solve this problem, we modify the conventional Euclidean distance shown in Fig. 4, using the health vector. When there exist missing data both in historical data and subject data in the various patterns like the three cases, we transform each historical data and subject data to only use the values available in both historical and subject data. This is done by multiplying them with health vectors of both historical and subject data. This "punches out" the historical and subject data for which any one of two have it missing.

Next, the distance between historical and subject data is given weight depending on the data availability, e.g., more weight to data with more available historical data. This reduces a chance of historical data with large missing data to be selected as a candidate for imputation. Refer to the Fig. 4 again. Because of the weight, the algorithm may recommend case 2 though case 3 has a closer pattern to the subject data. The modified Euclidean distance is formulated as follows:

$$d(X_{\text{subject}}, r, X_{\text{history,r}}) = \frac{\sqrt{\sum_{t=1}^{T}\sum_{s=1}^{S}\left(h_{t,s}^{\text{Hybrid},i}\cdot Z_{t,s}^{\text{subject}} - h_{t,s}^{\text{Hybrid},i}\cdot Z_{t,s}^{\text{history}}\right)^2}}{a^2} \tag{5}$$

$$H_{t,s}^{\text{Hybrid},i} = \begin{bmatrix} h_{t-\tau,1}^{\text{Hybrid},i} & \cdots & h_{t-\tau,S^{(r)}}^{\text{Hybrid},i} \\ \vdots & \ddots & \vdots \\ h_{t+\tau,1}^{\text{Hybrid},i} & \cdots & h_{t+\tau,S^{(r)}}^{\text{Hybrid},i} \end{bmatrix} \tag{6}$$

where a is the weight depending on the number of values used for distance calculation, $h_{t+\tau,S^{(r)}}^{\text{Hybrid},i}$ denotes the availability of detector data in both historical data and subject data of link s at time interval t in the section r, τ is the temporal range of a section, and $S^{(r)}$ is the spatial range of a section r.

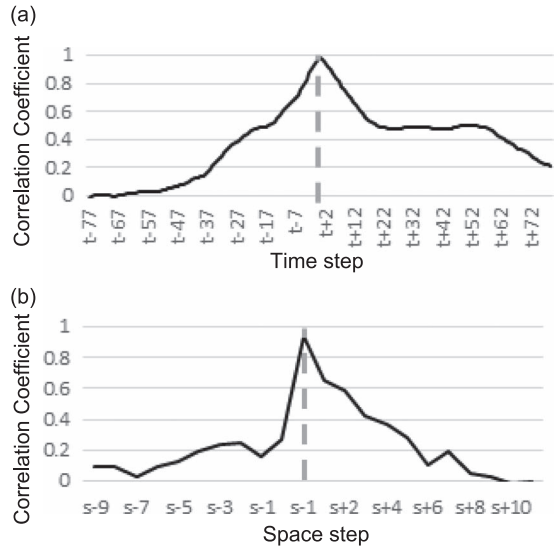

Fig. 5. Temporal and spatial correlation coefficient near a subject detector at subject time interval. (a) Temporal correlation. (b) Spatial correlation.

In addition to the weight from data availability, the spatial closeness of data near missing values are considered to give more weight, since it may be closely related to the missing values. The conventional Euclidean Distance gives equal weight to all distance, for example, speed difference at the first nearest detector and at the tenth nearest detector are calculated with same weight to calculate the distance with the historical data. However, the traffic condition of a subject link is shown to be significantly affected by the traffic condition of the upstream and downstream links, especially when congestion is generated or dissipated in the subject or adjacent links [3], [4], [7], [24]. Fig. 5 shows this phenomenon, with correlation coefficient trend with respect to the distance in time interval and in detector location. In both Fig. 5(a) and (b), first nearest time step or detector location show the highest correlation coefficient. As distance from the subject detector increases, the correlation coefficient gradually decreases.

To systematically represent the spatial and temporal dependencies to the subject detector location and subject time in distance calculation, we propose a weight matrix for the distance metric, formulated as follows:

$$C_i^{\text{time}} = \left[\rho_{(1,i)}\ \rho_{(2,i)}\ ,\ldots,\rho_{(1-1,i)}\ \rho_{(i,i)}\ \rho_{(1+1,i)}\ ,\ldots,\ \rho_{(T-1,i)}\ \rho_{T,i}\right]^T \tag{7}$$

$$C_j^{\text{space}} = \left[\rho_{1,j}\ \rho_{2,j}\ ,\ldots,\ \rho_{j-1,j}\ \rho_{(j,j)}\ \rho_{(j+1,j)}\ ,\ldots,\ \rho_{S-1,j}\ \rho_{(S,j)}\right] \tag{8}$$

where, $\rho_{(T,i)}$ is a temporal correlation coefficient between values at time T and values at time i, and $\rho_{(s,j)}$ is a spatial correlation coefficient between values at location S and values at location j. C_i^{time} is a vector for temporal correlation and C_j^{space} is a vector for spatial correlation

$$C^h = C_i^{\text{time}} C_j^{\text{space}} = \begin{bmatrix} \rho_{(1,i)}\rho_{(1,j)} & \cdots & \rho_{(1,i)}\rho_{(S,j)} \\ \vdots & \ddots & \vdots \\ \rho_{(T,i)}\rho_{(1,j)} & \cdots & \rho_{(T,i)}\rho_{(S,j)} \end{bmatrix} \quad h \in \{(i,j)\}, i \in I, j \in J \tag{9}$$

where h is a pairs of variables (i,j), i is a temporal location of the missing value in time space i, and j is spatial location of the missing value in the link location space J. By using the correlation matrices generated from k number of missing data, the weighting matrix is formulated as follows:

$$W = \begin{bmatrix} F\left(C^h_{1\,1}\right) & \cdots & F\left(C^h_{1\,S}\right) \\ \vdots & \ddots & \vdots \\ F\left(C^h_{T\,1}\right) & \cdots & F\left(C^h_{T\,S}\right) \end{bmatrix} \tag{10}$$

$$F\left(C^h_{m\,n}\right) = \max\left(C^1_{m\,n}, C^2_{m\,n}, C^3_{m\,n}, \ldots, C^h_{m\,n}\right) \tag{11}$$

where W is an $m \times n$ matrix, m is the temporal length of feature matrix and n is the spatial length of the feature matrix.

Based on this weighting matrix, the weighted Euclidean distance is formulated as follows:

$$d(X_{\text{subject},r}, X_{\text{history},r}) = \frac{\sqrt{\sum_{t=1}^{T}\sum_{s=1}^{S} W_{t,s} * \left(h_{t,s}^{\text{Hybrid},i} \cdot Z_{t,s}^{\text{subject}} - h_{t,s}^{\text{Hybrid},i} \cdot Z_{t,s}^{\text{history}}\right)^2}}{a^2} \tag{12}$$

where, $W_{t,s}$ is the weight for s-th value at time t.

D. Generation of Missing Data

The Local Estimation Method (LEM) calculates the imputation value for missing data from the nearest neighbors found with the distance metric and the k value [5], [24], [26]. Commonly the arithmetic mean of output $y(t)$ of the k nearest neighbors is used. However, this simple average method disregards the useful information, such as ranking of the historical data by distance and the Euclidean distance value [25]. To overcome this drawback, the weighted average method is used in this study, as shown in Equation (13). It uses the inverse of distance between historical and subject data as a weight

$$V^{\text{imputation}} = \frac{\sum_{k=1}^{K} \frac{V_k^{\text{history}}}{d_k}}{\sum_{k=1}^{K} \frac{1}{d_k}} \tag{13}$$

where d_i is the Euclidean Distance of the k-th neighbor, $V^{\text{imputation}}$ is the imputed speed, and K is the number of nearest neighbor.

III. Performance Evaluation

We evaluate the proposed imputation method by comparing to two other common imputation methods, the Bootstrap-based Expectation Maximization (B-EM) and Nearest Historical Average (NH). Various missing data type, missing ratio, day type and traffic states are observed in performance comparison between the three imputation methods by MAPE, RMSE, and PCV.

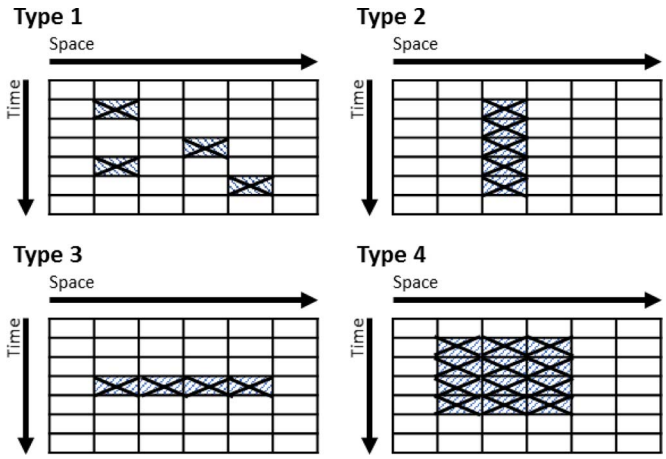

Fig. 6. Four types of missing data.

A. Generation of Missing Data

The missing data are intentionally generated to evaluate the performance of three methods with different missing data type, missing ratio, day type and traffic states because performance of imputation methods may significantly vary depending on these settings.

First, the missing ratio generated in this study ranged from 0.1% to 50%, to test if the imputation method is stable across various missing ratio [12], [26]–[29]. Second, various types of error are possible in data obtained from detectors, to test robustness of imputation method with reasonable accuracy. Refer to Fig. 6. Type 1 missing data is analogous to missing completely at random (MCAR) and may occur due to temporary power or communication failures [30]. Type 2 and 3 missing data are not random and may occur due to a prolonged physical damage, malfunction of communication device, and maintenance backlogs. Type 4 missing data may be due to the measurement noise, bias, and system failure over a large domain. Though type 2, 3, and 4 seem extreme in configuration, they are commonly observed in practice.

In addition, the traffic state is considered when evaluating the methods. Generally in the transportation data, missing values during free flow are more easily imputed with high accuracy than congestion and transition state. Missing values during transition state is more difficult to estimate than other states since it requires delicate estimation on which point the speed drops or recovers. In this paper, free flow is defined between 80 to 100 km/hr, transition state between 40 to 80 km/hr, and congestion below 40 km/hr. Also, day type must be mentioned in evaluation of imputation methods, since these data show significantly different trends in traffic demand, travel time, and speed [5], [33].

B. Imputation Techniques for Comparison Analysis

The first comparison method is NH, which fills a missing data with average historical data collected on the same detector at the same time but from a neighboring day [30]. This is the most common method because it shows a stable performance regardless of the missing data size with easy implementation. Therefore, many researchers use this method as a benchmark to compare imputation performance [1], [13], [23], [28]–[30], [32], [33].

NH technique is inherently based on the assumption that traffic pattern at the same detector at the same time is similar from day to day. The flexibility of this algorithm comes from the selection and integration of values taken over the recent historical days. In other words, we must control the size of nearest historical data, the day type of nearest historical data, e.g., weekend or weekday, and how you integrate them to calculate the final imputing value, e.g., weighted average or arithmetic average. In this study we categorize the day type into weekday and weekend data. The arithmetic average speed of the same time and same day of week over 20 historical days was used to impute the missing value.

The second comparison method is Bootstrap-based Expectation Maximization (B-EM), which is based on bootstrap sampling method and Expectation Maximization (EM) algorithm. Bootstrap sampling method is used to estimate the sample distribution of statistics and EM method is a popular tool in statistical missing data imputation in the various fields [13], [34]–[37]. In the B-EM method, the bootstrap sample is produced as random sample. Then using the EM algorithm, the missing data are regressed by computing the maximum likelihood estimate (MLE) in the presence of missing data. The procedure for the EM algorithm is as follows. First, calculate the sample means and covariance by using the sample data set with no missing value. For example, in type 4 of Fig. 6, sample means and covariance for 1st column, 5th column, and 6th column are calculated. Second, the mean is estimated to be the missing point, tentatively. Third, the maximum likelihood estimate and replace the missing values with new estimate. With initial inserted data, expecting likelihood using the initial estimate for the parameters is calculated and the likelihood is maximized by adjusting the estimate for the parameters. When the likelihood is maximized, the missing value is replaced with estimated value using parameters, which is maximizing likelihood. Fourth, estimate means and covariance matrix based on newly inserted estimates. Finally, repeat step one to four until means and covariance matrix converge [38]. The algorithm has reasonable accuracy in missing data imputation.

C. Quantitative Measures for Evaluation

To evaluate the performance of the imputation algorithms, error measures and change in natural variance were calculated. In this study, three measures were used—Mean Absolute Percent Error (MAPE), Root Mean Squared Error (RMSE), and Percent Change in Variance (PCV), formulated as follows:

$$\text{MAPE} = \frac{\sum \left| \frac{e_{i,t}}{V_{i,t}} \right|}{n} \times 100\% \tag{14}$$

$$\text{RMSE} = \sqrt{\frac{\sum e_{i,t}^2}{n}} \tag{15}$$

$$\text{PCV} = \frac{\text{var}(\hat{V}) - \text{var}(V)}{\text{var}(V)} \times 100\% \tag{16}$$

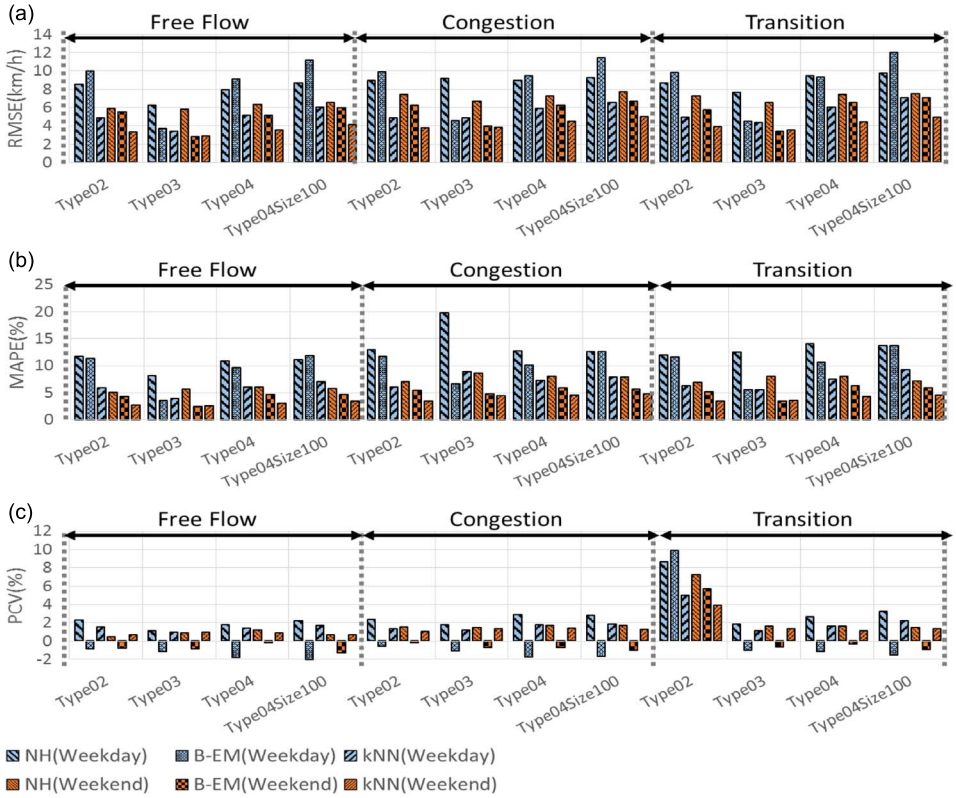

Fig. 7. (a) RMSE, (b) MAPE, and (c) PCV for three different imputation algorithms with varying missing types and traffic states during weekday and weekend. (a) RMSE. (b) MAPE. (c) PCV.

where $e_{i,j,t}$ is the error for detector i at time t, $V_{i,j,t}$ is the actual value for detector i at time t, $\hat{V}_{i,j,t}$ is the imputed value for detector i at time t, and n is the number of imputed values.

MAPE and RMSE are two of the most widely used measures to evaluate the accuracy of imputation techniques [4], [13], [29], [33], [39]. However, MAPE and RMSE cannot represent whether or not imputation techniques well maintain the detailed description of traffic pattern from historical data. For example, interpolation imputation or moving average imputation techniques with reasonable MAPE and RMSE values smoothen the transportation data excessively. This may lose information like sudden speed drop, which is a crucial phenomenon to estimate for an ITS service, such as travel time prediction. Therefore, PCV is also evaluated to observe this undesirable effect of each imputation technique, which measures the change in natural variance of the imputed data by calculating the difference between the variance in the actual data and the imputed data.

IV. Data

Speed data in this study was collected by the Dedicated Short Range Communications (DSRC) detectors on major highways in Korea over 472 days. The Korea Expressway Corporation currently employs approximately 1750 DSRC links on the highway system with a purpose of measuring link travel time, each detector observing a span of approximately 2 km [40]. These detectors provide link travel time of 5-minute interval and count data. Since these detectors are not complete survey but only sample data, the count data are discarded and only the link travel time has been assessed in this study. The highway network was divided into 21 sections by correlation in a moving window.

V. Results

The performance of the proposed kNN was evaluated with 472 days of historical data and a k value of 3, which is the number of selected nearest historical data as a result of kNN algorithm. Missing data in testing were intentionally produced from complete dataset and compared with the actual value for performance evaluation. We compared the proposed kNN to Nearest History (NH) and Bootstrap-based Expectation Maximization (B-EM), in terms of Mean Absolute Percent Error (MAPE), Root Mean Squared Error (RMSE), and Percent Change in Variance (PCV) for missing data with differing missing types, missing ratio, traffic state, and day type.

Fig. 7(a) and (b) show RMSE and MAPE of three imputation methods for different missing types, traffic states, and day types. First, observe RMSE and MAPE of the weekday imputation. The accuracy results of imputation represented by RMSE and MAPE show that in almost all traffic states, i.e., free flow, congestion, transition, and in almost all data types, i.e., type 2, 3, and 4, kNN outperforms NH and B-EM with lower RMSE and MAPE with relatively small variation. However, the B-EM method performs well in terms of MAPE in missing type 3 compared to other types of missing pattern, though it shows poor performance with type 2 and 4. Second, observe the weekend imputation. The accuracy results are similar in that

kNN method shows lowest MAPE and RMSE with small variance in all traffic states and day types, except B-EM performs similarly or better than kNN with type 3. We can also observe that NH performance drops for weekend imputation and that all imputation methods perform differently with different missing types.

The effect of the traffic state on imputation performance of three method is not as prominent as missing data. Regardless of the traffic state, the proposed kNN method shows the lowest RMSE. The notable finding is that the RMSE during the free flow state is smaller than the congestion or transition states in the all three imputation methods.

Though performing well with type 3, poor performance of B-EM with type 2 and 4 show that B-EM method may only be appropriate to impute missing values from only one detector rather than multiple neighboring detectors simultaneously. Also unlike the kNN method, NH method cannot impute missing data of weekday and weekend data equally well. The congestion pattern during weekdays is quite recurrent and repetitive during peak hours, but during weekends it is not very repetitive and may change by month, season and weather. Therefore, the NH method with average historical data cannot represent behavior of weekend and performs poorly.

From comparing the MAPE and RMSE of three imputation methods with different missing types and day types, we find that a) kNN method outperforms in almost all categories, b) performance of each imputation methods depend on the missing types, c) traffic state does not influence the performance of imputation for all three methods, d) B-EM method performs well with missing type 3 but poorly with other types, and e) NH performance better with weekday data than weekend data.

Refer to Fig. 7(c), with PCV for three imputation methods for varying missing types, traffic states, and day types. The PCV values of three imputation methods show quite different trends compared to RMSE and MAPE. The PVC value close to zero means that the imputation method has well conserved the detailed pattern of the historical values. As shown, three methods show consistent performance with varying missing types, though the PCV for each method is comparable. On one hand, the imputed values of B-EM method fluctuate more than the actual values and this increases the PCV value. On the other hand, both NH and the proposed kNN method have smaller PCV because these two methods in essence average the historical data for imputation. The averaging effect makes the variance of the imputed values decrease.

To summarize, results show that kNN method performs better than B-EM and NH methods in almost all missing types, traffic states, and day types, with exception for missing type 3 for which B-EM shows a better performance. When the missing type and day type can be identified before imputation, the proposed kNN method is appropriate for missing type 2 and 4, where the B-EM method is appropriate for missing type 3. However, in practice, the identification of missing type is difficult because various missing types may occur at the same time. Therefore, when the missing types is uncertain or mixed, the proposed kNN method have a strong advantage for its robustness and accuracy.

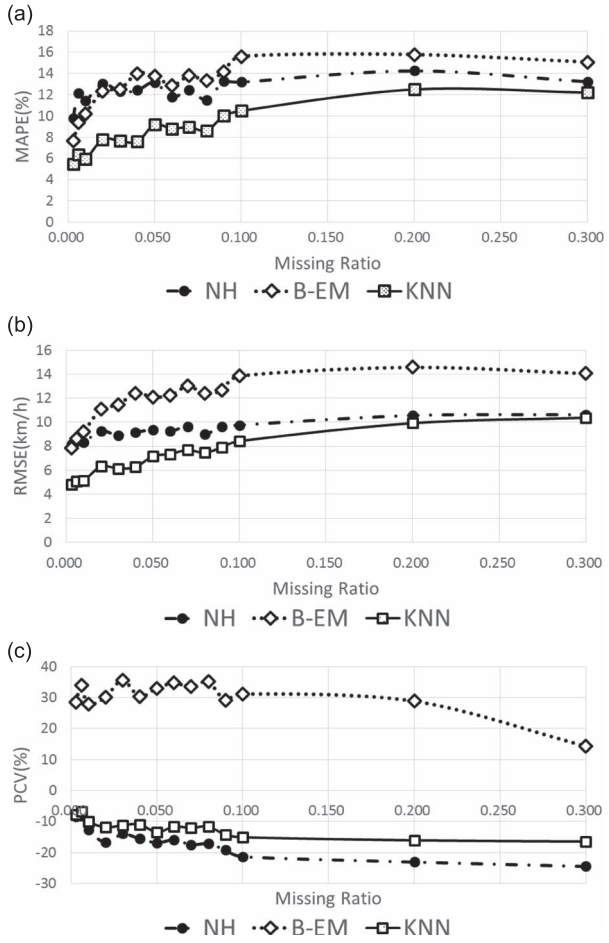

Fig. 8. (a) MAPE, (b) RMSE, (c) PCV for three imputation algorithms with varying missing ratios. (a) MAPE and missing ratio. (b) RMSE and missing ratio. (c) PCV and missing ratio.

Next, observe the impact of performance with varying missing ratios of weekday data as shown with MAPE, RMSE, and PCV of the three algorithms. Refer to Fig. 8(a) and (b). Both MAPE and RMSE of the proposed kNN method are much smaller than the other two methods. For instance, for missing ratio higher than 3%, MAPE of NH is around 33% and B-EM around 48%, both higher than that of kNN method. As shown in Fig. 8(a), the MAPE values of kNN method and B-EM method both show an increasing trend as the missing ratio increases. However, the MAPE of NH method is relatively constant because its calculation mechanism does not depend on neighboring data. The RMSE curve of three imputation method also show the similar results to the MAPE curve as shown in Fig. 8(b). The RMSE of NH and proposed kNN method are similar when the missing ratio is 30%. Note that the NH, B-EM and proposed kNN methods have computing speed of 0.001 second/missing point, 0.013 second/missing point, and 0.158 second/missing point, respectively.

The PCV values of three imputation methods for weekday data are quite different compared to MAPE and RSME as shown in Fig. 8(c). The PCV values of proposed kNN method are most constant and closest to zero than other methods, conserving the detailed pattern of the historical data the most.

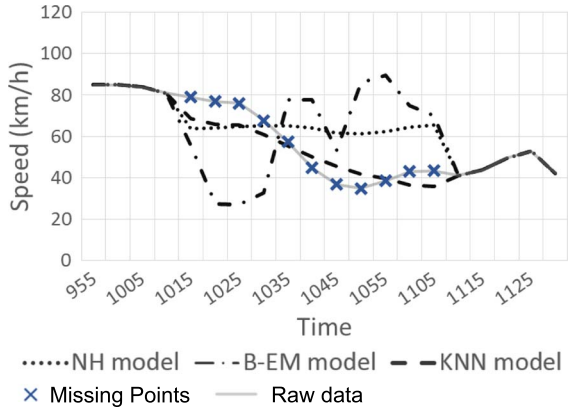

Fig. 9. Example of imputation performance of NH, B-EM, and kNN algorithms.

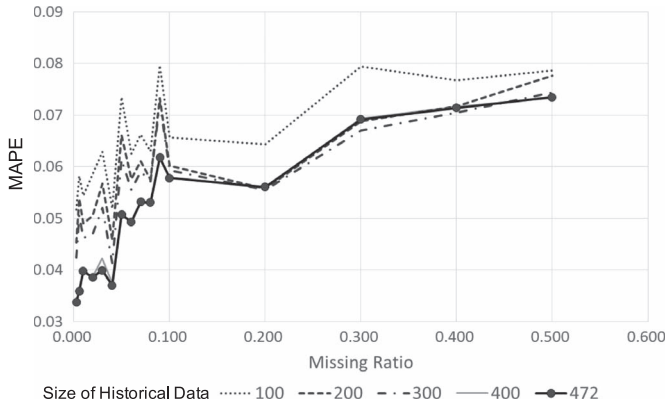

Fig. 10. RMSE for kNN algorithm with different historical data sizes at varying missing ratios.

The PCV of NH method is also relatively stable but its absolute value is larger than that of kNN method because of the smoothing of data. The PCV of B-EM method fluctuates largely with varying missing ratios and most far from zero. The B-EM model makes noise in the imputed values and this leads to the performance degradation.

To summarize, kNN shows better performance to NH and B-EM in terms of MAPE, RMSE, and PCV with all missing ratios. NH shows the higher performance than B-EM method when missing ratio is higher than 3% due to its stable performance. When the missing ratio is small, the accuracy of NH method is slightly lower than that of B-EM method. As the missing ratio increases, NH maintains the small fluctuation of accuracy and B-EM shows the sharply increasing trend of accuracy. Eventually, the accuracy of NH method is higher than that of B-EM method as the missing ratio increases.

Fig. 9 shows an example case of missing data imputation for three methods that was frequently observed in testing. This is a case of missing data at a transition state from free flow to non-recurrent congestion with sudden drop of speed. The proposed kNN method follows well the decreasing trend of the actual values in general. In comparison, the imputed values with B-EM show a large fluctuation. Because the traffic state is changing very slowly from free flow to congestion, the B-EM method cannot estimate the distribution during the transition phase or draw it smoothly. This not only reduces the imputation performance of B-EM but also creates noises in the imputed data. Low performance of B-EM method in a transition state is also shown in the Fig. 7(a) and (b). Also NH is not accurate for this non-recurrent case due to lack of occurrence of similar pattern in nearest historical data. The proposed kNN algorithm shows better imputation performance than both of NH and B-EM model.

Lastly, to guarantee the performance of proposed kNN method in practice, the appropriate historical data size should be investigated. Fig. 10 shows RMSE of the proposed kNN algorithm with varying historical data size. The 200 cases of missing data are randomly generated in different traffic states and different day types. Regardless of data size, RMSE is very low with missing ratio smaller than 10%. However, RMSE fluctuates largely in this region for data size smaller than 400. Therefore, minimum of 400 historical data for this proposed algorithm is required to assure the imputation quality.

VI. Conclusion and Future Work

In summary, this paper proposes a data-driven imputation method for sections of road based on their spatial and temporal correlation using a modified k-Nearest Neighbors (kNN) method. This method is different to the conventional algorithms in that 1) it defines a section of road links that share a similar traffic property by analyzing the correlation of neighboring links in a moving window, 2) it imputes missing data of multiple detectors in sectional units of road links and 3) tailor the kNN methodology to cope with practical challenges of transportation data, such as large missing ratio, incomplete historical data, different traffic states and day types, and to capture the traffic pattern in detail for missing data imputation.

The performance of the proposed kNN method for sectional imputation of road links was compared to two other commonly used imputation methods—Expectation Maximization and Nearly Historical Average, in terms of missing data type, day type, traffic states, and missing ratio. The proposed kNN method outperformed in almost all categories of missing scenario and provided a robust service with high accuracy with a reasonable historical data size of 400 days. Especially in practice, missing data type is unknown and day types of missing data are various and this makes the proposed kNN method a better choice than B-EM or NH. Also as the traffic data accumulates to form Big Data, the kNN method promises a successful performance together with distributed computing.

To improve this algorithm even further, it will be beneficial to study the performance of imputation methods with respect to unusual traffic patterns, i.e., accidents, which have uncertainty of occurrence time and severity of event. Also, it will be interesting to study how to apply this algorithm to arterial road network, since defining a section of spatially correlated road links will be different to non-linear arterial networks.

REFERENCES

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

[2] H. Chen, S. Grant-Muller, L. Mussone, and F. Montgomery, "A study of hybrid neural network approaches and the effects of missing data on traffic forecasting," *Neural Comput. Appl.*, vol. 10, no. 3, pp. 277–286, Dec. 2001.

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

[4] L. Li, Y. Li, and Z. Li, "Efficient missing data imputing for traffic flow by considering temporal and spatial dependence," *Transp. Res. C, Emerging Technol.*, vol. 34, pp. 108–120, Sep. 2013.

[5] Z. Liu, S. Sharma, and S. Datla, "Imputation of missing traffic data during holiday periods," *Transp. Plann. Technol.*, vol. 31, no. 5, pp. 525–544, Oct. 2008.

[6] E. Redfern, S. Watson, S. Clark, M. Tight, and G. Payne, "Modelling outliers and missing values in traffic count data using the ARIMA model," *Institute for Transport Studies Working Paper*, vol. 395. Leeds, U.K.: IEEE, 1993.

[7] H. Tan *et al.*, "A tensor-based method for missing traffic data completion," *Transp. Res. C, Emerging Technol.*, vol. 28, pp. 15–27, Mar. 2013.

[8] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," *Transp. Res. C, Emerging Technol.*, vol. 19, no. 3, pp. 387–399, Jun. 2011.

[9] J. W. C. van Lint, S. P. Hoogendoorn, H. J. van Zuylen, and J. Van Lint, "Accurate freeway travel time prediction with state-space neural networks under missing data," *Transp. Res. C, Emerging Technol.*, vol. 13, no. 5/6, pp. 347–369, Oct.–Dec. 2005.

[10] M. Zhong, S. Sharma, A. Dean, P. Lingras, and C. Science, "Genetically designed models for accurate imputation of missing traffic counts," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1879, no. 1, pp. 71–79, 2004.

[11] H. Tan, Y. Wu, B. Cheng, W. Wang, and B. Ran, "Robust missing traffic flow imputation considering nonnegativity and road capacity," *Math. Probl. Eng.*, vol. 2014, 2014, Art. ID 763469.

[12] D. Ni, J. J. D. Leonard, A. Guin, and C. Feng, "Multiple imputation scheme for overcoming the missing values and variability issues in ITS data," *J. Transp. Eng.*, vol. 131, no. 12, pp. 931–938, 2005.

[13] B. B. Smith, W. W. Scherer, and J. J. Conklin, "Exploring imputation techniques for missing data in transportation management systems," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1836, no. 1, pp. 132–142, 2003.

[14] Urban Crossroads, "PeMS Data extraction methodology and execution technical memorandum," Southern California Assoc. Gov., Irvine, CA, USA, 2006.

[15] X. Huang and Q. Zhu, "A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets," *Pattern Recognit. Lett.*, vol. 23, no. 13, pp. 1613–1622, Nov. 2002.

[16] D. Ni and J. D. Leonard, II, "Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1935, pp. 57–67, 2005.

[17] C. Chen, J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya, "Detecting errors and imputing missing data for single-loop surveillance systems," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1855, no. 1, pp. 160–167, 2003.

[18] K. Henrickson, Y. Zou, and Y. Wang, "Flexible and robust method for missing loop detector data imputation," in *Transp. Res. Board 94th Annu. Meet.*, 2015, pp. 29–36.

[19] M. Asif, N. Mitrovic, and L. Garg, "Low-dimensional models for missing data imputation in road networks," in *Proc. IEEE ICASSP*, 2013, pp. 3527–3531.

[20] X. G. X. Gong and F. W. F. Wang, "Three improvements on KNN-NPR for traffic flow forecasting," in *Proc. IEEE Intell. Transp. Syst.*, 2002, pp. 736–740.

[21] Y. Li, Z. Li, and L. Li, "Missing traffic data: Comparison of imputation methods," *IET Intell. Transp. Syst.*, vol. 8, no. 1, pp. 51–57, Feb. 2014.

[22] S. S. Robinson, "The development and application of an urban link travel time model using data derived from inductive loop detectors," M.S. thesis, Dept. Civil Environ. Eng., Imperial College London, London, U.K., 2006.

[23] S. Wu, Z. Yang, X. Zhu, and B. Yu, "Improved k-NN for short-term traffic forecasting using temporal and spatial information," *J. Transp. Eng.*, vol. 140, no. 7, Jul. 2014, Art. ID 04014026.

[24] M. Whitlock and C. Queen, "Modelling a traffic network with missing data," *J. Forecast.*, vol. 19, no. 7, pp. 561–574, Dec. 2000.

[25] B. Smith, B. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transp. Res. C, Emerging Technol.*, vol. 10, no. 4, pp. 303–321, Aug. 2002.

[26] Y. Zhang and Y. Liu, "Missing traffic flow data prediction using least squares support vector machines in urban arterial streets," in *Proc. IEEE CIDM*, 2009, pp. 76–83.

[27] Y. Zhang and Y. Liu, "Data imputation using least squares support vector machines in urban arterial streets," *IEEE Signal Process. Lett.*, vol. 16, no. 5, pp. 414–417, May 2009.

[28] J. Wang, N. Zou, and G. Chang, "Empirical analysis of missing data issues for ATIS applications: Travel time prediction," in *Proc. Transp. Res. Board 87th Annu. Meet.*, 2008, pp. 81–91.

[29] M. Chen, J. Xia, and R. Liu, "Developing a strategy for imputing missing traffic volume data," *J. Transp. Res. Forum*, vol. 45, no. 3, pp. 57–75, 2010.

[30] J. H. J. Conklin and W. W. T. Scherer, "Data imputation strategies for transportation management systems," Univ. Virginia, Charlottesville, VA, USA, UVACTS-13-0-80, 2003.

[31] J. Jun, "Understanding the variability of speed distributions under mixed traffic conditions caused by holiday traffic," *Transp. Res. C, Emerging Technol.*, vol. 18, no. 4, pp. 599–610, Aug. 2010.

[32] M. Zhong and S. Sharma, "Development of improved models for imputing missing traffic counts," *Open Transp. J.*, vol. 3, no. 1, pp. 35–45, Mar. 2009.

[33] F. Castrillon, A. Guin, R. Guensler, and J. Laval, "Comparison of Modeling Approaches for Imputation of Video Detection Data in Intelligent Transportation Systems," vol. 2308, pp. 138–147, 2012.

[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[35] X. Wu, J. Fan, and K. R. Subramanian, "B-EM: A classifier incorporating bootstrap with EM approach for data mining," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2002, pp. 670–675.

[36] T. Delleji, M. Zribi, and A. Hamida, "On the EM algorithm and bootstrap approach combination for improving satellite image fusion," *Int. J. Signal Process.*, vol. 4, no. 1, pp. 3796–3805, 2007.

[37] J. Honaker, G. King, and M. Blackwell, "Amelia II: A program for missing data," *J. Stat. Softw.*, vol. 45, no. 7, pp. 1–47, Dec. 2011.

[38] S. Borman, "The expectation maximization algorithma short tutorial," 2004. [Online]. Available: http://www.seanborman.com/publications/EM_algorithm.pdf

[39] Y. Li, Z. Li, L. Li, Y. Zhang, and M. Jin, "Comparison on PPCA, KPPCA and MPPCA based missing data imputing for traffic flow," in *Proc. 2nd Int. Conf. Transp. Inf. Safety, China*, 2013, pp. 1151–1156.

[40] OpenOASIS, Korea Expressway Corporation. [Online]. Available: http://data.ex.co.kr/

Sehyun Tak was born in Seoul, Korea, in 1982. He received the M.S. and Ph.D. degrees in civil and environmental engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2011 and 2015, respectively. Currently, he is a Postdoctoral Researcher with KAIST.

Soomin Woo received the bachelor's degree in civil and environmental engineering in 2014 from Korea Advanced Institute of Science and Technology, Daejeon, Korea, where she is currently working toward the master's degree. Her current research interests include prediction and analysis of transportation network.

Hwasoo Yeo received the Ph.D. degree in civil and environmental engineering from the University of California, Berkeley, CA, USA, in 2008. He is currently an Associate Professor with the Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea. His research interests include theoretical and simulation studies on traffic flow and traffic operations, traffic safety, and intelligent transportation systems.