International Journal of Geo-Information

MDPI

Article

Analysis of Spatiotemporal Data Imputation Methods for Traffic Flow Data in Urban Networks

Endra Joelianto [1,2,*], Muhammad Farhan Fathurrahman [3], Herman Yoseph Sutarto [3], Ivana Semanjski [4,5], Adiyana Putri [6] and Sidharta Gautama [4,5]

1 Instrumentation and Control Research Group, Faculty of Industrial Technology, Institut Teknologi Bandung, Bandung 40132, Indonesia
2 University Center of Excellence Artificial Intelligence on Vision, NLP and Big Data Analytics, Institut Teknologi Bandung, Bandung 40132, Indonesia
3 Department of Intelligent System, PT. Pusat Riset Energi, Bandung 40226, Indonesia; mf.fathur@rce.co.id (M.F.F.); hy.sutarto@rce.co.id (H.Y.S.)
4 Department of Industrial Systems Engineering and Product Design, Ghent University, 9052 Ghent, Belgium; ivana.semanjski@ugent.be (I.S.); sidharta.gautama@ugent.be (S.G.)
5 Industrial Systems Engineering (ISyE), Flanders Make, 9052 Ghent, Belgium
6 Artificial Intelligence, Control and Automation Laboratory, Faculty of Industrial Technology, Institut Teknologi Bandung, Bandung 40132, Indonesia; 23820005@mahasiswa.itb.ac.id
* Correspondence: ejoel@tf.itb.ac.id

Abstract: The increase in traffic in cities world-wide has led to a need for better traffic management systems in urban networks. Despite the advances in technology for traffic data collection, the collected data are still suffering from significant issues, such as missing data, hence the need for data imputation methods. This paper explores the spatiotemporal probabilistic principal component analysis (PPCA) based data imputation method that utilizes traffic flow data from vehicle detectors and focuses specifically on detectors in urban networks as opposed to a freeway setting. In the urban context, detectors are in a complex network, separated by traffic lights, measuring different flow directions on different types of roads. Different constructions of a spatial network are compared, from a single detector to a neighborhood and a city-wide network. Experiments are conducted on data from 285 detectors in the urban network of Surabaya, Indonesia, with a case study on the Diponegoro neighborhood. Methods are tested against both point-wise and interval-wise missing data in various scenarios. Results show that a spatial network adds robustness to the system and the choice of the subset has an impact on the imputation error. Compared to a single detector, spatiotemporal PPCA is better suited for interval-wise errors and more robust against outliers and extreme missing data. Even in the case where an entire day of data is missing, the method is still able to impute data accurately relying on other vehicle detectors in the network.

Keywords: urban traffic network; data imputation; spatiotemporal analysis; probabilistic PCA; traffic management

Citation: Joelianto, E.; Fathurrahman, M.F.; Sutarto, H.Y.; Semanjski, I.; Putri, A.; Gautama, S. Analysis of Spatiotemporal Data Imputation Methods for Traffic Flow Data in Urban Networks. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 310. https://doi.org/10.3390/ijgi11050310

Academic Editors: Wolfgang Kainz and Hartwig H. Hochmair

Received: 21 March 2022
Accepted: 7 May 2022
Published: 12 May 2022

1. Introduction

In cities world-wide, traffic is continuously increasing. Especially in urban environments, this makes the management of traffic more complex. In large-scale road networks, appropriate traffic management systems are needed to control traffic flows under varying conditions. Methods developed to deal with the traffic management system are, for example, SCATS [1–3], max pressure control [4–9] and other methods reviewed in the following literature [10–12]. To work with maximum efficiency, the methods require complete and reliable traffic flow data.

Despite the advances in technology for traffic data collection, resulting traffic flow data are not perfect, and important problems, such as missing data, are unavoidable [13,14].

For example, the missing data ratio of loop detectors collected by the performance measurement system (PeMS) in California is higher than 10% [15]. Missing data can occur due to a malfunction of the sensing, communication errors, power problems, occlusion, etc. Sometimes, detectors can fail for longer periods of time, creating gaps in the time signal. These issues pose a challenge for traffic management systems that often rely on complete and trusted data. Consequently, there is a clear need for methods that can accurately impute missing data.

The problem of missing data has received a lot of attention in the literature and has been studied across fields, such as engineering, medicine and economics [16–18]. Papers explore different categories of data imputation methods, such as prediction-based, interpolation-based, and statistical learning-based methods. Prediction-based methods, such as ARIMA, view missing data points as a value to predict based on the relation derived from historical data [19]. Interpolation-based methods, such as linear, spline, and nearest neighbor interpolation, impute missing data from a weighted average of known past data points that have a similar pattern as the current data point and assume the existence of seasonality in the data [20]. Statistical-learning-based methods assume that data follow some probability distribution and learn the statistical features of the data for data imputation. Examples of statistical-learning-based methods include mean imputation, hot-deck, and multiple imputation [21].

Various techniques have been proposed specifically to address missing data in traffic flow data. Most of them focus on temporal correlations on a day-to-day basis. Zhong et al. [22] studied data imputation methods used in several transportation agencies in North America and Europe. Most agencies mainly used simple factor and time series analysis methods based on historical data. These approaches assume strong daily traffic flow similarity over intervals. Experimental results show that these methods can have large errors up to 80% absolute percentage error in the morning peak hours. Ni and Leonard [23] proposed a data imputation method using a Bayesian network to learn from historical data and a Markov chain Monte Carlo technique to sample from probability distributions from the trained Bayesian network. Statistical-learning-based methods such as these try to derive a statistical model of traffic flow. They typically perform better compared to conventional methods. Another statistical method is called the probabilistic principal components analysis (PPCA) data imputation method, first proposed by Qu et al. [24]. PPCA imputes missing data based on a PCA-like analysis of historical data and has shown to improve performance up to 25%, compared to classical methods. Tan et al. [25] developed the RPCA method that exploits known limits of traffic volume and day mode similarity. The daily traffic similarity is used to impute missing data by the low-rank hypothesis of the constructed traffic flow matrix. The physical limits of the road capacity and nonnegativity are utilized in the optimization process as constraints.

Although all the mentioned data imputation methods above perform well compared to conventional methods, the weakness of these methods is that they only utilize temporal information from daily flow similarity, and hence, the change of the temporal pattern caused by traffic breakdown or burst [26] might degrade the performance. It also assumes that the traffic flow data are not spoiled by outliers which frequently occur in real-world settings. Traffic data collected by vehicle detectors contain spatiotemporal information, as vehicles pass through several vehicle detectors along their routes. Intuitively, the traffic flow data collected from vehicle detectors located close to each other will be highly spatially correlated; hence, the addition of spatial information in traffic data imputation method is useful and could improve its performance. In recent years, more research on traffic data imputation involves spatial correlation and shows promising results.

Smith et al. [27] studied both heuristic approaches and statistical approaches such as historical average and data augmentation. The research showed that it is possible to impute the data of a detector from other detectors surrounding it in a freeway context. Chen et al. [28] proposed a data imputation method that models the relationship between neighboring loops as a linear model, and linear regression is used to estimate the missing

data using historical data. The presented results showed better performance compared to conventional interpolation methods in freeways located in California. Li et al. [29] researched that the spatial information extracted from the information of multiple sensors help in reducing imputation error for the PPCA and KPPCA methods. Ran et al. [30] proposed tensor-based missing traffic data imputation that utilizes four-way tensors, consisting of day, week, time, and space information. The results have shown that the addition of spatial information could help to reduce imputing errors, even in extreme missing data cases. Laña et al. [31] developed the spatial-context sensing data imputation method that utilizes all vehicle detectors in the central area of Madrid with well-distributed vehicle detectors. The imputing method is built upon the predictions from an extreme learning machine (ELM) model. Li et al. [32] showed that the PPCA-based method using a single vehicle detector outperforms several data imputation methods, such as ARIMA, Bayesian network, k-NN, local least squares (LLS), and Markov chain Monte Carlo (MCMC).

The above research shows that utilizing spatial information can improve data imputation. However, the studies mainly focus on freeway settings or assume detectors to have similar behavior [29–32]. Urban traffic has very different characteristics compared to freeways. In freeways, the spatial correlation between vehicle detectors is straightforward, as detectors on the same link and close distance are usually directly related, so the characteristics of traffic flow between a vehicle detector and its upstream and downstream counterparts is similar, only affected by time lag. In an urban context, detectors are usually positioned to count vehicles that leave an intersection. This means that each detector is located on a different link, separated by traffic lights. Consequently, although detectors can be close together, they can measure very different flow behaviors, making the spatial correlation between detectors in urban networks not straightforward.

This paper investigates the performance of data imputation in an urban setting under different scenarios for spatial information, exploring single-detector, sub-network and citywide network definitions. Data imputation is performed using the spatiotemporal PPCA-based method, utilizing spatiotemporal correlation in an urban network by modifying the observed data matrix. The robustness of the method is explored by testing the method from small to severe error conditions. The comparison explores the effect of different definitions of spatial network related to the vehicle detectors on the data imputation performance.

The rest of this paper is organized as follows. Section 2 explains the theory behind PPCA-based data imputation methods, explanation of both the single detector PPCA-based method and network PPCA-based method, missing data classification used in this research work, case study used for experiment, and the data imputation performance metrics used. The experiment results are shown in Section 3 and discussed in Section 4. Section 5 concludes this research work and discusses future works.

2. Materials and Methods

In this paper, traffic flow data obtained from the area traffic control system (ATCS) located in the urban network of Surabaya city, Indonesia, are used for experiments. Artificial datasets with missing data were created by omitting data from the original dataset, and imputation methods are evaluated on these datasets. The following section describes the PPCA-based data imputation method and the extension toward spatiotemporal PPCA. The section presents the missing data scenarios used in the paper, explanation of the case study, and the performance metrics used for evaluation.

2.1. PPCA-Based Data Imputation Methods

PPCA-based data imputation methods for traffic data have been discussed in several papers [24,29,33]. PPCA is a reformulation of the well-known PCA as a maximum-likelihood estimation based on the data probability density model [34]. The PPCA method has demonstrated several advantages compared to PCA, such as the ability to handle missing data and better scalability. The idea behind the PPCA-based imputation method

is that missing data is treated as a random variable which is not observed. The model is trying to predict the probability function from the observed data so that missing data can be predicted from the probability function.

Supposing that the observed data are generated from PPCA model, the relation between observed data with its principal components can be described as a standard factor analysis mapping [35] as follows:

$$y = Wx + \mu + \varepsilon \tag{1}$$

where y is a d-dimensional vector of observed data, and x is a k-dimensional vector of latent variables. Generally, $k < d$ such that the latent variables reduce the dimension of the model and offer a parsimonious model. The $d \times k$ matrix W is a projection matrix that represents a linear mapping between observed data y and latent variables x. The mean matrix μ allows the model to have non-zero mean values, and ε is a matrix representing isotropic noise assumed to be independent and identically distributed normal with zero mean and σ^2 variance.

The number of principal components k is a design parameter of PPCA. Larger numbers of k lead to better preserved variance from the observed data and more accurately reconstructed data but it might cause the model to overfit. To balance generality and accuracy, k is usually calibrated using cross validation. The resulting model is defined as follows:

$$y \sim N(\mu, WW^T + \sigma^2 I) \tag{2}$$

For W and σ^2, there are no closed-form analytic formulations, and hence their estimates are determined by iterative maximization from the corresponding log likelihood using an expectation–maximization (EM) algorithm. An efficient EM algorithm for the estimation of these parameters was formulated in references [34,36,37].

The paper compares two approaches for PPCA-based data imputation methods. The first one, proposed by Qu et al. [24], the single detector PPCA-based method, is dependent solely on temporal correlation gathered from historical data of a single vehicle detector. The second one, proposed in this paper, is a spatiotemporal PPCA-based method that utilizes both temporal correlation and spatial correlation between vehicle detectors by modifying the observed data matrix used and using traffic count data from multiple vehicle detectors in an urban network.

2.1.1. Single Detector PPCA-Based Data Imputation Method

Assume traffic flow data at one vehicle detector is collected for one day and then gathered as series of data as $Y^1 = [y^1(1), \cdots, y^1(N)]$, where N denotes the number of data points per day. For example, if the vehicle detector sampling interval is 15 min, N equals 96. If traffic flow data is collected for D consecutive days, this yields D-dimensional row vectors. These row vectors are put together to result in a data matrix

$$Y_d \triangleq \begin{pmatrix} y^1(1) & \cdots & y^D(1) \\ \vdots & \ddots & \vdots \\ y^1(N) & \cdots & y^D(N) \end{pmatrix} \tag{3}$$

where each column represents traffic flow data collected in a single day. The resulting data matrix is $Y_d \in \mathbb{R}^{N \times D}$ for each vehicle detector.

This method assumes that the traffic flow values on the same sampling time but on different days are implicitly correlated through the PPCA model. It presumes that all elements in a particular row follow a joint distribution. This method also simultaneously uses the current-day flow fluctuation and its neighboring day traffic flow information, and hence it does not require strict similarity between all different days.

As discussed by Qu et al. [24], this method has two requirements. Firstly, data imputation results may be biased if the vehicle detector was malfunctioning for a long

time. Secondly, the reconstructed data from the model should preserve important aspects of the data, such as the distribution. This means that even though daily similarity of the flow is not strictly required, if the resulting model cannot preserve the distribution of the observed data, the imputation results might be inaccurate. In the remainder of this paper, this method is abbreviated as **Single PPCA**.

2.1.2. Spatiotemporal PPCA-Based Data Imputation Method

Suppose traffic flow data on all different vehicle detectors on a network are collected, and a series of data as $Y_1^1 = [y_1^1(1), \cdots, y_1^1(N)]$, $\cdots$, $Y_M^1 = [y_M^1(1), \cdots, y_M^1(N)]$ is acquired, where N is the number of data points per day and M is the number of vehicle detectors on a network. Assume the traffic flow data is gathered for consecutive days and all data points in a single detector are put together as $Y_1^1 = [y_1^1(1), \cdots, y_1^1(N)], \cdots, Y_1^D = [y_1^D(1), \cdots, y_1^D(N)]$. If all data points of a single vehicle detector are stacked as a single vector, then the traffic flow data can be arranged together into the following data matrix form $Y_t \in \mathbb{R}^{(N\times D)\times M}$ defined as

$$Y_t \triangleq \begin{pmatrix} y_1^1(1) & y_2^1(1) & \cdots & y_{M-1}^1(1) & y_M^1(1) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ y_1^1(N) & y_2^1(N) & \cdots & y_{M-1}^1(N) & y_M^1(N) \\ y_1^2(1) & y_2^2(1) & \cdots & y_{M-1}^2(1) & y_M^2(1) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ y_1^2(N) & y_2^2(N) & \ddots & y_{M-1}^2(N) & y_M^2(N) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ y_1^D(1) & y_2^D(1) & \cdots & y_{M-1}^D(1) & y_M^D(1) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ y_1^D(N) & y_2^D(N) & \cdots & y_{M-1}^D(N) & y_M^D(N) \end{pmatrix} \tag{4}$$

where each column represents data points from a single detector. The resulting matrix of observed data has dimensions $ND \times M$, where D denotes the number of consecutive days of traffic flow data on M vehicle detectors in the network. Figure 1 illustrates the construction of the matrix in Equation (4).

The proposed data matrix structure assumes that the traffic flow values in similar time slots over different detectors are implicitly correlated and follow a particular distribution. The method tries to recover the relationship between a group of vehicle detectors at different locations, utilizing both spatiotemporal information derived from traffic flow fluctuation of a particular vehicle detector and traffic flow information from other vehicle detectors in the network. As the method does not utilize traffic flow from neighboring days, there is no daily flow similarity requirement if the distribution derived from different vehicle detectors is preserved in the model. This can lead to better robustness in terms of traffic breakdown or burst. However, there are some caveats. The spatial correlation between different detectors typically decreases over distance. Additionally, in an urban network the vehicle detectors are not only separated by distance, but also separated by traffic lights located between detectors. Vehicle detectors can monitor different directions, even though their distance is close. Therefore, the choice of network and which detectors to put in the same network must be considered carefully. This is an important focus of this study. A comparison of the process flow for the single detector and spatiotemporal PPCA-based data imputation method is shown in Figure 2. The newly developed procedures are indicated in blue.

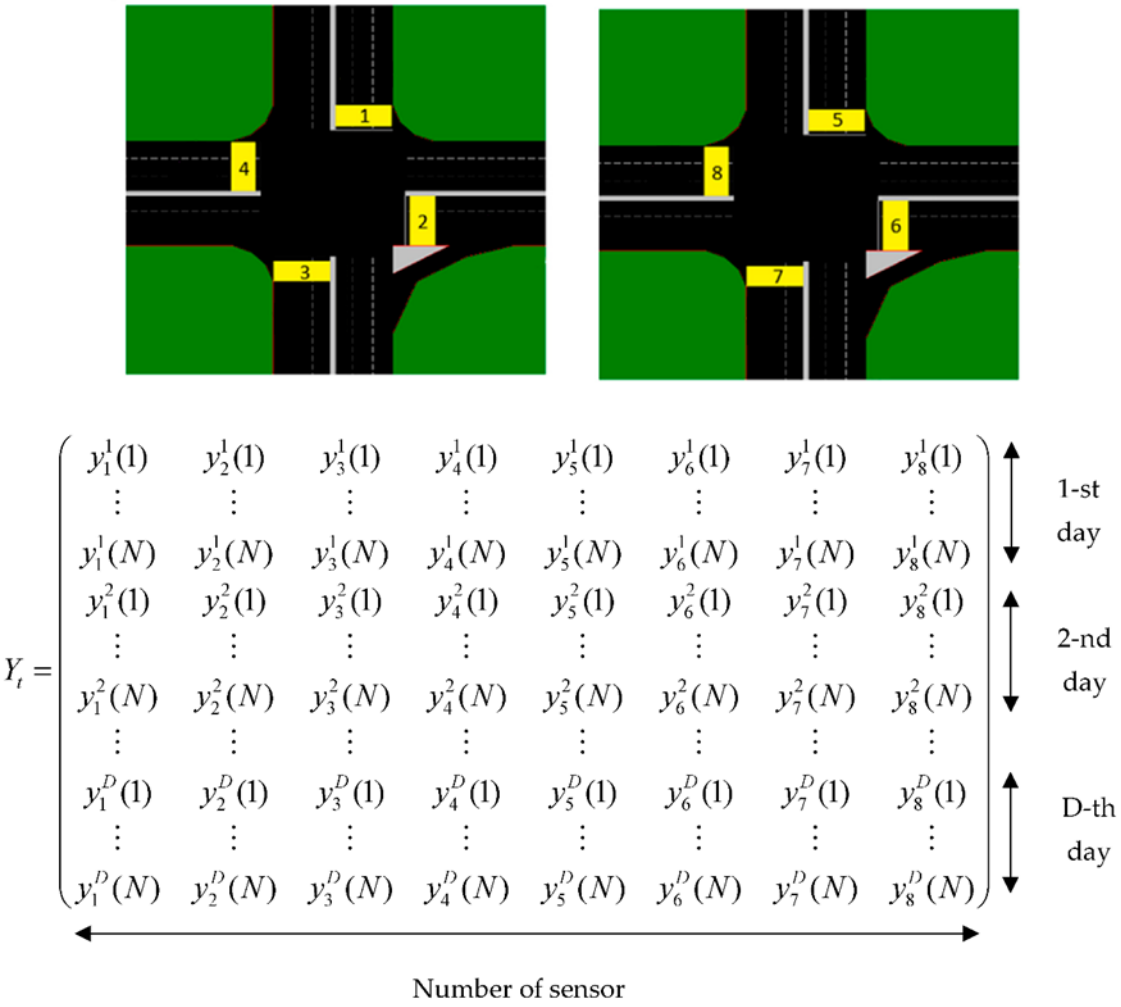


Figure 1. An example of construction of spatiotemporal data matrix (Equation (4)). The numbered yellow boxes on the top figure represent each vehicle detector.

Generally, a network usually refers to a large administrative area, such as a city-wide network. Unfortunately, in this definition of a network, vehicle detectors are not necessarily closely correlated because of the distance and different characteristics related to their position in the network. It is wiser to choose a subset of the network ('sub-network') that is more focused, where the nodes have similar properties (e.g., as defined by road class and land use). To show the impact of the network choice, two variants of this proposed method are considered: (1) a spatiotemporal PPCA-based method trained using data of an entire city-wide network termed **Network PPCA** and (2) a method trained using data of a hand-picked sub-network termed **Sub-Network PPCA**. The difference is the spatial correlation between detectors in the two approaches. **Network PPCA** is trained using traffic flow data that might have weak spatial correlation because of their vast area, while **Sub-Network PPCA** is trained using traffic flow data that have strong spatial correlation.

2.2. Missing Data

In general, there are three classes of missing data: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) [38]. Both MCAR and MAR have no underlying mechanism for the missing data, while NMAR assumes a dependence of the distribution of missing data on the complete dataset. These classifications of missing data have been used in different research [24,39,40].

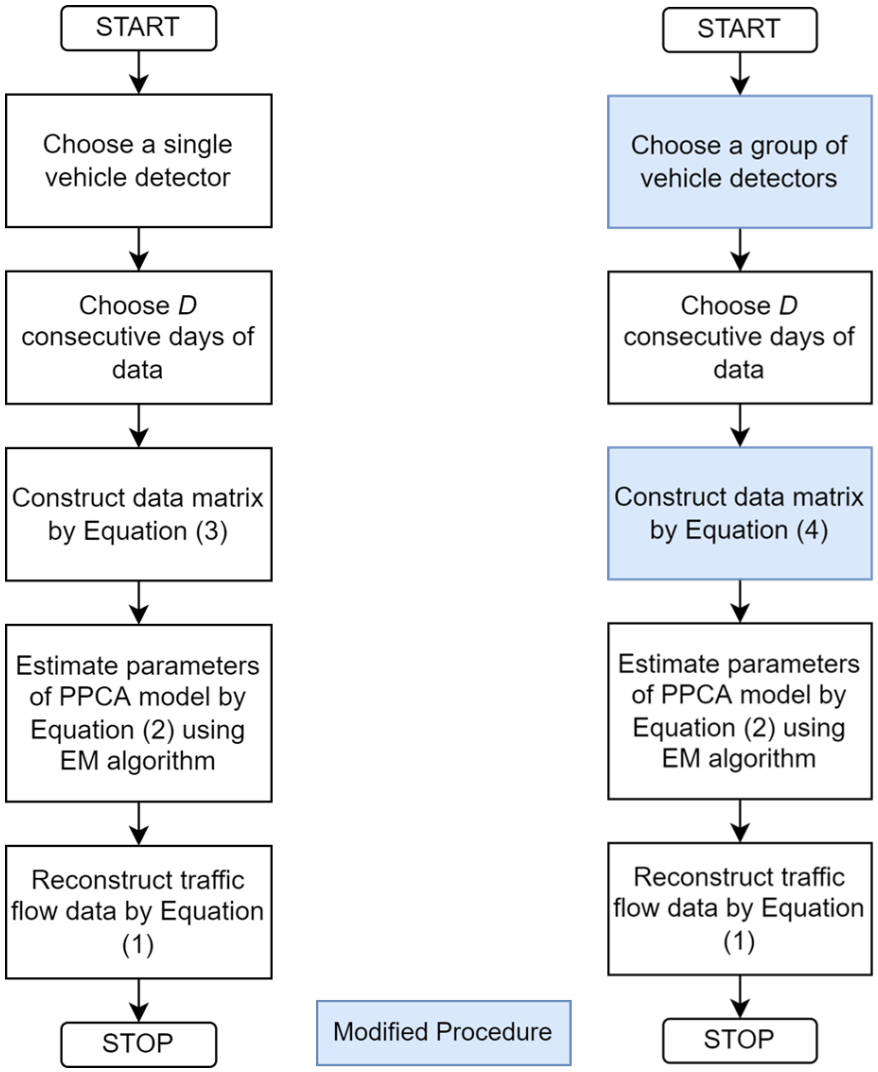


Figure 2. Single detector (**left**) and spatiotemporal (**right**) PPCA-based data imputation method flowchart comparison. EM algorithm for the estimation of PPCA model parameters was formulated in references [34,36,37].

In reality, missing data among the traffic flow observations may be a combination of MCAR, MAR, and NMAR. Because it is difficult to differentiate MAR and MCAR from NMAR based on data, Chiou et al. [41] suggested classifying the missing data as point-wise and interval-wise. Point-wise missing data are completely independent of the observed and unobserved value and also the missing points are randomly scattered. Point-wise missing data might be caused by short-term malfunction. Interval-wise missing data is missing data points that are grouped as an interval or a large group. Interval-wise missing data can be caused by a long-term malfunction in vehicle detectors, such as hardware malfunction, disconnected fiber optics, etc. Both point-wise and interval-wise missing data are illustrated in Figure 3, and both types of errors are considered in the experiments.

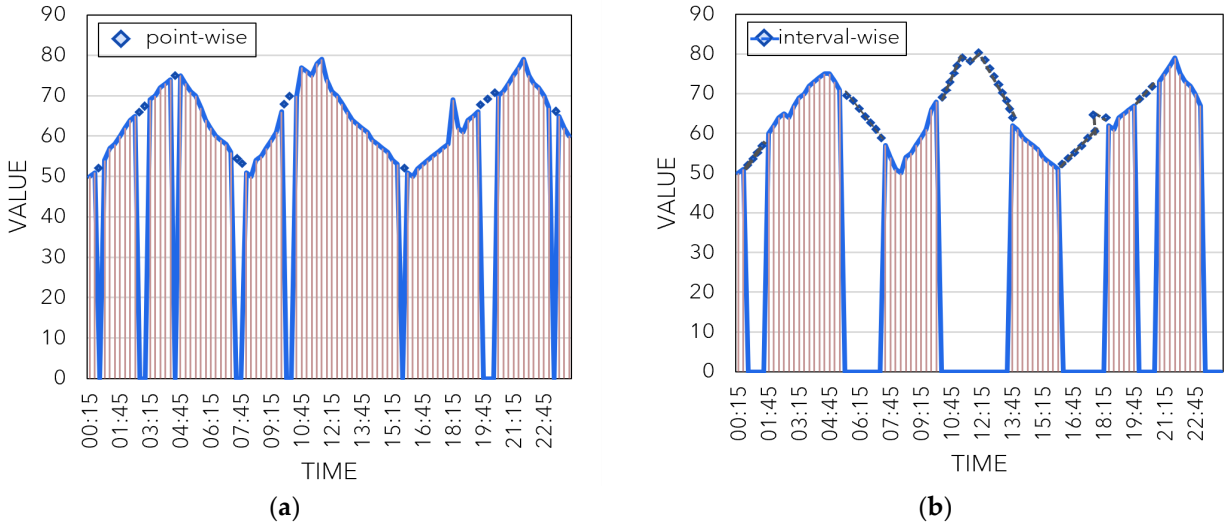


Figure 3. Illustration of (**a**) point-wise missing data and (**b**) interval-wise missing data.

2.3. Case Study: Urban Network of Surabaya, Indonesia

In this paper, traffic flow data from 438 vehicle detectors in the urban network of Surabaya, Indonesia are used in the experiments. The traffic count data are collected using video-based vehicle detectors and provided by the Surabaya area traffic control system, which uses the resulting data for the traffic control purposes. The traffic count data are aggregated every 15 min to obtain the traffic flow information. In this paper, the data were collected from 1 January 2020 to 29 February 2020. During this period, it was found that only 285 out of 438 vehicle detectors are in working condition, while the rest of the detectors either have a lot of missing data or are not working at all.

These 285 detectors are located at 115 intersections around the urban network of Surabaya, Indonesia, covering an area around 200 km^2 as shown in Figure 4a. As the purpose of these detectors is traffic control, the detector counts vehicles that leave an intersection. An illustration of the placement of detectors in an intersection is displayed in Figure 4b. For the 285 detectors, the missing data ratio is on average 18.3%. There are more interval-wise missing data compared to point-wise missing data because the most common cause of missing data is communication problems caused by hardware problems or internet disconnection.

In the PPCA-based method, it is assumed that the temporal pattern of the dataset is similar on a day-to-day basis. The assumption is met by only using data collected on Monday from eight different weeks. Each day has 96 data points, so the total number of data points available is 768 data points for each of the 285 vehicle detectors, which leads to a total of 218,880 traffic flow data points for the entire network.

The Diponegoro Neighborhood

As described in Section 2.1.2, this subsection examines the influence of the choice of subset ("sub-network") on performance compared to a city-wide network. In this case, the paper assesses a sub-network of vehicle detectors that are close together and have the same road class. This does not mean that they will measure the same flow as in a freeway setting, as vehicles can enter or leave between detectors and measurements are done in different directions. However, the vicinity and similar road class lead to a potentially higher spatial correlation compared to a city-wide network and this might affect the imputation performance. The city-wide network and sub-network are illustrated in Figure 4c.

Figure 4. Maps of 115 intersections equipped with working vehicle detectors located in the urban network of Surabaya, Indonesia. Every traffic light symbol in (**a**) represents one intersection. Every intersection usually consists of 3–4 vehicle detectors. The placement of vehicle detectors in an intersection is shown in (**b**), where each yellow square represents one vehicle detector that counts vehicles leaving an intersection. Part (**c**) compares the city-wide network (blue dashed area) with the hand-picked sub-network (red dashed area).

The Diponegoro neighborhood is a corridor that spans around 2.7 km, and roads on which detectors lie are categorized as primary arterial roads [42]. If these detectors are grouped in a sub-network, there are four intersections labeled as Site ID 2, Site ID 3, Site ID 4, and Site ID 5 as shown in Figure 5. Vehicle detectors located at Site ID 34 and Site ID 112 are not considered in this case study, as vehicle detectors in both intersections are located in the link with a different road class, secondary arterial road class and secondary collector road class, respectively.

For this case study, the attention is put on vehicle detectors that have the same road class (primary arterial road class) on the Site ID 2, Site ID 3, Site ID 4, and Site ID 5 intersections. Each intersection consists of 4 vehicle detectors counting vehicles for each link, resulting in 16 vehicle detectors across all mentioned intersections. There are 8 out of 16 detectors located in links categorized as primary arterial road, while the rest of the vehicle detectors are located in different road class. One out of eight detectors malfunctioned at the time of data collection, so seven working vehicle detectors are considered for the case study. All seven vehicle detectors are located at different links, separated by traffic lights, and not all of the vehicle detectors have a similar direction. The length of each link is considered short in Surabaya, and there are not too many small roads that might contribute to sink and source noises. The paper uses this choice of subset of detectors and explores if this construction adds value compared to a single detector and a city-wide network method.

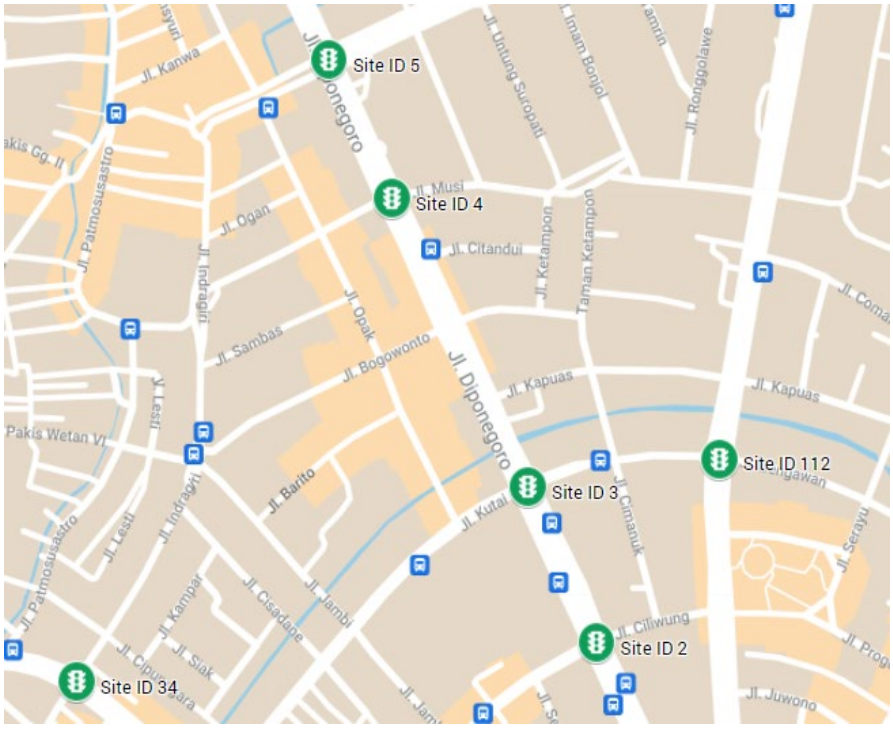


Figure 5. The Diponegoro neighborhood used in this paper as a case study. Each traffic light symbol represents a different intersection. Vehicle detectors in Diponegoro neighborhood consist of vehicle detectors in Site ID 2, Site ID 3, Site ID 4, and Site ID 5.

2.4. Data Imputation Performance Metrics

Generally, the performance of imputation methods is evaluated on the difference between imputed data and missing data. The popular performance metrics for data imputation are root mean square error (RMSE) and mean absolute percentage error (MAPE) [43–45]. RMSE is usually used as imputation performance metrics for single detector methods, as it is scale dependent. In this paper, data imputation methods impute missing data for multiple vehicle detectors simultaneously, so scale-invariant performance metrics are required. On the other hand, MAPE measures the percentage error of the imputed data in relation to the actual observed data, so it is scale invariant, and comparing data imputation performances for different vehicle detectors that have different mean values is possible. Unfortunately, traffic flow data may contain zero values data especially during midnight or dawn; hence, MAPE calculation of the traffic flow data imputation might have infinite error issues.

To solve these issues, the weighted mean absolute percentage error (WMAPE) [46–48] is considered to describe the imputation performance of each method. WMAPE is defined by the following formula

$$\text{WMAPE} = \frac{\sum_{i=1}^{I} |\hat{y}_i - y_i|}{\sum_{i=1}^{I} y_i} \times 100\% \tag{5}$$

where $\hat{y}_i$ are the i-th vectors of the imputed data, y_i are the i-th vectors of the known observed data, and I is the number of missing data. The total error between imputed data and known observed data is divided by the total values of known observed data, which removes the issue of having to divide by zero for traffic flow data that do not have negative values. The data points calculated in this performance metric are only at points where data are intentionally omitted.

3. Results

In the experiments, three methods were implemented and compared as follow:

1. **Single PPCA**: Trained using traffic flow data of a single detector collected during Monday for 8 weeks, leads to a 96×8 dimensional matrix of the observed data. In this approach, missing data in every vehicle detector in the Diponegoro neighborhood are imputed separately.
2. **Sub-Network PPCA**: Trained using traffic flow data of 7 vehicle detectors located at Diponegoro neighborhood and collected during Monday for 8 weeks, leads to a 768×7 dimensional matrix of the observed data. Missing data are imputed simultaneously for all vehicle detectors.
3. **Network PPCA**: Trained using traffic flow data of 285 vehicle detectors located at the urban network of Surabaya, Indonesia and collected during Monday for 8 weeks, leads to a 768×285 dimensional matrix of the observed data. Missing data are imputed simultaneously for all vehicle detectors, but only vehicle detectors located at Diponegoro neighborhood are considered.

All mentioned methods were evaluated for various types and amounts of missing data. The missing data are generated by intentionally omitting data from the observed data. The defined ratio for point-wise missing data is denoted by $\xi \in \{10, 25, 50, 75\}\%$, and the defined interval for interval-wise missing data is $\psi \in \{8, 16, 32, 64\}$ intervals per day. For the point-wise missing data, ξ % of traffic flow data are omitted individually at random across all the observed data, while for the interval-wise missing data, the ψ-interval of traffic flow data is omitted randomly in every one-day data.

Three different scenarios were also considered in this paper to showcase the performance and robustness of each method for various scenarios of missing data. Below is the explanation of each scenario.

1. **Scenario A**: Missing data points are distributed uniformly across all vehicle detectors and days.
2. **Scenario B**: Missing data points appear only in a number of vehicle detectors. The purpose of this scenario is to examine the case when there is a mix of functioning and malfunctioning detectors.
3. **Scenario C**: Scenario when several links suffer missing data for a day or more. The purpose of this scenario is to examine the case when several vehicle detectors suffer long-term malfunction.

a. Scenario A

Scenario A examines the performance where missing data is uniformly distributed across all links and all days. After that, WMAPE is calculated for all detectors in the Diponegoro neighborhood and the average error across all detectors. The results are shown in Figure 6 for point-wise missing data and Figure 7 for interval-wise missing data. The method fails to impute missing data in extreme cases, e.g., a ratio of 75% of point-wise missing data and 64 interval-wise missing data in Figure 8. This failure happens when an entire row of training data is missing, which happens at high levels of missing data.

b. Scenario B

Scenario A is not fully realistic, as it is unlikely that all vehicle detectors have missing data at the same time. Usually only a number of detectors are suffering from missing data in a network at a given time. In Scenario B, the performance between **Single PPCA** and **Sub-Network PPCA** is compared for cases where only some links in the network are suffering from missing data problems. **Network PPCA** is not included for this scenario, as both **Sub-Network PPCA** and **Network PPCA** are variants of the spatiotemporal PPCA-based method, and it is clear from Scenario A that **Network PPCA** performs worst. A performance comparison is made with different varieties of missing data and number of vehicle detectors malfunctioning.

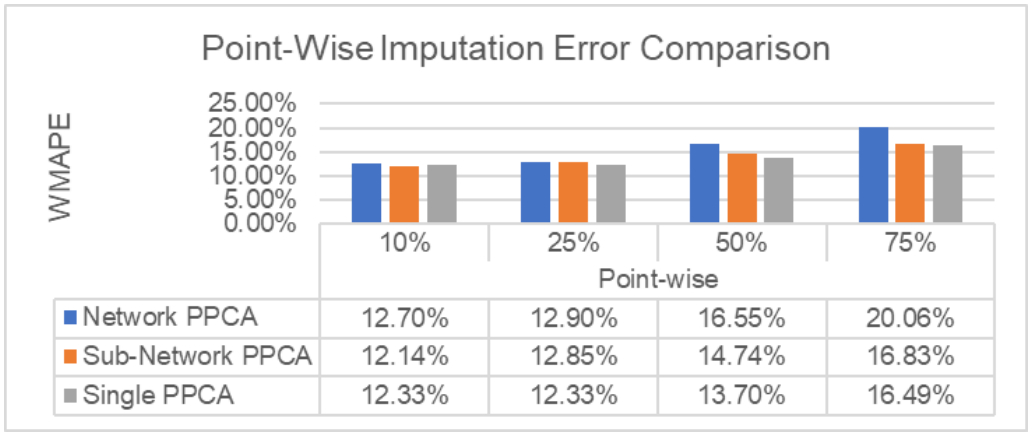


Figure 6. Imputation performance comparison for **Network PPCA**, **Sub-Network PPCA**, and **Single PPCA** for **Scenario A** with different point-wise missing data ratios.

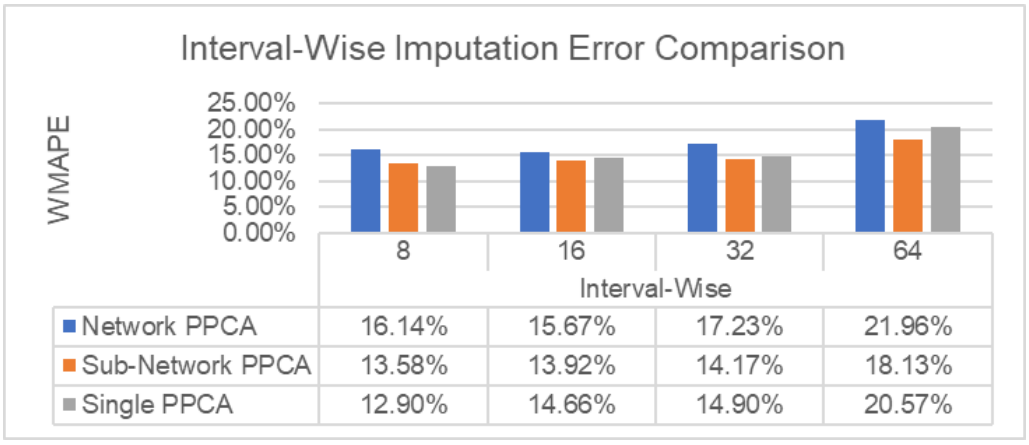


Figure 7. Imputation performance comparison for **Network PPCA**, **Sub-Network PPCA**, and **Single PPCA** for **Scenario A** with different interval-wise missing data intervals.

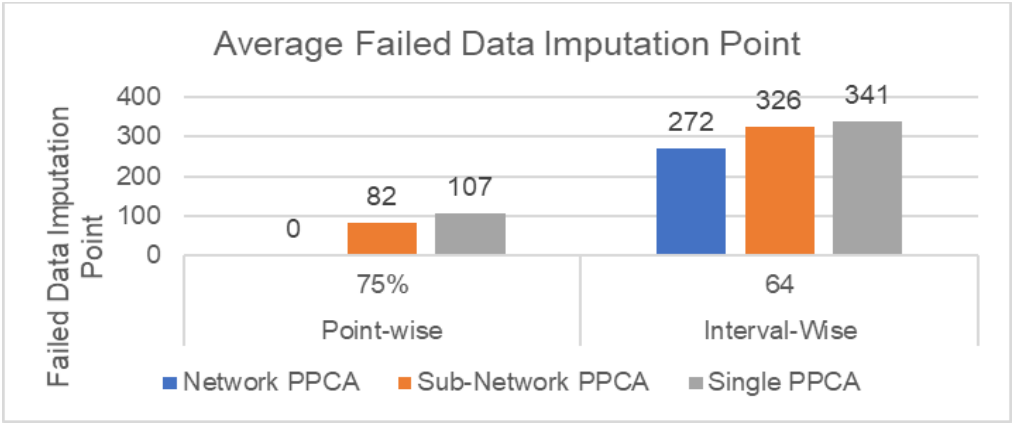


Figure 8. Average failed data imputation points for **Network PPCA**, **Sub-Network PPCA**, and **Single PPCA** for **Scenario A** with 75% ratio of point-wise missing data and 64 intervals of interval-wise missing data.

c. Scenario C

In the Surabaya network traffic flow data, there are cases where some of the vehicle detectors suffer from interval-wise missing data, which can last more than a day. This can happen because of internet connection problems, disconnected fiber optics or hardware problems and can take days to be repaired. In scenario C, missing data is imputed for such cases when 24 h of data is missing in the 8th week for a number of detectors.

d. Robustness against Outlier Vehicle Detectors

A robustness analysis is carried out by analyzing the impact of outlier vehicle detectors by comparing a sub-network PPCA trained using six vehicle detectors in the Diponegoro neighborhood plus an outlier detector, and a sub-network PPCA 2, using the same six vehicle detectors, excluding the outlier. The robustness analysis results are explained in the Discussion section.

4. Discussion

In Scenario A, for both point-wise and interval-wise missing data, Network PPCA performs worst, with a similar performance as the other methods for point-wise errors with ratios up to 25%, but performing worse on higher ratios and interval-type errors in general. Although it performs worst for WMAPE performance, the Network PPCA, however, is more robust in extreme missing data cases, as it can impute all missing data, even though the missing data ratio is 75%. This robustness comes from the amount of data used for training, as PPCA typically fails when an entire row of training data is missing due to combined errors. This is less likely to happen for larger networks. The reduced accuracy for Network PPCA can be explained by the heterogeneous characteristics of the detectors in large networks that reduce the focus of the generalized distributions. The results show that the choice of a better-defined network can have a significant impact on the performance of the spatiotemporal PPCA-based method.

The performance comparison of Sub-Network PPCA and Single PPCA indicates that both methods are close, with an average difference around 1% WMAPE. Single PPCA tends to perform better in point-wise missing data cases, while Sub-Network PPCA performs better in interval-wise missing data cases. Factors that reduce the performance of Single PPCA in interval-wise missing data cases might be the limitation of Single PPCA in that the method may be biased if the vehicle detector malfunctions for a long period of time as explained in Section 2.1.1. As explained in Section 2.3, missing data in vehicle detectors in urban networks are often interval-wise, making the proposed method, Sub-Network PPCA, more performant in these conditions. The results show that Sub-Network PPCA is able to impute missing data over a neighborhood, where the detectors are not necessarily fully spatially correlated, with a performance that is similar or better compared to Single PPCA.

For Scenario B, Figure 9 shows the performance of both methods tested against point-wise missing data for different numbers of malfunctioning detectors. In general, the performance between both methods is close, which is a similar result as in Scenario A Sub-Network PPCA indicates less influence by the missing data ratio, compared to Single PPCA from the steepness of the plot. One other important point is that Sub-Network PPCA successfully imputes all missing data, even when the missing data ratio is 75% as shown in Figure 10. Sub-Network PPCA is able to achieve this because the method can impute missing data from other healthy detectors, while in Single PPCA, each detector can only rely on its own historical data, resulting in around 80–88 failed data imputation points for each malfunctioning detector.

Figure 11 shows the performance of both methods tested against point-wise missing data for different numbers of malfunctioning detectors. For interval-wise missing data, the Sub-Network PPCA imputation performance is better than that of Single PPCA for all missing data intervals and number of malfunctioning detectors. These results are also more or less comparable to the results in Scenario A. This finding means that Sub-Network PPCA performs better in an urban network, where the majority of its missing data is interval-wise.

Figure 12 shows a similar result to the previous results tested on point-wise missing data, where, even in extreme case of missing data, Sub-Network PPCA can still successfully impute all missing data accurately, while Single PPCA fails more in interval-wise missing data, failing to impute around 304–360 data points. These experiments show that the Sub-Network PPCA imputation performance and robustness is better in Scenario B, and the traffic flow information received from neighboring healthy vehicle detectors gives Sub-Network PPCA the advantage.

In Scenario C, Single PPCA is unable to impute missing data if the traffic count data is missing for an entire day because one entire column of the dataset is missing. To fix this singular case, four data points (1 h of traffic flow data) are imputed based on the historical average to enable Single PPCA to work. The number of malfunctioning detectors were tested ranging from one to four out of seven vehicle detectors to see the impact of the number of vehicle detectors malfunctioning. In this case, Sub-Network PPCA performs significantly better compared to Single PPCA as shown in Table 1 because the Sub-Network PPCA is able to impute data based on the spatial correlation derived from other vehicle detectors in the sub-network. Single PPCA is unable to impute data accurately because the resulting data from historical average are not accurate enough, thus resulting in inaccurate data imputation. The results also show that the imputation error of Sub-Network PPCA increases with the number of malfunction vehicle detectors as expected but continues to outperform Single PPCA. It shows the robustness of Sub-Network PPCA against extreme missing data, as it is able to impute missing data for vehicle detectors that malfunction for longer periods, and a robustness against the number of detectors failing at the same time.

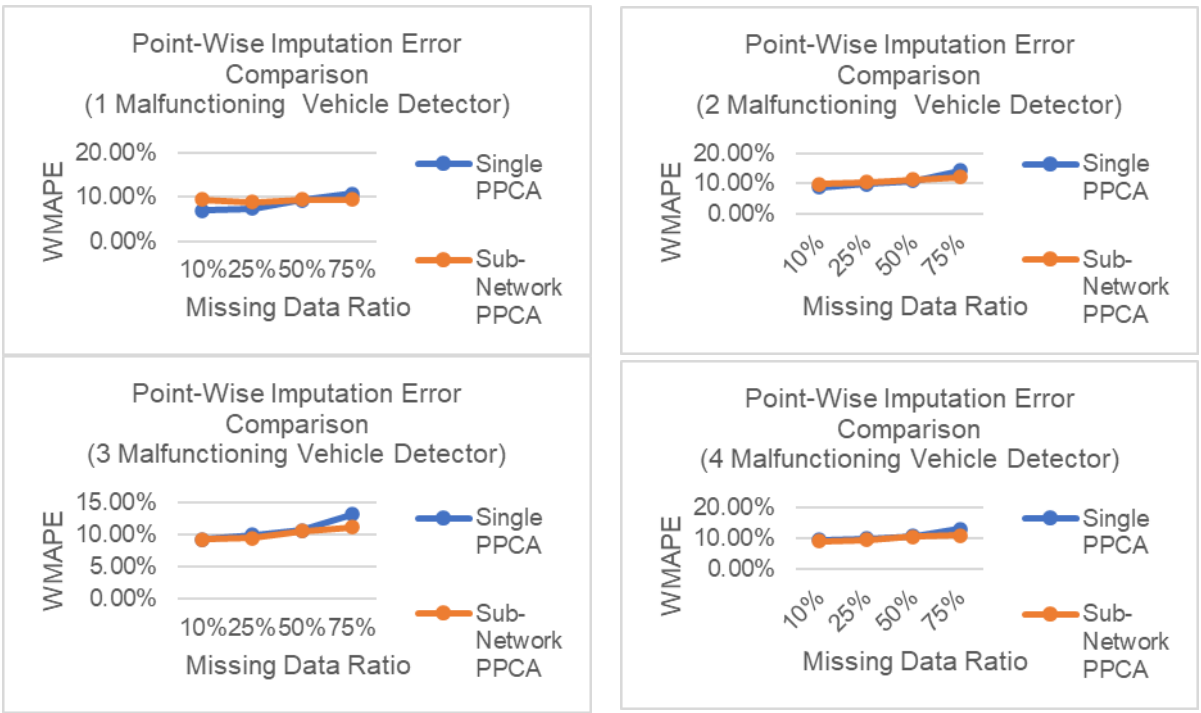


Figure 9. Scenario B Imputation error comparison for point-wise missing data.

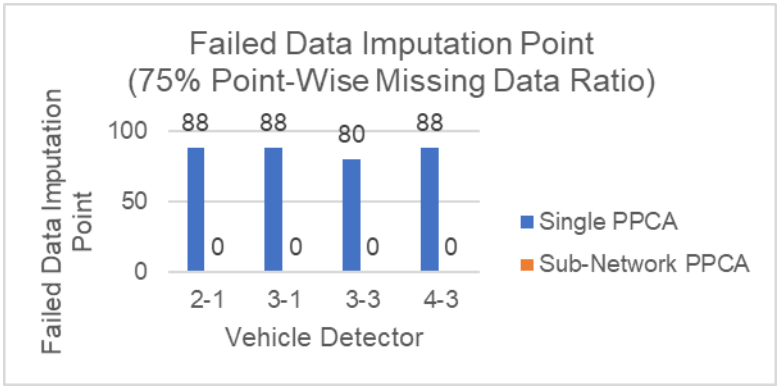


Figure 10. Failed data imputation points for in **Scenario B** for 75% point-wise missing data ratio.

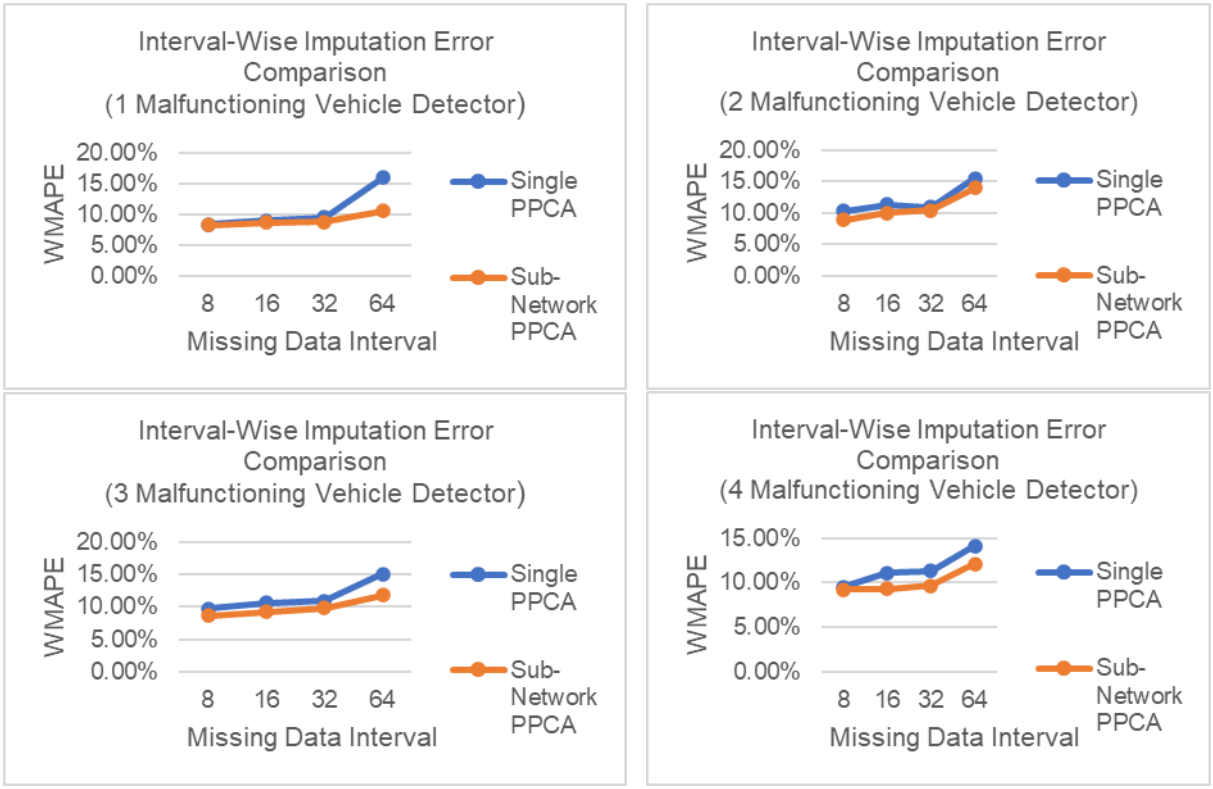


Figure 11. Scenario B Imputation error comparison for interval-wise missing data.

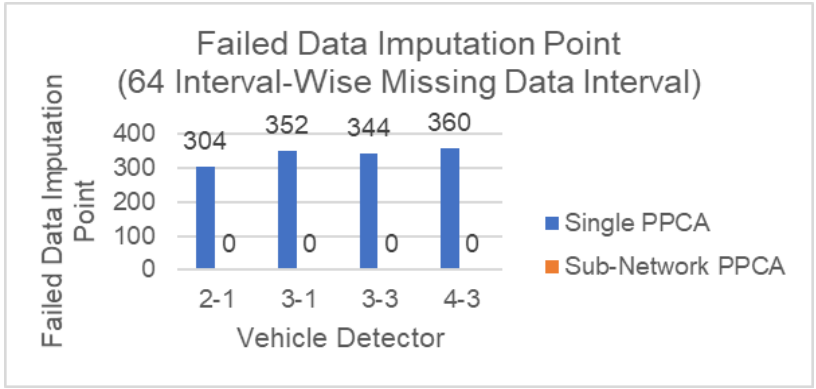


Figure 12. Failed data imputation points for in **Scenario B** for 64 interval-wise missing data intervals.

Table 1. Imputation error comparison between **Sub-Network PPCA** and **Single PPCA** for **Scenario C**.

Malfunctioning Vehicle Detector	Sub-Network PPCA				Single PPCA
	1 Malfunctioning Detector	2 Malfunctioning Detector	3 Malfunctioning Detector	4 Malfunctioning Detector	
2-1	8.67%	10.42%	11.52%	11.52%	23.47%
3-1		12.80%	14.40%	14.47%	46.02%
3-3			8.38%	10.14%	20.55%
4-3				9.13%	30.86%

For the analysis of the robustness against the outlier, in Table 2, it is found that one of the vehicle detectors in the Diponegoro site, namely detector 4-1, has a large imputation error compared to other detectors in all methods. The reason is because the temporal pattern of detector 4-1 fluctuates over the weeks, which results in large imputation errors as shown in Figure 13. All the methods show this error, confirming that this particular vehicle detector is an outlier. Figure 14 shows that the performance results between Sub-Network PPCA and Sub-Network PPCA 2 are quite similar, and it shows that the network PPCA method is able to impute data accurately, even in the presence of outliers. This would give a margin of error when constructing good subsets of detectors to include in one sub-network.

Table 2. Imputation error of each vehicle detector in Diponegoro neighborhood for 10% point-wise missing data and 8 interval-wise missing data. Vehicle Detector 4-1 is considered an outlier, as it has huge imputation error compared to other links.

Vehicle Detector	10% Point-Wise Missing Data Imputation Error		
	Network PPCA	Sub-Network PPCA	Single PPCA
2-1	7.39%	10.28%	7.89%
3-1	9.96%	9.86%	10.16%
3-3	7.28%	9.79%	8.55%
4-1	40.95%	29.55%	31.26%
4-3	8.97%	8.88%	10.84%
5-1	7.78%	8.63%	8.94%
5-3	6.59%	8.01%	8.68%
Average	**12.70%**	**12.14%**	**12.33%**
Vehicle Detector	**8 Interval-Wise Missing Data Imputation Error**		
	Network PPCA	**Sub-Network PPCA**	**Single PPCA**
2-1	8.38%	15.21%	12.96%
3-1	15.62%	11.88%	12.05%
3-3	10.55%	9.71%	8.62%
4-1	51.06%	28.68%	26.30%
4-3	8.88%	7.62%	9.63%
5-1	9.82%	12.65%	10.83%
5-3	8.70%	9.35%	9.90%
Average	**16.14%**	**13.58%**	**12.90%**

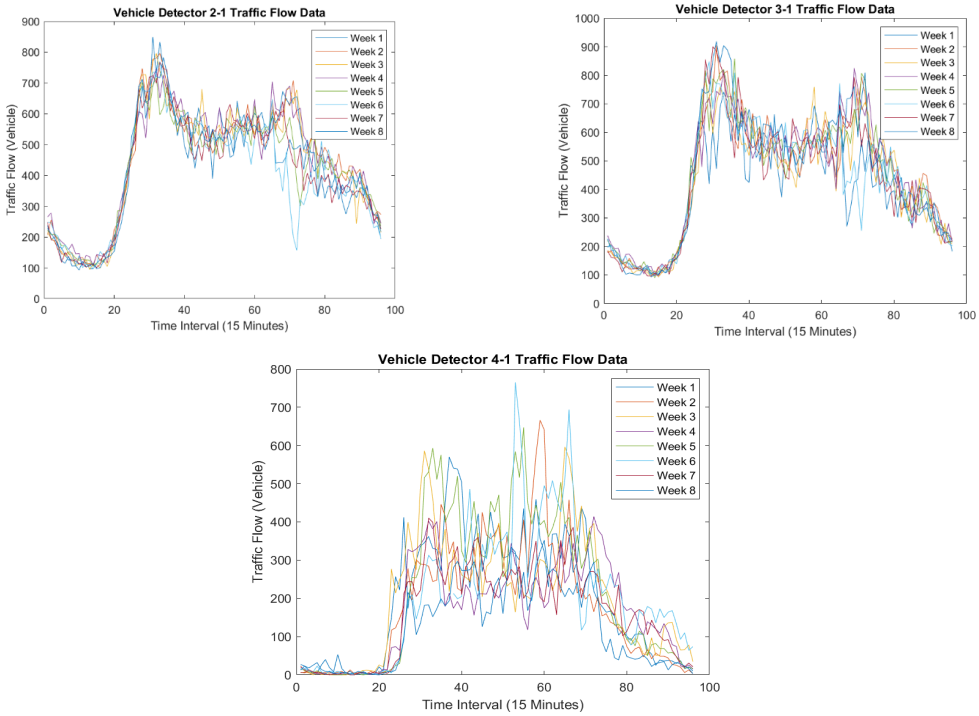
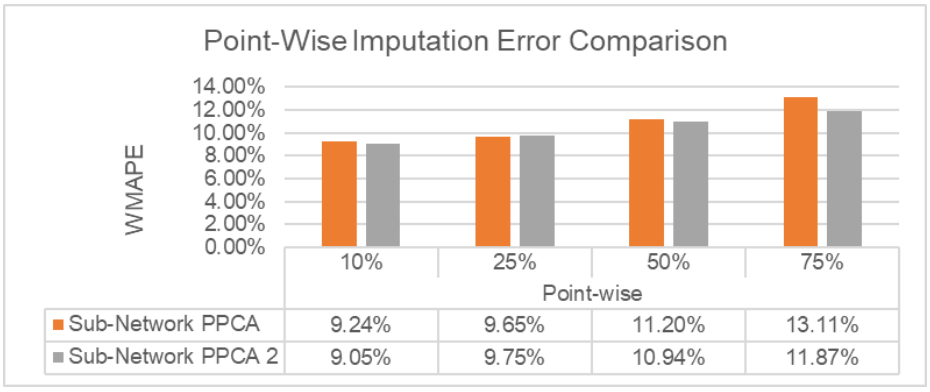
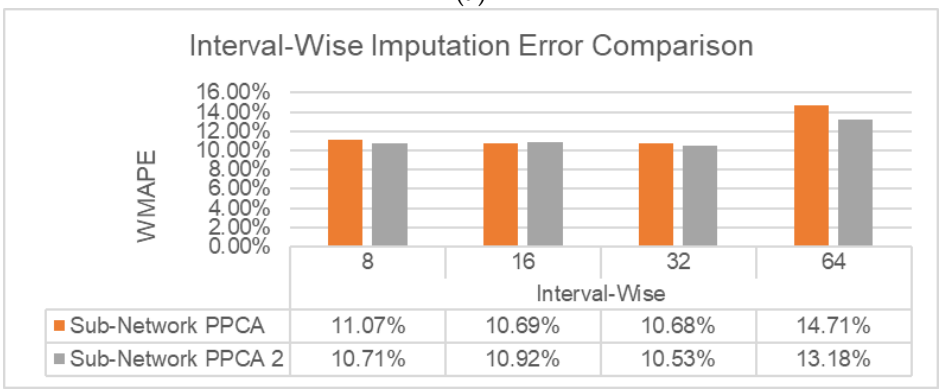


Figure 13. Plot of traffic count data of vehicle detector 4-1 for each week show huge fluctuation compared to other vehicle detectors in Diponegoro neighborhood. This vehicle detector shows a fluctuating pattern, which results in huge imputation error.



(**a**)



(**b**)

Figure 14. Imputation error comparison between **Sub-Network PPCA** and **Sub-Network PPCA 2** for (**a**) point-wise missing data and (**b**) interval-wise missing data.

5. Conclusions

In this paper, the spatiotemporal PPCA-based data imputation method was analyzed by utilizing both temporal and spatial information from multiple vehicle detectors. Two different choices of spatial networks were considered, namely, a city-wide network, or **Network PPCA,** and a neighborhood-based network, or Sub-Network PPCA. Both networks were compared with **Single PPCA,** relying only on temporal information. The methods were tested against point-wise and interval-wise missing data. The results established that **Network PPCA** has the lowest accuracy of the three methods but devises a better robustness in extreme cases of missing data. Both **Single PPCA** and **Sub-Network PPCA** performed similarly when missing data were uniformly distributed across all days and all vehicle detectors. **Sub-Network PPCA** achieved better for interval-wise missing data, while the **Single PPCA** was better for point-wise missing data.

In the more realistic case where only some vehicle detectors suffer missing data problems, **Sub-Network PPCA** resulted in better performance for all types of missing data compared to **Single PPCA**, exploiting the information derived from healthy neighboring vehicle detectors. The neighboring healthy vehicle detectors in **Sub-Network PPCA** also helped to impute all missing data without failure in extreme cases of missing data, up to 75% point-wise missing data and 64 time-intervals of missing data.

When several vehicle detectors were malfunctioning for an entire day, **Sub-Network PPCA** was still able to impute data accurately by relying on other vehicle detectors information. **Single PPCA** was unable to impute missing data without using other methods such as historical average to impute current-day data points and was out-performed by **Sub-Network PPCA.** In the experiments, it was determined that one vehicle detector showed a fluctuating temporal pattern, resulting in a large local imputation error. The effect of this outlier was examined, and it was found that **Sub-Network PPCA** is robust against its presence. The result indicated that there was an error margin when constructing good subsets of detectors to include in one sub-network.

Overall, the experiments confirmed good results and indicated that the spatial information of a sub-network can lead to a precise and more robust performance. The choice of which vehicle detectors to include in a sub-network is still an open problem, but results show that a good choice leads to improved performance. Currently, the choice has been made manually based on road class and vicinity. Future work will focus on automated constructions of subsets of detectors. Furthermore, the results can be used for, to illustrate, the verification and validation of sensor selection problems in traffic management.

Author Contributions: Conceptualization, Endra Joelianto, Muhammad Farhan Fathurrahman, Sidharta Gautama, Herman Yoseph Sutarto and Ivana Semanjski; methodology, Endra Joelianto, Muhammad Farhan Fathurrahman, Sidharta Gautama, Herman Yoseph Sutarto and Ivana Semanjski; software, Muhammad Farhan Fathurrahman and Herman Yoseph Sutarto; validation, Endra Joelianto, Sidharta Gautama and Ivana Semanjski; formal analysis, Endra Joelianto, Sidharta Gautama, Herman Yoseph Sutarto and Ivana Semanjski; investigation, Endra Joelianto, Sidharta Gautama, Herman Yoseph Sutarto and Ivana Semanjski; resources, Herman Yoseph Sutarto and Sidharta Gautama; data curation, Muhammad Farhan Fathurrahman and Adiyana Putri; writing—original draft preparation, Muhammad Farhan Fathurrahman and Adiyana Putri; writing—review and editing, Endra Joelianto, Muhammad Farhan Fathurrahman, Sidharta Gautama, Herman Yoseph Sutarto, Ivana Semanjski and Adiyana Putri; visualization, Muhammad Farhan Fathurrahman and Adiyana Putri; supervision, Endra Joelianto and Herman Yoseph Sutarto; project administration, Endra Joelianto; funding acquisition, Endra Joelianto. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the World Class Professor (WCP) Program, the Ministry of Education, Culture, Research and Technology, Republic Indonesia, No. 2817/E4.1/KK.04.05/2021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The work was supported by the World Class Professor (WCP) Program, the Ministry of Education, Culture, Research and Technology, Republic Indonesia, No. 2817/E4.1/KK.04.05/2021. The authors would like to thank Kusumo Purnomoputro, PT. Newtel, for helpful discussion and providing traffic count data of Surabaya, Indonesia. We thank the anonymous reviewers for their insightful and constructive remarks that helped us improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lowrie, P.R. *SCATS, Sydney Co-ordinated Adaptive Traffic System: A Traffic Responsive Method of Controlling Urban Traffic*; Roads and Traffic Authority NSW: Darlinghurst, NSW, Australia, 1999.
2. Xu, D.; Dai, H.; Wang, Y.; Peng, P.; Xuan, Q.; Guo, H. Road traffic state prediction based on a graph embedding recurrent neural network under the SCATS. *Chaos Interdiscip. J. Nonlinear Sci.* **2019**, *29*, 103125. [CrossRef] [PubMed]
3. Bing, Q.; Qu, D.; Chen, X.; Pan, F.; Wei, J. Arterial travel time estimation method using SCATS traffic data based on KNN-LSSVR model. *Adv. Mech. Eng.* **2019**, *11*, 1687814019841926. [CrossRef]
4. Varaiya, P. Max pressure control of a network of signalized intersections. *Transp. Res. Part C Emerg. Technol.* **2013**, *36*, 177–195. [CrossRef]
5. Mercader, P.; Uwayid, W.; Haddad, J. Max-pressure traffic controller based on travel times: An experimental analysis. *Transp. Res. Part C Emerg. Technol.* **2020**, *110*, 275–290. [CrossRef]
6. Boukerche, A.; Zhong, D.; Sun, P. A Novel Reinforcement Learning-based Cooperative Traffic Signal System through Max-pressure Control. *IEEE Trans. Veh. Technol.* **2021**, *71*, 1187–1198. [CrossRef]
7. Ramadhan, S.A.; Sutarto, H.Y.; Kuswana, G.S.; Joelianto, E. Application of area traffic control using the max-pressure algorithm. *Transp. Plan. Technol.* **2020**, *43*, 783–802. [CrossRef]
8. Levin, M.W.; Rey, D.; Schwartz, A. Max-pressure control of dynamic lane reversal and autonomous intersection management. *Transp. B Transp. Dyn.* **2019**, *7*, 1693–1718. [CrossRef]
9. Joelianto, E.; Utami, F.P.; Sutarto, H.Y.; Gautama, S.; Semanjski, I.; Fathurrahman, M.F. Performance Analysis of Max-Pressure Control System for Traffic Network using Macroscopic Fundamental Diagram. *Int. J. Artif. Intell.* 2022; *in press*.
10. Lei, T.; Hou, Z.; Ren, Y. Data-driven model free adaptive perimeter control for multi-region urban traffic networks with route choice. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 2894–2905. [CrossRef]
11. Lee, W.H.; Chiu, C.Y. Design and implementation of a smart traffic signal control system for smart city applications. *Sensors* **2020**, *20*, 508. [CrossRef]
12. Durmusoglu, A.; Durmusoglu, Z.D.U. Traffic Control System Technologies for Road Vehicles: A Patent Analysis. *IEEE Intell. Transp. Syst. Mag.* **2020**, *13*, 31–41. [CrossRef]
13. Lin, P.W.; Chang, G.L. Modeling measurement errors and missing initial values in freeway dynamic origin–destination estimation systems. *Transp. Res. Part C Emerg. Technol.* **2006**, *14*, 384–402. [CrossRef]
14. El Faouzi, N.E.; Leung, H.; Kurian, A. Data fusion in intelligent transportation systems: Progress and challenges—A survey. *Inf. Fusion* **2011**, *12*, 4–10. [CrossRef]
15. PeMS. California Performance Measurement System. Available online: https://pems.dot.ca.gov/ (accessed on 20 November 2021).
16. Mockus, A. Missing data in software engineering. In *Guide to Advanced Empirical Software Engineering*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 185–200.
17. Sterner, J.A.; White, I.R.; Carlin, J.B.; Spratt, M.; Royston, P. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ Br. Med. J. Int. Ed.* **2009**, *339*, 157–160.
18. Vroomen, J.M.; Eekhout, I.; Dijkgraaf, M.G.; van Hout, H.; de Rooij, S.E.; Heymans, M.W.; Bosmans, J.E. Multiple imputation strategies for zero-inflated cost data in economic evaluations: Which method works best? *Eur. J. Health Econ.* **2016**, *17*, 939–950. [CrossRef] [PubMed]
19. Elshenawy, M.; El-Darieby, M.; Abdulhai, B. Automatic imputation of missing highway traffic volume data. In Proceedings of the 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Athens, Greece, 19–23 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 373–378.
20. Junninen, H.; Niska, H.; Tuppurainen, K.; Ruuskanen, J.; Kolehmainen, M. Methods for imputation of missing values in air quality data sets. *Atmos. Environ.* **2004**, *38*, 2895–2907. [CrossRef]
21. Jerez, J.M.; Molina, I.; García-Laencina, P.J.; Alba, E.; Ribelles, N.; Martín, M.; Franco, L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif. Intell. Med.* **2010**, *50*, 105–115. [CrossRef]
22. Zhong, M.; Sharma, S.; Liu, Z. Assessing robustness of imputation models based on data from different jurisdictions: Examples of Alberta and Saskatchewan, Canada. *Transp. Res. Rec.* **2005**, *1917*, 116–126. [CrossRef]
23. Ni, D.; Leonard, J.D. Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data. *Transp. Res. Rec.* **2005**, *1935*, 57–67. [CrossRef]
24. Qu, L.; Li, L.; Zhang, Y.; Hu, J. PPCA-based missing data imputation for traffic flow volume: A systematical approach. *IEEE Trans. Intell. Transp. Syst.* **2009**, *10*, 512–522.

25. Tan, H.; Wu, Y.; Cheng, B.; Wang, W.; Ran, B. Robust missing traffic flow imputation considering nonnegativity and road capacity. *Math. Probl. Eng.* **2014**, *2014*, 763469. [CrossRef]
26. Chen, C.; Wang, Y.; Li, L.; Hu, J.; Zhang, Z. The retrieval of intra-day trend and its influence on traffic prediction. *Transp. Res. Part C Emerg. Technol.* **2012**, *22*, 103–118. [CrossRef]
27. Smith, B.L.; Scherer, W.T.; Conklin, J.H. Exploring imputation techniques for missing data in transportation management systems. *Transp. Res. Rec.* **2003**, *1836*, 132–142. [CrossRef]
28. Chen, C.; Kwon, J.; Rice, J.; Skabardonis, A.; Varaiya, P. Detecting errors and imputing missing data for single-loop surveillance systems. *Transp. Res. Rec.* **2003**, *1855*, 160–167. [CrossRef]
29. Li, L.; Li, Y.; Li, Z. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transp. Res. Part C Emerg. Technol.* **2013**, *34*, 108–120. [CrossRef]
30. Ran, B.; Tan, H.; Wu, Y.; Jin, P.J. Tensor based missing traffic data completion with spatial–temporal correlation. *Phys. A Stat. Mech. Its Appl.* **2016**, *446*, 54–63. [CrossRef]
31. Laña, I.; Olabarrieta, I.I.; Vélez, M.; Del Ser, J. On the imputation of missing data for road traffic forecasting: New insights and novel techniques. *Transp. Res. Part C Emerg. Technol.* **2018**, *90*, 18–33. [CrossRef]
32. Li, Y.; Li, Z.; Li, L. Missing traffic data: Comparison of imputation methods. *IET Intell. Transp. Syst.* **2014**, *8*, 51–57. [CrossRef]
33. Li, Y.; Li, Z.; Li, L.; Zhang, Y.; Jin, M. Comparison on PPCA, KPPCA and MPPCA based missing data imputing for traffic flow. In *ICTIS 2013: Improving Multimodal Transportation Systems-Information, Safety, and Integration*; 2013; pp. 1151–1156. Available online: https://www.semanticscholar.org/paper/Comparison-on-PPCA%2C-KPPCA-and-MPPCA-Based-Missing-Li-Li/ddb4ea2090f90a77882d6773da7dbb52d3306e17 (accessed on 1 May 2022).
34. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611–622. [CrossRef]
35. Bartholomew, D.J.; Knott, M.; Moustaki, I. *Latent Variable Models and Factor Analysis: A Unified Approach*; John Wiley & Sons: Hoboken, NJ, USA, 2011; Volume 904.
36. Porta, J.M.; Verbeek, J.J.; Kröse, B.J. Active appearance-based robot localization using stereo vision. *Auton. Robot.* **2005**, *18*, 59–80. [CrossRef]
37. Ilin, A.; Raiko, T. Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res.* **2010**, *11*, 1957–2000.
38. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Hoboken, NJ, USA, 2019; Volume 793.
39. Tang, J.; Zhang, G.; Wang, Y.; Wang, H.; Liu, F. A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. *Transp. Res. Part C Emerg. Technol.* **2015**, *51*, 29–40. [CrossRef]
40. Henrickson, K.; Zou, Y.; Wang, Y. Flexible and robust method for missing loop detector data imputation. *Transp. Res. Rec.* **2015**, *2527*, 29–36. [CrossRef]
41. Chiou, J.M.; Zhang, Y.C.; Chen, W.H.; Chang, C.W. A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transp. B Transp. Dyn.* **2014**, *2*, 106–129. [CrossRef]
42. Peraturan Daerah Kota Surabaya Nomor 07 Tahun 2003 (Regional Regulation of The City of Surabaya Number 07 of 2003). Available online: https://jdih.surabaya.go.id/pdfdoc/perda_50.pdf (accessed on 1 May 2022).
43. Chen, X.; He, Z.; Sun, L. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. *Transp. Res. Part C Emerg. Technol.* **2019**, *98*, 73–84. [CrossRef]
44. Luo, X.; Meng, X.; Gan, W.; Chen, Y. Traffic data imputation algorithm based on improved low-rank matrix decomposition. *J. Sens.* **2019**, *2019*, 7092713. [CrossRef]
45. Velasco-Gallego, C.; Lazakis, I. Real-time data-driven missing data imputation for short-term sensor data of marine systems. *A comparative study. Ocean Eng.* **2020**, *218*, 108261. [CrossRef]
46. Wang, C.; Hou, Y.; Barth, M. Data-driven multi-step demand prediction for ride-hailing services using convolutional neural network. In *Advances in Computer Vision. CVC 2019. Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2020; pp. 11–22.
47. De Medrano, R.; Aznarte, J.L. A spatio-temporal attention-based spot-forecasting framework for urban traffic prediction. *Appl. Soft Comput.* **2020**, *96*, 106615. [CrossRef]
48. Chen, D.; Yan, X.; Liu, X.; Wang, L.; Li, F.; Li, S. Multi-Task Fusion Deep Learning Model for Short-Term Intersection Operation Performance Forecasting. *Remote Sens.* **2021**, *13*, 1919. [CrossRef]