sensors

MDPI

Article

Missing Traffic Data Imputation with a Linear Generative Model Based on Probabilistic Principal Component Analysis

Liping Huang *, Zhenghuan Li, Ruikang Luo and Rong Su *

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore
* Correspondence: liping.huang@ntu.edu.sg (L.H.); rsu@ntu.edu.sg (R.S.)

Abstract: Even with the ubiquitous sensing data in intelligent transportation systems, such as the mobile sensing of vehicle trajectories, traffic estimation is still faced with the data missing problem due to the detector faults or limited number of probe vehicles as mobile sensors. Such data missing issue poses an obstacle for many further explorations, e.g., the link-based traffic status modeling. Although many studies have focused on tackling this kind of problem, existing studies mainly focus on the situation in which data are missing at random and ignore the distinction between links of missing data. In the practical scenario, traffic speed data are always missing not at random (MNAR). The distinction for recovering missing data on different links has not been studied yet. In this paper, we propose a general linear model based on probabilistic principal component analysis (PPCA) for solving MNAR traffic speed data imputation. Furthermore, we propose a metric, i.e., Pearson score (p-score), for distinguishing links and investigate how the model performs on links with different p-score values. Experimental results show that the new model outperforms the typically used PPCA model, and missing data on links with higher p-score values can be better recovered.

Keywords: missing data; urban traffic sensing; probabilistic; principal component analysis

Citation: Huang, L.; Li, Z.; Luo, R.; Su, R. Missing Traffic Data Imputation with a Linear Generative Model Based on Probabilistic Principal Component Analysis. *Sensors* **2023**, *23*, 204. https://doi.org/10.3390/s23010204

Academic Editor: Felipe Jiménez

Received: 28 November 2022
Revised: 21 December 2022
Accepted: 23 December 2022
Published: 25 December 2022

1. Introduction

Traffic data generated by loop detectors or floating cars in urban road networks serve as the foundation for various data-driven applications in intelligent transportation systems, including traffic forecasting and traffic control [1–3]. However, even with ubiquitous sensing data, the missing data problem is almost inevitable due to either detector faults or a limited number of probe vehicles operating as mobile sensors in road networks, which means not each road in the network is covered by a detector or traveled by a probe vehicle in each minute [4–6]. Such an issue of missing traffic data poses obstacles for many further data-driven explorations in both academic and industrial fields, e.g., the link-based traffic status modeling, and network-wise traffic dynamics capturing [7,8]. Hence, accurate and reliable imputation is a basic need for such kind of incomplete data for the downstream explorations.

Many efforts have been made for estimating the missing traffic data on multiple traffic datasets, resulting the generative probabilistic model [9], the matrix decomposition and tensor factorization models [10–12], the autoencoder model [13], the fusion models [14]. Some basic mathematical models are also adopted, including the autoregressive integrated moving average (ARIMA) model, the Bayesian networks (BNs) method, the Markov chain Monte Carlo (MCMC) method, and the K-nearest neighbors (KNN) model, which are all tested in [15] for traffic missing data imputation.

The studies in [16] have validated that the matrix decomposition-based method is not capable for recovering missing data when the missing percentage large. The tensor models are based on the global structure capturing, and it is faced with challenging to permutation in the spatial and temporal dimension [17]. The probabilistic principal component analysis (PPCA) model [18] also plays a major role in missing data completion due to its generative

feature [19]. Observations are assumed to be generated from a low dimensional space, with which the missing data can be recovered by optimizing the generative parameters using the observations [20].

Although many studies have focused on tackling this kind of problem, existing studies focus on the situation that the data are missing at random. Specifically, missing data can be classified into missing at random, missing completely at random, and missing not at random (MNAR) [16]. MNAR data always exist in the practical scenarios, and it is more challenging to estimate the missing values, which is the target of this paper.

The studies in [15] demonstrate that the PPCA model yields best performance among ARIMA, BNs, MCMC, and the KNN models, and in the research in [21], it has been certified that the PPCA model outperforms the basic tensor decomposition method. Hence, in this study, we set the PPCA model as a basis and further improve the PPCA model for tackling the MNAR traffic data. Additionally, the missing data on different links or sensors may be of different levels of challenges for data completion. Hence, there is also a need to distinguish different scenarios that missing data are on different links or sensors. Instead of the centrality of a sensor in the network, we utilize the time series correlations to define a metric for distinguishing the role of a link in the traffic network. Such a metric is adaptive to the scenario that sensors or links are anonymous. Contributions of this work are summarized as below:

- We design a metric, p-score to denote the relative importance of links in terms of time series observations, which is used to distinguish the links with missing values.
- We propose a linear model for the MNAR traffic data imputation, which is based on the probabilistic principal component analysis.
- We conduct experiments on a real-world traffic dataset using the model and the proposed metric. Experimental results show missing data on links with higher p-score values can be better recovered. Moreover, testing on the real-world dataset, the results of the proposed model on links with the lowest p-score value also outperforms the typically used PPCA model.

The remainder of the paper is structured as follows. Section 2 presents the problem statement of the missing traffic data imputation. Section 3 the details of the proposed model. Section 4 shows the outcome of the experimental evaluation results, Section 5 presents a short discussion of the potential application scenarios of the proposed method, and finally, Section 6 gives the conclusions of this paper and the directions for future studies.

2. Problem Statement

Let $Y \in \mathbb{R}^{n \times p}$ be a traffic data organization matrix with each element Y_{ij} denoting the i observation of a link j.

We assume that the traffic data are missing and links with missing values are organized as $Y_{\cdot m_1}, Y_{\cdot m_2}, \ldots, Y_{\cdot m_d}$, which is indexed by $\mathcal{M} := \{m_1, m_2, \ldots, m_d\} \subset \{1, \ldots, p\}$ with $d < p$ are supposed to have missing values. Here, $\mathcal{M}$ is the link set that has missing values. Other values in Y are observed.

We label the missing status of Y_{ij} with another variable, written as

$$\Omega_{ij} = \begin{cases} 0, & Y_{ij} \textit{ is missing} \\ 1, & \textit{otherwise} \end{cases} \tag{1}$$

Traffic sensing in urban road networks is faced with the missing data, or data sparsity problem. We construct the traffic matrix Y with all missing values in columns $\mathcal{M}$. The missing data imputation problem is to estimate these missing values, i.e., $\hat{Y}_{im}$, $m \in \mathcal{M}$, where $\Omega_{im} = 0$.

3. Methodology

3.1. PPCA

Assuming that the target variable is organized as a matrix Y, and it can be drawn from X of a low rank by linear combination, written as

$$Y = 1\alpha + XA + \epsilon \tag{2}$$

here, $Y \in \mathbb{R}^{n\times p}$, where n is the sample number and p is the number of variables in the determination system. Specifically, in our link-based missing data imputation problem, p is the total number of links.

$X = (X_{1\cdot}|\cdots|X_{n\cdot})^T$ is the latent variable. $X \in \mathbb{R}^{n\times r}$, and the row is drawn from a Gaussian distribution with zero mean, i.e., $X_{i\cdot} \sim \mathcal{N}(0_r, \mathrm{I}d_{r\times r})$. Here, $r < \min\{n, p\}$, indicating a lower dimension. A is the loading matrix of rank r, $A \in \mathbb{R}^{r\times p}$. $\epsilon = (\epsilon_{1\cdot}|\cdots|\epsilon_{n\cdot})^T$ is a model error, and each row $\epsilon_{i\cdot} \sim \mathcal{N}(0_p, \sigma^2\mathrm{I}d_{p\times p}) \in \mathbb{R}^p$, which also has a zero mean. $1 = (1, \ldots, 1)^T \in \mathbb{R}^n$, $\alpha = (\alpha_1, \cdots, \alpha_p) \in \mathbb{R}^p$. Given the linear expression above, the mean value of each sample of Y is α.

3.2. Missing Variables Differentiation Based on Time Series

Assume that we have missing data on two different variables, $Y_{\cdot j}, Y_{\cdot k}$, with the same percent, the imputation accuracy can be different due to the variable's role in the whole variable set. In the traffic missing data imputation problem, two links, $Y_{\cdot j}, Y_{\cdot k}$, may have different correlations to other links. In this section, we prose a metric to differentiate the role of each link.

The observation of each link is also a time series. We first adopt the Pearson correlation coefficient to estimate the correlation between each pair of time series, which is calculated as

$$\rho(Y_{\cdot j}, Y_{\cdot k}) = \frac{Cov(Y_{\cdot j}, Y_{\cdot k})}{\sigma_{Y_{\cdot j}}\sigma_{Y_{\cdot k}}} \tag{3}$$

By calculate the Pearson correlation among each pair of variables, we can obtain a correlation matrix, written as

$$P = \begin{bmatrix} 1 & \cdots & \rho(Y_{\cdot 1}, Y_{\cdot n}) \\ \vdots & \ddots & \vdots \\ \rho(Y_{\cdot n}, Y_{\cdot 1}) & \cdots & 1 \end{bmatrix} \tag{4}$$

We define a Pearson score (p-score) for each variable to differentiate the variables in $\mathcal{M}$, which is calculated as

$$P_{score}(Y_{\cdot j}) = \sum_{k\in\{1,\ldots,n\}} P_{jk} \tag{5}$$

The variable $Y_{\cdot j}$ that obtains a higher p-score value than $Y_{\cdot k}$ denotes it has higher correlation to other links. Such a metric can differentiate the variables in terms of the time series observations. When $P_{score}(Y_{\cdot j}) > P_{score}(Y_{\cdot k})$, and the two links have the same missing data percentage, the imputation accuracy for $Y_{\cdot j}$ should be higher than that of $Y_{\cdot k}$.

3.3. Preliminaries and Assumptions

Assume that we have missing data on two different variables, $Y_{\cdot j}, Y_{\cdot k}$ with the same percent, the imputation accuracy can be different due to the variable's role in the whole variable set. In the traffic missing data imputation problem, two links $Y_{\cdot j}, Y_{\cdot k}$ may have different correlations to other links. In this section, we prose a metric to differentiate the role of each link.

Assume that we have a D dimensional Gaussian distribution, written as

$$\mathcal{N}(x|u, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-u)^T\Sigma^{-1}(x-u)\right) \tag{6}$$

where u is a D-dimensional mean vector, Σ is a $D \times D$ covariance matrix, $|\Sigma|$ denotes the determinant of Σ. Then, we partition the D-dimensional vector into two parts, written as

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \tag{7}$$

Correspondingly, the mean vector and the covariance matrix are, respectively partitioned into two parts and four parts, written as below.

$$u = \begin{pmatrix} u_a \\ u_b \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \tag{8}$$

We further utilize another variable Λ to denote the inverse matrix of the covariance matrix, defined as

$$\Lambda \equiv \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \tag{9}$$

Note that, we have the theory of matrix inverse as

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}CMBD^{-1} \end{pmatrix} \tag{10}$$

$$M = \left(A - BD^{-1}C\right)^{-1} \tag{11}$$

Hence, for the inverse of the covariance matrix, we have

$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \tag{12}$$

where we care about the expression of Λ_{aa} and Λ_{ab}, written as

$$\Lambda_{aa} = \left(\Sigma_{aa} - \Sigma_{ab}{\Sigma_{bb}}^{-1}\Sigma_{ba}\right)^{-1} \tag{13}$$

$$\Lambda_{ab} = -\left(\Sigma_{aa} - \Sigma_{ab}{\Sigma_{bb}}^{-1}\Sigma_{ba}\right)^{-1}\Sigma_{ab}{\Sigma_{bb}}^{-1} \tag{14}$$

For the Gaussian distribution, the exponent parts can be expanded as

$$-\frac{1}{2}(x-u)^T\Sigma^{-1}(x-u) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}u + const \tag{15}$$

when we partition the D-dimensional vector into two parts $x = (x_a, x_b)^T$, then the exponent part of the Gaussian distribution can be expanded into

$$\begin{aligned} &-\tfrac{1}{2}(x-u)^T\Sigma^{-1}(x-u) \\ &\quad = -\tfrac{1}{2}(x_a-u_a)^T\Lambda_{aa}(x_a-u_a) - \tfrac{1}{2}(x_a-u_a)^T\Lambda_{ab}(x_b-u_b) \\ &\quad -\tfrac{1}{2}(x_b-u_b)^T\Lambda_{ba}(x_a-u_a) - \tfrac{1}{2}(x_b-u_b)^T\Lambda_{bb}(x_b-u_b) \end{aligned} \tag{16}$$

We assume that x_b is known in advance, so it can be regarded as a constant. Hence, the first order of x_a is written as

$${x_a}^T\{\Lambda_{aa}u_a - \Lambda_{ab}(x_b-u_b)\} \tag{17}$$

which should have the same expression of the original expression for the first order part written as $x^T\Sigma^{-1}u$. For $x^T\Sigma^{-1}u$, when we consider the x_b is known, then $\Sigma^{-1}u$ can be written as ${\Sigma_{a|b}}^{-1}u_{a|b}$, which should be equal to $\Lambda_{aa}u_a - \Lambda_{ab}(x_b-u_b)$, written as

$$\Lambda_{aa}u_a - \Lambda_{ab}(x_b-u_b) = {\Sigma_{a|b}}^{-1}u_{a|b} \tag{18}$$

Hence, we have the expression the estimated value of $u_{a|b}$ written as conditional Gaussian distribution

$$u_{a|b} = \Sigma_{a|b}\{\Lambda_{aa}u_a - \Lambda_{ab}(x_b-u_b)\} \tag{19}$$

where Λ_{aa} and Λ_{ab} are already known as above.

Based on the conditional Gaussian distribution, we replace the x_b part with $\left((Y_{\cdot k})_{k\in\overline{\mathcal{M}}}\right)$, which is assumed to known observations, and replace the x_a part as the unknown part $Y_{\cdot m}$, which is to be estimated because that the data are missing. Then, we have the expectation of the estimation as

$$\mathbb{E}\left[Y_{\cdot m}\middle|\left((Y_{\cdot k})_{k\in\overline{\mathcal{M}}}\right)\right] = \alpha_m + \Sigma_{m,\overline{\mathcal{M}}}\Sigma^{-1}_{\overline{\mathcal{M}},\overline{\mathcal{M}}}\left(Y^T_{\cdot\overline{\mathcal{M}}} - \alpha_{\overline{\mathcal{M}}}\right) \tag{20}$$

Then, the estimation of the missing data is calculated as

$$\hat{Y}_{im} = \hat{\alpha}_m + \hat{\Sigma}_{m,\overline{\mathcal{M}}}\hat{\Sigma}^{-1}_{\overline{\mathcal{M}},\overline{\mathcal{M}}}\left(Y^T_{\cdot\overline{\mathcal{M}}} - \alpha_{\overline{\mathcal{M}}}\right) \tag{21}$$

Hence, the missing data estimations depend on the estimations of $\hat{\alpha}_m$, $\hat{\Sigma}_{m,\overline{\mathcal{M}}}$ and $\hat{\Sigma}^{-1}_{\overline{\mathcal{M}},\overline{\mathcal{M}}}$. Below are assumptions for estimating the model parameters.

Assumption 1: $\forall m \in \mathcal{M}, \forall j \in \mathcal{I}$, $\left(A_{\cdot m}\left(A_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right)$ *is invertible.* $\mathcal{I}_{-j} = \mathcal{I}\backslash\{j\}$.

Assumption 2: $\forall m \in \mathcal{M}, \forall j \in \mathcal{I}, Y_{\cdot j} \perp \Omega_{\cdot m}\left|\left((Y_{\cdot k})_{k\in\overline{\{j\}}}\right)\right.$.

Assumption 1 denotes that the matrix $\left(A_{\cdot m}\left(A_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right)$ is of full rank. **Assumption 2** denotes that, given the values in $(Y_{\cdot k})_{k\in\overline{\{j\}}}$, the column $Y_{\cdot j}$ is independent with the column $\Omega_{\cdot m}$.

The missing data imputation for MNAR is to estimate the value of Y_{im} with $m \in \mathcal{M}$ for i such that $\Omega_{im} = 0$. Assumption 2 leads to

$$\mathbb{E}\left[Y_{\cdot j}\middle|\left((Y_{\cdot k})_{k\in\overline{\{j\}}}\right)\right] = \mathbb{E}\left[Y_{\cdot j}\middle|\left((Y_{\cdot k})_{k\in\overline{\{j\}}}\right),\ \Omega_{im} = 1\right] \tag{22}$$

3.4. Estimation of α

We first define the regression coefficients of $Y_{\cdot j}$ on $Y_{\cdot m}$ and $Y_{\cdot k}$, for $k \in \mathcal{I}_{-j}$ in the complete case, that will be used to express the mean of a variable with MNAR values.

Considering the model $Y = 1\alpha + XA + \epsilon$, with an assumption that matrix $A \in \mathbb{R}^{r\times p}$ is of full rank r. Therefore, the expression of $Y_{\cdot j}$ can be reduced to the following linear system.

$$\left(Y_{\cdot m}\left(Y_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right) = 1\alpha_{|r} + (X_{\cdot 1}, \cdots, X_{\cdot r})A_{|r} + \epsilon_{|r} \tag{23}$$

Here, $\alpha_{|r}$ and $\epsilon_{|r}$ are the reduced matrix of α and ϵ. $A_{|r} \in \mathbb{R}^{r\times r}$ denotes the reduced matrix of $\left(A_{\cdot m}\left(A_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right)$.

Given Assumption 1, the $A_{|r}$ is invertible, and the inverted matrix is denoted as $\acute{A}^{-1}$. The latent matrix of full rank r can be written as

$$(X_{\cdot 1}, \cdots, X_{\cdot r}) = \left(\left(Y_{\cdot m}\left(Y_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right) - 1\alpha_{|r} - \epsilon_{|r}\right)\acute{A}^{-1}. \tag{24}$$

Using the original model $Y = 1\alpha + XA + \epsilon$, the expression of $Y_{\cdot j}$ is then can be written as

$$\begin{aligned} Y_{\cdot j} = 1\alpha_j + \left(\left(Y_{\cdot m}\left(Y_{\cdot j'}\right)_{j'\in\mathcal{I}_{-j}}\right) - 1\alpha_{|r} - \epsilon_{|r}\right)\acute{A}^{-1}A_{j\cdot} + \epsilon_{\cdot j} = \textstyle\sum_{k\in\{m\}\cup\mathcal{I}_{-j}}\left(\sum_{l\in m\cup\mathcal{I}_{-j}}\acute{A}^{-1}{}_{lk}A_{jl}\right)Y_{\cdot k} - \\ \textstyle\sum_{k\in\{m\}\cup\mathcal{I}_{-j}}\left(\sum_{l\in m\cup\mathcal{I}_{-j}}\acute{A}^{-1}{}_{lk}A_{jl}\right)(1\alpha_k + \epsilon_{\cdot k}) + 1\alpha_j + \epsilon_{\cdot j} \end{aligned} \tag{25}$$

where we can get the intercept and the coefficients of $Y_{\cdot j}$ on $\left(Y_{\cdot m}, (Y_{\cdot k})_{k\in\mathcal{I}_{-j}}\right)$.

For $j \in \mathcal{I}$, $k \in \mathcal{I}_{-j}$, let $A^c_{j\to m,\mathcal{I}_{-j}[0]}$ be the intercept, and $A^c_{j\to m,\mathcal{I}_{-j}[m]}$ $A^c_{j\to m,\mathcal{I}_{-j}[k]}$ be the coefficients standing for the effects of $Y_{\cdot j}$ on $\left(Y_{\cdot m}, (Y_{\cdot k})_{k\in\mathcal{I}_{-j}}\right)$ in the complete case, i.e., when $\Omega_{\cdot m} = 1$. Then, we have

$$\left(Y_{\cdot j}\right)_{|\Omega_{\cdot m}=1} = A^c_{j\to m,\mathcal{I}_{-j}[0]} + \sum_{j'\in\mathcal{I}_{-j}} A^c_{j\to m,\mathcal{I}_{-j}[j']} Y_{\cdot j'} + A^c_{j\to m,\mathcal{I}_{-j}[m]} Y_{\cdot k} + \zeta^c \tag{26}$$

where the coefficients are calculated as below equations.

$$A^c_{j\to m,\mathcal{I}_{-j}[j']} = \sum_{l\in\{m\}\cup\mathcal{I}_{-j}} \acute{A}^{-1}{}_{lj'} A_{jl},\ j' \in \mathcal{I}_{-j} \tag{27}$$

$$A^c_{j\to m,\mathcal{I}_{-j}[m]} = \sum_{l\in\{m\}\cup\mathcal{I}_{-j}} \acute{A}^{-1}{}_{lm} A_{jl} \tag{28}$$

$$A^c_{j\to m,\mathcal{I}_{-j}[0]} = 1\alpha_j - \sum_{k\in\{m\}\cup\mathcal{I}_{-j}} (\sum_{l\in\{m\}\cup\mathcal{I}_{-j}} \acute{A}^{-1}{}_{lk} A_{jl}) 1\alpha_k \tag{29}$$

$$\zeta^c = -\sum_{k\in\{m\}\cup\mathcal{I}_{-j}} A^c_{j\to m,\mathcal{I}_{-j}[k]} \epsilon_{\cdot k} + \epsilon_{\cdot j} \tag{30}$$

Here, the arrow $j \to m, \mathcal{I}_{-j}$ indicates the regression model of $Y_{\cdot j}$ on $Y_{\cdot(m,\mathcal{I}_{-j})}$, and the squared bracket $[k]$ indicates the coefficient. Based on the model setting, we have $\mathbb{E}[\epsilon_{\cdot k}] = 0$, hence $\mathbb{E}[\zeta^c] = 0$.

Assumption 2 leads to

$$\begin{aligned}
\mathbb{E}\left[Y_{\cdot j}\middle|\left((Y_{\cdot k})_{k\in\overline{\{j\}}}\right)\right] &= \mathbb{E}\left[Y_{\cdot j}\middle|\left((Y_{\cdot k})_{k\in\overline{\{j\}}}\right),\ \Omega_{im} = 1\right] \\
&= \mathbb{E}\left[A^c_{j\to m,\mathcal{I}_{-j}[0]} + \textstyle\sum_{k\in\{m\}\cup\mathcal{I}_{-j}} A^c_{j\to m,\mathcal{I}_{-j}[k]} Y_{\cdot k} + \zeta^c \middle|\left((Y_{\cdot k})_{k\in\overline{\{j\}}}\right)\right] \\
&= A^c_{j\to m,\mathcal{I}_{-j}[0]} + \sum_{k\in\{m\}\cup\mathcal{I}_{-j}} A^c_{j\to m,\mathcal{I}_{-j}[k]} Y_{\cdot k} + \mathbb{E}\left[\zeta^c\middle|\left((Y_{\cdot k})_{k\in\overline{\{j\}}}\right)\right]
\end{aligned} \tag{31}$$

By taking the expectation of the left and right parts of the equality above given $\mathbb{E}[\epsilon_{\cdot k}] = 0$ for $\forall k \in \{m\} \cup \mathcal{I}_{-j}$, we have

$$Left = \mathbb{E}\left[\mathbb{E}\left[Y_{\cdot j}\middle|\left((Y_{\cdot k})_{k\in\overline{\{j\}}}\right)\right]\right] = \mathbb{E}\left[Y_{\cdot j}\right] = \alpha_j \tag{32}$$

$$\begin{aligned}
Right &= \mathbb{E}\left[A^c_{j\to m,\mathcal{I}_{-j}[0]} + \textstyle\sum_{k\in\{m\}\cup\mathcal{I}_{-j}} A^c_{j\to m,\mathcal{I}_{-j}[k]} Y_{\cdot k} + \mathbb{E}\left[\zeta^c\middle|\left((Y_{\cdot k})_{k\in\overline{\{j\}}}\right)\right]\right] \\
&= A^c_{j\to m,\mathcal{I}_{-j}[0]} + \textstyle\sum_{j'\in\mathcal{I}_{-j}} A^c_{j\to m,\mathcal{I}_{-j}[j']}\alpha_k + A^c_{j\to m,\mathcal{I}_{-j}[m]}\alpha_m + \mathbb{E}[\zeta^c]
\end{aligned} \tag{33}$$

Above two equalities are identical. So, we have

$$\hat{\alpha}_m = \frac{\alpha_j - A^c_{j\to m,\mathcal{I}_{-j}[0]} - \sum_{j'\in\mathcal{I}_{-j}} A^c_{j\to m,\mathcal{I}_{-j}[j']}\alpha_k}{A^c_{j\to m,\mathcal{I}_{-j}[m]}} \tag{34}$$

3.5. Estimation of Variance and Covariance

Let $Z = (Y_{\cdot k})_{k\in\overline{\{j\}}}$, for the variance $\Sigma_{\overline{\mathcal{M}},\overline{\mathcal{M}}}$, we have

$$\Sigma_{\overline{\mathcal{M}},\overline{\mathcal{M}}} = Var\left(Y_{\cdot j}\right) = \mathbb{E}\left[Var\left(Y_{\cdot j}\middle|Z\right)\right] + Var(\mathbb{E}[Y_{\cdot j}\Big|Z]). \tag{35}$$

Assumption 2 leads to $Var\left(Y_{\cdot j}\middle|Z\right) = Var\left(Y_{\cdot j}\middle|Z, \Omega_{\cdot m} = 1\right)$. According to the conditional variance for a Gaussian vector, we have

$$Var\left(Y_{\cdot j}\middle|Z\right) = Var\left(Y_{\cdot j}\right) - Cov\left(Z, Y_{\cdot j}\right) Var(Z)^{-1} Cov\left(Z, Y_{\cdot j}\right)^T. \tag{36}$$

Then, we have the first term of $Var\left(Y_{\cdot j}\right)$ as

$$\mathbb{E}\left[Var\left(Y_{\cdot j}\middle|Z\right)\right] = Var\left(Y_{\cdot j}\right) - Cov\left(Z, Y_{\cdot j}\right) Var(Z)^{-1} Cov\left(Z, Y_{\cdot j}\right)^T \Big|\Omega_{\cdot m} = 1 \tag{37}$$

For the second term of $Var(Y_{\cdot j})$, we have

$$\begin{aligned} &Var\big(\mathbb{E}\big[Y_{\cdot j}|Z\big]\big) = Var\big(\mathbb{E}\big[Y_{\cdot j}|Z,\Omega_{\cdot m}=1\big]\big) \\ &= Var\Big(\textstyle\sum_{k\in\{m\}\cup\mathcal{I}_{-j}} A^{c}_{j\to m,\mathcal{I}_{-j}[k]} Y_{\cdot k} - \sum_{k\in\{m\}\cup\mathcal{I}_{-j}} A^{c}_{j\to m,\mathcal{I}_{-j}[k]}\mathbb{E}[\epsilon_{\cdot k}|Z] + A^{c}_{j\to m,\mathcal{I}_{-j}[0]} + \mathbb{E}\big[\epsilon_{\cdot j}\big]\Big) \end{aligned} \tag{38}$$

where $\mathbb{E}[\epsilon_{\cdot k}|Z] = \sigma^2\big(Var(Z)^{-1}\big)_{k\cdot}(Z - \mathbb{E}[Z])$.

For the covariances $\hat{\Sigma}_{m,\overline{\mathcal{M}}}$ between $Y_{\cdot m}$, $Y_{\cdot k}$, $k \in \mathcal{I}$, let $Z = (Y_{\cdot l})_{\,l\in\overline{\{\,k\,\}}}$, we have

$$\begin{aligned} \hat{\Sigma}_{m,\overline{\mathcal{M}}} &= Cov(Y_{\cdot m}\, Y_{\cdot k}) = \mathbb{E}[Y_{\cdot m}\, Y_{\cdot k}] - \mathbb{E}[Y_{\cdot m}]\mathbb{E}[\,Y_{\cdot k}] \\ &= \mathbb{E}[\mathbb{E}[Y_{\cdot m}\, Y_{\cdot k}|Z]] - \mathbb{E}[Y_{\cdot m}]\mathbb{E}[\,Y_{\cdot k}] \\ &= \mathbb{E}[Y_{\cdot m}\mathbb{E}[Y_{\cdot k}|Z]] - \mathbb{E}[Y_{\cdot m}]\mathbb{E}[\,Y_{\cdot k}]. \end{aligned} \tag{39}$$

For the first term, we have

$$\begin{aligned} \mathbb{E}[Y_{\cdot m}\mathbb{E}[Y_{\cdot k}|Z]] &= \mathbb{E}[Y_{\cdot m}\mathbb{E}[Y_{\cdot k}|Z,\Omega_{\cdot m}=1]] \\ &= \mathbb{E}\left[Y_{\cdot m}\left(A^{c}_{j\to m,\mathcal{I}_{-j}[0]} + \sum_{j'\in\{m\}\mathcal{I}_{-j}} A^{c}_{j\to m,\mathcal{I}_{-j}[j']} Y_{\cdot j'} + \mathbb{E}(\zeta^{c}|Z)\right)\right] \\ &= A^{c}_{j\to m,\mathcal{I}_{-j}[0]}\mathbb{E}[Y_{\cdot m}] + A^{c}_{j\to m,\mathcal{I}_{-j}[m]}\mathbb{E}\big[Y^2_{\cdot m}\big] + \sum_{j'\in\mathcal{I}_{-j}} A^{c}_{j\to m,\mathcal{I}_{-j}[j']}\mathbb{E}\Big[Y_{\cdot m}Y_{\cdot j'}\Big] + \mathbb{E}[Y_{\cdot m}\mathbb{E}(\zeta^{c}|Z)] \end{aligned} \tag{40}$$

According to the derivation in [22], $\mathbb{E}[Y_{\cdot m}\mathbb{E}(\zeta^{c}|Z)]$ is calculated as

$$-\sigma^2\Big(\sum_{l\in\mathcal{I}_{-k}}\sum_{s\in\mathcal{I}_{-k}} Var(Z)^{-1}A^{c}_{j\to m,\mathcal{I}_{-j}[l]}Cov(Y_{\cdot m}Y_{\cdot l}) + A^{c}_{j\to m,\mathcal{I}_{-j}[m]}\Big). \tag{41}$$

Note that for the second term, $\mathbb{E}[Y_{\cdot m}]\mathbb{E}[\,Y_{\cdot k}]$, it can be directly calculated.

4. Experiment

4.1. Dataset and Preprocessing

We utilize a road traffic speed dataset published by [23]. Road segments are anonymous, covering the main urban expressways within two months from 1 August to 30 September 2016, (a total of 61 days).

The time interval is 10 min. From the original dataset, we select twenty links whose speed are generated in the morning rush hours (i.e., 7:00 A.M. to 9:00 A.M.) for evaluating the proposed method. The speed of each link is transformed to the congestion index, which is calculated as $v_{ij}/\max(v_{\cdot j})$, v_{ij} denotes the i speed value of link j and $v_{\cdot j}$ denotes all observations of link j. For each link j, the time series length of speed observations is 732 (12 observations in two hours 61 days). Hence, the dimension of Y is $n = 732$, p is the number of links.

The basic assumption of the proposed model is that the observations of each link are drawn from a Gaussian distribution, Hence, we adopt the quantile–quantile plot (QQ Plot) to display the quantiles of the data (after normalizing) versus the theoretical quantile values from a normal distribution. If the distribution of the data is normal, then the data plot is linear. As shown in the Figure 1, the plot closely follows the straight lines, suggesting that the data after normalizing the congestion data have an approximately normal distribution.

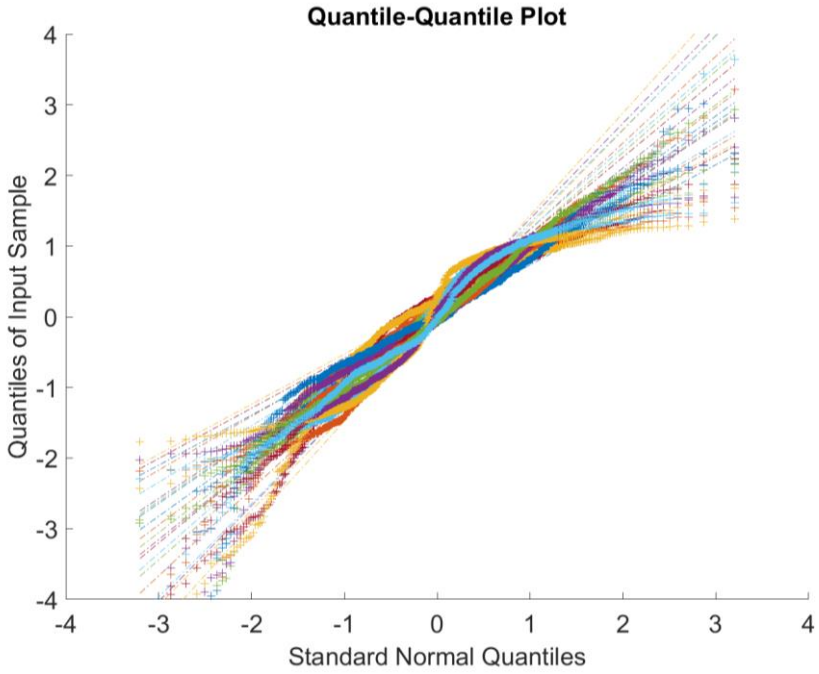

Figure 1. Plot of the Data Quantitles and Standard Normal Quantitles.

4.2. Metrics for Missing Data Imputation Accuracy

For evaluating the performance of missing data imputation, we adopt the below four metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), and R^2. Note that a higher R^2 value denotes better accuracy.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

4.3. Benchmark and Experiment Settings

We compare the new model with the typically used PPCA model, where σ^2 and A is estimated by an expectation-maximum (EM) algorithm. We name it ppca-em in this section. We further use the estimated σ^2 by EM as the known inputs of the new model in this study. As to the rank r in the model, the best value is determined by cross-validation on the dataset. In this section, we further detail the experiment settings in terms of the MNAR data generation and the settings of link set $\mathcal{M}$.

4.3.1. Generating MNAR

Note that the model targets at solving the imputation for MNAR data. We utilize the mechanism of generating MNAR in [22]. Specifically, a logistic regression function is adopted as $f(x) = 1/(-a(x-b))$, where x is an observation, and (a, b) is set for selecting different missing percentage. The function transforms the observation x to a value in $(0, 1)$. The observation x with $f(x) > \mu$, is set to be the MNAR data, where μ is a random threshold. We set the parameters (a, b) as below Table 1, which is corresponding to a specific missing percentage.

Table 1. Settings for Generating MNAR Data in the Experiments.

a	b	Missing Percentage
−1	−1.3	25%
3	0	50%
1	−1.3	75%

4.3.2. Settings of Link Set $\mathcal{M}$

Missing data on different links may obtain different recover accuracy, even with the same missing percentage. For evaluating this proposition, we first test the missing data imputation accuracy with different p-score values of the links. When a link observation $Y_{\cdot j}$ is set to be $\mathcal{M} = \{j\}$, all other links are set to be $\mathcal{I}_{-j} = \mathcal{I} \backslash \{j\}$, where $\mathcal{I} = \{1, 2, \ldots, 20\}$. Further, we test the missing data imputation accuracy of several select links (or link combination) compared with the ppca-em model, to evaluate the advantage of the new model.

4.4. Results and Analysis

We first examine the relationship between the missing data imputation accuracy and the proposed metric, i.e., p-score value. We select two links with the highest p-score value and the lowest p-score value in the dataset. The p-score values of two selected links, i.e., link 6 and link 18, are shown in Figure 2, where the color map denotes the ρ values between the selected link and all links in link set $\mathcal{I}$.

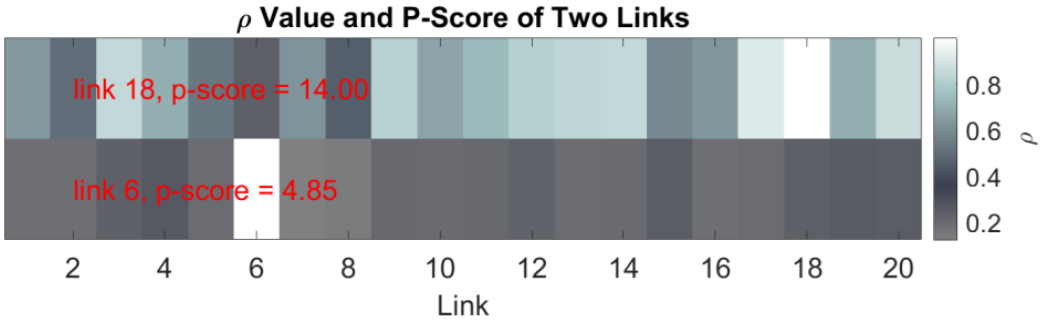

Figure 2. ρ Value of Two Links with the Highest and Lowest p-score.

Accordingly, we calculate the absolute errors of the model on these two selected links. Figure 3 shows the results missing data imputation results on these links in terms of different missing data percentages. We can see that missing data on link 18, which is with a higher p-score value than that of link 6, are better recovered regarding all scenarios of missing data percentages (25%, 50%, 75%).

We further examine the relationship between the p-score value and the accuracy metrics on the traffic dataset, which is shown in below Figure 4. The accuracy measured by four metrics presents a positive correlation with the p-score value on different links, meaning that missing data on the links with higher p-score values can be better recovered.

We also compare the new model with the ppca-em model on other links or link combinations in the dataset. The settings and corresponding missing data imputation results measured by the four metrics are shown in Table 2. Except the above four metrics, we further added the accuracy as another metric for directly representing the estimation accuracy results and better understanding the accuracy comparison between the new model and the baseline. Here, the accuracy is calculated as

$$\text{Accuracy} = 1 - 100\% * \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$

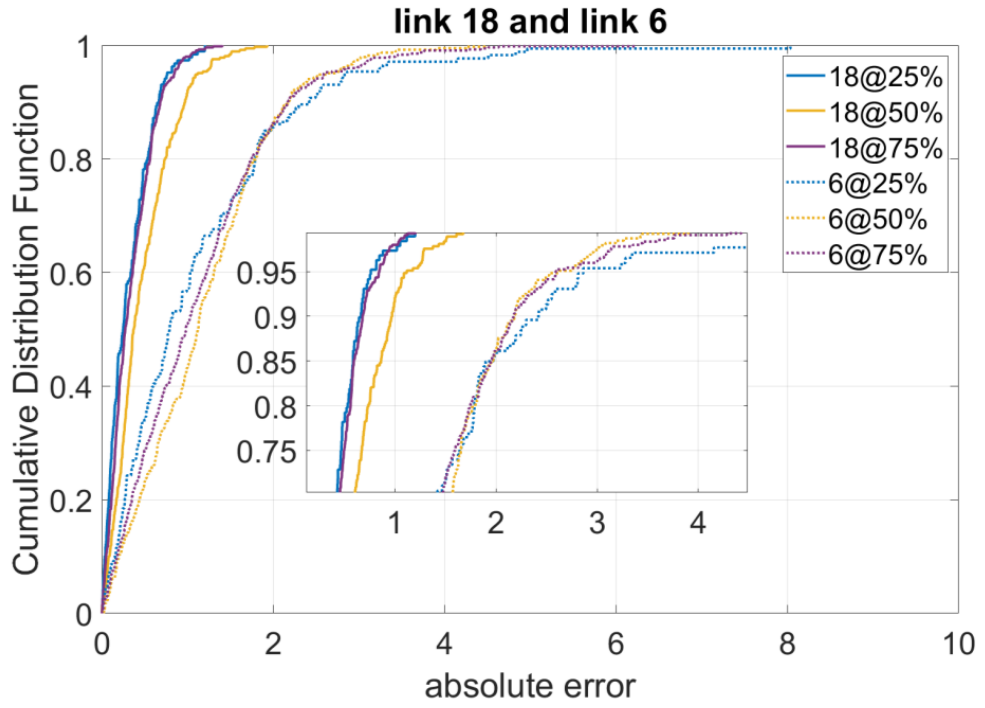

Figure 3. Performance of the Algorithm for Links with Highest p-Score and Lowest p-Score.

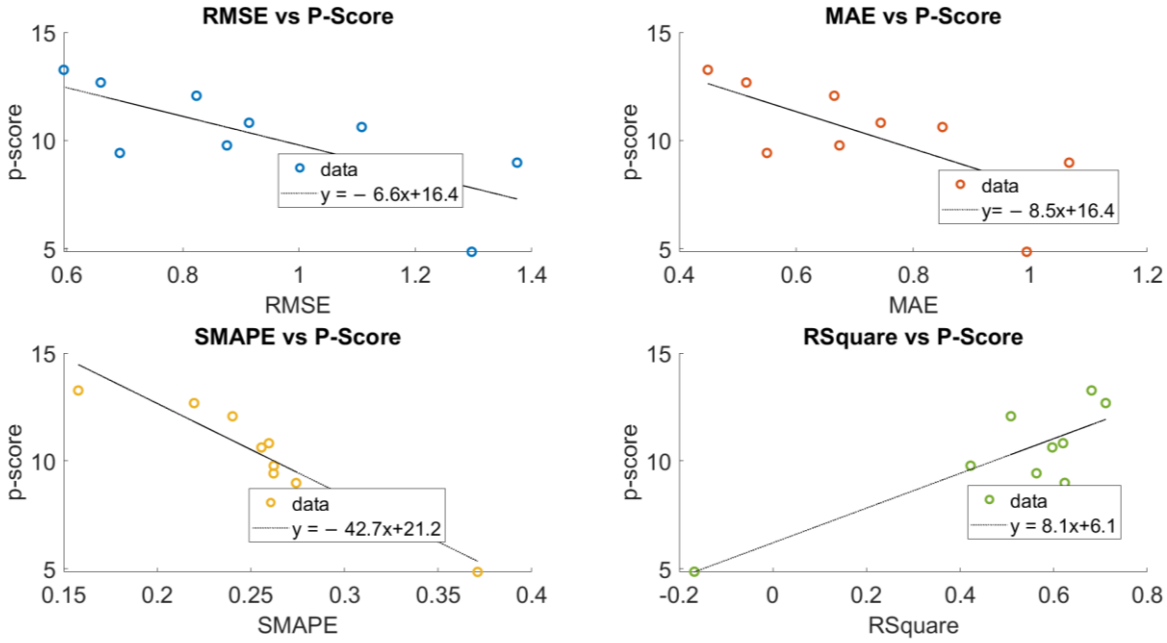

Figure 4. Scatter Plot between the Accuracy Metrics and the p-score Values.

Table 2. Experiment Setting and performance of the algorithms with different Percent of MNAR Data on Links.

Experiment Setting: Missing Rate (%) @ $\mathcal{M}$										
	50 @{1}		50 @{3}		75 @{1}		75 @{1,3}		75 @{3,5}	
p-score	10.62@{1}		13.26@{3}		10.62@{1}		–		9.42@{5}	
Performance Comparison										
Metrics	ppca-em	New	ppca-em	New	ppca-em	New	ppca-em	New	ppca-em	New
RMSE	0.992	0.746	0.559	0.595	1.069	0.746	0.835	0.871	0.942	0.627
MAE	0.810	0.564	0.458	0.448	0.789	0.564	0.598	0.625	0.665	0.468
SMAPE	0.340	0.223	0.216	0.157	0.289	0.223	0.231	0.228	0.253	0.201
R^2	0.150	0.688	0.595	0.681	0.545	0.688	0.208	0.677	0.115	0.740
Accuracy	83.0%	88.9%	89.2%	92.2%	85.5%	88.9%	88.4%	88.6%	87.3%	89.9%
Computing Time										
Sec	6.54	2.03	6.29	2.03	6.73	2.64	6.06	4.06	11.32	4.11

Note that Figure 3 already shows that the new model obtains the worst accuracy on link 6. Hence, we further compare two models on this link to compare the new model with the ppca-em model. The results are shown in Figure 5. It shows that even on link 6, the absolute errors of the new model are still lower than the ppca-em model for three missing ratios.

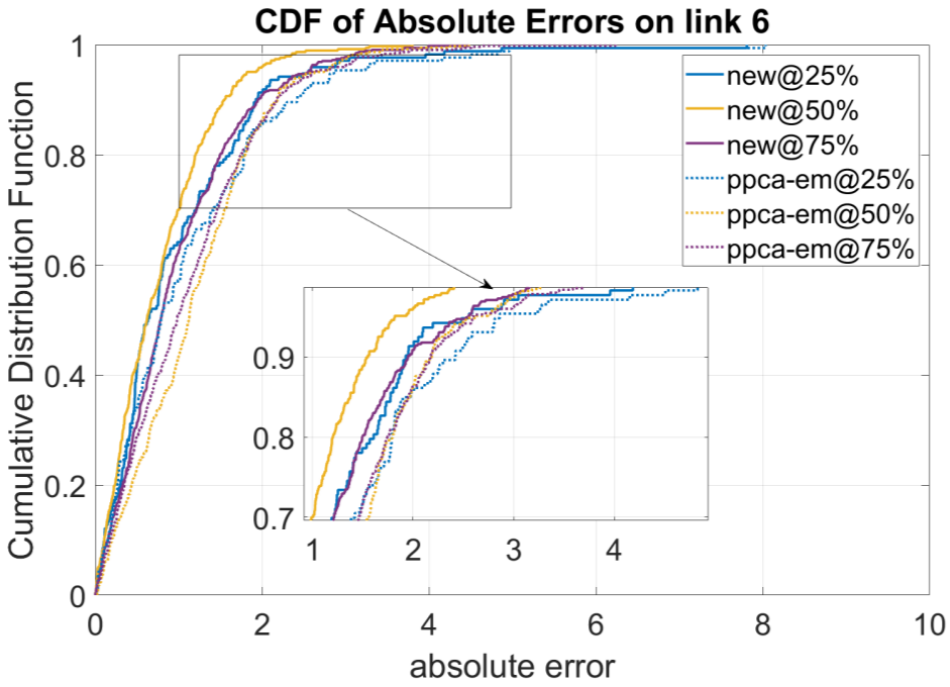

Figure 5. Performance of Models on the Link with Lowest p-Score Values.

The experiment results in Table 2 and Figure 5 demonstrate that the new model performs better than the typically used ppca-em model in terms of four accuracy metrics and computing time. The results indicate that the new model is more effective and efficient for the MNAR traffic data imputation problem, which is the target of this study. The typical ppca-em method is usually used for imputation of data missing at random, whereas the new model is more general and is capable of MNAR data imputation.

5. Discussion

Our improved linear probabilistic principal component analysis method can be applied to a variety of missing traffic data imputation applications, such as missing traffic speed estimation, or other traffic indicators. Notably, because the proposed missing data imputation method is a linear and interpretable model, which is naturally of high computing efficiency, it can be utilized in the systems where real-time missing data estimation is required. Additionally, the time-series based metric, p-Score value, is proposed to distinguish variables, e.g., links with missing traffic speed data, for estimating the missing values. Such a method can be applied to the applications of traffic surveillance systems to identify which sensors should be of high priority to maintained in the systems to ensure the full surveillance, or which links should be equipped with sensor for traffic surveillance.

6. Conclusions

In this study, we propose a general linear model based on the PPCA to tackle the MNAR traffic data imputation problem. We also propose a time series-based metric, i.e., the p-score, to distinguish links that are of missing data. Experimental results on a real-world traffic dataset show that the proposed model performs better than the typically used ppca-em model in terms of missing data imputation accuracy and computing time. Furthermore, we test the model on links with different p-score values. The experiment results show that the missing data on links with higher p-score values are better recovered. Such an observation helps us understand the data recovering distinction for different links in the road network, which has not been studied in any research to our best knowledge. In future work, we will further compare the model with other methods on more traffic datasets.

Author Contributions: Conceptualization, L.H. and R.S.; methodology, L.H.; software, Z.L.; validation, R.L.; formal analysis, R.S.; investigation, L.H.; resources, R.S.; data curation, Z.L.; writing—original draft preparation, L.H.; writing—review and editing, L.H. and R.S.; visualization, L.H.; supervision, R.S.; project administration, R.S.; funding acquisition, R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported under the RIE2020 Industry Alignment Fund—Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s), and A*STAR under its Industry Alignment Fund (LOA Award I1901E0046).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this paper is published by [23], which can be found at https://doi.org/10.5281/zenodo.1205229, Access on 18 February 2022.

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

References

1. Yuan, H.; Li, G. A survey of traffic prediction: From spatio-temporal data to intelligent transportation. *Data Sci. Eng.* **2021**, *6*, 63–85. [CrossRef]
2. Neelakandan, S.; Berlin, M.A.; Tripathi, S.; Devi, V.B.; Bhardwaj, I.; Arulkumar, N. IoT-based traffic prediction and traffic signal control system for smart city. *Soft Comput.* **2021**, *25*, 12241–12248. [CrossRef]
3. Tan, H.C.; Wu, Y.K.; Feng, J.S.; Wang, W.H.; Ran, B. Traffic missing data completion with spatial-temporal correlations. In Proceedings of the 93rd Annual Meeting of the Transportation Research Board, Washington, DC, USA, 12–16 January 2014.
4. Li, H.P.; Wang, Y.H.; Li, M. Modified GAN Model for Traffic Missing Data Imputation. In *CICTP 2020, Proceedings of the 20th COTA International Conference of Transportation Professionals, Xi'an, China, 14–16 August 2020*; American Society of Civil Engineers: Reston, VA, USA, 2020; pp. 3013–3023.
5. Yang, F.; Liu, G.; Huang, L.; Chin, C.S. Tensor Decomposition for Spatial—Temporal Traffic Flow Prediction with Sparse Data. *Sensors* **2020**, *20*, 6046. [CrossRef] [PubMed]
6. Huang, L.P.; Zhao, S.D.; Luo, R.K.; Su, R.; Sindhwani, M.; Chan, S.K.; Dhinesh, G.R. An incremental map matching approach with speed estimation constraints for high sampling rate vehicle trajectories. In Proceedings of the IEEE 17th International Conference on Control & Automation (ICCA), Naples, Italy, 27–30 June 2022; pp. 758–765.

7. Huang, L.P.; Yang, Y.J.; Chen, H.C.; Zhang, Y.; Wang, Z.; He, L. Context aware road travel time estimation by coupled tensor decomposition based on trajectory data. *KBS* **2022**, *245*, 108596. [CrossRef]
8. Huang, L.; Li, Z.; Zhao, S.; Luo, R.; Su, R.; Guan, Y. Coupling Urban Road Travel Time and Traffic Status from Vehicle Trajectories by Gaussian Distribution. In Proceedings of the IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 4056–4061.
9. Huang, L.P.; Yang, Y.J.; Zhao, X.H.; Ma, C.; Gao, H. Sparse data-based urban road travel speed prediction using probabilistic principal component analysis. *IEEE Access* **2018**, *6*, 44022–44035. [CrossRef]
10. Asif, M.T.; Mitrovic, N.; Garg, L.; Dauwels, J.; Jaillet, P. Low-dimensional models for missing data imputation in road networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013.
11. Jia, X.; Dong, X.; Chen, M.; Yu, X. Missing data imputation for traffic congestion data based on joint matrix factorization. *Knowl.-Based Syst.* **2021**, *225*, 107114. [CrossRef]
12. Asif, M.T.; Mitrovic, N.; Dauwels, J.; Jaillet, P. Matrix and tensor-based methods for missing data estimation in large traffic networks. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 1816–1825. [CrossRef]
13. Jiang, B.; Siddiqi, M.D.; Asadi, R.; Regan, A. Imputation of missing traffic flow data using denoising autoencoders. *Procedia Comput. Sci.* **2021**, *184*, 84–91. [CrossRef]
14. Shang, Q.; Yang, Z.; Gao, S.; Tan, D. An imputation method for missing traffic data based on FCM optimized by PSO-SVR. *J. Adv. Transp.* **2018**, *2018*, 2935248. [CrossRef]
15. Li, Y.B.; Li, Z.H.; Li, L. Missing traffic data: Comparison of imputation methods. *IET Intell. Transp. Syst.* **2018**, *8*, 51–57. [CrossRef]
16. Wu, P.; Xu, L.; Huang, Z. Imputation methods used in missing traffic data: A literature review. In Proceedings of the International Symposium on Intelligence Computation and Applications, Guangzhou, China, 20–21 November 2019.
17. Chen, X.; Lei, M.; Saunier, N.; Sun, L. Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 12301–12310. [CrossRef]
18. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **1999**, *61*, 611–622. [CrossRef]
19. Ilin, A.; Raiko, T. Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res.* **2010**, *11*, 1957–2000.
20. Audigier, B.; Husson, F.; Josse, J. Multiple imputation for continuous variables using a Bayesian principal component analysis. *J. Stat. Comput. Simul.* **2016**, *86*, 2140–2156. [CrossRef]
21. 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.
22. Sportisse, A.; Boyer, C.; Josse, J. Estimation and imputation in probabilistic principal component analysis with missing not at random data. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7067–7077.
23. Chen, X.; Yang, J.; Sun, L. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation. *Transp. Res. Part C Emerg. Technol.* **2020**, *117*, 102673. [CrossRef]