mathematics

MDPI

Article

A Novel Truncated Normal Tensor Completion Method for Multi-Source Fusion Data

Yongmei Zhao [1,2]

[1] School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China; zhaoyongmei@mail.nwpu.edu.cn
[2] School of Equipment Management and Unmanned Aerial Vehicle Engineering, Air Force Engineering University, Xi'an 710051, China

Abstract: Completing traffic data is a basic requirement for intelligent transportation systems. However, completing spatiotemporal traffic data poses a significant challenge, especially for high-dimensional data with complex missing mechanisms. Various completion methods targeting different missing mechanisms have showcased the superiority of tensor learning by effectively characterizing intricate spatiotemporal correlations. In this study, a novel tensor completion framework, known as the multi-source tensor completion method for data fusion, is proposed. This framework incorporates passenger transfer relationships between buses and subways into subway data completion, enhancing the data completion accuracy. Moreover, by combining bus transfer passenger flow data with other data dimensions, such as the different road sections, time intervals, and days, an innovative 4D low-rank tensor completion data framework was obtained. In addition, to boost the completion accuracy, a truncated $\ell_{2,p}$ norm optimization model was derived. This model enhances the non-convex performance of the objective function throughout the tensor completion process. The results highlight the superiority of the proposed completion method, leveraging fused subway/bus data over other completion methods that rely solely on subway data.

Keywords: low-rank tensor completion; missing data; intelligent transportation system; data fusion

MSC: 14N07

Citation: Zhao, Y. A Novel Truncated Normal Tensor Completion Method for Multi-Source Fusion Data. *Mathematics* **2024**, *12*, 223. https://doi.org/10.3390/math12020223

Academic Editors: Bo Wang and Faheim Sufi

Received: 4 December 2023
Revised: 4 January 2024
Accepted: 7 January 2024
Published: 9 January 2024

1. Introduction

In data-driven applications, robust and accurate data completion is of great importance [1]. Data fusion, achieved by combining data from multiple sources, can decrease the uncertainty of results [2–5]. For instance, Roy et al. improved vehicle tracking and detection performance under non-line-of-sight (NLOS) image and non-image modalities by combining multiple state-of-the-art fusion strategies [6]. Moreover, Senel et al. fused data on the object list level from distributed vehicle sensors using multiple sensors cooperatively [7]. In addition, Chen et al. fused traffic flow, occupancy, and recreational vehicle speed data and dynamically trained the fused data using a dynamic backpropagation (PB) fusion method [8]. Moreover, Lin et al. improved road traffic speed prediction by fusing traditional speed sensor data with cross-domain novel "sensor" data [9]. As for Zißner et al., they introduced the data fusion on intelligent traffic system (DataFITS), which collected four data types from seven sources over nine months and fused them in a spatiotemporal domain, leading to significant improvements in road coverage [10]. The results showed that DataFITS significantly increased the road coverage by 137% and improved the road information quality by up to 40%. Moreover, these studies demonstrated that the multi-source data could provide more information from different dimensions, thereby enhancing the quality of the data analysis.

Considering that traffic data include various multi-source and heterogeneous data, more data-fusion methods are required. Currently, classical data-fusion methods include

Kalman filtering, Bayesian networks, Dempster–Shafer (D–S) evidence theory, neural networks (NNs), fuzzy theory, and voting methods. These techniques have already been applied in vehicle positioning, traffic event recognition, traffic event prediction, and traffic guidance in the transport sector. Referring to existing studies, Kalman filtering and D–S evidence theory are more commonly used. In recent years, with the advancement of artificial intelligence (AI), some machine learning methods, such as recurrent neural networks, convolutional neural networks, and graph neural networks, have also been widely applied in traffic analysis, achieving high success rates in predicting traffic flow [11].

For instance, Ji et al. encoded each feature in the spatial and temporal dimensions and considered data directly related to flow as the inputs for a graph convolutional network (GCN) [12]. Furthermore, Satish et al. proposed a multidimensional k-means++ data-fusion method, merging attribute values into a single or very few dimensions [13]. In addition, Zhang et al. introduced a data-fusion method based on Bayesian fusion rules [14]. Considering the relationships between multiple traffic factors from a system perspective, this method fuses traffic speeds from different data sources, based on prior probabilities, while using the high-order multivariate Markov model. Nevertheless, the long training time of NNs, the model-building complexity, and the dimensionality issues are still difficulties existing within fusion methods [15].

Moreover, a tensor serves as a natural mathematical representation of high-dimensional data. Drawing inspiration from compressive sensing theory, which considers sparse structures in tensors, accurate recovery of unknown information can be achieved in a small number of observations if the target tensor exhibits a low-rank structure or can be well approximated by a low-rank tensor. Recently, the tensor completion theory has shown promising applications in traffic data completion, serving as a theoretical foundation for the accurate achievement of multidimensional, multimodal, and multi-structured traffic data.

Recently, low-dimensional fusion of multimodal data has attracted great attention. Tensor-based decomposition is able to map multimodal data into a unified low-dimensional subspace, enabling clustering, classification, and other data-fusion analyses within that subspace [16]. For instance, Pan et al. utilized the tensor nuclear norm, derived from tensor singular value decomposition (t-SVD), to optimally leverage high-dimensional information, thereby improving the precision of multi-view data processing [17]. Shen et al. introduced a comprehensive method for analyzing and processing multi-source traffic data [18]. Their method incorporated data-analysis techniques rooted in spatiotemporal regression models and data-fusion methods based on confidence tensors and evidence theory. Moreover, Xing et al. established a novel tensor decomposition data-fusion framework, demonstrating that the method combining license-plate recognition (LPR) data and cellphone location (CL) data yielded significantly better results compared with the interpolation methods using only LPR data [19]. In addition, Respati et al. proposed an adaptive weight-based fusion technique (ABAFT), deploying data space coverage and quality or confidence as factors for building weights [20]. The completion methods for the different missing mechanisms demonstrated the superiority of the tensor learning method by effectively characterizing complicated spatiotemporal correlations [21]. Therefore, in this study, based on the tensor representation of high-dimensional data, subway data and bus data were fused to achieve efficient multi-source data completion.

In summary, the contributions of this study are as follows:

(1) Subway inbound passenger flow and bus transfer time-series data were constructed, and the completion performance of single-source data containing only subway data was compared to that of multi-source data containing subway and bus transfer data;
(2) A truncated $\ell_{2,p}$ norm optimization model was derived to enhance the non-convex performance of the objective function during the tensor completion process;
(3) A multi-source data-fusion tensor completion framework was established, and ADMM was applied to achieve an effective completion algorithm.

2. Characteristics of Multi-Source Traffic Data

In practical transportation, complex traffic data from distinct sources and with different structures are present. For example, data gathered from road traffic speed, subway passenger flow, bus passenger flow, shared bike-riding, and cellphones are combined together. The data structure varies significantly among the different types of transportation,. Even for the same type of transportation, such as buses, the data structure differs due to varying data sources and collection focuses (refer to Table 1). Given that the transportation system is a complex and dynamically interactive network, additional data fusion is required to effectively capture global interactions within the system.

Table 1. Structure of multi-source bus data [18].

Data Source	Key Fields
Passenger card swiping data	IC card no., card swiping time, license plate no., and route no.
Vehicle station announcement data	License plate no., route no., station name, station inbound time, and station outbound time
Vehicle coordinates and route tables	Route name, direction, station list, station coordinates, and route trajectory
Vehicle dispatch records	Vehicle dispatch record, departure schedule, license plate no., departure time, and stops

3. Low-Rank Completion Model of Truncated $\ell_{2,p}$

3.1. Truncated $\ell_{2,p}$ Norm (P2TN)

As an alternative approach to solving rank-minimization problems, the truncated $\ell_{2,p}$ norm incorporates the advantages of the truncated nuclear norm and the $\ell_{2,p}$ norm. Moreover, it is more compact than a single norm during the approximation process of rank-minimization problems. Therefore, the truncated $\ell_{2,p}$ norm preserves the complete features of large singular values in a truncated manner and applies $\ell_{2,p}$ contraction only to the singular values of $\min(n_1, n_2) - r$ after achieving the truncation. During the contraction, smaller singular values incur larger threshold penalties, ensuring the retention of a sufficient amount of internal strong correlation information to the maximum extent. Thus, the $\ell_{2,p}$ norm is defined as follows [22]:

Given a matrix $X \in \mathbb{R}^{m\times n}$, for any $p \in \mathbb{R}$,

$$\|X\|_p^2 = \frac{1}{2}\left(\sum_{i=1}^{\min(m,n)} |\sigma_i|^{\frac{p-2}{2}}\right) \tag{1}$$

where $\| \cdot \|_p^2$ refers to the $\ell_{2,p}$ norm of the matrix. Assuming that the singular value decomposition of matrix X is $U\Sigma V^T$, then one can write $\Sigma = Diag(\sigma_i(X))$, where σ_i represents the singular value of the matrix, i.e., the nuclear norm of the matrix. Moreover, the singular values are ranked in descending order as $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_{\min}\{m, n\} \geqslant 0$, and p indicates the contraction parameter.

Then, the $\ell_{2,p}$ norm is expanded to the tensor mode. For a given $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times n_3}$, $\ell_{2,p}$ can be expressed as follows [23]:

$$\|\mathcal{X}\|_p^2 = \frac{1}{n_3}\sum_{k=1}^{n_3} \| \overline{X}^{(i)} \|_p^2 = \frac{1}{n_3}\sum_{k=1}^{n_3}\sum_{i=1}^{\min\{n_1,n_2\}} |\overline{\sigma}_i|^{\frac{p-2}{2}} \tag{2}$$

According to the tensor singular value decomposition principle, a Fourier transform is performed for X along the third dimension, n_3. In more detail, $\overline{X}^{(i)}$ refers to the forward slice after transformation in matrix form, while $|\overline{\sigma}_i|$ indicates the singular value obtained by singular value decomposition for the forward slice $\overline{X}^{(i)}$, e.g., the tensor nuclear norm.

The truncated nuclear norm of the tensor is defined as follows [24]:

Given a third-order tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and a positive integer $t = \min(n_1, n_2)$, the truncated nuclear norm represents the sum of the minimum singular values of the tensor:

$$\begin{aligned} \| \mathcal{X} \|_{r,*} &= \frac{1}{n_3} \| \overline{X} \|_{r,*} \\ &= \frac{1}{n_3} \sum_{j=1}^{n_3} \sum_{i=r+1}^{t} \sigma_i\left(\overline{X}^{(j)}\right) \\ &= \frac{1}{n_3} \sum_{j=1}^{n_3} \sum_{i=1}^{t} \sigma_i\left(\overline{X}^{(j)}\right) - \frac{1}{n_3} \sum_{j=1}^{n_3} \sum_{i=1}^{r} \sigma_i\left(\overline{X}^{(j)}\right) \\ &= \sum_{i=1}^{t} \sigma_i(\mathcal{X}) - \sum_{i=1}^{r} \sigma_i(\mathcal{X}) \end{aligned} \tag{3}$$

where $\| \cdot \|_{r,*}$ refers to the truncated nuclear norm of the tensor, $\sigma_i(\mathcal{X})$ refers to the i-th singular value of $\mathcal{X}$, and the following constraint is to be respected: $r < \min(n_1, n_2)$.

Combining both tensor norm definitions, the truncated $\ell_{2,p}$ norm (P2TNN) of the tensor is defined as follows:

$$\| \mathcal{X} \|_{p,r}^2 = \frac{1}{n_3} \sum_{j=1}^{n_3} \sum_{i=r+1}^{t} |\overline{\sigma}_i|^{\frac{p-2}{2}} \tag{4}$$

where $\| \cdot \|_{p,r}^2$ refers to the truncated $\ell_{2,p}$ norm, and in the tensor mode, for any given $\mathcal{X}$, $\mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $\mu > 0$, $p \in \mathbb{R}$, and $r < \min(n_1, n_2)$.

In the context of a given parameter τ, for a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, the tensor singular value decomposition (t-SVD) can be expressed as $\mathcal{X} = \mathcal{U}\mathcal{S}\mathcal{V}^\top$. The singular value threshold for the tensor $\mathcal{X}$ is consequently defined.

$$\mathcal{D}_\tau(\mathcal{A}) = \mathcal{U}\mathcal{S}_\tau\mathcal{V}^\top \tag{5}$$

The contraction threshold operator is defined for the truncated $\ell_{2,p}$ as $\mathcal{D}_{p,r}^{\mu}(.)$. Then, the following equation is obtained:

$$\mathcal{D}_{p,r}^{\mu}(\mathcal{Y}) = \underset{\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}}{\operatorname{argmin}} \| \mathcal{X} \|_{p,r}^2 + \frac{1}{2\mu} \| \mathcal{X} - \mathcal{Y} \|_F^2 \tag{6}$$

According to the t-SVD theorem, $\mathcal{D}_{p,r}^{\mu}(\mathcal{Y})$ represents the solution to the $\mathcal{X}$ optimization problem, μ denotes the threshold parameter for singular value thresholding decomposition, and p indicates the contraction parameter. When $-\infty < p < 1$, the larger the truncated $\ell_{2,p}$ contraction threshold, the smaller the received penalty. Moreover, $\| \cdot \|_F$ refers to the Frobenius norm. Regarding $\mathcal{D}_{p,r}^{\mu}(\mathcal{Y})$, we have $\mathcal{D}_{p,r}^{\mu}(\mathcal{Y}) = \mathcal{U}\mathcal{S}_r(\Sigma)\mathcal{V}^T$; therefore, one can write

$$\begin{aligned} \left(\mathcal{D}_{p,r}^{\mu}(\Sigma)\right)_{ijk} &= \mathcal{D}_{p,r}^{\mu}(\Sigma_{ijk}) \\ &= Diag\left(\left(\begin{array}{c} \sigma_1, \cdots, \sigma_r, \left[\sigma_{r+1} - \mu|\sigma_{r+1}|^{\frac{p-2}{2}}\right]_+ \\ , \cdots, \left[\sigma_{\min\{m,n\}} - \mu\left|\sigma_{\min\{m,n\}}\right|^{\frac{p-2}{2}}\right]_+ \end{array}\right)^\top\right) \end{aligned} \tag{7}$$

For any value $a \in \mathbb{R}$, the addition function $(.)_+$ is defined as $a_+ = \max\{a, 0\}$.

3.2. LRTC-P2TN Model

The NWLRTC model [25] was further optimized using the $\ell_{2,p}$ norm to develop the LRTC-P2TN model:

$$\begin{aligned} &\min_{\mathcal{X} \geq 0, \mathcal{B} \geq 0, \mathcal{Z} \geq 0, \mathcal{W} \geq 0} : \alpha_1 \| \mathcal{B} \|_{p,r}^2 + \alpha_2 \| \mathcal{Z} \|_{p,r}^2 + \alpha_3 \| \mathcal{W} \|_{p,r}^2 \\ &\text{s.t.} : \mathcal{P}_\Omega(\mathcal{X}_i) = \mathcal{P}_\Omega(\mathcal{M}_i) \\ &\quad \mathcal{X}_1 = \mathcal{B}, \mathcal{X}_2 = \mathcal{Z}, \mathcal{X}_3 = \mathcal{W} \end{aligned} \tag{8}$$

where $\mathcal{B} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, $\mathcal{Z} \in \mathbb{R}^{n_2 \times n_3 \times n_1}$, and $\mathcal{W} \in \mathbb{R}^{n_3 \times n_1 \times n_2}$ refer to the spatiotemporal traffic data input in three different directions, whereas the truncated nuclear norm of the tensor is introduced to avoid dependencies on the input direction. Moreover, α_1, α_2, and α_3 represent the weights of each direction.

Equation (8) denotes a convex optimization problem that incorporates equality constraints. To convert this problem into one without such constraints, the augmented Lagrangian function is employed. The definition of the augmented Lagrangian function, which takes into account only the $\mathcal{X}_1 = \mathcal{B}$, $\mathcal{X}_2 = \mathcal{Z}$, $\mathcal{X}_3 = \mathcal{W}$ constraint conditions, is formulated as follows:

$$\begin{aligned} &\mathcal{L}(\mathcal{X}, \mathcal{B}, \mathcal{Z}, \mathcal{W}, \mathcal{Y}) \\ &= \alpha_1 \parallel \mathcal{B} \parallel_{p,r}^2 + \alpha_2 \parallel \mathcal{Z} \parallel_{p,r}^2 + \alpha_3 \parallel \mathcal{W} \parallel_{p,r}^2 \\ &+ < \mathcal{Y}_1, \mathcal{X}_1 - \mathcal{B} > + \frac{\rho}{2} \parallel \mathcal{X}_1 - \mathcal{B} \parallel_F^2 + < \mathcal{Y}_2, \mathcal{X}_2 - \mathcal{Z} > \\ &+ \frac{\rho}{2} \parallel \mathcal{X}_2 - \mathcal{Z} \parallel_F^2 + < \mathcal{Y}_3, \mathcal{X}_3 - \mathcal{W} > + \frac{\rho}{2} \parallel \mathcal{X}_3 \\ &- \mathcal{W} \parallel_F^2 \end{aligned} \tag{9}$$

where $\mathcal{Y}_1 \in \mathbb{R}^{n_1 \times n_2 \times n_3}, \mathcal{Y}_2 \in \mathbb{R}^{n_2 \times n_3 \times n_1}, \mathcal{Y}_3 \in \mathbb{R}^{n_3 \times n_1 \times n_2}$ are the defined Lagrange multipliers and ρ is the penalty parameter. The Frobenius norm is an added regularization term that can enhance the convergence of the function.

The Frobenius norm of tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is defined as follows:

$$\|\mathcal{X}\|_F^2 = \frac{1}{2} \left(\sum_{k=1}^{n_3} \sum_{i=1}^{n2} \sum_{j=1}^{n1} |x_{i,j}|^2 \right)^{\frac{1}{2}} \tag{10}$$

Equation (9) is addressed using the alternating direction method of multipliers (ADMM), an algorithm that simplifies multiparameter optimization challenges by reducing them to problems concerning a single optimization variable. By leveraging the ADMM approach, the objective function of the primary problem can be divided into a set of manageable subproblems, which may then be simultaneously solved. The resolution of these sub-problems subsequently converges through coordination, yielding the comprehensive solution to the initial problem.

When solving the model via the ADMM algorithm, it updates one variable at a time, maintaining the others as constants. Specifically, for Equation (9), it sequentially employs gradient descent to update the variables $\mathcal{B}$, $\mathcal{Z}$, $\mathcal{W}$, $\mathcal{X}$, and $\mathcal{Y}$ in an alternating fashion.

Step 1: Updating $\mathcal{B}^{l+1}$. Keeping $\mathcal{X}_i^l$, $\mathcal{Y}_i^l$, $\mathcal{Z}^l$, and $\mathcal{W}^l$ fixed, let $\frac{\partial \mathcal{L}}{\partial \mathcal{B}} = 0$; the update is as follows:

$$\begin{aligned} \mathcal{B}^{l+1} &:= \underset{\mathcal{B}}{\operatorname{argmin}} \mathcal{L}\left(\mathcal{X}_i^l, \mathcal{B}, \mathcal{Z}^l, \mathcal{W}^l, \mathcal{Y}_1^l\right) \\ &= \alpha_1 \parallel \mathcal{B} \parallel_{p,r}^2 + < \mathcal{Y}_1^l, \mathcal{X}_1^l - \mathcal{B} > + \frac{\rho}{2} \parallel \mathcal{X}_1 - \mathcal{B} \parallel_F^2 \\ &= \alpha_1 \parallel \mathcal{B} \parallel_{p,r}^2 + \frac{\rho}{2} \parallel \mathcal{B} - < \mathcal{X}_1^l + \frac{1}{\rho} \mathcal{Y}_1^l > \|_F^2 \\ &= \mathcal{D}_{\frac{\alpha_1}{\rho}, r, p}\left(\mathcal{X}_1^l + \frac{1}{\rho} \mathcal{Y}_1^l\right) \end{aligned} \tag{11}$$

According to the definition of tensor singular value thresholding, it can be inferred that $\mathcal{D}_{\frac{\alpha_i}{\rho}, r, *}\left(\mathcal{X}_i^l + \frac{1}{\rho} \mathcal{Y}_i^l\right)$ is the threshold applied to truncate the tensor $\left(\mathcal{X}_i^l + \frac{1}{\rho} \mathcal{Y}_i^l\right)$ after performing singular value decomposition, with the threshold $\tau = \frac{\alpha_i}{\rho}$.

Step 2: Updating $\mathcal{Z}^{l+1}$. Keeping $\mathcal{X}_i^l$, $\mathcal{Y}_i^l$, $\mathcal{B}^{l+1}$, and $\mathcal{W}^l$ fixed, let $\frac{\partial \mathcal{L}}{\partial \mathcal{Z}} = 0$; the update is as follows:

$$\begin{aligned} \mathcal{Z}^{l+1} &= \underset{\mathcal{Z}}{\operatorname{argmin}} \mathcal{L}\left(\mathcal{X}^l, \mathcal{B}^{l+1}, \mathcal{Z}, \mathcal{W}^l, \mathcal{Y}_2^l\right) \\ &= \alpha_2 \parallel \mathcal{Z} \parallel_{p,r}^2 + < \mathcal{Y}_2^l, \mathcal{X}_2^l - \mathcal{Z} > + \frac{\rho}{2} \parallel \mathcal{X}_2^l - \mathcal{Z} \parallel_F^2 \\ &= \alpha_2 \parallel \mathcal{Z} \parallel_{p,r}^2 + \frac{\rho}{2} \parallel \mathcal{Z} - < \mathcal{X}_2^l + \frac{1}{\rho} \mathcal{Y}_2^l > \|_F^2 \\ &= \mathcal{D}_{\frac{\alpha_2}{\rho}, r, p}\left(\mathcal{X}_2^l + \frac{1}{\rho} \mathcal{Y}_2^l\right) \end{aligned} \tag{12}$$

Step 3: Updating $\mathcal{W}^{l+1}$. Keeping $\mathcal{X}_i^l$, $\mathcal{Y}_i^l$, $\mathcal{B}^{l+1}$, and $\mathcal{Z}^{l+1}$ fixed, let $\frac{\partial \mathcal{L}}{\partial \mathcal{W}} = 0$; the update is as follows:

$$\begin{aligned}\mathcal{W}^{l+1} &= \underset{\mathcal{W}}{\operatorname{argmin}} \mathcal{L}\left(\mathcal{X}_3^l, \mathcal{B}^{l+1}, \mathcal{Z}^{l+1}, \mathcal{W}, \mathcal{Y}_3^l\right) \\ &= \alpha_3 \parallel \mathcal{W} \parallel_{p,r}^2 + < \mathcal{Y}_3^l, \mathcal{X}_3^l - \mathcal{W} > + \frac{\rho}{2} \parallel \mathcal{X}_3^l - \mathcal{W} \parallel_F^2 \\ &= \alpha_3 \parallel \mathcal{W} \parallel_{p,r}^2 + \frac{\rho}{2} \parallel \mathcal{W} - < \mathcal{X}_3^l + \frac{1}{\rho}\mathcal{Y}_3^l > \parallel_F^2 \\ &= \mathcal{D}_{\frac{\alpha_3}{\rho},p,r}\left(\mathcal{X}_3^l + \frac{1}{\rho}\mathcal{Y}^l\right)\end{aligned} \tag{13}$$

Step 4: Updating $\mathcal{X}_i^{l+1}$. Keeping $\mathcal{W}^{l+1}$, $\mathcal{Y}_i^l$, $\mathcal{B}^{l+1}$, and $\mathcal{Z}^{l+1}$ fixed, let $\frac{\partial \mathcal{L}}{\partial \mathcal{X}} = 0$; the update is as follows:

$$\begin{aligned}\mathcal{X}_i^{l+1} &:= \underset{\mathcal{X}}{\operatorname{argmin}} \mathcal{L}\left(\mathcal{X}_i, \mathcal{B}^{l+1}, \mathcal{Z}^{l+1}, \mathcal{W}^{l+1}, \mathcal{Y}_i^l\right) \\ &= < \mathcal{Y}_1^l, \mathcal{X}_1 - \mathcal{B}^{l+1} > + \frac{\rho}{2} \parallel \mathcal{X}_1 - \mathcal{B}^{l+1} \parallel_F^2 + \\ &< \mathcal{Y}, \mathcal{X}_2 - \mathcal{Z}^{l+1} > + \frac{\rho}{2} \parallel \mathcal{X}_2 - \mathcal{Z}^{l+1} \parallel_F^2 + < \mathcal{Y}_3^l, \\ &\mathcal{X}_3 - \mathcal{W}^{l+1} > + \frac{\rho}{2} \parallel \mathcal{X}_3 - \mathcal{W}^{l+1} \parallel_F^2 \\ &= \sum_{i=1}^{3} \left((\mathcal{B}^{l+1} + \mathcal{Z}^{l+1} + \mathcal{W}^{l+1}) - \frac{1}{\rho}\mathcal{Y}_i^l \right)\end{aligned} \tag{14}$$

Moreover, the remaining constraint $\mathcal{P}_\Omega(\mathcal{X}_i) = \mathcal{P}_\Omega(\mathcal{M}_i)$ is added. Thus, $\mathcal{X}_i^{l+1}$ can be expressed as follows:

$$\begin{aligned}\mathcal{X}_i^{l+1} = \sum_{i=1}^{3} \left((\mathcal{B}^{l+1} + \mathcal{Z}^{l+1} + \mathcal{W}^{l+1}) - \frac{1}{\rho}\mathcal{Y}_i^l \right) \\ \text{s.t. } : \mathcal{P}_\Omega(\mathcal{X}_i) = \mathcal{P}_\Omega(\mathcal{M}_i)\end{aligned} \tag{15}$$

Consequently, the constraint is removed, and elements at the positions of the observation points in the completion tensor are replaced by elements at the positions of the original observation points:

$$\mathcal{X}_i^{l+1} = \sum_{i=1}^{3} \left((\mathcal{B}^{l+1} + \mathcal{Z}^{l+1} + \mathcal{W}^{l+1}) - \frac{1}{\rho}\mathcal{Y}_i^l \right)_{\overline{\Omega}} + \mathcal{P}_\Omega(\mathcal{M}) \tag{16}$$

Step 4: Updating $\mathcal{Y}_i^{l+1}$. Keeping $\mathcal{W}^{l+1}$, $\mathcal{X}_i^{l+1}$, $\mathcal{B}^{l+1}$, and $\mathcal{Z}^{l+1}$ fixed, the value $\mathcal{Y}_i^{l+1}$ is updated based on the following process:

$$\mathcal{Y}_i^{l+1} := \mathcal{Y}_i^l + \rho \left\langle \begin{array}{l} a_1\left(\mathcal{B}^{l+1} - \mathcal{X}_1^{l+1}\right) + \\ a_2\left(\mathcal{Z}^{l+1} - \mathcal{X}_2^{l+1}\right) + \\ a_3\left(\mathcal{W}^{l+1} - \mathcal{X}_3^{l+1}\right) \end{array} \right\rangle \tag{17}$$

The number of iterations is set and updated for each iteration according to the above steps. Moreover, the optimal solution for the model is computed. The steps of the LRTC-P2TN algorithm are defined as follows (Algorithm 1):

Algorithm 1 LRTC-P2TN

Input: $\mathcal{X} \geq 0$, A,$\mathcal{P}_\Omega(\mathcal{M}_i) = \mathcal{P}_\Omega(\mathcal{X}_i)$, $\mathcal{Y}_i = 0$, ρ, k, λ
Output: the recovered tensor $\mathcal{X}$
Calculate W using Formula (1)
for $i = 1$ **to** k
  Update $\mathcal{B}^{l+1}$ using Formula (11)
  Update $\mathcal{Z}^{l+1}$ using Formula (12)
  Update $\mathcal{W}^{l+1}$ using Formula (13)
  Update $\mathcal{X}_i^{l+1}$ using Formula (16)
  Update $\mathcal{Y}_i^{l+1}$ using Formula (17)
  If $\frac{\|\mathcal{X}^{l+1}-\mathcal{X}^l\|_F^2}{\|\mathcal{P}_\Omega(\mathcal{X})\|_F^2} < \mathcal{E}$ then break
end for

During the construction and solution process of the entire mathematical model, the influence of noise was not considered. In more detail, the LRTC-P2TNN algorithm only considered the reconstruction accuracy of the given data.

3.3. Multi-Source Data-Fusion Tensor Completion Framework

The core of the multi-source data-fusion completion framework represents the LRTC-P2TN module. As depicted in Figure 1, the correlated data, gathered from different sources, were represented as tensors and were set with different deletion rates. After completion with the LRTC-P2TN module, the final completion data were acquired. The truncated $\ell_{2,p}$ norm integrated into the model played a crucial role in ensuring the convergence of the algorithm during the completion process.

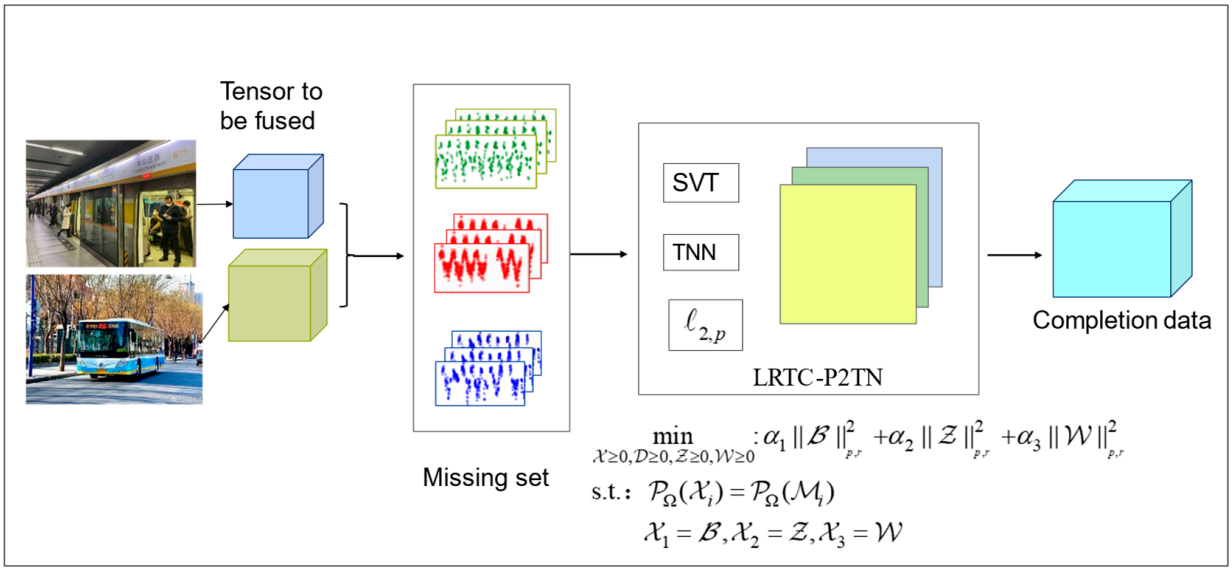

Figure 1. Multi-source data-fusion tensor completion framework.

4. Results and Discussion

4.1. Data Preparation

The experimental data included two datasets: one set of Shenzhen Tong data, referred to as SZ, and one set of Beijing Transportation Smart Card data, known as BJ.

The Shenzhen Tong dataset contained Shenzhen Tong Smart Card data for 22 October 2013, covering data gathered from 117 bus and subway stations. The methodology included both single-source and multi-source data completion techniques. Single-source data adopted a three-dimensional array formulated as day × time interval × subway station, whereas multi-source data extended to a four-dimensional array featuring passenger flow

type × day × time interval × station. Analysis of bus flows at subway stations leveraged user ID sequencing to unravel transfer dynamics, with subway flows exclusively accounting for entry volumes, tallied at five-minute intervals. Shenzhen's data footprint for a single day manifested as a three-dimensional array ($1 \times 288 \times 117$) and a four-dimensional array ($2 \times 1 \times 288 \times 117$) for single-source and multi-source contexts, respectively. By examining RMSE and MAPE metrics, we evaluated the augmentation of subway data completion accuracy via the incorporation of bus data.

Conversely, the BJ dataset encompassed entries from 239 stations, spanning 1 and 10 May 2019. It mirrored the SZ data's analytical approach, yielding a four-dimensional structure of day × passenger flow type × time interval × station, quantified in five-minute intervals. The resulting dataset, encapsulating a ten-day span, consisted of a sizably expansive array at $2 \times 10 \times 288 \times 239$.

The time-series datasets that we analyzed and constructed can be downloaded from the website https://github.com/yongmeizhao/smartcard-data/blob/main/data.rar, accessed on 2 December 2023.

4.2. Single-Source Data Completion

During this experiment, we used Beijing subway data for 7 May 2019 and Shenzhen smart card data for 22 October 2013. Figures 2 and 3 illustrate the visualization results of both datasets, where the horizontal axis represents the station and the vertical axis is the passenger flow at each time interval. Based on these figures, it can be observed that the Beijing subway data for 7 May showed a more obvious periodicity, with clearer peaks in the morning and evening.

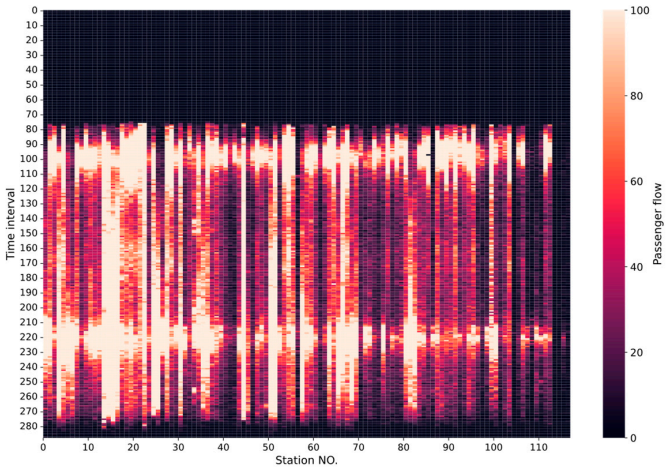

Figure 2. Passenger flow data of the Shenzhen subway on 22 October 2013.

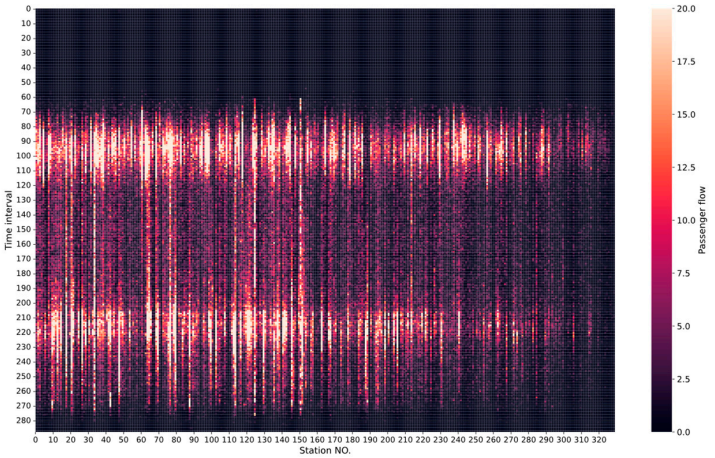

Figure 3. Passenger flow data of the Beijing subway on 7 May 2019.

Moreover, the single-source data experiments were considered for subway flow data. The experimental data used three-dimensional data consisting of station × time interval × subway flow (data can be downloaded from the address). The truncation value was set to 4.5, and ρ represents the iteration step size. The initial iteration step size ρ was set to 1.15, and the maximum value $\rho_{\max}$ of ρ was equal to 1×10^{-5}. In each iteration, ρ was updated through the equation $\rho = \min\{1.05\rho, \rho_{\max}\}$, and the number of iterations was set to 150. For convenience of recording and accuracy, the MAPE values in the table below were multiplied by ten. Moreover, Table 2 illustrates the completion results under deletion rates of 10%, 30%, 50%, and 70%, using LRTC-P2TN.

Table 2. RMSE/MAPE of single-source data completion under random deficiency.

	10%	**30%**	**50%**	**70%**
SZ	30.52/37.01	26.13/40.16	32.81/62.08	50.37/83.48
BJ	3.25/5.75	3.25/5.82	3.31/5.90	3.93/6.05

4.3. Multi-Source Data Completion

In this study, we used the Beijing subway and bus data for 7 May 2019 and Shenzhen smart card data for 22 October 2013. The experimental data were constructed as a four-dimensional tensor dataset of subway station name, time, subway station passenger flow, and bus station passenger flow. For missing data processing, deletion rates were only set for the subway flow data, while the bus flow data were complete. The experimental parameters were the same as those specified for the single-source data completion.

According to Tables 2 and 3, the completion performance with the fusion of multi-source data of buses led to better behavior compared to the single-source subway data completion.

Table 3. RMSE/MAPE of multi-source data completion under random deficiency.

	10%	**30%**	**50%**	**70%**
SZ	22.74/32.23	20.39/36.83	25.12/47.74	26.77/48.17
BJ	3.15/5.76	3.20/5.73	3.28/5.87	3.86/6.02

Furthermore, Figures 4 and 5 display the data completion effects of single- and multi-source data on Route No. 7 at a deletion rate of 0.5. As indicated, the fusion of bus transfer data enhanced the algorithm's capability to capture the overall trend of the data.

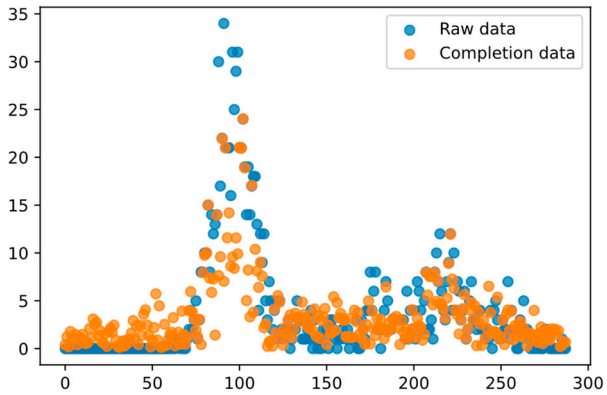

Figure 4. Results of single-source data completion.

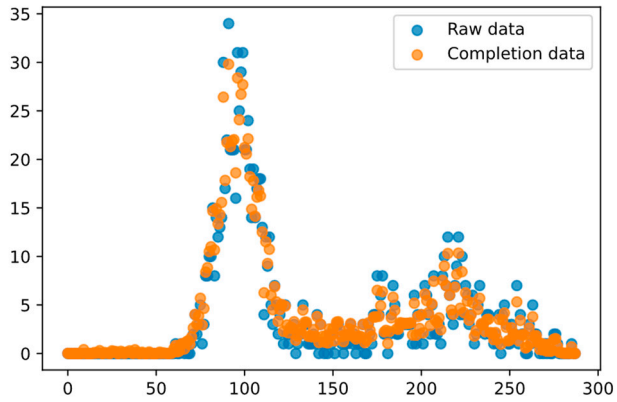

Figure 5. Results of multi-source data completion.

4.4. Comparison of the LRTC-P2TN Model with Other Models

The proposed LRTC-P2TN model was compared to the LATC [26], BGCP [27], NWLRTC [25], and BTTF [28] models. In more detail, LATC is an autoregressive low-rank tensor completion algorithm, whereas BGCP is a Bayesian tensor decomposition model that learns latent factor matrices (i.e., low-rank structures) using the Markov chain Monte Carlo method. Moreover, the Bayesian temporal tensor factorization (BTTF) method merges low-rank matrix/tensor decomposition with vector autoregressive (VAR) processes within a probabilistic graphical model framework to facilitate problem-solving. Finally, NWLRTC is a non-negative multidirectional weighted low-rank tensor completion method.

On the BJ dataset, each baseline algorithm was executed. The deletion rate was set to 20%, 30%, 40%, 50%, 60%, 70%, and 80%; ρ represents the iteration step size, which was initially set to 1×10^{-5} and updated by $\rho = \min\{1.05\rho, \rho_{\max}\}$, while $\mathcal{E}$ is the convergence accuracy, which was initially set to 1. The experimental results showed that, under random deficiency, setting $\rho = 1$ and $\mathcal{E} = 1$ could achieve higher convergence speed and completion accuracy. Moreover, Figure 6 shows the RMSE values for the different algorithms at various deletion rates under random deficiency. A comparison of the computed results indicates that, compared to the LATC, BGCP, NWLRTC, and BTTF models, the LRTC-P2TN model achieved better imputation accuracy at deletion rates between 20% and 60%. Furthermore, at deletion rates of 70% and 80%, its MAPE values surpassed those of the other algorithms, although the RMSE values yielded higher values (as depicted in Table 4). Additionally, the LRTC-P2TN model achieved higher completion accuracy than LATC, BGCP, NWLRTC, and BTTF at deletion rates of 20–60%; when the deletion rates were 70% and 80%, our model achieved higher MAPE values than the other algorithms, but its RMSE values were relatively higher (see Table 4).

Table 4. RMSE/MAPE of different models under random deficiency.

Missing Rate	LATC	BGCP	BTTF	NWLRTC	LRTC-P2TN
20%	3.47/6.50	3.15/5.53	3.22/5.55	3.17/5.73	3.15/5.76
30%	3.48/6.69	3.31/5.65	3.35/5.55	3.34/5.84	3.18/5.63
40%	3.56/6.68	3.28/5.55	3.49/5.67	3.25/5.87	3.17/5.78
50%	3.64/6.92	3.33/5.72	3.36/5.74	3.41/5.95	3.19/5.83
60%	3.64/7.03	3.41/5.82	3.48/5.96	3.47/6.10	3.26/5.85
70%	3.77/7.32	3.51/6.13	3.64/6.17	3.86/6.05	3.78/5.99
80%	3.79/7.31	3.93/6.93	4.12/6.95	5.29/5.96	5.06/5.79

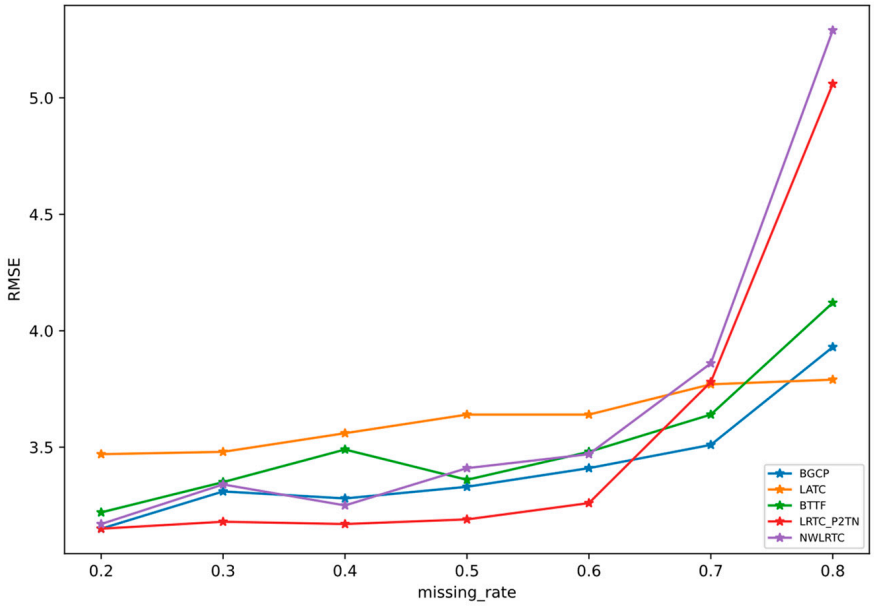

Figure 6. RMSE of different models under different missing rates.

Table 4 displays the RMSE and MAPE values for a range of algorithms using the BJ data with an input orientation of $1 \times 288 \times 239$. A minimal variation can be observed in the RMSE and MAPE metrics for the LATC, BGCP, NWLRTC, and BTTF algorithms when the input orientation shifted to $239 \times 288 \times 1$. The incorporation of directional weighting factors in the LRTC-P2TN and NWLRTC algorithms resulted in negligible changes to their RMSE and MAPE values.

5. Conclusions

Multi-source data can assist in the completion of single-source data, especially when the single-source data have a lot of missing values where correlated data could effectively contribute to data completion. In this study, a four-dimensional tensor dataset of subway and bus transfer was constructed, and experiments were conducted to demonstrate the fusion of multi-source bus transfer data that could achieve higher completion accuracy than single-source subway data alone. Since the tensor is a natural representation of high-dimensional data, tensor-based multi-source data fusion has the advantages of simple data representation and high completion accuracy. Low-rank tensor completion uses the truncated $\ell_{2,p}$ norm as an alternative approach for rank minimization, integrating the advantages of the truncated nuclear norm and the $\ell_{2,p}$ norm; therefore, it achieves more compact results than a single norm during the approximation process of rank minimization.

Funding: This work was supported by the National Natural Science Foundation of China (No. 6200238).

Data Availability Statement: (1) The numerical data analyzed during the present study are available at: https://github.com/yongmeizhao/smartcard-data/blob/main/data.rar (accessed on 2 December 2023). (2) The datasets generated during this study are included in this published article.

Conflicts of Interest: The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Liang, J.; Härdle, W.K.; Tian, M. Imputed quantile tensor regression for near-sited spatial-temporal data. *Comput. Stat. Data Anal.* **2023**, *182*, 107713. [CrossRef]
2. Chen, G.; Liu, Z.; Yu, G.; Liang, J. A new view of multisensor data fusion: Research on generalized fusion. *Math. Probl. Eng.* **2021**, *2021*, 5471242. [CrossRef]
3. Mitchell, H.B. *Data Fusion: Concepts and Ideas*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.

4. Ma, X.; Ye, Q.; Yan, H. L2P-norm distance twin support vector machine. *IEEE Access* **2017**, *5*, 23473–23483. [CrossRef]
5. Muravyov, S.V.; Khudonogova, L.I.; Dai Ho, M. Analysis of heteroscedastic measurement data by the self-refining method of interval fusion with preference aggregation–IF&PA. *Measurement* **2021**, *183*, 109851. [CrossRef]
6. Roy, D.; Li, Y.; Jian, T.; Ioannidis, S. Multi-modality sensing and data fusion for multi-vehicle detection. *IEEE Trans. Multimedia* **2023**, *25*, 2280–2295. [CrossRef]
7. Senel, N.; Kefferpütz, K.; Doycheva, K.; Elger, G. Multi-Sensor Data Fusion for Real-Time Multi-Object Tracking. *Processes* **2023**, *11*, 501. [CrossRef]
8. Chen, R.; Ning, J.; Lei, Y.; Cheng, N. Mixed Traffic Flow State Detection: A Connected Vehicles Assisted Roadside Radar and Video Data Fusion Scheme. *IEEE Open J. Intell. Transp. Syst.* **2023**, *4*, 360–371. [CrossRef]
9. Lin, L.; Li, J.; Chen, F.; Ye, J.; Huai, J. Road traffic speed prediction: A probabilistic model fusing multi-source data. *IEEE Trans. Knowl. Data Eng.* **2017**, *30*, 1310–1323. [CrossRef]
10. Zißner, P.; Rettore, P.H.; Santos, B.P.; Loevenich, J.F.; Lopes, R.R.F. DataFITS: A Heterogeneous Data Fusion Framework for Traffic and Incident Prediction. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 11466–11478. [CrossRef]
11. Wang, P.; Hao, W.; Jin, Y. Fine-grained traffic flow prediction of various vehicle types via fusion of multisource data and deep learning approaches. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 6921–6930. [CrossRef]
12. Ji, B.; Chen, Z.; Mumtaz, S.; Liu, J.; Zhang, Y.; Zhu, J.; Li, C. SWIPT enabled intelligent transportation systems with advanced sensing fusion. *IEEE Sens. J.* **2020**, *21*, 15643–15650. [CrossRef]
13. Satish, V.; Kumar, P.A.R. Variance based data fusion for k-means++. In Proceedings of the 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, India, 7–9 April 2017.
14. Zhang, W.; Qi, Y.; Zhou, Z.; Biancardo, S.A.; Wang, Y.H. Method of Speed Data Fusion Based on Bayesian Combination Algorithm and Markov Model. *IET Intell. Transp. syst.* **2018**, *12*, 1312–1321. [CrossRef]
15. Sofuoglu, S.E.; Aviyente, S. Graph Regularized Tensor Train Decomposition. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3912–3916.
16. Zhao, L. *Research on Multimodal Data Fusion Algorithms*; Dalian University of Technology: Dalian, China, 2018.
17. Pan, B.; Li, C.; Che, H.; Leung, M.F.; Yu, K. Low-Rank Tensor Regularized Graph Fuzzy Learning for Multi-View Data Processing. *IEEE Trans. Consum. Electron.* **2023**, 1. [CrossRef]
18. Shen, G.; Han, X.; Zhou, J.; Ruan, Z.; Pan, Q. Research on Intelligent Analysis and depth Fusion of Multi-source Traffic data. *IEEE Access* **2018**, *6*, 59329–59335. [CrossRef]
19. Xing, J.; Liu, R.; Anish, K.; Liu, Z. A Customized Data Fusion Tensor Approach for Interval-Wise Missing Network Volume Imputation. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 12107–12122. [CrossRef]
20. Respati, S.; Chung, E.; Zheng, Z.; Bhaskar, A. ABAFT: An adaptive weight-based fusion technique for travel time estimation using multi-source data with different confidence and spatial coverage. *J. Intell. Transp. Syst.* **2023**, 1–14. [CrossRef]
21. Liu, Y.X. *Research on Urban Public Transit System Passenger Travel Pattern Mining and Its Application Based on Multi-Source Data Fusion*; South China University of Technology: Guangzhou, China, 2018.
22. Li, Y.F.; Shang, K.; Huang, Z.H. A singular value p-shrinkage thresholding algorithm for low rank matrix recovery. *Comput. Optim. Appl.* **2019**, *73*, 453–476. [CrossRef]
23. Liu, C.; Shan, H.; Chen, C. Tensor p-shrinkage nuclear norm for low-rank tensor completion. *Neurocomputing* **2020**, *387*, 255–267. [CrossRef]
24. Zhang, D.; Hu, Y.; Ye, J.; Li, X. Matrix completion by Truncated Nuclear Norm Regularization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2192–2199.
25. Zhao, Y.M.; Tuo, M.; Zhang, H.; Zhang, H.; Wu, J.; Gao, F. Nonnegative low-rank tensor completion method for spatiotemporal traffic data. *Multimed. Tools Appl.* **2023**, 1–16. [CrossRef]
26. Chen, X.; Cheng, Z.; Saunier, N.; Sun, L. Laplacian convolutional representation for traffic time series imputation. *arXiv* **2022**, arXiv:2212.01529.
27. Chen, X.; He, Z.; Sun, L. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. *Transp. Res. Pt. C-Emerg. Technol.* **2019**, *98*, 73–84. [CrossRef]
28. Chen, X.; Sun, L. Bayesian Temporal Factorization for Multidimensional Time Series Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4659–4673. [CrossRef] [PubMed]