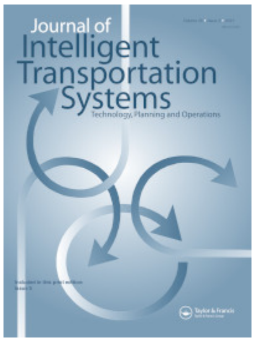

Journal of Intelligent Transportation Systems
Technology, Planning and Operations

Journal of Intelligent Transportation Systems

Technology, Planning, and Operations

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gits20

Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory

Jinjun Tang, Xinshao Zhang, Weiqi Yin, Yajie Zou & Yinhai Wang

To cite this article: Jinjun Tang, Xinshao Zhang, Weiqi Yin, Yajie Zou & Yinhai Wang (2021) Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory, Journal of Intelligent Transportation Systems, 25:5, 439-454, DOI: 10.1080/15472450.2020.1713772

To link to this article: https://doi.org/10.1080/15472450.2020.1713772

Published online: 27 Jan 2020.

JOURNAL OF INTELLIGENT TRANSPORTATION SYSTEMS
2021, VOL. 25, NO. 5, 439–454
https://doi.org/10.1080/15472450.2020.1713772

Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory

Jinjun Tang[a], Xinshao Zhang[a], Weiqi Yin[a], Yajie Zou[b], and Yinhai Wang[c]

[a]Smart Transport Key Laboratory of Hunan Province, School of Traffic and Transportation Engineering, Central South University, Changsha, China; [b]Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, Shanghai, China; [c]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA

ABSTRACT

Currently, accurate traffic flow analysis and modeling are important key steps for intelligent transportation system (ITS). Missing traffic flow data are one of the most critical issues in the application of ITS. In this study, a hybrid method combining fuzzy rough set (FRS) and fuzzy neural network (FNN) is proposed for imputation of missing traffic data. Firstly, FNN is used for data classification, then the K-Nearest Neighbor (KNN) method is used to determine the optimal number of data used to estimate missing data in each category, and finally the fuzzy rough set is used to impute missing values. In order to validate the imputation performance of the proposed hybrid method, the traffic flow data collected from the loop detectors at different time intervals on roadway network are used in model calibration and validation. Three common indicators, including RMSE (root mean square error), R (correlation coefficient) and RA (relative accuracy), are used to evaluate the imputation performance under different data missing ratios. A model comparison is conducted between proposed imputation method and several widely used models including average-based and regression-based methods. The results show that the proposed method is superior to the traditional method for the traffic flow data collected at different time intervals with different missing ratios, which also further demonstrate its effectivity and validity.

ARTICLE HISTORY
Received 22 July 2019
Revised 6 January 2020
Accepted 7 January 2020

KEYWORDS
fuzzy neural network; fuzzy rough set; imputation; K-nearest neighbor; missing values; traffic flow

1. Introduction

Intelligent transportation system combines advanced technologies in the practical application of transportation system. It can effectively utilize existing transportation facility and infrastructures, relieve traffic congestion and environmental pollution, enhance traffic safety and improve transportation efficiency. Currently, one challenging issue for intelligent transportation system is accurate collection of real-time traffic flow data. Due to various reasons including weather effect, detector fault and artificial error etc., the missing traffic flow data become one of the most critical problems. The incomplete traffic flow data can result in the quality of model application, data analysis and traffic flow prediction.

At present, many researchers proposed various methods on the missing data imputation in the transportation system and other fields. The most commonly used methods are based on interpolation, such as historical or adjacent interpolation (Chen & Shao, 2000; Nguyen & Scherer, 2003; Ni, Leonard, Guin, & Feng, 2005, Tang, Zhang, Wang, Wang, & Liu, 2015), spline (including linear) and regression interpolation method (Allison, 2001; Baharaeen & Masud, 1986; De Boor, 1978; Holt, 2004).

Other typical methods include Matrix-based interpolation techniques (Qu, Hu, Li, & Zhang, 2009; Qu, Zhang, Hu, Jia, & Li, 2008), Time Series (Ahmed & Cook, 1979; Ghosh, Basu, & O'Mahony, 2005), Machine learning techniques (Tang, Liu, Zou, Zhang, & Wang, 2017; Tang et al., 2019) and other prediction models (Chen et al., 2019; Yan, Zhang, Tang, & Wang, 2017) are also applied in traffic forecasting.

On the basis of historical interpolation method, Gold, Turner, Gajewski, and Spiegelman (2000) introduced the method to complete imputation process, and they used the "factorization and linear" interpolation strategy to estimate the missing traffic data. The missing data was estimated by using average values which were given by calculating various impact factors. Zhong, Lingras, and Sharma (2004) tested the factor method on the permanent flow count (PTC) to

CONTACT Yajie Zou ✉ yajiezou@hotmail.com Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, Shanghai 201804, China.

detect the ability of imputation for the missing values. In terms of the regression interpolation method, Al-Deek, Venkata, and Chandra (2004) described and established various regression models and applied the regression model to estimate the missing data. They concluded that the optimal model was quadratic regression model with selective median. Boyles (2011) compared the accuracy of 11 different methods by simulating three types of missing data, including some regression models: simple linear regression models, multiple linear regression models, and non-normal Bayesian linear regression models. The three missing data types are: random, continuous for a long time, and missing data from the same system every day. Wilby, Díaz, Rodríguez GonzĺLez, and Sotelo (2014) proposed a lightweight method which based on a pairing linear regression to carry out "on-line" occupancy imputation and the estimation. In general, interpolation methods such as the historical interpolation method and the regression interpolation method are easy to develop and apply, but the imputation accuracy of this kind method depends on the similarity assumption of traffic flow data within a certain range of time period, that is, the time correlation characteristics of traffic flow data. However, for the random variation in traffic flow data caused by traffic incidents, the imputation of interpolation related models may face unstable or inaccurate estimation results.

Matrix-based interpolation technique, an improved interpolation, is adopted and applied to solve missing data issues in transportation system recently. Qu et al. (2008) and (2009) proposed a missing value estimation algorithm based on Bayesian principal component analysis (BPCA) and Probabilistic Principal Component Analysis (PPCA) through establishing traffic data matrix. Although the experiments showed that the BPCA and PPCA methods performed better and have higher precision than the traditional methods (Qu et al., 2009), they were only applied in the scenario for the missing data imputation of a single loop detector. At present, Li, Li, and Li (2013) extended the kernel probability principal component analysis (KPPCA) method to the imputation of missing data from multi-loop detectors, and the results showed that the KPPCA method expressed better imputation performance by considering the spatio-temporal correlation of traffic flow data. Tan et al. (2013) proposed a tensor decomposition based estimation method for the missing traffic flow data. Ran et al. (2016) presented a new imputation method for traffic flow based on tensor model considering spatial-temporal features in traffic flow and obtained higher imputation accuracy.

Time series based imputation method is also a direction focused by many transportation practitioners. These methods generally impute missing traffic flow data through prediction, such as the Auto-regressive Integrated Moving Average (ARIMA) (Ahmed & Cook, 1979) or seasonal ARIMA (Ghosh et al., 2005; Ramsey & Hayden, 1994; Redfern, Waston, Tight, & Clark, 1993; Williams & Hoel, 2003). The basic idea of those methods is firstly to explore the future changing trend or pattern according to the historical traffic flow data, and then predict the future trend, which is used as estimation for missing data (Zhong et al., 2004). A research team from the University of Leeds conducted a series of studies (Clark, Watson, & Redfern, 1993; Redfern et al., 1993; Tight, Redfern, & Watson, 1993). They used exponentially weighted moving averages, based on auto-correlation impact functions and ARIMA models to finish missing traffic data imputation. The comparing results showed that the ARIMA model was effective in detecting and estimating missing traffic flow data. In this kind of model, the premise of the time series method is that the regular changes of pattern in the historical data.

The machine learning techniques are an effective estimation algorithm which uses regulation data to predict unknown data and develop rapidly in the field of traffic data imputation recently. For instance, Hofleitner, Herring, and Bayen (2012) proposed a hybrid modeling framework which combined the theory of traffic flow and a machine learning network. Xie and Huynh (2010) introduced two kernel-based supervised machine learning methods to estimate data. Zhang et al. (Zhang, Zou, Tang, Ash, and Wang 2016; Zhang, Yu, Qi, Shu, & Wang, 2019) applied artificial neural network and support vector machine to estimate waiting queue lengths and combined them to obtain a superior prediction performance. Nowicki (2009, 2010) incorporated the rough fuzzy sets into fuzzy-neuro structures to derive rough fuzzy-neuro classifier which presented a new approach to fuzzy classification in the case of missing data. This method provided an effective way for classifying missing data. Kumar, Vanajakshi, and Subramanian (2018) combined the similarity of the input data and the historic data to establish a k-neural network classifying algorithm to recognize significant inputs and caught an improvement performance.

Regarding the research of using rough set to impute the data, it can effectively analyze and process various information systems that are inaccurate, inconsistent or incomplete, and can make the missing attributes and other attributes of the system as consistent as possible. Gao and Fasheng (2009) proposed a method of missing data complementation based on rough set theory, which mainly uses ROUSTDIA algorithm from rough set theory to complete the data, and then a large number of researchers have modified the rough set to improve the accuracy (Duan, Liu, Chen, Jiang, & Li, 2011; Fan, 2013). Rough set only uses the information provided by the data set being mined and does not depend on other models or parameter assumptions. But many missing data cannot be repaired by the rough set, and finally need to use interpolation and other methods to complete. Jensen and Cornelis (2011) demonstrated that the accuracy and robustness of fuzzy-rough set-based nearest neighbor methods are high than their non-fuzzy-rough counterparts. Amiri and Jensen (2016) proposed three missing value imputation methods based on fuzzy-rough sets and its recent extensions: namely, implicator/t-norm based fuzzy-rough sets, vaguely quantified rough sets and also ordered weighted average based rough sets. The procession of missing data imputation is similar to deal with the uncertainty and ambiguity data. The method based on fuzzy rough set is an effective tool to deal with the problem of uncertainty with high accuracy and robustness. Nevertheless, above methods adopting rough set or fuzzy-rough set cannot finish parameters optimization and missing data classification, which results in low precision of data completion.

However, based on the above research in the field of missing traffic flow data imputation, there are still several limitations or challenges in the current studies: 1) most imputation methods are used for modeling and data recovery under the assumption of complete data; 2) although traditional methods such as historical average and ARIMA have simple calculation structure and high calculation speed, the imputation performance is not good when the traffic flow data express irregular variation and changes; 3) each individual model has its own advantages, and has its specific scope of application. Combining several models with different characteristics could improve imputation performance.

To address those above problems, this paper proposes a hybrid imputation method for missing traffic flow data, which combines rough set and fuzzy neural network to finish imputation process. This method fuses the characteristics of different models: 1) use FNN (fuzzy neural network) to complete classification of the original traffic flow data. FNN can combine the adaptive learning ability of neural network system and the fuzzy inference of FRS and expresses advantages in dealing with the clustering issue of traffic flow data with missing values; 2) use KNN (K-nearest neighbor) to determine the number of data samples related to the imputation object for each cluster or category; 3) apply fuzzy rough set theory to estimate the missing data.

The reminder of this study is organized as follows. In section 2, related methods are introduced, including the definition of fuzzy rough set, the procedure of classification with FNN and the procedure to impute missing data with FRS based on KNN. The case study of missing data imputation for loop detectors in the road network is introduced in section 3 and the results are further discussed. The last section summarizes the conclusions of this study.

2. Methodology

The framework of missing data imputation in this study is shown as follow:

This framework includes three sections. The first section is the skeleton of the frame which is the procedure of classify and impute missing flow data. The second section is the process of classification with FNN and the third section is the step of imputation missing flow data with FRS. All the details about the calculation process in three sections will be introduced in the following sub-sections.

2.1. Classification based on fuzzy neural network

In this paper, the FNN is used to classify the traffic flow before the imputation of missing traffic flow data. The specific steps are shown as follows:

Step 1: use the FNN to train the traffic data and obtain the optimized values of parameters: m, b and w. The parameter m is the central value of the membership function of the second layer. The parameter b is the width of the membership function of the second layer. The parameter w is the connection weight of input layer;

Step 2: use the calibrated FNN to finish classification process under following five specific calculation steps:

Layer 1: input a eigenvector which include five attribute values of a time period, such as $\boldsymbol{V_i}=\{a_1(x_i), a_2(x_i), a_3(x_i), a_4(x_i), a_5(x_i)\}$;

Layer 2 (blur layer): define the membership function. Gaussian function is used as the membership function in this study for its generality:

$$\mu_i^j = \lambda^{-\frac{(v_i - m_{ij})^2}{b_{ij}{}^2}} \tag{1}$$

where, v_i presents traffic flow data as shown in layer 1, and j is the number of membership functions set.

Layer 3 (rule antecedent layer): use the MAMDANI reasoning method:

$$\mu_A(\bar{v}) = \underset{i=1}{\overset{n}{T}} \mu_{A_i}(\bar{v_i}) \tag{2}$$

where $A = A_1 \times A_2 \times \ldots \times A_n$, n is the number of input attribute and is set to 5 in this study. A_n is the membership function vector of n attributes, and T is any t-norm.

Layer 4: calculate the membership degree of object x:

$$\bar{z} = \sum_{r=1}^{n} \bar{p^r} \mu_{A^r}(\bar{v}) \tag{3}$$

$$p^r = \frac{w^r}{\sum w^r} \tag{4}$$

Layer 5: set classifying ranges according to the results of the fourth layer and make the data samples fall in the same range into the same category.

Step 3: the traffic flow data is divided into several categories by implementing the step from 1 to 2, and then estimate the values of missing data by using the hybrid model combining the FRS and KNN, which will introduce in the next section.

2.2. Imputation process based on combination of fuzzy rough set and KNN

Rough set theory is a widely used tool for dealing with uncertainty. In the rough set theory, the information system is expressed as (*X*, *A*), where *X* is a limited set of data instance and *A* is a set of non-empty finite attributes. For $B \subseteq A$, the indistinguishable relation of *B* is:

$$R_B = \{(x, y) \in X^2 | \forall a \in B \Rightarrow a(x) = a(y)\} \tag{5}$$

where, R_B is the equivalence relation. This particular information system contains decision attributes that contain classes for each data instance. If $A \subseteq X$, then the upper and lower approximations of set *A* are defined as:

$$R_B \uparrow A = \{x \in X | [x]_{R_B} \cap A \neq \varphi\} \tag{6}$$

$$R_B \downarrow A = \{x \in X | [x]_{R_B} \subseteq A\} \tag{7}$$

In set *X*, any member of $R_B \uparrow A$ may be a member of set *A* and any member of $R_B \downarrow A$ must be a member of set *A*. If *R* is a fuzzy tolerance relation, $(X, \cup \{d\})$ is the decision system and *X* is a subset of *A*, then the fuzzy rough upper approximation set and low approximation set of *A* are defined as:

$$(R_B \downarrow A)(y) = \inf_{x \in X} I\Big(R_B(x, y), A(x)\Big) \tag{8}$$

$$(R_B \uparrow A)(y) = \sup_{x \in X} T\Big(R_B(x, y), A(x)\Big) \tag{9}$$

where, *I* is the implication operator, *T* is a *t*-modulus and *B* is a subset of set *A*. Attribute set *A* can be a ordinary set or a fuzzy set.

In this study, the upper and low approximation set of the fuzzy rough set are defined as follows:

$$(R \downarrow R_d^Z)(y) = \inf_{t \in N} I\Big(R(y, t), R_d(t, z)\Big) \tag{10}$$

$$(R \uparrow R_d^Z)(y) = \sup_{t \in N} T\Big(R(y, t), R_d(t, z)\Big) \tag{11}$$

where, R_d^Z is the fuzzy tolerance relation and is used to calculate the similarity of two instances of decision features. $R_d(t, z)$ is the similarity or decision feature of element *z* and *t* relative to feature *d* and calculated by fuzzy tolerance relation. The fuzzy set *A* can be represented by $R_d(t, z)$. In generally, $R_a^Z(t)$ refers to the similarity of element *z* and t of attribute *a*.

The detailed calculation process of the FRS based on KNN for imputation of missing traffic flow data is summarized as follows：

Step 1: construct the missing data group $(x_0, a_i(x_0))$ according to the missing data table, in which x_0 indicates a time period of missing data, and $a_i(x_0)$ represents the value of missing data and be defaulted to zero;

Step 2: calculate the distance between the complete attribute value in the missing data group. Euclidean distance formula is used to evaluate the distance in this study and shown as follows:

$$distance(x_m, x_0) = \sqrt{\Big(a_1(x_m) - a_1(x_0)\Big)^2 + \Big(a_2(x_m) - a_2(x_0)\Big)^2 + \ldots\Big(a_r(x_m) - a_r(x_0)\Big)^2} \tag{12}$$

where $a_r(x_0)$ represents the attribute value of r in the missing data group in addition to the missing value, $a_r(x_m)$ represents the attribution value of r in data group x_m which is apart from the missing data group x_0.

Step 3: set the nearest neighbor parameter k for the KNN. According to the results of the second step, the set contains k data group with the smallest Euclidean distance is obtained, and sort k data in terms of distance from smallest to largest to obtain the new set $\{x_1, x_2, ..., x_k\}$;

Step 4: according to the k data group sorting in the third step, set the similarity example $t = x_j$, $z = \{x_1, x_2, ...x_k\}, k \neq j$, and carry out the calculation of the following ①—⑥ sub-steps to finish missing data imputation, and set $y = x_0$:

① Calculate the similarity characteristics of y and t in terms of the attribute value $a_r(r \neq i)$:

$$R_{a_r}(y,t) = 1 - \frac{a_r(y) - a_r(t)}{a_{r\max} - a_{r\min}} \tag{13}$$

And then calculate the similarity features of y and t for the attribute value $A(A = \{a_1, a_2, ..., a_r\}, r \neq i)$:

$$R(y,t) = \min_{a \in A} R_a(y,t), A = \{a_1, a_2, ...a_r\}, r \neq i \tag{14}$$

② Calculate the similarity characteristics of y and t considering the attribute value a_i:

$$R_{a_i}(t,z) = 1 - \frac{a_i(t) - a_i(z)}{a_{i\max} - a_{i\min}} \tag{15}$$

③ Obtain the implication operator of I for similar features as follows:

$$I\Big(R(y,t), R_{a_i}(t,z)\Big) = \max\Big(1 - R(y,t), R_{a_i}(t,z)\Big) \tag{16}$$

Then, calculate t-modulus of similar features as follows:

$$T\Big(R(y,t), R_{a_i}(t,z)\Big) = \min\Big(R(y,t), R_{a_i}(t,z)\Big) \tag{17}$$

④ Estimate the upper and lower approximate values of the fuzzy rough set of the missing values under setting different similarity instance t and z:

$$(R \downarrow R_a^Z)(y) = \inf_{t \in N} I\Big(R(y,t), R_{a_i}(z,t)\Big) \tag{18}$$

$$(R \uparrow R_a^Z)(y) = \sup_{t \in N} T\Big(R(y,t), R_{a_i}(z,t)\Big) \tag{19}$$

⑤ Calculate the average approximate value of the fuzzy rough set of the missing value according to the result of step ④:

$$\tau_1 = \sum \frac{(R \downarrow R_a^Z)(y) + (R \uparrow R_a^Z)(y)}{2} \tag{20}$$

Then, obtain the fuzzy rough weighted average approximation of the missing value:

$$\tau_2 = \sum \left(a(z) * \frac{(R \downarrow R_a^Z)(y) + (R \uparrow R_a^Z)(y)}{2} \right) \tag{21}$$

⑥ Obtain the estimated value for the missing dataset:

$$a(x_0) = {}^{\tau_2}/_{\tau_1} \tag{22}$$

Step 5: Repeat the calculation process from the step 1 to step 4 to impute all the missing traffic flow in the data set, then conduct and analyze the imputation performance according to the results.

3. Experimental testing and result discussion

3.1. Data description

In the experiment, we use traffic flow data to evaluate the imputation performance of the proposed method in this study, and the data were collected from the loop detectors on expressway in Harbin City, China, during January to April in 2011 from 00:00 to 24:00. The original traffic flow data were aggregated in four different time intervals: 5, 10, 30 and 60 min. The data were randomly deleted at different missing ratios: 5%, 10%, 15%, 25%, 30% and then be used as testing data set to evaluate the imputation accuracy of the different methods. The missing ratio is the number of missing data divided by the total number of data. In this study, we consider two types of missing values, random and continuous type. The random type means that the traffic flow data are missing at random, and the continuous type means that the flow is missing for a relative long time periods. Table 1 shows a section of traffic flow data collected in 5 days at interval of 5 min. The symbol "?" indicates missing values, and it can be imputed from the values or information around it. In the experiment, the testing data were collected at different intervals and were deleted at different ratios, and then we conduct model comparison to evaluate imputation performance.

3.2 Performance evaluation criteria

In this study, three performance criteria were used to evaluate the accuracy of data imputation as follows:

1. Root Mean Square Error (RMSE)

RMSE is the error between missing data y and imputed data $\bar{y_i}$:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2} \tag{23}$$

2. Correlation Coefficient (R)
   R is the approximate degree between the missing data y and imputed data$\bar{y_i}$:

$$R = \frac{n \cdot \sum_{i=1}^{n}(y_i \cdot y_i) - \sum_{i=1}^{n}y_i \sum_{i=1}^{n} y_i}{\sqrt{[n \cdot \sum_{i=1}^{n}y_i^2 - (\sum_{i=1}^{n}y_i)^2]} \cdot \sqrt{[n \cdot \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2]}} \tag{24}$$

3. Relative Accuracy (RA)
   RA is an indicator to measure the quantity within a certain tolerance error range. In this study, the tolerance error range is set as ±10%. It is defined as follows:

$$PAE = \frac{|y_i - y_i|}{y_i} \times 100 \tag{25}$$

$$RA = \frac{n_p}{n} \times 100 \tag{26}$$

   where PAE represents the percentage absolute error, n_p represents the number of PAE value limited within the range of ±10%, and n represents the total number of data. The larger of the RA value is, the higher percentage of errors limited in small ranges will be.

Table 1. Traffic flow in 5 consecutive days at interval of 5 minutes.

Time	Monday	Tuesday	Wednesday	Thursday	Friday
00:05	8	10	13	14	16
00:10	10	7	?	7	7
00:15	?	10	8	10	14
00:20	13	12	7	?	12
…	…	…	…	…	…
…	…	…	…	…	…
00:05	12	6	10	12	10
00:10	?	10	6	11	9
…	…	…	…	…	…

3.3. Result comparisons and discussions

Different from other imputation methods, which use the complete data in model training and then removing the part of data samples as to testify imputation performance, the method proposed in this study is based on the deleted database for training and imputation. In order to measure the accuracy of the method proposed in this paper, a model comparison is made among several imputation methods. As the traditional average model and regression model express good imputation performance under missing data set, we select them as model candidate in comparison. (1) The average method. Use the Euclidean distance to obtain k group data set nearest to the missing data group and then average the values of k group data sets as imputed value; (2) The regression method. Use the Euclidean distance to calculate k group data set closest to the missing data set and then construct regress model based on the k group data sets to impute missing values; (3) The fuzzy rough set method based on KNN. The algorithm is shown in 2.2 of this paper; (4) The KNN fuzzy rough set method based on the fuzzy neural network shown in the section 2. The missing data is classified by fuzzy neural network before the imputation of the KNN-based fuzzy rough set.

3.3.1. Imputation performance under random missing data

Firstly, we compare the imputation performance under completely random missing at different missing rations: 5%, 10%, 15%, 20%, 25% and 30%. The imputation results of four methods using traffic flow data collected at 5 min are shown in Table 2. In order to make a complete comparison, the imputation performances of four methods at different time intervals are shown in Figures 1–4.

3.3.2. Imputation performance under continuous missing data

We further compare the imputation performance under continuous missing type at different missing rations: 5%, 10%, 15%, 20%, 25% and 30%. The imputation results of four methods at different data collection time intervals are shown in Table 3, Figures 5–7.

From the model comparison between different models, several conclusions can be summarized as follows:

1. The accuracy of data imputation for the data missing at random type is higher than that at continuous type.

Taking data collected at the interval of 5 min as example, the imputation performance evaluated using

Table 2. Comparison of imputation results of datasets at intervals of 5 min.

Missing rate		5%	10%	15%	20%	25%	30%
Average model	RMSE	7.5468	8.7561	9.1631	10.2720	10.3573	11.7276
	R	0.9517	0.9398	0.9403	0.9260	0.9283	0.9170
	RA	0.3667	0.3154	0.3093	0.2631	0.2467	0.2264
Regression model	RMSE	7.5376	8.5268	9.1568	10.1837	10.3369	12.0006
	R	0.9515	0.9421	0.9398	0.9294	0.9288	0.9191
	RA	0.3488	0.3519	0.2889	0.2891	0.2613	0.2190
KNN-fuzzy rough set	RMSE	7.3297	8.4082	8.7813	9.9594	9.8568	10.8914
	R	0.9558	0.9502	0.9474	0.9371	0.9409	0.9388
	RA	0.3779	0.3370	0.3169	0.2960	0.2837	0.2469
KNN-fuzzy rough set with FNN	RMSE	6.9766	8.0304	8.4075	9.0160	9.4112	10.4998
	R	0.9578	0.9509	0.9518	0.9446	0.9413	0.9396
	RA	0.3924	0.3430	0.3297	0.2999	0.2854	0.2593

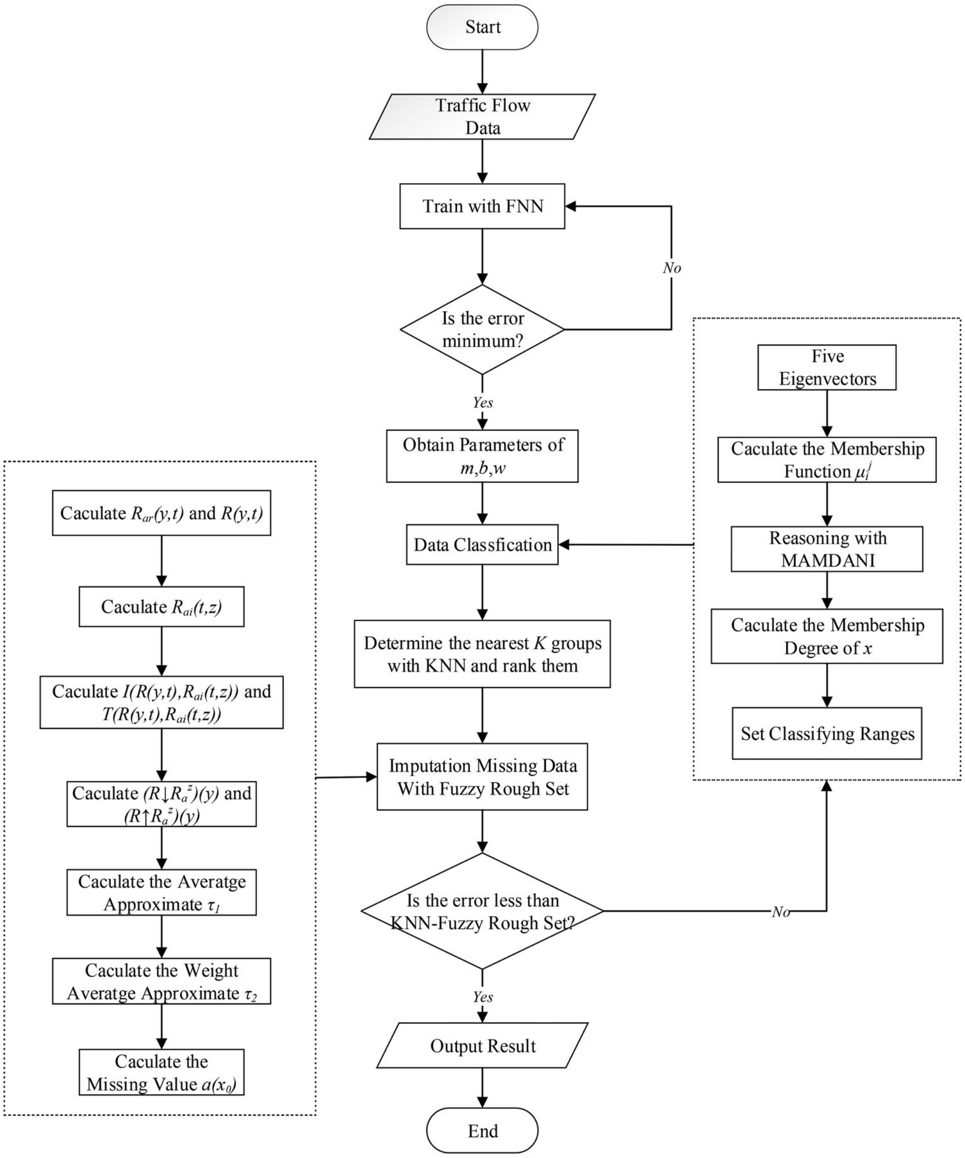
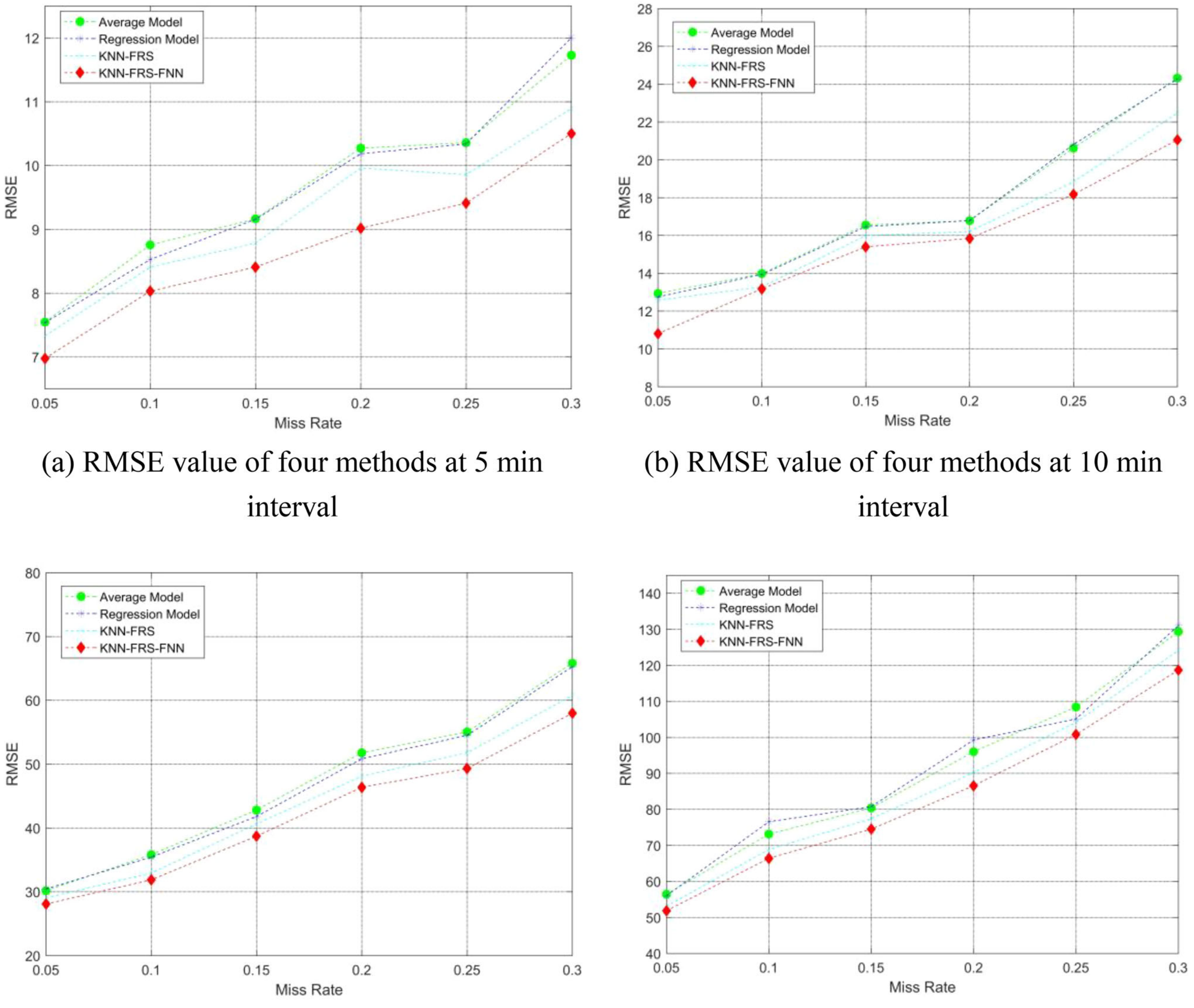

Start
Traffic Flow Data
Train with FNN
No
Is the error minimum?
Yes
Obtain Parameters of m,b,w
Data Classfication
Determine the nearest K groups with KNN and rank them
Imputation Missing Data With Fuzzy Rough Set
Is the error less than KNN-Fuzzy Rough Set?
No
Yes
Output Result
End
Five Eigenvectors
Caculate the Membership Function μ_i^j
Reasoning with MAMDANI
Caculate the Membership Degree of x
Set Classifying Ranges
Caculate $R_{ar}(y,t)$ and $R(y,t)$
Caculate $R_{ai}(t,z)$
Caculate $I(R(y,t),R_{ai}(t,z))$ and $T(R(y,t),R_{ai}(t,z))$
Caculate $(R\downarrow R_a^z)(y)$ and $(R\uparrow R_a^z)(y)$
Caculate the Averatge Approximate τ_1
Caculate the Weight Averatge Approximate τ_2
Caculate the Missing Value $a(x_0)$

Figure 1. Missing data imputation framework of this study.



(a) RMSE value of four methods at 5 min interval

(b) RMSE value of four methods at 10 min interval

(c) RMSE value of four methods at 30 min interval

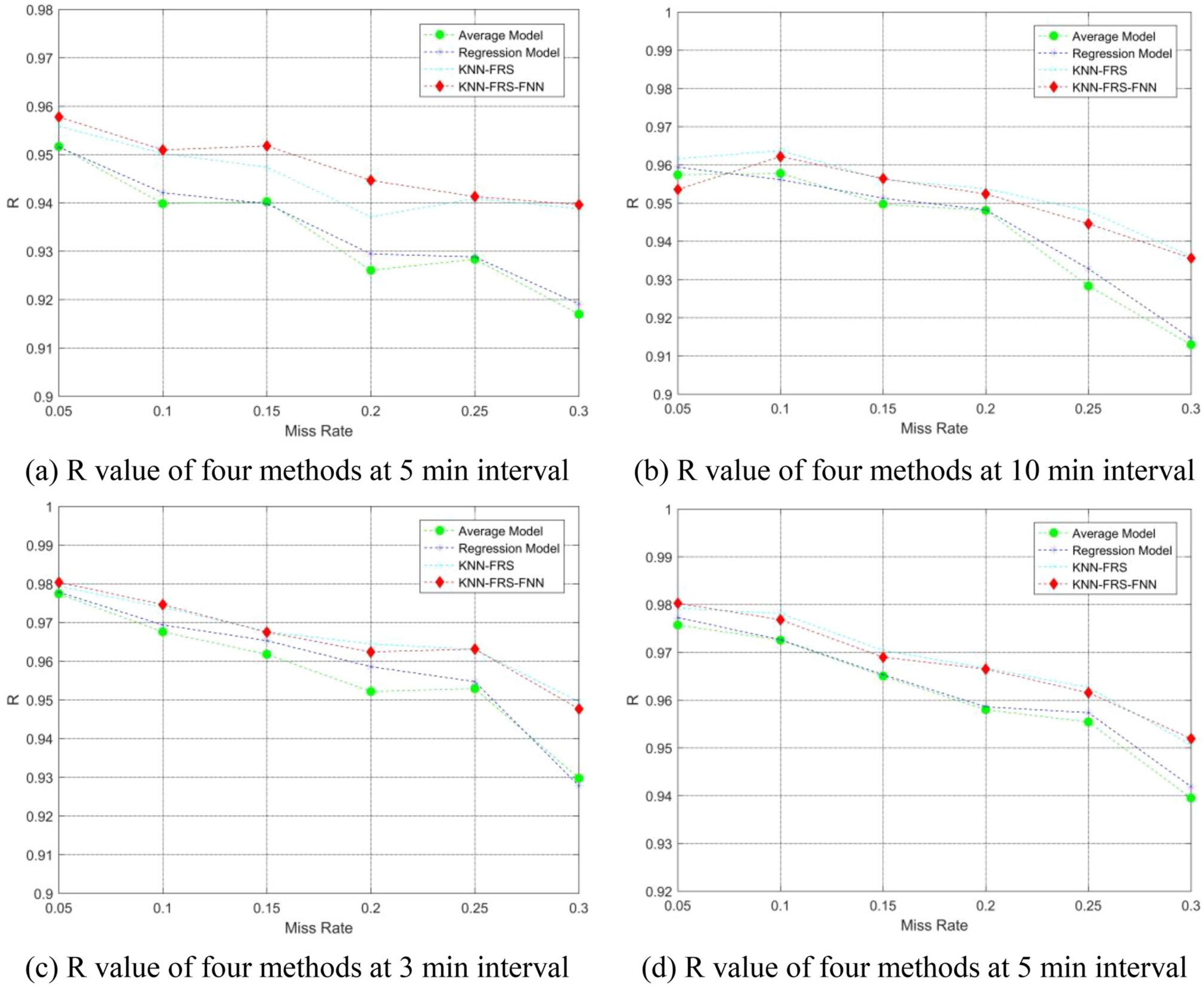
(d) RMSE value of four methods at 60 min interval

Figure 2. RMSE values for four methods at different data aggregation intervals.

three indicators of four methods are shown in Tables 2 and 3, when the missing ratio is 5%, the RMSE values of four methods under random missing are 7.5468、7.5376、7.3297、6.9766, and the RMSE values under continuous missing are 10.0556、18.7820、9.2406、8.8138. It is obviously that the RMSE values for four methods under random type are lower than that under continuous type. Furthermore, the other two indicators such as R and RA values also express similar results under two different missing types. And similar regulation can be obtained with different missing ratios and different data aggregation intervals, see Figures 5–7. The reason is that, compared with the imputation under random missing data, the adjacent section of traffic flow data set will be lost under the continuous missing type. In this condition, it is difficult to use the surrounding complete data to effectively estimate the missing data values, which will cause the information used in imputation is limited, and then result in low imputation performance.

2. The accuracy of imputation increases with the increase of data aggregated interval.

As show in Figures 4 and 7, it is evident that the values of RA raise with the increase of data aggregated interval under the same missing rate. In addition, through the results analysis in Figures 5 and 6, the same conclusion can be obtained for the values distribution R: with the values of R and RA grows, the



(a) R value of four methods at 5 min interval

(b) R value of four methods at 10 min interval

(c) R value of four methods at 3 min interval

(d) R value of four methods at 5 min interval

Figure 3. R values for four methods at different data aggregation intervals.

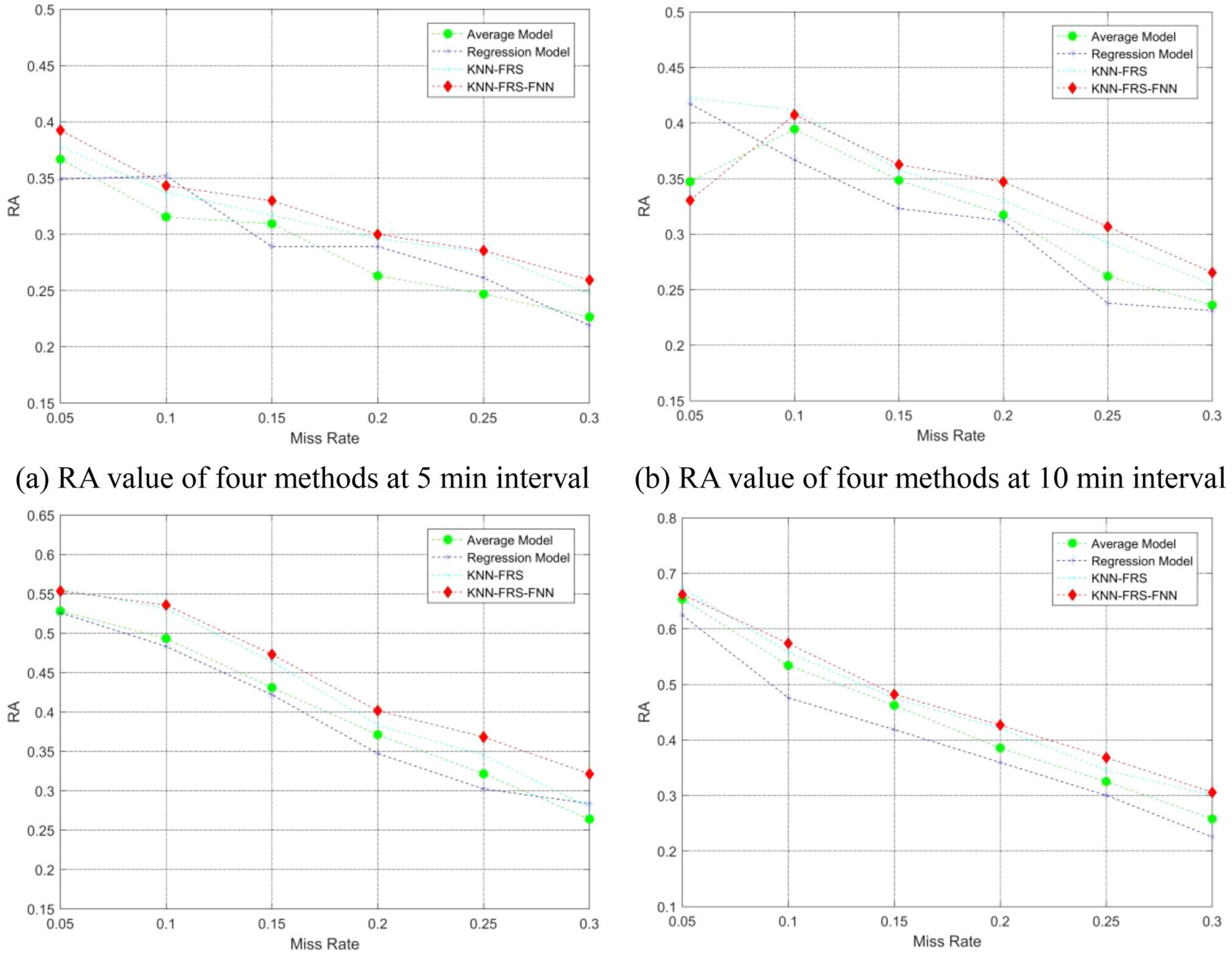
imputation accuracy of model increases. The reason is that, when traffic flow was aggregated with small interval, the data is not smooth and express fluctuation patterns frequently, which will lead to difficulty of modeling in imputation and results in the decrease of performance. As the aggregated interval of traffic flow data increases, the intense variation of traffic flow will weaken and the data will becomes smooth and stable, eventually, the imputation performance of the model gradually improves under same missing ratios.

3. The accuracy of data imputation decreases with the increase of data missing ratio.

The imputation performance of four methods under different data aggregated intervals and different data missing ratio are shown from Figures 2 to 7. For example, as expressed in Figures 2 and 5, it is observed that with the increase of data missing ratio, the RSEM values increase under different missing types: random or continuous missing. It can also find that the values of R decline with the raise of data missing ratio through Figures 3 and 6, and similar regulation can be observed with the changes of RA in Figures 4 and 7. The reason is that, with the increase of data missing ratio, the amount of data that can be used in missing data imputation will decrease gradually and the imputation performance of the model will deteriorate.

4. The imputation performance of the KNN fuzzy rough set based model is better than that of the traditional average and regression methods

As shown in Figures 2 and 5, when the data is aggregated at different interval, with the increase of missing ratio, the RSEM values of the KNN Fuzzy rough set method based on the classification with Fuzzy neural network (KnnFrsFnn) and the KNN



(a) RA value of four methods at 5 min interval

(b) RA value of four methods at 10 min interval

(c) RA value of four methods at 30 min interval

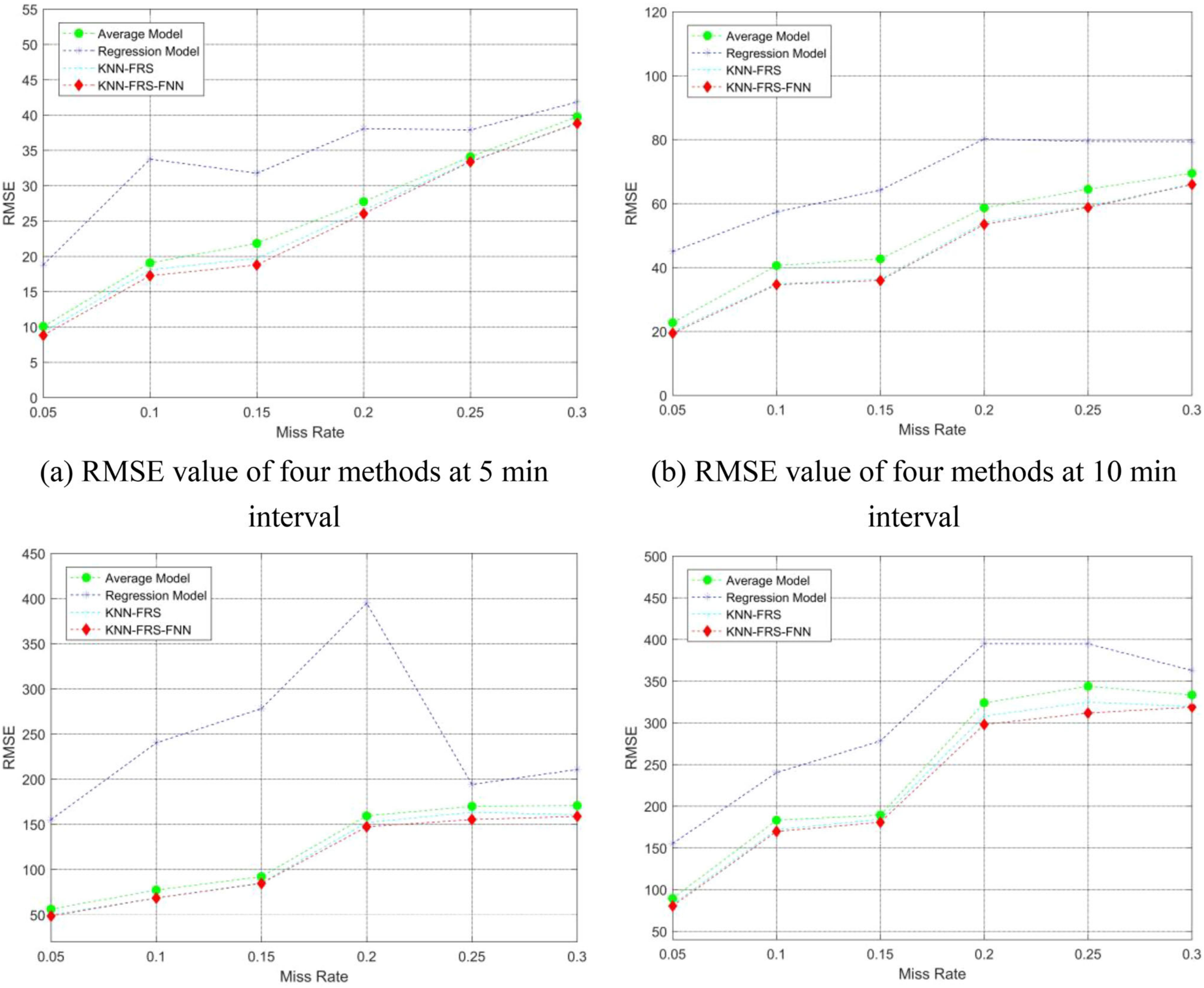
(d) RA value of four methods at 60 min interval

Figure 4. RA values for four methods at different data aggregation intervals.

Table 3. Comparison of imputation results of datasets at intervals of 5 min.

Missing ratio		5%	10%	15%	20%	25%	30%
Average model	RMSE	10.0556	19.0648	21.8378	27.7339	34.0898	39.7483
	R	0.8016	0.7721	0.7537	0.5800	0.3282	0.2152
	RA	0.1028	0.0708	0.0426	0.0676	0.0347	0.0257
Regression model	RMSE	18.7820	33.7526	31.7600	38.0581	37.9052	41.8343
	R	0.4968	0.1474	0.3043	0.2197	0.0679	0.1151
	RA	0.0389	0.0111	0.0259	0.0119	0.0073	0.0110
KNN-fuzzy rough Set	RMSE	9.2406	18.0858	19.7060	26.5323	33.4526	38.8846
	R	0.8349	0.8083	0.7849	0.6149	0.3230	0.2338
	RA	0.0944	0.0556	0.0481	0.0572	0.0374	0.0239
KNN-fuzzy rough set with FNN	RMSE	8.8138	17.2430	18.7738	26.0185	33.3846	38.8050
	R	0.8375	0.8002	0.7814	0.6253	0.3291	0.2374
	RA	0.0917	0.0792	0.0769	0.0564	0.0391	0.0297

combined with Fuzzy rough set (KnnFrs) are steadily lower than that of the average and regression methods, and the KnnFrsFnn expresses better imputation performance than KnnFrs. Similarly, as the distribution of R shown in Figures 3 and 6 and RA shown in Figures 4 and 7, with the increase of missing rate when the data is aggregated at different interval, the R and RA values of KnnFrsFnn and KnnFrs are higher than that of average and regression methods, which indicates the better imputation



(a) RMSE value of four methods at 5 min interval

(b) RMSE value of four methods at 10 min interval

(c) RMSE value of four methods at 30 min interval

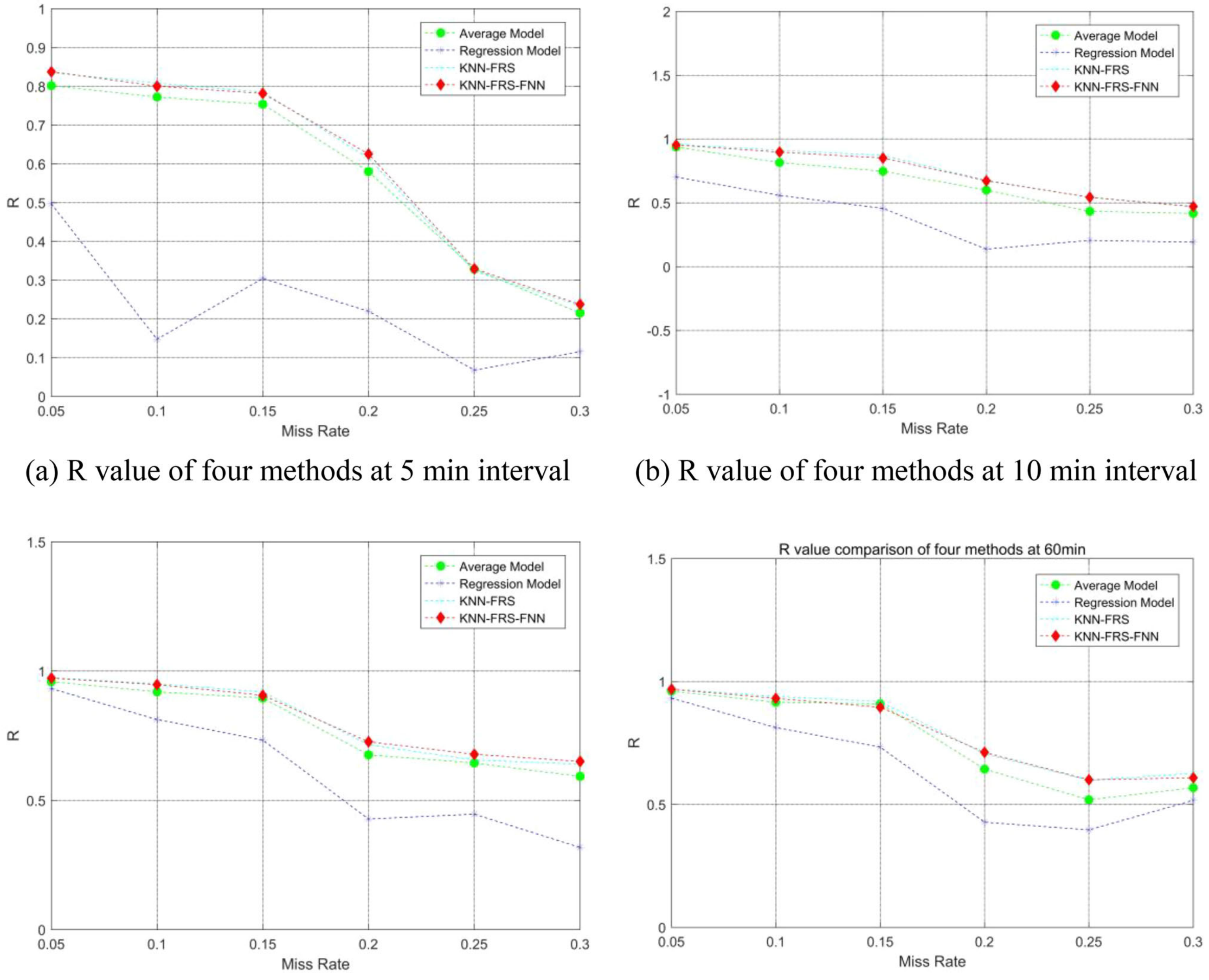
(d) RMSE value of four methods at 60 min interval

Figure 5. RMSE distribution for four methods at different data aggregation intervals.

performance of hybrid models. The FRS can effectively analyze and dealing with uncertain issues in the systems with inaccurate, inconsistent or incomplete data structure and condition. Furthermore, the R and RA values of KnnFrsFnn are higher than that of KnnFrs, and this also indicates the superiority of the proposed model by considering adaptive classification strategy based on FNN. This also shows that the classification of fuzzy neural networks can play an important role before the imputation process is conducted, for the reason that the classification of data by using fuzzy neural network allows the original dataset categorized into similar clusters, which make more reasonable information used to estimate missing values and improve the efficiency and the accuracy of the imputation.

3.3.3 Imputation performance under high data missing ratios

In this section, in order to testify the imputation performance of proposed model under high missing ratios, we further compare the imputation performance under random and continuous missing types with missing ratio including: 40%, 50%, 60%, 70%, and 80%. The imputation results evaluated by RMSE of four methods using traffic flow at 5 min interval are shown in Figure 8.



(a) R value of four methods at 5 min interval

(b) R value of four methods at 10 min interval

(c) R value of four methods at 30 min interval

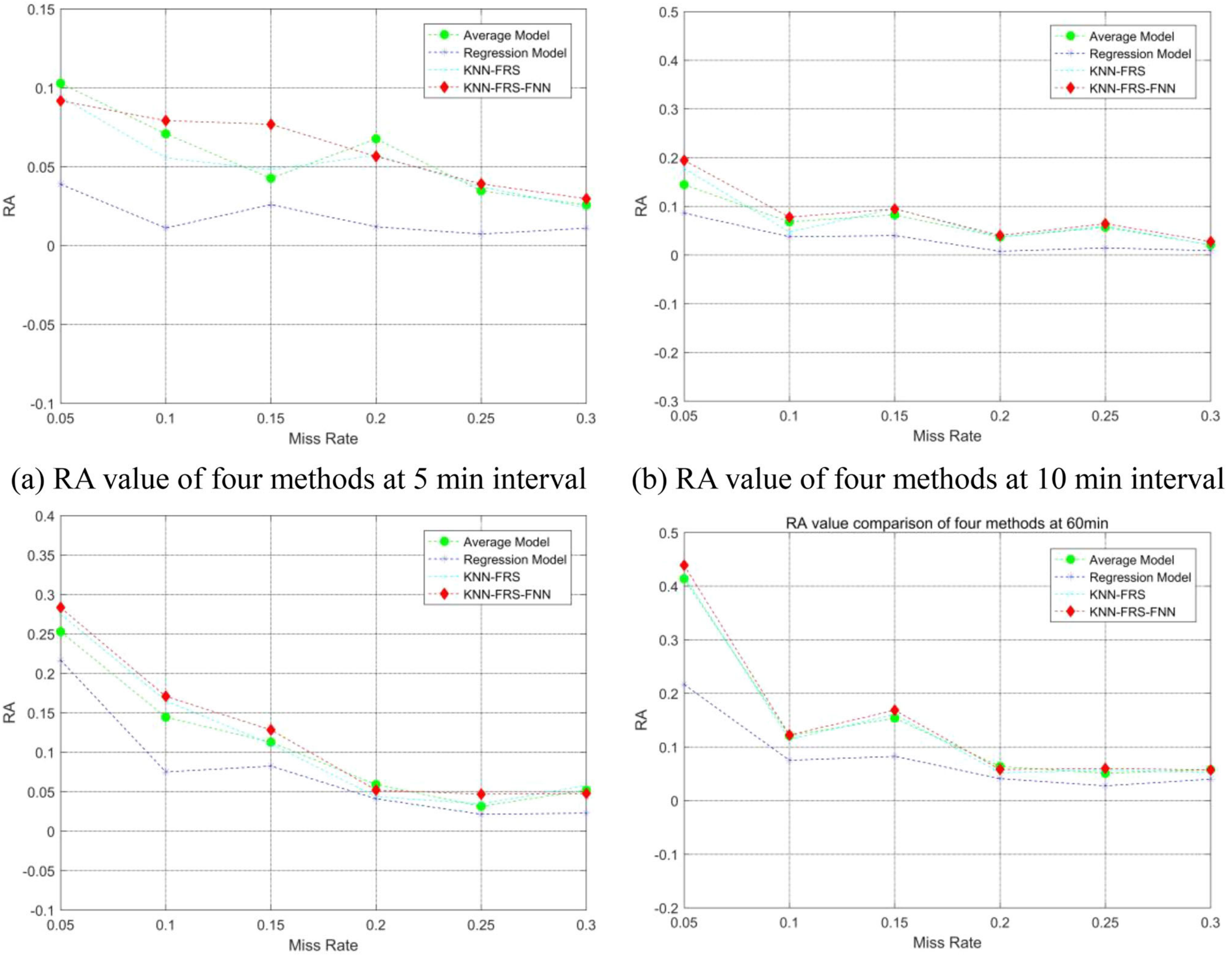
(d) R value of four methods at 60 min interval

Figure 6. R distribution for four methods at different data aggregation intervals.

For high data missing ratio, we can obtain similar comparing results as before. For example, the accuracy of imputation decreases with the increase of missing ratio, and the accuracy of imputation for the data missing at random type is higher than that at continuous type. As shown in Figure 8(a,b), with the increase of missing ratio, imputation errors of the KnnFrsFnn and the KnnFrs are steadily lower than that of the average and regression methods, and the KnnFrsFnn expresses better imputation performance than KnnFrs. For the distribution of R shown in Figure 8(c,d) and RA shown in Figure 8(e,f), with the increase of missing rate when the data is aggregated at different interval, the R and RA values of KnnFrsFnn and KnnFrs are higher than that of average and regression methods, which indicates the hybrid models have better imputation performance. Furthermore, the R and RA values of KnnFrsFnn are higher than that of KnnFrs, and this also indicates the superiority of the proposed model by considering adaptive classification strategy based on FNN. It can be also seen that, in the case of high data missing ratio, the imputation accuracy of proposed hybrid method in this study is still superior to the other three methods. This further indicates that the KnnFrsFnn can still represents stable imputation effect and application feasibility in high missing ratios.

4. Conclusion

In this study, FNN and FRS based on KNN are combined to impute the missing traffic data. First, this study use FNN to classify the original traffic flow data, and then use KNN to determine the number of data group



(a) RA value of four methods at 5 min interval

(b) RA value of four methods at 10 min interval

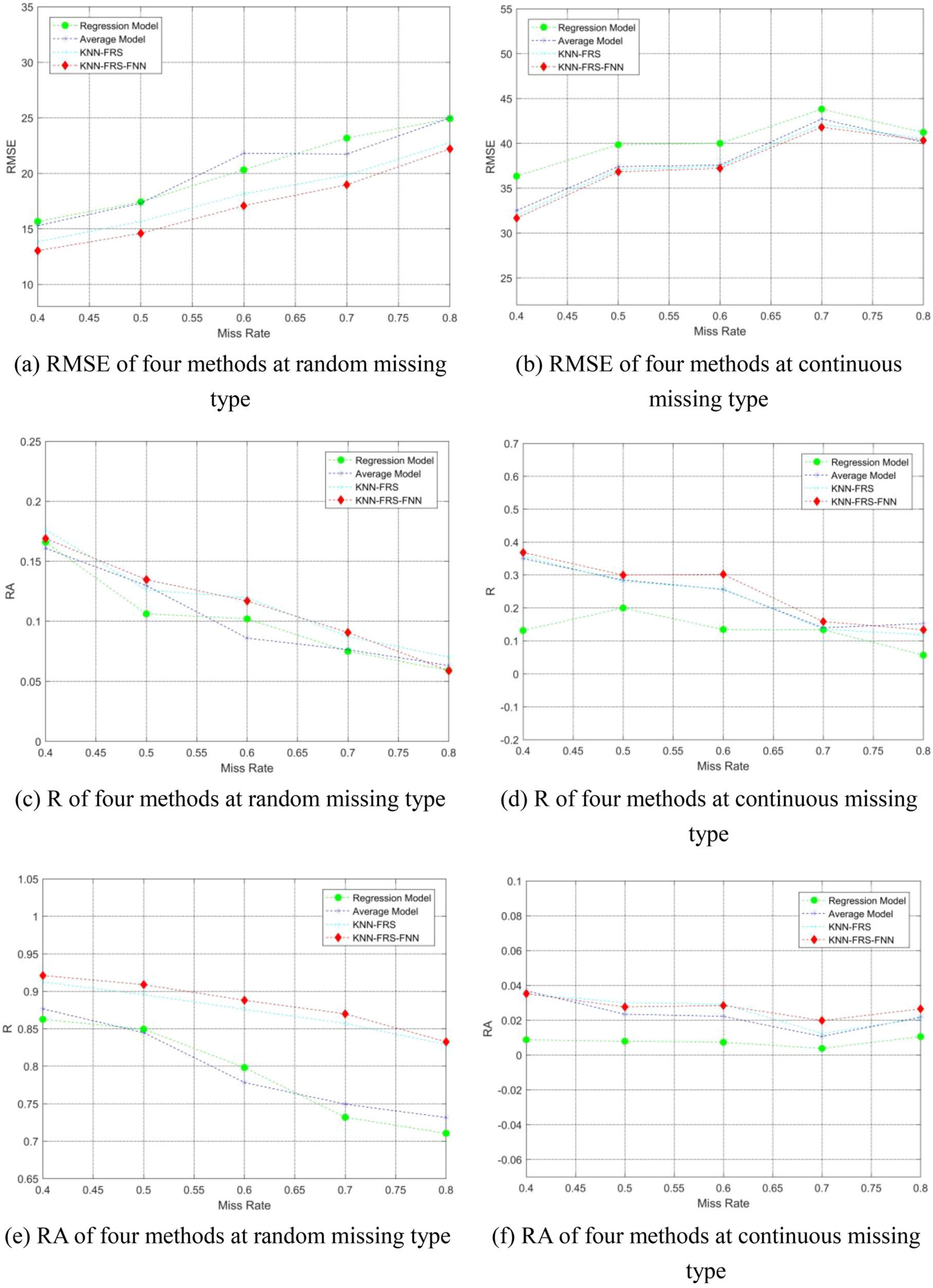
(c) RA value of four methods at 30 min interval

(d) RA value of four methods at 60 min interval

Figure 7. RA distribution for four methods at different data aggregation intervals.

related to the imputation object. Finally, this study uses FRS to impute missing traffic flow data. In order to evaluate the imputation performance, the original traffic flow data, which were aggregated at different time intervals and random or continuous missing at different ratios, are used as test data. Three different evaluation indicators, which are RMSE, R and RA, were used to calculate the imputation performance. Then we compare the imputation performance with conventional methods such as the average model and the regression model. It shows that method in this study is superior to other traditional methods and the imputation performance can be promoted under the classification of FNN. It also fully reflects the potential of the proposed method in the imputation of missing traffic flow data.

However, in this study, there are still some limitations for future study. For instance, this research only considers the case of data imputation under random and continuous types, but it did not consider the data imputation under special circumstance, such as accident condition and the influence of weather. Under these circumstances, the detectors cannot record the data normally and the traffic data will be lost in the system. Furthermore, the proposed model can only estimate for a single detector without considering the spatio-temporal correlation between multiple detectors to improve the imputation accuracy. Finally, with the increase of the missing ratio, the deterioration degree of three evaluation indicators increases quickly, in the future, great efforts should



(a) RMSE of four methods at random missing type

(b) RMSE of four methods at continuous missing type

(c) R of four methods at random missing type

(d) R of four methods at continuous missing type

(e) RA of four methods at random missing type

(f) RA of four methods at continuous missing type

Figure 8. Imputation performances for four methods at high missing ratio.

be made to improve the stability of model under high missing ratios.

Funding

This research was funded in part by the National Natural Science Foundation of China (No. 71701215), Foundation of Central South University (No. 502045002), Postdoctoral Science Foundation of China (No. 2018M630914 and 2019T120716), Innovation-Driven Project of Central South University (No. 2020CX041).

References

Ahmed, M. S., & Cook, A. R. (1979). Analysis of freeway traffic time-series data by using Box-Jenkins techniques. *Transportation Research Record: Journal of the Transportation Research Board*, *722*, 1–9.

Al-Deek, H. M., Venkata, C., & Chandra, S. R. (2004). New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse. *Transportation Research Record: Journal of the Transportation Research Board*, *1867*(1), 116–126. doi:10.3141/1867-14

Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage Publications.

Amiri, M., & Jensen, R. (2016). Missing data imputation using fuzzy-rough methods. *Neurocomputing*, *205*, 152–164. doi:10.1016/j.neucom.2016.04.015

Baharaeen, S., & Masud, A. S. (1986). A computer program for time series forecasting using single and double exponential smoothing techniques. *Computers & Industrial Engineering*, *11*(1), 151–155. doi:10.1016/0360-8352(86)90068-9

Boyles, S. (2011, January). A comparison of interpolation methods for missing traffic volume data. Proceedings of the 90th Annual Meeting of the Transportation Research Board, Washington, DC, pp. 23–27.

Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, *16*(2), 113.

Chen, X., Wang, S., Shi, C., Wu, H., Zhao, J., & Fu, J. (2019). Robust Ship Tracking via Multi-view Learning and Sparse Representation. *Journal of Navigation*, *72*(1), 176–192. doi:10.1017/S0373463318000504

Clark, S. D., Watson, S., & Redfern, E. (1993). *Application of outlier detection and missing value estimation techniques to various forms of traffic count data*. Leeds, UK: Institute of Transport Studies, University of Leeds.

De Boor, C. (1978). *A practical guide to splines*. New York: Springer-Verlag.

Duan, G., Liu, P., Chen, P., Jiang, Q., & Li, N. (2011). *Short-term traffic flow prediction based on rough set and support vector machine*. Eighth International Conference on Fuzzy Systems & Knowledge Discovery.

Fan, A. (2013). The traffic prediction and control based on rough set theory. *Advanced Materials Research*, *756–759*, 632–635. doi:10.4028/www.scientific.net/AMR.756-759.632

Gao, H., & Fasheng, L. (2009). *Combination prediction model of traffic flow based on rough set theory*. International Conference on Information Technology & Computer Science.

Ghosh, B., Basu, B., & O'Mahony, M. (2005, January). Time-series modeling for forecasting vehicular traffic flow in Dublin. *Proceedings of the 84th Annual Meeting of Transportation Research Board*, Washington, DC.

Gold, D. L., Turner, S. M., Gajewski, B. J., & Spiegelman, C. (2000, January). Imputing missing values in its data archives for intervals under 5 minutes. *Proceedings of the 80th Transportation Research Board Meeting*, Washington, DC.

Hofleitner, A., Herring, R., & Bayen, A. (2012). Arterial travel time forecast with streaming data: A hybrid approach of flow modeling and machine learning. *Transportation Research Part B: Methodological*, *46*(9), 1097–1122. doi:10.1016/j.trb.2012.03.006

Holt, C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5–10. doi:10.1016/j.ijforecast.2003.09.015

Jensen, R., & Cornelis, C. (2011). Fuzzy-rough nearest neighbour classification and prediction. *Theoretical Computer Science*, *412*(42), 5871–5884. doi:10.1016/j.tcs.2011.05.040

Kumar, B. A., Vanajakshi, L., & Subramanian, S. C. (2018). A hybrid model based method for bus travel time estimation. *Journal of Intelligent Transportation Systems*, *22*(5), 390–406. doi:10.1080/15472450.2017.1378102

Li, Z., Li, L., & Li, Y. (2013). Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies*, *34*, 108–120.

Nguyen, L. H., & Scherer, W. T. (2003). *Imputation techniques to account for missing data in support of intelligent transportation systems applications*. Technical Report UVACTS-13-0-78, University of Virginia, Center for Transportation Studies.

Ni, D., Leonard, J. D., Guin, A., & Feng, C. (2005). Multiple imputation scheme for overcoming the missing values and variability issues in ITS data. *Journal of Transportation Engineering*, *131*(12), 931–938. doi:10.1061/(ASCE)0733-947X(2005)131:12(931)

Nowicki, R. (2009). Rough neuro-fuzzy structures for classification with missing data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics))*, *39*(6), 1334–1347. doi:10.1109/TSMCB.2009.2012504

Nowicki, R. (2010). On classification with missing data using rough-neuro-fuzzy systems. *International Journal of Applied Mathematics and Computer Science*, *20*(1), 55–67. doi:10.2478/v10006-010-0004-8

Qu, L., Hu, J., Li, L., & Zhang, Y. (2009). PPCA-based missing data imputation for traffic flow volume: A systematical approach. *IEEE Transactions on Intelligent Transportation Systems*, *10*(3), 512–522.

Qu, L., Zhang, Y., Hu, J., Jia, L., & Li, L. (2008). A BPCA based missing value imputing method for traffic flow volume data. Paper presented at the 985-990.

Ramsey, B., & Hayden, G. (1994). AutoCounts: A way to analyse automatic traffic count data. *Traffic Engineering & Control*, *35* (4), 245.

Ran, B., Tan, H., Feng, J., Wang, W., Cheng, Y., & Jin, P. (2016). Estimating missing traffic volume using low multilinear rank tensor completion. *Journal of Intelligent Transportation Systems*, *20*(2), 152–161. doi:10.1080/15472450.2015.1015721

Redfern, E. J., Waston, S. M., Tight, M. R., & Clark, S. D. (1993). A comparative assessment of current and new techniques for detecting outliers and estimating missing values in transport related time series data. Proceedings of Highways and Planning Summer Annual Meeting, Institute of Science and Technology, University of Manchester, England.

Tan, H., Feng, J., Feng, G., Wang, W., Zhang, Y., & Li, F. (2013). A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies*, *28*, 15–27. doi:10.1016/j.trc.2012.12.007

Tang, J., Chen, X., Hu, Z., Zong, F., Han, C., & Li, L. (2019). Traffic flow prediction based on combination of support vector machine and data denoising schemes. *Physica A: Statistical Mechanics and Its Applications*, *534*, 120642, 1–19. doi:10.1016/j.physa.2019.03.007

Tang, J., Liu, F., Zou, Y., Zhang, W., & Wang, Y. (2017). An improved fuzzy neural network for traffic speed prediction considering periodic characteristic. *IEEE Transactions on Intelligent Transportation Systems*, *18*(9), 2340–2350. doi:10.1109/TITS.2016.2643005

Tang, J., Zhang, G., Wang, Y., Wang, H., & Liu, F. (2015). A hybrid approach to integrate fuzzy c-means based imputation method with genetic algorithm for missing traffic volume data estimation. *Transportation Research Part C: Emerging Technologies*, *51*, 29–40. doi:10.1016/j.trc.2014.11.003

Tight, M. R., Redfern, E. J., & Watson, S. M. (1993). *Outlier detection and missing value estimation in time series traffic count data*. Leeds, UK: Institute of Transport Studies, University of Leeds.

Wilby, M. R., Díaz, J. J. V., Rodríguez GonzĹez, A. B., & Sotelo, M. Á. (2014). Lightweight occupancy estimation on freeways using extended floating car data. *Journal of Intelligent Transportation Systems*, *18*(2), 149–163. doi:10.1080/15472450.2013.801711

Williams, B. M., & Hoel, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of Transportation Engineering*, *129*(6), 664–672. doi:10.1061/(ASCE)0733-947X(2003)129:6(664)

Xie, Y., & Huynh, N. (2010). Kernel-based machine learning models for predicting daily truck volume at seaport terminals. *Journal of Transportation Engineering*, *136*(12), 1145–1152.

Yan, Y., Zhang, S., Tang, J., & Wang, X. (2017). Understanding characteristics in multivariate traffic flow time series from complex network structure. *Physica A: Statistical Mechanics and Its Applications*, *477*, 149–160. doi:10.1016/j.physa.2017.02.040

Zhang, W., Yu, Y., Qi, Y., Shu, F., & Wang, Y. (2019). Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning. *Transportmetrica A: Transport Science*, *15*(2), 1688–1711. doi:10.1080/23249935.2019.1637966

Zhang, W., Zou, Y., Tang, J., Ash, J., & Wang, Y. (2016). Short-term prediction of vehicle waiting queue at ferry terminal based on machine learning method. *Journal of Marine Science and Technology*, *21*(4), 729–741. doi:10.1007/s00773-016-0385-y

Zhong, M., Lingras, P., & Sharma, S. (2004). Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. *Transportation Research Part C*, *12*(2), 139–166. doi:10.1016/j.trc.2004.07.006