Hindawi
Journal of Sensors
Volume 2018, Article ID 7465026, 11 pages
https://doi.org/10.1155/2018/7465026

Research Article

A Novel Method for Air Quality Data Imputation by Nuclear Norm Minimization

Xiaobo Chen[1,2] **and Yan Xiao**[3]

[1]*Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China*
[2]*School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China*
[3]*School of Chemistry and Chemical Engineering, Jiangsu University, Zhenjiang 212013, China*

Correspondence should be addressed to Xiaobo Chen; xbchen82@gmail.com

Received 23 December 2017; Accepted 10 April 2018; Published 26 April 2018

Academic Editor: Fanli Meng

Missing data is a frequently encountered problem in environment research community. To facilitate the analysis and management of air quality data, for example, $PM_{2.5}$ concentration in this study, a commonly adopted strategy for handling missing values in the samples is to generate a complete data set using imputation methods. Many imputation methods based on temporal or spatial correlation have been developed for this purpose in the existing literatures. The difference of various methods lies in characterizing the dependence relationship of data samples with different mathematical models, which is crucial for missing data imputation. In this paper, we propose two novel and principled imputation methods based on the nuclear norm of a matrix since it measures such dependence in a global fashion. The first method, termed as global nuclear norm minimization (GNNM), tries to impute missing values through directly minimizing the nuclear norm of the whole sample matrix, thus at the same time maximizing the linear dependence of samples. The second method, called local nuclear norm minimization (LNNM), concentrates more on each sample and its most similar samples which are estimated from the imputation results of the first method. In such a way, the nuclear norm minimization can be performed on those highly correlated samples instead of the whole sample matrix as in GNNM, thus reducing the adverse impact of irrelevant samples. The two methods are evaluated on a data set of $PM_{2.5}$ concentration measured every 1 h by 22 monitoring stations. The missing values are simulated with different percentages. The imputed values are compared with the ground truth values to evaluate the imputation performance of different methods. The experimental results verify the effectiveness of our methods, especially LNNM, for missing air quality data imputation.

1. Introduction

During the last decades, a large amount of air quality data which reflect significant pollutant concentrations have been collected by air quality monitoring stations distributed over a certain area. Due to the adverse effects of pollutants on the environment and human health, the analysis of air quality data plays an important role for environment protection and pollution treatment. However, because of many uncontrollable factors, such as instrument faults, communication, and processing errors, these data often suffer from missing values or incomplete samples [1, 2] with different proportions, thus causing serious difficulties for subsequent data analysis and decision making. For instance, many standard data analysis methods, such as neural networks [3, 4] and support vector machines [5, 6], are not applicable since they can only work on complete data.

According to [7, 8], the missing data mechanism can be categorized into three cases: (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) missing not at random (MNAR). For MCAR, the missing values are completely independent of each other and thus appear as a few isolated points. For MAR, the missing values are related to each other in a neighborhood and thus appear as a group of values lost at a time. For MNAR, the occurrence of missing values has specific patterns, for example, the

pattern caused by a long time malfunction of monitoring station. In this paper, we mainly focus on the first two cases since the last case is too restrictive in realities [9].

Traditional method to handle missing data is discarding those samples with missing values, which is generally called listwise deletion. However, this method will cause information loss because the observed values in the incomplete samples are actually informative. Moreover, the analytical results drawn from listwise method may be biased since the data distribution might be altered after deletion, especially in the case of high proportion of incomplete samples. Therefore, instead of deleting all incomplete samples, data imputation which replaces the missing values with probable values estimated by different methods has attracted much attention of air quality research community recently.

The imputation methods can be roughly divided into two categories: single imputation [10, 11] and multiple imputation [12, 13]. Single imputation methods estimate a single value for each missing element whereas multiple imputation methods generate multiple possible values for each missing element, thus reflecting the uncertainty of estimation results. In this study, we concentrate on single imputation because it is more convenient to integrate with popular data analysis tools. So far, many single imputation methods have been explored for various application fields, such as probabilistic principal component analysis (PPCA) [14, 15], expectation-maximization (EM) [16], and neural networks [17, 18]. PPCA is based on probabilistic latent variable model, making an assumption that the observed high-dimensional data are sampled from a low-dimensional intrinsic subspace. Both the intrinsic subspace and the missing values can be jointly solved through maximum likelihood estimation (MLE). Neural networks are a regression-based model missing value imputation approach where the relationship between the observed values and the missing value is characterized by neural network model. Among these methods, station mean (SM) [19] is a typical single imputation method, which imputes the missing value by computing the mean of observed values measured at the same time by the other monitoring stations. This method is actually based on the spatial correlation, which utilizes the relatedness of air quality samples measured at different monitoring stations. On the other hand, the nearest neighbor- (NN-) based imputation let the missing value equal the value of the closest sample in time. It assumes that the time series of daily air quality data have strong local temporal correlation, which can be used to impute the missing values. The extensions of NN imputation include linear interpolation and cubic spline imputation, which characterize the locally temporal relationship between nearby air quality samples using more complex models.

It can be observed from the existing works that the missing data imputation closely depends on certain prior assumption about air quality data. For example, NN imputation method supposes that the local pattern variation of air quality time series is constant, linear, or cubic with time. In this paper, we suppose that the air quality samples measured at different time points or different stations in a certain area should be dependent to each other as in NN and SM methods. This dependence also imposes a prior structure on air quality data. In the case of missing data, we can recover such a prior structure and missing data based on the data which are observable. Therefore, we aim to impute the missing values by making the linear dependence of the resulting complete data matrix in terms of rows and columns as large as possible. From this perspective, characterizing the linear dependence of the rows and columns of a data matrix plays an important role for missing data imputation.

As is well known, the rank of a matrix [20] characterizes the number of linearly independent rows or columns of a matrix. The lower the rank, the more the rows and columns are linearly dependent. Usually, the matrix with rank less than the number of rows and columns is said to be rank deficient or not full rank. Therefore, rank is an interesting quantity when comparing different low-rank matrices but it lacks sufficient description of dependence strength for general matrices. In addition, minimizing the rank of a matrix is generally a NP-hard problem [21, 22] difficult to solve. Different from the rank, the nuclear norm of a matrix also provides a well measurement of the linear dependence of the rows and columns. Moreover, the nuclear norm is the best convex approximation of the matrix rank over the unit ball of matrices, thus leading to efficient optimization algorithm and preferable globally optimal solution [21, 23, 24]. Due to these advantages, the nuclear norm has been widely applied in various fields, such as image processing [25, 26] and bioinformatics [27].

Inspired by the above discussion, in this paper, we propose two new and principled imputation methods for air quality data by utilizing the nuclear norm of a matrix to measure the inherent dependence of samples. The first method, called global nuclear norm minimization (GNNM), minimizes the nuclear norm directly on the whole air quality data in order to impute the missing values. This method is relatively simple but may produce suboptimal estimation of missing values, especially when the data is not strongly dependent in a global way. To deal with this problem, we further propose local nuclear norm minimization (LNNM) by introducing the local similarity in order to improve imputation accuracy of air quality data. Specifically, LNNM consists of two steps. The first step aims to gain a rough estimation of missing values by using the above GNNM. Then, we concentrate more on each air quality sample and select its k most similar samples, that is, k nearest neighbors, based on the above rough estimation. The nuclear norm minimization is performed on the highly correlated sample subset comprising of this sample and its k most similar samples so as to refine the estimation of missing values for this sample. The above refinement procedure is conducted for each air quality sample, thus resulting in final estimation of the whole missing data. Note that although nuclear norm minimization has been employed for some missing value imputation problems, such as traffic flow data [28], it was seldom investigated for air quality data imputation.

The paper is organized as follows: in Section 2, the real-world $PM_{2.5}$ concentration data is described and the proposed GNNM and LNNM methods are presented. In Section

3, we report and analyze the imputation performance of different methods. Finally, we conclude this paper in Section 4 by summarizing the main results of this study.

2. Data and Methods

2.1. Data. In this study, we consider $PM_{2.5}$ concentration measured every 1 h by 22 air quality monitoring stations distributed over the metropolitan area of Beijing, China, for 19 days, April 2013 [29, 30]. The whole data matrix consists of 22 rows and 456 columns, thus generating totally 10,032 measures. Each row i denotes the ith monitoring stations, and each column j denotes the readings of all monitoring stations at a particular time point j. The structure of data is shown in Table 1. The $PM_{2.5}$ concentration measured at a station in one day is viewed as an air quality sample, thus comprising 24 measurements.

2.2. Global Nuclear Norm Minimization. First of all, we give some notations that will be used. Let the air quality sample, that is, $PM_{2.5}$ concentration, at the ith $(1 \le i \le 22)$ monitoring station for the jth $(1 \le j \le 19)$ day be denoted by $\mathbf{x}_{i,j} \in R^{24}$, since the concentration is measured on a time scale of one per hour. Notice that some elements in $\mathbf{x}_{i,j}$ are missing. All air quality samples $\{\mathbf{x}_{i,j}\}$ recorded by different monitoring stations and at different times can be organized in matrix form as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{1,1} & \cdots & \mathbf{x}_{1,19} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_{22,1} & \cdots & \mathbf{x}_{22,19} \end{bmatrix} \in R^{22\times 456}. \quad (1)$$

Table 1 illustrates the specific structure of $\mathbf{X}$ used in this study. In terms of missing value estimation problem, let Ω be the set of indexes (i, j) with observable $PM_{2.5}$ concentration $\{(i, j) \in \Omega \mid \mathbf{X}_\Omega \text{ is observable}\}$. The complementary set of Ω is denoted by $\bar{\Omega}$, where the values $\{\mathbf{X}_{\bar{\Omega}} \mid (i, j) \in \bar{\Omega}\}$ are missing. Thus, we have $\mathbf{X} = \mathbf{X}_\Omega \cup \mathbf{X}_{\bar{\Omega}}$.

As discussed in the Introduction, the nuclear norm is a convex surrogate of the matrix rank, which can characterize the linear dependence of the rows and columns of general matrix and at the same time is computationally efficient. Therefore, we propose the following nuclear norm minimization problem to recover a complete matrix $\mathbf{X}^{(1)}$ from incomplete data $\mathbf{X}$.

$$\begin{aligned} \mathbf{X}^{(1)} &= \arg\min_{\mathbf{X}} \|\mathbf{X}\|_*, \\ \text{s.t. } & \mathbf{X}_\Omega = \mathbf{D}_\Omega, \end{aligned} \quad (2)$$

where $\|\mathbf{X}\|_*$ denotes the nuclear norm (the sum of the singular values) of matrix $\mathbf{X}$ and $\mathbf{D}_\Omega$ denotes the observed values. Through solving (2), we can get optimal solution $\mathbf{X}^{(1)}$, where $\mathbf{X}_\Omega^{(1)} (= \mathbf{D}_\Omega)$ is the observed data, and thus, $\mathbf{X}_{\bar{\Omega}}^{(1)}$ gives the estimation of missing data $\mathbf{X}_{\bar{\Omega}}$. In this paper, we use singular value thresholding (SVT) algorithm [23] to solve (2).

Table 1: Data matrix structure.

Station	1	2	3	4	5	6	7	⋯	456
St1	223	227	233	233	224	224	210	⋯	62
St2	178	187	187	185	185	185	188	⋯	39
St3	193	200	200	198	198	198	192	⋯	111
St4	181	186	186	188	188	188	185	⋯	158
St5	183	185	185	191	201	201	202	⋯	163
St6	198	206	218	218	225	225	231	⋯	67
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
St22	193	193	200	203	229	233	244	⋯	108

2.3. Local Nuclear Norm Minimization. The effectiveness of (2), in terms of recovering missing data $\mathbf{X}_{\bar{\Omega}}$, closely depends on the strength of linear dependence of the rows and columns of $\mathbf{X}$. As a result, the recovered values $\mathbf{X}_{\bar{\Omega}}^{(1)}$ from (2) may be suboptimal if some irrelevant air quality samples $\mathbf{x}_{i,j}$ exist in the whole set $\mathbf{X}$. Therefore, estimating the correlation between samples is significant for imputation. However, due to a large number of missing values in $\mathbf{X}$, it is impossible to obtain an exact estimation of correlation between two air quality samples $\mathbf{x}_{i,j}$ and $\mathbf{x}_{p,q}$ a priori. Nevertheless, through solving (2), we obtain a rough estimation $\mathbf{X}^{(1)}$ of $\mathbf{X}$, that is, $\mathbf{X}_{\bar{\Omega}}^{(1)}$ of $\mathbf{X}_{\bar{\Omega}}$. Subsequently, the estimation of correlation between $\mathbf{x}_{i,j}$ and $\mathbf{x}_{p,q}$ can be inferred based on similarity metric on $\mathbf{X}^{(1)}$. Finally, we can refine the missing value estimation for each air quality sample $\mathbf{x}_{i,j}$ by only selecting those samples most similar to $\mathbf{x}_{i,j}$ and performing (2) on only those highly similar samples.

Specifically, suppose the estimation of $\mathbf{x}_{i,j}$ in $\mathbf{X}$ is corresponding to $\mathbf{x}_{i,j}^{(1)}$ in $\mathbf{X}^{(1)}$, then a similarity between $\mathbf{x}_{i,j}^{(1)}$ and $\mathbf{x}_{p,q}^{(1)}$ is denoted by

$$\mathrm{SIM}\left(\mathbf{x}_{i,j}^{(1)}, \mathbf{x}_{p,q}^{(1)}\right) = \left(\left\|\mathbf{x}_{i,j}^{(1)} - \mathbf{x}_{p,q}^{(1)}\right\|_2\right)^{-1}, \quad (3)$$

where $\|\bullet\|_2$ denotes the l_2 norm of a vector. Equation (3) means that if two air quality samples estimated from (2) are close to each other in distance, they will have large similarity and the missing values of the highly correlated samples can be imputed more reliably. Therefore, for each sample $\mathbf{x}_{i,j}^{(1)}$, we can compute its similarities with all the remaining samples $\mathbf{x}_{p,q}^{(1)}$, $p \ne i, q \ne j$, and sort these sample in order of magnitude from large to small, that is,

$$\mathrm{SIM}\left(\mathbf{x}_{i,j}^{(1)}, \mathbf{x}_{i_1,j_1}^{(1)}\right) \ge \mathrm{SIM}\left(\mathbf{x}_{i,j}^{(1)}, \mathbf{x}_{i_2,j_2}^{(1)}\right) \ge \mathrm{SIM}\left(\mathbf{x}_{i,j}^{(1)}, \mathbf{x}_{i_3,j_3}^{(1)}\right) \ge \cdots. \quad (4)$$

Suppose the similarity between $\mathbf{x}_{i,j}^{(1)}$ and $\mathbf{x}_{p,q}^{(1)}$ is consistent with the similarity between $\mathbf{x}_{i,j}$ and $\mathbf{x}_{p,q}$ to a certain degree, the top k most similar samples $\mathbf{G} = [\mathbf{x}_{i_1,j_1}, \mathbf{x}_{i_2,j_2}, \ldots, \mathbf{x}_{i_k,j_k}]$

Table 2: Descriptive statistics.

Statistics	Percentage of missing values				
	10%	20%	30%	40%	50%
Number of valid values	9043	8056	7058	6010	4989
Number of missing values	989	1976	2974	4022	5043
Mean	128.62	128.08	128.23	127.63	128.20
Standard deviation	69.69	69.72	69.54	69.39	69.42
Skewness	0.2581	0.2611	0.2477	0.2467	0.2448
Kurtosis	2.4551	2.4469	2.4409	2.4347	2.4372
Range	338	338	338	338	338
Maximum value	348	348	348	348	348
Minimum value	10	10	10	10	10

Table 3: Performance of different methods for 10% missing values.

Method		Mean	STD	p value
SM	RMSE	22.77	0.92	$9.75e-13$
	R^2	0.8941	0.0083	$7.98e-13$
NN	RMSE	21.11	1.16	$5.00e-10$
	R^2	0.9107	0.0097	$1.14e-09$
GNNM	RMSE	19.12	0.99	$7.52e-08$
	R^2	0.9288	0.0081	$5.92e-07$
LNNM	RMSE	15.24	0.90	—
	R^2	0.9536	0.0058	—

Table 4: Performance of different methods for 20% missing values.

Method		Mean	STD	p value
SM	RMSE	22.82	0.52	$1.14e-14$
	R^2	0.8921	0.0045	$7.05e-16$
NN	RMSE	21.42	0.77	$7.39e-12$
	R^2	0.9071	0.0058	$1.38e-12$
GNNM	RMSE	19.57	0.60	$7.99e-10$
	R^2	0.9249	0.0045	$1.06e-09$
LNNM	RMSE	15.73	0.78	—
	R^2	0.9501	0.0048	—

Table 5: Performance of different methods for 30% missing values.

Method		Mean	STD	p value
SM	RMSE	22.92	0.43	$8.79e-16$
	R^2	0.8914	0.0036	$4.25e-17$
NN	RMSE	21.98	0.79	$2.47e-12$
	R^2	0.9025	0.0064	$1.26e-12$
GNNM	RMSE	20.08	0.32	$5.29e-12$
	R^2	0.9217	0.0026	$8.80e-12$
LNNM	RMSE	16.52	0.60	—
	R^2	0.9454	0.0038	—

Table 6: Performance of different methods for 40% missing values.

Method		Mean	STD	p value
SM	RMSE	22.88	0.45	$5.13e-15$
	R^2	0.8918	0.0041	$1.28e-15$
NN	RMSE	22.87	0.61	$7.68e-14$
	R^2	0.8948	0.0055	$1.01e-13$
GNNM	RMSE	20.59	0.26	$4.12e-12$
	R^2	0.9185	0.0029	$1.21e-10$
LNNM	RMSE	17.57	0.50	—
	R^2	0.9388	0.0037	—

Table 7: Performance of different methods for 50% missing values.

Method		Mean	STD	p value
SM	RMSE	23.09	0.34	$2.26e-14$
	R^2	0.8899	0.0034	$1.88e-15$
NN	RMSE	24.31	0.62	$5.36e-14$
	R^2	0.8814	0.0060	$3.70e-14$
GNNM	RMSE	21.43	0.26	$9.53e-08$
	R^2	0.9127	0.0027	$2.65e-09$
LNNM	RMSE	19.01	0.45	—
	R^2	0.9295	0.0033	—

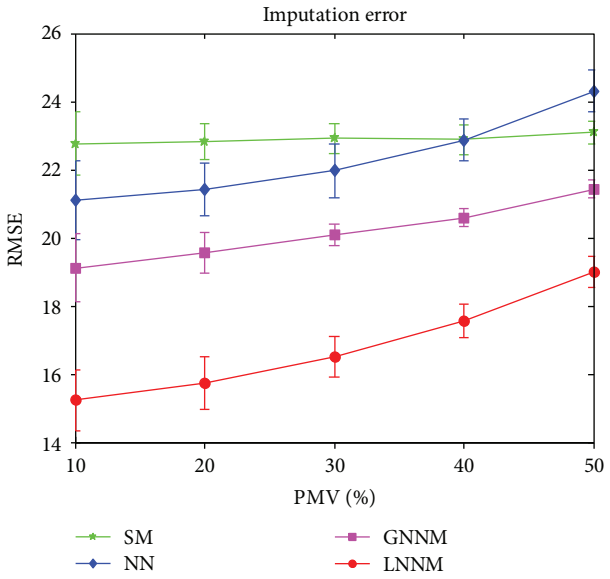

Figure 1: The variation of imputation RMSE versus the percentage of missing values.

$\in R^{k\times 24}$ in terms of (4) are selected and combined with $\mathbf{x}_{i,j}$ to form a new sample matrix $\mathbf{X}'$ with missing values.

$$\mathbf{X}' = \begin{bmatrix} \mathbf{x}_{i,j} \\ \mathbf{G} \end{bmatrix} \in R^{(k+1)\times 24}. \quad (5)$$

Finally, we can solve the following nuclear norm minimization problem on the incomplete data matrix $\mathbf{X}'$ instead of

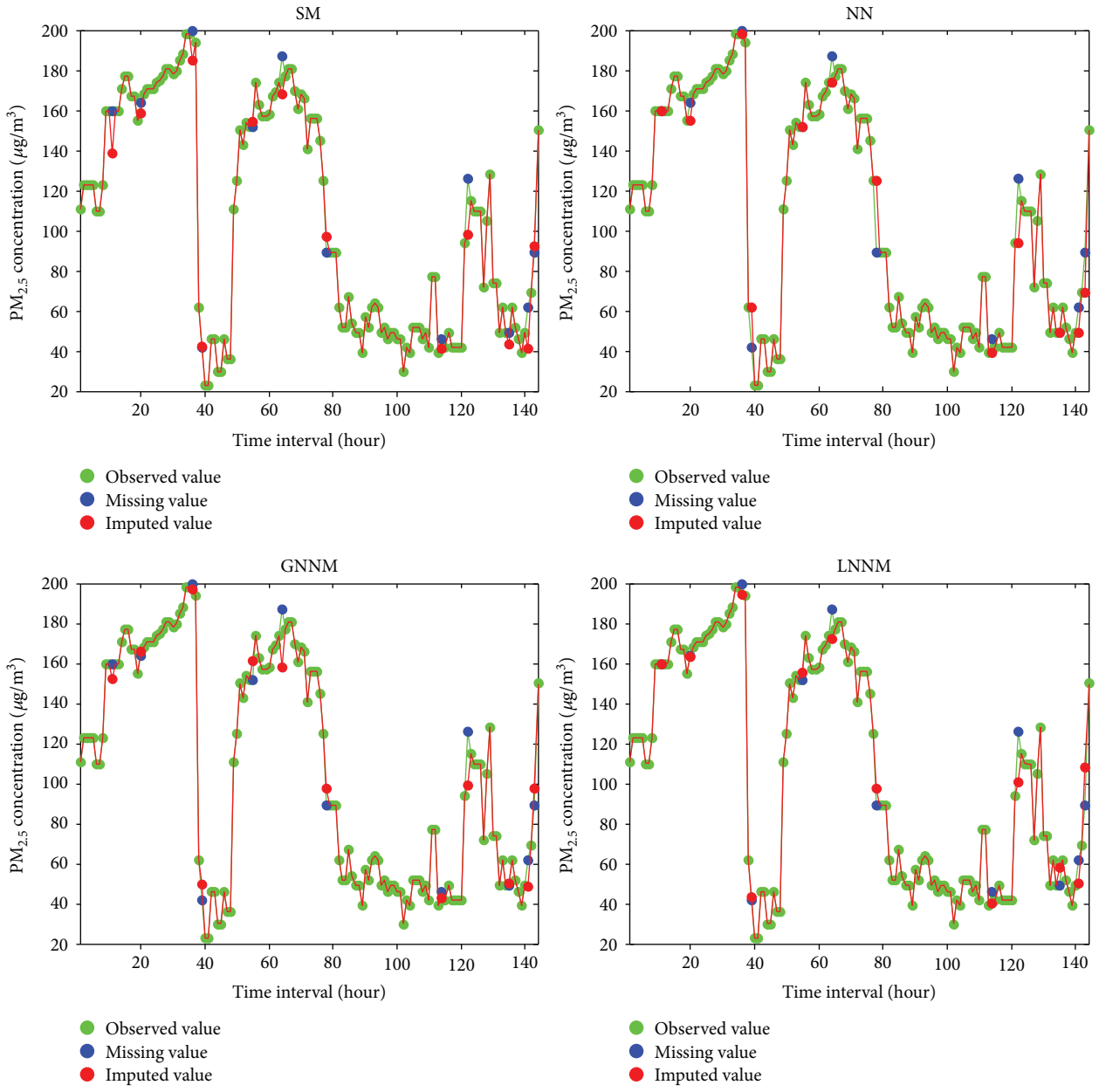

Figure 2: Imputed $PM_{2.5}$ concentration when the percentage of missing values equals 10%.

the original whole data matrix $\mathbf{X}$ to get a refined imputation for $\mathbf{x}_{i,j}$.

$$\mathbf{X}^{(2)} = \underset{\mathbf{X}'}{\operatorname{argmin}} \left\| \mathbf{X}' \right\|_*, \quad \text{s.t. } \mathbf{X}_{\Omega'}' = \mathbf{D}'_{\Omega}, \tag{6}$$

where $\mathbf{D}'_{\Omega}$ refers to the observed values in the combined matrix $\mathbf{X}'$, respectively. The refined estimation $\mathbf{x}_{i,j}^{(2)}$ for $\mathbf{x}_{i,j}$ is thus given by the first row of $\mathbf{X}^{(2)}$. The above procedure is performed for each air quality sample $\mathbf{x}_{i,j}$, and finally, we can obtain the refined estimation of all missing values in $\mathbf{X}$.

3. Experimental Results and Analysis

3.1. Experimental Configuration. The proposed GNNM and LNNM are compared with typical SM and NN imputation methods. To evaluate the imputation performance of different methods, the complete data described in Section 2 is used as ground truth test. We randomly generate missing data in which the percentage of missing values (PMV) changes from 10% to 50%. Obviously, the larger PMV, the harder the imputation problem. Different imputation method is then applied on the incomplete data such that the missing values could be estimated based on the observed values. Finally, the imputed values and the ground truth values are compared in order to evaluate the imputation performance. To reduce the possible bias, the above

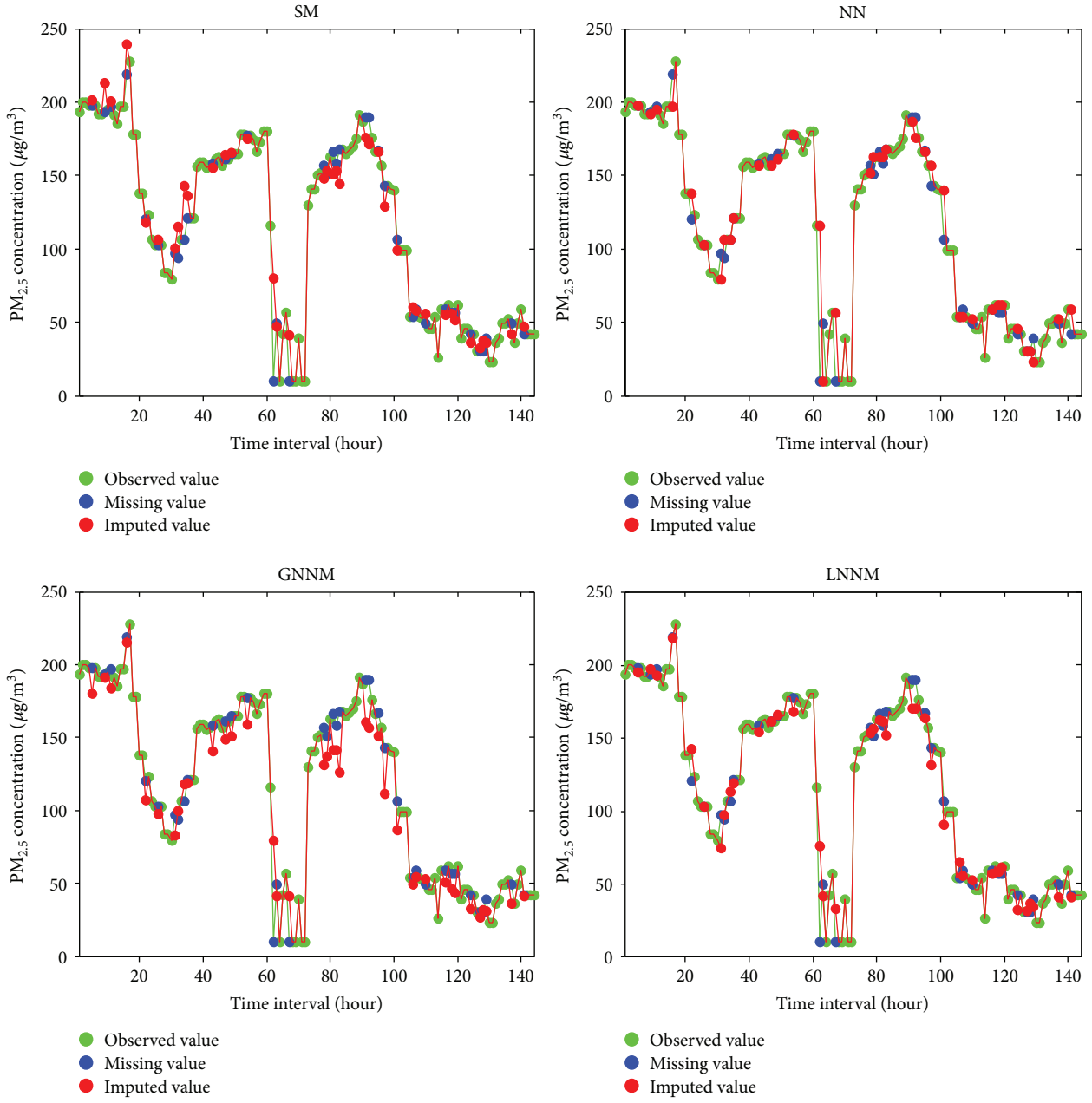

Figure 3: Imputed $PM_{2.5}$ concentration when the percentage of missing values equals 30%.

evaluation procedure is repeated 10 times and the averaged results are recorded.

In the experiments, we adopt two indices [10, 18, 31] which are widely used to compare different imputation methods. Let M denote the number of missing values and P_i and Q_i $(1 \le i \le M)$ stand for the ith imputed and observed value, respectively. Then we have the following:

(1) Root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(P_i - Q_i)^2}, \tag{7}$$

(2) Coefficient of determination (R^2):

$$R^2 = \left(\frac{1}{M}\frac{\sum_{i=1}^{M}(P_i - \bar{P})(O_i - \bar{O})}{\sigma_P \sigma_Q}\right)^2, \tag{8}$$

where $\bar{P} = 1/M\sum_{i=1}^{M}P_i$ and $\bar{O} = 1/M\sum_{i=1}^{M}O_i$ are the average of imputed and observed data, respectively, σ_P and σ_Q are the standard deviation of imputed and observed data, respectively.

Index (1) measures the discrepancy between the imputed and observed values; thus, a small RMSE is preferable when comparing different imputation methods. In contrast, index (2) characterizes the correlation between the imputed and

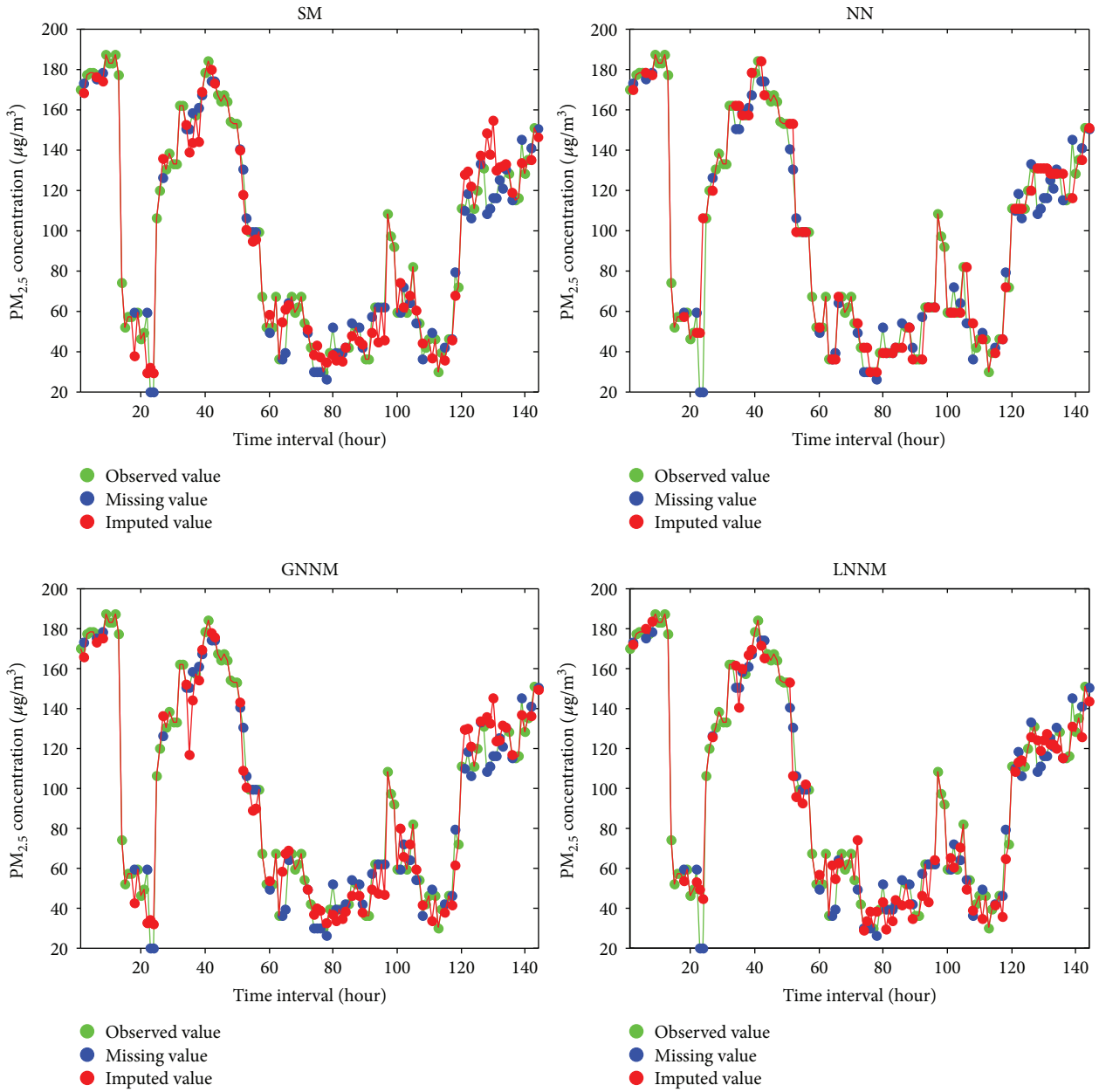

Figure 4: Imputed $PM_{2.5}$ concentration when the percentage of missing values equals 50%.

observed values; thus, a large R^2 is preferable for imputation methods.

The descriptive statistics of the data under different percentages of missing values are shown in Table 2. We can see that these statistics vary very little with respect to the percentage of missing data.

3.2. Imputation Error Comparison. The imputation performance of different methods when PMV equals 10% is shown in Table 3 where we report the mean value (mean) and the corresponding standard deviation (STD) of RMSE and R^2 across 10 tests. As we can see, GNNM achieves competitive performance in comparison with SM and NN. Note that GNNM is based on correlation between different stations and time periods, thus implying *nuclear norm* provides an interesting and effective global measure for spatial and temporal correlation of $PM_{2.5}$ pollution concentrations. More importantly, the proposed LNNM, which elaborately integrates both global and local correlation in a unified framework, consistently outperforms all other competing methods. In addition, we perform paired t-test on the 10 imputation results to show if there is statistically significant differences between LNNM and other competitive methods. The p values of t-test at 5% significance level are also reported in Table 3 where a greater difference between LNNM and the other method exists for a p value less than 0.05. As we can see, the imputation performance difference between LNNM and all other methods is significant since the p values are all very

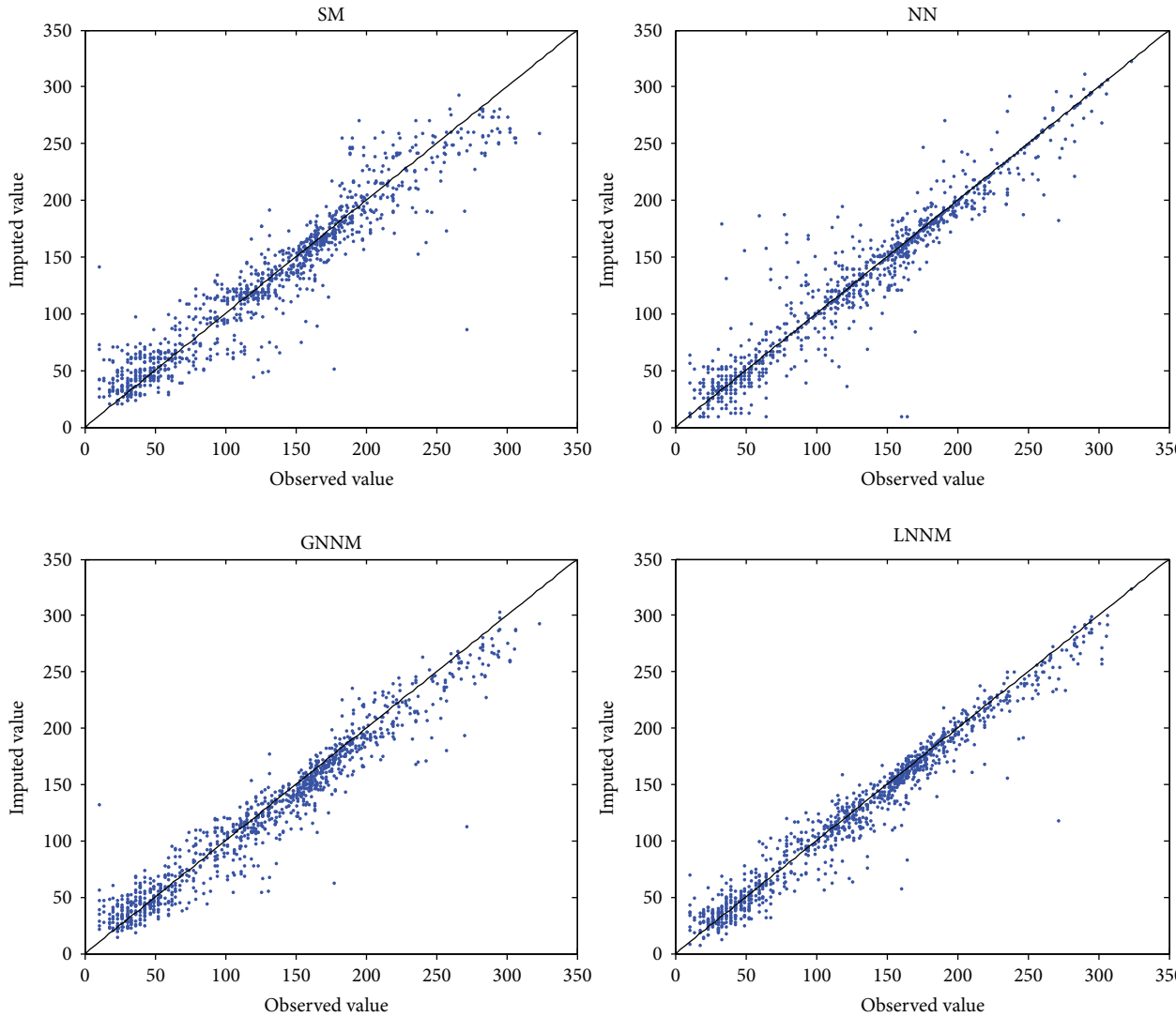

Figure 5: Imputation error scatter plots when the percentage of missing values equals 10%.

close to zero. These results show that performing nuclear norm minimization on highly correlated data across different stations and time periods can sufficiently make use of such high correlation, thus being able to impute missing values with better accuracy.

The experimental results when PMV equals 20%, 30%, 40%, and 50% are shown in Tables 4–7, respectively. We can see from the results that, as the percentage of missing values continues to increase, the estimation errors increase at the same time. This can be explained because the available information for imputation reduces in the case of high percentage of missing data. This variation of imputation error against the percentage of missing values is shown in Figure 1 for better comparison. We can observe that SM is more insensitive to the percentage of missing data, although its estimation error is usually larger than the other methods. NN is superior to SM when the number of missing data is not very large. However, when that percentage exceeds 40%, NN becomes the worst one. This may due to the fact that NN is purely based on local temporal correlation which is unreliable when large amount of data is missing. The proposed methods, especially LNNM, consistently achieve comparable or smaller estimation errors even in the case of large quantity of missing data. It confirms the effectiveness of the proposed methods.

3.3. Illustration of Imputation Results. In this section, we illustrate some imputation examples of SM, NN, GNNM, and LNNM in case that PMV equals 10%, 30%, and 50% in Figures 2, 3, and 4, respectively. As we can see, due to the inherent principle of a specific method, it may produce better estimation on some missing data, that is, the difference between the imputed value and observed value is small, but worse estimation on other missing data. In other words, no method can consistently outperform the other methods on all missing values. Nevertheless, statistically speaking, the proposed GNNM and LNNM can give better estimation of missing values in most cases.

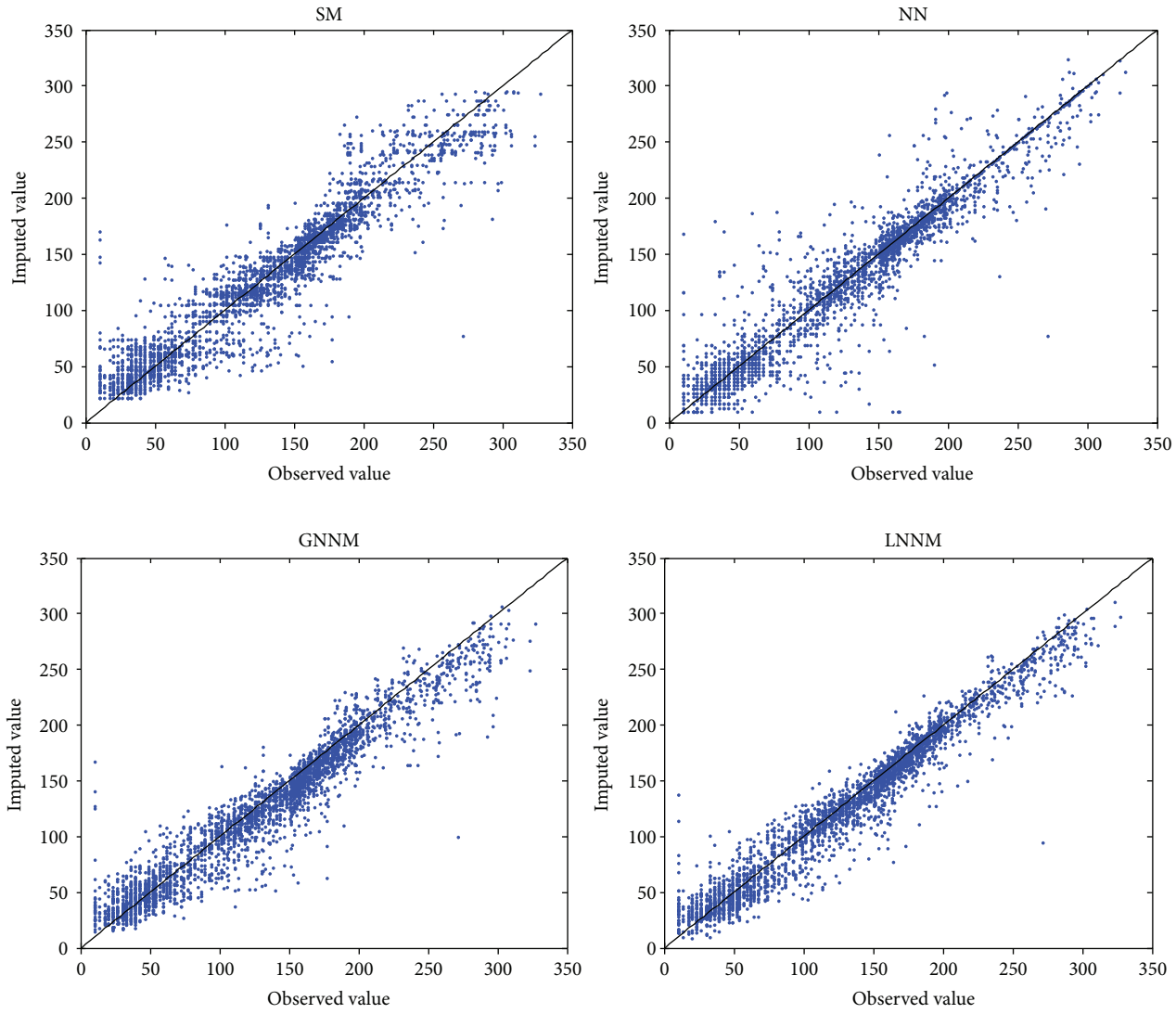

Figure 6: Imputation error scatter plots when the percentage of missing values equals 30%.

Figures 5, 6, and 7, respectively, show the scatter plots of the observed and imputed $PM_{2.5}$ pollutant concentration when PMV equals 10%, 30%, and 50%. We can see that the scatter of imputation results of our proposed methods, especially LNNM, is smaller than the other methods. It means the errors between observed and imputed values are generally smaller.

4. Conclusions

Imputing missing values is an important preprocessing task for air quality data analysis. How to make use of the inherent structure underlying data is closely related to the imputation performance. In this paper, we propose two new methods for air quality data imputation. The motivation is the row-wise or column-wise correlation that provides a prior structure of air quality data matrix. As a result, we can naturally use the nuclear norm to characterize such correlation and implement data imputation. In the first method (GNNM), we directly minimize the nuclear norm on the whole samples to maximize the global correlation. In the second method (LNNM), we tend to perform nuclear norm minimization on those highly correlated samples for improving imputation performance. The experiments on real-world $PM_{2.5}$ concentration data set verify the effectiveness of our proposed method.

It should be emphasized that despite better recovery performance, the proposed LNNM will suffer from higher computational burden since nuclear norm minimization has to be conducted on each sample and its nearest neighbors. Nevertheless, there are some possible ways to deal with this problem. For example, it is clear that the imputation on each sample and its nearest neighbors can be well parallelized so as to reduce the computational time. The proposed algorithm can be implemented on specific hardware, such as graphic processing unit (GPU) consisting of thousands of cores, to further improve the computational efficiency. The comparison of our method with other air quality data

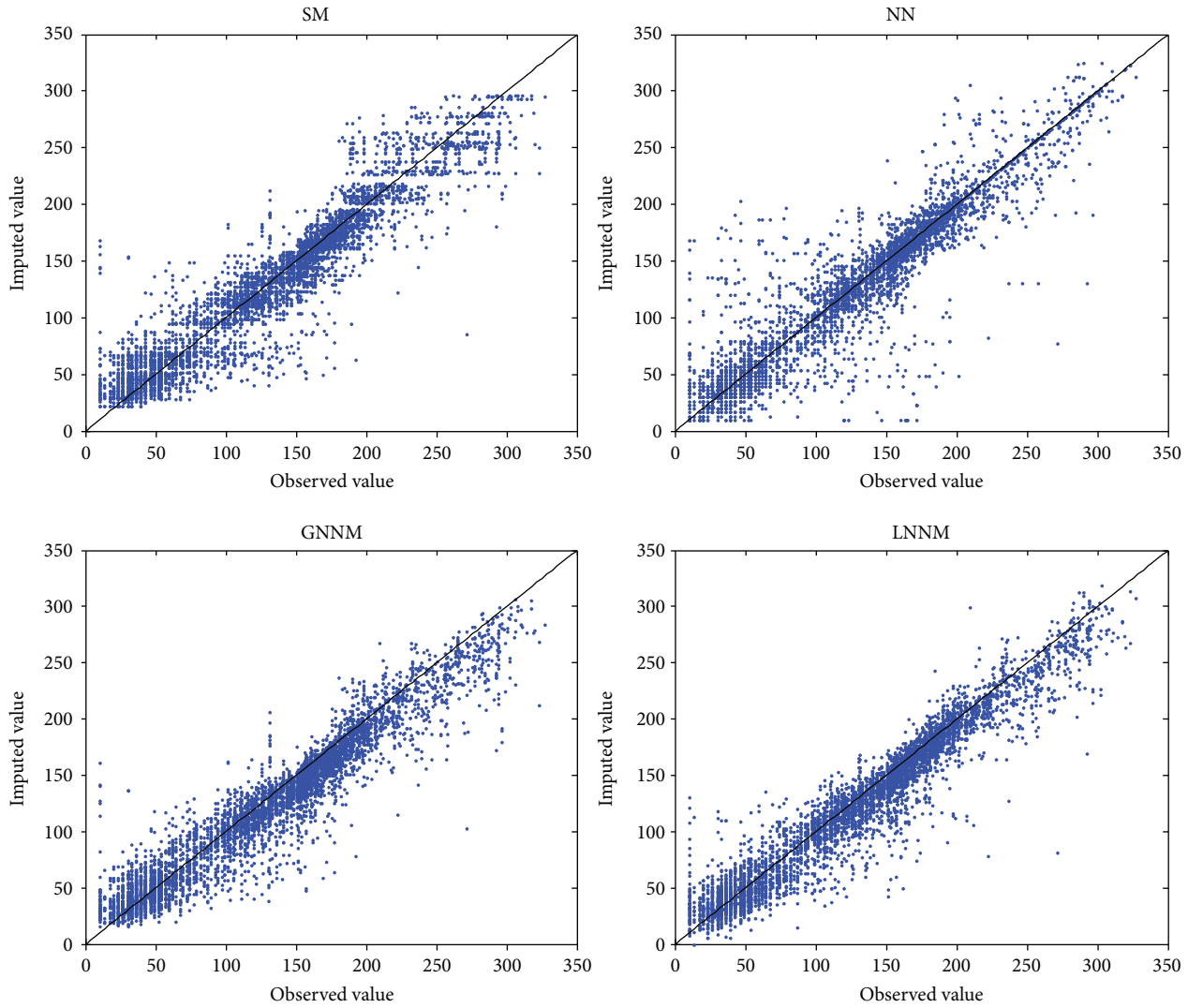

Figure 7: Imputation error scatter plots when the percentage of missing values equals 50%.

imputation methods [32–34] is also an interesting research topic in the future.

Conflicts of Interest

The authors declare that there is no conflict of interests.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant no. 61773184), Six Talent Peaks Project of Jiangsu Province (Grant no. 2017-JXQC-007), and the Talent Foundation of Jiangsu University, China (no. 14JDG066).

References

[1] P. S. Pooler, "Handling missing data: applications to environmental analysis," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 400-401, 2006.

[2] T. Schneider, "Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[3] X. Sun, L. Chen, Z. Yang, and H. Zhu, "Speed-sensorless vector control of a bearingless induction motor with artificial neural network inverse speed observer," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 4, pp. 1357–1366, 2013.

[4] X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, "Artificial neural networks forecasting of PM 2.5 pollution using air mass trajectory based geographic model and wavelet transformation," *Atmospheric Environment*, vol. 107, pp. 118–128, 2015.

[5] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, NY, USA, 1998.

[6] X. Chen, J. Yang, Q. Ye, and J. Liang, "Recursive projection twin support vector machine via within-class variance minimization," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2643–2655, 2011.

[7] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[8] J. L. Schafer, *Analysis of Incomplete Multivariate Data*, CRC press, New York, DC, USA, 1997.

[9] A. Pollice and G. J. Lasinio, "Two approaches to imputation and adjustment of air quality data from a composite monitoring network," *Journal of Data Science*, vol. 7, no. 1, pp. 43–59, 2009.

[10] A. Plaia and A. L. Bondì, "Single imputation method of missing values in environmental pollution data sets," *Atmospheric Environment*, vol. 40, no. 38, pp. 7316–7330, 2006.

[11] C. Real, J. Á. Fernández, J. R. Aboal, and A. Carballeira, "Substituting missing data in compositional analysis," *Environmental Pollution*, vol. 159, no. 10, pp. 2797–2800, 2011.

[12] M. P. Gómez-Carracedo, J. M. Andrade, P. López-Mahía, S. Muniategui, and D. Prada, "A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets," *Chemometrics and Intelligent Laboratory Systems*, vol. 134, pp. 23–33, 2014.

[13] J. L. Schafer and M. K. Olsen, "Multiple imputation for multivariate missing-data problems: a data analyst's perspective," *Multivariate Behavioral Research*, vol. 33, no. 4, pp. 545–571, 1998.

[14] P. R. C. Nelson, P. A. Taylor, and J. F. MacGregor, "Missing data methods in PCA and PLS: score calculations with incomplete observations," *Chemometrics and Intelligent Laboratory Systems*, vol. 35, no. 1, pp. 45–65, 1996.

[15] D. T. Andrews and P. D. Wentzell, "Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer," *Analytica Chimica Acta*, vol. 350, no. 3, pp. 341–352, 1997.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[17] E.-L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, and M.-D. Cubiles-de-la-Vega, "Missing value imputation on missing completely at random data using multilayer perceptrons," *Neural Networks*, vol. 24, no. 1, pp. 121–129, 2011.

[18] L. Folguera, J. Zupan, D. Cicerone, and J. F. Magallanes, "Self-organizing maps for imputation of missing data in incomplete data matrices," *Chemometrics and Intelligent Laboratory Systems*, vol. 143, pp. 146–151, 2015.

[19] J. M. Engels and P. Diehr, "Imputation of missing longitudinal data: a comparison of methods," *Journal of Clinical Epidemiology*, vol. 56, no. 10, pp. 968–976, 2003.

[20] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 2012.

[21] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[22] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[23] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Los Angeles, CA, USA, 2004.

[25] F. Shi, J. Cheng, L. Wang, P.-T. Yap, and D. Shen, "Low-rank total variation for image super-resolution," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. MICCAI 2013. Lecture Notes in Computer Science, vol 8149*, pp. 155–162, Springer, Berlin, Heidelberg, 2013.

[26] S. Wang and Z. Zhang, "Colorization by matrix completion," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, Toronto, Ontario, Canada, 2012.

[27] X.-Y. Pan, Y. Tian, Y. Huang, and H.-B. Shen, "Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach," *Genomics*, vol. 97, no. 5, pp. 257–264, 2011.

[28] X. Chen, Z. Wei, Z. Li, J. Liang, Y. Cai, and B. Zhang, "Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation," *Knowledge-Based Systems*, vol. 132, pp. 249–262, 2017.

[29] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: when urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1436–1444, Chicago, IL, USA, August 2013.

[30] Y. Zheng, X. Chen, Q. Jin et al., "A cloud-based knowledge discovery system for monitoring fine-grained air quality," preparation, Microsoft Tech Report, 2014, http://research.microsoft.com/apps/pubs/default.aspx.

[31] X. Chen, J. Yang, J. Liang, and Q. Ye, "Recursive robust least squares support vector regression based on maximum correntropy criterion," *Neurocomputing*, vol. 97, pp. 63–73, 2012.

[32] Á. Arroyo, Á. Herrero, V. Tricio, E. Corchado, and M. Woźniak, "Neural models for imputation of missing ozone data in air-quality datasets," *Complexity*, vol. 2018, Article ID 7238015, 14 pages, 2018.

[33] L. C. Larsen and M. Shah, "A context-intensive approach to imputation of missing values in data sets from networks of environmental monitors," *Journal of the Air & Waste Management Association*, vol. 66, no. 1, pp. 38–52, 2016.

[34] N. A. Zainuri, A. A. Jemain, and N. Muda, "A comparison of various imputation methods for missing values in air quality data," *Sains Malaysiana*, vol. 44, no. 3, pp. 449–456, 2015.