대한지리학회지 제53권 제3호 2018(427~444)

An Ensemble Machine Learning from Spatio-temporal Kriging for Imputation of PM_{10} in Seoul, Korea

Insang Song* · Changro Lee** · Key-Ho Park***

서울 미세먼지 데이터 결측대치를 위한 시공간 크리깅의 앙상블 머신러닝

송인상* · 이창로** · 박기호***

Abstract : Missing values in spatio-temporal data presumably cause defects, such that contaminate the results of spatio-temporal analyses. However, imputation methods for spatio-temporal data considering the inherent nature of spatio-temporal dependence have been neglected. We suggest an imputation algorithm based on ensemble spatio-temporal kriging for particulate matter measurement data for the period 2010-2014 at 54 monitoring stations near the metropolitan city of Seoul, Korea. We review previous studies on imputation methods for spatio-temporal data, then shed light on the necessity of our approach. Our approach implements resampling techniques on limited spatio-temporal data for a short-term period, then aims to enhance the imputation accuracy by taking the ensemble of the imputation results of resampled sub datasets. To examine such enhancement, we apply different conditions in experiments, including the number of resampling, neighborhood ratios, and ratios of artificially generated missing values. Results show that our approach outperforms both spatio-temporal kriging with the whole dataset (1.32~11.36%) and the linear regression-based imputation algorithm (52% in average). Our results show that the learning approach by resampling is still effective in spatio-temporal kriging in a limited environment as well as the spatio-temporal algorithm considering the inherent dependence among the data. But the considerable underperformance compared to the accuracy of the machine learning-based algorithm indicates the necessity of further examination of the effect of spatio-temporal dependence in such an algorithm.

Key Words : Imputation, Spatio-temporal kriging, Resampling, Particulate matter

요약 : 시공간 데이터의 결측치는 그 자체로 데이터의 결함으로서 시공간 분석 결과를 왜곡시킬 수 있다. 그러나 시공간 데이터에 내재된 시공간 의존성을 이용한 결측대치 방법은 덜 주목받아 왔다. 이에 본 연구에서는 서울특별시 및 근방의 54개 측정소로부터 2010년부터 2014년까지 5년간 측정된 시간별 미세먼지(PM_{10}) 데이터의 결측치를 대치하기 위하여 앙상블 시공간 크리깅 모형에 기초한 결측대치 모형을 제안하였다. 기존 연구들을 검토한 결과, 본 연구에서 이용된 접근법의 필요성이 발견되었다. 본 연구가 제안하는 앙상블 결측대치 모형은 단기간의 시공간 데이터에서 재표집(resampling)된 하위 데이터셋으로 복수의 시공간 크리깅 모형들을 적합하고, 이들을 앙상블하여 결측대치 정확도를 높이고자 한다. 향상 여부를 실증하기 위하여 측정 데이터에 대해 결측대치 실험을 실시하였다. 실험에서는 재표집 횟수, 시공간 크리깅 적합 시 이웃 비율, 결측 생성 비율 등 3요소에 대해 서로

Parts of this study are from the Master's thesis of the first author.

* Researcher, The Institute for Korean Regional Studies, Seoul National University, South Korea (e-mail: henry385@snu.ac.kr)

** Researcher, The Institute for Korean Regional Studies, Seoul National University, South Korea (e-mail: spatialstat@naver.com)

*** Professor and Adjunct researcher, Department of Geography and the Institute for Korean Regional Studies, Seoul National University, South Korea (e-mail: khp@snu.ac.kr) (Corresponding Author)

다른 조건들을 적용하였다. 실험 결과, 제안된 앙상블 모형은 단일 시행 시공간 크리깅 모형(1.32~11.36%)과, 선형 앙상블 모형(평균 52%)보다 높은 정확도로 결측치를 대치하였다. 본 결과는 제한된 환경에서 시공간 크리깅 모형 앙상블이 결측 대치 정확도를 높이는 데 효과가 있음을 입증한다. 다만 제안된 알고리즘의 정확성은 머신러닝 기반의 결측대치 알고리즘에 비해서 덜 우수했는데, 이 결과는 머신러닝 알고리즘에서 시공간 의존성 효과가 어떻게 나타나는지에 대한 추가 연구 필요성을 제기한다.

주요어 : 결측대치, 시공간 크리깅, 재표집, 미세먼지

1. Introduction

Missing values in spatio-temporal data are an obstacle to space-time analysis, as breaking the data makes it difficult to examine the change of spatial phenomenon[1]. This makes statistical reasoning for the whole study area more likely to be inaccurate than when using complete data. Also, the spatial analysis results may be biased when the spatial sample at a period with incomplete data contains properties that are different from those obtained from spatial samples. Most importantly, all observed values of spatio-temporal data are unique to a specific place and time, so even if the ratio of the missing values is small, the effect on the structure of the entire data could be significant (Kondrashov and Ghil, 2006).

In this study, we applied a spatio-temporal kriging model for imputation to fill in the missing values in the particulate matter with aerodynamic diameters less than or equals to 10 μm (PM_{10}). Missing measurements in air pollution data, possibly due to arbitrary deletion or functional failure of the measuring instrument, are limiting factors in assessing the effects of short-term exposure to air pollution in health geography (Karahalios *et al*., 2013). Spatio-temporal kriging is a method of modeling the temporally-augmented kriging that implements the space-time covariance function, which incorporates the temporal variability of variables. It has been widely applied since the 2000s (박노욱, 2011). The detailed methods will be described further in later sections of this paper.

The purpose of the imputation is to restore the original data as accurately as possible. Therefore, by adopting a strategy of model reinforcement, it is possible to improve the imputation accuracy. One of the methods that can be applied is the ensemble of models by resampling. The resampling method divides or randomly extracts the sampled data currently used, which is usually used for cross-validation or bootstrap aggregation (Breiman, 1996). Focusing on the effect of an ensemble, we hypothesized that the ensembled spatio-temporal kriging model would show better accuracy than both the kriging model without ensemble and other imputation techniques. Therefore, the main purpose of this study is to identify two sub-hypotheses by using a resampling-based learning and a spatio-temporal kriging method to implement it into the imputation. First, we examine whether the spatio-temporal kriging model could be reinforced through the plug-in of the resampling method to spatio-temporal kriging. Second, we compare the effects of spatio-temporal kriging of resampled data in comparison to existing methods.

2. Literature review

Bennett *et al*.(1984) comprehensively reviewed the missing value problems in spatial data. The authors classified the causes of missing values in spatial data into three categories: spatial, temporal, and deletion

processes. The problems of missing data in spatial data were divided into spatial and statistical aspects. Also, they proposed interpolation methods and spatial filtering techniques to fill the missing values in the spatial data. However, in their study, interpolation and imputation were used without any clear distinction, so the imputation methods were treated as a type of interpolation method.

Haining *et al*.(1989) suggested the processing of missing values using spatial models. They applied the first-order conditional autoregressive model to the lattice data to perform the imputation. The main results showed that more effective dislocations are possible when reflecting the autocorrelation inherent in spatial data. This paper describes the necessity and validity of the method to impute spatial data. Until the 1990s, missing data studies of spatial data were often done on raster data (Wilson *et al*., 2012).

Several studies have used an array of spatial statistical techniques to impute missing values in data measured continuously in certain locations, such as climate or air pollution. Although the data used in the studies are spatio-temporal, studies have adopted the regression models or simple averaging (Smith *et al*., 2003; Xia *et al*., 1999). A study applied multi-layer perceptron (MLP) and artificial neural network to combine the methodology of artificial intelligence into the context of imputation (Junninen *et al*., 2004).

Schneider(2001) focused on the spatio-temporal dimensions of missing values, then suggested the EM (Expectation-Maximization) algorithm which outperforms the existing imputation methods. 김병식 등 (2011) reported that the inverse distance weighting and the correlation weighting method yielded reliable results compared to other imputation methods in the rainfall data. Feng *et al*.(2014) developed a CUTOFF algorithm, which imputes the missing values in hydrologic measurement data by filtering the measurement values according to their correlation. Li and Parker(2014) applied the k-nearest neighbor algorithm as a method to impute the measurement data from sensor networks installed in various locations. They reported that the imputation errors were lower than those of the EM algorithm. Deng *et al*.(2016) imputed the missing values in the temperature and precipitation data with a method combining the cluster analysis and the space-time weighting. They considered spatio-temporal inverse distance weighting, spatio-temporal kriging, and spatio-temporal hybrid covariance model.

In previous studies, models reflecting the inherent dependence in spatio-temporal data tended to show better imputation performances than existing imputation models which do not consider spatial and/or temporal dimensions of data. However, as mentioned above, previous studies focused on comparing the performances of various algorithms. We aim to examine the effectiveness of training the spatio-temporal kriging model as a base learner in real-world data.

3. Methods

1) Spatio-temporal kriging

Spatio-temporal kriging is a model which incorporates components of purely temporal and interaction between spatial and temporal variations in kriging. When the variables show similar values according to spatial and temporal adjacency, the structure can be predicted and explained by the space-time covariance function. By using spatio-temporal kriging, we can estimate the possible values and their uncertainties at the unknown locations and time points based on the space-time structure of the sample (Cressie and Wikle, 2011). The spatio-temporal kriging is decomposed into five elements, which are expressed as follows (Wackernagel, 2003; Heuvelink and Griffith, 2010).

$$Y(s, t)=\mu(s, t)+\beta(s)+\gamma(t)+\kappa(s, t)+\delta(s, t) \quad (1)$$

$Y(s, t)$: value at spatial position s and time point t

$\mu(s, t)$: mean estimate at spatial position s and time point t

$\beta(s)$: random variation that depends on the site

$\gamma(t)$: random variation that depends on the time point

$\kappa(s, t)$: spatio-temporal interaction term

$\delta(s, t)$: spatio-temporal variability in a fine scale

$s \in D_s$, $t \in D_t$ where D_s and D_t are spatial and temporal data set, respectively.

The formula (1) represents the decomposability of the value at time point t and spatial entity s into the sum of spatial, temporal, and residual processes. Assuming $\mu(s, t)$ is variable, the formula below corresponds to the formulation of ordinary kriging as follows (Cressie and Wikle, 2011).

$$\hat{Y}(\boldsymbol{s}_0, t_0)=\hat{\mu}_{gls}+\mathbf{c}_0'\mathbf{C}_\mathbf{z}^{-1}(\mathbf{Z}-\hat{\mu}_{gls}\mathbf{1})=\boldsymbol{\lambda}'\mathbf{Z} \quad (2)$$

$\hat{Y}(\boldsymbol{s}_0, t_0)$: the estimated value(s) at spatial coordinate vector $\boldsymbol{s}_0$ and time point t_0

$\hat{\mu}_{gls}$: the mean estimate by generalized least squares

$\mathbf{Z}$: a matrix with attribute values, coordinates, and time points

$\mathbf{C}_\mathbf{z}=\text{var}(\mathbf{Z})$

$\mathbf{c}_0'=\text{cov}(Y(\boldsymbol{s}_0; t_0), \mathbf{Z})$

$\boldsymbol{\lambda}$: column vector with kriging weights

As the best fit procedure for kriging, spatio-temporal kriging requires theoretical spatio-temporal variograms, which are derived from the empirical spatio-temporal variograms. Theoretical spatio-temporal variograms are constructed by accounting for the interaction structure of spatial and temporal variability after fitting marginal variograms. Marginal variograms stand for variograms of each separate dimension when time lag and distance are respectively zero.

Types of spatio-temporal semivariogram functions are listed in Table 1. The separable covariance function represents the spatio-temporal function as the product of spatial and temporal covariance functions. Thus, the spatial and temporal dimensions are fully independent. Other types of spatio-temporal covariance functions are suggested to substantiate the cases without fulfilling the independence across two dimensions. The product-sum covariance function adds the covariance functions of each dimension to the product with scale parameter k. The metric covariance function requires the common

Table 1. Types of spatio-temporal covariance and semivariogram functions and their theoretical formulae

Form	Covariance and semivariogram formula
Separable	$C_{separable}(\mathbf{h}, u)=C_s(\mathbf{h})C_t(u)$ $\gamma_{separable}(\mathbf{h}, u)=\text{sill}\cdot(\bar{\gamma}_s(\mathbf{h})+\bar{\gamma}_t(u)-\bar{\gamma}_s(\mathbf{h})\bar{\gamma}_t(u))$
Product-sum	$C_{product\text{-}sum}(\mathbf{h}, u)=kC_s(\mathbf{h})C_t(u)+C_s(\mathbf{h})+C_t(u)$ $\gamma_{product\text{-}sum}(\mathbf{h}, u)=(k\cdot\text{sill}_t+1)\gamma_s(\mathbf{h})+(k\cdot\text{sill}_s+1)\gamma_t(u)-k\gamma_s(\mathbf{h})\gamma_t(u)$
Metric	$C_{metric}(\mathbf{h}, u)=C_{joint}\left(\sqrt{\mathbf{h}^2+(\kappa\cdot u)^2}\right)$ $\gamma_{metric}(\mathbf{h}, u)=\gamma_{joint}\left(\sqrt{\mathbf{h}^2+(\kappa\cdot u)^2}\right)$
Sum-metric	$C_{sum\text{-}metric}(\mathbf{h}, u)=Cs(\mathbf{h})+C_t(u)+C_{joint}\left(\sqrt{\mathbf{h}^2+(\kappa\cdot u)^2}\right)$ $\gamma_{sum\text{-}metric}(\mathbf{h}, u)=\gamma_s(\mathbf{h})+\gamma_t(u)+\gamma_{joint}\left(\sqrt{\mathbf{h}^2+(\kappa\cdot u)^2}\right)$

$C(.,.)$, $C_s(.,.)$, $C_t(.,.)$: spatio-temporal, spatial, and temporal covariance functions; $\gamma(.\ ,\ .)$: spatio-temporal (semi)variogram function; $\mathbf{h}$: spatial distance or lag; u: time lag; sill: spatio-temporal sill; sill_s, sill_t: sills for spatial and temporal dimensions; $\bar{\gamma}_s$, $\bar{\gamma}_t$: spatial and temporal (semi)variogram functions which are scaled to have the bound [0,1].

metric which is interpretable for the integrated covariance function, and thus it would model the spatio-temporal covariance in a function rather than the three listed covariance models. The sum-metric covariance function incorporates the sums of spatial and temporal covariance functions and the metric covariance function. It should be noted that metric covariance functions could present by defining the combined spatio-temporal metric κ(kappa). The detailed process in calculating the metric is diversified by selecting a specific type of function translating the temporal lags into spatial distances, i.e., variogram matching or calculating ratios of variogram values.

We defined all spatio-temporal variograms used to fit spatio-temporal kriging model as sum-metric. This is because sum-metric variograms are available to model

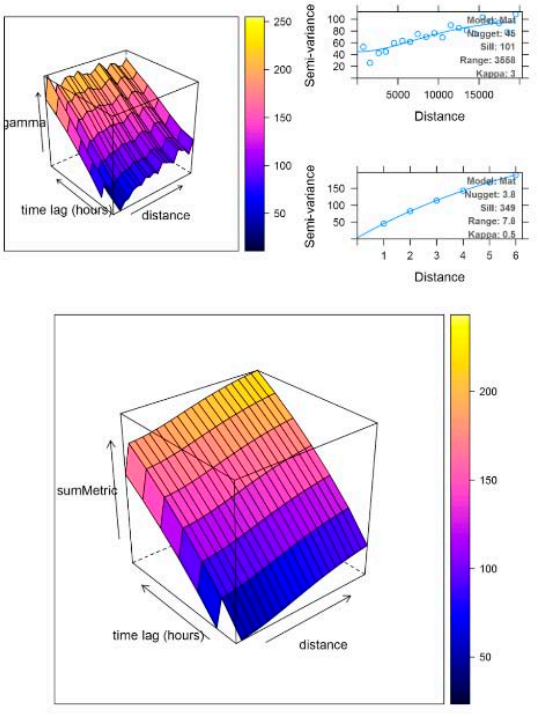

Figure 1. Graphical example of fitting a spatio-temporal variogram model

Upper left: empirical spatio-temporal variogram
Upper right: spatial and temporal marginal variograms (as ordered)
Below: fitted sum-metric spatio-temporal variogram

empirical spatio-temporal variograms in a more flexible manner. Gräler *et al.*(2016) observed that sum-metric variograms show lower errors than variograms following other forms, especially in air quality monitoring data. For time and space marginal variograms of test sets, we chose theoretical variogram formulations that belong to Matérn or exponential families and showed the lowest sum of squared errors (for example, Figure 1). Also, we set the form of the composite spatio-temporal variogram as Gaussian for ease of computation. Finally, we placed regular spatial lag and its maximum as 1,000 meters and 20,000 meters. To compare the imputation performance, we chose multiple imputation by chained equations (MICE) and missForest, which employs the ensemble machine learning algorithm random forest to examine the imputation accuracy. The imputation accuracy was evaluated with root mean squared error (RMSE) as shown in the formula below.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (3)$$

where $\hat{y}_i$ is i^{th} estimate and y_i is the i^{th} actual value in the dataset.

To note, the spatio-temporal anisotropy ratio (notated as κ; following Gräler *et al.*, 2016) is determined with marginal variograms[2]. As we calculated the anisotropy ratio with spatial and temporal variograms, we gained computation efficiency by limiting ranges of candidate solutions with the minimum and maximum using empirical variograms.

The proposed spatio-temporal kriging-based imputation approach follows as below:

a. Take the whole or partial spatio-temporal dataset with missing values;

b. Determine the theoretical spatio-temporal variogram by the empirical spatio-temporal variogram of the dataset taken at the step a;

c. Split the dataset into N subdatasets after row-randomization;

d. Initialize i as 1;

e. For $i \leq N$, iterate:

f. Calculate the theoretical spatio-temporal variogram from N-1 subdatasets except for i^{th} subdataset;

g. With the theoretical variogram of the step f, save the kriged estimates at places of the missing values;

h. Save the average of N kriged values as the imputed values.

The third step should be noted. The data becomes sequentially randomized, as this enables us to exclude the influence of temporal sequence embedded in the spatio-temporal dataset. To explain, we could avoid the possibility of including or excluding data points at specific spatial or temporal extents. In step g, we calculate the average of imputed values from the resampled datasets, which is the ensemble approach for model estimates based on continuous variables (Opitz and Maclin, 1999; Kuhn and Johnson, 2013)[3]. Also, to emulate the actual perfor-

Table 2. Conditions of experiments

Conditions	Values
Ratios of random missing generation	As data, 5%, 10%, 20%
Maximum time lag	4-12 hours, 1 hour
Number of resampling	3, 5, 10 times
Neighborhood ratio	75%, 100%

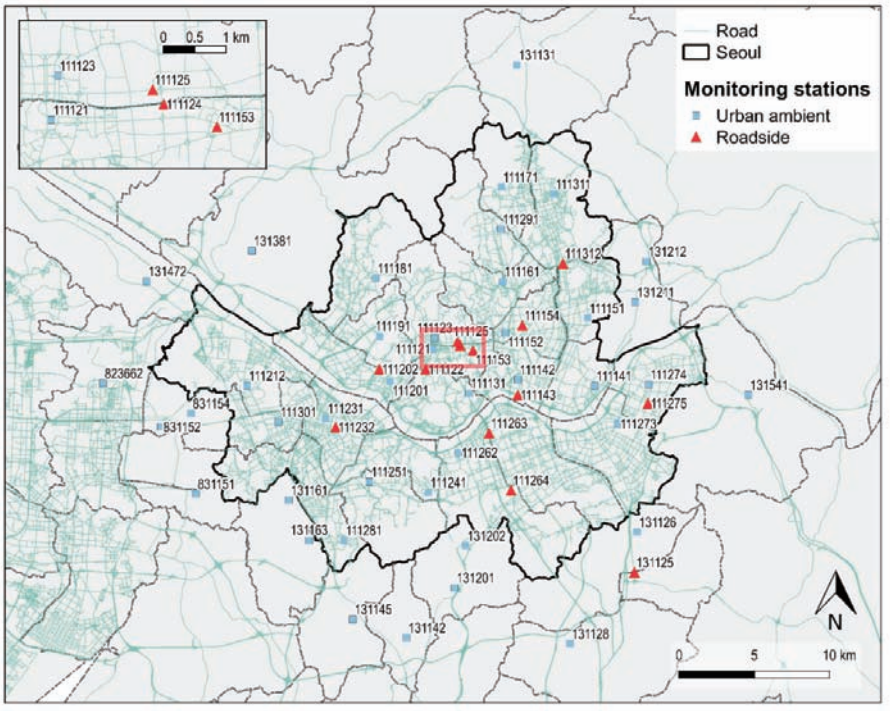

Figure 2. Locations of 54 monitoring stations near Seoul

Tabular source of locations from 국립환경과학원(2015); labels are station identifiers.

mance of our approach, we considered two conventional algorithms which represent either linear (simple, multiple imputation by chained equations)[4] and machine learning (complex, missForest)[5] approaches to compare the performances of each algorithm. Lastly, we tried applying several values for conditions of experiments to compare the efficacy of each conditional value. Conditional values are listed in Table 2.

2) Data

We used the PM_{10} measurement data from Air Korea administered by the National Institute of Environmental Research. The data was automatically measured at 54 monitoring stations around Seoul and was distributed after quality assessment (Figure 2, Korea Environment Corporation, 2018)[6]. The data includes monitoring stations adjacent to the administrative boundary of Seoul to avoid a possible edge effect. We checked the data consistency regarding the change in the location of monitoring stations to coherently analyze continuous spatio-temporal distribution.

To apply the proposed imputation model, we extracted subdatasets from the five-year measurement data by the following standards: (a) temporal length equal to or greater than 24 hours; (b) no missing values in any of the 54 monitoring stations. There were 17 sub datasets subject to such standards. Durations are between 24 to 60 hours.

There are 1260.3 missing hours on average among 43,824 hours in total for the study period, and their median was 1,191 hours per station. However, there is a significant decrement in the number of samples from all 54 stations due to the missing values contained in the air pollution data. Only 73,291 (3.09%) out of 2,366,496 hours are missing in the data, which is ignorable considering the studies on missing data (Harrell, 2015)[7]. When the spatial dimension of the dataset is considered, the ratio of missing values is significantly inflated such that 10,981 (25.06%) out of 43,824 hours include complete measurement data from 54 monitoring stations. Accordingly, it can be presumably expected that the results of spatial analysis such as interpolation using the incomplete data will be biased compared to those not using the incomplete data.

4. Results

1) Descriptive analysis

The missing values in the whole dataset show short and dispersed patterns, which comprise one or two hours except for some continuous missing values. The matrix plot illustrates the missing pattern visually by highlighting the missing elements by color (see Figure S1 in the supplementary material). The data includes a few continuous missing values in the time series, and missing values with lengths ranging from a few hours to several days are mainly seen in the period of device calibration done by the administrative authority. Test datasets show varied characteristics as observing their values (1-269), ranges (54-248), duration (24-60 hours), and standard deviations (6.34-33.80) (Table 3). To measure the proximity among multiple time series in each dataset, we used the temporal correlation coefficient suggested by Chouakaria and Nagabhushan(2007), using the formula (4). We calculated the sum of 1,431 (the number of 2-combinations from 54 time series, $\binom{54}{2}=\frac{54\times53}{2}$) pairwise temporal dissimilarity of each dataset and normalized it by dividing the range of values of each dataset to control the amount of values. We termed the final index mCORT (abbreviation of 'mean dissimilarity coefficient of time series'), which values show 0.28-2.21 (Table 3).

Table 3. List of test datasets and their summary statistics

Set	Begin	End	Duration	Minimum	Median	Mean	Maximum	Range	Stdev†	CV‡	mCORT¶
1	12/31/2010 15:00	01/01/2011 15:00	25	22	49	49.48	94	72	11.78	0.24	1.61
2	03/12/2011 15:00	03/13/2011 22:00	32	26	64	65.71	146	120	16.40	0.25	1.58
3	07/16/2011 02:00	07/17/2011 07:00	30	1	17	17.77	69	68	8.31	0.47	1.93
4	08/14/2011 14:00	08/15/2011 19:00	30	9	49	48.81	104	95	14.67	0.30	0.88
5	01/22/2012 09:00	01/23/2012 20:00	36	6	31	31.26	74	68	8.70	0.28	1.53
6	03/03/2012 20:00	03/05/2012 01:00	30	2	17	17.61	63	61	6.34	0.36	1.33
7	03/09/2012 20:00	03/11/2012 17:00	46	5	36	39.06	91	85	17.27	0.44	0.91
8	03/24/2012 21:00	03/25/2012 22:00	26	10	42	55.61	185	175	33.80	0.61	0.28
9	05/04/2012 23:00	05/07/2012 10:00	60	24	61	71.54	219	195	31.06	0.43	0.77
10	02/02/2013 16:00	02/03/2013 15:00	24	5	27	28.00	59	54	8.56	0.31	1.37
11	02/23/2013 16:00	02/25/2013 01:00	34	4	47	47.99	110	106	16.47	0.34	0.99
12	05/12/2013 09:00	05/13/2013 10:00	26	55	103	102.80	177	122	18.13	0.18	1.44
13	07/06/2013 09:00	07/07/2013 10:00	26	29	65	66.17	119	90	14.41	0.22	1.36
14	02/22/2014 16:00	02/24/2014 11:00	44	51	107	112.46	269	248	29.54	0.26	1.15
15	04/06/2014 10:00	04/07/2014 11:00	26	17	60	56.61	115	98	18.71	0.33	0.73
16	04/12/2014 23:00	04/14/2014 09:00	35	40	76	78.51	135	95	14.34	0.18	2.21
17	07/12/2014 13:00	07/14/2014 03:00	39	22	59	59.85	119	97	13.26	0.22	1.32

† standard deviation; ‡ coefficient of variation; ¶ average pairwise dissimilarity of 54 time series in each dataset (following Chouakria and Nagabhushan, 2007)

Note: units of minimum, median, mean, maximum, and standard deviation are micrograms per cubic meter (μg/m^3); unit of duration is hours.

$$D(S_1, S_2)=2/(1+\exp(2\cdot \mathrm{CORT}(S_1, S_2))\cdot \delta_{\mathrm{conv}}(S_1, S_2) \quad (4)$$

where $\mathrm{CORT}(S_1, S_2)=\sum_{i=1}^{p-1}(u_{i+1}-u_i)(v_{i+1}-v_i) \,/\, \sqrt{\sum_{i=1}^{p-1}(u_{i+1}-u_i)^2}\sqrt{\sum_{i=1}^{p-1}(v_{i+1}-v_i)^2}$

$\delta_{\mathrm{conv}}(S_1, S_2)=\sqrt{\sum_{i=1}^{p-1}(u_i-v_i)^2}$

u_i and v_i are the i^{th} element of time series S_1 and S_2, respectively.

2) Imputation results by missing ratios

To examine changes in model performances as proportions of missing values vary, we experimented with varied missing ratios as the main data observed in 2010-2014, 5, 10, 20, and 30 percent. We fixed the maximum time lag at six hours and the number of resampling at ten times in the experiment. Root mean squared errors (RMSE) of imputations tend to increase as the ratio becomes higher, and the increment of RMSE becomes smaller as the ratio exceeds 20 percent (Figure 3). Comparing the accuracy as the number of sampling increases, we observed that all results with resampling showed better accuracy than those of kriging without resampling. Also, three or five times of resampling was the most effectiveness for 16 out of 17 test datasets (Table 4).

The average RMSE was the minimum of the data with missing ratios observed in the entire dataset over a period of five years. This is because the average ratio of randomly generated missing values in all monitoring stations was about 3 percent, which is far lower than in other experimental conditions. However, because of the variation in empirical missing ratios across monitoring stations, the

Table 4. Performance improvement in the change in RMSE compared to the single imputation by spatio-temporal kriging

unit: percent

Set	Number of resampling		
	3	5	10
1	7.38	**7.54**	6.56
2	1.65	**2.40**	1.79
3	**2.73**	1.64	0.93
4	**5.00**	4.83	4.02
5	**11.36**	9.71	5.70
6	**4.68**	4.15	2.45
7	2.55	**3.13**	1.70
8	**3.99**	3.67	3.46
9	2.25	**2.55**	1.24
10	**7.08**	6.81	5.21
11	1.07	**1.66**	1.10
12	**1.32**	1.19	0.74
13	0.88	**1.37**	1.28
14	1.04	**1.85**	0.87
15	**3.98**	3.53	1.32
16	**1.54**	1.26	0.66
17	5.97	**6.43**	4.39

The maximum improvements by sets are highlighted; missing ratio is ten percent.

result shows variability in errors for each imputation. On the other hand, the range of RMSEs tends to decrease as the missing ratio increases except for in set 6. Along with observing the decrease in differences of RMSE, it can be suggested that the missing values in the spatio-temporal data change the spatio-temporal structure of the data and the transition of the structure to the new state as the error exceeds a certain level.

When the characteristics of each set shown in Table 2 are considered together, it is observed that the decrease in the predictive performance against the missing ratio tends to increase as the time series patterns of the monitoring stations are different. In the case of mCORTs such as sets 1, 2, 3, 5, 10, 12, and 16, the increase in the RMSE was large as the error rate increased from 5% to 20%. On the other hand, the correlations between measurement points such as sets 8, 9, 14, and 15 are high, so even if the mCORT is small, the increase of the RMSE appears to be insignificant.

3) Comparison to conventional imputation algorithms

The spatio-temporal kriging model with resampling is more accurate than the linear imputation model and has similar accuracy to the machine-learning algorithm (Figure 5). The variability of RMSE indicates the stability of the imputation algorithm. The variability of RMSEs for every imputation trial tended to show similar accuracy as that of the imputation, and the variation gradually increased in the order of missForest, spatio-temporal kriging (denoted as ST-kriging in Figure 5), and MICE. On the other hand, results of spatio-temporal kriging in the sets 1 and 5 showed high RMSEs in a few trials. This shows that the combined effect of the locations and time points of missing values affects the fitting of kriging models and their accuracy of the imputations (for detailed results by datasets, see Table S1 in the supplementary material).

5. Discussion

To sum up all imputation results, the resampling-based spatio-temporal kriging imputation approach showed applicable performance. This can be conceptually explained in two ways. First, the model takes into account the spatio-temporal dependency, which is implicit in the data. The spatio-temporal kriging-based imputation model performed predictions on a level comparable with the nonparametric machine learning-based algorithm in terms of accuracy. We could find that taking unique structure into consideration helps to improve spatio-temporal analyses and modeling proce-

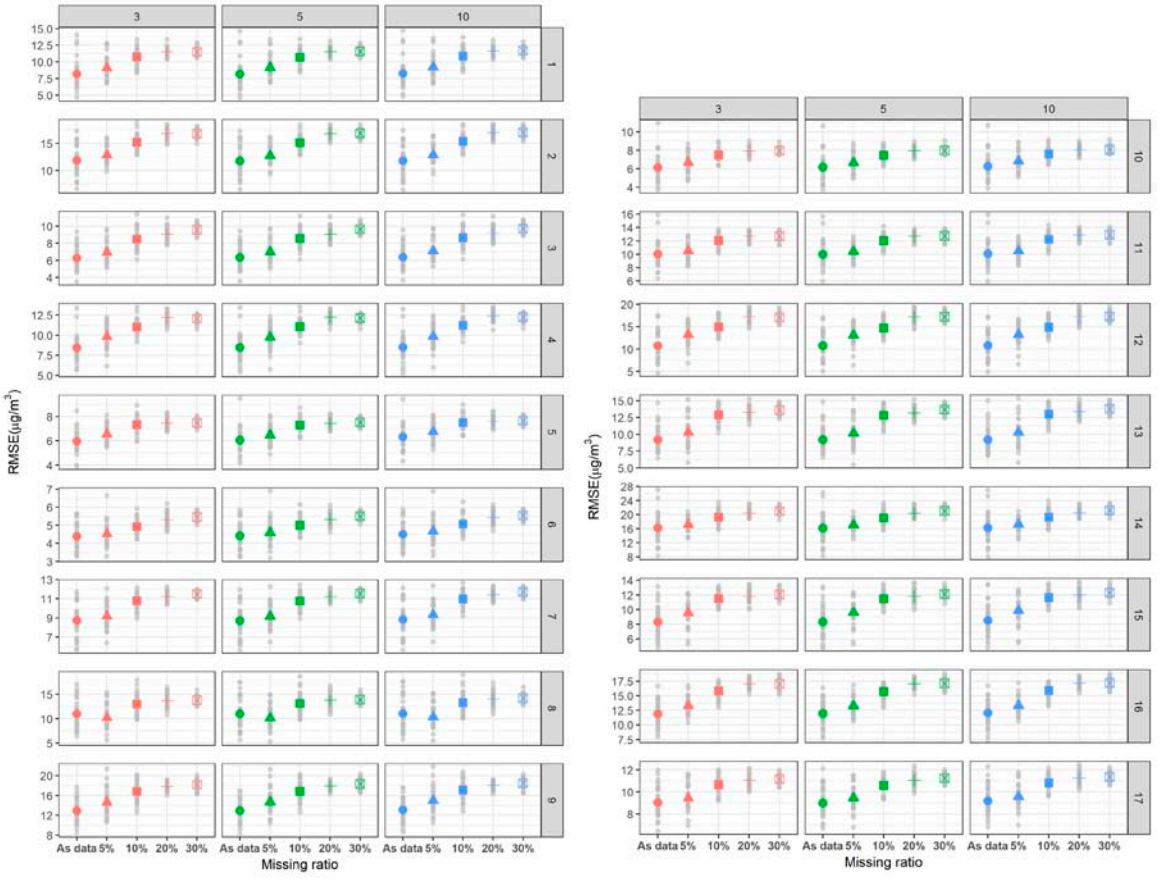

Figure 3. Ratio of missing data generation and RMSEs of imputation results

Numbers in the title of each vertical panel represent set numbers, whereas the title of horizontal panel shows the number of resampling; large symbols per missing ratio in each panel represents the average RMSE for 30 trials; small symbols are RMSEs of 30 trials.

dures. The second conceptual explanation is the role of spatio-temporal kriging as a tool for understanding the dependence in the spatio-temporal data. Spatio-temporal variogram, which is a base tool for conducting spatio-temporal kriging, can visualize the dependence in spatio-temporal data. Accordingly, the proposed model is useful in understanding the model compared to the machine learning counterpart, of which the mechanism difficult to comprehend.

However, the results suggest the need to clarify the ways in which the spatio-temporal aspects of the data affect the performance of machine learning algorithms. Since the results of missForest showed higher imputation accuracy and lower variation, it should be examined whether spatial and/or temporal dependence affects the model performance. This is because such examination could show the efficacy of consideration for implicit spatio-temporality in the data. When no spatial and/or temporal effects were observed in machine learning algorithms, it could be hard to stress that the spatio-temporal dependencies should be accounted for modeling with data of spatial and/or temporal dimensions. In summation, results of this study demand further explanation of possible contributions of spatio-temporality in machine

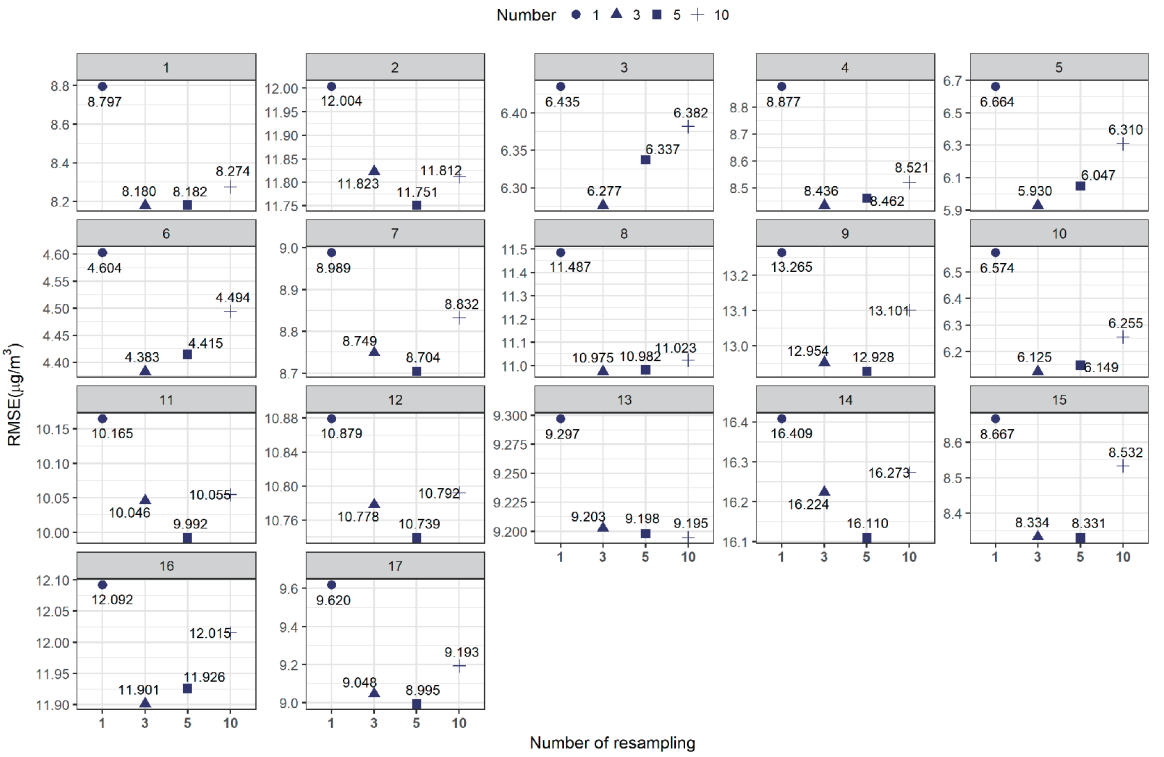

Figure 4. Average RMSE of 30 imputation results by the numbers of resampling

Missing ratio is ten percent.

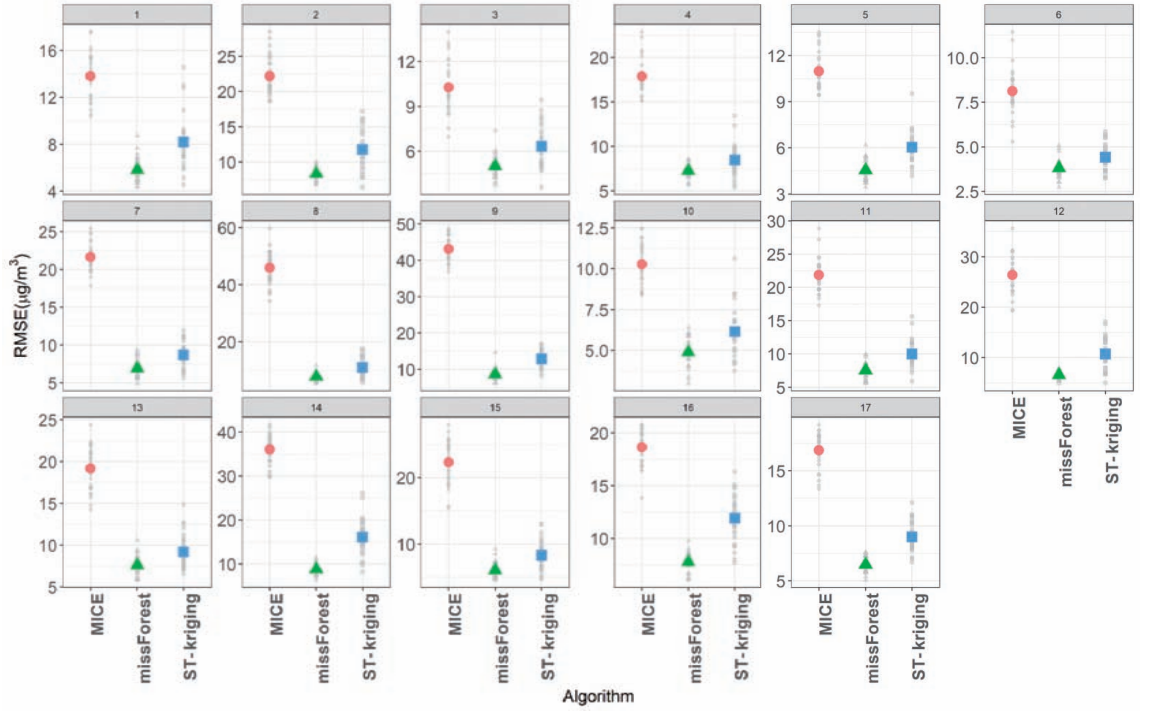

Figure 5. Comparison of RMSEs from the imputation results of three algorithms

For spatio-temporal kriging, the number of resampling for each dataset is five and the missing ratio is ten percent.

learning algorithms.

Lastly, we explored the residual space-time cube to check whether there is a spatio-temporal pattern in the imputation results. Figures 6 and 7 display the results of sets 8 and 16, which have the lowest and the highest temporal correlations measured by mCORT, respectively. The results show that errors increase in outer points of the space-time cubes, which present the spatio-temporal edge effect. It would be partially explained by the fact that the available information around proximal points becomes scarce at the external positions in the space-time cube.

The result of the imputation accuracy evaluation, based on the number of resampling and the ratio of missing values, has significant implications for both the analysis of the effect of missing values and the application of the imputation model for spatio-temporal data. First, the variance of the imputation accuracy gradually decreased as the number of resampling increased (Figure 3). This shows that the shape of spatio-temporal variogram of the data is gradually transformed as the missing rate increases by adding higher variability on spatio-temporal variogram. Second, scale dependency should be accounted for. It can be inferred from the applied conditions when we

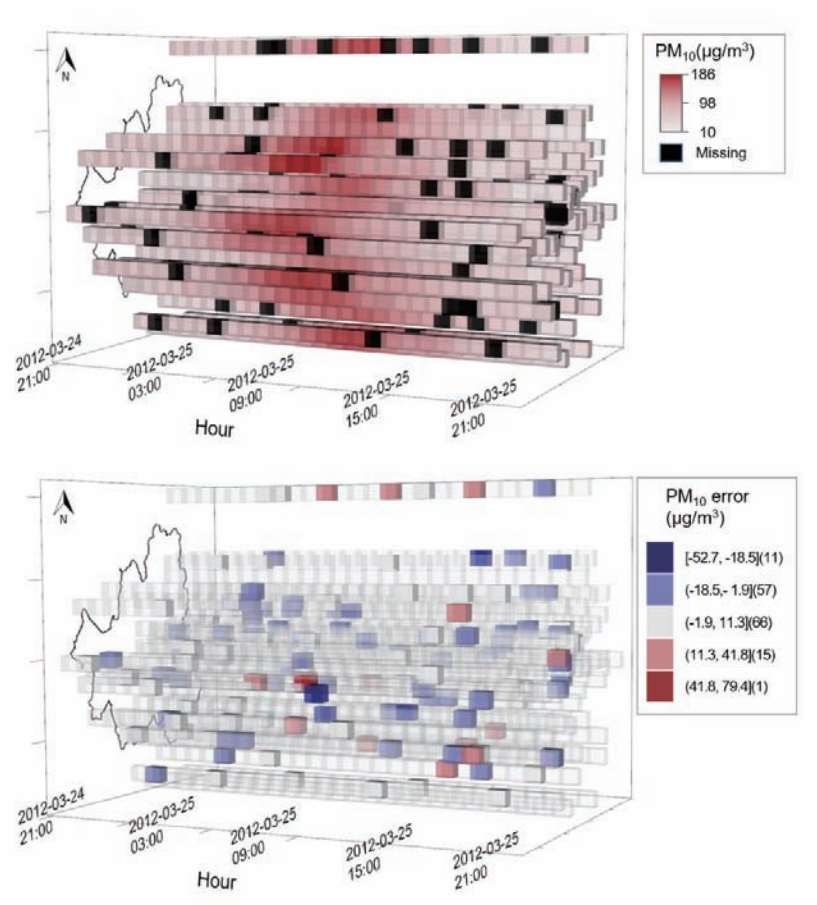

Figure 6. Space-time cubes of the set 8-3-1

Upper panel: the fitted data after generating missing values, Lower panel: residuals after the imputation
Note: the number of resampling was five; [ ] symbols represent 'higher/lower than or equal to', whereas ( ) symbols represent 'higher or lower than'; the numbers next to the ranges of residuals are the numbers of elements subject to each range.

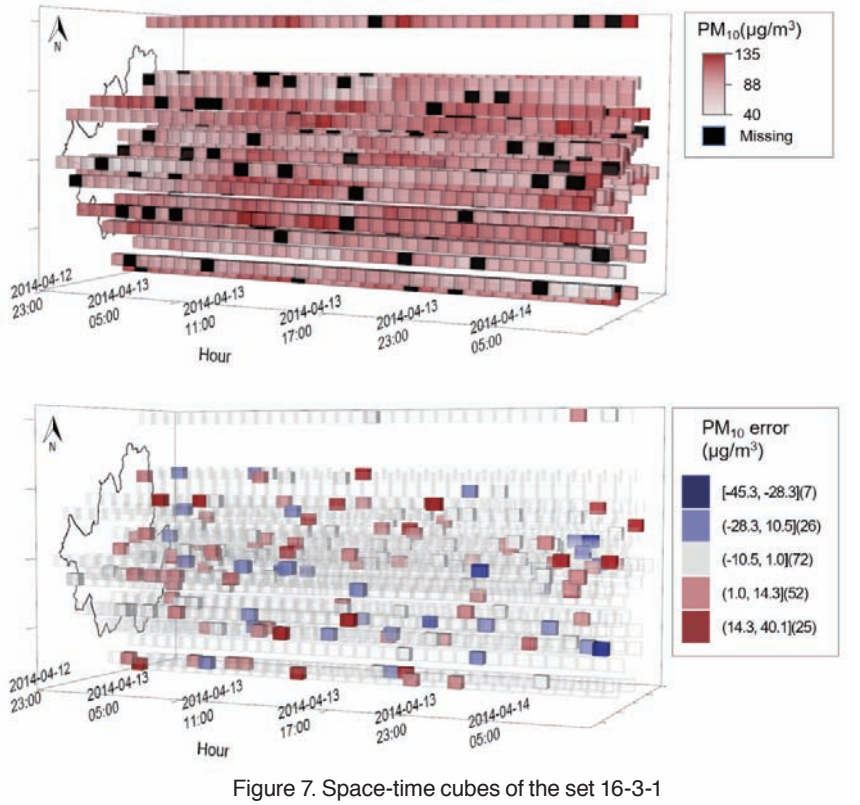

Figure 7. Space-time cubes of the set 16-3-1

Upper panel: the fitted data after generating missing values, Lower panel: residuals after the imputation
Note: the number of resampling is five.

design spatio-temporal kriging models. We arbitrarily set spatial lag distance as 1,000 meters. Thus, this is one of the limitations of parametric theoretical spatio-temporal variogram models. We can see that further research is required to account for local variations flexibly, i.e. variogram fitting with the piecewise linear model (Shapiro and Botha, 1991) or spatio-temporal hierarchical models, and Bayesian techniques, which consider variogram parameters to be distributions (Cressie and Wikle, 2011; Montero *et al.*, 2015). The last to be noted is results of this study suggest the applicability of spatio-temporal kriging for imputation in the short-term and data with dispersed missing values. As the test sets have 24 to 60 hours of time span, the performance gains as the models are fitted in other time scales (i.e., monthly or yearly scale) may not be observed. Also, as seen in Figure S1 in the supplementary material, continuous missing values in PM_{10} measurement data are present for longer than a few hours to even for a few weeks. This might destabilize spatio-temporal variograms for resampled data such that cause underperformance of spatio-temporal kriging. We see that further studies will scrutinize such issues in different data and study problems.

6. Conclusion

This study suggested an imputation approach that hybridizes resampling techniques and spatio-temporal

kriging for spatio-temporal data. The proposed approach was applied to impute missing values in 17 selected datasets from the PM_{10} measurement data of five years to test the effectiveness. Results are summarized as follows. First, our approach confirmed the accuracy gain by the ensemble of spatio-temporal kriging models for resampled sub datasets. It showed that RMSE decreased the most on average when each dataset was resampled 3-5 times. This demonstrates that the resampling approach for spatio-temporal kriging can contribute to deal with the problems of missing values in spatio-temporal data. Second, our approach outperformed MICE, which is based on linear regression models, and showed similar performance to the machine learning based algorithm missForest. Moreover, our approaches gained an advantage over the compared algorithms in both the explanation of inherent spatio-temporal variability of the dataset and the enhancement of the imputation accuracy with resampling.

This study offers two contributions. First, we examined the effect of the combination of spatio-temporal kriging and the learning method, thus observing the expandability of spatio-temporal kriging by reinforcing the model from the data itself. Second, this study provided a perspective on the missing values in spatio-temporal data which have been undermined in geography and called forth the need for subsequent studies related to missing values. Nevertheless, there are several limitations in this study. We succeeded in suggesting a working example of our approach. However, there are challenges such as providing generalized examples that employ simulation approaches and mathematical foundations of the availability of our approach. In addition, as the accuracy of the missForest was slightly higher than that of our approach, further research is needed to find how the machine learning algorithm reflects the inherent dependence of spatio-temporal data. As for the kriging methodology, it is necessary to follow up on the recent kriging method using the Bayesian approach, nonparametric approach, and efficient kriging algorithm for huge data.

Notes

1) Missing values mean the unmeasured values in data. In a wide sense, missing values are 'no measured or reported values' at which there are indices such as row and column numbers that had values therein, whereas the data has partial missing values which are directly related to the phenomena of interest in a narrow sense (Little and Rubin, 2002; McKnight *et al.*, 2007).
2) There are four approaches to estimating the spatio-temporal anisotropy: ratio-linear, range, variogram, and metric. Such approaches vary in their ways of finding the optimal scaling method, which converts temporal units into spatial units and vice versa by matching the values of spatial and temporal marginal variograms (Gräler *et al.*, 2016).
3) The procedure was compiled by adding functions for fitting spatio-temporal variograms automatically based on R (R Core Team, 2017) packages `gstat` (Pebesma, 2004), `spacetime` (Pebesma, 2012), and `automap` (Hiemstra *et al.*, 2009).
4) Multiple imputation by chained equations is an imputation algorithm employing chained equations to fill missing values in multivariate data (Schafer, 1997; van Buuren and Groothuis-Oudshoorn, 2011).
5) missForest is an imputation algorithm which implements random forest, one of the machine learning algorithms (Steckhoven and Bühlmann, 2012).
6) 41 out of 54 monitoring stations are urban ambient stations whereas 13 stations are roadside stations. Urban ambient stations are established to measure the background level of air pollutants in cities, while roadside stations are installed near the major roads to monitor the effect of road traffic to the concentrations of air pollutants (국립환경과학원, 2015).
7) There were 1,826 days (365 days × 5 years +1 leap day) for the study period, such that there were 43,824 hours (24 hours × 1,826 days) per monitoring station and 2,366,496 hours for 54 monitoring stations (54 stations × 43,824 hours) in total.

References

국립환경과학원, 2015, 대기환경연보 2014.

김병식 · 노희성 · 김형수, 2011, "시공간적 변동성을 고려한 강우의 결측치 추정 방법의 비교," 한국습지학회지, 13(2), 189−197.

박노욱, 2011, "시계열 환경변수 분포도 작성 및 불확실성 모델링: 미세먼지(PM_{10}) 농도 분포도 작성 사례연구," 한국지구과학회지, 32(3), 249−264.

한국환경공단, 2018, 에어코리아. http://www.airkorea.or.kr

Bennett, R. J., Haining, R. P. and Griffith, D. A., 1984, The problem of missing data on spatial surfaces, *Annals of the Association of American Geographers*, 74(1), 138-156.

Breiman, L., 1996, Bagging predictors, *Machine Learning*, 24, 123-140.

Chouakria, A. D. and Nagabhushan, P. N., 2007, Adaptive dissimilarity index for measuring time series proximity, *Advances in Data Analysis and Classification*, 1(1), 5-21.

Cressie, N. and Wikle, C. K., 2011, *Statistics for Spatio-temporal Data,* Wiley, Hoboken.

Deng, M., Fan, Z., Liu, Q. and Gong, J., 2016, A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets, *International Journal of Geo-Information*, 5(13), 1-14.

Feng, L., Nowak, G., O'Neill, T. J. and Welsh, A. H., 2014, CUTOFF: a spatio-temporal imputation method, *Journal of Hydrology*, 519, 3591-3605.

Gräler, B., Pebesma, E. and Heuvelink, G., 2016, Spatio-temporal interpolation using `gstat`, *R Journal*, 8(1), 204-218.

Haining, R., Griffith, D. and Bennett, R., 1989, Maximum likelihood estimation with missing spatial data and with an application to remotely sensed data, *Communications in Statistics - Theory and Methods*, 18(5), 1875-1894.

Harrell, F. E. J., 2015, *Regression Modeling Strategies - With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis,* Springer, New York.

Heuvelink, G. B. M. and Griffith, D. A., 2010, Space-time geostatistics for geography: a case study of radiation monitoring across parts of Germany, *Geographical Analysis*, 42(2), 161-179.

Hiemstra, P. H., Pebesma, E. J., Twenhöfel, C. J. W. and Heuvelink, G. B. M., 2009, Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network, *Computers and Geosciences*, 35(8), 1711-1721.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M., 2004, Methods for imputation of missing values in air quality data sets, *Atmospheric Environment*, 38(18), 2895-2907.

Karahalios, A., Baglietto, L., Lee, K. J., English, D. R., Carlin, J. B. and Simpson, J. A., 2013, The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: a simulation study, *Emerging themes in epidemiology*, 10(1), 6.

Kondrashov, D. and Ghil, M., 2006, Spatio-temporal filling of missing points in geophysical data sets, *Nonlinear Processes in Geophysics*, 13(2), 151-159.

Kuhn, M. and Johnson, K., 2013, *Applied Predictive Modeling,* Springer, New York.

Li, Y. and Parker, L. E., 2014, Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks, *Information Fusion*, 15, 64-79.

Little, Roderick J. A. and Rubin, Donald B., 2002, *Statistical Analysis with Missing Data*, 2nd Edition, Wiley, Chichester.

McKnight, P. E., McKnight, K. M., Sidani, S. and Figueredo, A. J., 2007, *Missing Data - A Gentle Introduction.* The Guilford Press, New York.

Montero, J. M., Fernández-Avilés, G. and Mateu, J., 2015, *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging,* John Wiley and Sons, Chichester.

Opitz, D. and Maclin, R., 1999, Popular ensemble methods: an empirical study, *Journal of Artificial Intelligence Research*, 11, 169-198.

Pebesma, E., 2004, Multivariate geostatistics in S: the `gstat` package, *Computers and Geosciences*, 30(7), 683-691.

Pebesma, E., 2012, `spacetime`: spatio-temporal data in R, *Journal of Statistical Software*, 51(7), 1-30.

R Core Team, 2017, *R: A language and environment for statistical computing,* version 3.3.3, R Foundation for Statistical Computing, Vienna, Austria.

Schafer, J. L., 1997, *Analysis of Incomplete Multivariate Data,* Chapman and Hall/CRC, Boca Raton.

Schneider, T., 2001, Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values, *Journal of Climate*, 14(5), 853-871.

Shapiro, A. and Botha, J. D., 1991, Variogram fitting with a general class of conditionally nonnegative definite functions, *Computational Statistics and Data Analysis*, 11(1), 87-96.

Smith, R. L., Kolenikov, S. and Cox, L. H., 2003, Spatiotemporal modeling of $PM_{2.5}$ data with missing values: modeling of $PM_{2.5}$ data with missing values, *Journal of Geophysical Research: Atmospheres*, 108(D24), 9004.

Stekhoven, D. J. and Bühlmann, P., 2012, MissForest-nonparametric missing value imputation for mixed-type data, *Bioinformatics*, 28(1), 112-118.

van Buuren, S. and Groothuis-Oudshoorn, K., 2011, Multivariate imputation by chained equations, *Journal of Statistical Software*, 45(3), 1-67.

Wackernagel, H., 2003, *Multivariate Geostatiatics - An Introduction with Applications,* Springer Verlag, New York.

Wilson, B. T., Lister, A. J. and Riemann, R. I., 2012, A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data, *Forest Ecology and Management*, 271, 182-198.

Xia, Y., Fabian, P., Stohl, A. and Winterhalter, M., 1999, Forest climatology: estimation of missing values for Bavaria, Germany, *Agricultural and Forest Meteorology*, 96(1999), 131-144.

교신: 박기호, 08826, 서울특별시 관악구 관악로 1, 서울대학교 지리학과 (이메일: khp@snu.ac.kr, 전화: 02-880-6453)

Correspondence: Key-Ho Park, Department of Geography, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Korea (e-mail: khp@snu.ac.kr, phone: +82-2-880-6453)

Received April 23, 2018
Revised May 15, 2018
Accepted May 31, 2018

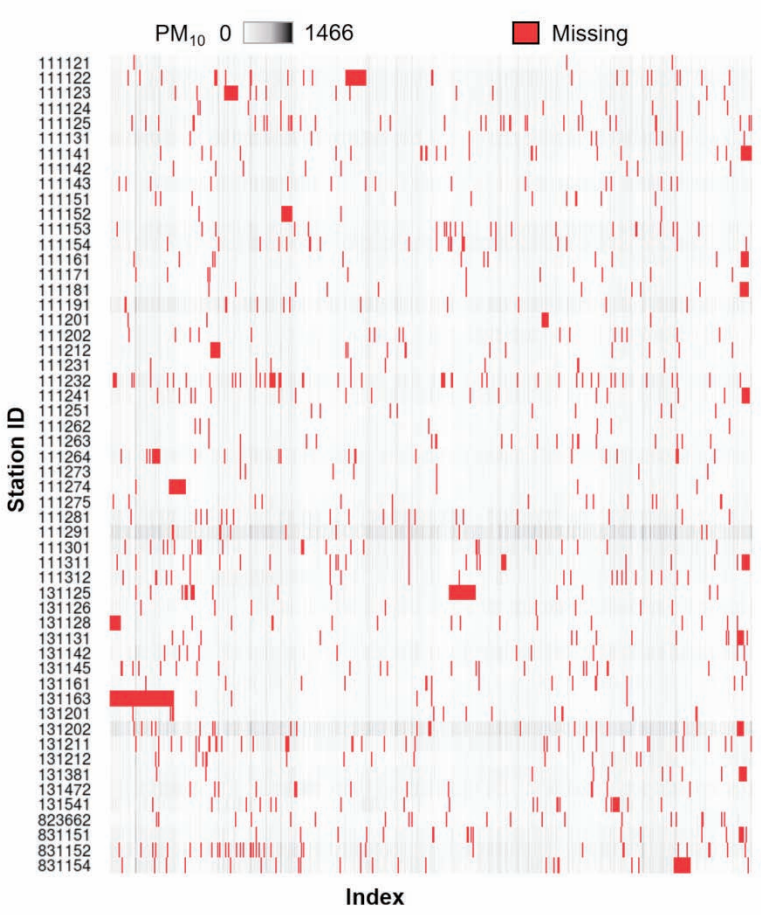

Figure S1. Matrix plot of missing values in the measurement data of 54 monitoring stations

Table S1. Summary statistics of RMSEs by sets and imputation algorithms

unit: μg/m^3

Set	Algorithm	Mean	Minimum	Maximum	SD
1	ST-kriging*	8.18	4.56	14.56	2.25
	missForest	5.82	4.34	8.73	1.02
	MICE	13.80	10.35	17.57	1.94
2	ST-kriging	11.75	6.44	17.22	2.79
	missForest	8.34	6.79	10.07	0.92
	MICE	22.17	18.59	28.46	2.70
3	ST-kriging	6.34	3.61	9.43	1.42
	missForest	5.01	3.75	7.39	0.77
	MICE	10.26	6.98	13.97	1.71
4	ST-kriging	8.46	5.44	13.43	1.74
	missForest	7.27	5.69	8.58	0.78
	MICE	17.87	15.16	22.88	1.87
5	ST-kriging	6.05	4.18	9.52	1.01
	missForest	4.55	3.45	6.18	0.66
	MICE	10.97	9.42	13.53	1.23
6	ST-kriging	4.41	3.25	5.86	0.75
	missForest	3.81	2.75	5.08	0.53
	MICE	8.12	5.30	11.44	1.25
7	ST-kriging	8.70	5.64	11.94	1.70
	missForest	6.90	4.93	9.33	1.13
	MICE	21.64	17.84	25.43	1.76
8	ST-kriging	10.98	5.70	17.46	3.33
	missForest	7.84	5.43	11.67	1.44
	MICE	45.90	34.33	59.77	5.73
9	ST-kriging	12.93	8.23	16.95	2.33
	missForest	8.63	6.15	14.68	1.61
	MICE	43.07	36.88	48.78	3.45
10	ST-kriging	6.15	3.74	10.62	1.46
	missForest	4.90	2.98	6.40	0.89
	MICE	10.27	8.38	12.46	1.25
11	ST-kriging	9.99	5.92	15.66	2.02
	missForest	7.54	5.57	9.93	1.23
	MICE	21.87	17.31	28.86	2.51
12	ST-kriging	10.74	5.05	17.12	3.09
	missForest	6.60	4.91	7.85	0.84
	MICE	26.36	19.31	35.61	3.61
13	ST-kriging	9.20	6.62	14.87	2.01
	missForest	7.60	5.85	10.58	1.18
	MICE	19.18	14.26	24.40	2.59
14	ST-kriging	16.11	8.25	26.16	4.10
	missForest	8.86	6.41	11.55	1.25
	MICE	36.01	29.81	41.63	3.21
15	ST-kriging	8.33	4.75	13.02	2.23
	missForest	6.07	4.62	9.25	1.16
	MICE	22.29	15.42	27.94	3.26
16	ST-kriging	11.93	7.72	16.32	2.20
	missForest	7.79	6.08	9.78	0.93
	MICE	18.64	13.83	20.77	1.68
17	ST-kriging	8.99	6.71	12.11	1.38
	missForest	6.48	5.10	7.62	0.66
	MICE	16.85	13.35	19.17	1.54

*ST-kriging: resampling-based spatio-temporal kriging, SD: standard deviation