Single imputation method of missing air quality data for i-Tree Eco analyses in the conterminous United States

Satoshi Hirabayashi[1] and Charles N. Kroll [2]

Version 1.0

January 9, 2017

[1]The Davey Tree Expert Company, 5 Moon Library, State University of New York, Syracuse, New York 13210, United States

[2]Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry, 1 Forestry Drive, 423 Baker Laboratory, Syracuse, New York 13210, United States

Abstract

Air pollution is a major environmental and human health concern in urban areas where urban forests can play an important role to remove air pollutants through dry deposition processes. Employing the US EPA's country-wide hourly air quality data set, the USDA Forest Service's i-Tree Eco assesses annual impacts of urban forests on air quality improvement. When applying i-Tree Eco, missing values in the air quality data degrades the assessments. The goal of this study is to develop a new single imputation method to fill gaps in the hourly air quality data set to enhance the applicability of i-Tree Eco across the conterminous United States. Considering weekly, daily and hourly time effects of air pollutant level, the developed method can estimate missing values with only on-site data. Compare to other standard single imputation methods, the performance of the developed method was the best for gaps greater than 4 hours, and as good as the other methods for smaller gaps.

1 Introduction

Air pollution is a major environmental and human health concern around the world, particularly in cities where pollutant sources are concentrated (Serageldin, 2002). Air pollutant can partially be removed from the air by dry deposition processes in which urban forests play an important role. Due to the relatively large surface area of a tree canopy compared to grass or bare earth, ecological engineering designs use trees as biological filters to remove air pollutant and improve air quality (Beckett et al., 1998). Based on hourly air quality measurements, the United States Department of Agriculture (USDA) Forest Service's i-Tree Eco (i-Tree, 2013) simulates hourly dry deposition rates of six criteria air pollutants (CAPs) to assess the annual impact of trees on environmental health risks in urban areas (Currie and Bass, 2008; Deutsch et al., 2005; Hirabayashi et al., 2011, 2012; Nowak et al., 1998, 2006; Nowak and Crane, 2000). Hourly air quality data employed by i-Tree Eco are provided by the United States Environmental Protection Agency (US EPA) repository of ambient air quality data, Air Quality System (AQS). AQS monitoring sites are located across the entire conterminous United States, which enables country-wide applications of i-Tree Eco; however, missing data occur at most sites.

Missing data is a very common problem in air quality studies. The major causes for missing air pollutant data includes monitor malfunctions and errors, power outages, computer system crashes, pollutant levels lower than detection limits, and filter changes (Imtiaz and Shah, 2008; Li et al., 1999). Missing data mechanisms can be generally classified into three types: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Little and Rubin, 2002; Schafer, 1997). Consider the complete data $X=x_i$ and the missing data indicator $M=m_i$, where $m_i=1$ if x_i is missing and $m_i=0$ otherwise. The missing data mechanism can be characterized by the function $f(M|X,\varphi)$, where φ denotes unknown parameters. For MCAR, $f(M|X,\varphi)=f(M|\varphi)$ for all X and φ. That is, missing of data does not depend on the value of X. For MAR, $f(M|X,\varphi)=f(M|X_{obs},\varphi)$ for all X_{miss} and φ, where X_{obs} and X_{miss} represent observed and missing components of X, respectively. Missing of data depends only on the components of X that are observed, and not on the components that are missing. For NMAR, $(M|X,\varphi)=f(M|X_{miss},\varphi)$. Missing of data depends on X_{miss}. In general, the mechanism of missing air quality data is MAR as the probability that a value is missing is not dependent on the missing values themselves (Junninen et al., 2004; Marwala, 2009).

Missing data imputation essentially means dealing with missing data. The most commonly used approach to deal with missing data is called case deletion, in which those cases with missing data are merely left out and analyses are performed on the remaining data (complete-case analysis). For instance, the statistical software package R automatically eliminates all cases in which any of the inputs are missing and runs analyses on the remaining data for classical regression and other models (Gelman and Hill, 2006).

Another common approach to deal with missing data is to estimate missing values with a variety of techniques that range from extremely simple to rather complex. These techniques can be categorized based on the number of imputed values (single or multiple) used to replace each missing value and the absence or presence of covariates (univariate or multivariate) (Little and Rubin, 2002). With single imputation, precisely one value is filled in for each missing value, and thus imputation efforts need to be carried out only once. Multiple imputation methods generate multiple simulated values for each missing data value to retain the uncertainty of the missing data, and may be a more statistically sound approach (Schafer, 1997). In univariate imputation, missing values of a single variable are estimated as a function of other observed values of the same variable. These methods are generally simple and straightforward and thus easy to implement, though the performance of missing data estimation may be poor. In multivariate imputation, missing values are estimated with concurrent records of the same or other variables, and the performance may be improved over univariate imputation. However, when a number of concurrent variables are missing, it can be difficult to replicate data patterns (Junninen et al., 2004). Univariate single imputation schemes commonly employed include using the mean or median of measured values, carrying the last observation forward or the next observation backward, or the average of the last and next observations (Engels and Diehr, 2003; Gelman and Hill, 2006; Marwala, 2009). Multivariate single imputation schemes commonly used are using the mean or median of the concurrently measured values, hot-deck, cold-deck, regression, regression with error (Engels and Diehr, 2003), neural networks, decision trees (Marwala, 2009), multivariate nearest neighbor (Junninen et al., 2004), expectation maximization (EM) (Dempster et al., 1977), spatio-temporal filling (Kondrashov and Ghil, 2006), and Kalman filters (Moffat et al., 2007).

In air quality studies, covariates such as meteorological data are often used. Junninen et al. (2004) employed air quality data as well as weather data that were concurrently measured at two sites to evaluate several imputation methods including univariate single imputation (linear, spline and nearest neighbor), multivariate single imputation (regression-based imputation, nearest neighbor, self-organizing map, multi-layer perceptron), hybrids of the above mentioned univariate and multivariate single imputation methods, and multi imputation methods (computed as a the mean of multivariate methods and hybrid methods). In air quality studies, covariates may be same air quality variables at different sites. Plaia and Bondi (2006) employed air quality data at eight neighboring sites and tested several imputation methods including univariate single imputation (hourly mean, mean of last and next), multivariate single imputation (mean of concurrently measured values at neighboring sites, site-dependent effect method (SDEM)), and multivariate multiple imputation methods (model-based multiple imputation).

In some large cities, the AQS monitoring site network is rather dense, and concurrent records are available at nearby stations to impute missing values. In more remote areas, on the other hand, monitoring sites are sparsely located and concurrent records are not available. In addition, for some monitoring sites, a dominant air pollutant emission source such as plants or other facilities may be specified. In such cases, nearby monitoring sites without a dominant source may have different air quality patterns. These conditions have led us to explore methods that only require on-site data to fill data gaps in AQS data (i.e. univariate single imputation method).

In the current study, a single imputation method developed to impute missing values at AQS air pollutant monitors across the conterminous United States is described. In this new method, site-dependent weekly, daily, and hourly patterns of the air pollutant concentrations are used to fill in hourly missing values of carbon monoxide (CO), nitrous oxide (NO_2), ozone (O_3), and sulfur dioxide (SO_2). The developed method is compared with other standard single imputation methods to evaluate the suitability of the method for AQS monitoring data.

2 Air Quality System (AQS) Data

The AQS national database contains ambient air quality data collected by US EPA, State, local, and tribal air pollution control agencies from thousands of monitoring sites (US EPA, 2010a). While the database contains values from 1957 through present, in this study we used AQS data from the 2010 calendar year to analyze missing data and develop a single imputation method of missing data for hourly measurements of CO, NO_2, O_3, and SO_2 across the conterminous United States. Table 1 presents the total number of AQS monitoring sites for these air pollutants in 2010. Among these sites, no monitor had a complete record of hourly observations for 2010. For each AQS monitoring site, an inter- or intra-state regional boundary called an Air Quality Control Region (AQCR) is assigned. The AQCRs were established by the US EPA based on jurisdictional boundaries, urban-industrial concentrations, and other factors such as air sheds, to provide adequate implementation of the Clean Air Act (US EPA, 1972). 247 AQCRs cover all fifty states, the District of Columbia, Puerto Rico, and the Virgin Islands, of which 239 AQCRs cover the contiguous United States.

AQS hourly pollutant concentrations vary spatially across the contiguous United States. Figure 1 provides box plots of the four pollutants for 2010 within selected AQCRs. For each plot, the AQCRs with the maximum and minimum medians for a specific pollutant are presented along with eight AQCRs with medians evenly spaced between the two. Larger medians were found in the AQCR containing major cities (i.e. 36: Metropolitan Denver, 67: Metropolitan Chicago, 94: Metropolitan Kansas City), while smaller medians were found in more remote areas (i.e. 88: Northeast Iowa, 206: South Dakota).

Spatial variation in pollutant data was influenced by pollutant sources. As two major emission sources of CO and NO_2 are transportation (motor vehicles) and industries (chemical plants, metal processing, electric utility, etc.), higher levels were found in urban and industrial areas. O_3 is a secondary pollutant that is not usually emitted directly into the air, but at ground-level is formed through chemical reactions between nitrogen oxides (NO_x) and volatile organic compounds (VOCs) in the presence of sunlight and heat (Sillman, 1999). Both NO_x and VOCs are emitted by motor vehicle exhaust, industrial sources, gasoline vapors, chemical solvents, and natural sources. There was relatively little spatial variation for O_3 because a secondary pollutant is less impacted by local sources (Ito el al., 2005; Sarnat et al., 2010). Wind may carry NO_x and VOC hundreds of miles away from their original source (US EPA, 2010b), which may smooth O_3 variation across space. Ambient SO_2 is emitted largely from stationary sources such as coal and oil combustion, steel mills, refineries, and

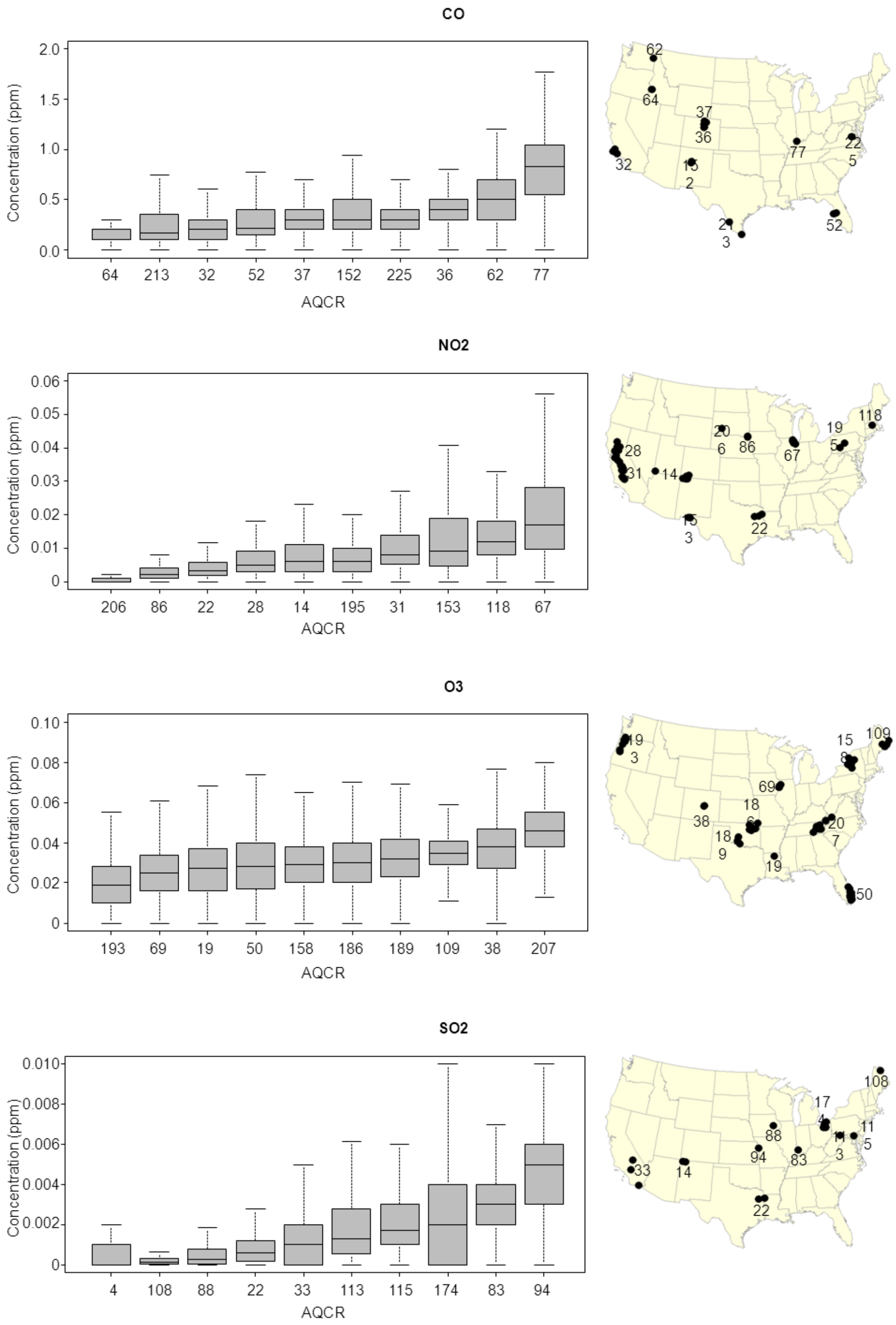

Figure 1 Spatial distribution of air pollutant concentration across the conterminous United States

pulp and paper mills (US EPA, 2013). Maximum SO_2 concentrations are generally observed within a few hundred kilometers of the major SO_2 sources, such as the Ohio River Valley (Seinfeld, 1986). AQCRs with larger SO_2 median can be found in 83 and 174, located within Ohio River Valley.

Temporal variation in air pollutant concentrations is typically influenced by seasonal and diurnal factors including transportation volume, industrial activities, meteorological cycles, and interactions between these factors (Capilla, 2007). In addition, there often are distinct differences in air pollutant concentrations due to human activities on weekdays and weekends. For the four air pollutants, the monitoring site with the longest length of available observations in 2010 was selected from each of the 10 AQCRs presented in Figure 1 (hereafter, selected monitoring sites are referred to as monitoring sites 1 to 10 in the order of AQCRs in Figure 1 for each pollutant). Figure 2 presents week-of-year, day-of-week, and hour-of-day means calculated for monitoring sites 1 to 10 of each pollutant. In most cases, all of the sites exhibit similar patterns for each pollutant, yet the patterns are different between pollutants. CO levels are sensitive to cold season temperature inversions which trap the gas beneath a layer of warm air, leading to higher concentrations (US EPA, 2010b). The seasonal influence of inversions is shown in Figure 2 where CO concentrations were higher during the autumn and winter weeks. Lower concentrations in the weekends and during off-peak commuting hours can be explained by reduced traffic. NO_2 levels also peak in the winter during inversions that reduce mixing and entrainment (Atkins and Lee, 1995; Hargreaves et al., 2000). Lower NO_2 values in summer weeks, weekends, and mid-day may be attributed to dry deposition on vegetative surfaces (Nowak et al., 2006). Due to transportation and industrial activities NO_2 levels were higher during weekdays around the commuting hours and the peak of industrial activity in the afternoon. The photochemical production of O_3 reaches peak values during warmer weeks and hours. Although SO_2 shows no distinct seasonal or week-of-day pattern, the peak levels tend to occur near noon when the boundary layer is deep and SO_2 aloft could be more effectively mixed down to the surface (Chen et al., 2001).

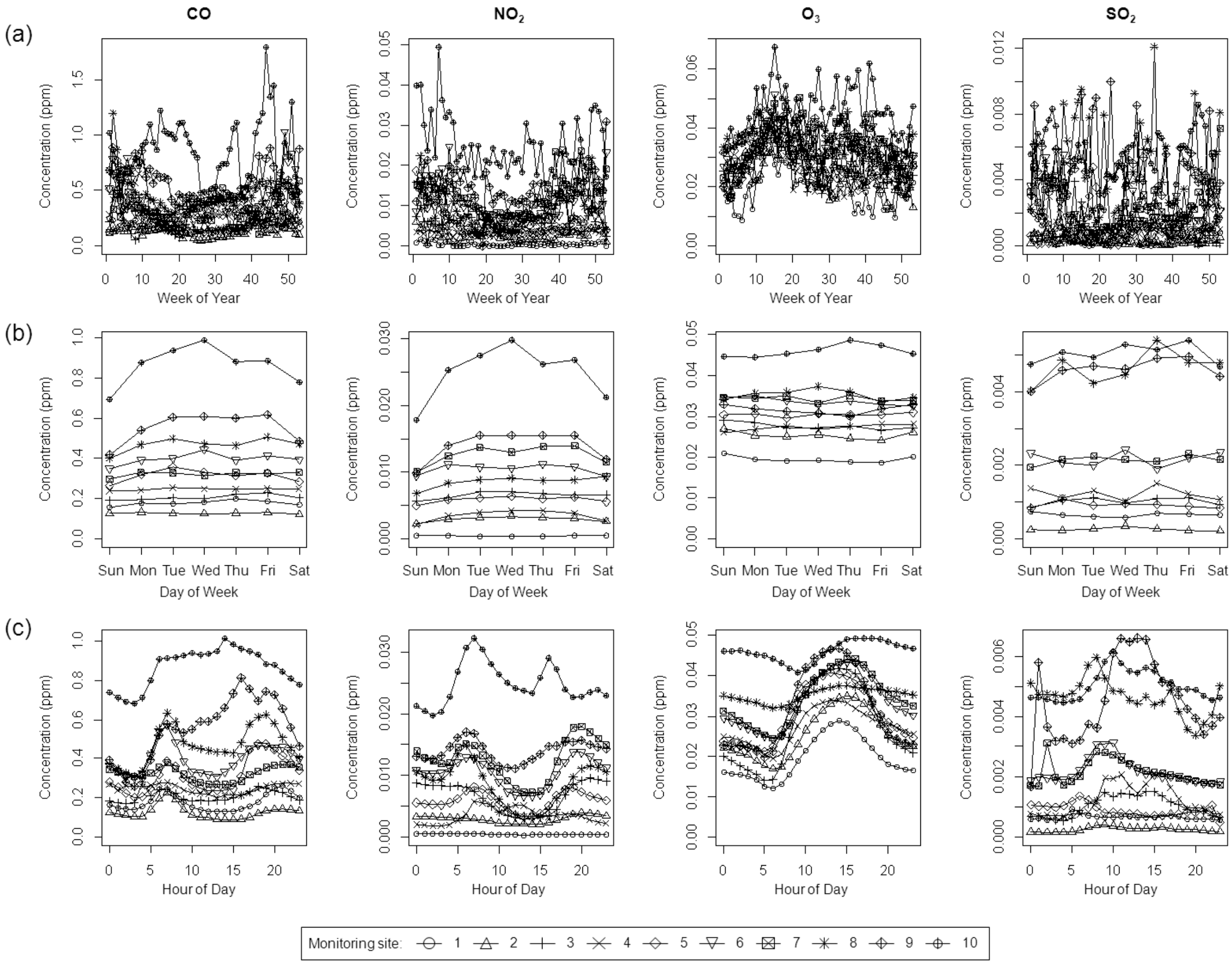

Figure 2 (a) week-of-year, (b) day-of-week and (c) hour-of-day means of pollutant concentration for monitoring sites 1 to 10 for the four pollutants

Week-of-year, day-of-week and hour-of-day means were calculated for each monitoring sites. Throughout the year, 53 week-of-year, 7 day-of-week, and 24 hour-of-day means needed to be calculated. Note that the week 1 and 53 may include less than seven days as the week assignment starts on Sundays and ends on Saturdays. This assignment was employed because human activities may be more tightly tied to weeks starting from Sunday. If no data was available for an averaging period at a monitoring site, the criteria for the number of means were not satisfied and thus this site was excluded from the further analyses (Table 1). Missing data for O_3 tend to present a distinct temporal pattern, since many monitoring sites discontinue their O_3 monitoring during the cold season, and resume during the warm “ozone season” from spring to fall (CFR, 2010). As a result, for 2010 only 31.5% of O_3 monitoring sites have measurement records where all means can be calculated. While the AQS monitor sites are distributed nationally, density of coverage varies, and sites included in the further analyses are concentrated in major cities in the Midwest as well as coastal regions.

For each of the sites included in the analyses, percentage of missing values as well as percentage of the 1-hour, 2-hour, 3- to 24-hour, and over 24-hour gap occurrences were calculated, and averaged for all monitors (Table 2). For each air pollutant, on average 2.6% - 4.7% of the total 8760-hour records was missing. In most cases, monitoring was interrupted for only a few hours, and over 97% of missing data were gaps less than 24 successive hours.

Table 1 Number of monitoring sites in the conterminous United States in 2010 (sites included are those for which week-of-year, day-of-week and hour-of-day means can be calculated, while sites excluded are those for which these means are unable to calculate due to a large number of missing values).

Air pollutant	Sites included		Sites excluded		Total
CO	198	(59.5%)	135	(40.5%)	333
NO_2	220	(54.7%)	182	(45.3%)	402
O_3	391	(31.5%)	849	(68.5%)	1240
SO_2	268	(61.2%)	170	(38.8%)	438

3 Imputation Methods

3.1 Modified Site-Dependent Effect Method (SDEMm)

The new method proposed in the current study (which will be referred to as SDEMm) is a modification of Plaia and Bondi (2006)'s SDEM. SDEMm uses week-of-year, day-of-week and hour-of-day means at the target site. Missing values at time, *t* specified with week-of-year, *w* (=1,2,…, 53), day-of-week, *d* (=1,2,…,7), and hour-of-day, *h* (=0,1,…23), $\hat{x}_{t_{wdh}}$ are estimated as the week-of-year mean multiplied by the day-of-week effect and the hour-of-day effect, and then adjusted with the last observation before the missing values and the next observation after the missing values:

$$\hat{x}_{t_{wdh}} = \bar{x}_w \times \frac{\bar{x}_d}{\frac{1}{7}\sum_{d=1}^{7} \bar{x}_d} \times \frac{\bar{x}_h}{\frac{1}{24}\sum_{h=0}^{23} \bar{x}_h} \times F \qquad (1)$$

where $\bar{x}_w$ is week-of-year mean, $\bar{x}_d$ is day-of-week mean, $\bar{x}_h$ is hour-of-day mean, and *F* is an adjustment factor:

$$F = \frac{\frac{x_{t_{last}}}{\hat{x}_{t_{last}}} + \frac{x_{t_{next}}}{\hat{x}_{t_{next}}}}{2} \qquad (2)$$

where $x_{t_{last}}$ and $x_{t_{next}}$ are the last and next observations for missing values at time, t_{wdh}, respectively ($t_{last} < t_{wdh} < t_{next}$), and $\hat{x}_{t_{last}}$ and $\hat{x}_{t_{next}}$ are values estimated with Equation 1. Preliminary studies showed when the hourly variation of the observed data was deviated a large amount from $\bar{x}_h$, this method could produce extremely large or small $\hat{x}_{t_{wdh}}$. To avoid this, $\hat{x}_{t_{wdh}}$ is capped with minimum and maximum of observed values.

3.2 Hour Mean (HM)

This method assigns the mean of all known hourly observations at the same monitoring site at the same hour over the entire year to the missing value:

$$\hat{x}_{t_{wdh}} = \bar{x}_{\mathrm{h}} \qquad (3)$$

3.3 Linear Interpolation (LIN)

This method fits a straight line between the last and next observations of a gap and estimates the missing values based on linear interpolation:

$$\hat{x}_{t_{wdh}} = x_{t_{last}} + a(\mathrm{t}_{wdh} - t_{last}) \qquad (4)$$

where $a = \frac{x_{t_{next}} - x_{t_{last}}}{t_{next} - t_{last}}$, and $t_{last} < t_{wdh} < t_{next}$

3.4 Last & Next (LN)

This method fills in missing values with the mean of the last and next observations of a gap:

$$\hat{x}_{t_{wdh}} = \frac{x_{t_{last}} + x_{t_{next}}}{2} \qquad (5)$$

where $t_{last} < t_{wdh} < t_{next}$

3.5 Nearest Neighbor (NN)

In this method, the last or next observations of the gaps are used as the estimates for all the missing values, depending on the temporal proximity of the missing value to these know observations:

$$\hat{x}_{t_{wdh}} = x_{t_{last}} \quad \text{if} \quad t_{wdh} \leq t_{last} + \frac{t_{next} - t_{last}}{2}$$

$$\hat{x}_{t_{wdh}} = x_{t_{next}} \quad \text{if} \quad t_{wdh} > t_{last} + \frac{t_{next} - t_{last}}{2} \qquad (6)$$

4 Evaluation of imputation methods

4.1 Missing value simulations

The imputation methods were applied to simulated missing values for monitoring sites 1 to 10 for each of the four pollutants. For the 10 CO monitoring sites, on average 2.8% of records were missing, while for NO_2, O_3 and SO_2 monitors, 3.1%, 2.6% and 3.9%, respectively, were missing. To simulate missing values in these monitor records, the missing values originally in the records were deleted (case deletion) to get the complete case. As shown in Table 2, more than 97% of the gap occurrences were less than 24 successive hours. To simulate these gap characteristics, three simulations were performed for each of the monitoring sites, in which 1-, 2- and 24-hour gaps were created from the beginning through the end of the records by sequentially moving the start hour of the gap by 1 hour throughout the record. This way, our test can thoroughly simulate possible patterns of 1-, 2- and 24-hour of missing data.

Table 2 Gap length statistics

Air pollutant	Gap (%)	1-hour gap (%)	2-hour gap (%)	3- to 24-hour gap (%)	Over 24-hour gap (%)
CO	3.3	64.9	21.9	11.4	1.9
NO_2	4.7	41.1	33.8	22.6	2.5
O_3	2.6	63.2	22.4	12.6	1.9
SO_2	2.8	51.8	29.7	16.2	2.4

4.2 Performance metrics

For the simulations performed, the estimated missing values and their original values were compared to evaluate the performance of the imputation methods. Following the recommendations on model validation by Willmott (1981) and Legates and McCabe (1999), we considered two dimensionless measures (index of agreement (d_2) and coefficient of efficiency (E_2)) and one measure quantifying errors in terms of the units of the variable (mean absolute error (MAE)) along with the observed and estimated mean and standard deviations. In addition, we employed normalized mean bias (NMB) to assess average model bias percent relative to observed mean; that is, average over- or under-estimation.

d_2 was developed by Willmott (1981) as:

$$d_2 = 1.0 - \frac{\Sigma\left(x_{t_{wdh}} - \hat{x}_{t_{wdh}}\right)^2}{\Sigma\left(\left|\hat{x}_{t_{wdh}} - \bar{x}_{t_{wdh}}\right| + \left|x_{t_{wdh}} - \bar{x}_{t_{wdh}}\right|\right)^2} \qquad (7)$$

d_2 ranges from 0.0 to 1.0, with higher values indicating better agreement.
E_2 can be calculated as:

$$E_2 = 1.0 - \frac{\sum\left(x_{t_{wdh}} - \hat{x}_{t_{wdh}}\right)^2}{\sum\left(x_{t_{wdh}} - \bar{x}_{t_{wdh}}\right)^2} \tag{8}$$

E_2 ranges from negative infinity to 1.0 with the higher values indicating better agreement (Nash and Sutcliffe, 1970). Physically, E_2 is the ratio of the mean square error (numerator) to the variance in the observed data (denominator), subtracted from 1.0. MAE can be calculated as:

$$MAE = \frac{1}{N}\sum\left|x_{t_{wdh}} - \hat{x}_{t_{wdh}}\right| \tag{9}$$

where N is the number of tested values at a single site. NMB can be calculated as:

$$NMB = \frac{\sum\left(\hat{x}_{t_{wdh}} - x_{t_{wdh}}\right)}{\sum x_{t_{wdh}}} \times 100 = \frac{\bar{\hat{x}}_{t_{wdh}} - \bar{x}_{t_{wdh}}}{\bar{x}_{t_{wdh}}} \times 100 \tag{10}$$

5 Results and Discussions

5.1 Dimensionless Measures

Figure 3 presents d_2 obtained from 1-, 2- and 24-hour gap tests for each of the ten monitoring sites of the four pollutants. Overall, the imputation methods worked the best for O_3, and the worst for SO_2. The performances for NO_2 and CO were between those for O_3 and SO_2. The performance decreased as the gap length increased and in the following order: SDEMm, LIN, LN, NN and HM. The difference of d_2 between the best four methods increased as the gap length increased.

d_2 of HM was the worst across the tests, and almost identical across gaps for each site since hour means used to fill gaps were almost identical for 1-, 2-, and 24-hour gap tests. For 1-hour gaps, LIN and LN both estimated the missing value by taking average of the last and next observations, thus resulted in the exactly same performance. As shown in Figure 2, the concentration changes almost linearly between daily maxima and minima, and thus LIN and LN worked very well for short gaps (1- and 2-hour gaps). SDEMm performed as well as

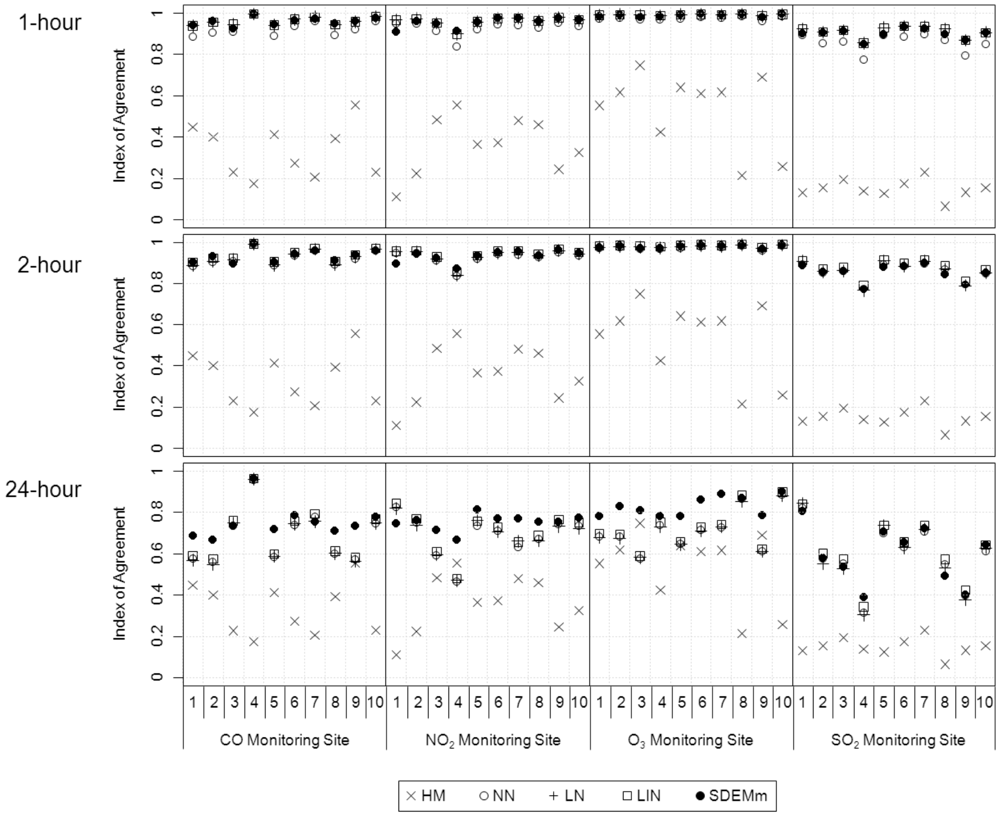

Figure 3 Index of agreement (d_2) for the 1-, 2- and 24-hour gap tests for monitoring sites 1 to 10 of the four pollutants

these two methods for short gaps, while the NN exhibited a slightly worse performance. In the 24-hour gap tests, SDEMm outperformed other methods for CO, NO_2 and O_3, while for SO_2, SDEMm was slightly worse than LIN (mean d_2 across 10 sites was 0.59 for SDEMm, LN and NN, while 0.61for LIN), but the variation of d_2 across the ten monitoring sites for SDEMm was the smallest (standard deviation was 0.14 for SDEMm, while > 0.15 for LIN, LN and NN). Compared to the other three pollutants, d_2 across the 10 sites varied the most for SO_2 in the 24-hour gap tests (standard deviations of SDEMm were 0.08, 0.04, and 0.05 for CO, NO_2, and O_3 respectively).

The above results indicated that SDEMm generally worked the best, but performance varied across the monitoring sites. The model estimates missing values based on three time effects: week-of-year, day-of-week, and hour-of-day means. We investigated how the variations in these time-averaged values affected the variations in d_2. No relationship between d_2 and week-of-year and day-of-week means was identified, while the negative correlation between d_2 and coefficient of variation (CV) of hour-of-day mean was observed for CO, O_3 and SO_2 (Table 3). This indicates that the smaller the variation in hour-of-day means at a monitoring site, the better SDEMm works for that site. Since the gap simulated was inserted in consecutive hours, the performance was influenced by hour-of-day effects.

Table 3 Correlation coefficient between coefficient of variation (CV) of hour-of-day averages and d2

Air pollutant	1-hour gap	2-hour gap	24-hour gap
CO	-0.44	-0.58[a]	-0.66[a]
NO_2	-0.17	-0.36	-0.44
O_3	-0.70[a]	-0.73[a]	-0.66[a]
SO_2	-0.55[a]	-0.64[a]	-0.67[a]

[a] Significantly different than zero at a 10% level.

Coefficient of efficiency (E_2) showed similar results as d_2: the SDEMm, LIN and LN exhibited comparable E_2 for 1- and 2-hour gaps, and NN had a slightly worse E_2. Imputation for O_3 worked the best (SDEMm's E_2 for 1-hour gaps ranged from 0.93-0.98 (mean 0.96), and for 2-hour gaps ranged from 0.88-0.95 (0.92)). For 24-hour gaps, SDEMm had a greater E_2 than the other methods for CO, NO_2 and O_3 (Figure 4). It should be noted that E_2 was negative for some monitoring sites tested with 24-hour gaps. Negative E_2 indicates that the mean square error (numerator) exceeds the variance in the observed data (denominator), and thus the observation mean is a better predictor than the model (Wilcox et al., 1990; Legates and McCabe, 1999). This situation was observed for LIN, LN, and NN applied to some monitoring sites for the four air pollutants as well as the SDEMm applied to some SO_2 monitors.

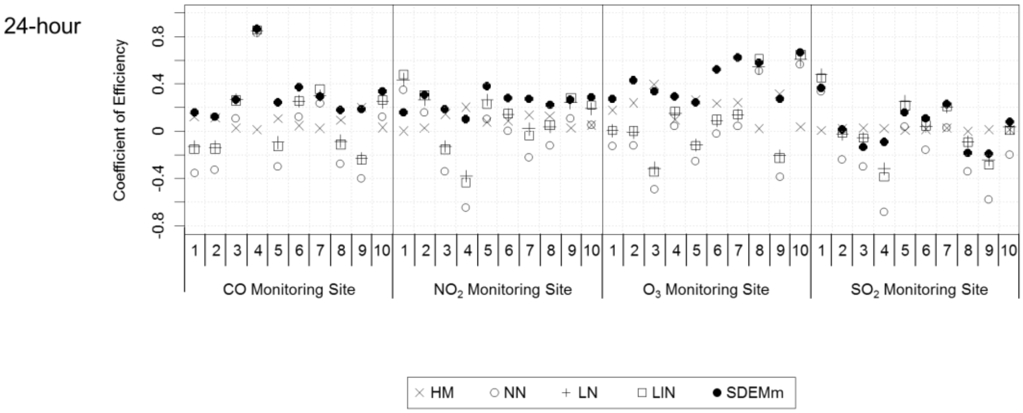

Figure 4 Coefficient of efficiency (E_2) for 24-hour gap tests for monitoring sites 1 to 10 of the four pollutants

5.2 Error analyses

MAE provides the magnitude of imputation errors in terms of the units of the variable. In the 24-hour gap test (worst case in the three gap length tests), SDEMm had the smallest MAE for CO, NO_2 and O_3, which were on average 0.132 ppm, 0.004 ppm and 0.0077 ppm (37%, 53% and 26% of the observation mean, i.e. the normalized MAE), respectively (Figure 5). For SO_2, the MAE was similar for the five methods and was on average about 0.015 ppm (70% of the observation mean).

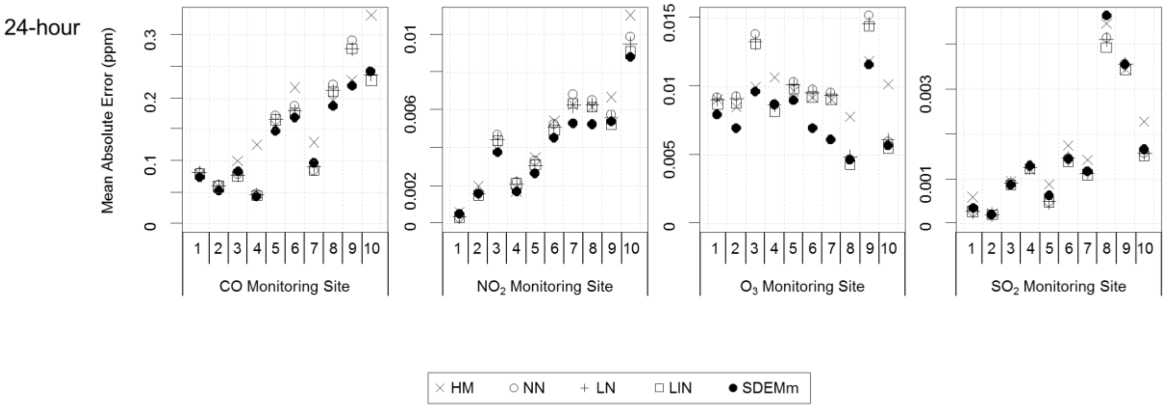

Figure 5 Mean absolute error for the 24-hour gap tests for monitoring sites 1 to 10 of the four pollutants

NMB indicates average over- or under-prediction of the model. Regardless of pollutants and gap hours, HM had nearly 0% NMB. This is due to the fact that this method imputed missing values with the mean of the observations at the target hour over the year. The mean of the estimates is thus almost identical to the observed mean throughout the year. For the entire year, the mean bias (estimated mean - observed mean) was almost 0, leading nearly 0% NMB. The standard deviation of estimates was much smaller than observed for HM as it used a limited number of values to estimate missing values throughout the year.

For a single imputation of a 24-hour gap, the mean of the concentrations estimated by the LIN was the mean of the last and next observations, LN used this mean value for all imputed values, while the NN used the last value to fill the first 12-hour gap and the next value to fill the second 12-hour gap. Therefore, when averaged, hourly concentrations estimated by LIN, LN and NN are identical. As a result, for the entire year comparison at each monitoring site, observed mean and estimated mean by these three methods were identical. This is true for 2-hour gaps as well. For 1-hour gaps, LIN and LN produced the same estimates, while NN used the last observation to fill the gap, which produced a slight difference in estimated mean between NN and LIN. Regardless, throughout the year, estimated mean by the three methods and observed mean were almost identical, leading to a NMB of nearly 0%. In addition, estimates from these three methods never exceeded the last and next observed values, and the standard deviations of these estimates were almost the same as the observed value.

For SDEMm, the week-of-year, day-of-week and hour-of-day effects were all rounded by taking means, and thus large values tended to be under-estimated and small values tended to be over-estimated. Throughout the year, SDEMm generally over-estimated the observed concentration for the four pollutants. For the 24-hour gap tests, NMB for SDEMm was on average 1.7%, 6.1%, 0.7%, and 10.1% for CO, NO_2, O_3, and SO_2, respectively.

To further investigate the relationship of the gap length and the imputation methods, gap lengths of 4 to 48 hours (increasing by 2) were additionally tested using data at one O_3 monitoring site in AQCR186. Figure 6 presents the four performance measures against the gap length tested. Based on d_2, E_2 and MAE, the results of the LIN, LN, NN and SDEMm were equally good for gaps of 1 to 4 hours, but the performance of the LIN, LN and NN declined faster as the gap length increased. It can be observed that the performance kept degrading as the gap length increased up to 20 to 24 hours, and then began improving up to

30 to 32 hours for LIN and LN. This may be due to the autocorrelated diurnal variation of the pollution level with approximately a 24-hour lag. The SDEMm method had the best performance across all gap lengths tested. Over- or under-estimation of the models indicated by NMB were in a small range (-0.1 to 0.1 %). SDEMm generally over-estimated and the other four methods under-estimated the measurements. As the gap hours increased, the magnitude of over- or under-estimation increased. The improvement observed from 12 to 20 hours for SDEMm may be again due to the autocorrelation of the pollution level.

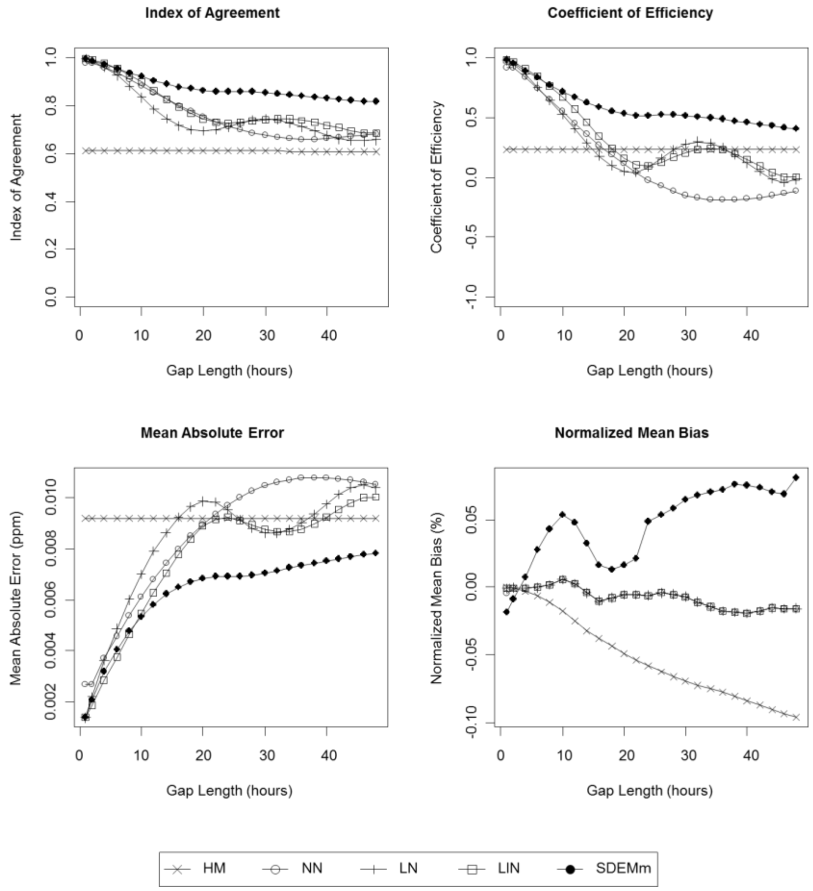

Figure 6 Performance of simple imputation methods as a function of gap length. The test was performed using single O_3 monitor in AQCR186

6 Conclusions

This study developed a single imputation method to fill missing CO, NO_2, O_3, and SO_2 hourly values for the US EPA's AQS data. The newly developed method, called SDEMm, estimates missing values for sites with no covariates such as weather data or concurrent air quality data. SDEMm can estimate missing values with only on-site data for monitors where week-of-year, day-of-week and hour-of-day means can be calculated. Comparison with other standard single imputation methods (i.e. hour mean, linear, last & next and nearest neighbor) revealed that the performance of SDEMm was the best for gaps greater than 4 hours, and as good as any of the other methods for smaller gaps.

Using SDEMm, we processed hourly concentrations of CO, NO_2, O_3 and SO_2 for the conterminous United States from 2005 to 2010. These data are stored in the Davey Tree Expert Company's i-Tree Eco server to provide United States users the ability to run i-Tree Eco in any cities, counties and states in the conterminous United States. Moreover, as the AQS data is often used by epidemiologic studies (Ito et al., 2005; Eder and Yu, 2006; Liao et al., 2006; Sarnat et al. 2010), the dataset completed by this study may also help improve these analyses.

Results from this study indicate that additional experiments should be conducted in the future. Monitoring sites with a long interruption, especially O_3 monitors that are active only during the "ozone season," were excluded in this study. Due to the long gaps in these records, the SDEMm method was unable to calculate time mean concentrations for these monitors. Filling gaps with the time means at these monitoring sites will allow us to apply the developed method to those excluded monitors, which will provide improvements to i-Tree Eco's applicability. Other exclusions in this study are PM_{10} and $PM_{2.5}$, which are typically measured on a daily or longer duration for most of the year as opposed to the hourly measurements required by i-Tree Eco.

References

Atkins, D.H.F., Lee, D.S., 1995. Spatial and temporal variation of rural nitrogen dioxide concentrations across the United Kingdom, Atmospheric Environment 29, 223-239.

Beckett, K.P., Freer-Smith, P.H., Taylor, G., 1998. Urban woodlands: their role in reducing the effects of particulate pollution. Environmental Pollution 99, 347-360.

Capilla, C., 2007. Analysis of the trend and seasonal cycle of carbon monoxide concentrations in an urban area. Environmental Science and Pollution Research 14, Special Issue 1, 19 – 22.

Chen, L.-W.A., Doddridge, B.G., Dickerson, R.R., Chow, J.C., Mueller, P.K., Quinn, J. Butler, W.A., 2001. Seasonal variations in elemental carbon aerosol, carbon monoxide and sulfur dioxide: Implications for sources. Geophysical Research Letters 23, 1711-1714.

Code of Federal Regulations (CFR). 2010. Electronic Code of Federal Regulations (e-CFR), Title 40: Protection of Environment, Part 58 - Ambient Air Quality Surveillance, Appendix D to Part 58 – Network Design Criteria for Ambient Air Quality Monitoring. http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&sid= 39a729e6ac9d1dd8249bc270f56f2e5e&rgn=div9&view=text&node=40:5.0.1.1.6.7. 1.3.32&idno=40 [accessed 10 December, 2010].

Currie, B.A., Bass, B., 2008. Estimates of air pollution mitigation with green plants and green roofs using the UFORE model. Urban Ecosystems 11, 409-422.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood for incomplete data via the EM algorithm, Journal of Royal Statistic Society B39, 1-38.

Deutsch, B., Whitlow, H., Sullivan, M., Savineau, A., 2005. Re-greening Washington, DC: A green roof vision based on quantifying storm water and air quality benefits. http://www.greenroofs.org/resources/greenroofvisionfordc.pdf [accessed 12 February, 2009].

Eder, B., Yu, S., 2006. A performance evaluation of the 2004 release of Models-3 CMAQ. Atmospheric Environment 40, 4811-4824.

Engels, J.M., Diehr, P., 2003. Imputation of missing longitudinal data: a comparison of methods. Journal of Clinical Epidemiology 56, 968-976.

Gelman A., Hill, J., 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, 648 pp.

Hargreaves, P.R., Leidi, A., Grubb, H.J., Howe, M.T., Mugglestone, M.A., 2000. Local and seasonal variations in atmospheric nitrogen dioxide levels at Rothamsted, UK, and relationships with meteorological conditions. Atmospheric Environment 34, 843-853.

Hirabayashi, S., Kroll, C.N., Nowak, D.J., 2011. Component-based development and sensitivity analyses of an air pollutant dry deposition model. Environmental Modelling & Software 26, 804-816.

Hirabayashi, S., Kroll, C.N., Nowak, D.J., 2012. Development of a distributed air pollutant dry deposition modeling framework. Environmental Pollution 171, 9-17.

Imtiaz, S.A., Shah, S.L., 2008. Treatment of missing values in process data analysis. The Canadian Journal of Chemical Engineering 86, 838-858.

Ito, K., De Leon, S., Thurston, G.D., Nadas, A., Lippmann, M., 2005. Monitor-to-monitor temporal correlation of air pollution in the contiguous US. Journal of Exposure Analysis and Environmental Epidemiology 15, 172-184.

i-Tree, 2013. i-Tree - Tools for Assessing and Managing Community Forests. http://www.itreetools.org [accessed 6 June, 2013].

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

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

Legates, D.R., McCabe Jr., G.J., 1999. Evaluating the use of “goodness-of-fit” measures in hydrologic model validataion. Water Resource Research 35, 233-241.

Li, K.H., Le, N.D., Sun, L, Zidek, J.V. 1999. Spatial-temporal models for ambient hourly PM_{10} in Vancouver. Environmetrics 10, 321-338.

Liao, D., Peuquet, D.J., Duan, Y., Whitsel, E.A., Dou, J., Smith, R.L., Lin, H.-M., Chen, J.-C., Heiss, G., 2006. GIS approaches for the estimation of residential-level ambient PM concentrations. Environmental Health Perspectives 114, 1374-1380.

Little, R.J.A, Rubin, D.B., 2002. Statistical analysis with missing data, second edition. John Wiley & Sons, Hoboken, NJ, 381 pp.

Marwala, T., 2009. Computational intelligence for missing data imputation, estimation and management: Knowledge optimization techniques, Hershey, New York, 306 pp.

Moffat, A.M., Papale, D. Reichstein, M., Hollinger, D.Y., Richardson, A.D., Barr, A.G., Beckstein, C., Braswell, B.H., Churkina, G., Desai, A.R., Falge, E., Gove, J.H., Heimann, M., Hui, D., Jarvis, A.J., Kattge, J., Noormets, A., Stauch, V.J., 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agricultural and forest meteorology 147, 209-232.

Nash, J.E. Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, I, A discussion of principles. Journal of Hydrology 10, 282-290.

Nowak, D.J., Crane, D.E., 2000. The Urban Forest Effects (UFORE) Model: quantifying urban forest structure and functions. In: Hansen, M., Burk, T. (eds.) Proceedings: Integrated Tools for Natural Resources Inventories in the 21st Century. IUFRO Conference, Boise, ID. Gen. Tech. Report NC-212. US Department of Agriculture, Forest Service, North Central Research Station, St. Paul, MN, pp. 714–720.

Nowak, D.J., Crane, D.E., Stevens, J.C., 2006. Air pollution removal by urban trees and shrubs in the United States. Urban Forestry & Urban Greening 4, 115-123.

Nowak, D.J., McHale, P.J., Ibarra, M., Crane, D., Stevens, J., Luley, C., 1998. Modeling the effects of urban vegetation on air pollution. In: Gryning, S.E., Chaumerliac, N. (Eds.) Air Pollution Modeling and its Application XII, Plenum Press, New York, pp. 399-407.

Plaia, A., Bondi, A.L., 2006. Single imputation method of missing values in environmental pollution data sets. Atmospheric Environment 40, 7316-7330.

Sarnat, S.E., Klein, M., Sarnat, J.A., Flanders, W.D., Waller, L.A., Mulholland, J.A., Russell, A.G., Tolbert, P.E., 2010. An examination of exposure measurement error from air pollutant spatial variability in time-series studies. Journal of Exposure Science and Environmental Epidemiology 20, 135-146.

Schafer, J.L., 1997. Analysis of incomplete multivariate data. Monographs on Statistics and Applied Probability, 72. Chapman & Hall, London.

Seinfeld, J.H., 1986. Atmospheric Chemistry and Physics of Air Pollution, John Wiley & Sons, Hoboken, NJ, 738pp.

Serageldin, I., 2002. World poverty and hunger - the challenge for science. Science, 54-58.

Sillman, S., 1999. The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments. Atmospheric Environment 33, 1821-1845.

United States Environmental Protection Agency (US EPA). 1972. Federal air quality control regions. U.S. Environmental Protection Agency, Office of Air Programs, Office of the Assistant Commissioner for Regional Activities. Rockville, MD, 202pp.

United States Environmental Protection Agency (US EPA). 2010a. AQS Data Dictionary Version 2.25. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Information Transfer and Program Integration Division, Information Management Group, Research Triangle Park, NC, 410pp.

United States Environmental Protection Agency (US EPA). 2010b. AirData: About the AQS database. http://www.epa.gov/air/data/aqsdb.html [accessed 08 January, 2010].

United States Environmental Protection Agency (US EPA). 2013. Air Emission Sources. http://www.epa.gov/air/emissions/index.htm [accessed 18 March, 2013].

Wilcox, B.P., Rawls, W.J., Brakensiek, D.L. Wight, J.R., 1990. Predicting runoff from rangeland cathments: A comparison of two models. Water Resource Research 26, 2401-2410.

Willmott, C.J., 1981. On the validation of models. Physical Geography 2, 184-194.