Available online at www.sciencedirect.com

ScienceDirect

Atmospheric Environment 40 (2006) 7316–7330

ATMOSPHERIC ENVIRONMENT

www.elsevier.com/locate/atmosenv

Single imputation method of missing values in environmental pollution data sets

A. Plaia*, A.L. Bondì

Department of Mathematical and Statistical Sciences, University of Palermo, Viale delle Scienze, Building 13, 90128 Palermo, Italy

Received 3 May 2006; received in revised form 26 June 2006; accepted 27 June 2006

Abstract

Missing data represent a general problem in many scientific fields above all in environmental research. Several methods have been proposed in literature for handling missing data and the choice of an appropriate method depends, among others, on the missing data pattern and on the missing-data mechanism. One approach to the problem is to impute them to yield a complete data set. The goal of this paper is to propose a new single imputation method and to compare its performance to other single and multiple imputation methods known in literature. Considering a data set of PM_{10} concentration measured every 2 h by eight monitoring stations distributed over the metropolitan area of Palermo, Sicily, during 2003, simulated incomplete data have been generated, and the performance of the imputation methods have been compared on the correlation coefficient (ρ), the index of agreement (d), the root mean square deviation (RMSD) and the mean absolute deviation (MAD). All the performance indicators agree to evaluate the proposed method as the best among the ones compared, independently on the gap length and on the number of stations with missing data.

Keywords: Missing at random; Time series; Environmental pollutant; Single imputation

1. Introduction

Missing data is a very frequent problem in many scientific fields above all in environmental research (Xia et al., 1999), usually due to faults in data acquisition (Latini and Passerini, 2004). The impact of the missing data on the result of statistical analysis depends on the mechanism that made the data to be missing and on the way the data analyst deals with them. This making the aim of scientist to find suitable criteria to replace data holes with appropriate values. The choice of an appropriate method for handling missing data depends, among others, on the missing data pattern and on the missing-data mechanism. The standard classification of missing data mechanism (Little and Rubin, 1987; Schafer, 1997), considers data: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). Considering the complete data $Y = Y_j$ and the missing data indicator $M = m_j$, where $m_j = 1$ if Y_j is missing and $m_j = 0$ otherwise, the missing data mechanism is characterized by the conditional distribution of M given Y, say $f(M|Y, \phi)$, where ϕ denotes unknown parameters. Let Y_{obs} and Y_{miss}

*Corresponding author. Tel.: +39 091 6626244; fax: +39 091 485726.

E-mail addresses: plaia@unipa.it (A. Plaia), bondi@dssm.unipa.it (A.L. Bondi).

1352-2310/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2006.06.040

denote the observed and the missing components of Y, respectively. If

$$f(M|Y,\phi)=f(M|\phi) \tag{1}$$

for all Y, the data are called MCAR; if

$$f(M|Y,\phi)=f(M|Y_{\text{obs}},\phi) \tag{2}$$

for all Y_{miss}, ϕ, the missing data mechanism is MAR; but if the distribution of them depends on Y_{miss} the mechanism is called NMAR. Usually, the data mechanism of air quality data is at least MAR, as, being due to the monitoring site down, the probability that a value is missing does not depend on the missing value (as it would be if the monitoring site could not measure values below a given threshold). In presence of missing data, a possibility is to discard units whose information is incomplete, considering a *listwise deletion*, that involves the complete discard of the units with missing data, or a *pairwise deletion*, where units are excluded from any calculation involving variables with missing data. On the other hand, statistical methods are available that take the missing data into account at the time of analysis. These methods include likelihood-based approaches such as generalized linear models and the expectation-maximization (E-M) algorithm when data are MAR. A different approach is to impute the missing values so that the resulting data set is complete. This third possibility is to be followed if the data set will be used for many different types of analysis by a number of researchers. Methods available for creating complete data matrices can be divided into two main categories: *single imputation* and *multiple imputation* methods. Single imputation methods fill in one value for each missing one; they have many appealing features, because standard complete-data methods can be applied directly and because imputation need to be carried out only once. Multiple imputation methods generate multiple simulated values for each missing value, in order to reflect the uncertainty attached to missing data (Schafer, 1997). Generally a multiple imputation method requires a full specification of the distributional form of Y in order to derive the conditional distribution of the missing data given the observed data. Besides, multiple imputation is generally used to estimate some parameter θ of the distribution of the Y.

The aim of this paper is to propose a new single imputation method that, considering the particular structure of the data set, creates a "complete" data set that can be analyzed by any researcher on different occasions and using different techniques.

2. Data and methods

2.1. Data

In Bondì and Plaia (2005) the authors analyzed the space–time variability of PM_{10} concentration via meteorological variables. In the cited paper, as well as in almost all the researches about the effects of PM_{10} on human health, daily mean concentrations are considered. But a daily mean cannot be computed if more than 25% of bi-hour (or hour) daily data is missing. Therefore, in the present paper we will consider the bi-hour data set in order to 'complete' it by imputing missing values. In particular, we will consider PM_{10} concentration measured every 2 h by eight monitoring stations distributed over the metropolitan area of Palermo, Sicily, during 2003 (but whatever pollutant values could be considered, assuming that its concentration has been measured over time in different sites), that is the data matrix consists of 4380 cases (measures over time) and eight columns (monitoring sites). These longitudinal data show a multilevel structure with monitoring sites as second-level units and single measures as first level units. Table 1 shows the data set arranged in a four way data table, whose entries are: the week of the year (W), assuming values between 1 and 53, the day of the week (W-D), assuming values between 1 and 7, the hour of the day (H), assuming values between 2 and 24—step2, and the monitoring site (St1–St8). The data set shows a certain percentage of missing data (see Table 2), going from a 3.4% of Station 3 to a 13% of Station 2. Categorizing the gap length (l) in PM_{10} air concentration data into four classes, namely 1 observation gaps (bi-hour), from 1 to 3 observation gaps, from 3 to 12 observation gaps and more than 12 observation gaps, Table 2 shows the distribution of missing values according to this categorization: averaging over the eight monitoring sites about 74% of missing data belongs to gaps of missing values with length 1, 18.7% to gaps with a length between 1 and 3, 5.1% to gaps with a length between 3 and 12 and only 2.2% to gaps longer than 12 (i.e. longer than one day). Fig. 1 shows the frequency distribution of the gap length for a specific monitoring station (Station 7).

Table 1
Data set structure

W^a	W-D	H	St1	St2	St3	St4	St5	St6	St7	St8
1	3	2	23.32	12.47	108.02	30.46	NA	4.29	45.05	50.91
1	3	4	34.68	27.09	28.17	9.34	20.10	26.66	NA	26.36
1	3	6	26.85	11.10	34.70	24.17	29.76	21.94	NA	44.46
1	3	8	19.80	24.62	31.51	27.23	34.96	26.20	NA	45.89
1	3	10	24.38	14.84	23.91	18.50	15.21	20.10	20.94	NA
1	3	12	22.47	16.46	30.72	38.54	35.87	16.76	15.20	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

[a] W = week of the year; W-D = day of the week; H = hour of the day; St1−St8 = monitoring sites.

Table 2
Missing data percentage and descriptive statistics by stations

	% Missing	Gap lengths (l)				Raw data	
		$l=1$	$1<l\leqslant 3$	$3<l\geqslant 12$	$l>12$	Mean	Sd
St1	10.6	75.2	20.3	2.8	1.7	38.9	23.2
St2	13.0	72.4	19.2	6.2	2.2	28.9	18.1
St3	3.4	69.3	16.2	12.9	1.6	40.3	22.1
St4	4.8	81.5	15.9	2.6	–	45.8	28.3
St5	8.4	66.7	25.4	4.4	3.5	44.4	25.1
St6	10.2	69.5	25.4	3.5	1.6	34.3	20.9
St7	6.6	80.2	14.9	3.7	1.2	30.8	18.4
St8	5.9	77.5	12.4	4.5	5.6	41.0	22.5
Mean	7.8	74.0	18.7	5.1	2.2		

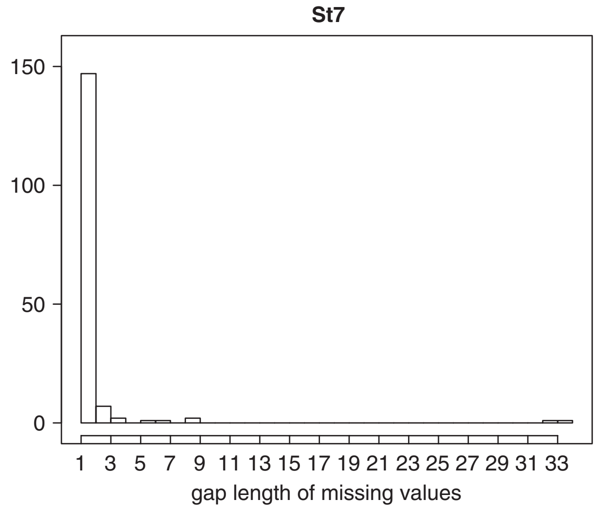

Fig. 1. Distribution of gap length in a monitoring station: St7.

2.2. Imputation methods

A number of methods are available in literature to impute missing air quality data. Among the single imputation methods for longitudinal data we can distinguish methods based on the information on the same subject (e.g. *last observation carried forward*, *next observation carried backward*, *last & next* (Engels and Diehr, 2003)), methods that borrow information from other subjects (*row mean/median*, referring to Table 1 borrow information from others stations) and methods that use both pieces of information (e.g. *conditional mean imputation*, *hot-deck imputation*). Considering the multilevel structure of our data set (with monitoring sites as second level units and single measures as first level units), we suppose it is important to consider both the row and column information (in Table 1) to impute a missing value (that is to take advantage of both the spatial and temporal correlation among data). Actually, differently from what can happen for a climatological data set (that can shows the same structure as our data set) here it is a main point to consider site specific effects, for example week, week-day or day-hour site specific effects, that is to distinguish, for example, between an hour mean (averaged over the eight monitoring sites) and a site specific hour mean.

Referring to Fig. 3, which shows the week-day site specific mean, computed as the average of all the values measured in a site on each week-day, together with the overall week-day mean, computed as shown in Table 3, we can see the importance of considering a week-day site specific effect, computed as the difference between a week-day site mean and the overall week-day mean. The new imputation method we will propose in the next section considers these particular characteristics of the data.

2.2.1. SDEM method

The new single imputation method proposed in this paper uses space–time information on PM_{10}

Table 3
Week-day mean $\bar{x}_{s.d.}$

W-D	Stations								$(\sum_{s=1}^{S}\bar{x}_{s.d.}/S)$
	St1	St2	St3	St4	St5	St6	St7	St8	
1	36.64	26.67	38.68	44.85	42.85	34.67	30.30	38.58	36.66
2	40.48	26.62	41.19	48.55	43.97	36.09	30.25	42.00	38.64
3	42.29	30.27	42.29	45.06	44.98	35.54	31.71	43.23	39.42
4	42.19	31.19	42.49	47.83	47.74	34.71	32.40	45.83	40.55
5	41.99	32.88	43.98	53.19	49.56	37.22	33.33	44.89	42.13
6	39.42	27.90	40.49	45.10	43.78	32.31	30.23	40.27	37.44
7	32.64	23.96	34.04	34.91	35.09	28.55	25.25	32.39	30.85

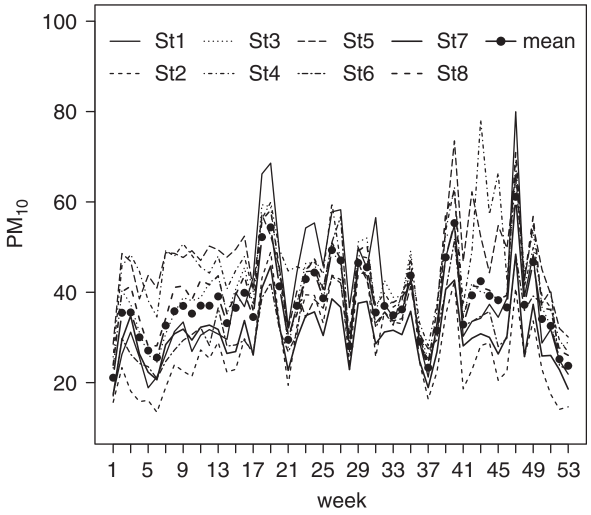

Fig. 2. Week site specific and mean effect in eight monitoring sites.

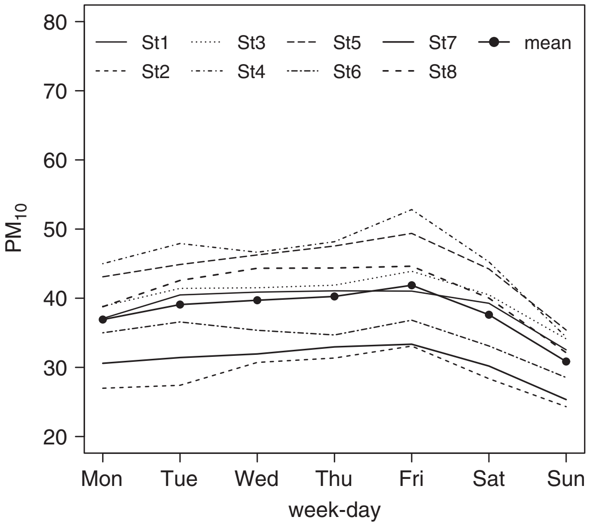

Fig. 3. Week-day site specific and mean effect in eight monitoring sites.

level concentration in eight monitoring sites. It can be defined as follow. Denoting the generic element of the data set (Table 1) as x_{swdh}, where s refers to the monitoring site ($s = 1, 2, \ldots, S$), w to the week ($w = 1, 2, \ldots, 53$), d to the week-day ($d = 1, 2, \ldots, 7$) and h to the hour ($h = 2, 4, \ldots, 24$) we consider the:

(1) Week site mean matrix: it is a 53×8 matrix whose generic element $\bar{x}_{sw\cdot\cdot}$ is the mean of the values observed on week w in site s (each column of the matrix corresponds to a line plot in Fig. 2). The difference between $\bar{x}_{sw\cdot\cdot}$ and its marginal row mean, $\sum_{s=1}^{S}(\bar{x}_{sw\cdot\cdot}/S)$ (the dot plot in Fig. 2), can be considered the specific week effect of site s.
(2) Week-day site mean matrix: as shown in Table 3 and in Fig. 3 (whose line plots correspond to the columns of Table 3), it is a 7×8 matrix whose generic element $\bar{x}_{s\cdot d\cdot}$ is the mean of the values observed on week-day d in site s. The specific week-day effect of site s is obtained computing the difference between $\bar{x}_{s\cdot d\cdot}$ and its marginal row mean $\sum_{s=1}^{S}(\bar{x}_{s\cdot d\cdot}/S)$ (the dot plot in Fig. 3).
(3) Hour site mean matrix: it is a 12×8 matrix whose generic element $\bar{x}_{s\cdot\cdot h}$ is the mean of the values observed on day hour h ($h = 2, 4, \ldots, 24$) in site s (see Fig. 4). The specific day hour effect of site s is obtained computing the difference between $\bar{x}_{s\cdot\cdot h}$ and its marginal row mean $\sum_{s=1}^{S}(\bar{x}_{s\cdot\cdot h}/S)$.

The method we propose, called the Site-Dependent Effect method (SDEM), considers explicitly a week effect, a day effect and an hour effect (all site-dependent), assuming their additivity, and estimates

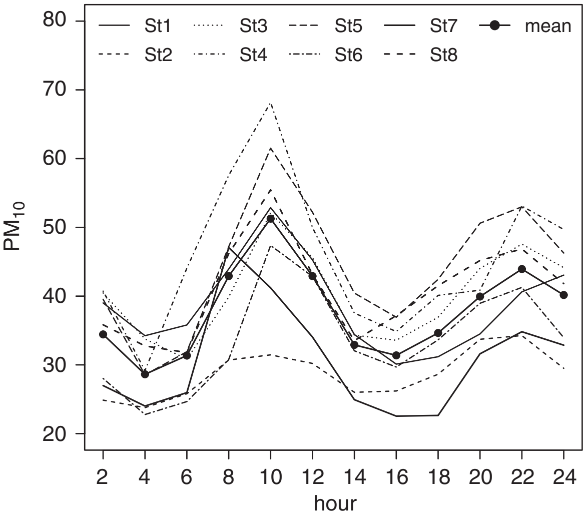

Fig. 4. Hour site specific and mean effect in eight monitoring sites.

a missing value as

$$\hat{x}_{swdh} = \bar{x}_{\cdot wdh} + \frac{1}{2}\left(\bar{x}_{sw\cdot\cdot} - \sum_{s=1}^{S}\frac{\bar{x}_{sw\cdot\cdot}}{S}\right) + \frac{1}{2}\left(\bar{x}_{s\cdot d\cdot} - \sum_{s=1}^{S}\frac{\bar{x}_{s\cdot d\cdot}}{S}\right) + \frac{1}{2}\left(\bar{x}_{s\cdot\cdot h} - \sum_{s=1}^{S}\frac{\bar{x}_{s\cdot\cdot h}}{S}\right). \qquad (3)$$

Figs. 2–4 justify the use of a site specific week, week-day and hour effect, respectively, showing, for each monitoring site, the difference between a time mean value and a site specific time mean value.

The performance of this method will be compared to other single imputation methods in literature, and in particular to the *hour mean* method (Li et al., 1999), to *last & next* method (Engels and Diehr, 2003), to the *row-mean* method (Engels and Diehr, 2003) and to a *model-based multiple imputation* (MI) method (Schafer, 1997).

2.2.2. *Hour mean method*

This method use hourly information known on PM_{10} concentration levels in the same monitoring site. According to (Li et al., 1999) this single imputation method fill in missing hourly observations computing the mean of all known hourly observations on the same monitoring site at the same hour over the whole year:

$$\widehat{x}_{swdh} = \bar{x}_{s\cdot\cdot h}, \qquad (4)$$

where $s = 1, 2, \ldots, S$, $h = 2, 4, \ldots, 24$.

2.2.3. *Row mean method*

This method use hourly information known on PM_{10} concentration levels in the other monitoring sites (Engels and Diehr, 2003). This single imputation method fill in a missing observation computing the mean of all known observations on the same row of the data matrix, that is computing the mean of the values registered at the same time by the other monitoring sites:

$$\widehat{x}_{swdh} = \bar{x}_{\cdot wdh}, \qquad (5)$$

where $w = 1, 2, \ldots, 52$, $d = 1, 2, \ldots, 7$, $h = 2, 4, \ldots, 24$.

2.2.4. *Last & next method*

This method (Engels and Diehr, 2003) considers information on the same subject, here monitoring site, assigning the average of the station's last known and next known observations to the missing value of PM_{10}:

$$\widehat{x}_{swdh} = \frac{x_{swd(h-1)} + x_{swd(h+1)}}{2}, \qquad (6)$$

where $s = 1, 2, \ldots, S$, $w = 1, 2, \ldots, 52$, $d = 1, 2, \ldots, 7$, $h = 2, 4, \ldots, 24$.

2.2.5. *Model-based multiple imputation method* (MI)

A single imputed value will never be able to represent all the uncertainty about which value to implement. Multiple imputation tries to solve this problem by imputing several values instead of only one, and then analyzing each data set with standard methods. (Rubin, 1996) described multiple imputation as a three-step process. First, sets of plausible values for missing observations are created that reflect uncertainty about data. Each of these sets of plausible values can be used to "fill in" the missing values and create a "completed" data set. Second, each of these data sets can be analyzed using complete-data methods. Finally, the results are combined, which allows the uncertainty regarding the imputation to be taken into account. We will use a model-based multiple imputation that improves results by introducing uncertainty into the model and using that uncertainty to model the natural variability of incomplete records, solving the problem of underestimation of the error variance (Schafer, 1997). In this paper we use multiple imputation by chained equations (MICE). This method generates multiple imputations for incomplete multivariate data by Gibbs sampling. The algorithm imputes an incomplete column (the target column) by generating appropriate imputation values given other columns in the data. Each incomplete column must act as a target column, and has its own specific set of predictors. The default predictor set, here used, consists of all other columns in the data. For predictors that are incomplete themselves, the most recently generated imputations are used to complete the predictors prior to imputation of the target column. Working to site specific level concentration of PM_{10}, numeric data column, the elementary imputation method predictive mean matching is used. Although computationally attractive, the chained equation approach implemented in MICE requires assumptions about the existence of the multivariate posterior distribution used for sampling, however, it is not always certain that such a distribution exists. MICE is a free library distributed for *R* (Van Buuren and Oudshoorn, 2005), a system for statistical computation and graphics, whose language and interface is very similar to S-Plus, and is freely available for download from the URL: http://www.multiple-imputation.com web site or URL: ⟨http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm⟩.

3. Performance indicators and missing data simulation

It is usually difficult to determine and compare the accuracy of different imputation methods because, being unable to retrieve the missing data, a method must be designed to create a data set that mimics real life data and missing data patterns: as a matter of fact, the imputation performance does not depend only on the amount of missing data but on the characteristics of the missing data mechanism. Some studies on imputation methods use real data sets and simulated missing data patterns by deleting values. Here, the true value is known, but the problem is to reproduce the missingness pattern. Otherwise, imputation can be performed on existing data sets with missing data, but the accuracy of the results cannot be determined. In this paper we decided to follow the first approach trying to reproduce, as better as possible, the missing pattern.

3.1. Performance indicators

Four performance indicators have been considered to assess the goodness of imputation. Denoting with O_i the ith observed data point, with $\overline{O}$ the average of observed data, with P_i the ith imputed data point, with $\overline{P}$ the average of imputed data, with σ_O the standard deviation of the observed data and σ_P the standard deviation of the imputed data, and finally with N the number of imputations (that is the number of missing data), we use (Junninen et al., 2004):

(1) *the coefficient of correlation* (ρ) *between observed and imputed*:

$$\rho = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} [(P_i - \overline{P})(O_i - \overline{O})]}{\sigma_P \sigma_O} \right], \tag{7}$$

(2) *the index of agreement* (d)

$$d = 1 - \left[\frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} (|P_i - \overline{O}| + |O_i - \overline{O}|)^2} \right], \tag{8}$$

Table 4
Setting of missing data simulation and resulting missing data statistics in the simulated patterns

Pattern type	Simulation settings				Resulting statistics				% Missing
	Columnwise area		Rowwise area		% of missing data in the gaps				
	Min	Max	Min	Max	$l = 1$	$1 < l \leqslant 3$	$3 < l \leqslant 12$	$l > 12$	
Expected missing data percentage: 5%									
A	1	72	1	4	11	25	36	28	5.7
B	1	72	1	8	10	22	30	38	5
Expected missing data percentage: 15%									
C	1	120	1	4	8	20	28	44	15.3
D	1	120	1	8	9	22	34	35	14.5

l—length of gap in time (columnwise area).

(3) *the root mean square deviation* (RMSD)

$$\mathrm{RMSD} = \left(\frac{1}{N}\sum_{i=1}^{N}[O_i - P_i]^2\right)^{1/2}, \tag{9}$$

(4) *the mean absolute deviation* (MAD)

$$\mathrm{MAD} = \frac{1}{N}\sum_{i=1}^{N}|(O_i - P_i)|. \tag{10}$$

Indices (8)–(10) with respect to the coefficient of correlation (7), are related to the sizes of the discrepancies between predicted and observed values. For all the indicators the standard deviation σ have been computed over the 100 matrix M of missing data patterns, as will be explained in Section 3.2.

3.2. Missing data simulation

In order to evaluate the performance of the imputation methods, simulated incomplete data have been generated, then the methods have been applied and the performance indicators computed. As already explained, the imputation performance depends both on the amount of missing data and on the missing data pattern. Fig. 1 shows a typical frequency distribution of the gap length (of a whole year) in a monitoring station. As it reveals, most of the sequences of missing values are very short (from 1 to 10 values long) with only 1 or 2 gaps longer

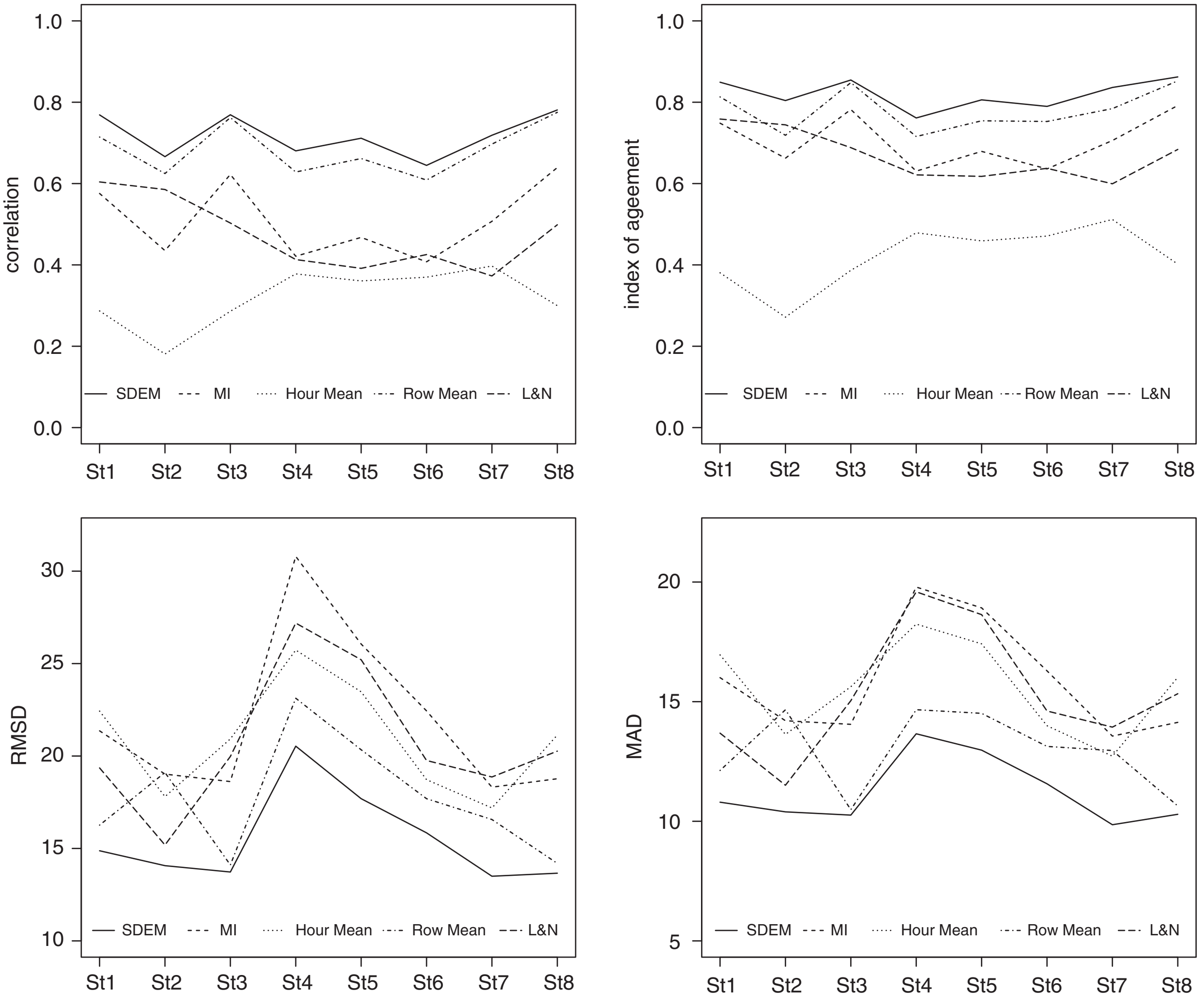

Fig. 5. Performance indicators for the five imputation methods: 5% missing and up to four missing values per row.

than one day (that is 12 consecutive missing values). In order to reproduce the actual pattern of missing data, a mixture of two distributions has been considered to randomly generate missing data indicator matrices M (differently from Section 1, here M is a matrix, having the same dimension as our data set, i.e. 4380×8) that applied to the observed data set create "artificially" missing data (actually real values are known); this allows to compute the value of the performance indicators (7)–(10) to assess the goodness of the imputation methods.

Referring to Table 1 we will consider four different missing data patterns that differ for the total percentage of missing data in the table, and for the maximum number of missing values per row. Two different total amount of missing data have been considered: about 5% with gaps not longer than 72, and about 15% with gaps up to 120 consecutive values long. Two maximum number of missing values per row have been considered, 4 and 8. For each of the four missing data pattern 100 missing data indicator matrices M have been generated (and this allows to compute the performance indicator standard deviations as already explained) according to the following procedure:

- the gap length is drawn from a mixture of two distributions, an exponential of parameter $\lambda = 0.5$ (that produces the short gaps) and a Uniform

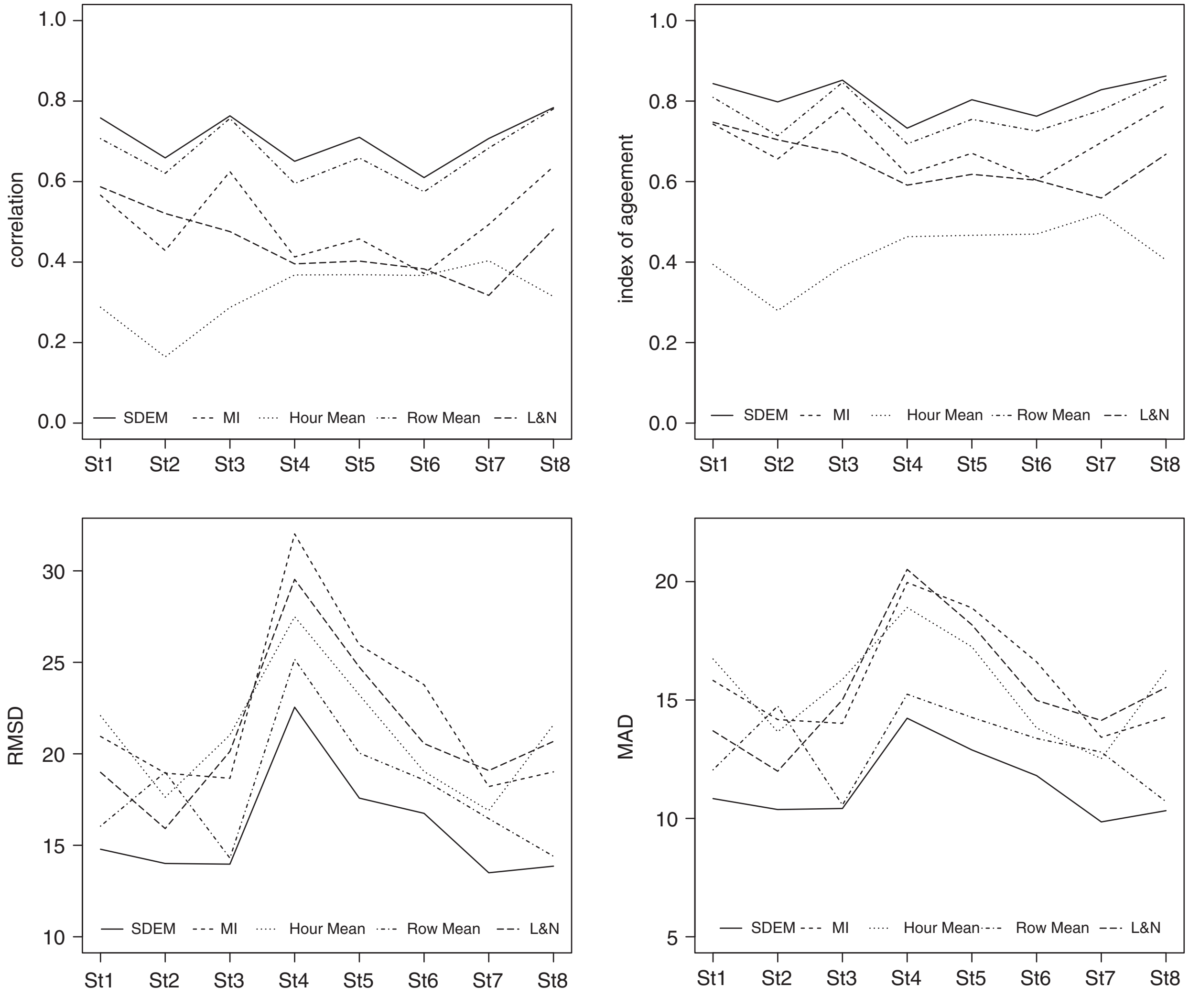

Fig. 6. Performance indicators for the five imputation methods: 5% missing and up to eight missing values per row.

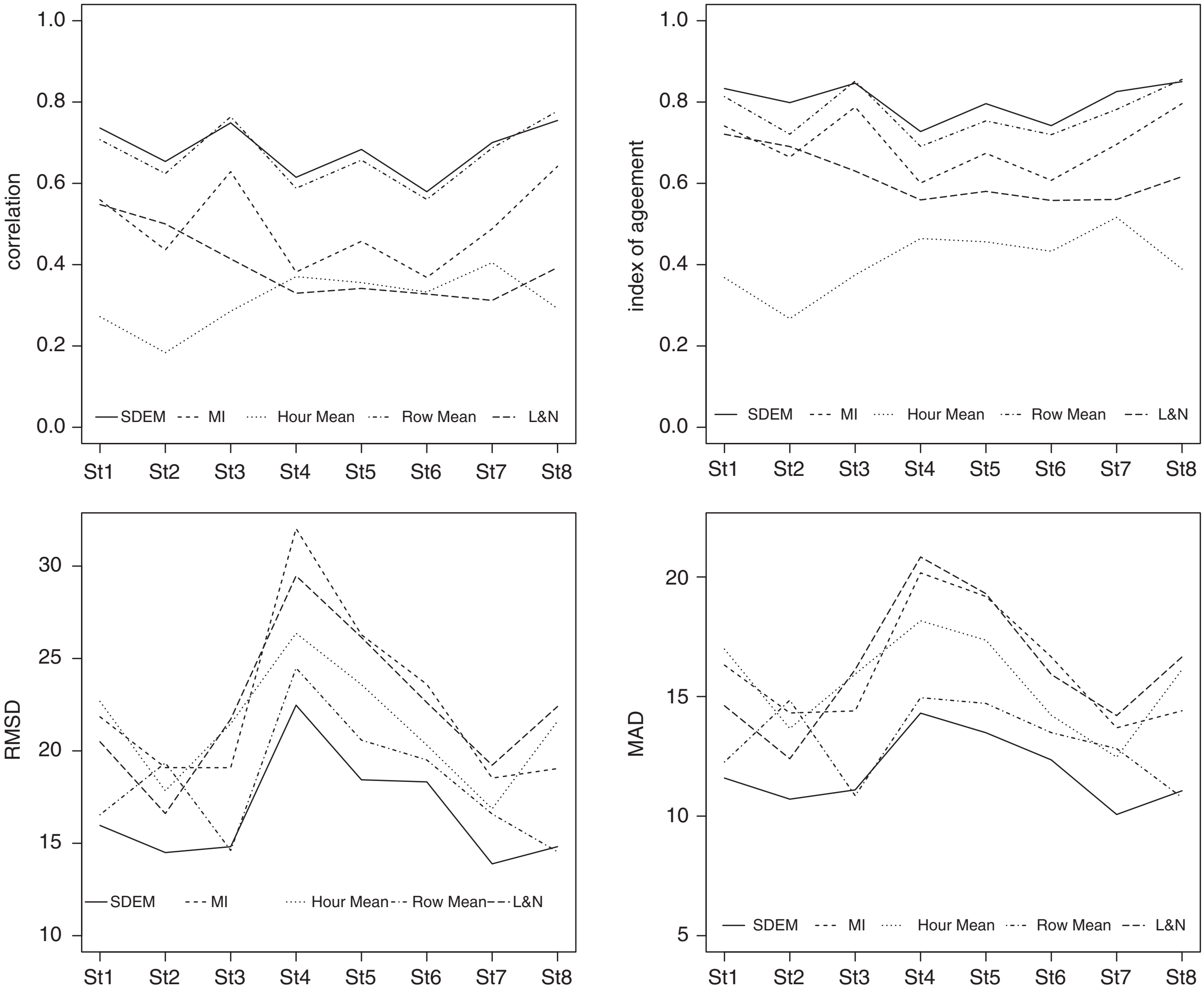

Fig. 7. Performance indicators for the five imputation methods: 15% missing and up to four missing values per row.

with parameters (20, 72) and (40, 120) for the 5% and 15% missing data patterns respectively (to produce the long gaps);

- the starting point of each gap is independently drawn from a Uniform (1, 4380);
- the number of missing data per row is drawn from a Uniform (0, 4) or a Uniform (0, 8).

Table 4 shows the setting of missing data simulations (on the left) and the resulting missing data statistics in the simulated missing data patterns (on the right).

4. Results and discussion

All the analysis, together with the generation of the 400 missing data indicator matrices M, have been carried out using the free software R (R Development Core Team, 2000–2005). Figs. 5–8 show the values of the performance indicators illustrated in Section 3.1 computed over the five imputation methods and the four missing data patterns (actually the average over the 100 indicator matrices M, with standard deviations shown in Table 5).

4.1. 5% missing and up to four missing values per row

Fig. 5 shows and compares the performance indicator values gained by the five imputation methods with an expected missing data percentage of 5 and up to four missing values per row. This is the simplest missing data pattern considered, as it consists in a 5.7% of actual missing

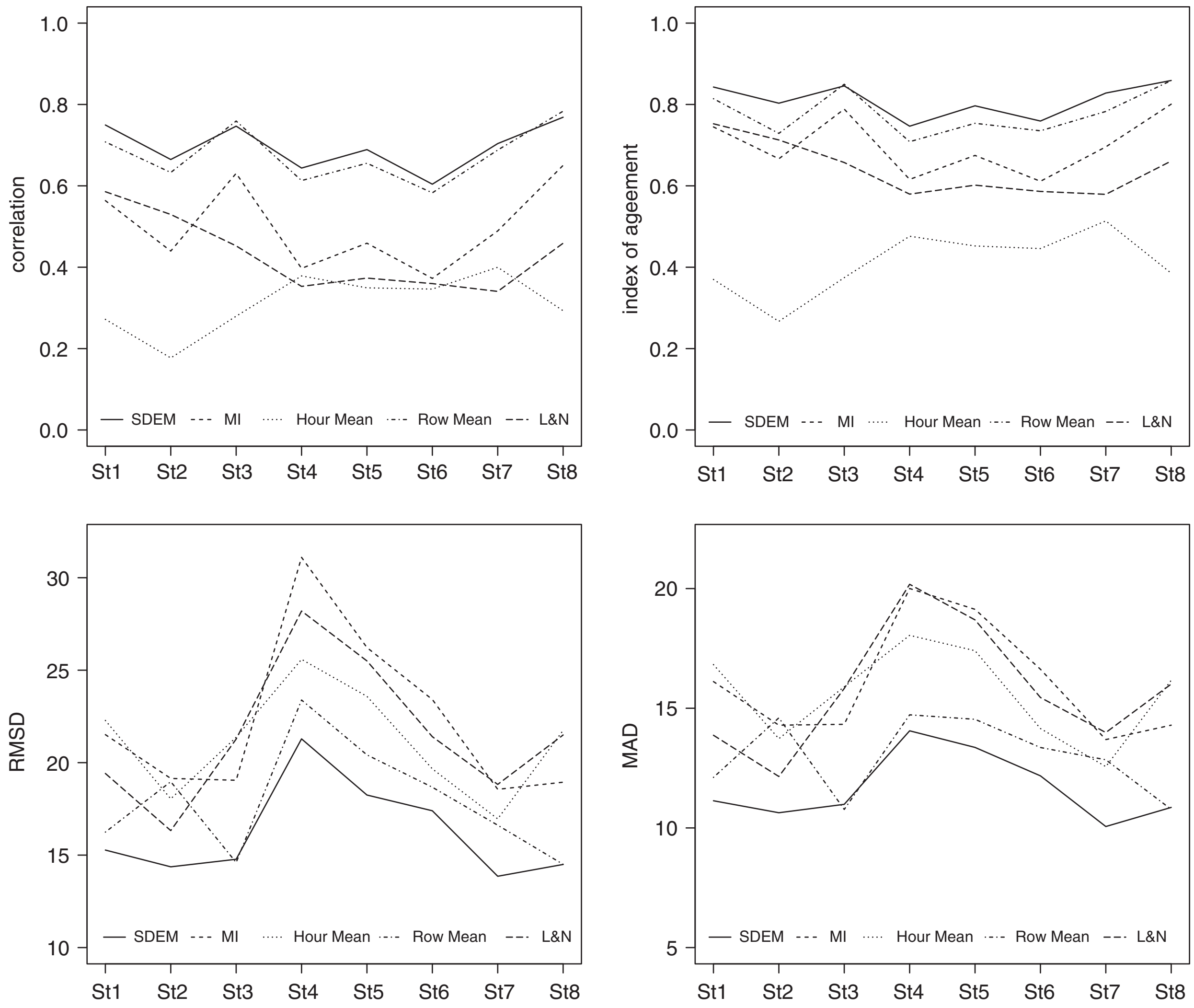

Fig. 8. Performance indicators for the five imputation methods: 15% missing and up to eight missing values per row.

data with frequency distribution of missing data per gap length shown in Table 4 (line A, right). All the performance indicators agree to assess SDEM as the best method among the ones compared (highest correlation and index of agreement, and lowest RMSD and MAD). Moreover SDEM shows the lowest variability in the performance among stations. *Hour-mean* performs very badly according both to the correlation and to the index of agreement, while its behavior is comparable to MI and *L&N* according to MAD and RMSD. *Row-mean* shows to perform almost comparably to SDEM according to correlation and index of agreement, but worse according to RMSD and MAD, and moreover displays a higher variability among sites.

Table 5
Performance indicator standard deviations

Method	σ_ρ	σ_d	σ_{RMSD}	σ_{MAD}
Last & next	0.11	0.08	3.48	2.10
Hour-mean	0.06	0.05	2.90	1.42
Row-mean	0.07	0.05	2.47	1.14
MI	0.07	0.05	2.41	1.11
SDEM	0.07	0.05	2.45	1.05

4.2. 5% missing and up to eight missing values per row

Fig. 6 shows the results gained by the five imputation methods according to the four performance indicators for the "5% missing and up to eight missing values per row": this pattern actually shows an average of 5% of missing values and a frequency distribution of missing values per gap length shown in Table 4 (line B, right). The behavior of the five imputation methods is similar to the one shown with the previous missing data pattern, and again SDEM shows the best performance with the lowest variability among stations. A comparison of each plot in Fig. 6 with the corresponding in Fig. 5 shows that the number of missing values per row does not seem to affect the method performance.

4.3. 15% missing and up to four missing values per row

Fig. 7 describes the average results gained with the third missing data pattern (whose average statistics are shown in Table 4—line C) which differs from the previous two because now the mean percentage of missing increases to 15% and the maximum gap length reach 120 consecutive values. The behavior of the imputation methods does not seem to be affected by this increase in the gap length except for *L&N* whose behavior, as was to be

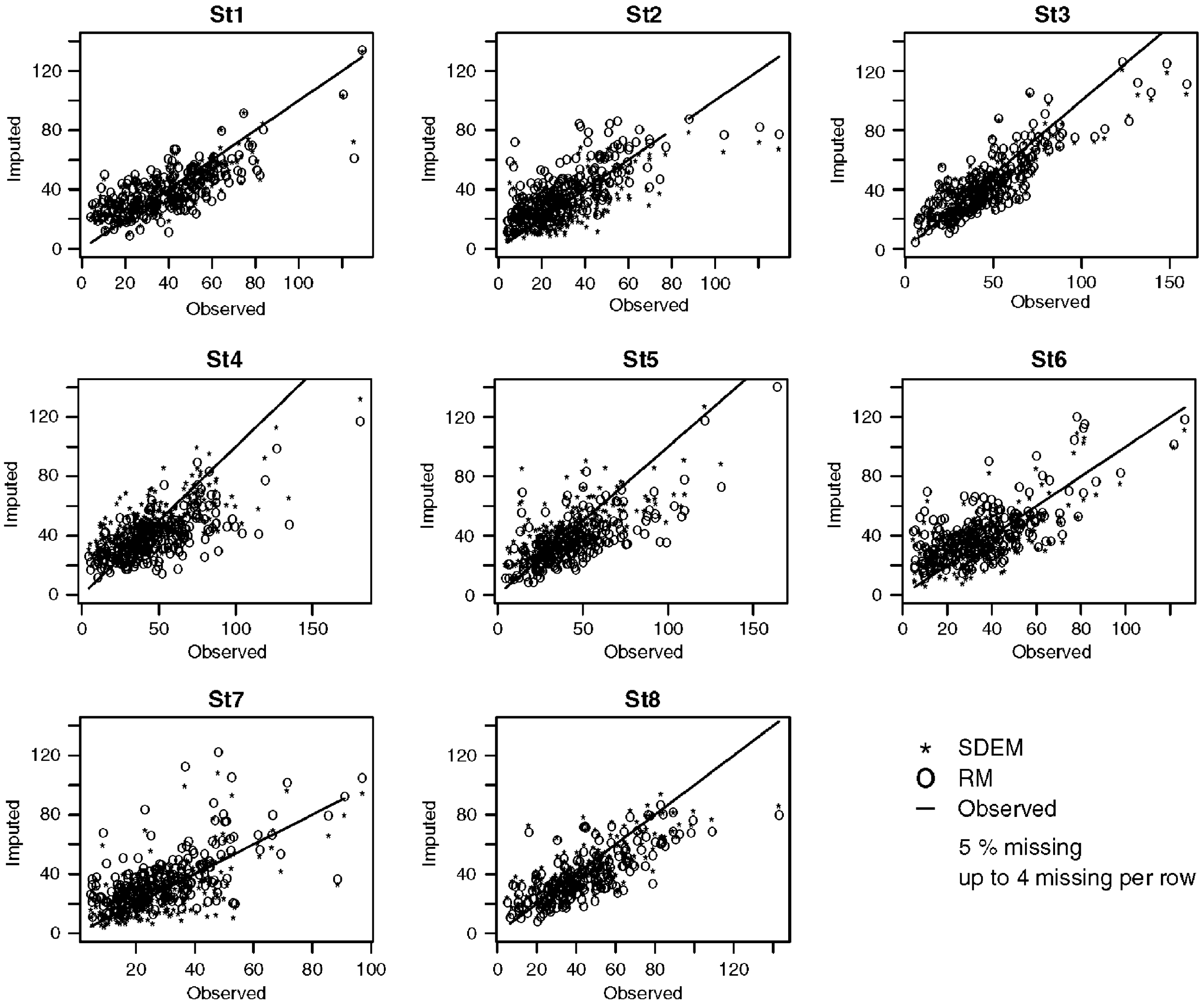

Fig. 9. Observed versus imputed data (SDEM and RM) 5% missing and up to four missing values per row.

expected, gets worse: the worsening can be noticed comparing each of the four plot with the corresponding plot in Fig. 5.

4.4. 15% missing and up to eight missing values per row

Fig. 8 describes the average results gained with the fourth missing data pattern (whose average statistics are shown in Table 4—line D). A comparison of each plot in Fig. 8 with the corresponding in Fig. 7 shows that the number of missing values per row does not seem to affect the method performance, while a comparison of each plot with the corresponding in Fig. 6 shows how *L&N* performs worse increasing the gap length.

4.5. Summary results

Figs. 5–8 shows SDEM as the best method among the ones compared, with only *row-mean* with a behavior comparable to it. Figs. 9–12 show the scatter plots of imputed against observed values, using this two best methods, *row-mean* and SDEM, for the four missing data patterns (a single missing data indicator matrix M has been considered for

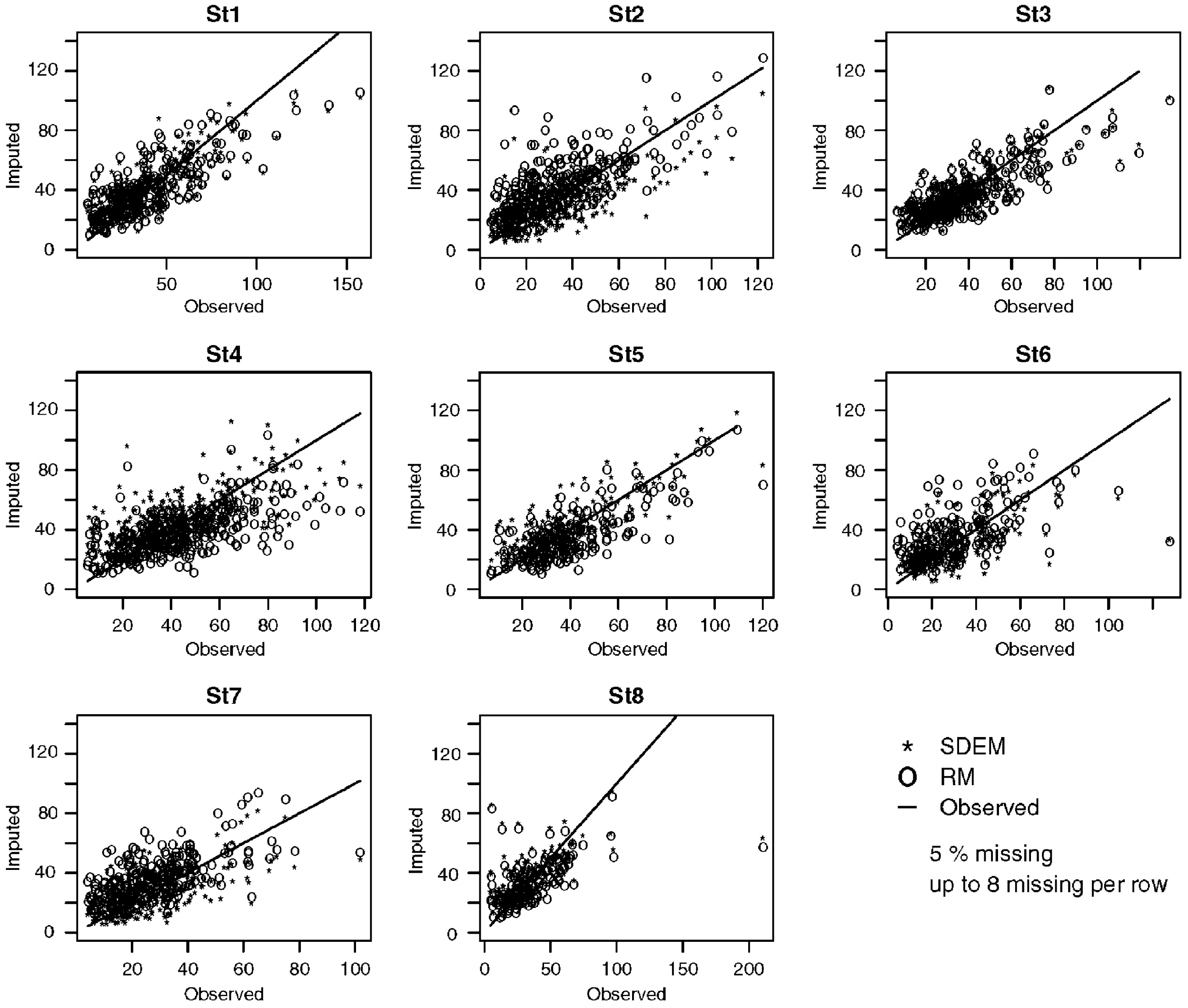

Fig. 10. Observed versus imputed data (SDEM and RM) 5% missing and up to eight missing values per row.

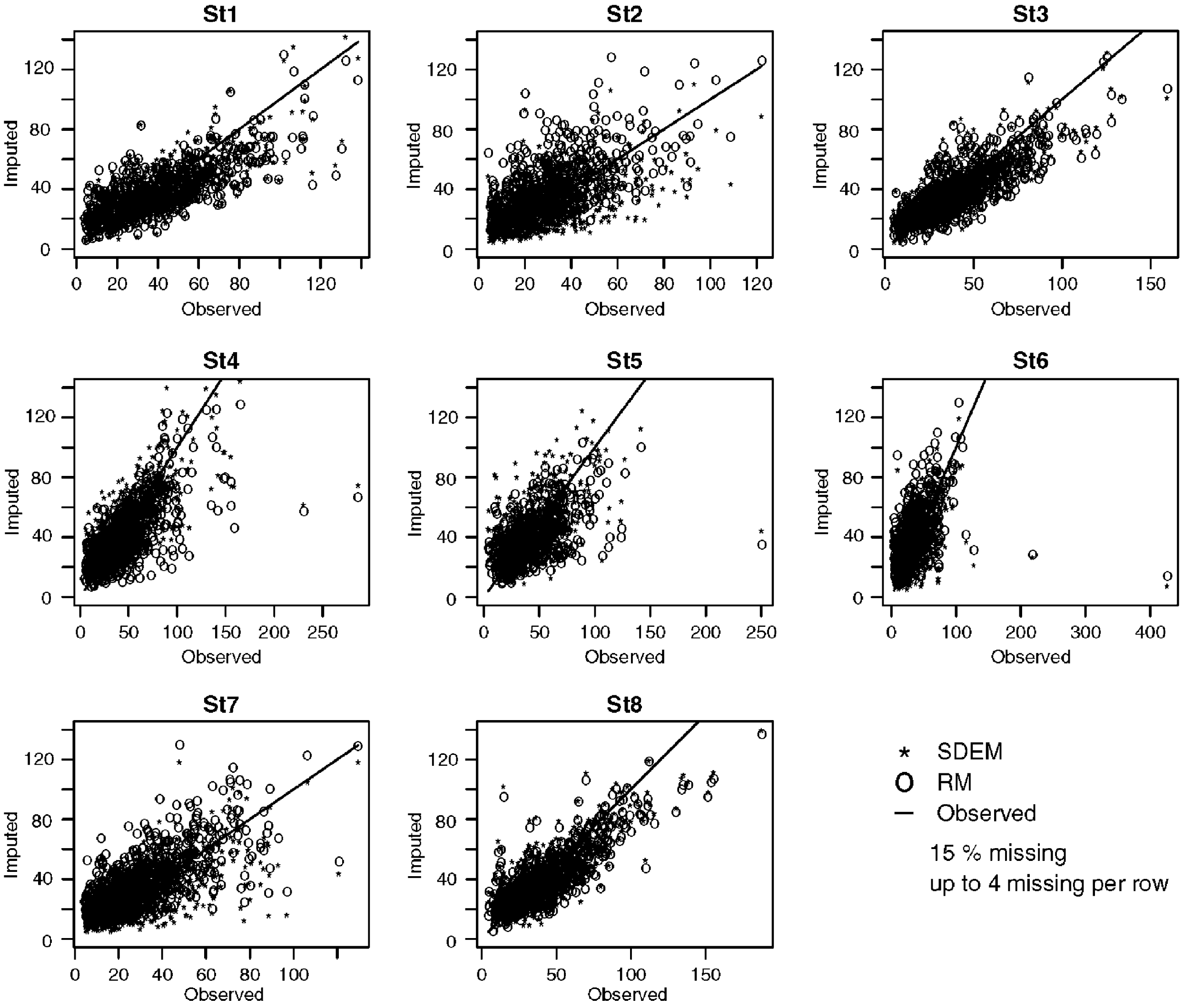

Fig. 11. Observed versus imputed data (SDEM and RM) 15% missing and up to four missing values per row.

each pattern). How it is possible to notice, *row-mean* method does not have an homogeneous behavior among stations, tending sometimes to overestimate missing values (see for example St2 and St7): this is due to St2 and St7 actually having values below the station averages. For these sites the week, week-day and hour effects in SDEM (3) correct the over-estimation given by the *row-mean*.

5. Conclusions

In a large longitudinal study where many analysis will be performed, it is important to have a complete (imputed) data set. If time information, like hour of the day, day of the week, week of the year (or month) are available, it is important to consider conditional mean imputation methods. But if the data shows a multilevel structure (as in our case) both the time dependent information and the site dependent ones are to be considered. The new imputation method proposed in this paper tries to make the best of all the information, considering, on one hand time effects averaged over the monitoring sites, on the other a site specific time effect. A combined strategy that imputes missing values considering both space and time information seems to have the best performance compared to single imputation methods, i.e. *hour mean*, *last & next*, *row-mean* methods and to multiple imputation method, i.e. MI. All the performance indicators

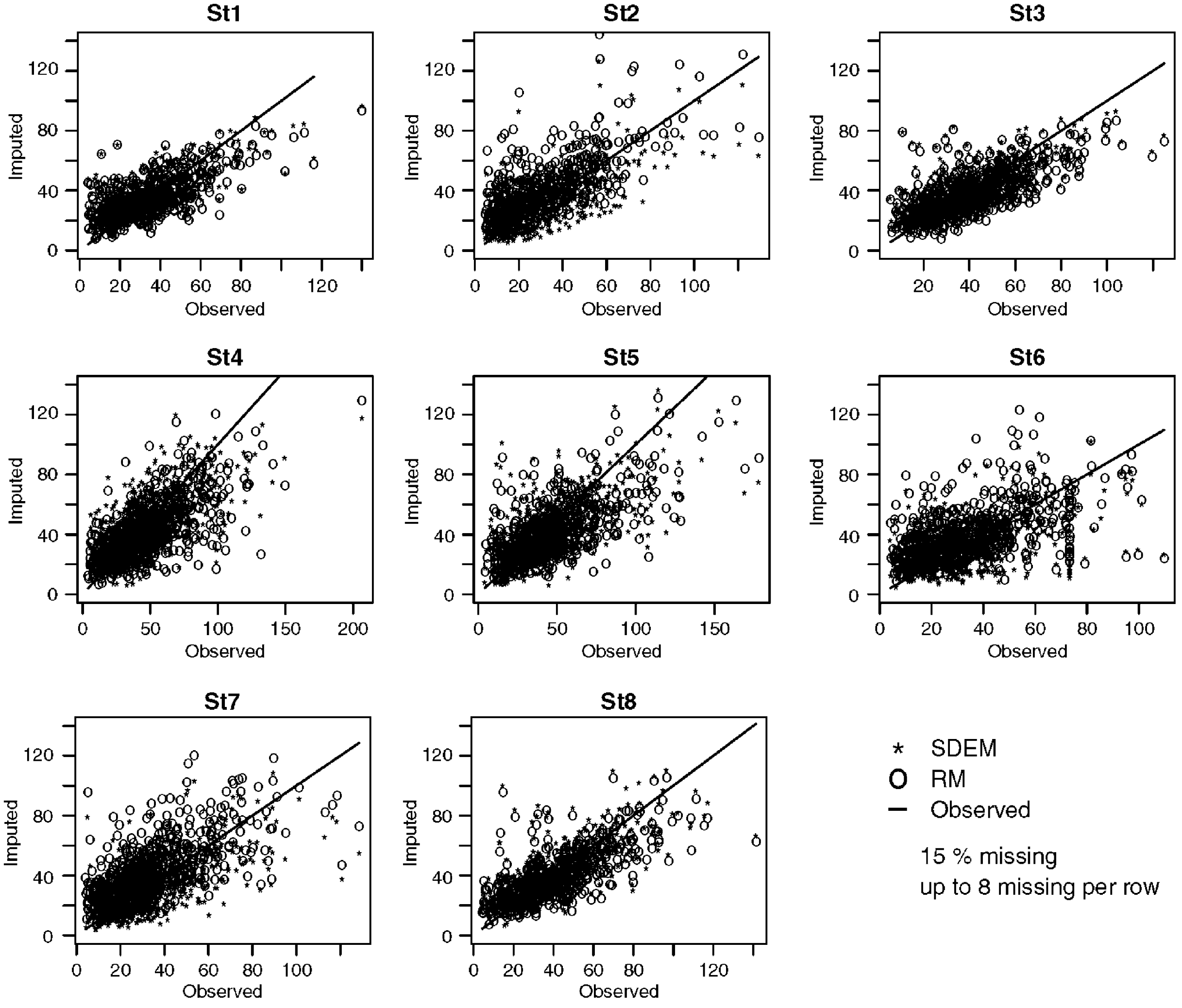

Fig. 12. Observed versus imputed data (SDEM and RM) 15% missing and up to eight missing values per row.

agree to evaluate SDEM as the best method among the ones compared in this paper, and independently on the gap length and on the number of stations with missing data. It is worthwhile noting that, even if, usually, air pollution data sets are multivariate, consisting in a set of pollutants registered over time, the proposed method does not use more than a variable at a time, and, despite of we tested it using PM_{10} data, it can be applied to whatever data set with the same structure.

Acknowledgments

This work is partially funded by the 2004 MIUR grant n. 2004137478_003 "MODELLI MULTIVARIATI E INDICATORI SINTETICI PER LANALISI DI SISTEMI AMBIENTALI COMPLESSI". We wish to thank Eng. M. Vultaggio, AMIA (Azienda Municipalizzata per l'Igiene Ambientale, Palermo) for data.

References

Bondì, A.L., Plaia, A., 2005. Weather variables and air pollution via hierarchical linear models. In: Proceedings of: SIS-2005 "Statistics and Environment" Messina, 21–23 September, Cleup, Padua, pp. 237–240.

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

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.

Latini, G., Passerini, G., 2004. Handling Missing Data: Applications to Environmental Analysis. WIT Press, Southampton, UK.

Li, K.H., Le, N.D., Sun, L., Zidek, J.V., 1999. Spatial–temporal models for ambient hourly PM_{10} in Vancouver. Environmetrics 10, 321–328.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.

R Development Core Team, 2000–2005. R language Definition. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-13-5 (URL ⟨http://www.R-project.org⟩).

Rubin, D.B., 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association 91, 473–489.

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

Van Buuren, S., Oudshoorn, C.G.M., 2005. Multivariate Imputation by Chained Equations: MICE Version 1.14 ⟨web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm⟩.

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