ELSEVIER

Atmospheric Environment 38 (2004) 2895–2907

ATMOSPHERIC ENVIRONMENT

www.elsevier.com/locate/atmosenv

Methods for imputation of missing values in air quality data sets

Heikki Junninen[a,b], Harri Niska[a,*], Kari Tuppurainen[c], Juhani Ruuskanen[a], Mikko Kolehmainen[a]

[a] *Department of Environmental Sciences, University of Kuopio, P.O.Box 1627, FIN-70211, Kuopio, Finland*
[b] *Institute for Environment and Sustainability, EC – Joint Research Centre, I-21020, Ispra (VA), Italy*
[c] *Department of Chemistry, University of Kuopio, P.O.Box 1627, FIN-70211, Kuopio, Finland*

Received 14 June 2002; accepted 11 February 2004

Abstract

Methods for data imputation applicable to air quality data sets were evaluated in the context of univariate (linear, spline and nearest neighbour interpolation), multivariate (regression-based imputation (REGEM), nearest neighbour (NN), self-organizing map (SOM), multi-layer perceptron (MLP)), and hybrid methods of the previous by using simulated missing data patterns. Additionally, a multiple imputation procedure was considered in order to make comparison between single and multiple imputations schemes. Four statistical criteria were adopted: the index of agreement, the squared correlation coefficient (R^2), the root mean square error and the mean absolute error with bootstrapped standard errors. The results showed that the performance of interpolation in respect to the length of gaps could be estimated separately for each variable of air quality by calculating a gradient and an exponent α (Hurst exponent). This can be further utilised in hybrid approach in which the imputation has been performed either by interpolation or multivariate method depending on the length of gaps and variable under study. Among the multivariate methods, SOM and MLP performed slightly better than REGEM and NN methods. The advantage of SOM over the others was that it was less dependent on the actual location of the missing values. If priority is given to computational speed, however, NN can be recommended. The results in general showed that the slight improvement in the performances of multivariate methods can be achieved by using the hybridisation and more substantial one by using the multiple imputations where a final estimate is composed of the outputs of several multivariate fill-in methods.

Keywords: Missing data; Air quality; Multivariate; Imputing; Neural networks

1. Introduction

Missing data, i.e. incomplete data matrices, are a problem that is repeatedly encountered in environmental research. The situation may be the result of insufficient sampling, errors in measurements or faults in data acquisition. Whatever the reasons, discontinuities pose a significant obstacle for time-series prediction schemes, which generally require continuous data as a condition for their use. Any relaxing of this condition will necessarily lead to the adoption of imputation techniques to cope with the problem.

The substitution of mean values for missing data is commonly suggested, and is still used in many statistical software packages. This can disrupt the inherent structure of the data considerably, however, leading to large errors in the covariance/correlation matrix and thereby degrading the performance of the statistical modelling. A slightly better approach is to impute the missing elements from an ANOVA model or something similar. More advanced imputation methods have been

*Corresponding author. Fax: +358-17-163191.
E-mail address: harri.niska@uku.fi (H. Niska).

1352-2310/$ - see front matter
doi:10.1016/j.atmosenv.2004.02.026

developed since the 1980s, and several methods and algorithms are now available (Little and Rubin, 1987; Schafer, 1997).

In broad outline, the methods available for creating complete data matrices can be divided into two main categories: single imputation and multiple imputation methods (Little and Rubin, 1987). Single imputation, i.e. filling in precisely one value for each missing one, intuitively has many appealing features, e.g. standard complete-data methods can be applied directly and the substantial effort required to create imputations needs to be carried out only once. Multiple-imputation is a method of generating multiple simulated values for each missing item, in order to reflect properly the uncertainty attached to missing data. This has been advocated as a statistically sound approach (Schafer, 1997), but so far its use has been limited mainly to the social and medical sciences.

In the present paper we compared a number of methods available for the imputation of missing air quality data: linear, spline and univariate nearest neighbour interpolations, regression-based imputation, multivariate nearest neighbour, self-organizing maps and multi-layer back-propagation nets. The univariate methods selected were simple and straightforward, and could thus be used for comparison. Additionally, we tested a multiple imputation scheme where multivariate methods were used together for estimating missing data. The emphasis of the work was particularly on the multivariate, the selection process being focused on ones that had been used before in air quality modelling. We then describe a 'hybrid' approach that we have developed in order to combine the best features of univariate and multivariate methods.

2. Materials and methods

2.1. Data

The performance criteria for an imputation method need to be defined somehow in terms of output. This is a difficult problem in practice, since there are hardly any complete data sets that could act as a reference. A perusal of all the data sets in the APPETISE[1] (Air pollution Episodes: Modelling Tools for Improved Smog Management) database (http://www.uea.ac.uk/env/appetise/) indicated that the Belfast 1998 and Helsinki 1998 data sets can provide feasible frame of reference, since only a small fraction of data was missing and the set of variables was fully uniform (see Table 1). The proposed data sets consisted of NO_x, NO_2, O_3, PM_{10}, SO_2 and CO concentrations, all on a time-scale of one per hour (hourly averaged), together with four meteorological parameters: wind speed (WS), wind direction (WD), temperature (T) and relative humidity (RH).

The reference data sets were completed by using the multivariate nearest neighbour method reported by Dixon (1979), after which the reliability of imputation was examined visually from imputed time-series and by statistical analysis (Table 1). The meteorological parameters of Helsinki 1998 were on a time-scale of one per three hours and therefore, were interpolated to the same time-scale with other parameters. Thus, the measurements resulted in a complete data matrix with 8758 lines, each described in terms of 10 measured responses (columns). In addition, there were four time variables, hour, day of the week, month and day of the year (1st January = 1, etc.), which were transformed to cyclic ones using sine and cosine components in order to avoid any non-physical discontinuities.

2.2. Simulation of missing data

Clearly, the imputation performance does not depend only on the amount of missing data but on the characteristics of missing data patterns. Moreover, the missing data mechanism of air quality data is generally random (MAR—missing at random), in the sense that the probability that a value is missing does not depend on the missing value (Rubin, 1976). Therefore, three randomly simulated missing data patterns were used for evaluating the methods in different missing data conditions. The patterns were different in complexity (Table 2) and where representing typical missing data patterns in air quality data. In addition, the blended pattern that includes simple, medium and complex patterns in proportion of $\frac{1}{4}, \frac{2}{4}$ and $\frac{1}{4}$ was constructed for examining the methods in more realistic condition reflecting the heterogeneity and dissimilarity of air quality data sets (Table 2). The patterns were simulated with missing data percentages of 10 and 25. Moreover, the patterns were mixed on the time series dimension so that each set formed ten new variants (the patterns were conserved) in order to minimise the effect of gap locations on performances.

2.3. Computational methods

2.3.1. Nearest neighbour, linear and cubic spline interpolation

Univariate nearest neighbour imputation is probably the simplest scheme available, in that the endpoints of the gaps are used as estimates for all the missing values (Eq.(1)).

$$y = y_1 \quad \text{if } x \leqslant x_1 + (x_2 - x_1)/2,$$

$$y = y_2 \quad \text{if } x > x_1 + (x_2 - x_1)/2, \tag{1}$$

where y is the interpolant, x is time point of the interpolant, y_1 and x_1 are the coordinates of the starting

[1] IST 1999-11764, EC framework V programme.

Table 1
The statistics of reference data sets by variables

Variable	Belfast						Helsinki					
	Md	Gap lengths			ΔStd	ΔMean	Md	Gap lengths (l)			ΔStd	ΔMean
		$l \leqslant 3$ h	3 h $< l \leqslant 24$ h	> 24 h				$l \leqslant 3$ h	3 h $< l \leqslant 24$ h	> 24 h		
NO_x	6.8	77.4	18.8	3.8	1.0	1.2	0.6	97.0	3.0	—	0.1	0.1
NO_2	6.8	77.4	18.8	3.8	0.5	0.5	0.6	97.0	3.0	—	0.1	0.1
O_3	6.1	80.0	16.0	4.0	0.0	1.2	0.2	71.4	28.6	—	0.1	0.0
PM_{10}	5.9	71.8	23.1	5.1	1.0	0.2	2.0	83.3	8.4	8.3	3.5	0.2
SO_2	10.9	71.4	21.5	7.1	0.0	0.9	4.4	100.0	—	—	3.6	0.0
CO	8.6	73.8	22.9	3.3	1.3	1	0.4	75.0	25.0	—	0.1	0.0
WS	0.3	50.0	50.0	—	0.0	0.0	66.7[a]	100.0[a]	—	—	—	—
WD	0.3	50.0	50.0	—	0.1	0.0	66.7[a]	100.0[a]	—	—	—	—
T	0.3	50.0	50.0	—	0.0	0.0	66.7[a]	100.0[a]	—	—	—	—
RH	0.3	50.0	50.0	—	0.0	0.0	66.7[a]	100.0[a]	—	—	—	—
Average	4.6	65.2	32.1	2.7	0.4	0.5	0.8	87.3	11.3	1.4	1.3	0.1

Incomplete rows (IR): 14.5% in Belfast data set and 7.2% in Helsinki data set.
Md—missing data (%), Gap lengths—missing data in the gap intervals (%), l—the length of gap in time, h—hour, ΔStd—change (%) in standard deviation after imputing, ΔMean—change (%) in mean after imputing.
[a] Linear interpolation from a time-scale of one per three hours to a time-scale of one per hour in Helsinki; these indices are not included in the calculation of averages.

Table 2
The settings of missing data simulation and resulting missing data statistics in the simulated patterns

Pattern type	Simulation settings				Resulting statistics					
	Column-wise area		Row-wise area		Missing data in the gaps (%)					IR (%)
	Min	Max	Min	Max	$l \leqslant 6$ h	6 h $< l <$ 24 h	24 h $< l <$ 3 d	3 d $< l <$ 7 d	$l >$ 7 d	
(a) *Missing data percentage of* 10										
Simple	3	10	1	5	37.2	62.8	—	—	—	31.7
Medium	50	200	1	5	—	—	14.9	75.3	9.8	32.5
Complex	50	200	8	10	—	—	—	83.2	16.8	11.4
Blended[a]	3	200	1	10	10.8	15.8	12.6	39.1	21.7	30.9
(b) *Missing data percentage of* 25										
Simple	3	10	1	5	35.7	64.3	—	—	—	83.1
Medium	50	200	1	5	—	—	7.8	56.8	35.4	75.0
Complex	50	200	8	10	—	—	12.1	64.4	23.5	27.5
Blended[a]	3	200	1	10	11.0	17.5	10.9	46.2	14.4	70.4

IR—incomplete rows, Column-wise area (l)—the length of gap in time, Row-wise area—missing values in 10 dimensional data vector, h—hour, d—day.
[a] Simple, medium and complex patterns are mixed in proportion of 25/50/25.

point of the gap, and y_2 and x_2 are the coordinates of the end point of the gap.

Linear interpolation (LIN) fits a straight line between the endpoints of the gap and enables the missing values to be calculated straightforwardly employing the line equation.

$$y = y_1 + k(x + x_1) \quad \text{where } k = (y_2 - y_1)/(x_2 - x_1); \\ x_1 < x < x_2 \text{ and } y_1 < y < y_2. \tag{2}$$

Cubic spline imputation is based on the fitting of cubic polynomials to a series of observed data points (x_i, y_i). The fitting is done so that at the knots (where piecewise portions joins) the function and its first two derivates are continuous. A cubic spline with knots at x_i, $i = 1, \ldots, n$ is defined as

$$f(x) = a_i + b_i x + c_i x^2 + d_i x^3 \quad \text{if } x_i \leqslant x < x_{i+1}. \tag{3}$$

2.3.2. Regression-based imputation

Regression-based imputation methods (REG) are based on estimated regression models between missing data and available data (as predictor). Here, the method

based on iterated analysis of linear regression by using EM algorithm (REGEM) (Schneider, 2001) was tested. For a detailed description of EM-algorithm, the reader is referred to Dempster et al. (1977).

In this method, the imputation procedure consists of iterations of the regularised EM algorithm where the means and covariance matrices of the incomplete data are estimated iteratively. Briefly, the iteration step of method can be described with three stages. First, for each incomplete record, the regression parameters of the variables are computed from the EM estimates of the mean and of the covariance matrix. Second, the missing values in a record are filled in with their conditional expectation values given the available values and the estimates of the mean and of the covariance matrix, the conditional expectation values being the product of the available values and the estimated regression coefficients. Third, the mean and the covariance matrix are re-estimated, the mean as the sample mean of the completed dataset and the covariance matrix as the sum of the sample covariance matrix of the completed dataset and an estimate of the conditional covariance matrix of the imputation error. For a detailed description of this method, the reader is referred to Schneider (2001).

2.3.3. Multivariate nearest-neighbour algorithm

The nearest neighbour (NN) imputation algorithm described by Dixon (1979), for instance, handles a row of N variables as a co-ordinate in an N-dimensional space and takes the missing values from the nearest neighbour (row) in that space where available, at the same time weighting the distances in proportion to the number of values missing in each row. Thus, rows having more missing values are under-weighted in order to compensate for their lesser reliability. The distance matrix is calculated by means of Eq. (4)

$$\mathrm{dist}_i^{\mathrm{ab}} = \mathrm{sqrt}(r_{\mathrm{a}}(i) - r_{\mathrm{b}}(i)),$$

$$\mathrm{dist}^{\mathrm{ab}} = \frac{N}{N - N^{\mathrm{md}}} \sum_{i=1}^{N} (\mathrm{dist}_i^{\mathrm{ab}})^2, \tag{4}$$

where $r_{\mathrm{a}}(i)$ and $r_{\mathrm{b}}(i)$ are ith elements of the N-dimensional feature vector a and b, respectively, $\mathrm{dist}_i^{\mathrm{ab}}$ is zero if either r_{a} or r_{b} or both are missing and N^{md} is the number of such pairs.

2.3.4. Self-organizing maps

Self-organizing map (SOM) neural networks (Kohonen, 1997) have been widely employed during the past 10–15 years, including applications to the atmospheric sciences (Kolehmainen et al., 2001). The basic idea of the SOM is to construct a mapping from the high dimensional input space R^n to a low dimensional output space consisted typically of a two-dimensional array of map units ('neurons'). A weight vector w_i for each neuron i is initialised from the discrete set $i = \{1, 2, \ldots, N\}$ and adapted by using an iterative unsupervised learning procedure in which usually the Euclidean distance $\|x - w_i\|$ is used for defining the best matching unit (BMU) b for the random data vector $x \in R^n$. The updates for the weight vectors of the BMU w_{b} and its neighbouring neurons i are made as follows:

$$w_i(t+1) = w_i(t) + h_{\mathrm{b}i}(t)[x(t) - w_i(t)], \tag{5}$$

where $h_{\mathrm{b}i}(t) = \alpha(t)\mathrm{H}((v_{\mathrm{b}} - v_i)/\sigma)$ defines the neighbourhood kernel $H(\cdot)$ over the map points (σ controls the width of the kernel, v_{b} and v_i are the coordinates of the units b and in the map) and the learning rate $(0 < \alpha(t) < 1)$ at iteration step t.

Missing values can be either ignored or included during the adaptation, and once the map is trained, the estimators for the missing values are taken from the nearest prototype vector represented by the weights of the map units. The prototype vector is found here by calculating the Euclidean distance between the prototype vector and incomplete data vector. For a detailed description of this method, the reader is referred to Häkkinen (2001).

2.3.5. Multi-layer back-propagation nets

The multi-layer perceptron (MLP) is probably the most widely known and successful neural network; for the theory and its applications in the atmospheric sciences, see Gardner and Dorling (1998, 1999). These networks employ a feed-forward architecture and are typically trained using a procedure called error back-propagation.

The MLP is used here in a combination of a number of separately trained two-layers MLP networks in which the learning algorithm of the scaled conjugated gradient and the transfer functions of sigmoid (hidden layer) and linear (output layer) were used. For each missing data pattern a network of its own is generated and trained by adopting early-stopping strategy (for preventing overfitting) with test set of training in proportion of $\frac{1}{5}$. The number of inputs corresponds with the number of available values and the number of outputs with the number of missing values in the data rows of missing pattern. The number of hidden neurons N_{mlp} is determined in an experimental way.

$$N_{\mathrm{o}} = \mathrm{round}(2\mathrm{x}I_{\mathrm{o}}) + 1,$$

$$N_{\mathrm{i}} = \mathrm{round}(2\mathrm{x}I_{\mathrm{i}}) + 1,$$

$$\text{if } N_{\mathrm{o}} < N_i, \text{ then } N_{\mathrm{mlp}} = N_i, \text{ else } N_{\mathrm{mlp}} = N_{\mathrm{o}}, \tag{6}$$

where N_{i} and N_{o} are the number of hidden neurons defined from the inputs (i) and outputs (o), I_{o} and I_{i} are the number of neurons in the input and output layers, and round is the rounding towards to nearest integer.

2.3.6. Hybrid model

In a 'hybrid' procedure considered here, short gaps are filled by the LIN and the rest of gaps by some multivariate method such as neural networks. The basic idea behind the method is an assumption that missing data in short gaps can be completed reliably by using the LIN, and further included to the training set of multivariate methods. This improves the performance of multivariate methods at least in case that a number of variables in a data row are missing and multivariate methods cannot derive enough information of a row pattern.

In this context, the critical length of gap depends on the variable under study, i.e. the maximum length of gap that can be replaced by the LIN must be estimated separately for each variable (see Fig. 1.), employing multiple linear regression with gradient and the exponent α as the independent variables (Eq. (7)):

$$\exp(d_i) = A_i \exp(\mathrm{grad}_i) + B_i \exp(\alpha) + C_i. \quad (7)$$

In this expression d_i is the index of agreement (see below), calculated iteratively for increasing hypothetical gap lengths i (range in the present material $1 \leqslant i \leqslant 20$), grad_i is the average gradient over the gap length i calculated for every available time point of the variable (real gaps in data were ignored), α is the exponent (Veitch, 2001; see also definition below), which is also calculated ignoring the real gaps, and A_i, B_i, and C_i are regression coefficients for the gaps i calculated from data sets.

The calculation of d was performed iteratively until the value of the index dropped below a chosen limit, in this case 0.90. The index was evaluated for gaps shorter than 20 values, the maximum gap length for replacement by the LIN in this 'hybrid' procedure.

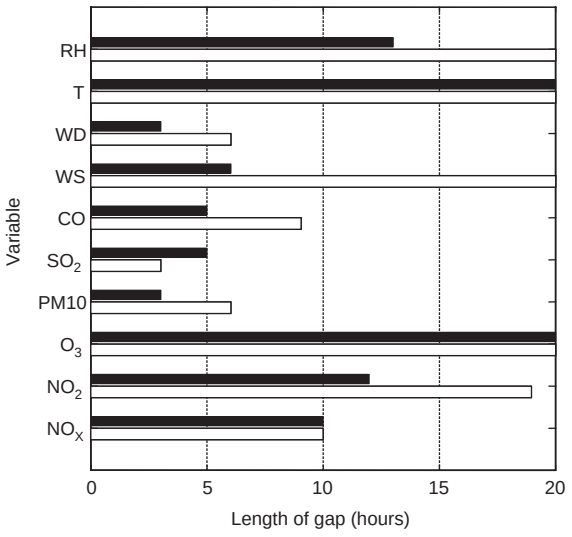

Fig. 1. Critical gap lengths of linear interpolation for variables of Helsinki (white bar) and Belfast (black bar) data sets where WS is wind speed, WD wind direction, T temperature and HUM relative humidity.

The exponent α can be defined as follows (Feder, 1988): if the power spectrum of a time series depends on the frequency (f), the fluctuation is said to be $1/f^{\alpha}$–like ('coloured' or flickering) noise. The spectral exponents α range approximately from 0 to 3, and the terms 'white' ($\alpha = 0$), 'pink' ($\alpha = 1$) and 'brown' ($\alpha = 2$) noise are in common use in the field of signal analysis. In practice, the slope of the best-fit line for the power spectrum as a function of frequency, both expressed on a logarithmic scale, gives a least-squares estimate for the spectral exponent α.

2.3.7. Multiple imputation scheme

The uncertainty attached to missing data may result into poor estimates due to insufficient sampling and disadvantages of the single imputing methods described above. Therefore, more accuracy can be attained by using model-based multiple imputations (MBMI) that solves the problem of underestimation of the error variance (Schafer, 1997). Probably, the most widespread MBMI approach for continuous multivariate data has been a normal model based method namely the NORM (Schafer, 1997), which can be loaded from the web address http://www.stat.psu.edu/~jls/misoftwa.html.

Recently, new promising MBMI-like methods have been developed. For example, Chiewchanwattana and Lursinsap (2002) used several different fill-in methods and generalised ensemble method (Perrone and Cooper, 1993) successfully in combination for improving the performance of neural networks in the incomplete time-series prediction. In this study we were interested to test straightforward approach where any weighting or more complex methods were not used but missing data estimate was derived directly as the mean output of multivariate methods (NN, SOM and MLP) and hybrid methods (H + NN, H + SOM and H + MLP) namely MI and H + MI.

2.4. Performance indicators

Several performance indicators were calculated for describing the goodness of imputation. The most common indicators of imputation ability are the correlation coefficient (R) and its square: coefficient of determination (R^2), i.e. the variance explained which is limited to a range between 0 and 1.

$$R^2 = \left[\frac{1}{N}\frac{\sum_{i=1}^{N}[(P_i - \bar{P})(O_i - \bar{O})]}{\sigma_P \sigma_O}\right]^2, \quad (8)$$

where N is the number of imputations, O_i the observed data point, P_i the imputed data point, $\bar{O}$ is the average of observed data, $\bar{P}$ average of imputed data, σ_P the

standard deviation of the imputed data and σ_O the standard deviation of the observed data.

However, as emphasised by Willmott et al. (1985), the values of these indicators may be unrelated to the sizes of the discrepancies between the predicted and observed values. To circumvent this problem, an index of agreement (d) has been developed (Willmott, 1982):

$$d = 1 - \left[\frac{\sum_{i=1}^{N}(P_i - O_i)^k}{\sum_{i=1}^{N}(|P_i - \bar{O}| + |O_i - \bar{O}|)^k}\right], \quad (9)$$

where k is either 1 or 2. The index has been employed throughout this work with k set to 2 (d_2).

The root mean squared error (RMSE) which summarises the difference between the observed and imputed concentrations was used to provide the average error of model.

$$\mathrm{RMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}[P_i - O_i]^2\right)^{\frac{1}{2}}. \quad (10)$$

Moreover, the mean absolute error (MAE) was included to the comparison as more sensitive measure of residual error as RMSE.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i|. \quad (11)$$

The cyclic nature of wind direction was taken account in the calculations of mean and distance values. The average value was calculated by using four quadrant arctangent function for sine and cosine components of wind direction and the distance between predicted and observed through the nearest direction.

Finally, in order to achieve more understanding of the accuracies of the indicators the standard errors (SE) were calculated by using the bootstrap method (Efron and Tibshirany, 1993) with 300 bootstrap samples. In this procedure, samples have been chosen randomly with replacement from an imputed data set, and then the performance of each sample has been analysed the same way. Lastly, SE is derived as a standard deviation of performance analysis.

2.5. Execution of the tests

First of all, the ability of the interpolation methods was examined in order to determine the most feasible interpolation method in the later stages of the work. For examining that, simulated incomplete data were generated with percentage 25% of items missing and varying gap lengths within the range $\{1, \ldots, 50\}$ to the Helsinki and Belfast data sets, after which the univariate methods were applied to the data. Lastly, the comparison of the interpolation methods was considered as a function of gap length against performance criterion d_2 (see Fig. 2). Based on these tests, the LIN was selected for further evaluation (see Section 3.1).

In the next phase, the main comparison was performed between the LIN, the multivariate methods (REGEM, NN, SOM and MLP), the hybrid methods (H + NN, H + SOM and H + MLP), the multiple imputation procedures (MI and H + MI) and reference methods, namely the substitution of mean value (MEAN) and random value (RAND). Four missing data patterns discussed in Section 2.2 (Table 2) were used for testing the methods in different missing data conditions. The performances were examined for each missing data pattern (see Section 3.2) by calculating the statistical indices (Section 2.4) as the average of data columns.

Moreover, the assessment was made for time series separately (see Section 3.3) in order to achieve a more

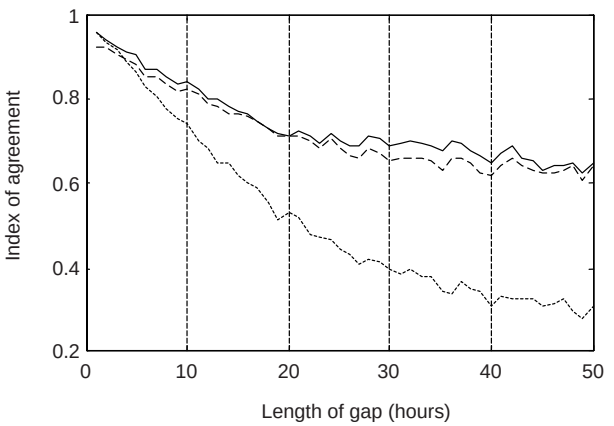

Fig. 2. Performance of different interpolation methods as a function of gap length. The test was performed using the Helsinki and Belfast data sets. Solid line—linear interpolation, dashed line—nearest neighbour interpolation and dotted line—cubic spline interpolation.

precise view about the quality of imputation. In this closer examination, we used the blended test patterns (Section 2.2) and missing data fraction of 25%, which comprised a good reference data for assessing the methods in more real-like situations. The imputation performance was then examined in the context of the index of agreement, scatter plots (predicted versus observed signals) and power spectral analysis. The spectral analysis comprised the analysis of time series frequencies and amplitudes by computing a Fourier transform and taking the portion of a signal's power as squared amplitude.

2.6. Software

Numerous in-house MATLAB^{TM} scripts were written to facilitate the computations in cases where no appropriate programs were available. These scripts can be downloaded for academic purposes as the Missing Data Toolbox at the web address www.uku.fi/envi/downloads.htm.

3. Results and discussion

3.1. Univariate methods

Univariate methods, namely nearest neighbour, linear and spline interpolations, were explored with simulated missing data (see Section 2.5). As might be expected, the performance of the missing data interpolation was limited, and in general they were able to fill only very short gaps (Fig. 2). The results of the linear and spline fitting were equally good for gaps of 1–2 values (hours), but the performance of splines declined faster as the length of gap increased, until they occasionally generated non-physical artefacts with gaps longer than 24 values, so that the unsystematic error increased drastically. Consequently, the use of splines cannot be recommended. Among the univariate methods linear interpolation is the method of choice, remembering its restrictions with regard to gap length. In general, the performance of any interpolation method will depend greatly on the variable under study.

3.2. Simulated test patterns

3.2.1. Simple, medium and complex patterns

The performance indices (d_2) in the simple, medium and complex patterns have been presented in Table 3. In the simple patterns the LIN performed better than the multivariate methods whereas in the medium ones the performance of the LIN decreased dramatically compared to the multivariate methods. The hybrid methods were capable of yielding high performance in both simple and medium conditions. Instead, the results in Table 3 show that all the methods degenerated in performance when the data included more complex patterns.

When considering the average performances $\bar{x}$ (see Table 3) both SOM and MLP apparently performed somewhat better than the REGEM and the multivariate NN method. This seems reasonable, since air quality time series are derived from complex, non-linear processes that neural networks might be able to capture without the limitations of more conventional statistical methods. Moreover, slight improvement in the average performances was obtained by combining the LIN with

Table 3
The comparison of methods versus missing data patterns in terms of index of agreement (d_2) with bootstrap estimates of standard errors where $\bar{x}$ is mean of patterns. The indices of the best model are in bold

Method	Helsinki				Belfast			
	Simple	Medium	Complex	$\bar{x}$	Simple	Medium	Complex	$\bar{x}$
RAND	0.22±0.01	0.23±0.01	0.24±0.01	0.23±0.01	0.22±0.01	0.22±0.01	0.23±0.01	0.22±0.01
MEAN	0.10±0.03	0.22±0.03	0.21±0.03	0.17±0.03	0.11±0.03	0.19±0.03	0.18±0.03	0.16±0.03
LIN	**0.85±0.02**	0.58±0.02	0.58±0.02	0.67±0.02	0.85±0.01	0.55±0.02	0.57±0.02	0.66±0.02
REGEM	0.68±0.02	0.68±0.02	0.39±0.02	0.58±0.02	0.73±0.02	0.72±0.02	0.49±0.02	0.64±0.02
NN	0.79±0.02	0.69±0.02	0.63±0.02	0.70±0.02	0.79±0.02	0.71±0.02	0.63±0.02	0.71±0.02
SOM	0.78±0.02	0.75±0.02	0.62±0.02	0.72±0.02	0.81±0.01	0.77±0.02	0.65±0.02	0.75±0.02
MLP	0.72±0.02	0.75±0.02	0.66±0.02	0.71±0.02	0.77±0.01	0.76±0.02	0.65±0.02	0.73±0.02
MI	0.81±0.02	0.77±0.02	0.67±0.02	0.75±0.02	0.83±0.01	0.78±0.02	0.68±0.02	0.77±0.02
H+NN	0.85±0.02	0.71±0.02	0.66±0.02	0.74±0.02	0.84±0.02	0.72±0.02	0.65±0.02	0.74±0.02
H+SOM	0.84±0.02	0.76±0.02	0.63±0.02	0.75±0.02	0.84±0.01	0.77±0.02	0.65±0.02	0.76±0.02
H+MLP	0.83±0.02	0.77±0.02	0.66±0.02	0.76±0.02	0.84±0.01	0.77±0.02	0.66±0.02	0.77±0.02
H+MI	**0.85±0.02**	**0.79±0.02**	**0.69±0.02**	**0.78±0.02**	**0.86±0.01**	**0.79±0.02**	**0.69±0.02**	**0.78±0.02**

RAND—random imputation, MEAN—mean imputation, LIN—linear interpolation, REGEM—regularised expectation maximisation, NN—nearest neighbour, SOM—self-organizing map, MLP—multi-layer perceptron, MI—multiple imputation procedure. Hybrid methods are marked with notation H+ method name.

the multivariate methods and more substantial one by the multiple imputation procedures.

Slight difference in the performances between locations could be detected (Table 3) only in terms of average error of the REGEM. However, surely the data interpolation (discussed in Section 2.1) has led to the loss of information, and thus, at least the LIN should perform slight better with interpolated data. This can be seen in Fig. 1 where the meteorological variables of Helsinki: HUM, WD and WS have longer critical gaps compared to the variables of Belfast due to the interpolation. On the other hand, the results suggest that relationships and phenomenal backgrounds of variables in the two test locations had similarities.

3.2.2. Blended pattern

The results of blended pattern (Table 4) showed that multiple imputation procedures (MI and H + MI) over performed the other methods in terms of d_2 and R^2 as it was observed with previous test patterns (see Table 3). Similar conclusions could be drawn from the mean MAE. However, also the hybrid methods were capable of yielding good performances, although when considering RMSE the advantage of hybridisation was not clear. This might be due to the sensitivity of RMSE for extreme values and thus, MAE could be more reliable indicator.

Also, we noticed that the increase in missing data percentage (10–25%) particularly decreased the ability of the NN, SOM and MLP whereas the performance of the LIN and the REGEM maintained almost at the same. The degeneration can be explained by the loss of training data, which leads to poorer generalisation ability in case of neural networks (SOM and MLP) and to smaller set of history data to be used for imputation in case of the NN. Obviously, the REGEM was capable of maintaining ability via the iterated EM-estimation of incomplete data.

3.3. Results by variable

In this paper only the scatter plots and the power spectrums for the hybrid SOM are presented (Figs. 4 and 5). The statistics depicted in Fig. 3 indicated that the general performance of imputation was fair good when considering the pollutants (NO_x, NO_2, O_3, PM_{10}, SO_2 and CO) which can be regarded valuable in terms of air quality modelling (Kukkonen et al., 2003). However, clear limitations were detected when considering the meteorological variables such as WS and WD direction that is largely due to complex (chaotic) phenomenon of these variables. Similar conclusions could be drawn from the scatter plots (Fig. 4).

Finally, we examined the effect of the different imputation methods on the spectral properties of the time series. In general, only slight changes between the power spectra of the imputed vs. original time series could be detected, as is exemplified with the hybrid SOM data (Fig. 5). The results with other imputation methods were similar (data not shown). However, it should be emphasised that the spectral analysis is not very sensitive in assessing the accuracy of the imputation. For example, both the statistics (Fig. 3) and scatter plots (Fig. 4) indicate that performance of the imputation is somewhat poor for the variable WS, but no dramatic changes in the power spectra could be detected.

3.4. Summary

The results suggested that SOM or MLP, improved with the hybrid approach, are the methods of choice for air quality data imputation and even better results can be

Table 4
The performance indices of the models when tested on the blended test patterns. The results differ by the missing data percentages and the best indices are in bold

Method	Missing data percentage of 10				Missing data percentage of 25			
	RMSE	d_2	R^2	MAE	RMSE	d_2	R^2	MAE
RAND	189.1±3.5	0.23±0.01	0.02±0.01	159.8±3.8	183.0±1.9	0.22±0.01	0.04±0.00	154.6±2.1
MEAN	29.2±1.2	0.20±0.03	0.06±0.00	22.7±0.7	28.5±0.9	0.13±0.02	0.07±0.00	21.6±0.4
LIN	29.4±1.3	0.66±0.02	0.28±0.03	21.4±0.8	29.0±0.9	0.66±0.02	0.29±0.02	20.3±0.5
REGEM	22.6±1.1	0.67±0.02	0.38±0.03	16.5±0.6	22.8±0.8	0.64±0.02	0.33±0.02	16.5±0.3
NN	27.6±1.5	0.73±0.02	0.37±0.03	19.3±0.8	29.3±1.0	0.66±0.02	0.27±0.02	20.0±0.5
SOM	24.6±1.3	0.76±0.02	0.44±0.03	17.7±0.7	24.7±0.9	0.74±0.02	0.41±0.02	17.0±0.4
MLP	23.3±1.1	0.76±0.02	0.49±0.03	17.7±0.6	22.8±0.8	0.72±0.01	0.46±0.02	16.6±0.3
MI	**22.2±1.1**	0.79±0.02	**0.54±0.03**	**16.2±0.6**	**22.1±0.8**	0.75±0.01	0.49±0.02	15.6±0.3
H + NN	26.8±1.4	0.76±0.02	0.41±0.03	18.9±0.8	26.5±1.0	0.73±0.02	0.36±0.02	17.5±0.5
H + SOM	25.3±1.3	0.78±0.02	0.46±0.03	18.0±0.7	24.2±0.9	0.75±0.02	0.43±0.02	16.2±0.4
H + MLP	25.1±1.3	0.78±0.02	0.50±0.03	18.2±0.7	23.1±0.9	0.77±0.02	0.48±0.02	16.0±0.4
H + MI	24.2±1.3	**0.80±0.02**	**0.54±0.03**	17.3±0.7	22.7±0.9	**0.78±0.01**	**0.50±0.02**	**15.4±0.4**

The notation is the same as that in Table 3.

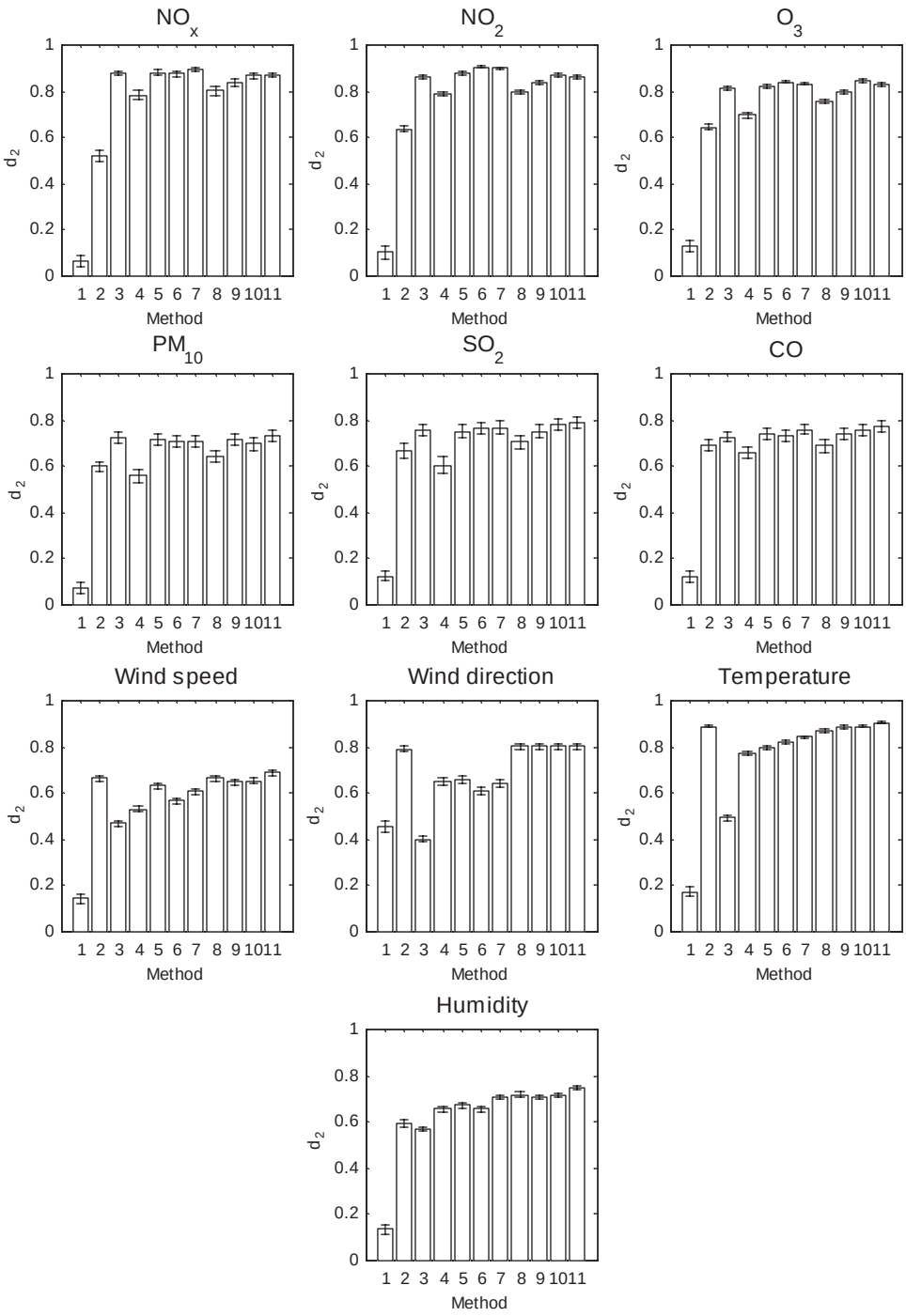

Fig. 3. The performance of imputation ($d_2 \pm$SE) by the methods in the Belfast data set when the blended test patterns and the missing data fraction of 25% were employed. The methods are listed as follows: 1—MEAN, 2—LIN, 3—REGEM, 4—NN, 5—SOM, 6—MLP, 7—MI, 8—H + NN, 9—H + SOM, 10—H + MLP and 11—H + MI.

achieved by using the multiple imputations. The advantage of SOM over other methods is that it is less dependent on the actual location of the missing values. If priority is given to computational speed, however, a hybrid application of the multivariate NN method can be recommended. As with univariate methods, the differences in performance between variables were large. The pros and cons can be found in more detail in Table 5.

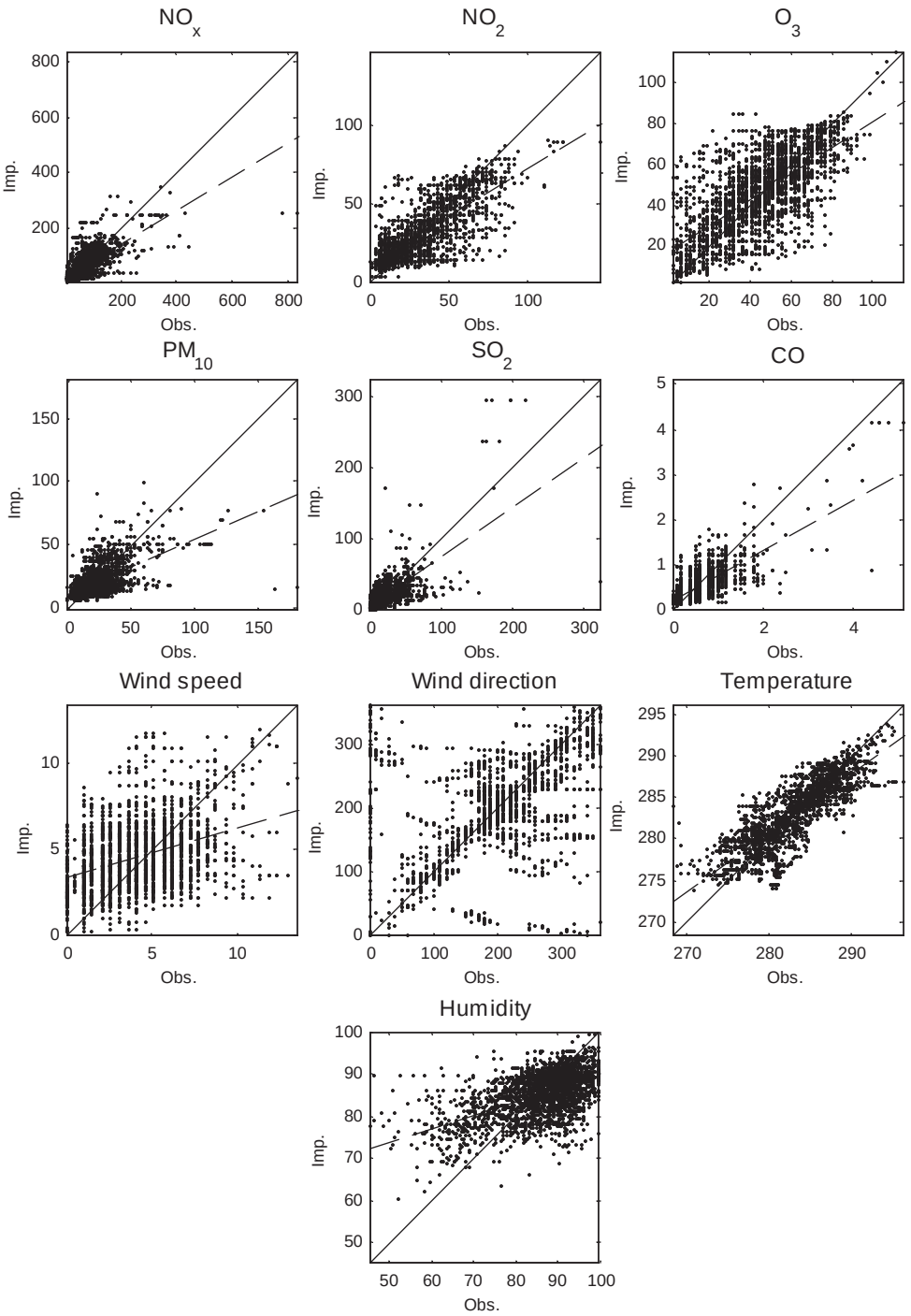

Fig. 4. The scatter plots of the observed versus imputed data of Belfast for the hybrid SOM when using the blended patterns and the missing data fraction of 25%. The plots are further enhanced with a least-squares fitting line (dotted) and a line showing perfect fit (solid) except the plot of wind direction where a least-square fitting has not been included due to the variable's cyclic nature.

It should be emphasised that single-imputed data can be misleading because the single values cannot reflect sampling variability around the actual (unknown) values—no matter how carefully the imputation has been done. Thus, for high variance models such as neural networks multiple imputations may therefore

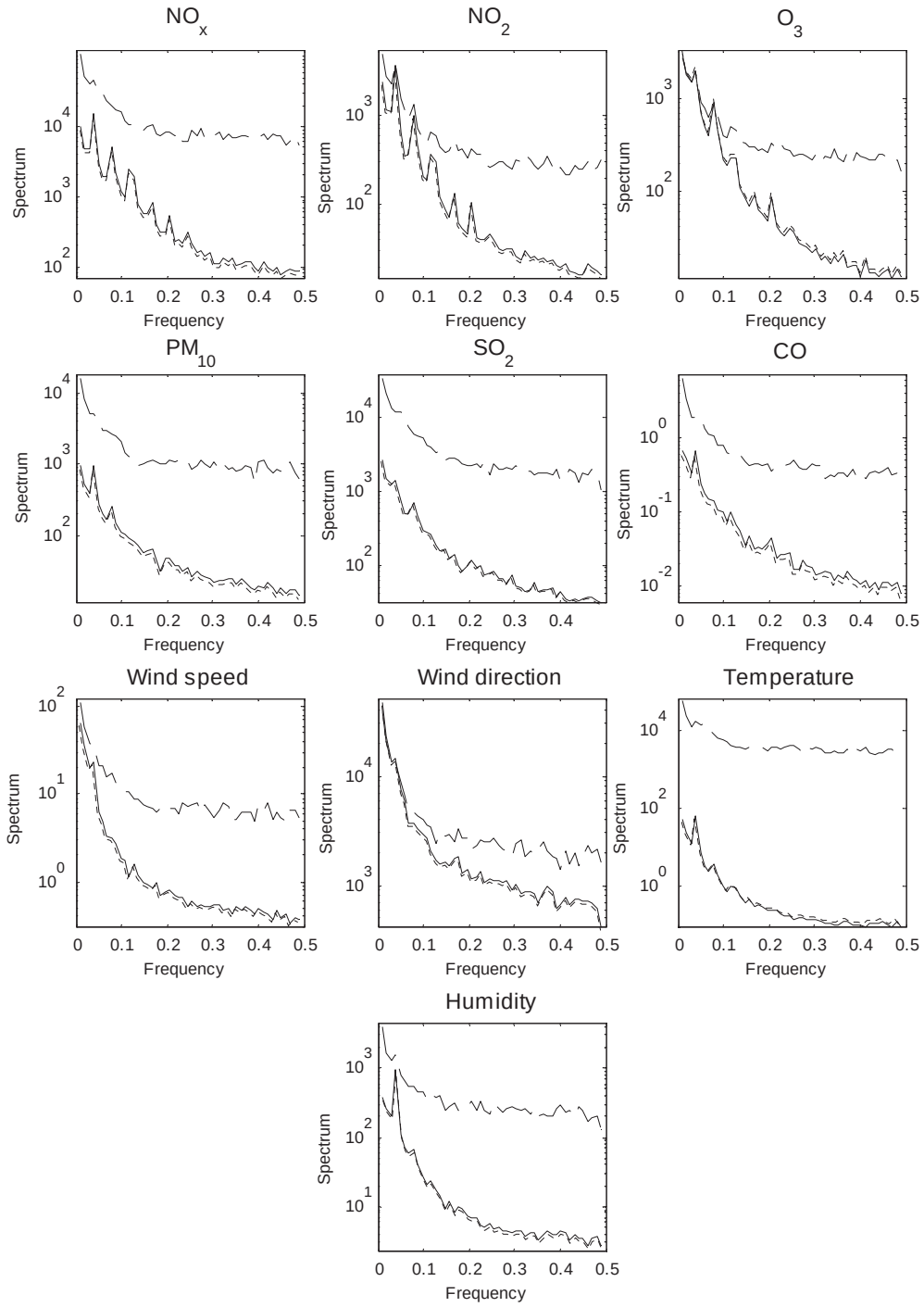

Fig. 5. The power spectra of the observed data (solid line) and imputed data sets of Belfast for the hybrid SOM (dotted line) and random imputation (dashed line) when using the blended patterns and the missing data fraction of 25%.

yield a substantial performance improvement. Consequently, the results should always be interpreted with due caution.

The results suggested also that the general missing data problem in air quality data may be divided into three main categories: (1) simple patterns, i.e. short gaps

Table 5
Pros and cons of the imputation methods studied in the work

Method	Performance		Speed	Reliability		Comments
	Short gaps	Long gaps		Short gaps	Long gaps	
Linear interpolation	+ + + +	+	+ + + +	+ + + + +	+ + +	Applicable for short gaps; utilises only local information
Linear regressions (REGEM)	+ + +	+ +	+ + +	+ + +	+ +	Limited only to linear relationships of variables; applicable for the variables having strong linear correlations
Multivariate nearest neighbour (NN)	+ + +	+ + +	+ + +	+ + +	+ + +	Employs existing data only; incomplete rows are utilized
Self-organizing map (SOM)	+ + +	+ + +	+ + +	+ + +	+ + +	Repeatable; incomplete data rows are utilized
Multi-layer perceptron (MLP)	+ +	+ +	+ +	+ +	+ +	Not repeatable; incomplete data rows are not utilized
Hybrid methods	+ + + +	+ + +	+ +	+ + +	+ + +	The best overall methods; performance depends on what methods are combined together
Multiple imputations	+ + + +	+ + + +	+	+ + +	+ + +	The best imputation procedure; reflect the uncertainty attached to missing data

+ Poor, + + + + + Best.

and relatively complete data rows, (2) medium patterns, i.e. long gaps and relatively incomplete data rows, and (3) complex patterns, i.e. long gaps with near or fully incomplete data rows. For categories 1 and 2, which fortunately cover the vast majority of cases, hybrid approaches work well, but for category 3, no imputation method discussed above will provide a complete solution, although partial success can be achieved by incorporating a large number of physically meaningful, independent, subsidiary constraints into the system, such as non-negativity, reasonable values for the imputed items, closure relations or coupling equations between the matrix elements, etc., but all these constraints are very difficult to incorporate and must always be corroborated by other information outside the original data set. This means that the problem will have no universal or automatic solution. Hence, the missing data problem is, in its most general form, more or less impossible to solve completely.

4. Conclusions

The aim of this work was to evaluate and compare univariate and multivariate methods for missing data imputation in air quality data sets. The univariate methods studied were linear interpolation, spline interpolation and nearest neighbour interpolation and the multivariate methods were the regression-based imputation, the multivariate nearest neighbour method, the self-organizing map and the multi-layer perceptron. Additionally, the model based multiple imputations scheme was tested.

The results showed that univariate methods are dependent on the length of the gap in time and that their performance also depends on the variable under study. Moreover, it was shown that the length of gap that can be replaced by the LIN can be estimated separately for each variable by calculating the gradient and the exponent α. The results obtained with the multivariate methods showed that both SOM and MLP perform slightly better than the multivariate NN method. The advantage of the SOM over other methods is that it is less dependent on the actual location of the missing values, while the advantages of the NN methods are particularly important in practical applications, i.e. it is computationally less demanding and does not generate new values in the data.

The results in general showed that slight increase in performances can be achieved by the hybridisation of the multivariate methods; and that the way this combining should be done is dependent on the variable to be inspected. Moreover, it was detected that the single imputation methods underestimates the error variance of missing data and accuracy can be improved substantially by using the multiple imputations and

thus, further work should be focused especially for that issue.

Finally, we wish to emphasise that imputation methods should not be regarded as a kind of 'statistical alchemy', which somehow could help scientists to generate information from nothing. Missing data are always lost, in their entirety and forever, but a proper imputation scheme can help to remedy the situation as much as possible. The method that performs best in each situation, in terms of the assessments made in this work (briefly summarised in Table 5), should be selected for wider dissemination and use.

Acknowledgements

The work was carried out within the framework of the project "Air Pollution Episodes: Modelling Tools for Improved Smog Management" (APPETISE), funded under the EU Information Society Technology Program Contract JST-99-11764 and co-ordinated by the School of Environmental Sciences, University of East Anglia, Earlham Rood, Norwich, NRY 7TS, UK.

References

Chiewchanwattana, S., Lursinsap, C., 2002. FI-GEM networks for incomplete time-series prediction. In: Proceedings of the 2002 International Joint Conference on Neural Networks, vol. 2, pp. 1757–1762.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society 39, 1–38.

Dixon, J.K., 1979. Pattern recognition with partly missing data. IEEE Transactions on Systems, Man, and Cybernetics 10 (SMC-9), 617–621.

Efron, B., Tibshirany, R., 1993. An Introduction to the Bootstrap. Chapman & Hall, London.

Feder, J., 1988. Fractals. Plenum Press, New York.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multiplayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment 32, 2627–2636.

Gardner, M.W., Dorling, S.R., 1999. Neural network modeling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmospheric Environment 33, 709–719.

Häkkinen, E., 2001. Design, implementation and evaluation of the neural data analysis environment. Ph.D. Thesis, University of Jyväskylä, Finland.

Kohonen, T., 1997. Self-Organizing Maps, 2nd Edition. Springer, Berlin.

Kolehmainen, M., Martikainen, P., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. Atmospheric Environment 35, 815–825.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G., 2003. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment 37, 4539–4550.

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

Perrone, M.P., Cooper, L.N., 1993. When neural networks disagree: ensemble methods for hybrid neural networks. In: Mammone, R.J. (Ed.), Neural Networks for Speech and Image Processing. Chapman & Hall, London, pp. 126–142.

Rubin, D.B., 1976. Inference and missing data. Biometrika 65, 581–592.

Schafer, J.L., 1997. Analysis of incomplete multivariate data. Monographs on Statistics and Applied Probability No. 72. Chapman & Hall, London.

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

Veitch, D., 2001. Matlab code for the estimation of scaling exponents (www.emulab.ee.mu.oz.au/~darryl/estimation_code.html)

Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society 63, 1309–1313.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. Journal of Geophysical Research 90 (C5), 8995–9005.