Missing Data Toolbox for Air Quality Datasets

Mikko Kolehmainen[1], Heikki Junninen[2], Harri Niska[1], Toni Patama[1], Anna Ruuskanen[3], Kari Tuppurainen[4] and Juhani Ruuskanen[1]

1.1. Abstract

Environmental databases and their effective use in many cases are needed to be uniterrupted by missing data values. The objective of our study was to find a useful missing data replacement method for air quality forecasting applications. A number of statistical methods were studied for their ability to replace the missing values in air quality data. Three computationally simple and fast univariate methods and four advanced multivariate methods were originally selected for the study.

The univariate methods studied were the linear interpolation, spline and nearest neighbour (univariate) interpolation. Linear interpolation fits a straight line between the endpoints of a gap formed by missing values, which can be calculated straightforwardly employing the line equation. Spline interpolation is based on the polynomials of different degrees. Univariate nearest neighbour is likely the simplest imputation scheme: the endpoints of a gap are used as estimates for all missing values.

Multivariate methods studied were multivariate nearest neighbour (NN), principal components analysis (PCA), Self-Organising Map (SOM) and Multi-Layer Perceptron (MLP). The NN imputation 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 they are available, weighting at the same time the distances proportionally to the number of missing values in each row. When PC

[1] Department of Environmental Sciences University of Kuopio, P.O.Box 1627, 70211 Kuopio, Finland

[2] Institute for Environment and Sustainability, EC – Joint Research Centre, I-21020, Ispra (VA), Italy

[3] Department of Applied Physics, University of Kuopio, P.O.Box 1627, 70211 Kuopio, Finland

[4] Department of Chemistry, University of Kuopio, P.O.Box 1627, 70211 Kuopio, Finland

is employed for imputation the approach is to simply skip the missing elements in the appropriate Gaussian normal equations used to calculate the score and loading matrices and then produce a complete data matrix employing a matrix multiplication. The goal of the SOM is to find vectors, which can represent the input data set with prototypes and at the same time realize a continuous mapping from input space to a lattice. The missing values can then be recovered using the lattice. Similarly with the PC and NN also the SOM is using a whole data set so that the information in incomplete rows is utilised. During the training process of the SOM missing elements are ignored and only values available in a vector (data row) are used for updating the weights of the map. The idea of the MLP network is to learn a black box like mapping from input variables to one or several output variables. The usage of MLP in this study is actually a combination of several MLP-nets so that for each missing data pattern network of its own was trained.

In our study, new approach was developed where univariate methods are combined with multivariate methods in order to utilise the best properties of both approaches. In this approach short gaps are filled with univariate interpolation which gives the best performance in the given situation and longer gaps are filled with a advanced multivariate method which actually has no dependence on the gap length in time space. In this context, the ‘shortness’ depends on the variable under study.

The results showed that univariate methods are dependent on the length of gap in time space and that the performance depends also on a variable under study. The univariate method selected for further evaluation combined with the multivariate method was chosen to be the linear interpolation.The length of gaps that can be replaced with linear interpolation can be estimated separately for each variable before the replacement. This is carried out by calculating the gradient and α-exponent by ignoring missing values and then employing multiple linear regressions with exponents of gradients and α-exponents as independent variables.

The results of the multivariate methods showed that both the SOM and MLP methods perform slightly better than multivariate NN method. The advantage of SOM compared with other methods is that it is less dependent on the actual location of the missing values. However, if computational speed is placed at first, multivariate NN is then recommended. It has also the advantage of not generating completely new values to the data. As with univariate methods, the differences in performance between variables were large.

The results with the PCA showed that it usually generates a small number of unphysical artefacts, including both bad outliers and negative values. Moreover, the algorithm was found to be very sensitive to the selected number of principal components, which makes the automation of the method difficult.

The results in general showed that the best overall performance can be achieved by combining univariate and multivariate methods and that the way of combining is dependent on the variable inspected. Based on these results a toolbox in Matlab environment was generated, which encapsulates the different algorithms and enables the treatment of missing data in a coherent way. The MDT (Missing Data Toolbox) is able to replace missing data in large air quality datasets. End-users can choose a method for the replacement from various linear and non-linear methods such as interpolation, nearest neighbour, SOM and MLP. Combinations of univariate and multivariate methods are available as hybrid methods, too. The toolbox was further enhanced with a graphical user interface (GUI) and the resulting tool can be distributed and used by average end-user with air quality data having missing values.

The MDT and GUI were tested on Windows and Linux environments. It was found to be usable in both environments but there are some restrictions on the Matlab version to be used with the package. The GUI was decided to be the best approach in order to bring the MDT for the end-users. The GUI has algorithms for loading and saving datasets in different file formats, calculating different statistics for missing data and illustrating missing data conditions with graphs. In the future the end-users will be able to download required Matlab codes for the MDT over Internet.