Missing Data Toolbox for Air Quality Datasets

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

Abstract

The objective of the study was to find a useful missing data imputing method for air quality forecasting applications. The univariate methods studied were the linear interpolation, spline and nearest neighbour (univariate) interpolation. Multivariate methods studied were multivariate nearest neighbour (NN), Self-Organising Map (SOM) and Multi-Layer Perceptron (MLP). Additionally, a new approach was developed where univariate methods were combined with multivariate methods in order to utilise the best properties of both approaches. 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 Missing Data Toolbox (MDT) with a Graphical User Interface (GUI) in Matlab environment was created. The MDT encapsulates the different algorithms and enables the treatment of missing data in a coherent way. The MDT and GUI were tested on Windows and Linux environments.

Departments of Environmental Sciences[1], Applied Physics[2] and Chemistry[3], University of Kuopio, P.O.Box 1627, FIN-70211 Kuopio, Finland
Institute for Environment and Sustainability[4], EC – Joint Research Centre, I-21020, Ispra (VA), Italy