Geophysical Research Abstracts,
Vol. 11, EGU2009-3302, 2009
EGU General Assembly 2009
© Author(s) 2009

Analysis of Pollution Patterns Using Unsupervised Machine Learning Algorithms

M. Kanevski (1), V. Timonin (1), A. Pozdnoukhov (2), and M. Maignan (1)

(1) Institute of Geomatics and Analysis of Risk, Lausanne, Switzerland (Mikhail.Kanevski@unil.ch), (2) National Centre for Geocomputation, National University of Ireland, Maynooth, Ireland

The research presents an application of Machine Learning Algorithms, mainly unsupervised learning techniques like self-organising Kohonen maps (SOM), to study spatial patterns of multivariate environmental spatial data. SOM are well-known neural networks widely used for high-dimensional data analysis, modelling (clustering and classification), and visualization.

Self-organising maps belong to the unsupervised machine learning algorithms providing solutions to clustering, classification or density modelling problems using unlabeled data. SOM are efficiently used for the dimensionality reduction and for the visualisation of high-dimensional data (projection into a two-dimensional space). Unlabeled data are points/vectors in a high-dimensional feature space that have some attributes (or coordinates) but have no target values, neither continuous (as in a regression problem) nor discrete labels (as in the case of classification problem). The main task of SOM is to “group” or to “range” in some manner these input vectors and to try to catch regularities (to find patterns) in data by preserving topological structure and by using some well defined similarity measures.

A generic methodology presented in this study consists of detailed spatial exploratory data analysis using statistical and geostatistical tools, analysis and modelling of spatial (cross)-correlations anisotropic structures, and application of SOM as a nonlinear modelling and visualisation tool. The case study considers multivariate data of sediments contamination by heavy metals (eight spatially distributes pollutants) in Geneva Lake. The most important modelling task is formulated as a problem of revealing structures or coherent clusters in this multivariate data set that would shed some light on the underlying phenomena of the contamination. Three major clusters, clearly spatially separated, were detected and explained by using the SOM technique.