Learning from environmental data: recent achievements and new challenges

Prof. M. Kanevski, IDYST, Faculty of Geosciences, University of Lausanne

Mikhail.Kanevski@unil.ch
Science Paradigms

- Thousand years ago: science was **empirical**
  describing natural phenomena
- Last few hundred years: **theoretical** branch
  using models, generalizations
- Last few decades: a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

\[
\left(\frac{\dot{a}}{a}\right)^2 = \frac{4 \pi G p}{3} - K \frac{c^2}{a^2}
\]
Our experience

- **Environmental data mining, machine learning**
- Geomatics
- Pollution data analyses and modelling (soil, water, air)
- Topo-climatic & meteodata modelling
- Natural hazards: landslides, forest fires, avalanches, earthquakes
- Fractals: MNO, spatial stat, local patterns…
- Time series, geophysical signals, chaos
- Socio-economic & geodemographic data
- Financial data, crime data, epidemiology
Environmental data

- Multi-scale (fractals?).
- High variability at several spatio-temporal scales
- Uncertainties: from raw data via **MODELLING** to the results
- Input spaces (features): high dimensional: >(10-100).
- Multivariante (pollution + environment ~ 100)
- Non-homogeneous data: discrete/classes, continuous variables, densities, etc…
- Environmental phenomena are **nonlinear**
- Data + science based models **integration**

*Detection of patterns, analysis of predictability: difficult problems in a high-dimensional space with noisy data for nonlinear phenomena*
Learning of spatio-temporal data in terms of patterns/structures:
- pattern recognition,
- pattern modelling,
- pattern predictions
A Generic Model of Learning from Data/Examples

MLA: (semi)(un)supervised, Active, transductive, manifold,…
Three fundamental tasks of statistical learning

Classification

Regression

Probability density modelling

Prof. M. Kanevski
Generic Methodology: from raw data via the validation/testing to the results with uncertainties

Data (real & simulated/shuffled) + Problems/Objectives

Input space preparation
Data splitting

Monitoring networks

EDA/ESDA
Data preprocessing

Modelling Tools
Training, validation
Feature selection

Modelling of Patterns
Detection/Analysis

Applications/Predictions
Uncertainties

Testing, EDA of Residuals
Spatial point pattern analysis of environmental data
(Marj Tonini)
Future studies in collaboration with WSL and the university of Lisbon

• Case studies and dataset: forest-fires and burnt areas

• Objectives: pattern distribution and influencing variables

• Methods:
  – Multivariate cluster analysis of forest fire events: spatio-temporal K-function, clustering, 3D-geovisualisation
  – Space-time permutation scan statistics: influence of the input dataset, new implementations: irregular shape
«New» challenges 1(2)

- Environmental data mining (to be continued).
- Fractal data mining. Intrinsic dimension estimation. Dimensionality reduction, features selection/extraction
- Scaling: dimensionality, number (BIG) & complexity of data
- Active learning and MNO
- From dependencies to cause-effect relationships
- Multi-task learning (learning many “relevant” tasks)
- Uncertainties!!! Risks and extremes
- (Geo)Visual analytics
- (Geo)Manifold learning. Space-embedded networks
- Integration of science-based and data-driven models. EDSS

**EDUCATION: courses, workshops, tutorials, reviews, new books, software tools!!!**
Computational sustainability:
“an interdisciplinary field that aims to apply techniques from computer science, information science, operations research, applied mathematics, and statistics for balancing environmental, economic, and societal needs for sustainable development”
Thank you for your attention!

Prof. M. Kanevski