UNIL | Université de Lausanne

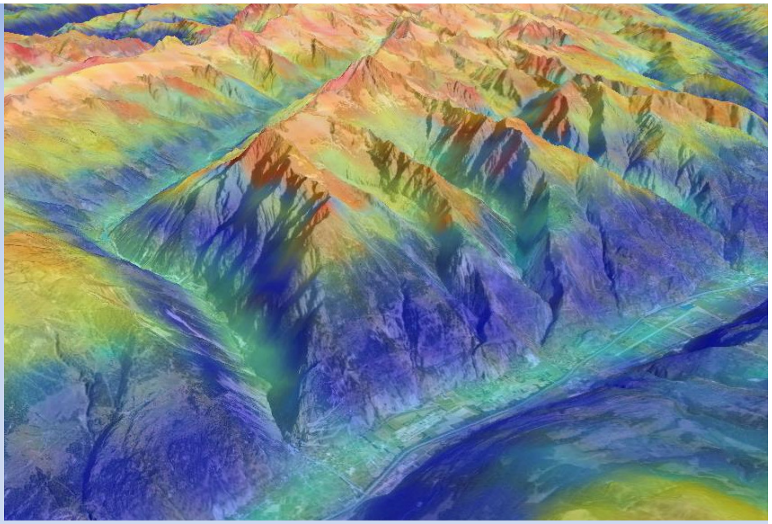
Advanced Analysis of Environmental Data Using Machine Learning

Mikhail Kanevski

Institute of Earth Surface Dynamics, University of Lausanne, Switzerland

Mikhail.Kanevski@unil.ch

Main Topics



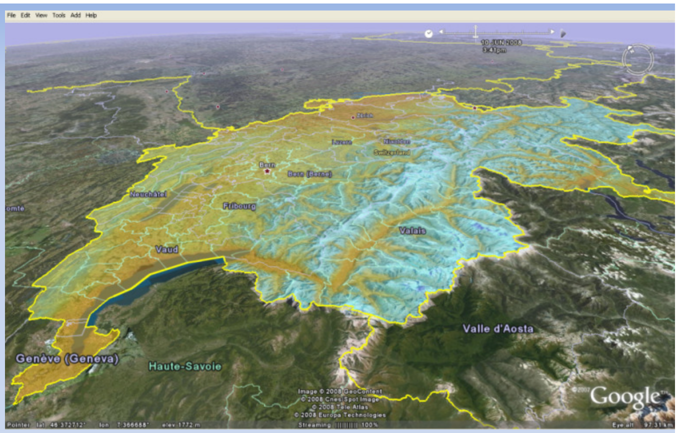
- Introduction. Environmental data
- Generic methodology
- Machine Learning for Environmental Data Modelling and Visualization
- Challenges
- Conclusions

Our data:

- *small, medium and big data*
- *multi-scale*
- *multivariate*
- *uncertain*
- *nonhomogeneous*
- *high dimensionality*
- *nonlinearity*
- *complexity*

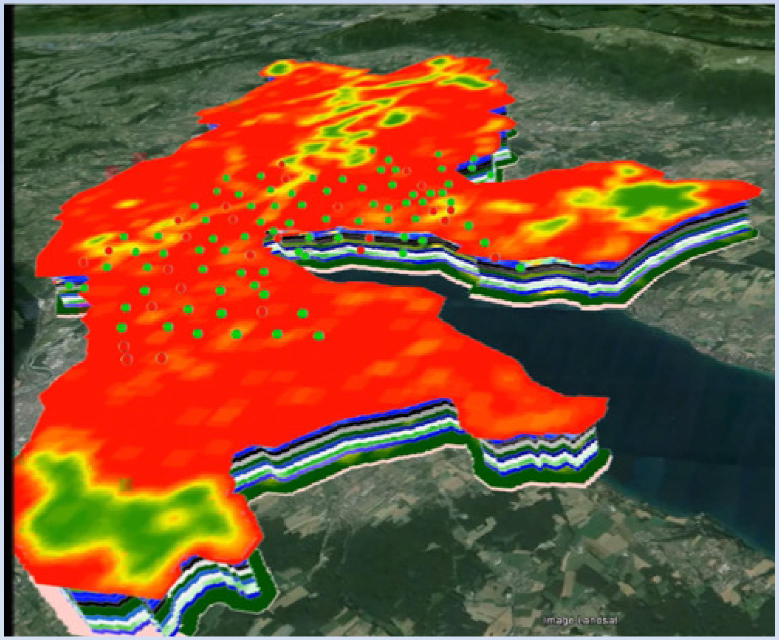
- *Integration of data and science-based models*

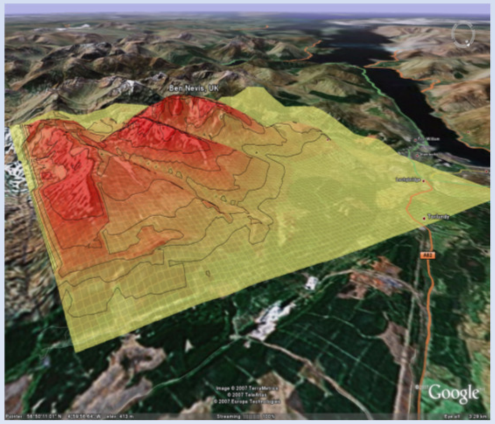
….

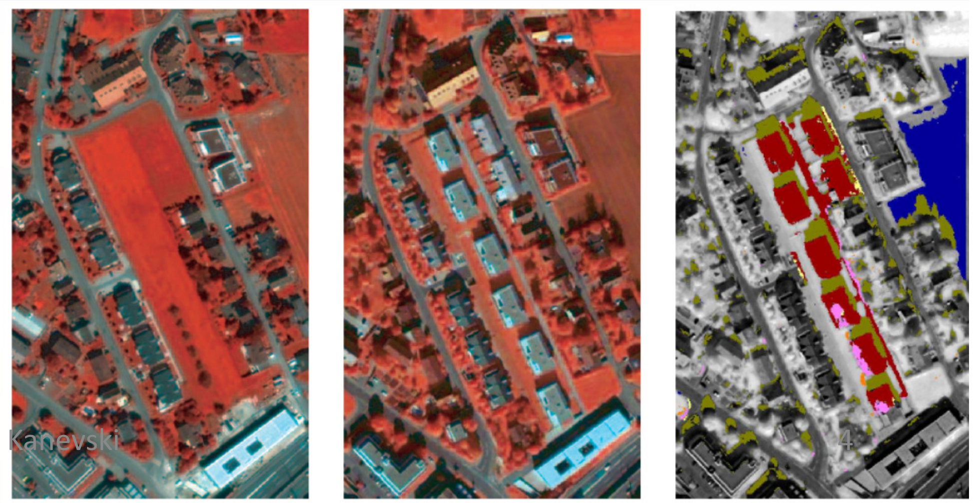


Cases Studies and Dimensionality

- *Monthly wind fields >13d,*
- *avalanches > 40d,*
- *landslides >18d,*
- *permafrost >20d,*
- *city pollution >50,*
- *remote sensing >100…*







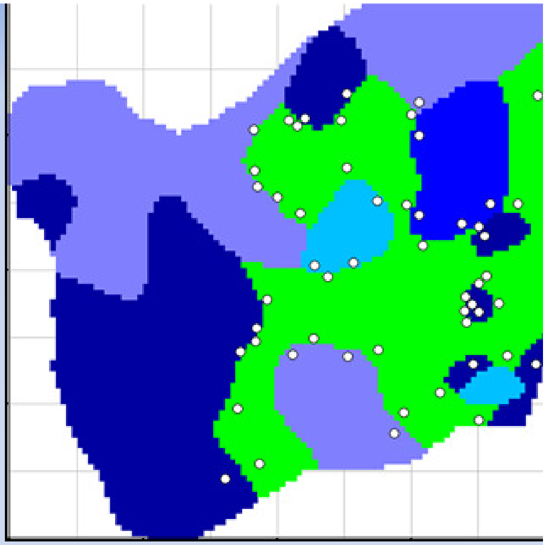
ML algorithms we have used for the environmental applications:

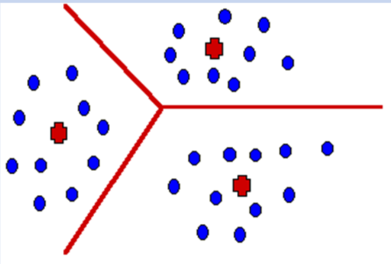
- Artificial neural networks of different architectures: Multilayer Perceptrons, Radial Basis Function Networks, General Regression Neural Networks, Probabilistic Neural Networks, Self-Organizing Maps, MDN, GMM,…

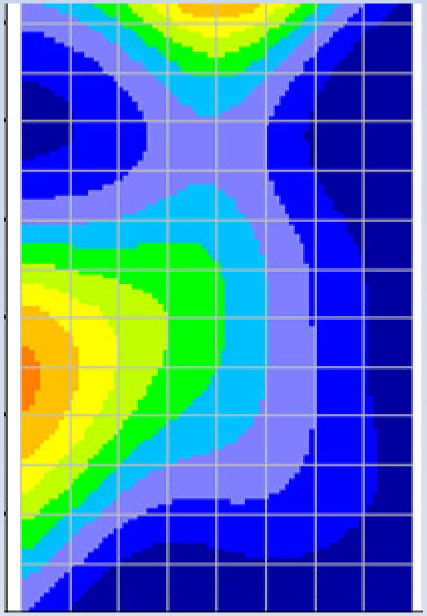
- Random Forests, Ensemble Learning

- Support Vector Machines; Support Vector Regression and many other Kernel-based models

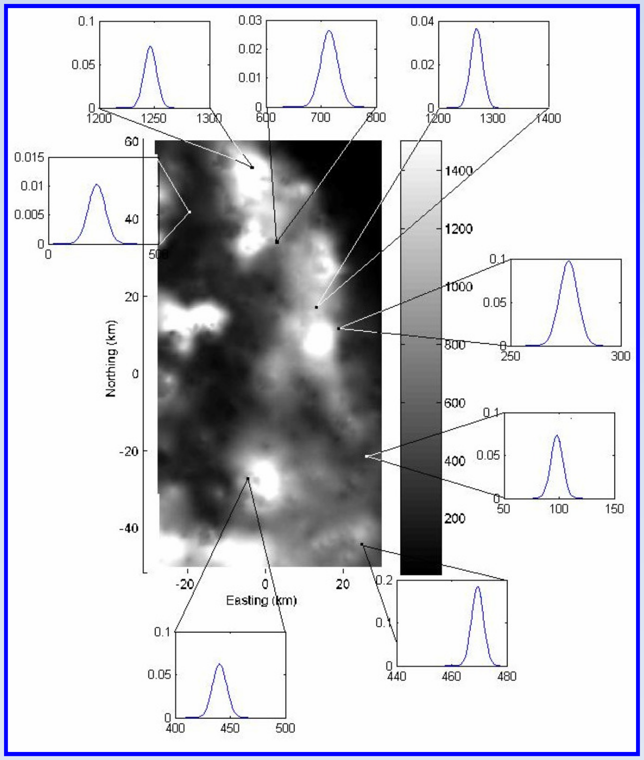
Major fundamental questions for data-driven modelling







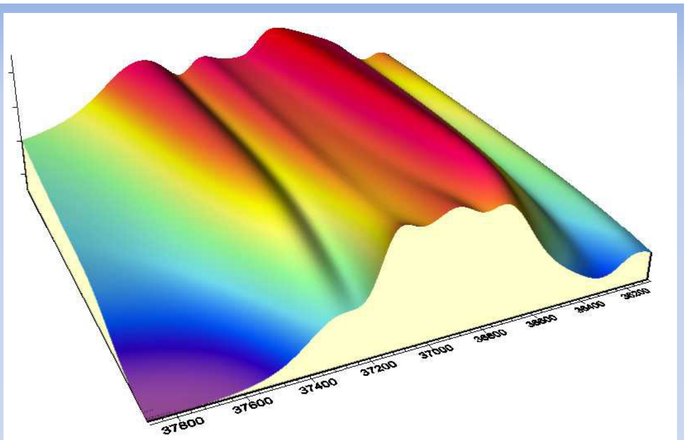
- Clustering
- Classification
- Regression
- Pdf modeling (risk analysis)



Approaches to geospatial environmental data

- Classical geostatistics: predictions/simulations
- Multi-point geostatistics
- Bayesian geostatistics, BME
- Machine Learning

Monitoring networks: clustering, preferential sampling, design, optimization…



<u>Learning</u> *of spatio-temporal data in terms of patterns/structures:*

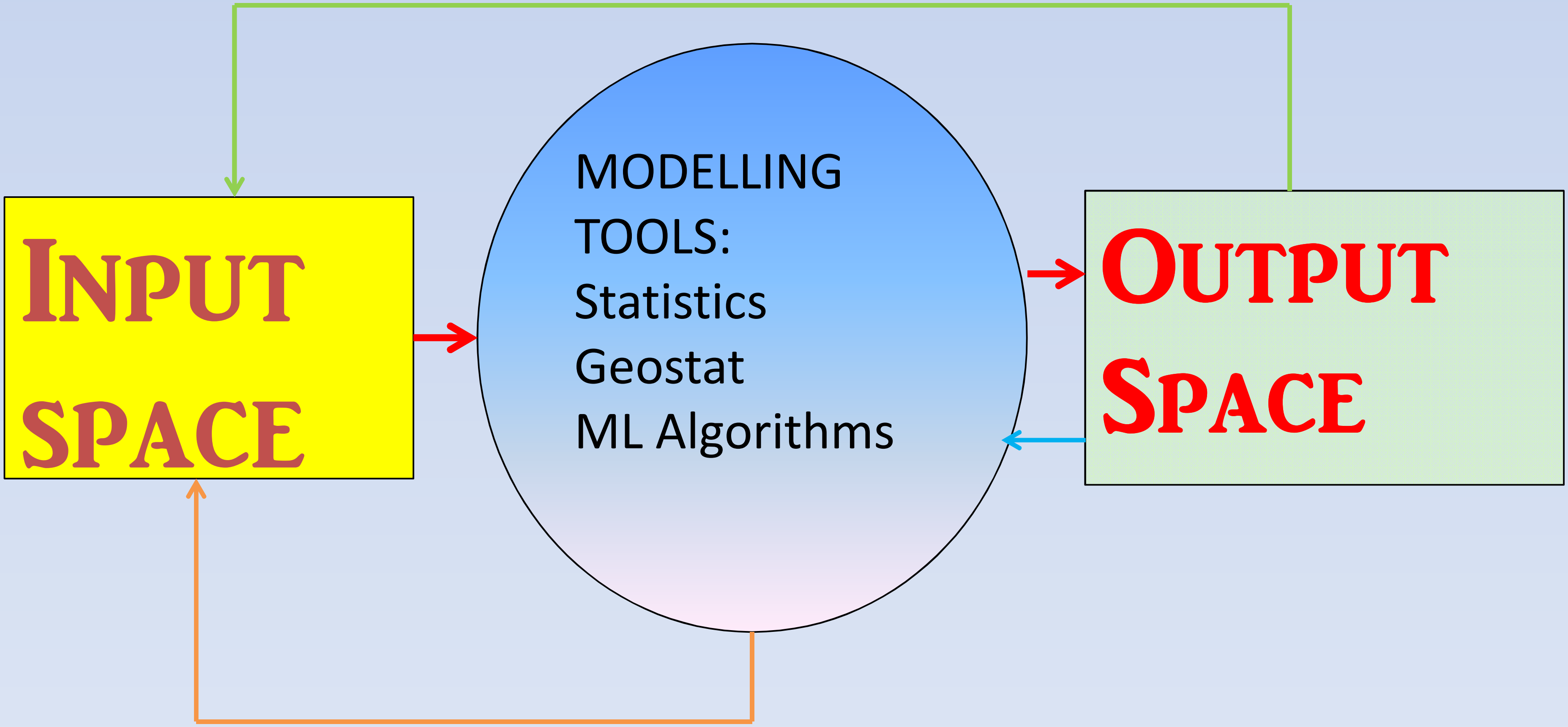
- pattern recognition,

- pattern modelling,

- pattern predictions

Generic Modelling Task

(classification, regression, density modelling)



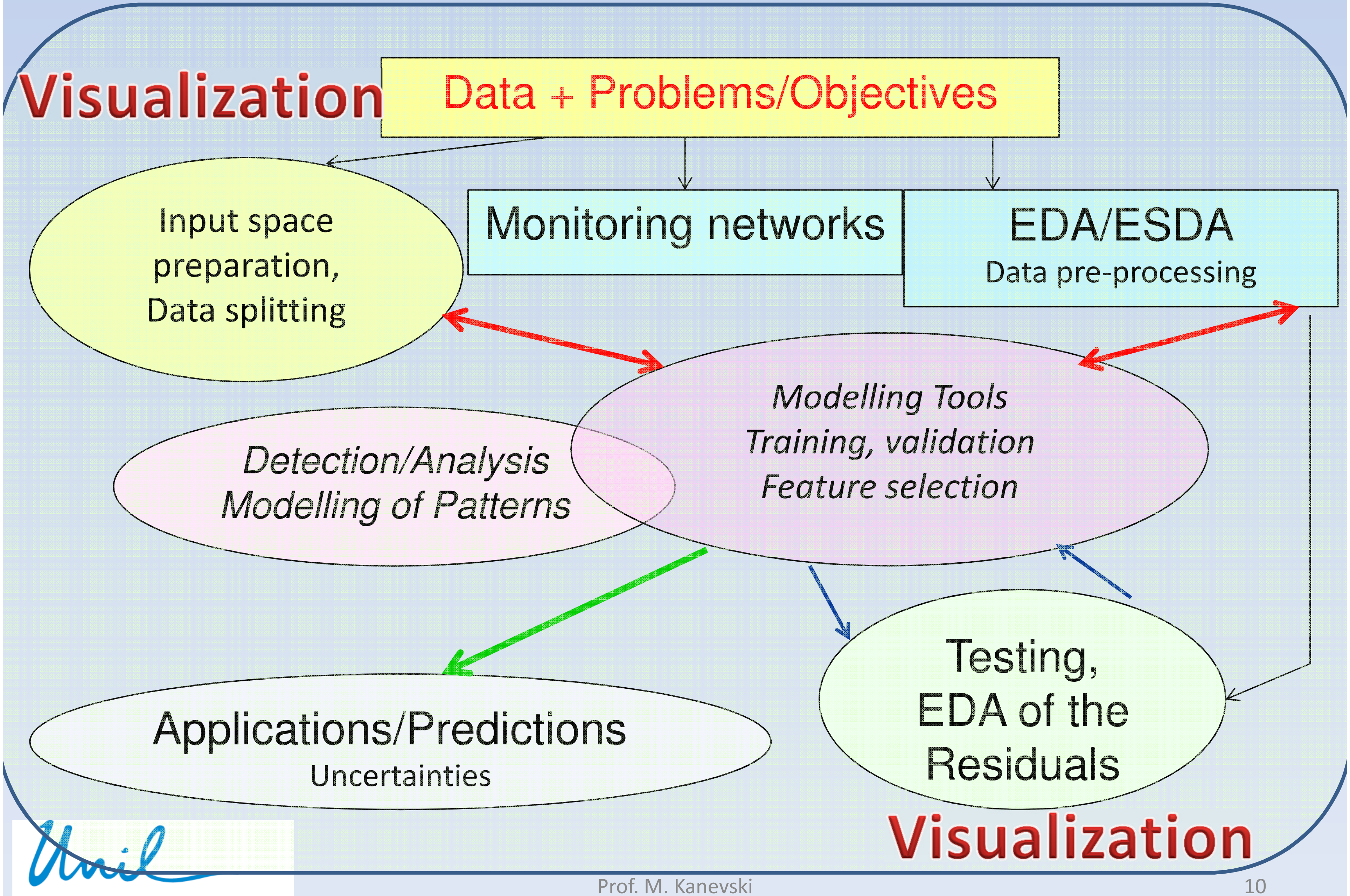
INPUT SPACE

MODELLING TOOLS:
Statistics
Geostat
ML Algorithms

OUTPUT SPACE

UNIL | Université de Lausanne

Generic methodology



Visualization

Data + Problems/Objectives

Input space preparation, Data splitting

Monitoring networks

EDA/ESDA

Data pre-processing

Modelling Tools

Training, validation

Feature selection

Detection/Analysis

Modelling of Patterns

Testing, EDA of the Residuals

Applications/Predictions

Uncertainties

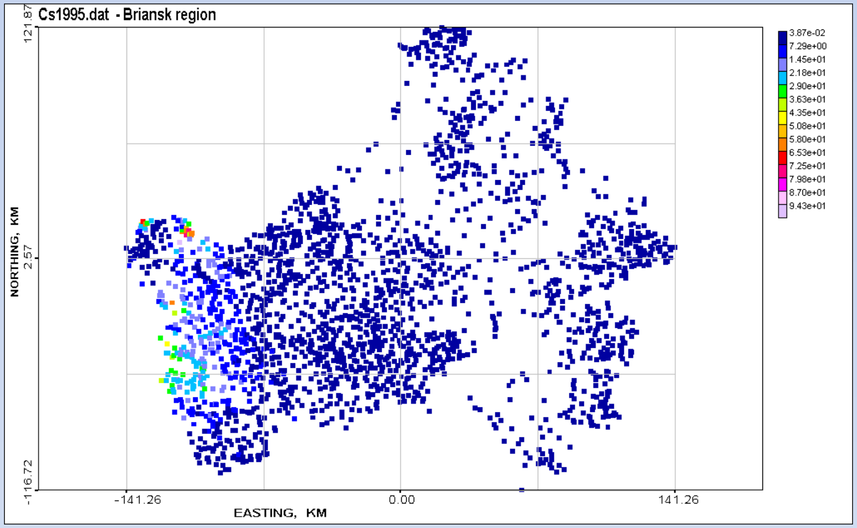
Visualization

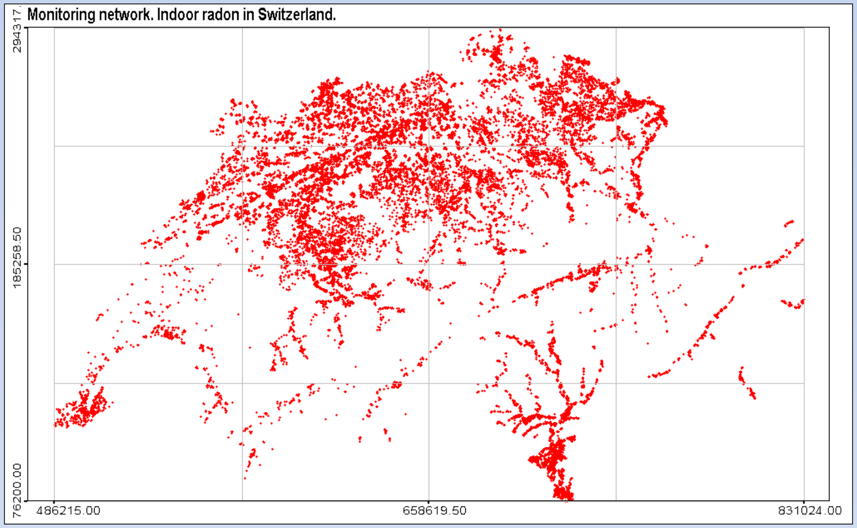
First important question. Monitoring networks:

Clustering and Preferential Sampling

Spatial and Dimensional Resolutions

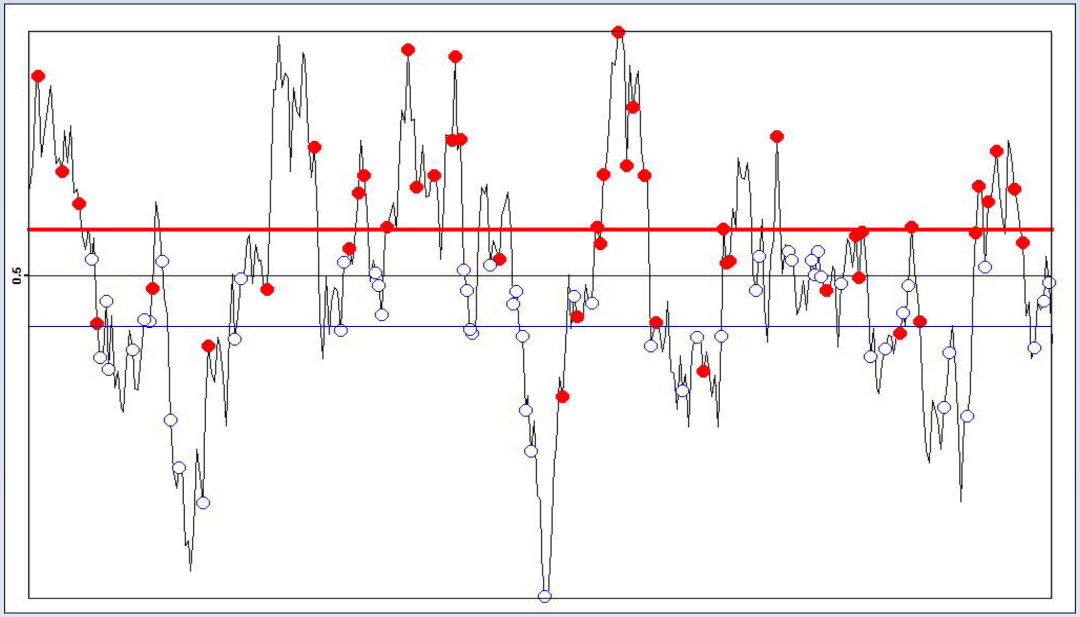
Monitoring networks





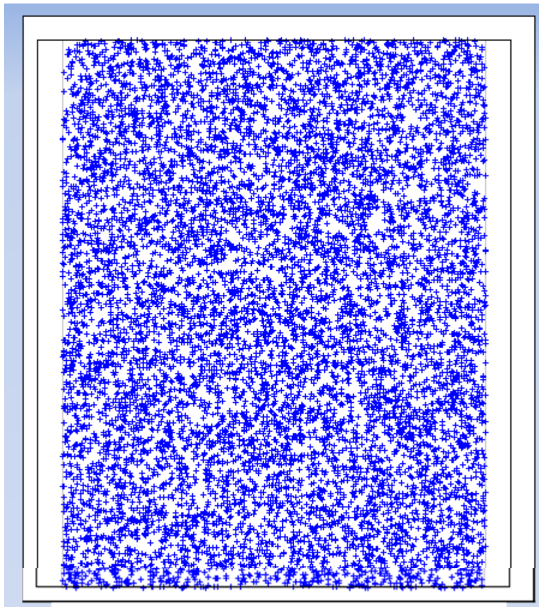
Measures of clustering:

Topological
Statistical
Fractal

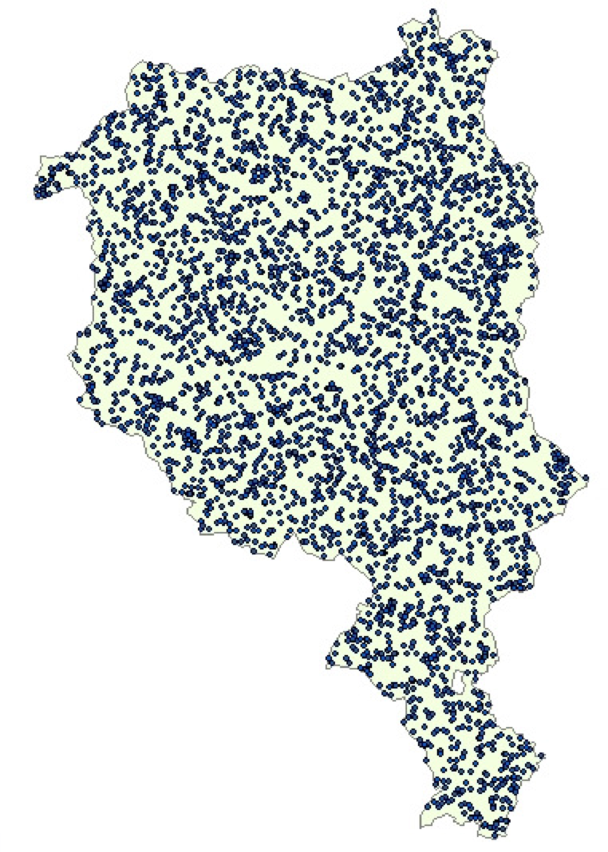


UNIL | Université de Lausanne

Validity domains: A- rectangular, B-admin, C- forest (points were randomly generated). d is estimated by a sand-box counting method



A: d=1.96

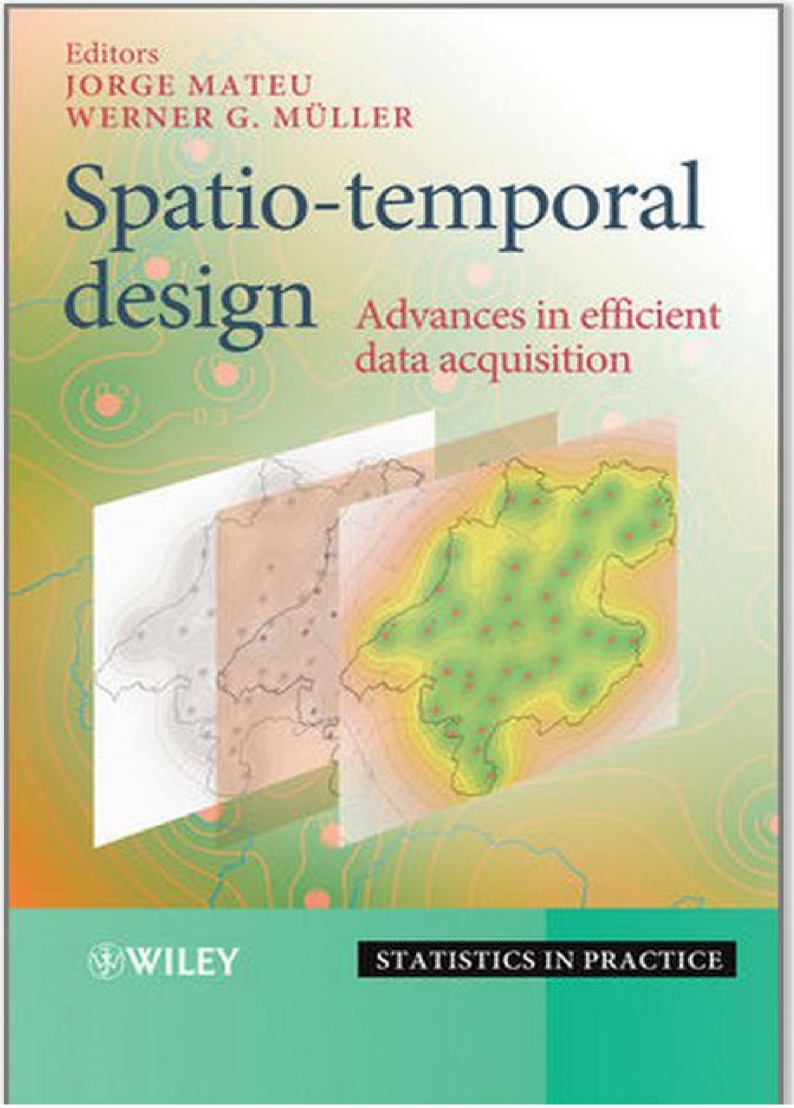


B: d=1.88

C: d=1.75

UNIL | Université de Lausanne

More on monitoring networks in:

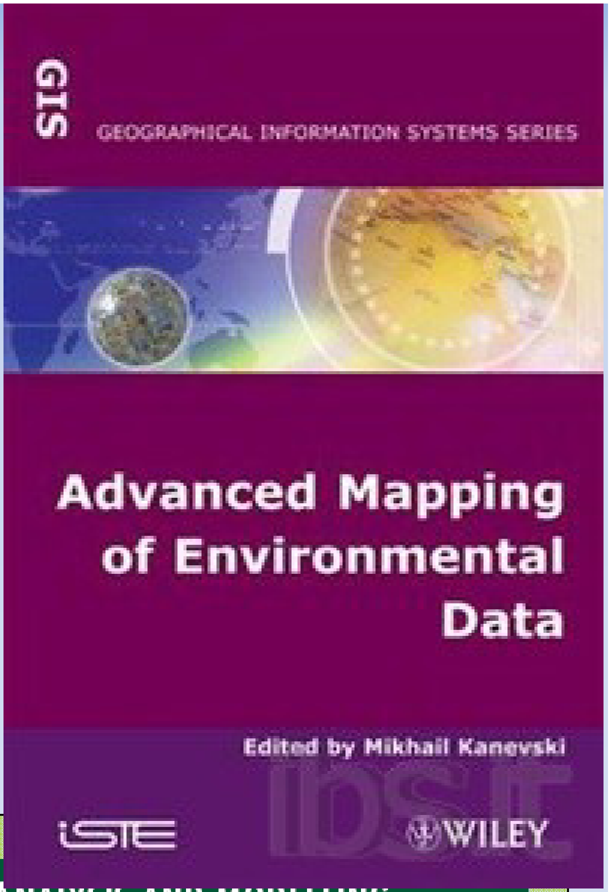


Editors
JORGE MATEU
WERNER G. MÜLLER

Spatio-temporal design
Advances in efficient data acquisition

WILEY

STATISTICS IN PRACTICE



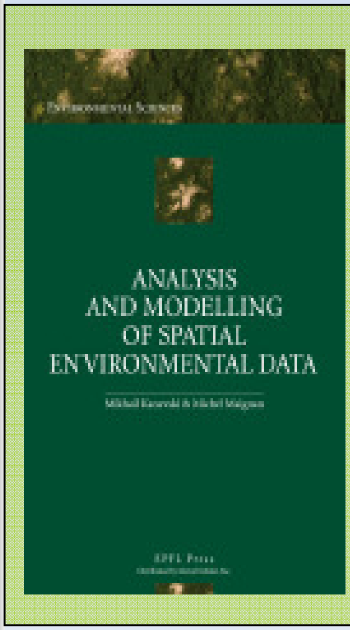
GIS GEOGRAPHICAL INFORMATION SYSTEMS SERIES

Advanced Mapping of Environmental Data

Edited by Mikhail Kanevski

ISTE

WILEY



Environmental Sciences

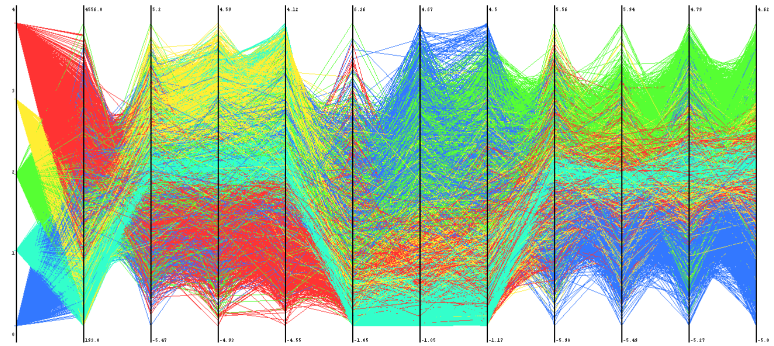
ANALYSIS
AND MODELLING
OF SPATIAL
ENVIRONMENTAL DATA

Mikhail Kanevski & Michel Maignan

EPFL Press

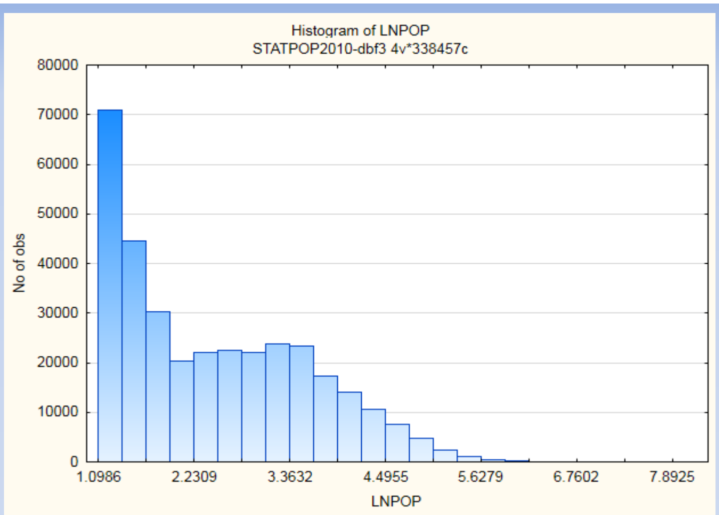
ANALYSIS AND MODELLING
OF SPATIAL ENVIRONMENTAL DATA

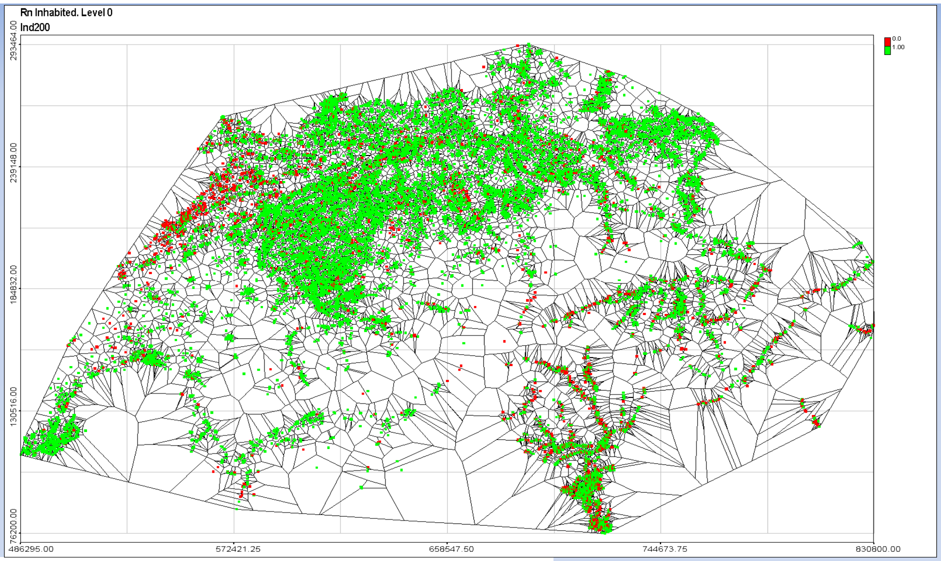
Mikhail Kanevski, Michel Maignan

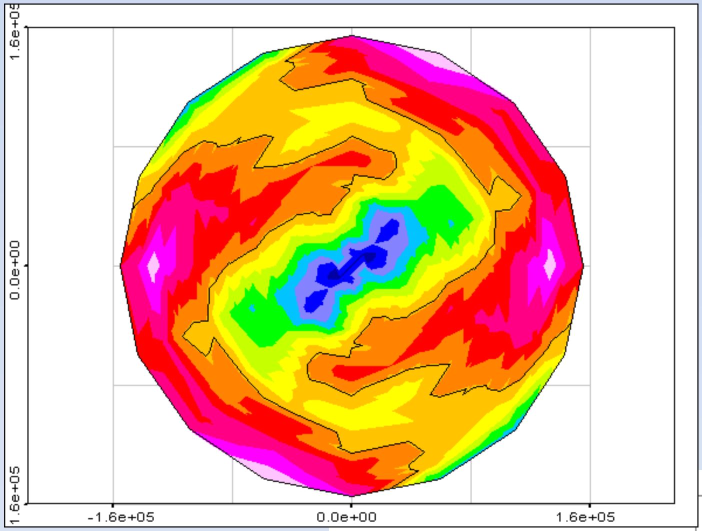


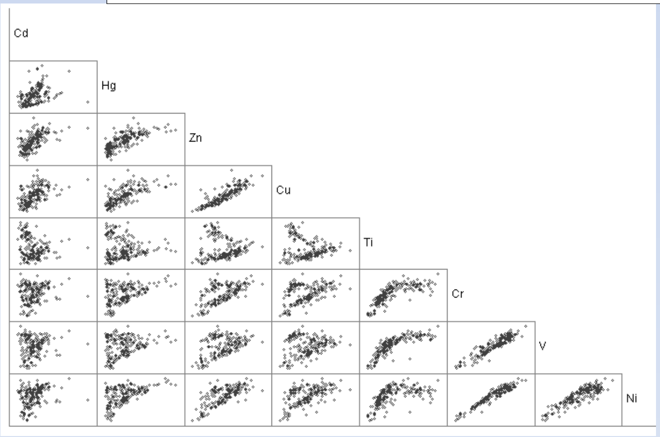
Advanced Exploratory Data Analysis and VISUALIZATION

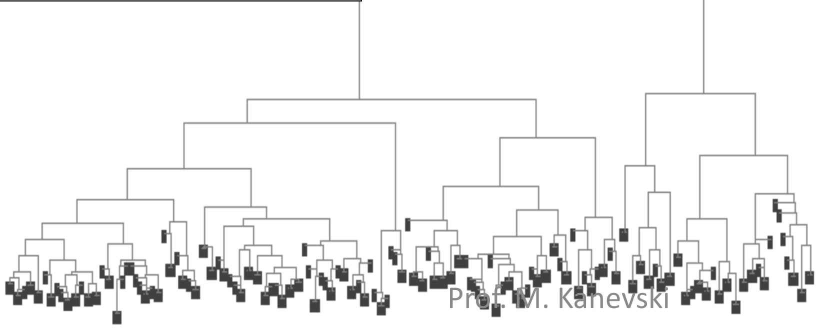
UNIL | Université de Lausanne

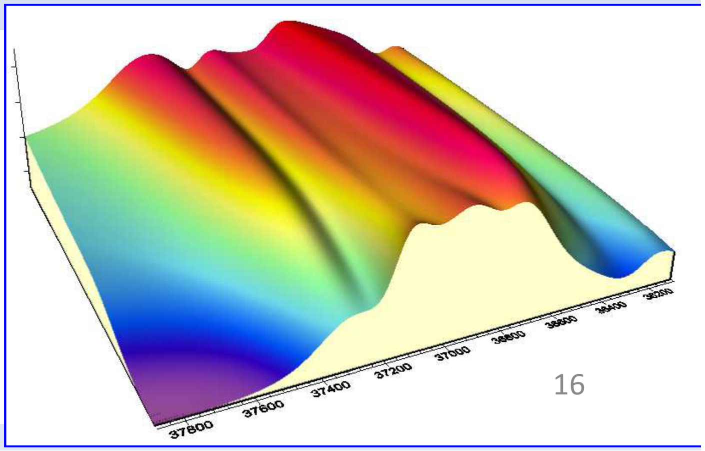

Histogram of LNPOP
STATPOP2010-dbf3 4v*338457c
No of obs
80000
70000
60000
50000
40000
30000
20000
10000
0
1.0986
2.2309
3.3632
4.4955
5.6279
6.7602
7.8925
LNPOP


Rn Inhabited. Level 0
Ind200
0.0
1.00
293464.00
239148.00
184832.00
130516.00
76200.00
486295.00
572421.25
658547.50
744673.75
830800.00

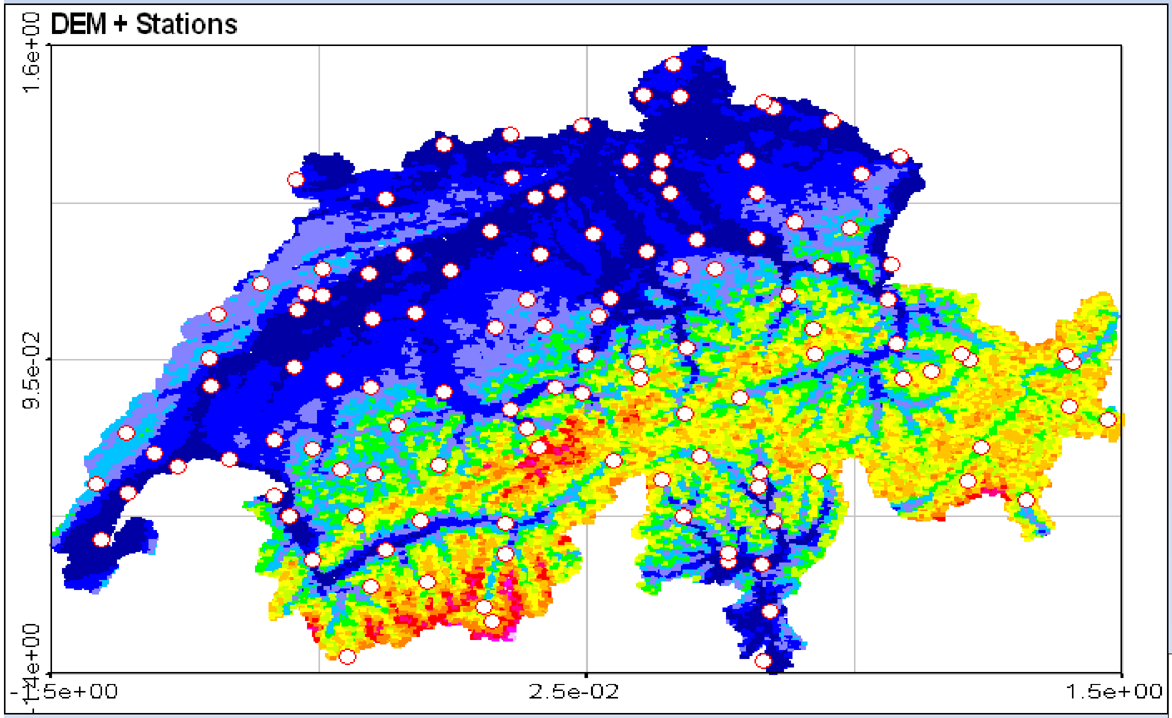

1.6e+05
0.0e+00
-1.6e+05
-1.6e+05
0.0e+00
1.6e+05

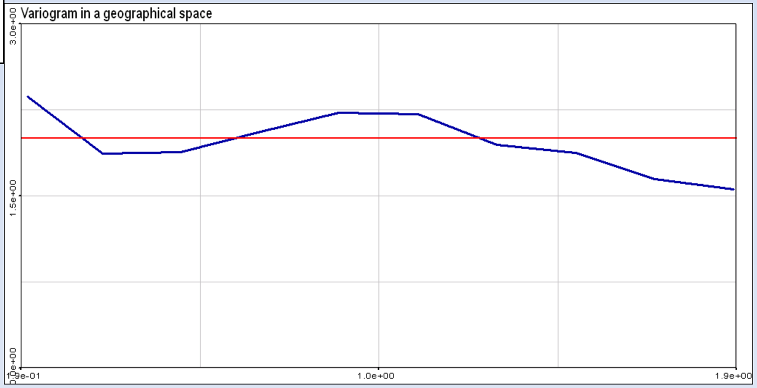

Cd
Hg
Zn
Cu
Ti
Cr
V
Ni




37800
37600
37400
37200
37000
36800
36600
36400
36200

Data: Monthly wind speed in Switzerland
(data prepared by S. Robert, ETHZ)





UNIL | Université de Lausanne

Embedding of raw data into a feature space

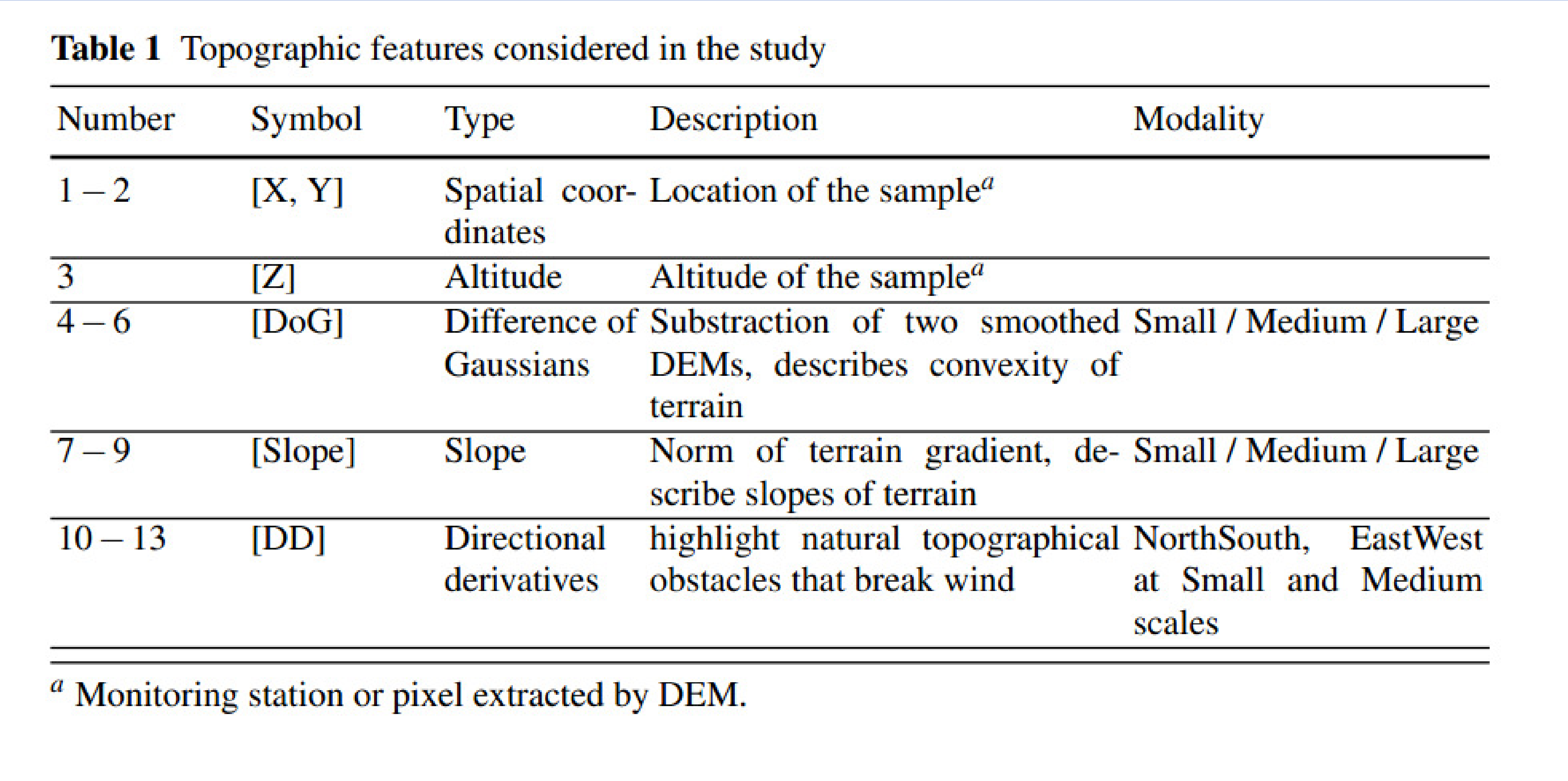
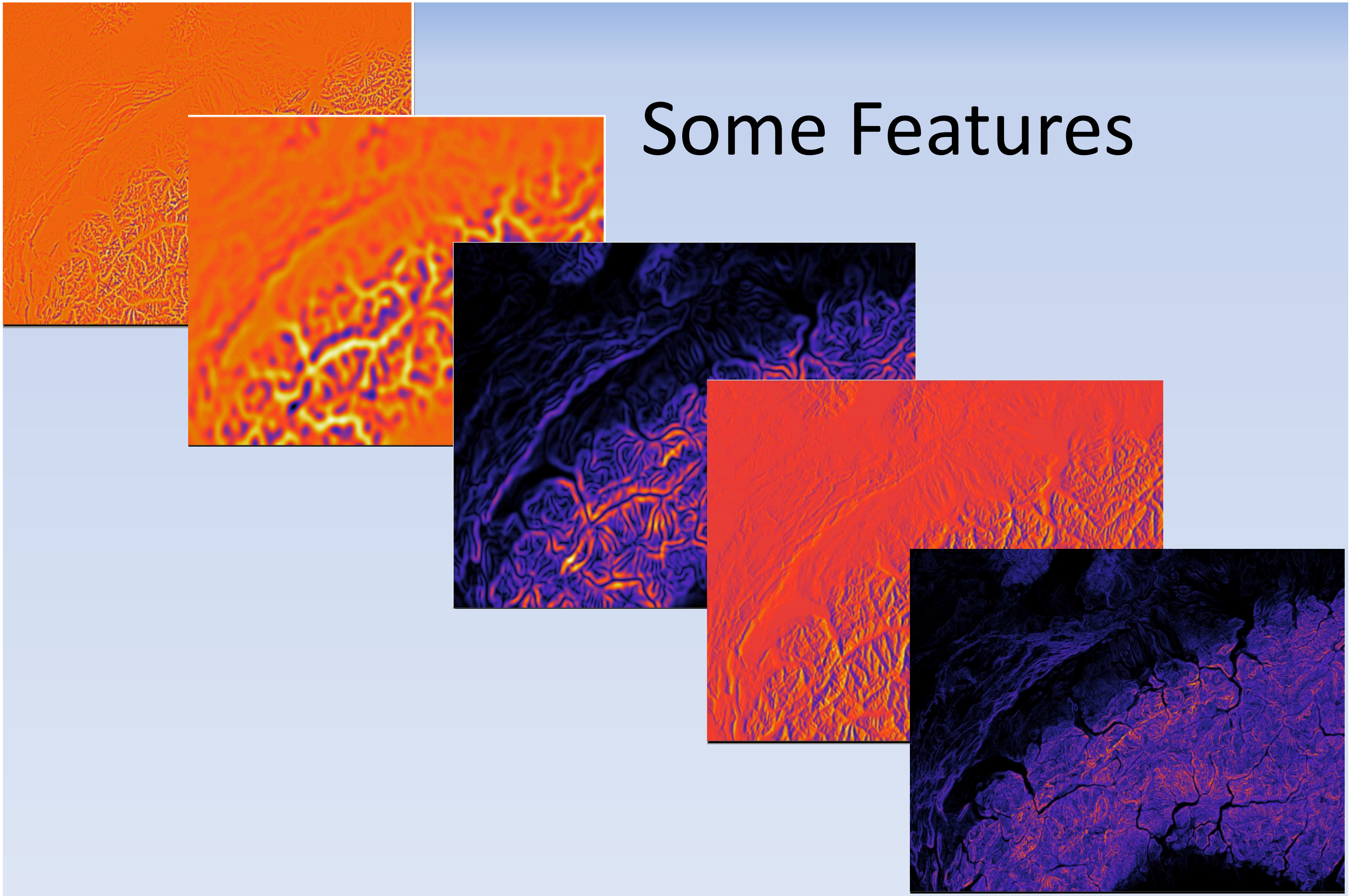


Table 1 Topographic features considered in the study

Number	Symbol	Type	Description	Modality
1 – 2	[X, Y]	Spatial coordinates	Location of the sample[a]	
3	[Z]	Altitude	Altitude of the sample[a]	
4 – 6	[DoG]	Difference of Gaussians	Substraction of two smoothed DEMs, describes convexity of terrain	Small / Medium / Large
7 – 9	[Slope]	Slope	Norm of terrain gradient, describe slopes of terrain	Small / Medium / Large
10 – 13	[DD]	Directional derivatives	highlight natural topographical obstacles that break wind	NorthSouth, EastWest at Small and Medium scales

[a] Monitoring station or pixel extracted by DEM.

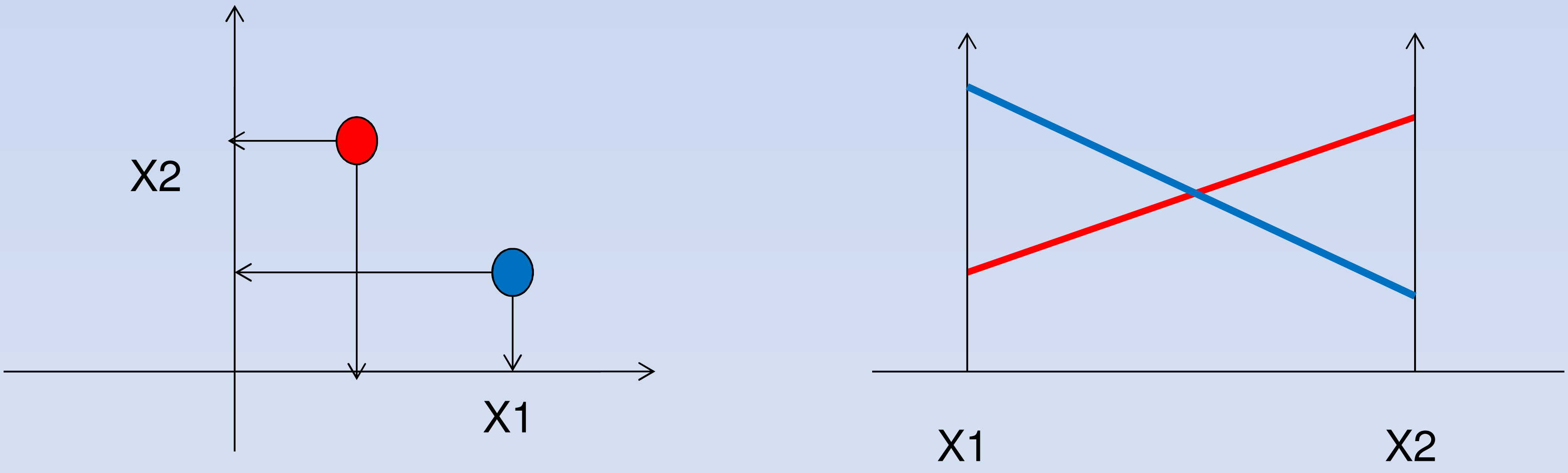
Some Features



Parallel coordinates (A. Inselberg)

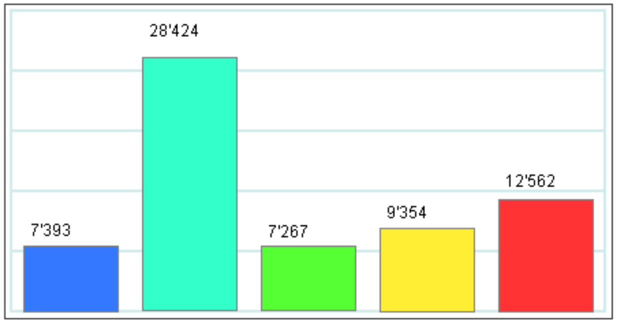
(very popular in high-dim data visualization)

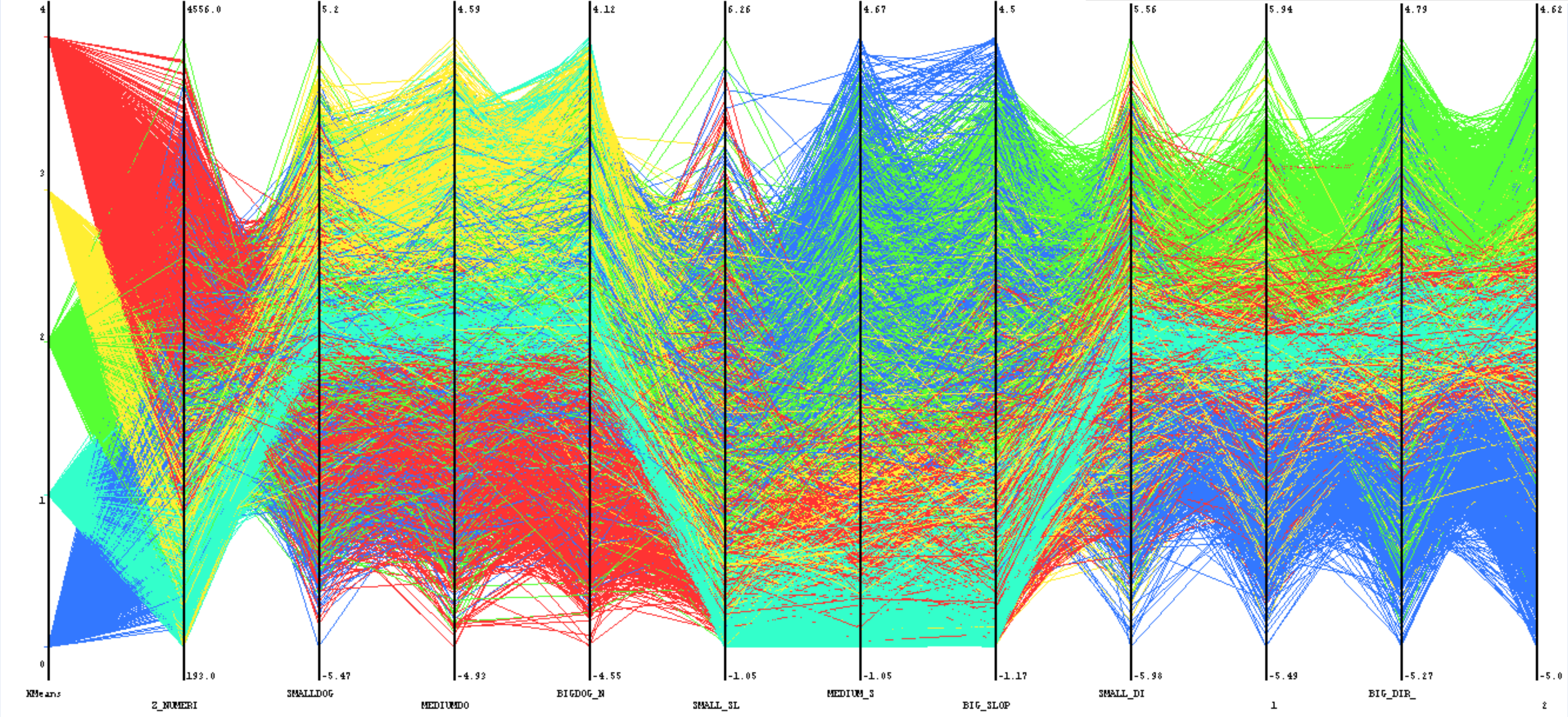
From orthogonal coord. to parallel coord.



The same but in d-dimensional space. Usually scaling is done between min-max for each coordinate.

11d input space: 5 classes



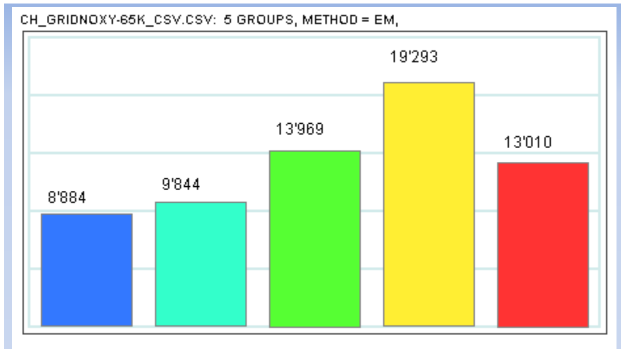


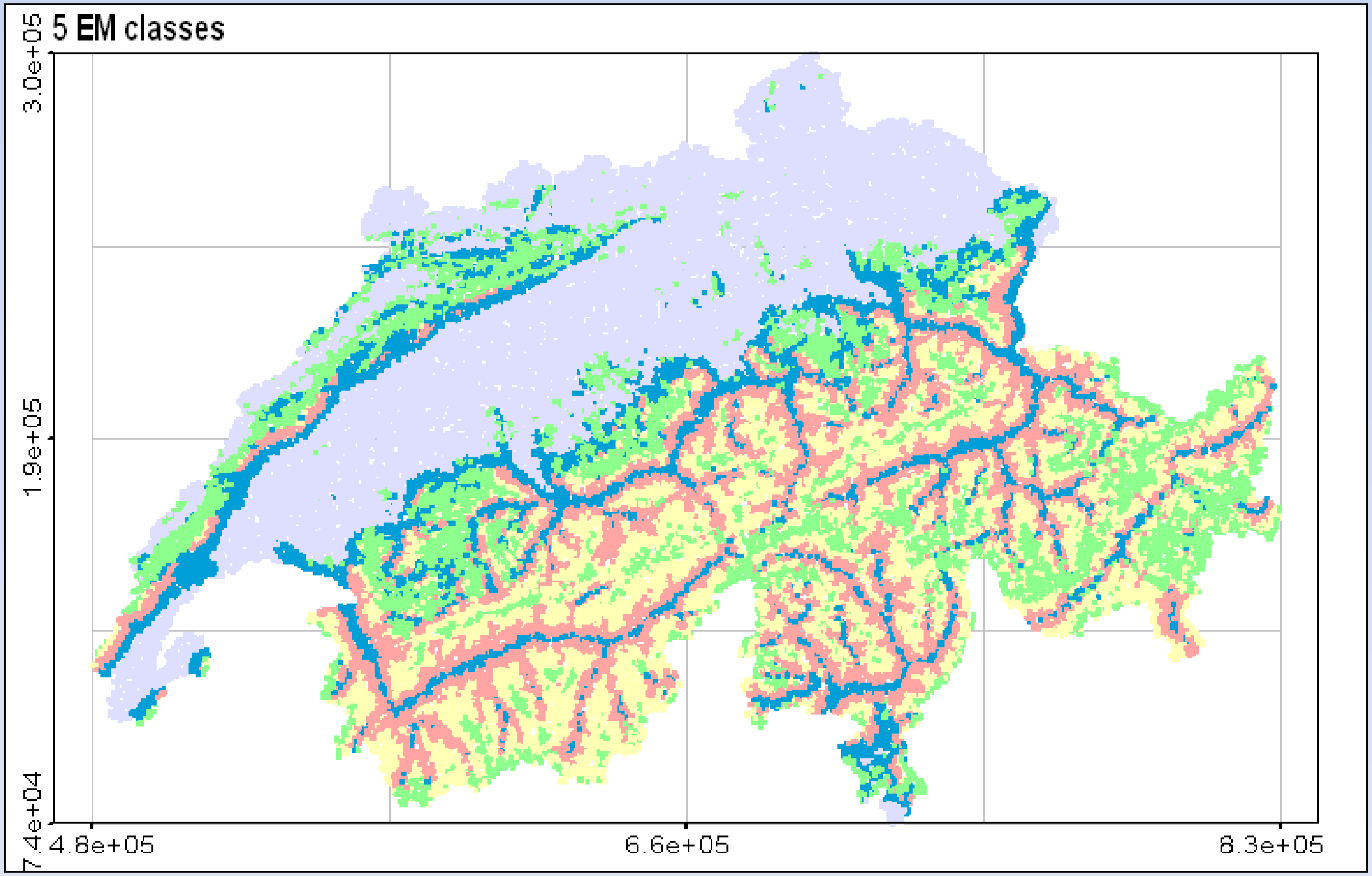
The same but in d-dimensional space. Usually scaling is done between min-max for each coordinate.

Number of possible different orderings = [(d+1)/2]!!!!

UNIL | Université de Lausanne

EM algorithm 5 classes

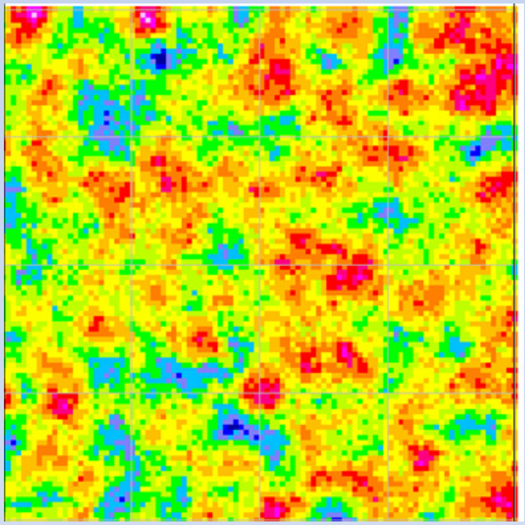




UNIL | Université de Lausanne

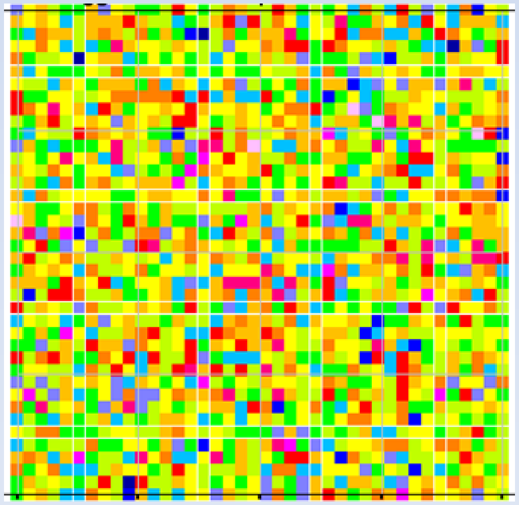
“General formula” of data (often used)

DATA=



Information/Structures/
Patterns

+



Noise

Information and noise

(very difficult problem)

Information/Structure/Pattern:

- To be defined, quantified, modelled (data/task/objective dependent)

Noise (difficult problem). To be estimated:

- *Before (independently of) modelling – the best;*
- *As a result (during) of modelling*

UNIL | Université de Lausanne

Gamma test

Antonia J. Jones. New tools in non-linear modelling and prediction. Computational Management Sciences 1: 109–149 (2004)

Suppose we are given a set of input-output data

$$\{x_1(i),\dots,x_m(i),y_i\}=\{(\mathbf{x}_i,y_i);1\le i\le M\}$$

if the underlying relationship is of the form

$$y=m(x_1,x_2,\dots x_m)+r$$

Gamma test

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M}\left[\mathbf{x}_{N(i,k)} - \mathbf{x}_i\right]^2$$

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M}\left\{y_{N(i,k)} - y_i\right\}^2$$

$$\gamma_M(k) = \Gamma + A\delta_M(k)$$

$$\Gamma \to Var(r) \ \ \text{when } M \ \to \infty$$

Calculating the regression line gradient can also provide helpful information on the complexity of the system under investigation (a steeper gradient indicates a model of greater complexity)

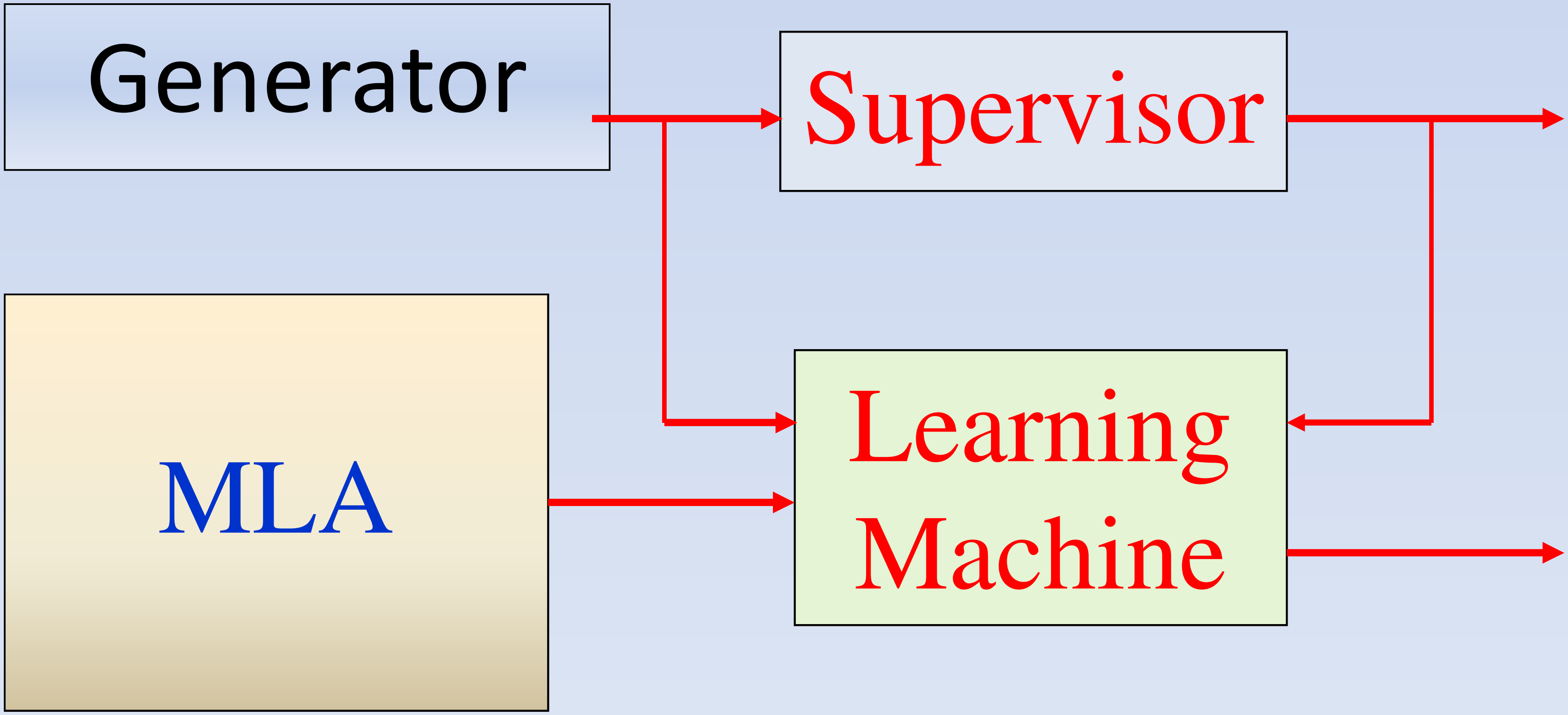
Unil
UNIL | Université de Lausanne

ML algorithms (WHY?)

- Universal modelling tools
- Nonlinear
- Robust
- Data adapted, data driven
- Easy data and knowledge integration
- Good for high dimensional spaces
- Good generalization properties

- *Uncertainties characterisation*
- *Interpretability*

A Generic Model of Supervised Learning from Data/Examples



Generator

Supervisor

MLA

Learning Machine

UNIL | Université de Lausanne

The Problem of Risk Minimization

*In order to choose the best available model to the supervisor's response, one measure the **LOSS** or discrepancy $L(y,f(x,\alpha))$ between the response y of the supervisor to a given input x and the response $f(x,\alpha)$ provided by the Loss Measure*.

Most of the ML problems are formulated in terms of Empirical Risk Minimization or Structural Risk Minimization principles

GRNN

General Regression

Neural Networks

UNIL | Université de Lausanne

Non-parametric kernel regression

Consider a non-linear regression problem, described by a model whose observable output z_i in response to an input vector $\boldsymbol{x}_i$ is defined by

$$z_i = f(\boldsymbol{x}_i) + \varepsilon_i \qquad i = 1, 2 \ldots, N$$

GRNN: short theory

Best prediction (Risk = MSE)

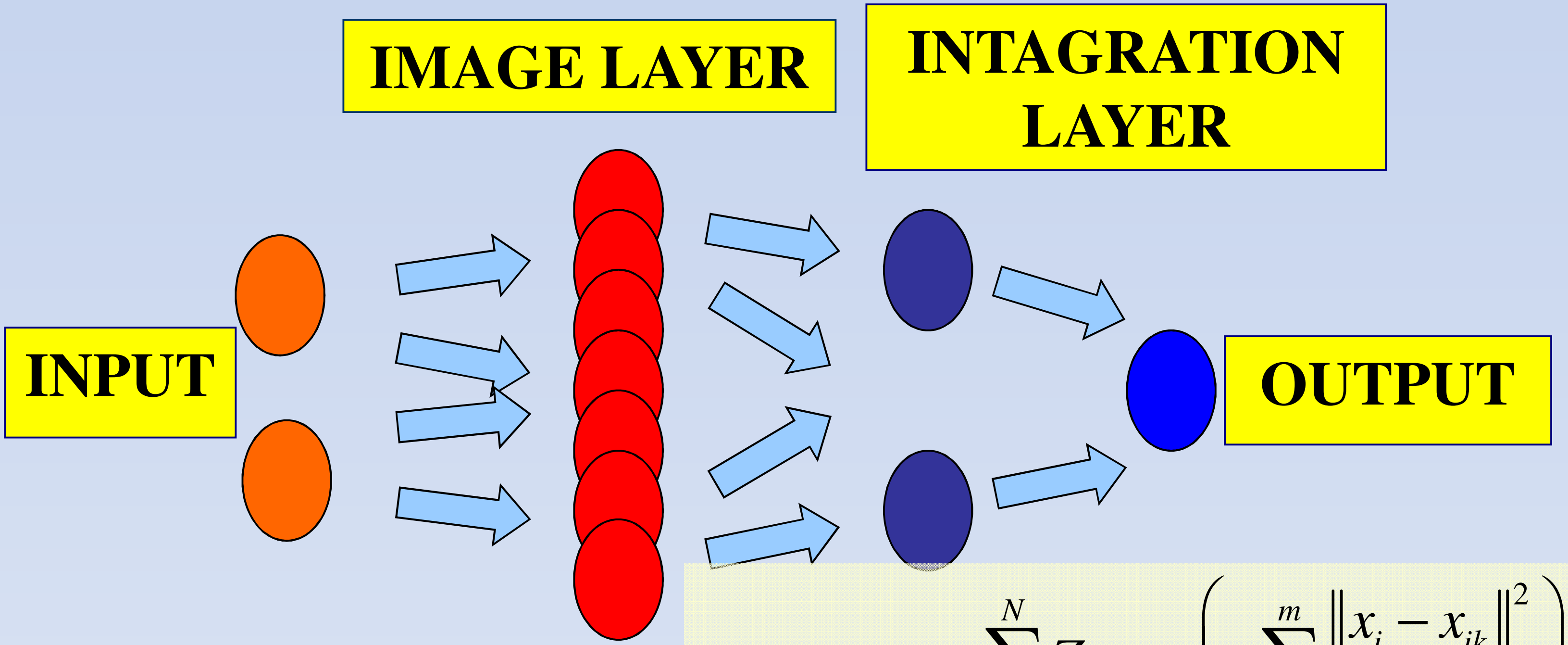
of f(x) is the conditional mean value :

$$f(\hat{\mathbf{x}}) = E\left\langle z(\mathbf{x}) \middle| \mathbf{x} \right\rangle] = \frac{\int_{-\infty}^{+\infty} z(\mathbf{x})\, p(\mathbf{x}, z)\, dz}{\int_{-\infty}^{+\infty} p(\mathbf{x}, z)\, dz}$$

where *p(x,y)* is a joint input-output distribution function and is estimated using kernel density estimator

(A)GRNN

(see nonparametric statistics, Nadaraya-Watson estimator).
N – number of data; m – number of features



GRNN estimate at a node D_i from samples Z_i:

$$Z(x_1,\ldots x_m)=\frac{\sum_{k=1}^{N} Z_k \exp\left(-\sum_{i=1}^{m}\frac{\left\|x_i-x_{ik}\right\|^2}{2\sigma_i^{\,2}}\right)}{\sum_{k=1}^{N} \exp\left(-\sum_{i=1}^{m}\frac{\left\|x_i-x_{ik}\right\|^2}{2\sigma_i^{\,2}}\right)}$$

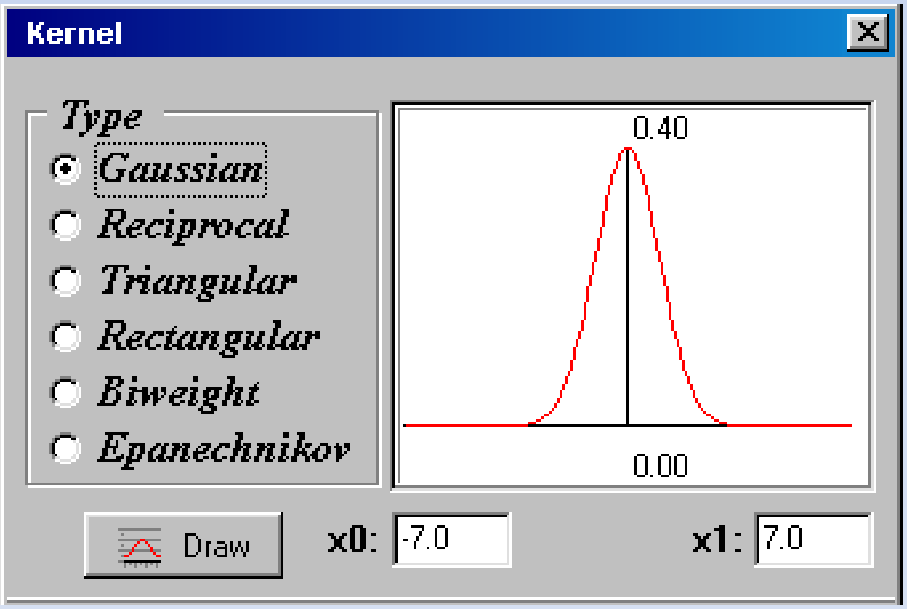
(Presented this way by Specht in 1991)

General Regression Neural Networks (GRNN)

Kernel Type:

The best values of

σ_i – kernel widths, are found with the help of a cross-validation by minimising the mean square error:



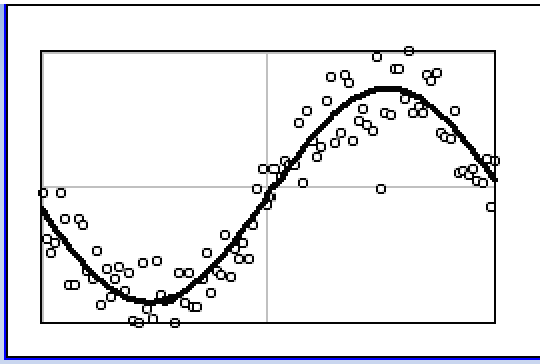
Type
● Gaussian
○ Reciprocal
○ Triangular
○ Rectangular
○ Biweight
○ Epanechnikov

x0: -7.0

x1: 7.0

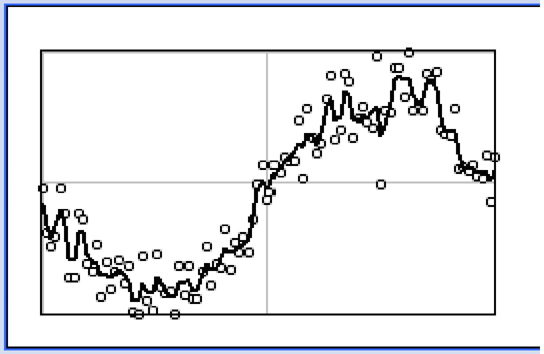
$$\sqrt{(1/N)\sum_i (Z_i^{NET} - Z_i)^2} \rightarrow \min$$

GRNN Training

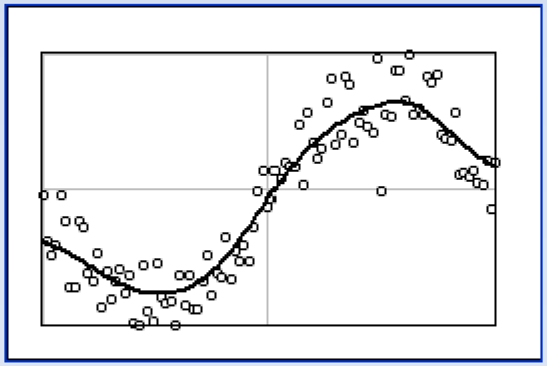


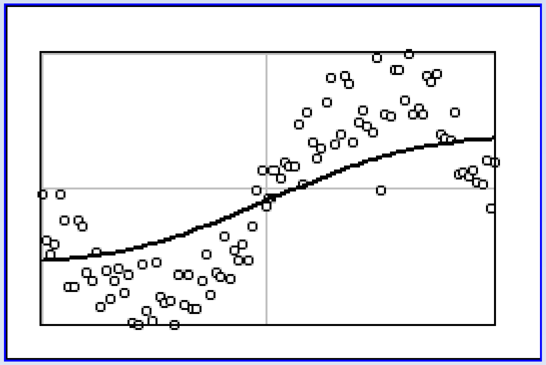
True function

Too small, overfitting



Optimal





Too large, oversmoothing

Some useful properties of GRNN

- When bandwidth is small:

  → nearest neighbour estimator

- When all bandwidths are larger than the region of the study:

  → there is no structure and

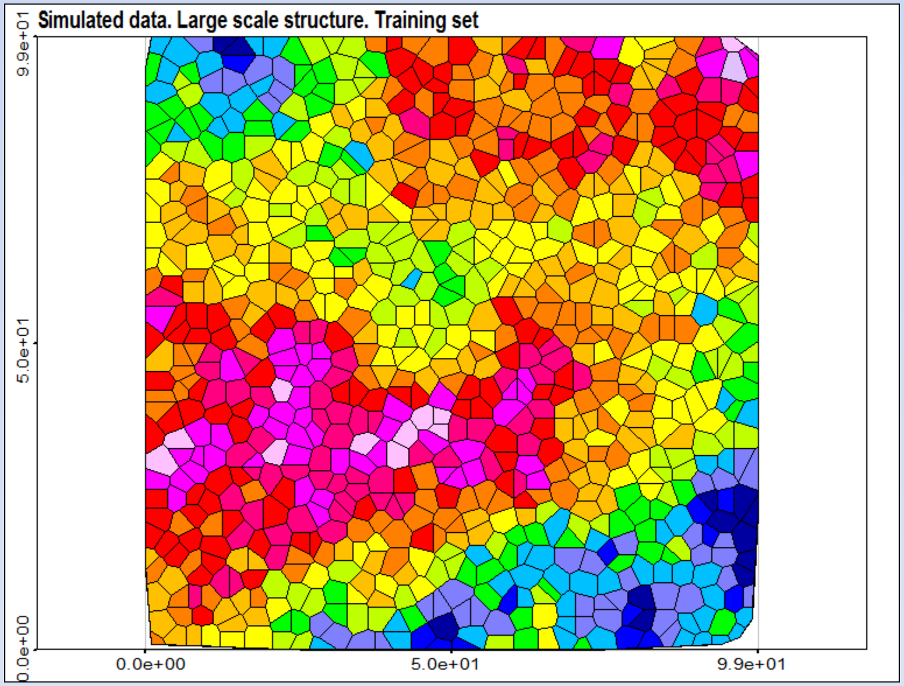
$$\hat{Z}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} Z_i$$

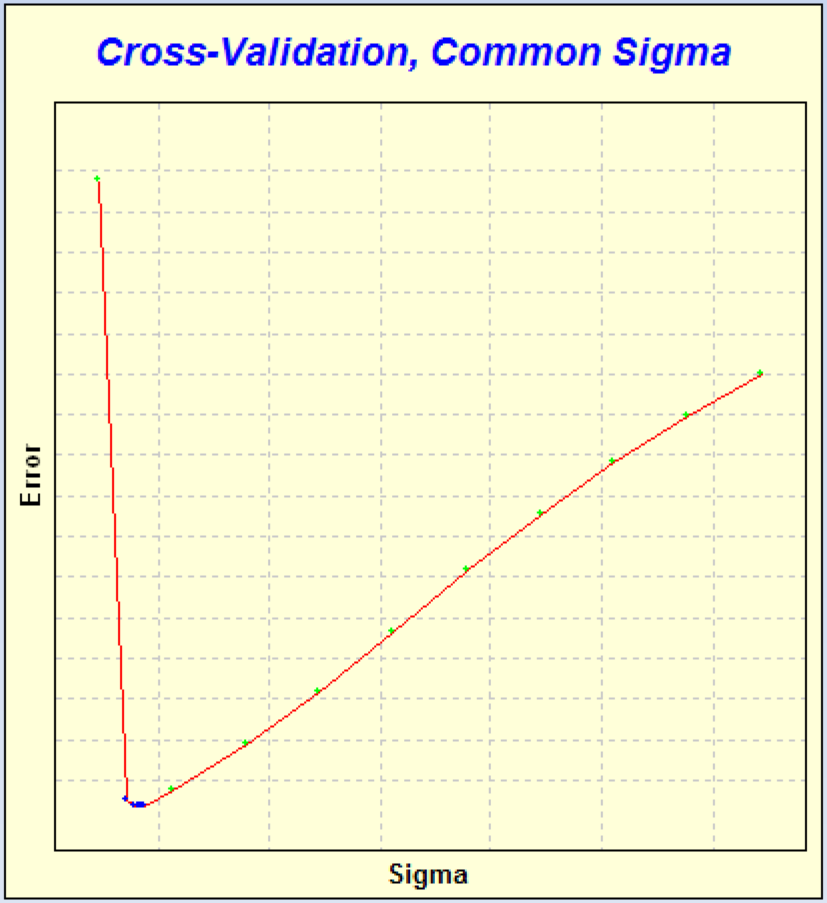
- When bandwidth for some coordinate/variable *i* is large, this coordinate will be filtered out (neglected) automatically:

$$\text{if} \quad \frac{\left\| x_i - x_{in} \right\|^2}{2\sigma_i^{\,2}} \ll 1, \quad \text{then} \quad \exp\left(-\frac{\left\| x_i - x_{in} \right\|^2}{2\sigma_i^{\,2}} \right) \approx 1$$

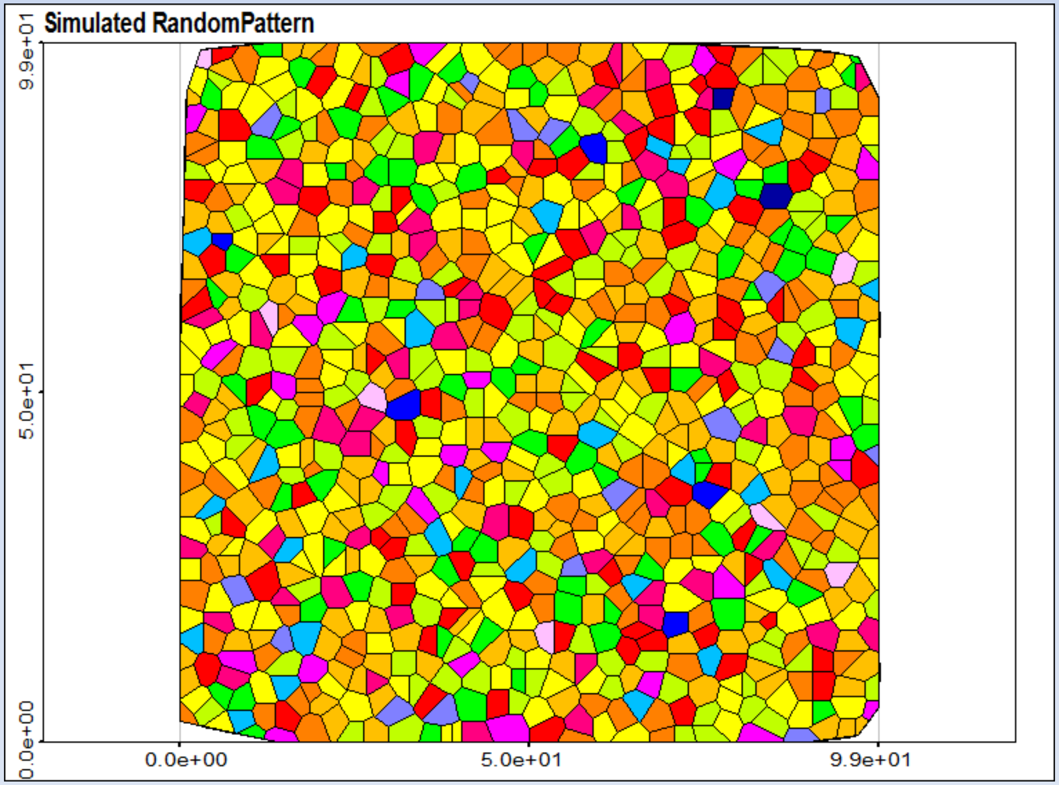
UNIL | Université de Lausanne

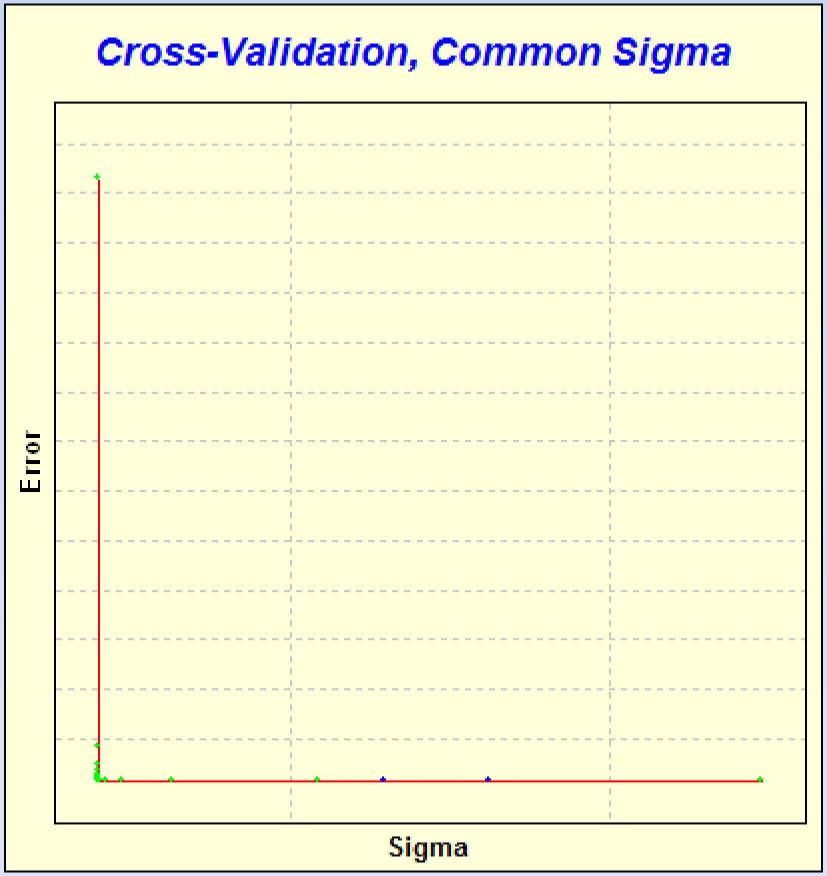
GRNN learning of a pattern (structured data)





GRNN learning of a random pattern (useful to know - to study the residuals!)



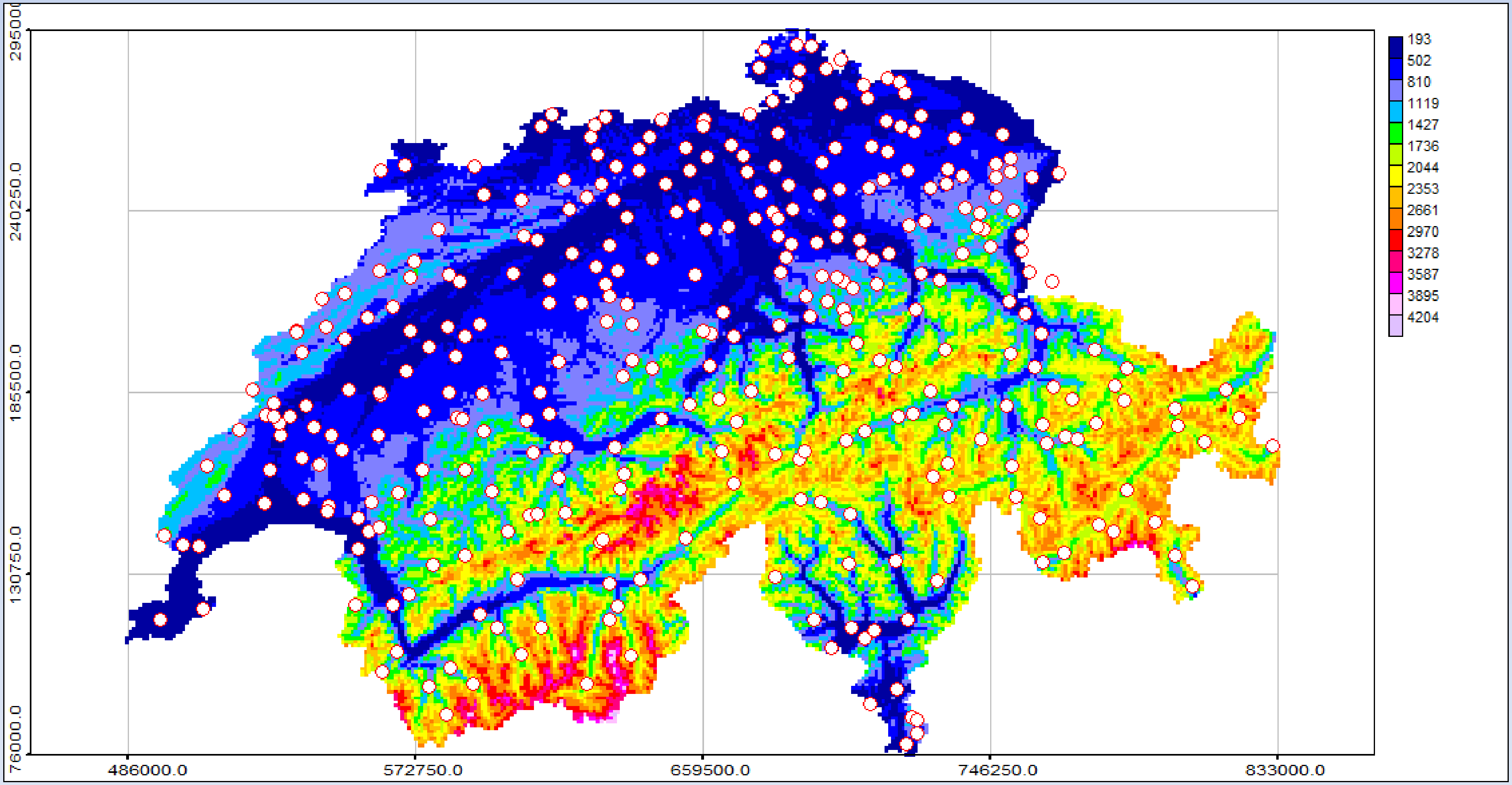


The same is valid for Adaptive GRNN!

Independent variables (features, inputs) which are irrelevant are "filtered out" automatically.

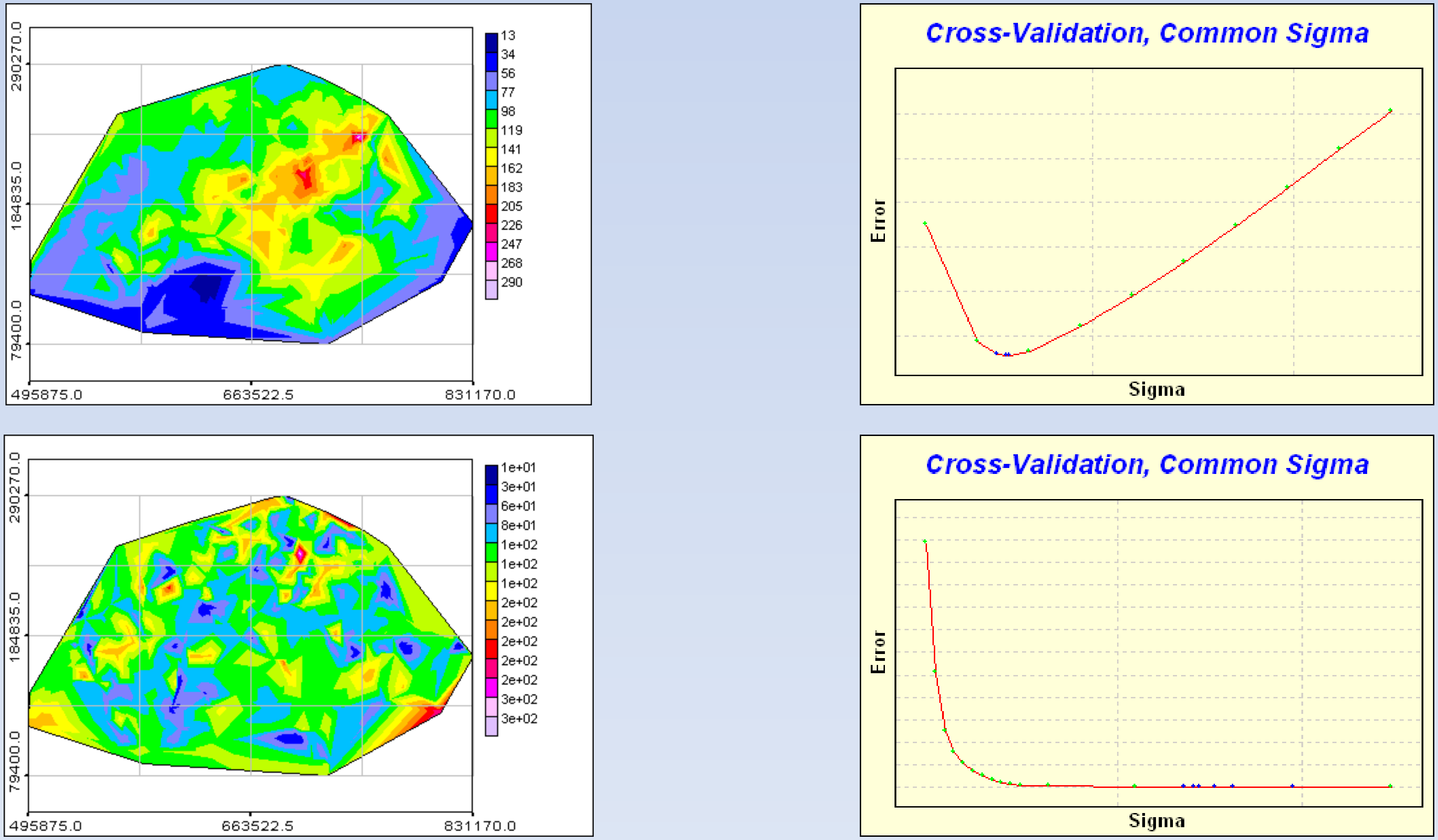
UNIL | Université de Lausanne

Swiss DEM and Precipitation Monitoring Network



UNIL | Université de Lausanne

Data: raw (top) and shuffled (down) and corresponding training curves



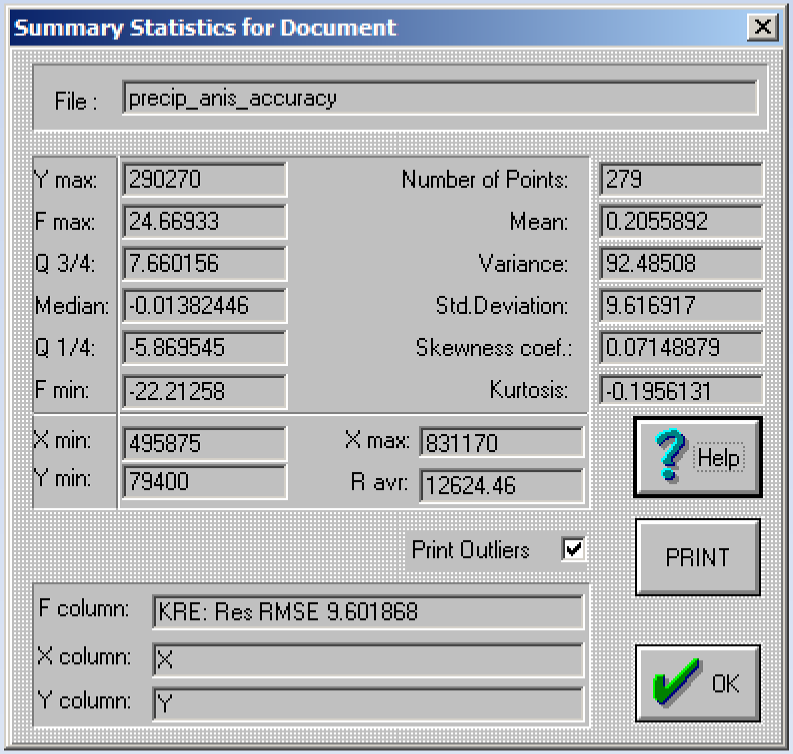
UNIL | Université de Lausanne

An example with added artificial coordinate: XYZ+ ZShuffled

Model	Cross-Validation error	Sigma values (metres)			
		σ_x	σ_y	σ_z	$\sigma_{Zshuffled}$
3D	419	7011	7601	192	
4D (3D+Noise)	420	6949	7474	191	*4135*

UNIL | Université de Lausanne

Quality of model? Analysis of the residuals using… GRNN!



Summary Statistics for Document

File : precip_anis_accuracy

Y max: 290270

F max: 24.66933

Q 3/4: 7.660156

Median: -0.01382446

Q 1/4: -5.869545

F min: -22.21258

Number of Points: 279

Mean: 0.2055892

Variance: 92.48508

Std.Deviation: 9.616917

Skewness coef.: 0.07148879

Kurtosis: -0.1956131

X min: 495875

Y min: 79400

X max: 831170

R avr: 12624.46

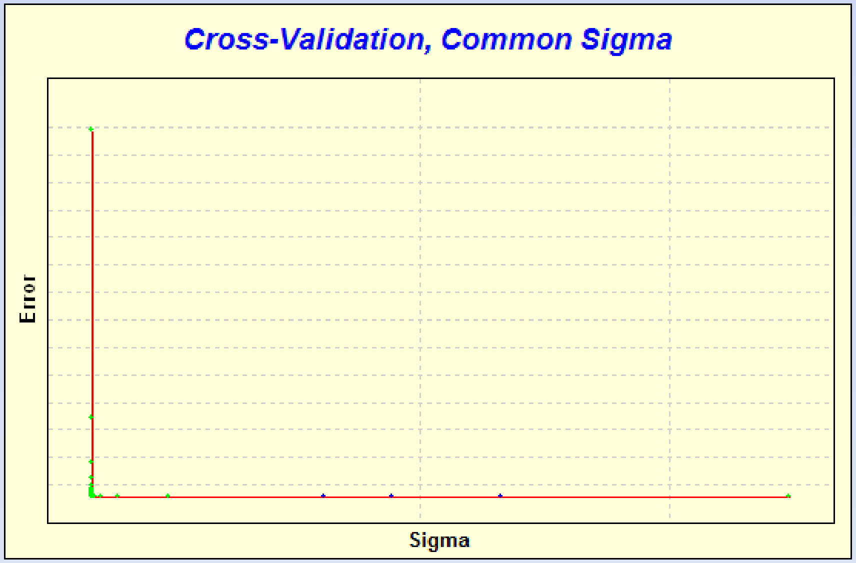
Print Outliers ☑

F column: KRE: Res RMSE 9.601868

X column: X

Y column: Y

CV error = 92.8; sigma=inf



Why feature selection

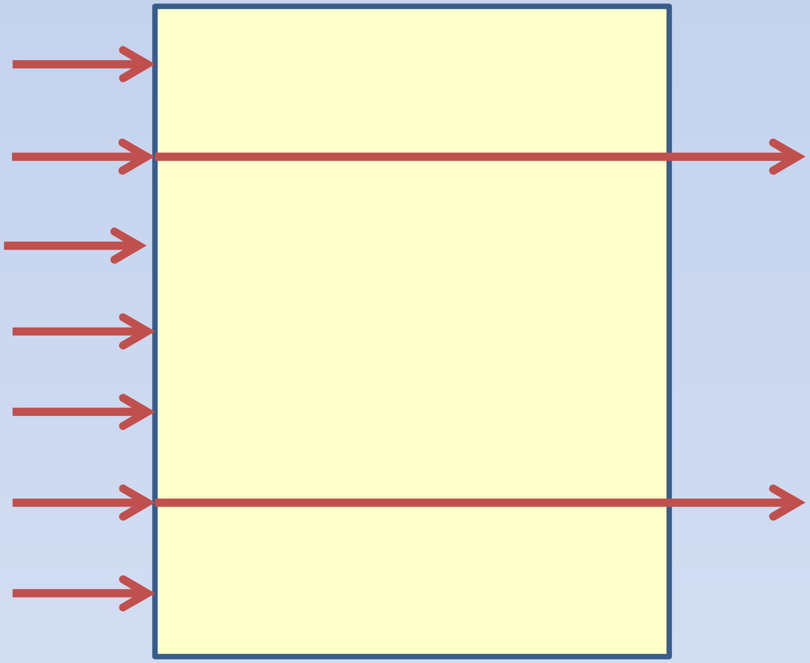
- Interpretability and data collection
  - Only relevant features
  - More parsimonious models
- Generalization (better predictability)
  - Lower dimensionality
- Computational efficiency
  - Faster
  - Number of parameters, overfitting
  - Scaling

Some models are sensitive to RD and IR variables (e.g. k-NN)

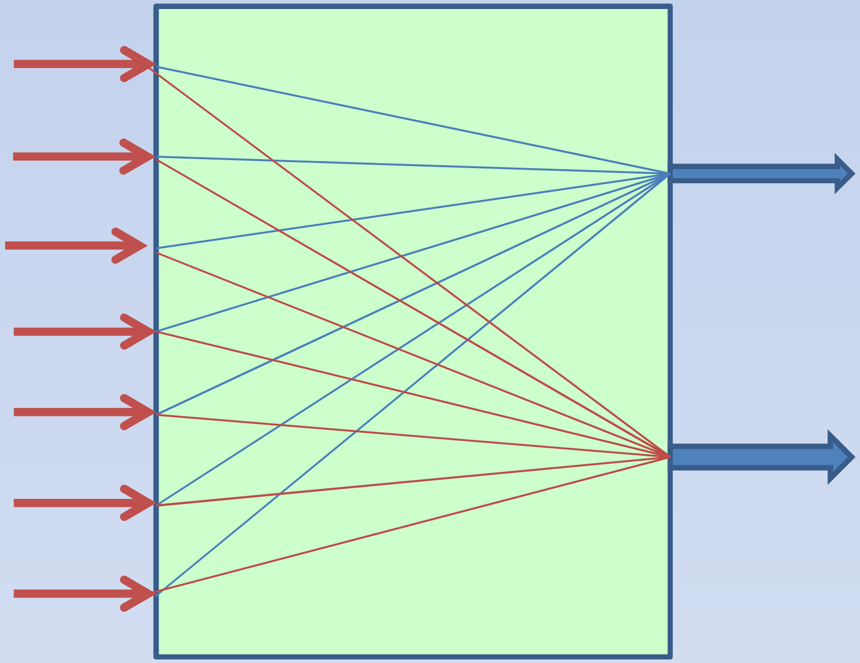
Features/Variables:

- *Relevant (RL) (contribute)*
- *Redundant (RD) (strongly correlated)*
- *Irrelevant (IR) (do not contribute)*

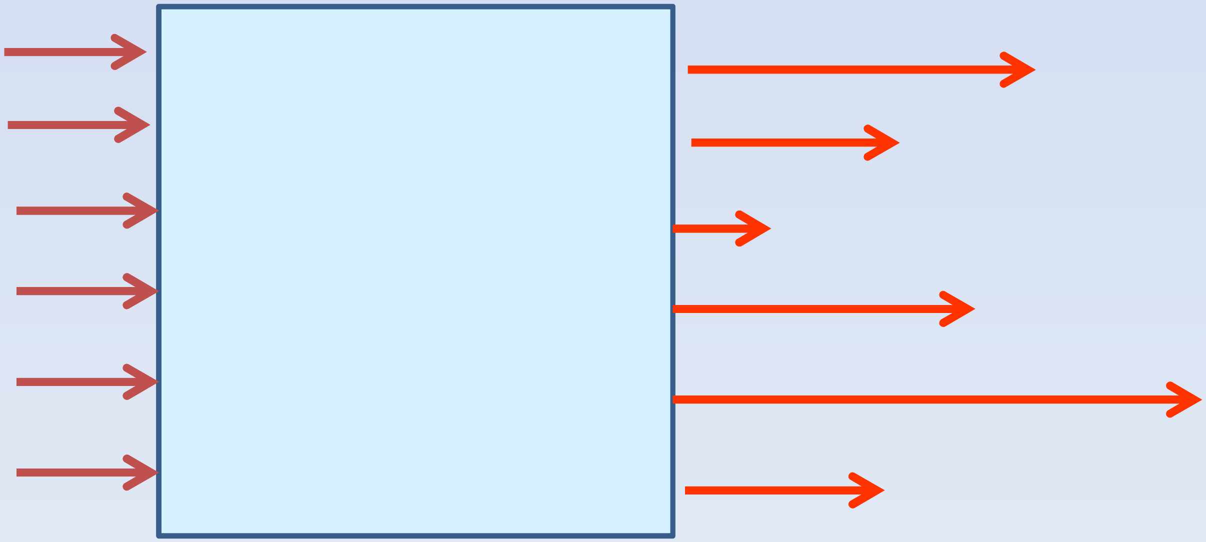
Feature Selection



Feature Extraction

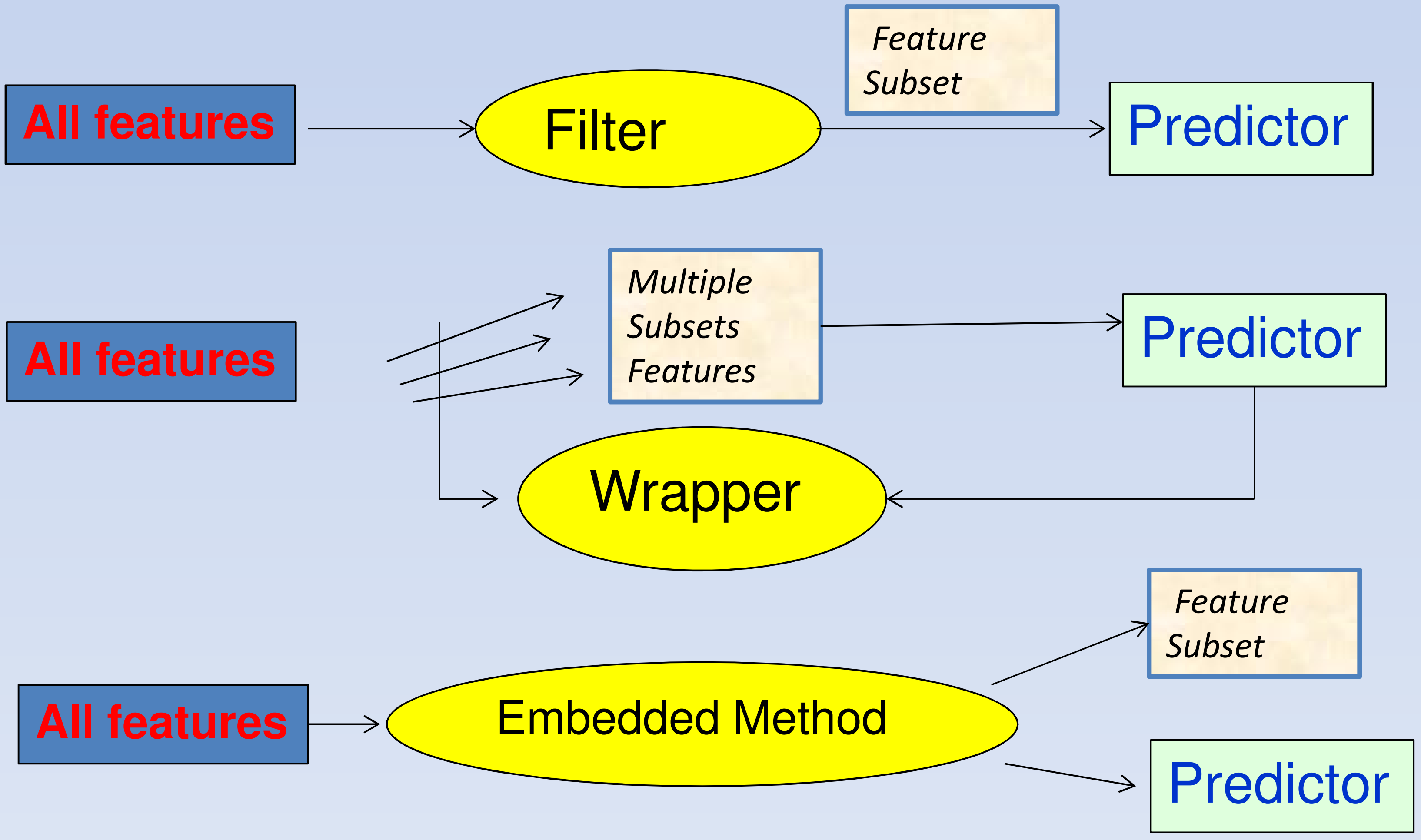


Feature Weighting



FS: fundamental approaches

(adapted from the lecture of I. Guyon and A. Eliseeff)



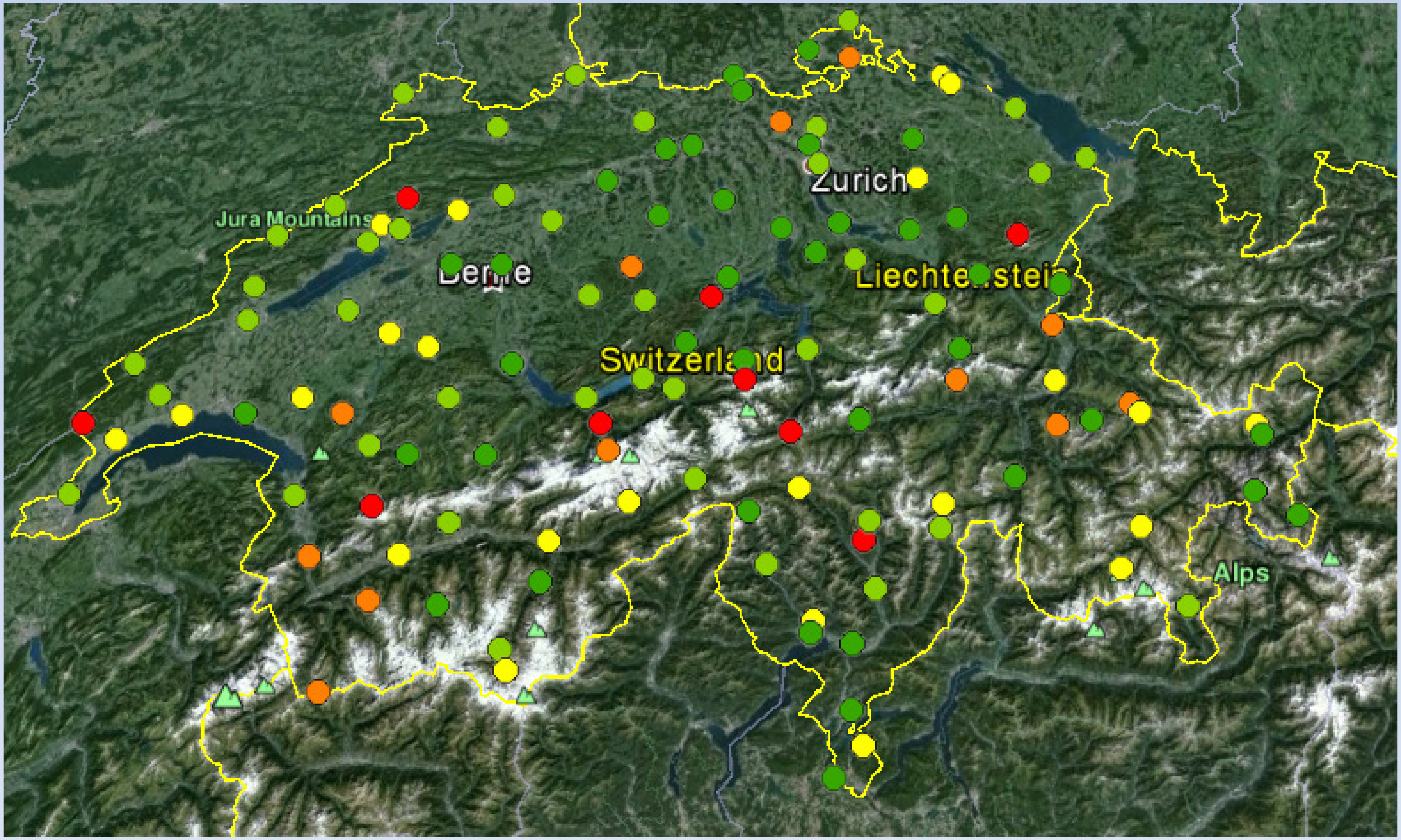
UNIL | Université de Lausanne

FS & Dimensionality reduction. Our experience

PCA, KPCA, MI, Gamma test, GRNN/AGRNN, ELM, SVM/RFE, MKL, ASVM, RF, Fractal-based, Morisita index

Case studies: topo-climatic modelling, natural hazards (landslides, avalanches, forest fires), pollution (air, water, soil), renewable resources (wind fields), remote sensing images

Adaptive GRNN (AGRNN) Analysis of wind speed



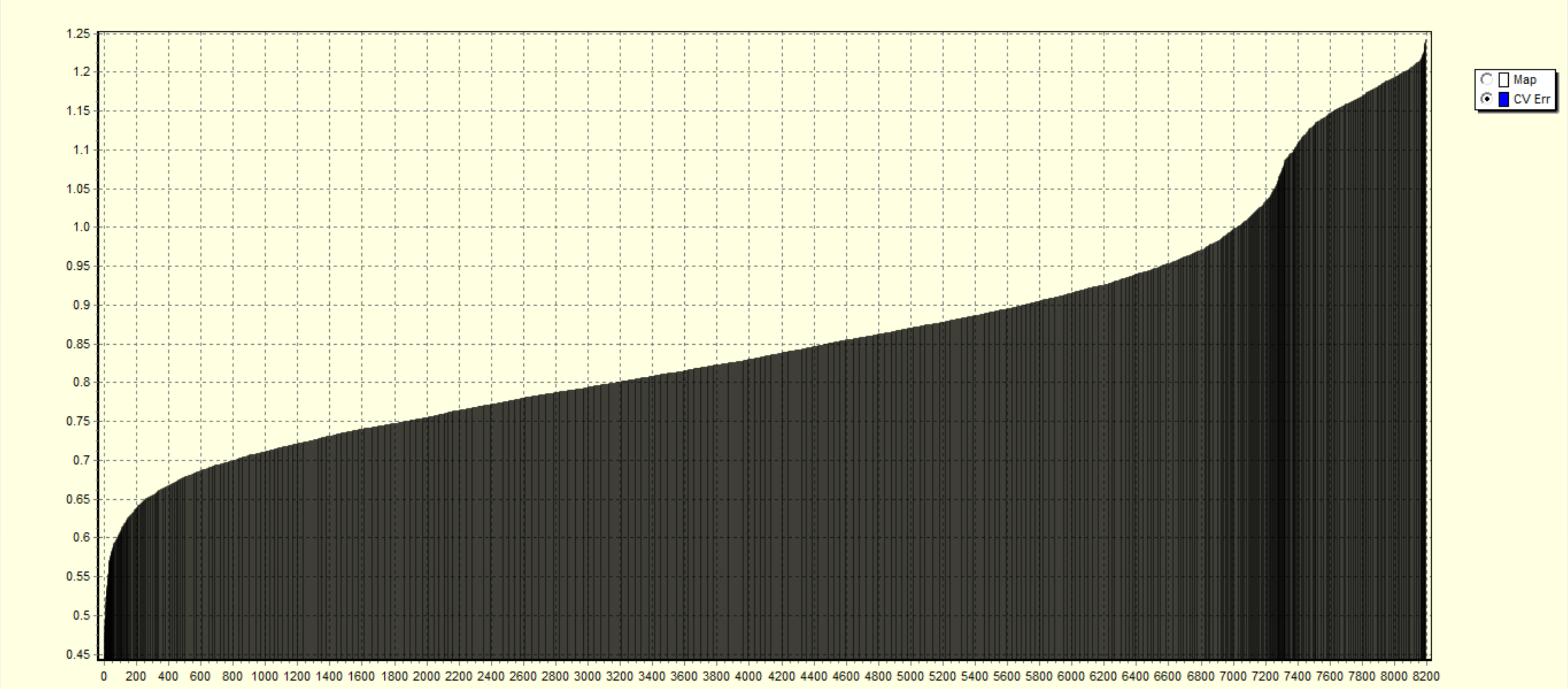
«Spatial prediction of monthly wind speeds in complex terrain with adaptive general regression neural networks». S. Robert, L. Foresti, M. Kanevski. International Journal of Climatology, 2013

Feature selection using comprehensive GRNN analysis of all possible models:

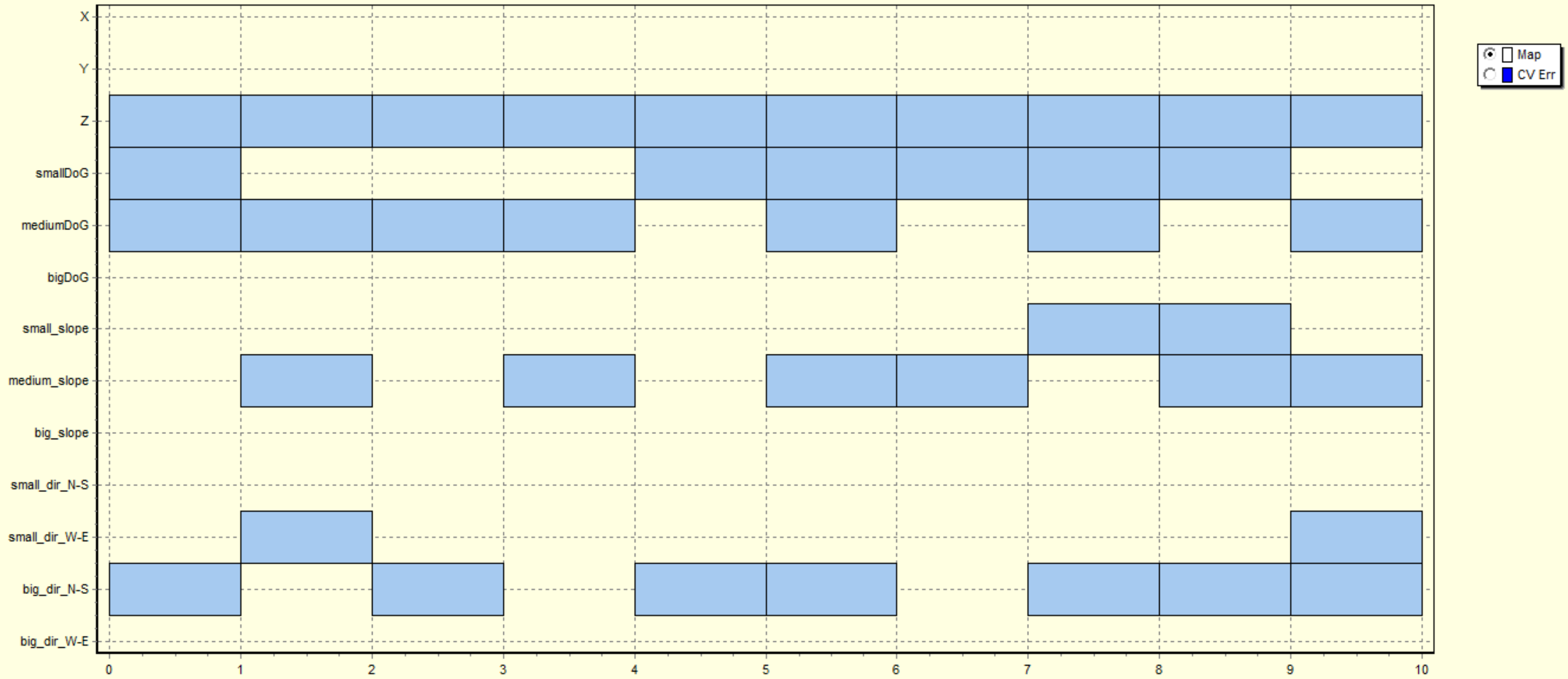
#Models = 2^d – 1 = 8191

UNIL | Université de Lausanne

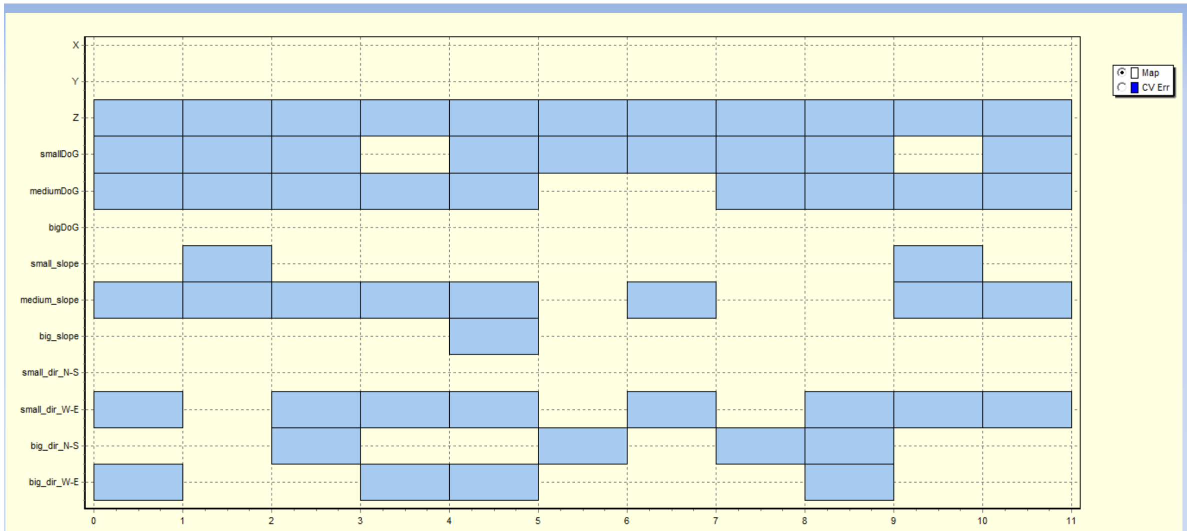
July 2008. CV error ordered from min to max

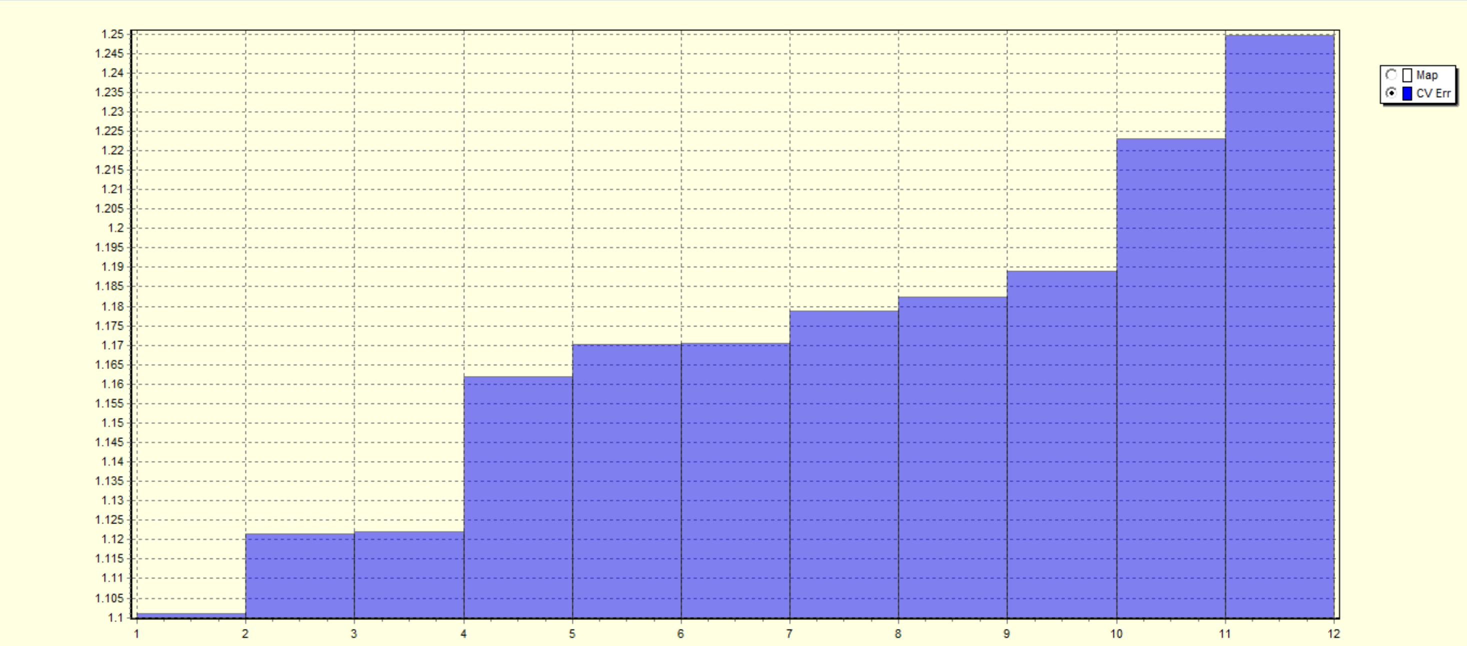


July 2008


X
Y
Z
smallDoG
mediumDoG
bigDoG
small_slope
medium_slope
big_slope
small_dir_N-S
small_dir_W-E
big_dir_N-S
big_dir_W-E
0
1
2
3
4
5
6
7
8
9
10
● □ Map
○ ■ CV Err

January 2008



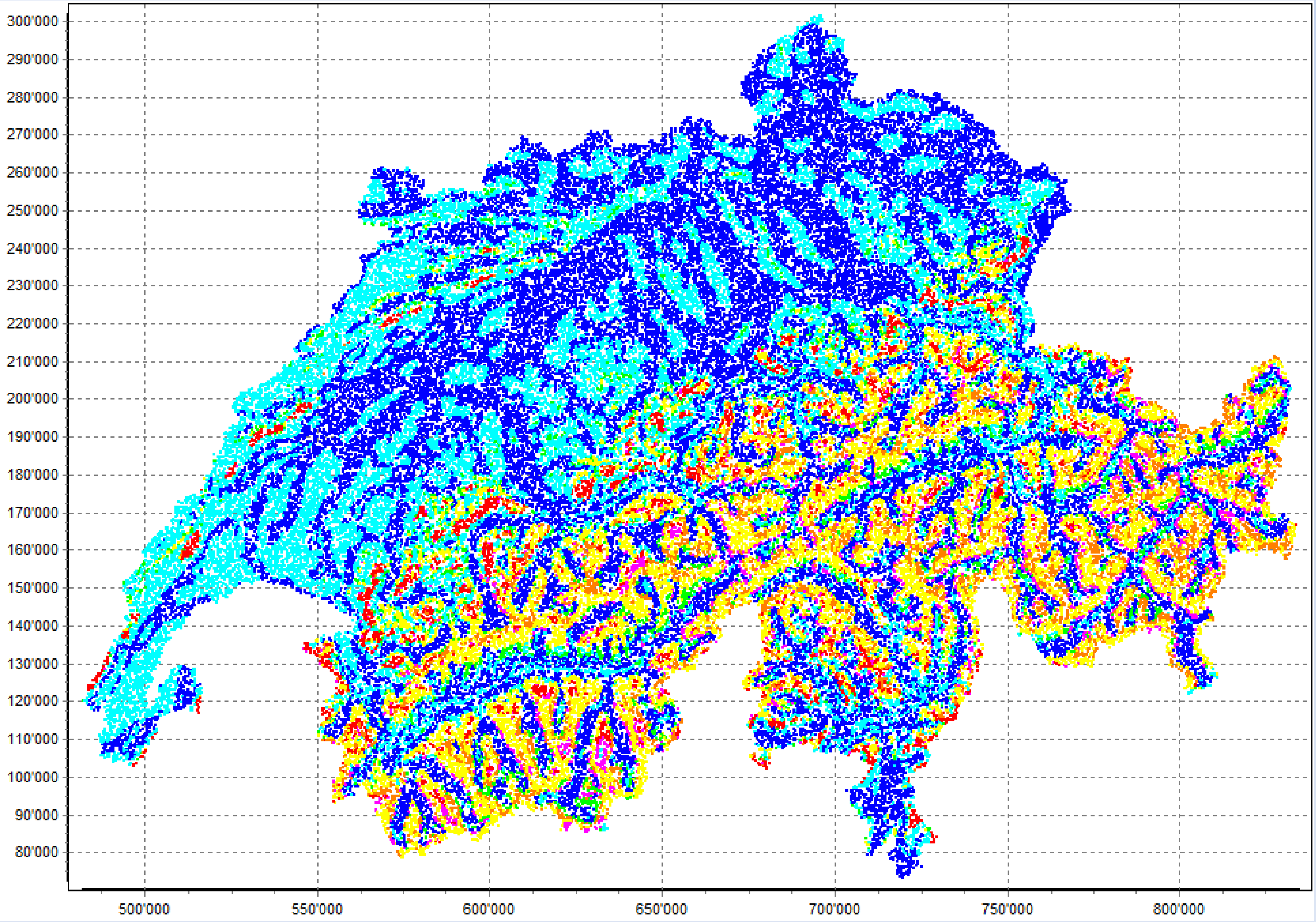


UNIL | Université de Lausanne

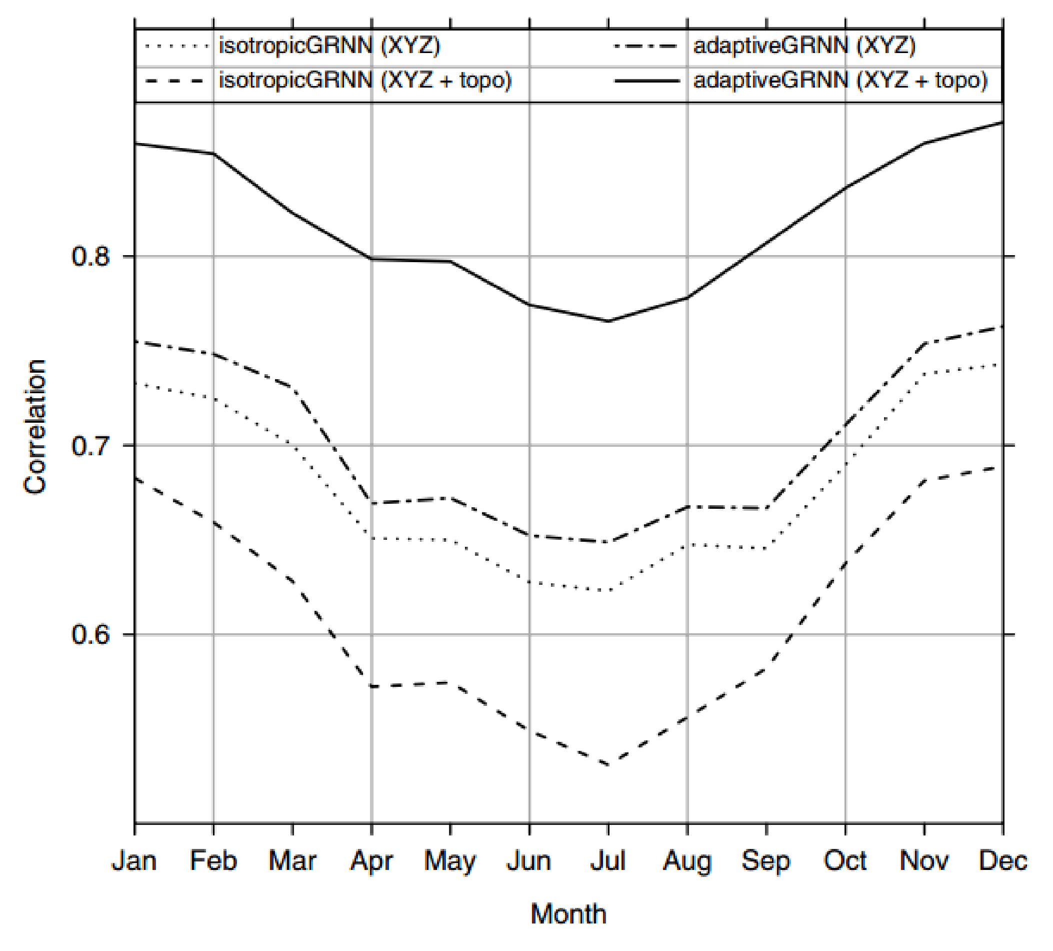
Comprehensive GRNN Analysis
Januaries

	1	2	3	4	5	6	7	8	9	10	11	12	13
Time\Features	X	Y	Z	smallDoG	mediumDoG	bigDoG	small_slope	medium_slope	big_slope	small_dir_N-S	small_dir_W-E	big_dir_N-S	big_dir_W-E
1968			🟨		🟨	🟨							
1973		🟨		🟨	🟨	🟨			🟨	🟨		🟨	
1978				🟨		🟨							
1983			🟨		🟨	🟨	🟨						
1988			🟨	🟨	🟨	🟨				🟨		🟨	
1993			🟨	🟨	🟨	🟨				🟨		🟨	
1998			🟨	🟨	🟨			🟨			🟨		
2003			🟨	🟨				🟨				🟨	
2008			🟨	🟨	🟨			🟨			🟨		🟨

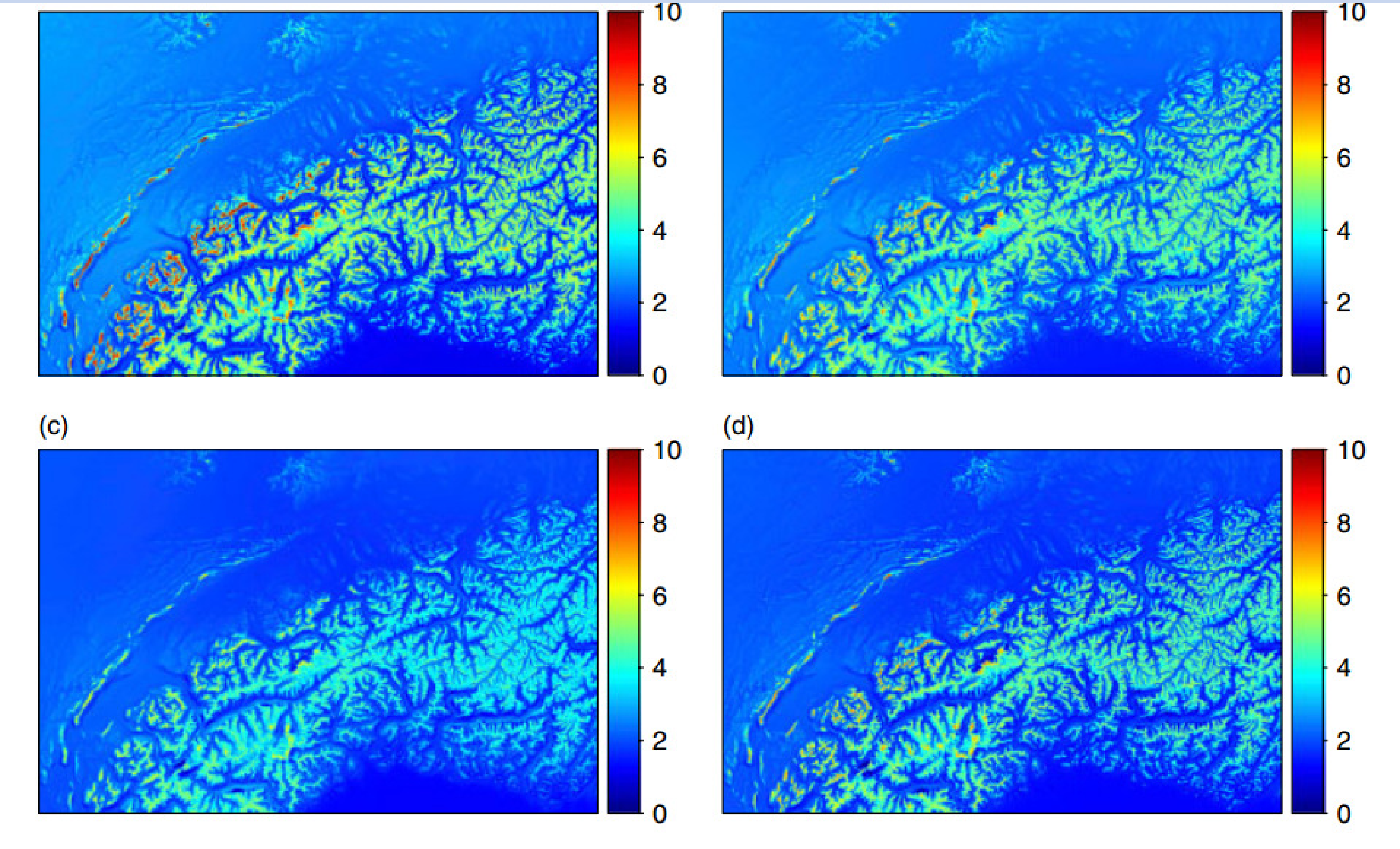
January 2008 (AGRNN)



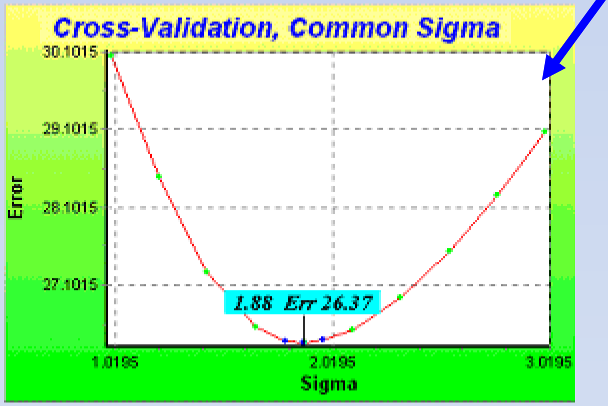
Monthly predictability

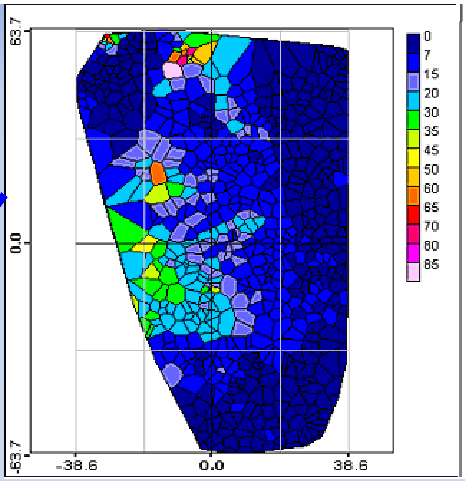


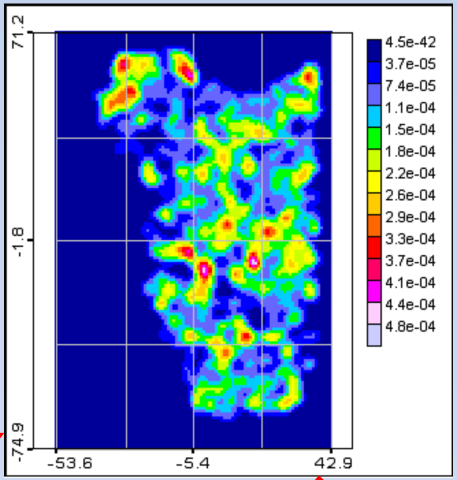
Seasonal maps

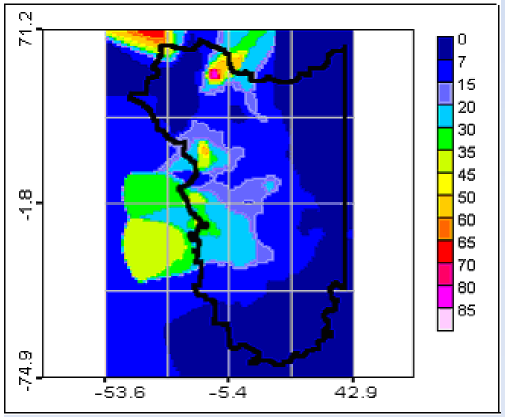


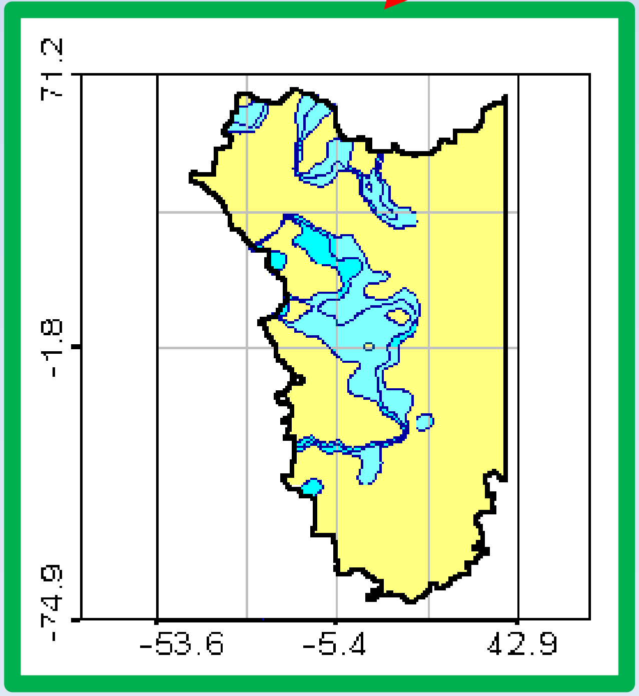
GRNN Mapping & uncertainties











Extreme Learning Machines for spatial environmental data
Michael Leuenberger & Mikhail Kanevski

Computers & Geosciences 85 (2015) pp. 64–73

UNIL | Université de Lausanne

Why ELM?

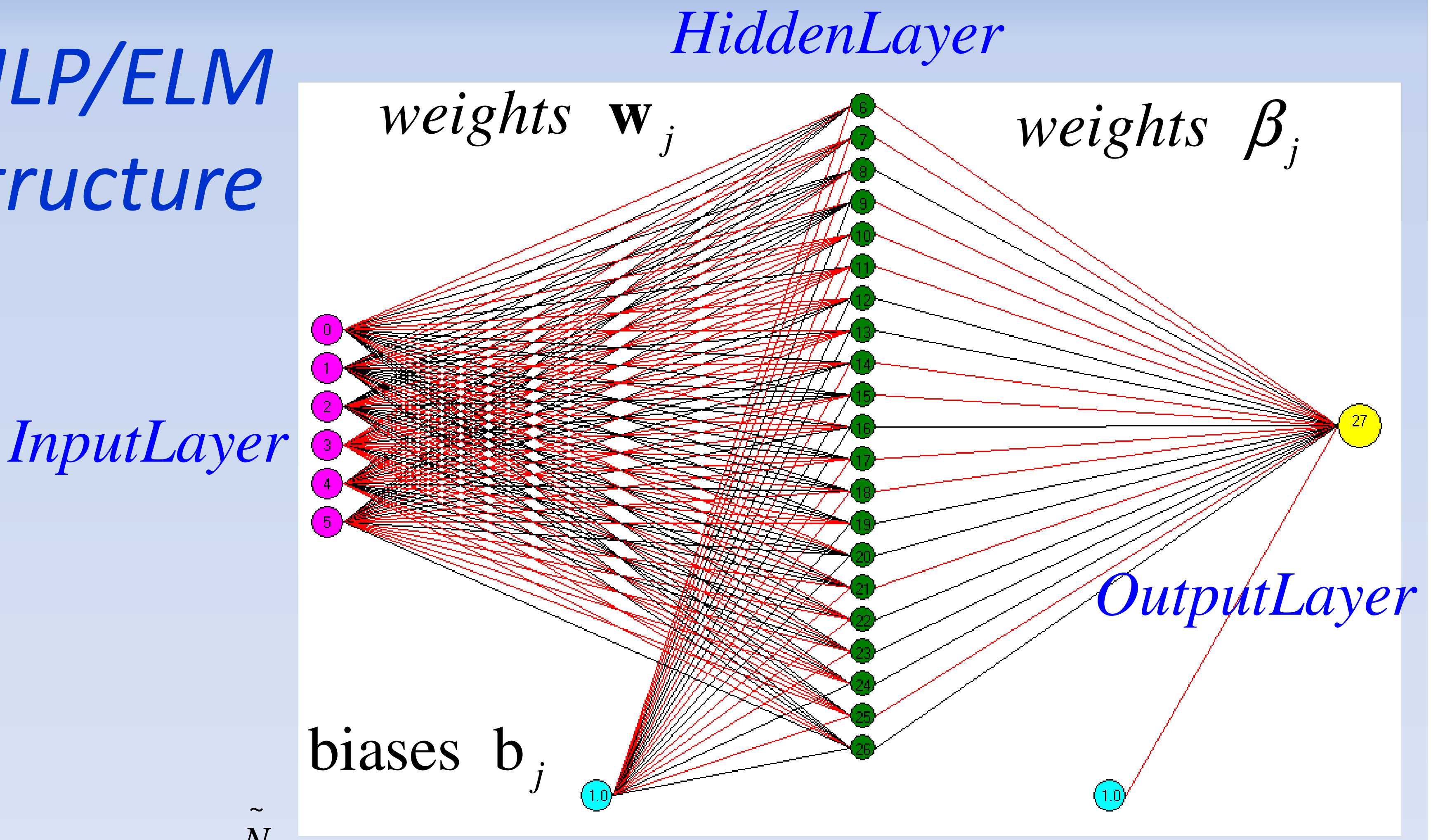
As the new proposed learning algorithm tends to

- reach the smallest training error,
- obtain the smallest norm of weights,
- the best generalization performance,
- and runs extremely fast,

in order to differentiate it from the other popular SLFN learning algorithms, it is called the Extreme Learning Machine (ELM)

Many successful applications in classification and regression tasks, extensions to semi-supervised and unsupervised learning,….

MLP/ELM structure



$$\sum_{j01}^{\tilde{N}} \beta_j g(\mathbf{x}_i \mathbf{w}_j + b_j) = \hat{y}_i \qquad \forall i = 1, \dots N$$

$H\beta = \hat{y}$, where $H_{ij} = g(\mathbf{x}_i \mathbf{w}_j + b_j)$ is the output matrix of a hidden layer

Sigmoid function

$$g(\mathbf{w}, b, \mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{wx} + \mathbf{b}))}$$

Gaussian function:

$$g(\mathbf{w}, b, \mathbf{x}) = \exp(-b\|\mathbf{x} - \mathbf{w}\|^2)$$

Fourier function

$$g(\mathbf{w}, b, \mathbf{x}) = \sin(\mathbf{wx} + b)$$

The ELM algorithm

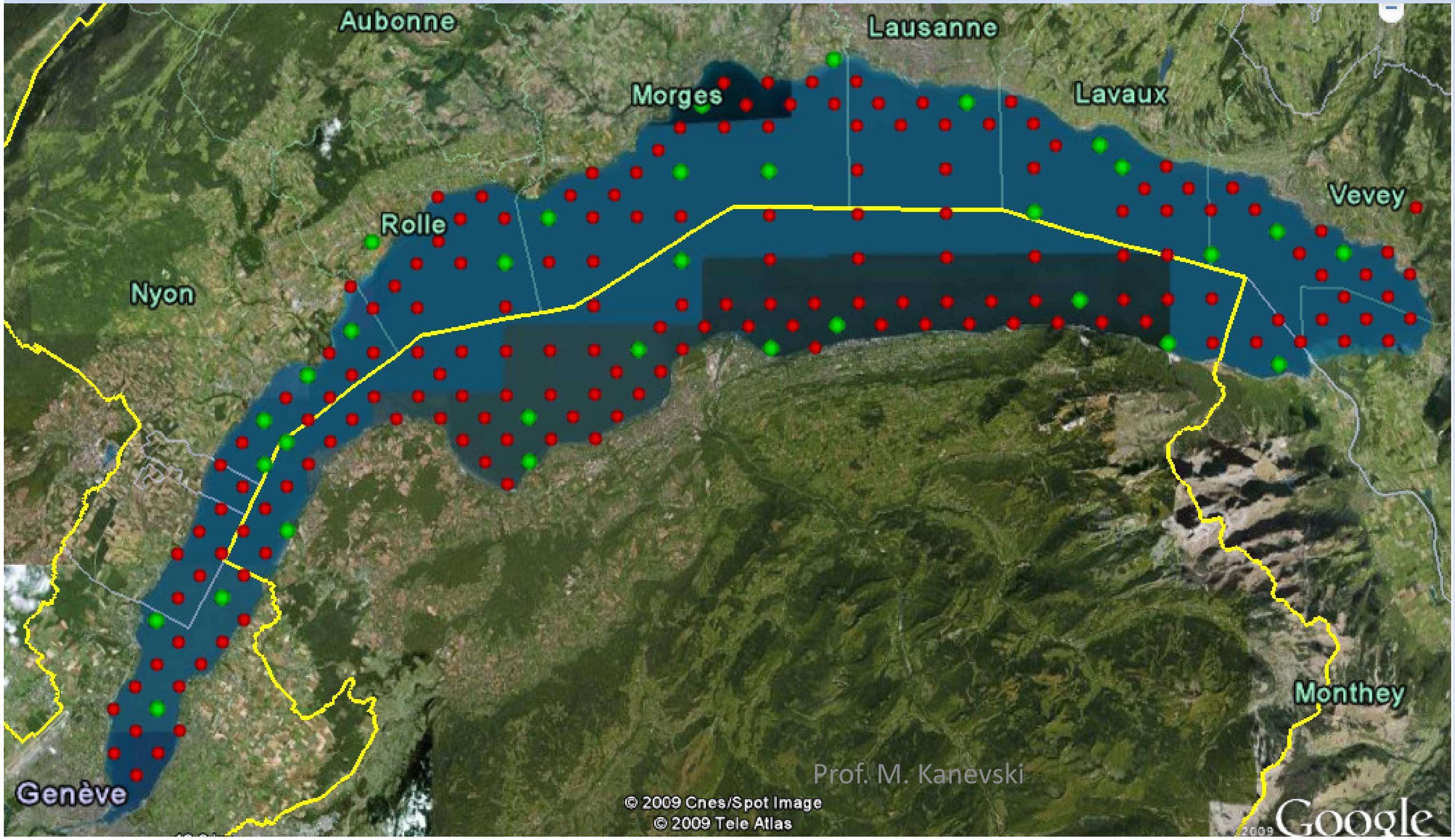
- Define a transfer function (e.g., sigmoid)
- Project data into [0,1] interval
- Split the data or use k-fold cross validation
- Select the hyper-parameter (the target of training): number of hidden neurons
- Generate randomly matrix weights and biases $\mathbf{w}_j, b_j$
- Calculate H, β_j
- Repeat for different number of hidden neurons and choose the solution with minimum CV error
- Test the model
- Make predictions

The ELM algorithm

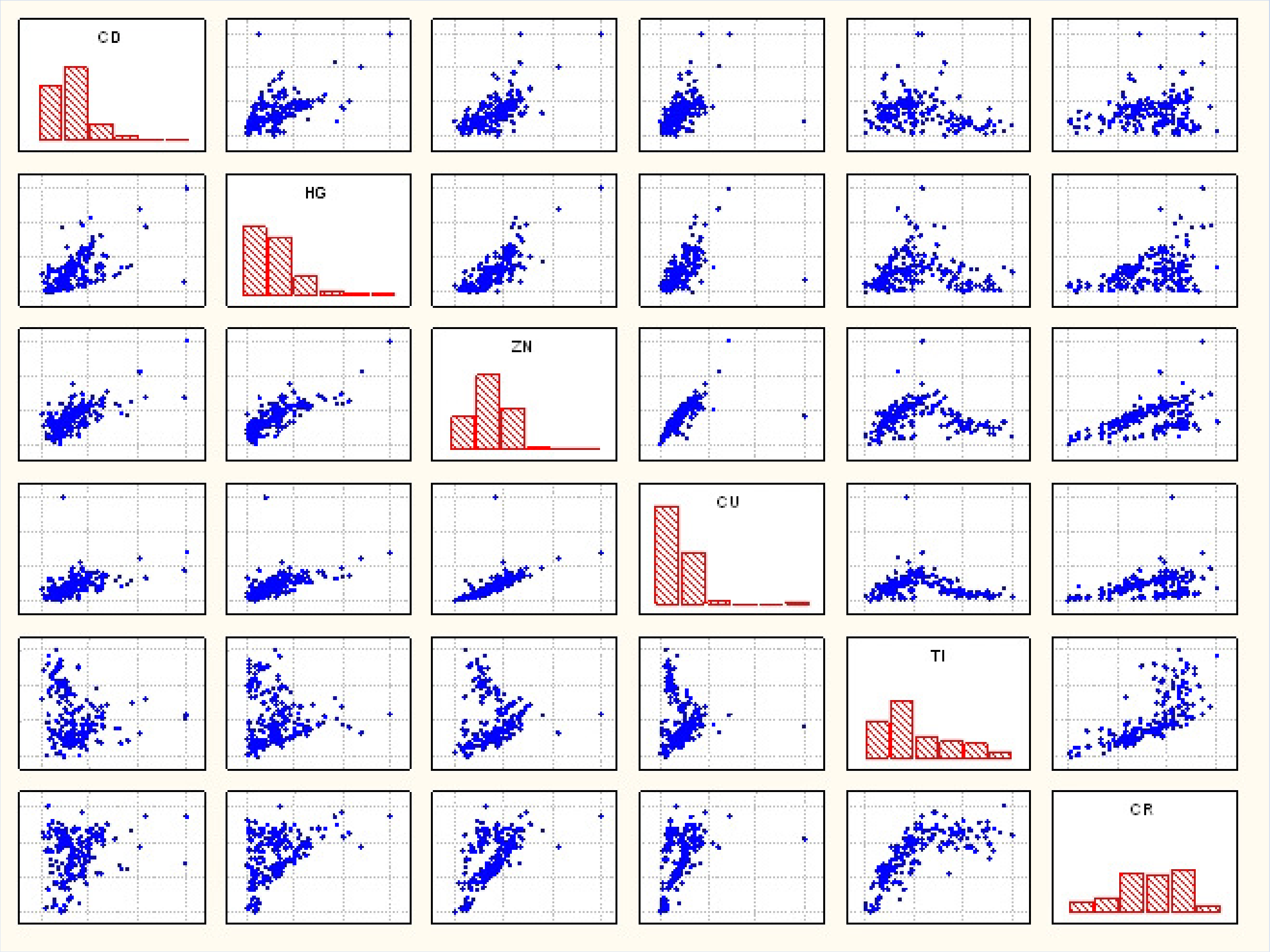
- Define a transfer function (e.g., sigmoid)
- Project data into [0,1] interval
- Split the data or use k-fold cross validation
- Select the hyper-parameter (the target of training): number of hidden neurons
- Generate randomly matrix weights and biases $\mathbf{w}_j, b_j$
- Calculate H, β_j
- Repeat for different number of hidden neurons and choose the solution with minimum CV error
- Test the model
- Make predictions

UNIL | Université de Lausanne

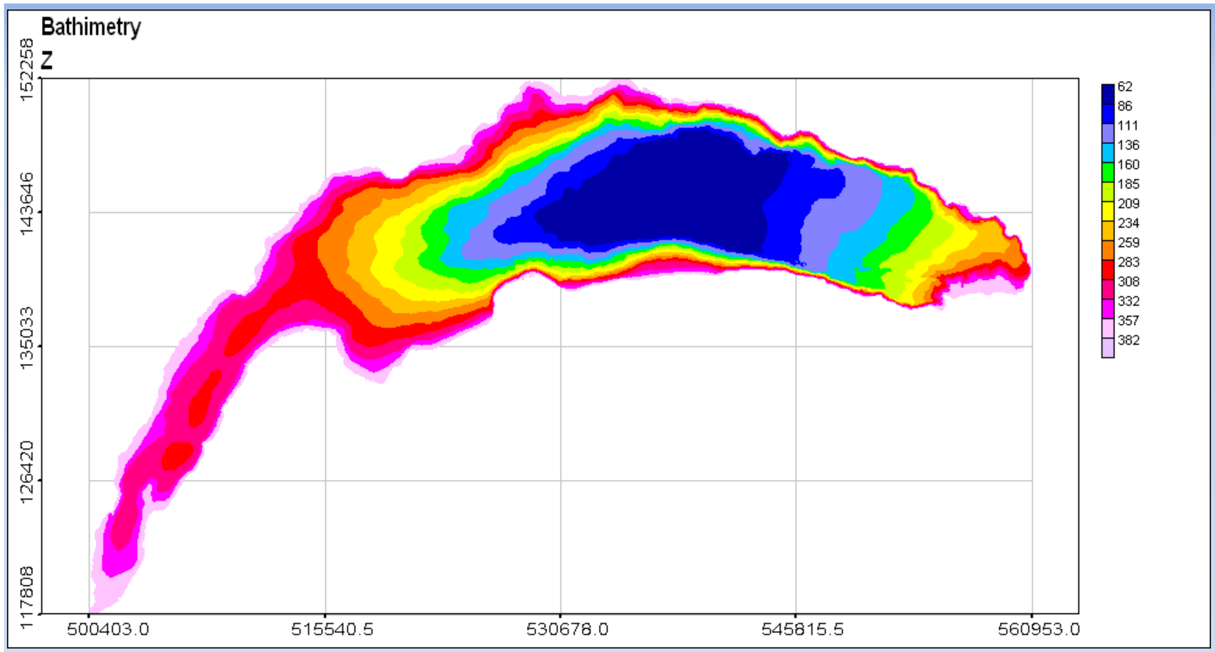
Geneva lake data



Matrix scatterplot

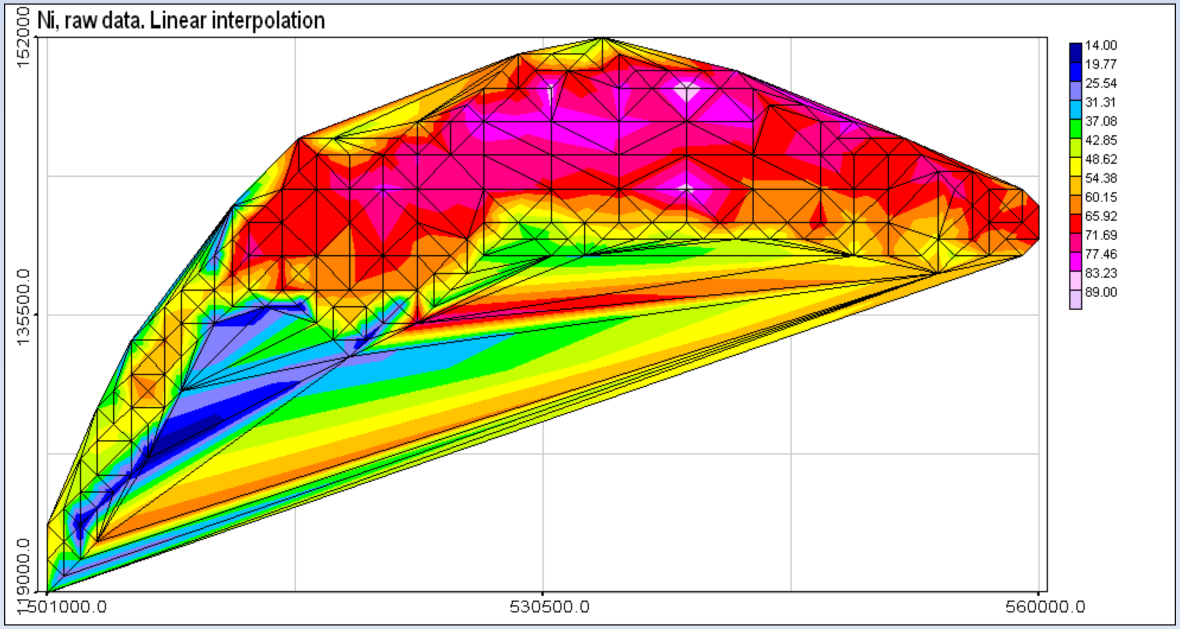


UNIL | Université de Lausanne



Geneva lake: bathymetry

Ni raw data



UNIL | Université de Lausanne

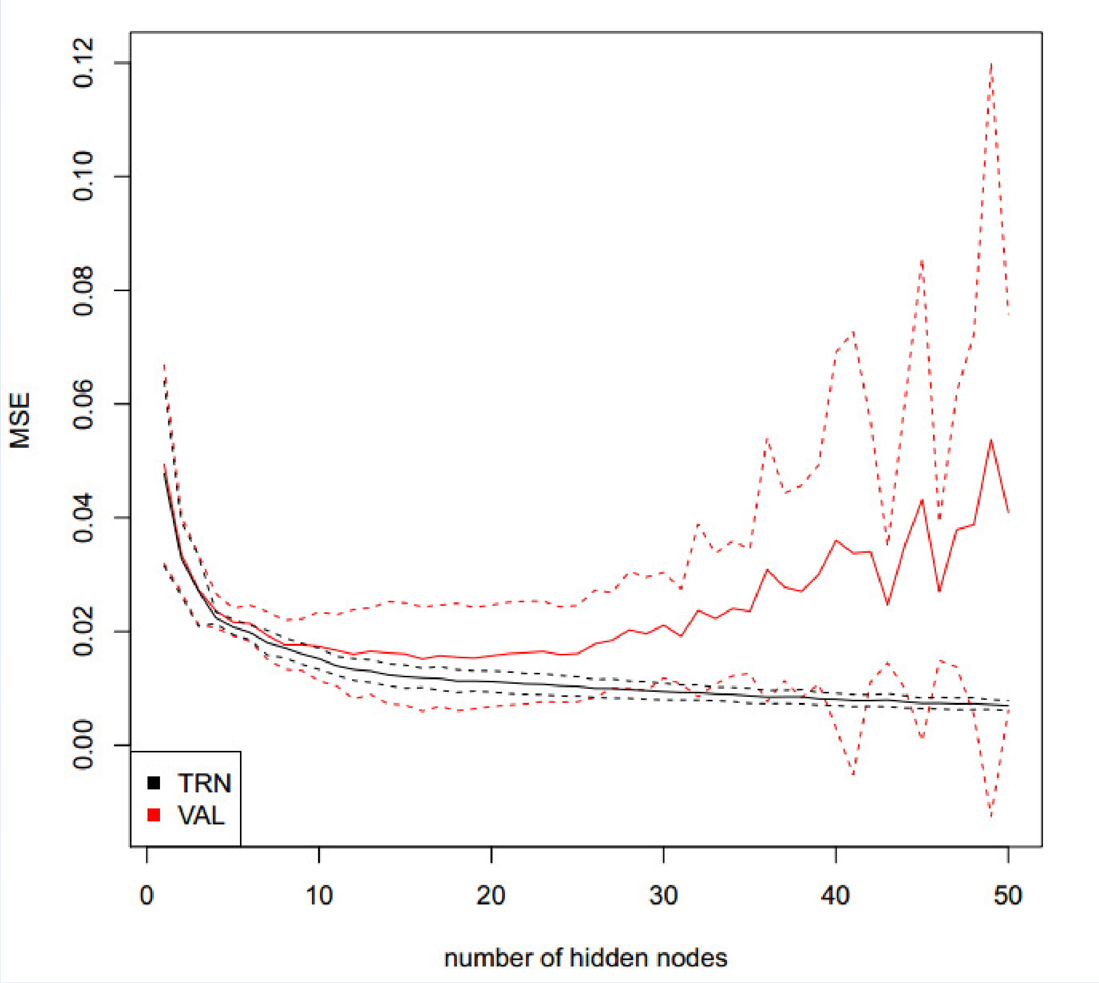
Geneva lake case study

- To find optimal number of hidden neurons, data were split into training (148) and testing (48) subsets
- Data were normalized to [0,1] interval
- Training was done by 5-fold cross validation restarted 20 times.

ELM training.
Optimal model: #hiddenNeurons = 16

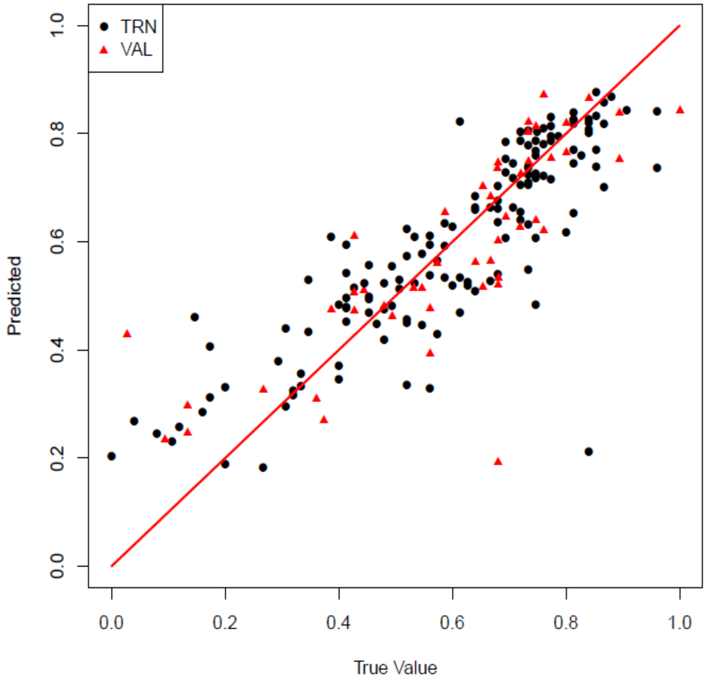
Training error: 0.013
Testing error: 0.017

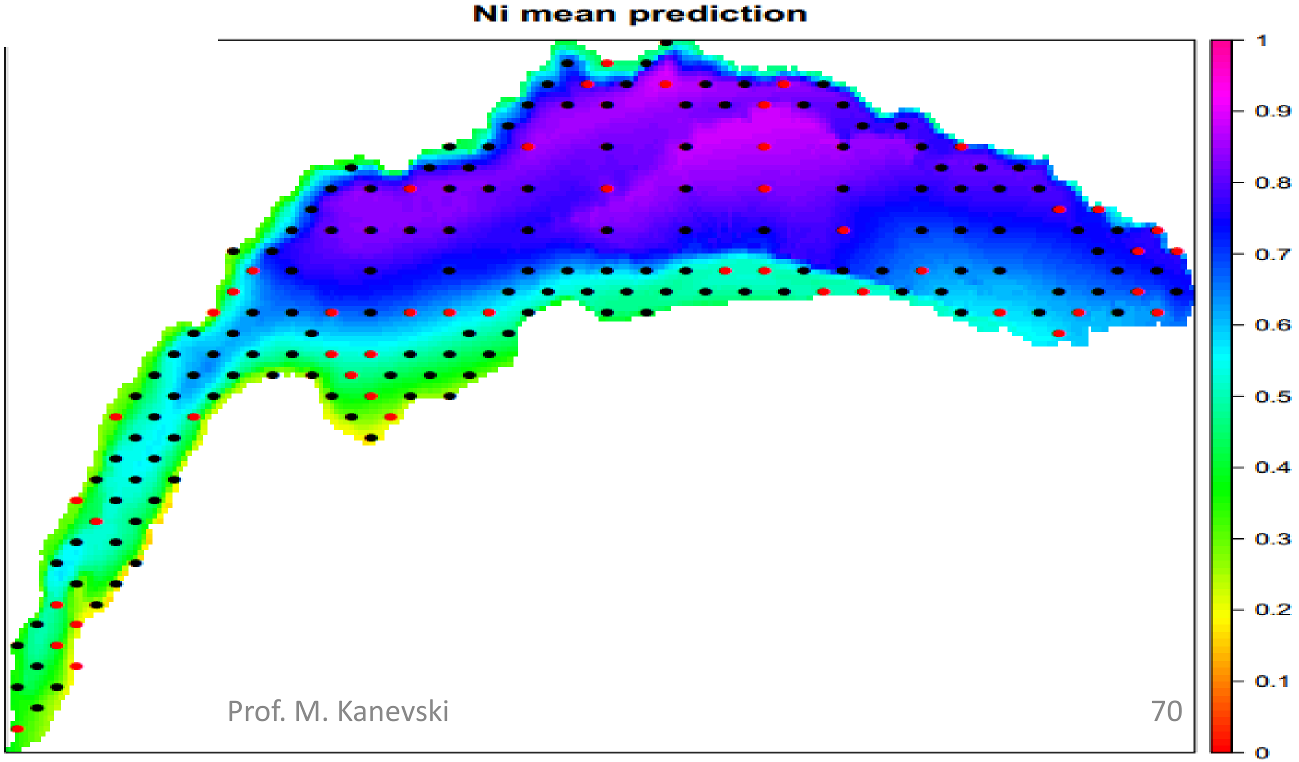
Variography of Ni:
Estimated nugget ~ 0.013



UNIL | Université de Lausanne

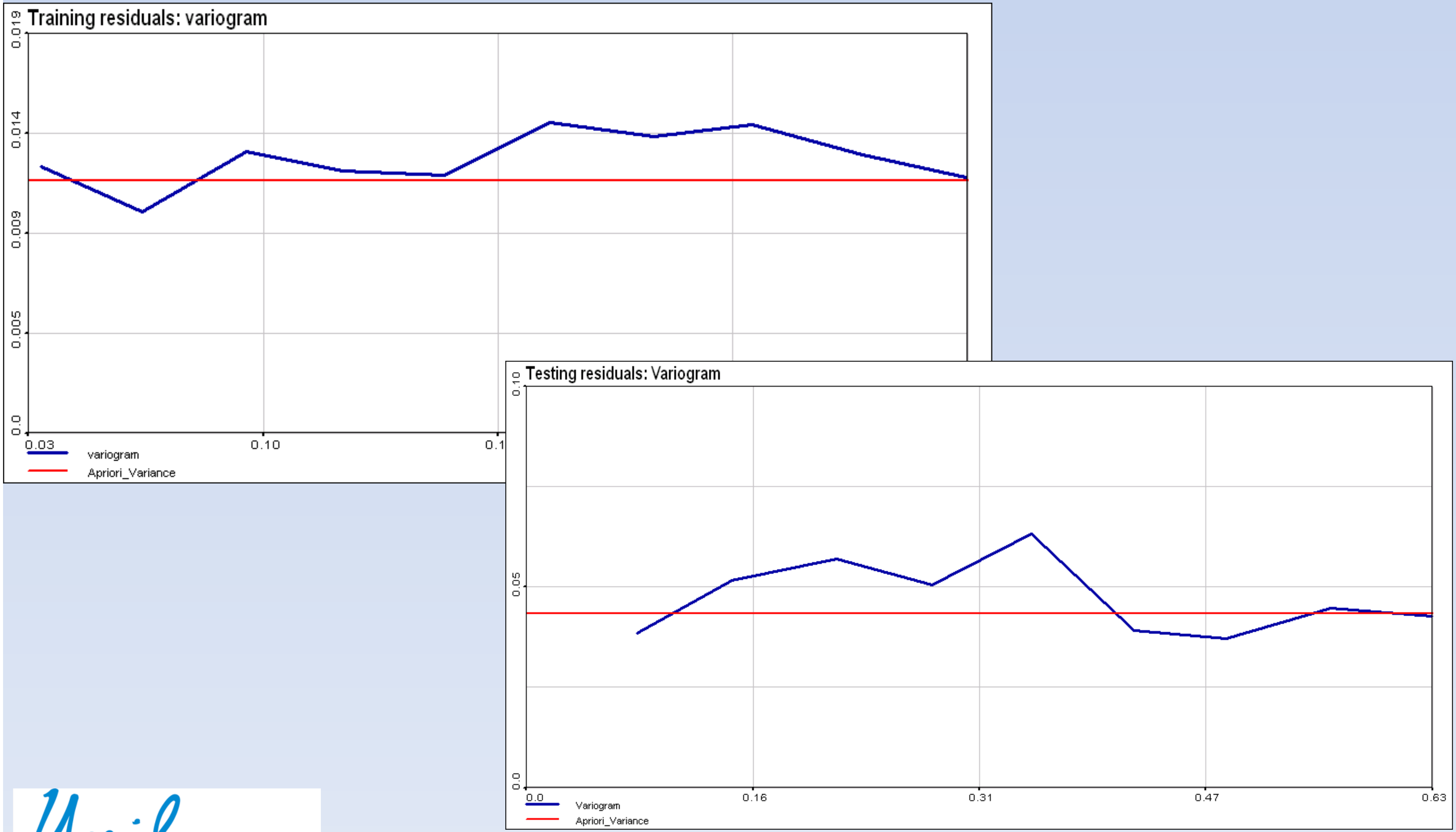
Ni 3d mapping. ELM(16)



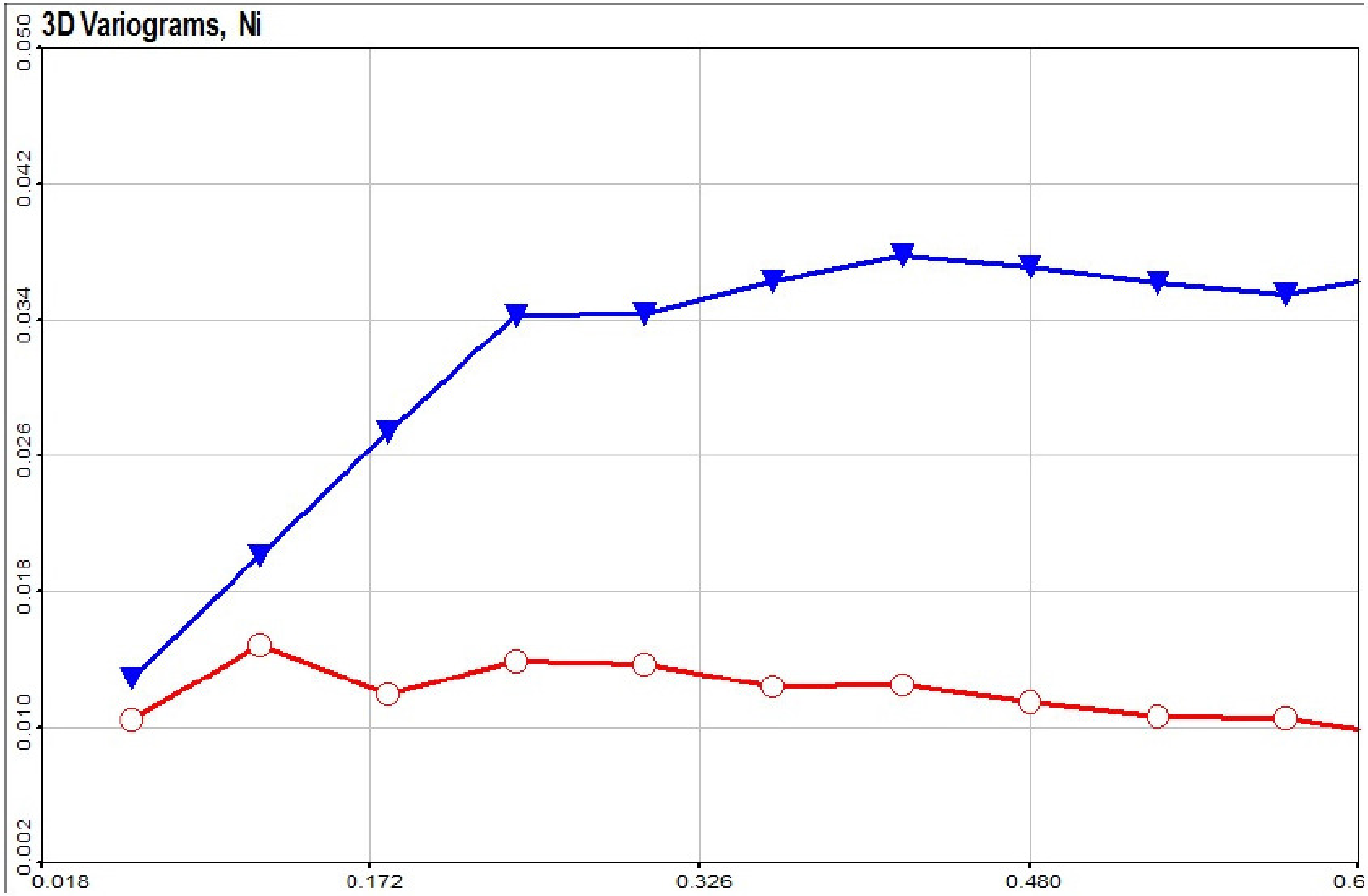


UNIL | Université de Lausanne

Analysis of the training (top) and testing (down) residuals



Variography: raw data and the residuals



UNIL | Université de Lausanne

Conclusions

MLA are efficient tools in environmental data analysis and modelling.

Recent our developments:

MLA for risk assessments (natural hazards,
pollution)

New methods for FS (ELM, Morisita index, AGRNN)

Applications: permafrost modelling, forest fires

Applications: Nonlinear Land Use Regression
models (air pollution in a city)

Unsupervised learning of local clustering

Deep understanding and methodologically correct use of ML are necessary

UNIL | Université de Lausanne

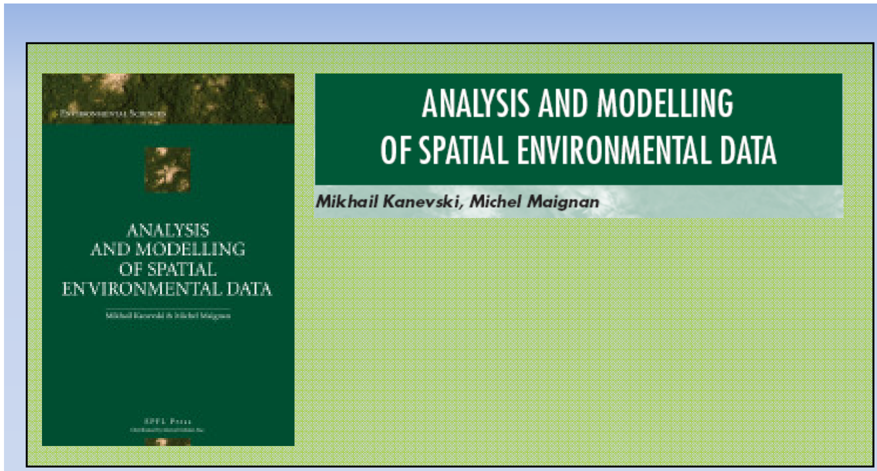
Challenges

- Environmental data mining (to be continued).
- "Fractal data mining": Intrinsic dimension estimation, Dimensionality reduction, Feature selection
- Scaling: dimensionality, volume (BIG) & complexity
- Active learning and Monitoring Networks Optimization
- From dependencies to cause-effect relationships
- Uncertainties. Risks and extremes
- **In**tegration of science-based and data-driven models.

From Advanced Analysis to ADVANCED THINKING!

UNIL | Université de Lausanne

Thank you for your attention!



ENVIRONMENTAL SCIENCES

ANALYSIS AND MODELLING OF SPATIAL ENVIRONMENTAL DATA

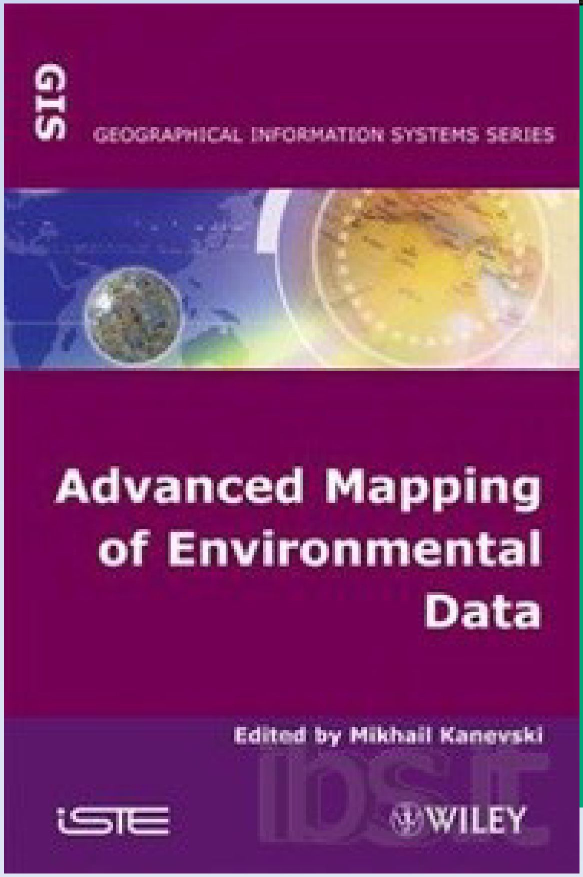
Mikhail Kanevski & Michel Maignan

EPFL Press

ANALYSIS AND MODELLING OF SPATIAL ENVIRONMENTAL DATA

Mikhail Kanevski, Michel Maignan

Environmental Data Science

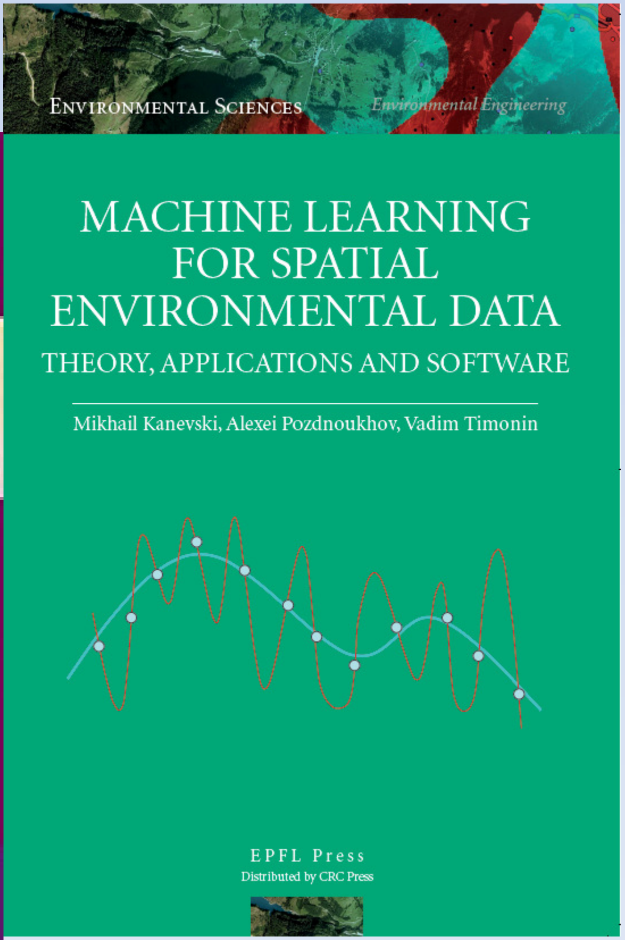


GIS GEOGRAPHICAL INFORMATION SYSTEMS SERIES

Advanced Mapping of Environmental Data

Edited by Mikhail Kanevski

iSTE WILEY



ENVIRONMENTAL SCIENCES *Environmental Engineering*

MACHINE LEARNING FOR SPATIAL ENVIRONMENTAL DATA

THEORY, APPLICATIONS AND SOFTWARE

Mikhail Kanevski, Alexei Pozdnoukhov, Vadim Timonin

EPFL Press

Distributed by CRC Press

UNIL | Université de Lausanne