

Motivation
oooooooo

GRNN
ooo

FS with GRNN
ooooo
ooo

Spatial wind speed prediction
oooooo
o
oo



Feature selection and data dependencies analysis: an operative tool for geospatial data,

Federico Amato, Fabian Guignard, and Mikhail Kanevski

Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment
University of Lausanne

Motivation
●○○○○○○○

GRNN
○○○

FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○
○
○○

The curse of dimensionality

Motivation

Environmental sciences and geomatics are experiencing a previously unseen growth of data collection and storage (big data). This results in **High Dimensional** *feature spaces*.

Typical real data case studies:

- ▶ Wind fields > 13d
- ▶ Avalanches > 40d
- ▶ Landslides > 18d
- ▶ Permafrost > 20d
- ▶ Wildfires > 25d
- ▶ City pollution > 50d
- ▶ Remote sensing > 100d
- ▶ ...



Motivation
○●○○○○○

GRNN
○○○

FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○○
○
○○

The curse of dimensionality

Why Feature Selection?

Strongly relevant

Features that cannot be removed without affecting the original conditional target distribution

Weakly relevant, but not redundant

Features that can become necessary under certain condition

Weakly relevant, but redundant

Features that can be completely replaced with a set of other features such that the target distribution is not disturbed

Irrelevant

Features which absence does not affect original conditional target distribution

- ▶ **Curse of dimensionality:** redundant or irrelevant features reduce the accuracy of prediction

Motivation
○○●○○○○○

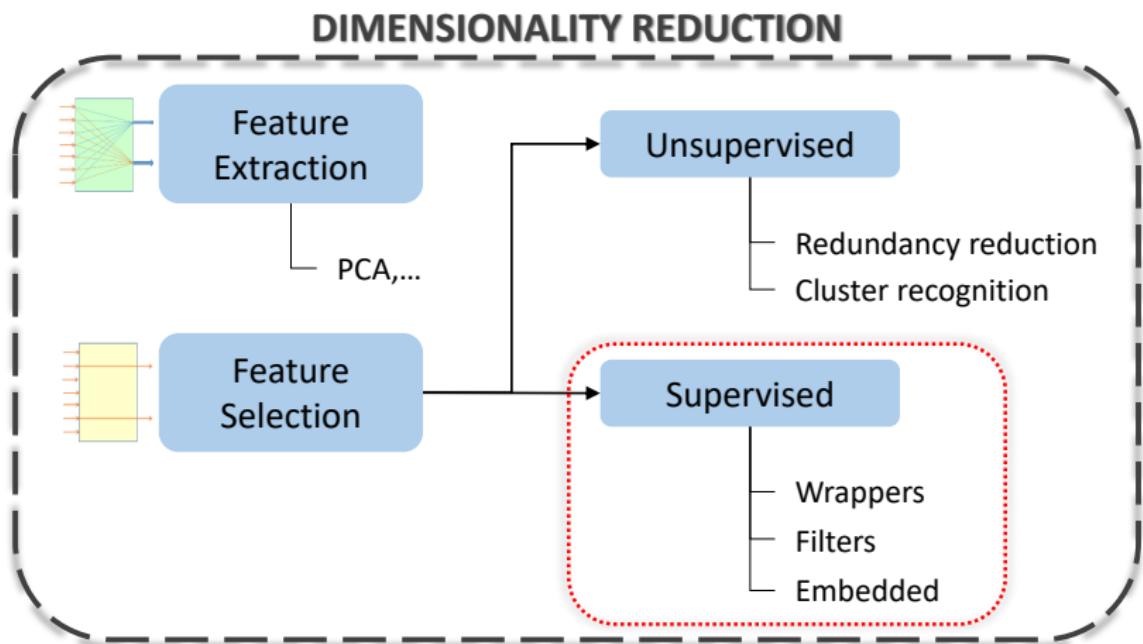
GRNN
○○○

FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○○
○
○○

The curse of dimensionality

From dimensionality reduction to feature selection



Motivation
○○○●○○○

GRNN
○○○

FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○○
○
○○

The curse of dimensionality

Wrappers

- ▶ Use the performance of a learning algorithm to determine the quality of the selected subset of feature
- ▶ With exhaustive search there are $2^p - 1$ models resulting from p -dimensional input space:
 - ▶ $d = 8 \rightarrow 255$ possible subsets
 - ▶ $d = 20 \rightarrow 1048575$ possible subsets
 - ▶ $d = 50 \rightarrow 1125899906842623$ possible subsets!!
- ▶ In most cases the exhaustive search is not computationally feasible. Several search strategies have been proposed, such as the sequential forward selection, the sequential backward selection, the randomized search...
- ▶ High computational time, high accuracy

Motivation
○○○●○○○

GRNN
○○○

FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○○
○
○○

The curse of dimensionality

Filters

- ▶ Use the intrinsic characteristics in data to rank features using dependency, distance or consistency measures
- ▶ Many filters have been proposed in literature (ReliefF, Correlation-based FS, ...)
- ▶ Faster than, but not as accurate as, wrappers

Motivation
oooooo●oo

GRNN
ooo

FS with GRNN
ooooo
ooo

Spatial wind speed prediction
oooooooo
o
oo

The curse of dimensionality

Embedded

- ▶ Integrate feature selection in the learning process itself (e.g. Lasso, Ridge, Random Forest - but not really -, ...)
- ▶ Depending on the algorithm there may be a strong bias

Motivation
oooooooo●○

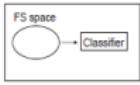
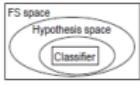
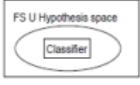
GRNN
○○○

FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○
○
○○

The curse of dimensionality

Different methods, different advantages and disadvantages

	Model search	Advantages	Disadvantages	Examples
Filter		Univariate Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	Chi-square Euclidean distance t-test Information gain, Gain ratio [6]
		Multivariate Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation based feature selection (CFS) [45] Markov blanket filter (MBF) [62] Fast correlation based feature selection (FCBF) [136]
Wrapper		Deterministic Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) [60] Sequential backward elimination (SBE) [60] Plus q take-away r [33] Beam search [106]
		Randomized Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing [110] Genetic algorithms [50] Estimation of distribution algorithms [52]
Embedded		Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies	Classifier dependent selection	Decision trees Weighted naive Bayes [28] Feature selection using the weight vector of SVM [44, 125]

Motivation
oooooooo●

GRNN
○○○

FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○○
○
○○

The curse of dimensionality

Objective

In this research we investigated the capability of General Regression Neural Networks:

- ▶ to fully explore the input space, recognizing relevant, irrelevant and redundant features ("Isotropic" wrapper approach);
- ▶ to perform an internal weighting/ranking of the features ("Anisotropic" embedded/filter approach)

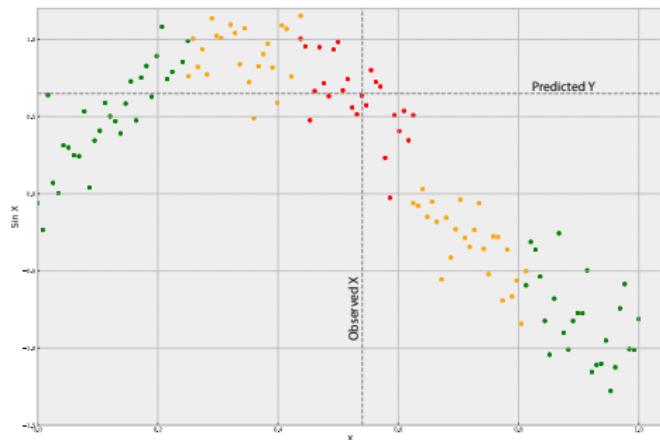
Motivation
○○○○○○○○

GRNN
●○○

FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○
○
○○

Nadaraya-Watson estimator



$$\hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^n y_i \exp(-D(\mathbf{x}, \mathbf{x}_i))}{\sum_{i=1}^n \exp(-D(\mathbf{x}, \mathbf{x}_i))}$$

$$D(\mathbf{x}, \mathbf{x}_i) = \sum_{j=1}^p \left(\frac{x_j - x_{ij}}{\sigma} \right)^2$$

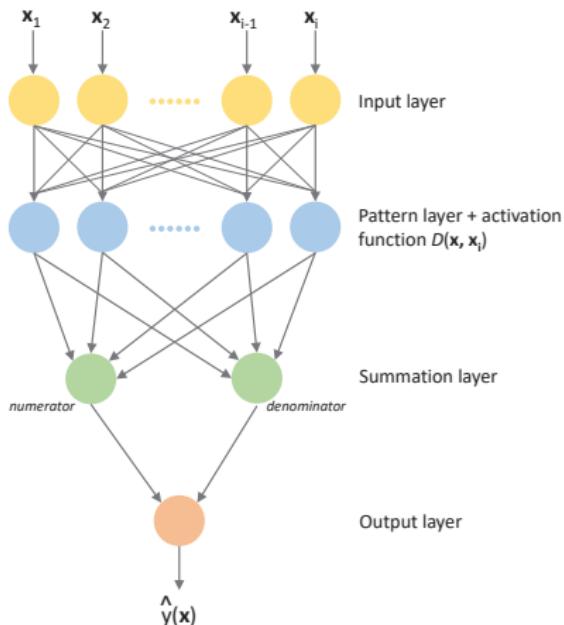
Motivation
○○○○○○○○

GRNN
○●○

FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○
○
○○

General Regression Neural Network



$$\hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^n y_i \exp(-D(\mathbf{x}, \mathbf{x}_i))}{\sum_{i=1}^n \exp(-D(\mathbf{x}, \mathbf{x}_i))}$$

$$D(\mathbf{x}, \mathbf{x}_i) = \sum_{j=1}^p \left(\frac{x_j - x_{ij}}{\sigma} \right)^2$$

- ▶ "**Isotropic**": one bandwidth for all the features
- ▶ "**Anisotropic**": different bandwidth for each feature

Motivation
oooooooo

GRNN
○○●

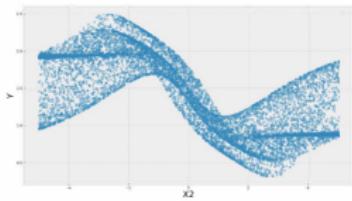
FS with GRNN
○○○○○
○○○

Spatial wind speed prediction
○○○○○
○
○○

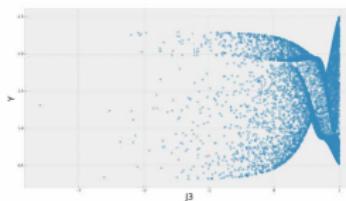
The butterfly dataset

The *butterfly* dataset is constituted by one target variable (Y) and eight features, of which two are relevant (namely X_1 and X_2), three are redundant (J_3 , J_4 and J_5) and three are irrelevant (I_6 , I_7 and I_8).

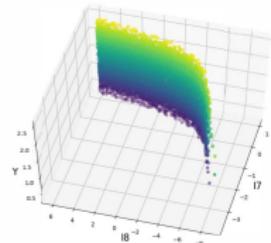
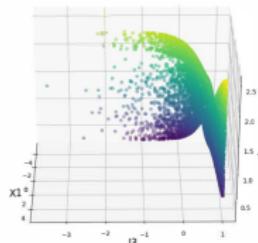
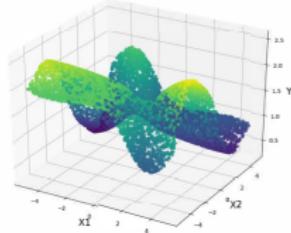
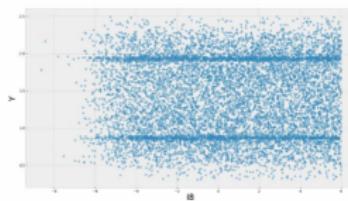
Relevant



Redundant



Irrelevant.



Motivation
oooooooo

GRNN
ooo

FS with GRNN
●ooooo
ooo

Spatial wind speed prediction
oooooooo
o
oo

Isotropic GRNN (wrapper feature selector)

Isotropic Selector (IS)

We use one bandwidth σ for all the variables (features)

- ▶ Plug-in bandwidth = $0.9\hat{\sigma}_j n^{\frac{-1}{(4+p)}}$
- ▶ Low training time!!!
- ▶ We can fastly perform an exhaustive search over the $2^P - 1$ possible combination of features to find the one minimizing the training MSE
- ▶ Limitation: not applicable in case studies with extremely large P

Motivation
oooooooo

GRNN
ooo

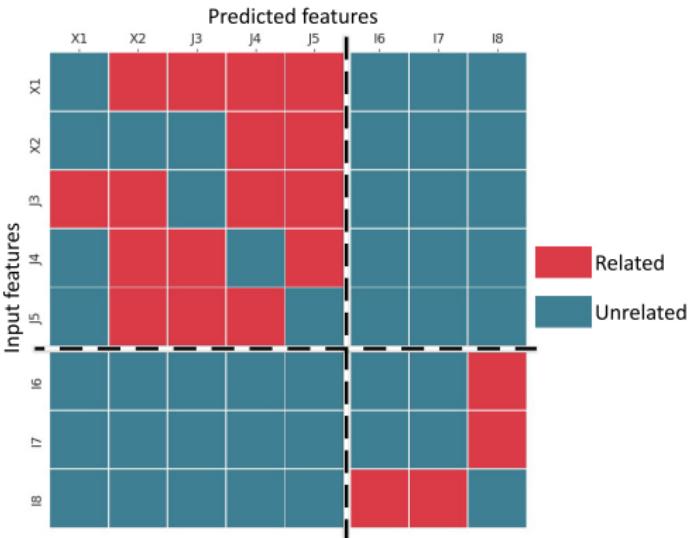
FS with GRNN
○●○○○
○○○

Spatial wind speed prediction
oooooo
○
○○

Isotropic GRNN (wrapper feature selector)

STEP 1: Analysis of dependencies in the input space

Relatidness: non-linear predictability of an input variable using the other features of the input space (without considering Y).



Motivation
oooooooo

GRNN
ooo

FS with GRNN
○○●○○
○○○

Spatial wind speed prediction
○○○○○
○
○○

Isotropic GRNN (wrapper feature selector)

STEP 2: Feature selection

Relevant

Features in the subset minimizing the CV-MSE in an exhaustive search (or a forward-backward feature selection);

Redundant

Features **related** to the relevant ones, i.e. included in the subset minimizing the CV-MSE in an exhaustive search (or a forward-backward feature selection) when using as output the relevant features;

Irrelevant

Features which are neither relevant nor redundant.

Motivation
oooooooo

GRNN
ooo

FS with GRNN
oooo●o
○○○

Spatial wind speed prediction
oooooo
○
○○

Isotropic GRNN (wrapper feature selector)

Feature selection, *butterfly* dataset (1)

% of noise	relevant	redundant	irrelevant
10%	(X_1, X_2)	(J_3, J_4, J_5)	(I_6, I_7, I_8)
20%	(X_1, X_2, J_4)	(J_3, J_4, J_5)	(I_6, I_7, I_8)
30%	(X_1, X_2, J_4)	(J_3, J_5)	(I_6, I_7, I_8)
40%	(X_1, X_2, J_4)	(J_3, J_5)	(I_6, I_7, I_8)
50%	(X_1, X_2, J_4)	(J_3, J_5)	(I_6, I_7, I_8)
75%	(X_1, X_2, J_4, J_5)	(J_3)	(I_6, I_7, I_8)
100%	(X_1, X_2, J_4, J_5)	(J_3, I_7)	(I_6, I_8)

Table: Result of the feature selection using IS on the butterfly dataset with **different percentages of noise injected in the input**

Motivation
oooooooo

GRNN
ooo

FS with GRNN
oooo●
ooo

Spatial wind speed prediction
oooooo
o
oo

Isotropic GRNN (wrapper feature selector)

Feature selection, *butterfly* dataset (2)

% of noise	relevant features [occurrences]
10%	$(X_1, X_2)[83], (X_1, X_2, J_4)[17]$
20%	$(X_1, X_2)[83], (X_1, X_2, J_4)[17]$
25%	$(X_1, X_2)[84], (X_1, X_2, J_4)[16]$
50%	$(X_1, X_2)[84], (X_1, X_2, J_4)[16]$
75%	$(X_1, X_2)[78], (X_1, X_2, J_4)[22]$
100%	$(X_1, X_2)[72], (X_1, X_2, J_4)[24], (X_1, X_2, J_3)[3], (X_2, J_4)[1]$

Table: Result of the feature selection using IS on the butterfly dataset with **different percentages of noise injected in the output**.

For each level of noise, 100 dataset have been generated and the feature selection has been repeated for each of them.

The number of times that each combination of features was recognised as relevant is also indicated.

Motivation
○○○○○○○

GRNN
○○○

FS with GRNN
○○○○○
●○○

Spatial wind speed prediction
○○○○○
○
○○

Anisotropic GRNN (embedded feature weighter)

Anisotropic selector (AS)

In the *anisotropic* case, we consider a different bandwidth for each dimension (feature).

$$\hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^n y_i \exp(-D(\mathbf{x}, \mathbf{x}_i))}{\sum_{i=1}^n \exp(-D(\mathbf{x}, \mathbf{x}_i))} \quad (1)$$

$$D(\mathbf{x}, \mathbf{x}_i) = \sum_{j=1}^p \left(\frac{x_j - x_{ij}}{\sigma_j} \right)^2 \quad (2)$$

Irrelevant

The j^{th} variable is irrelevant if σ_j is higher than a \sqrt{p} (i.e. the maximum euclidean distance between two points in the hyper-cube defined by the X_p inputs)

Motivation
○○○○○○○

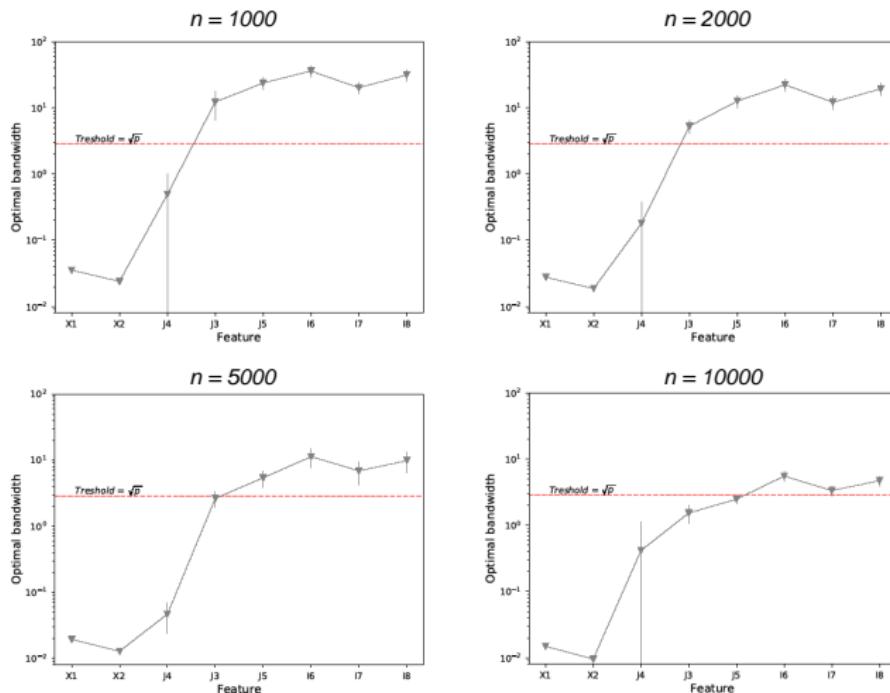
GRNN
○○○

FS with GRNN
○○○○○
○●○

Spatial wind speed prediction
○○○○○
○○
○○

Anisotropic GRNN (embedded feature weighter)

Feature weighting, *butterfly* dataset



Anisotropic GRNN (embedded feature weighter)

Feature shuffling, *butterfly* dataset

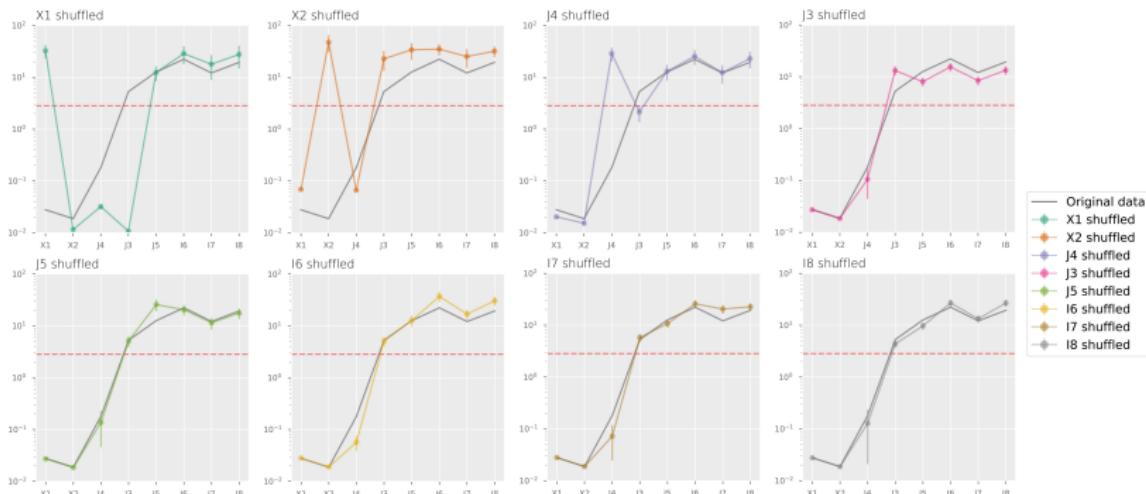


Figure: Mean and 95% confidence interval for the optimal bandwidth values after shuffling each feature on a butterfly dataset having $n = 2000$.

Motivation
oooooooo

GRNN
ooo

FS with GRNN
ooooo
ooo

Spatial wind speed prediction
●ooooo
○
○○

Case study: monthly wind speed spatial prediction in Switzerland

Spatial prediction of wind speed in Switzerland

Modelling 60 minutes frequency data from 118 stations in Switzerland from 2012 to 2013

- ▶ spatially-model the phenomena
- ▶ We will consider two frequency: hourly data and daily data
- ▶ Feature space includes 13 spatial features
- ▶ Is FS part of the preprocessing or do we have to include it dynamically in the modelling stage? (time-transferrability of the model)

Motivation
ooooooooo

GRNN
ooo

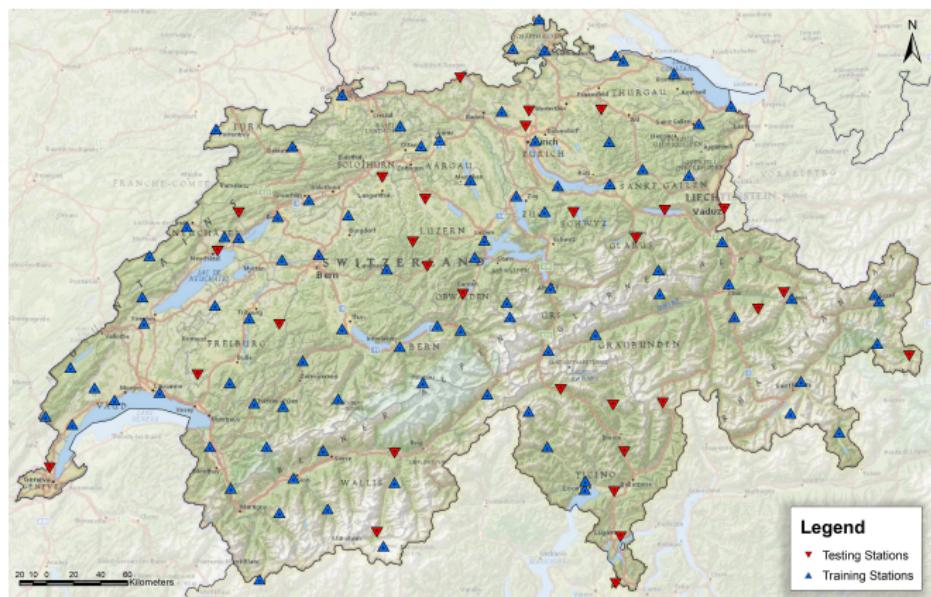
FS with GRNN
ooooo
ooo

Spatial wind speed prediction
○●oooo
○○○○○

Case study: monthly wind speed spatial prediction in Switzerland

The MeteoSwiss network

The 118 station were randomly splitted into training and testing



Motivation
ooooooooo

GRNN
ooo

FS with GRNN
ooooo
ooo

Spatial wind speed prediction
oooooo
o
o
oo

Case study: monthly wind speed spatial prediction in Switzerland

Spatial features

A 13-d input space has been designed including the geographical space (latitude, longitude and elevation) and features derived by applying filters and derivatives on the digital elevation model.

Feature	Bandwidth
Latitude	-
Longitude	-
Elevation	-
Difference of Gaussian (large scale)	$\sigma_1 = 3.75 \text{ km}, \sigma_2 = 5.55 \text{ km}$
Difference of Gaussian (medium scale)	$\sigma_1 = 1.75 \text{ km}, \sigma_2 = 2.25 \text{ km}$
Difference of Gaussian (small scale)	$\sigma_1 = 0.25 \text{ km}, \sigma_2 = 0.05 \text{ km}$
Slope (large scale)	$\sigma = 3.75 \text{ km}$
Slope (medium scale)	$\sigma = 1.75 \text{ km}$
Slope (small scale)	$\sigma = 0.20 \text{ km}$
Directional derivative E-W (medium scale)	$\sigma = 1.75 \text{ km}$
Directional derivative E-W (small scale)	$\sigma = 0.25 \text{ km}$
Directional derivative N-S (medium scale)	$\sigma = 1.75 \text{ km}$
Directional derivative N-S (small scale)	$\sigma = 0.25 \text{ km}$

Motivation
ooooooooo

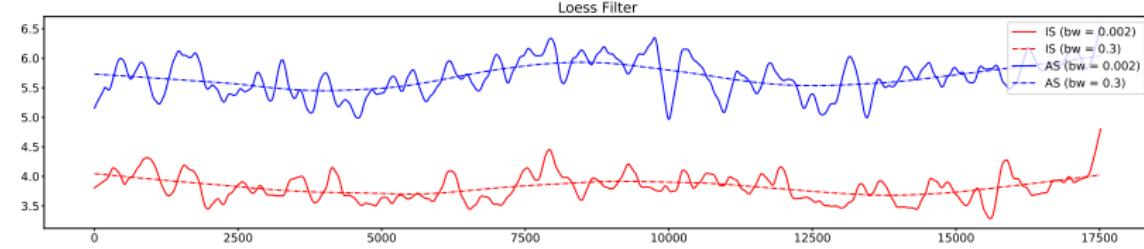
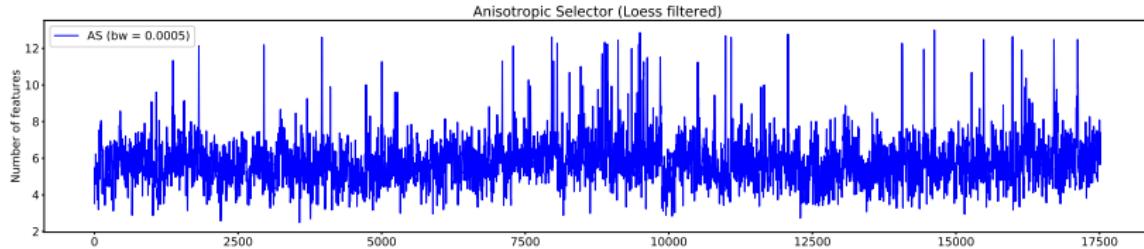
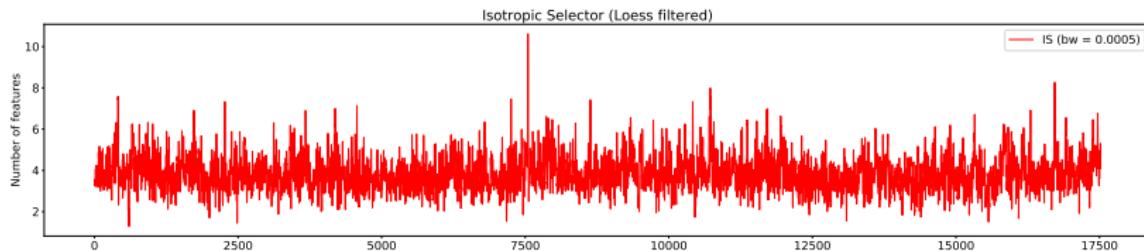
GRNN
ooo

FS with GRNN
ooooo
ooo

Spatial wind speed prediction
oooooooo
o
oo

Case study: monthly wind speed spatial prediction in Switzerland

Hourly data - Number of Selected Features



Motivation



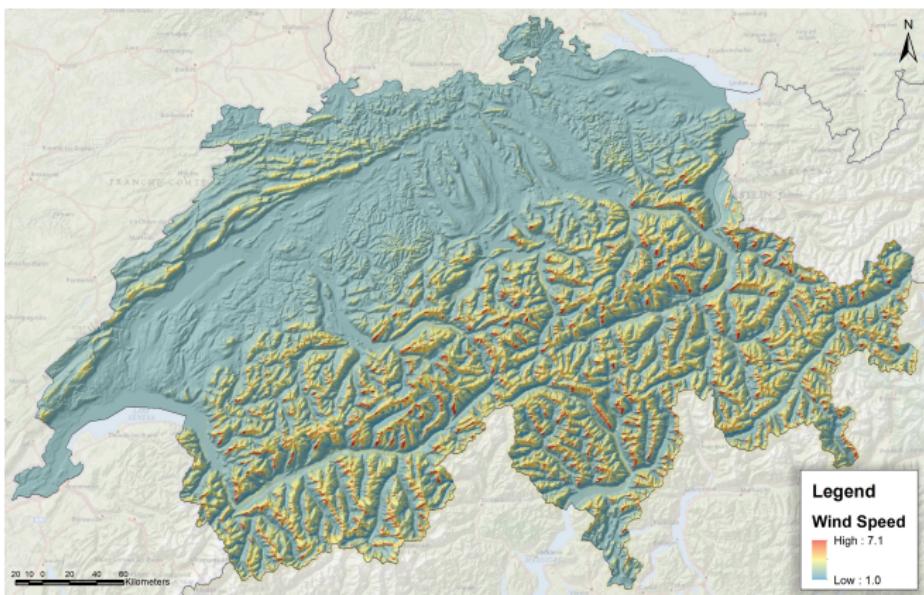
GRNN
000

Spatial wind speed prediction

○○○○●○
○
○○

Spatial prediction (automatic mapping)

Wind speed prediction for April, 1 2012 12:00 using the best subset of features obtained with the Isotropic selector.



Motivation
oooooooo

GRNN
ooo

FS with GRNN
ooooo
ooo

Spatial wind speed prediction
oooo●
○
○○

Case study: monthly wind speed spatial prediction in Switzerland

Conclusions

- ▶ The curse of dimensionality imposes to find new solution to reduce the dimensionality of our data
- ▶ We proposed two method based on GRNN to remove the non relevant variables from the input space (IS, AS)
- ▶ The application to wind speed showed how FS has to be intended as a dynamical stage in the modelling process

Motivation
oooooooo

GRNN
ooo

FS with GRNN
ooooo
ooo

Spatial wind speed prediction
oooooo
●
○○

Software availability

PyGRNN

The package *pyGRNN* has been implemented in Python and it is freely available.

- ▶ GitHub: <https://github.com/federhub/pyGRNN>
- ▶ Install in Python using *pip install PyGRNN*

A reference paper is available on Arxiv



Motivation
oooooooo

GRNN
ooo

FS with GRNN
ooooo
ooo

Spatial wind speed prediction
oooooo
o
●○

Software availability

Other dimensionality reduction tools

On the GeoKDD Lab webpage (<https://wp.unil.ch/geokdd>) you will find other resources for Dimensionality Reduction:

- ▶ **IDmining** - *Intrinsic Dimension for Data Mining*. Contains techniques for mining large and high-dimensional data sets by using the concept of Intrinsic Dimension. Author: Jean Golay
- ▶ **SFtools** - *Space Filling Based Tools for Data Mining*. Unsupervised dimensionality reduction tool to remove redundancy. Author: Mohamed Laib

Motivation
oooooooo

GRNN
ooo

FS with GRNN
ooooo
ooo

Spatial wind speed prediction
oooooo
o
o●

Software availability

