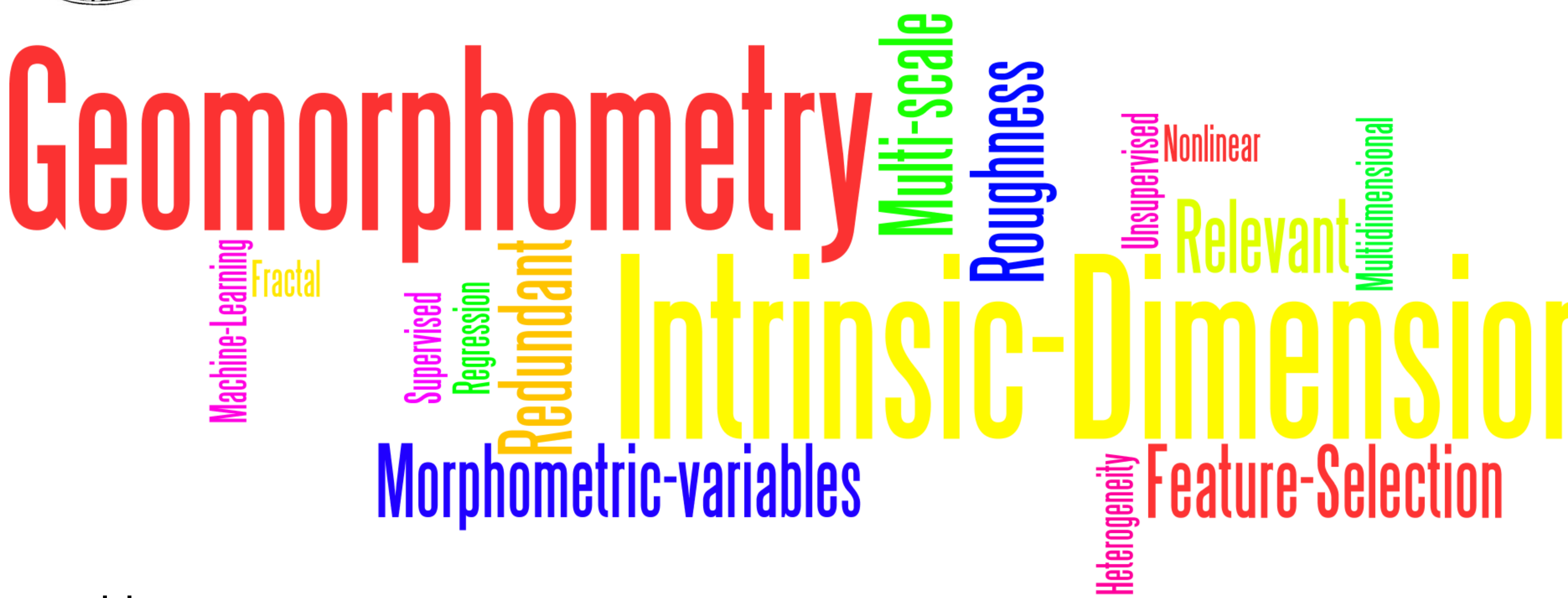
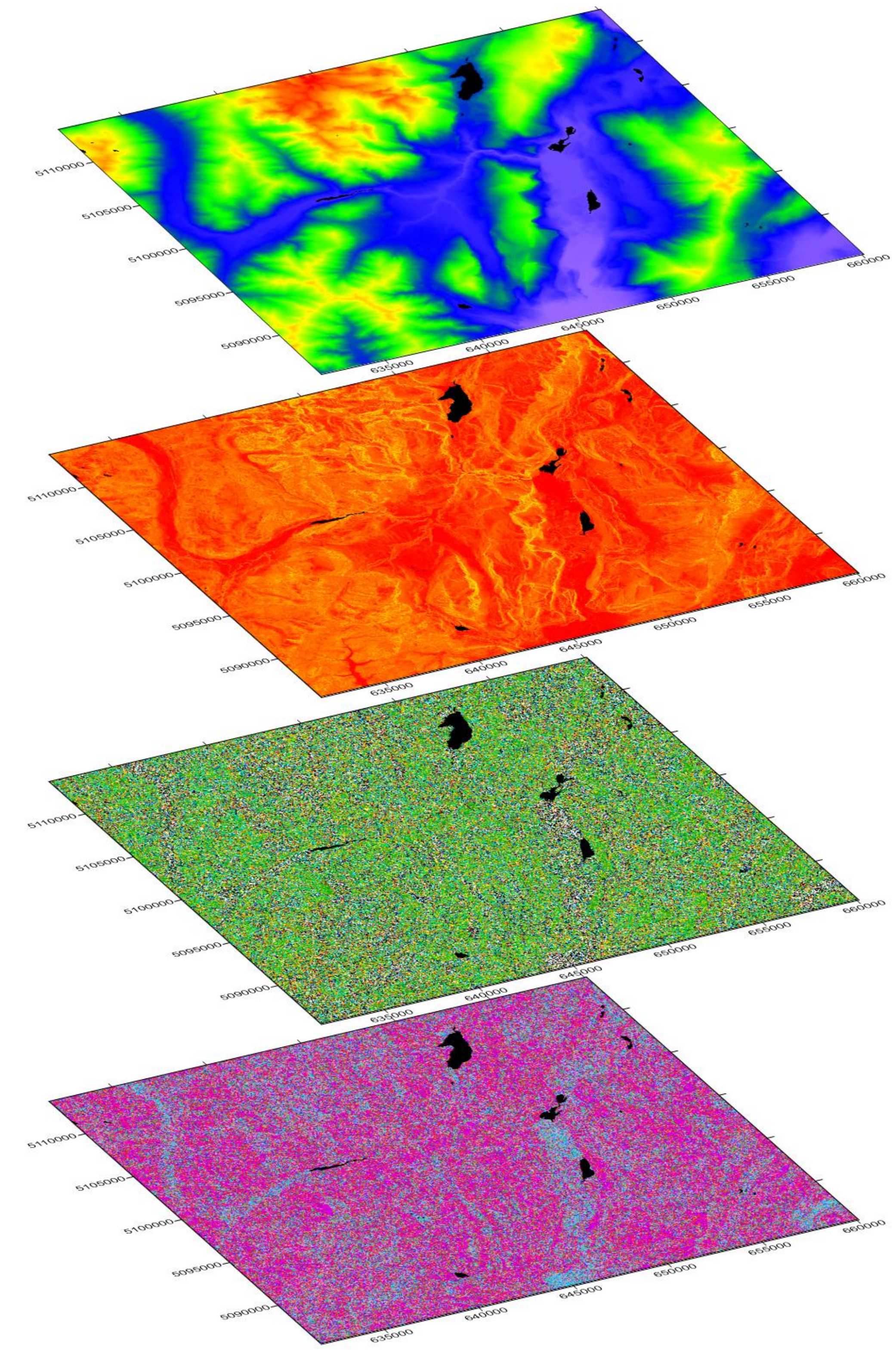




Preamble

The quantitative analysis of digital elevation models often generates **high-dimensional** datasets. This is related both to the high number of morphometric variables and local statistical metrics that can be computed as well as to the spatial-scale dependency (including directionality) inherent to geomorphometric analysis.

A too high number of **geomorphometric features** can impact **supervised** (e.g., landslide susceptibility mapping) and **unsupervised** approaches (e.g., landscape classification), increasing computational cost and reducing accuracy. Moreover, the detection of relevant features is particularly interesting for explorative analyses purposes.

Accordingly, the discrimination between **relevant, irrelevant** and **redundant** features is of fundamental importance in many geocomputational tasks

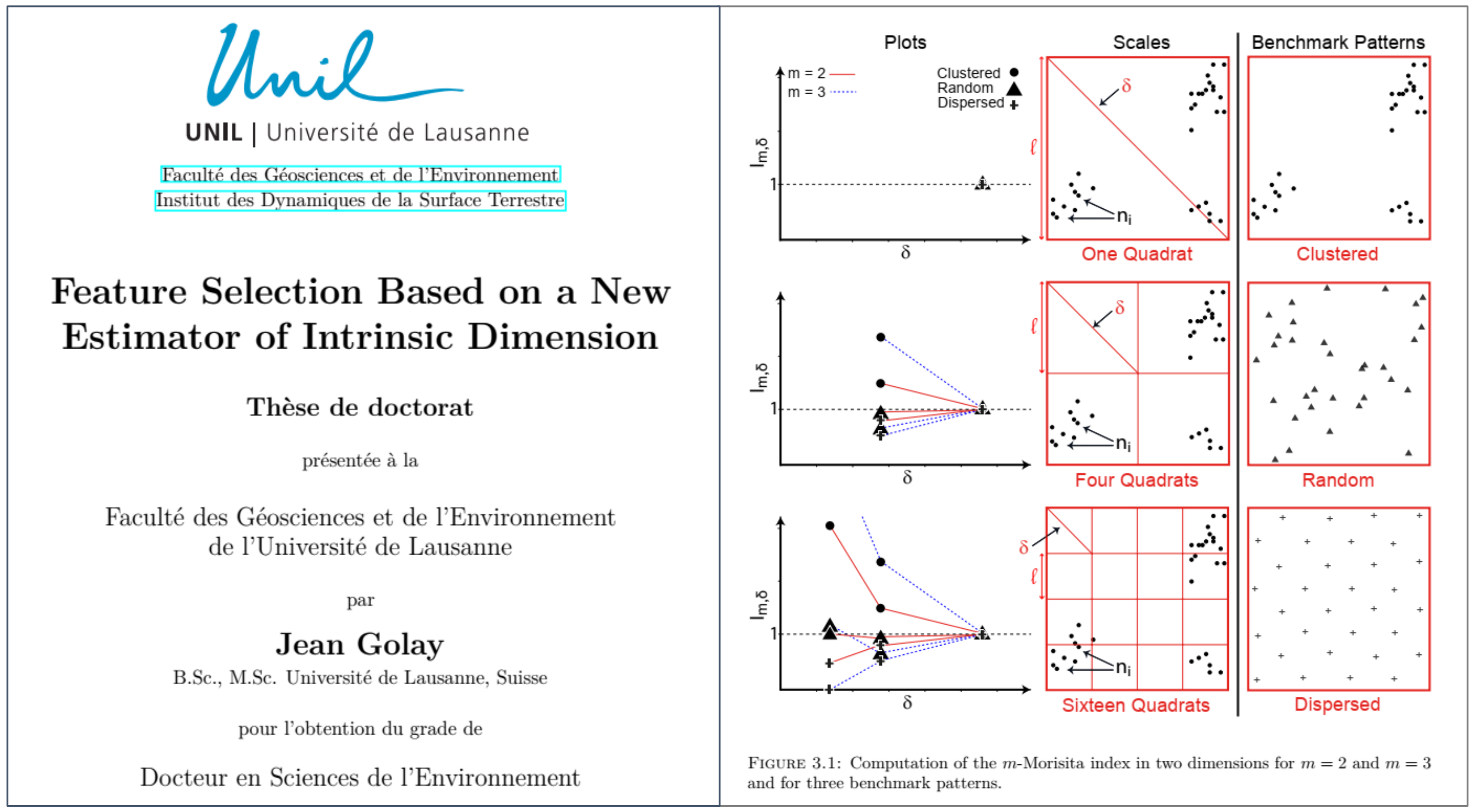
Sebastiano Trevisani - University IUAV of Venice – mail: strevisani@iuav.it

From Morisita index to Intrinsic Dimension

The recently developed **fractal-based** estimator of **Intrinsic Dimension** (**ID**), relying on a generalization of **Morisita Index** is promising in the context of feature selection

Morisita, M. 1962, "**Iσ-Index, a measure of dispersion of individuals**", Researches on Population Ecology, vol. 4, no. 1, pp. 1-7.

$$I_\delta = Q\,\frac{\sum_{i=1}^{Q} n_i(n_i-1)}{N(N-1)}$$


Unil
UNIL | Université de Lausanne
Faculté des Géosciences et de l'Environnement
Institut des Dynamiques de la Surface Terrestre

Feature Selection Based on a New Estimator of Intrinsic Dimension

Thèse de doctorat

présentée à la

Faculté des Géosciences et de l'Environnement
de l'Université de Lausanne

par

Jean Golay

B.Sc., M.Sc. Université de Lausanne, Suisse

pour l'obtention du grade de

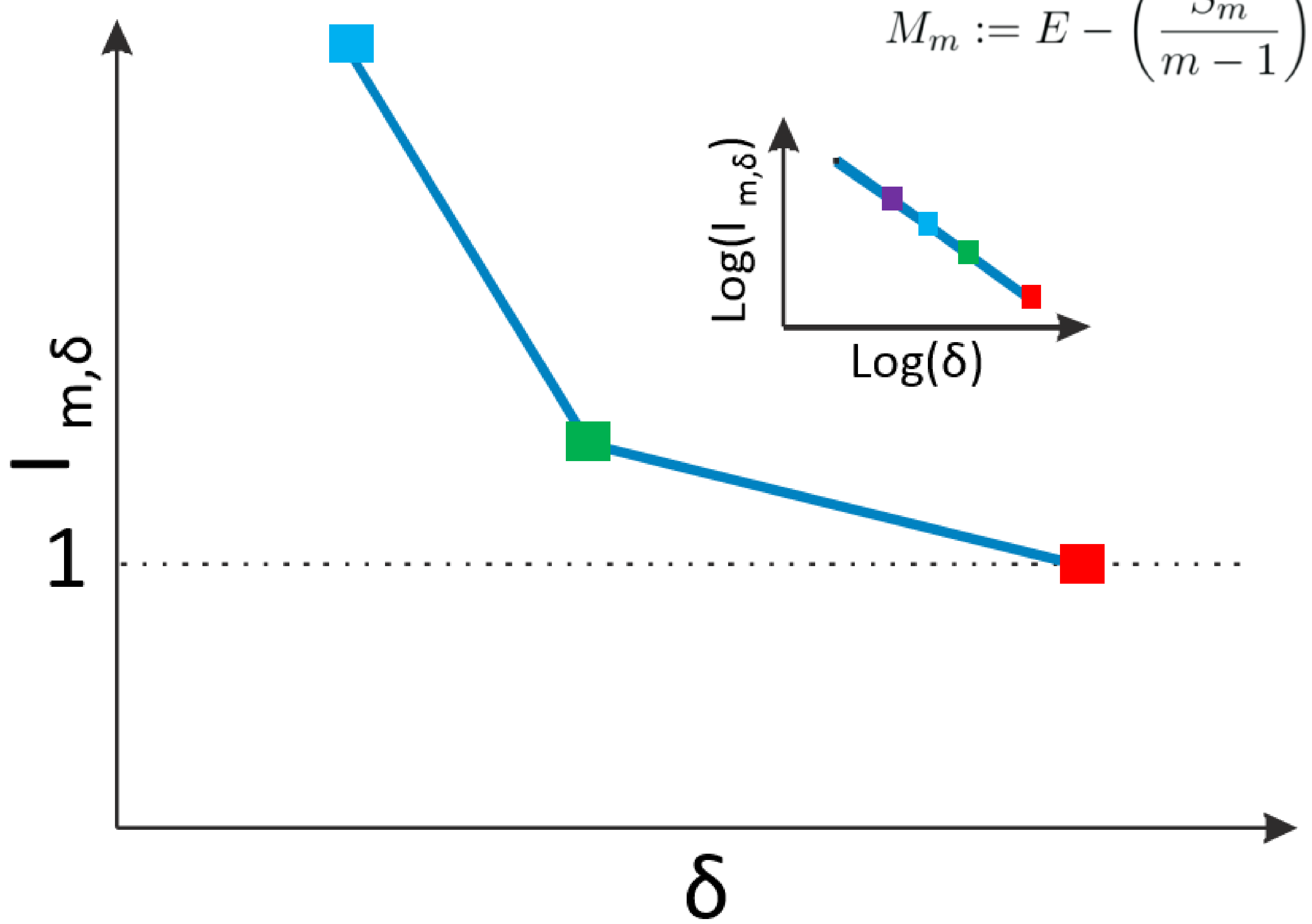
Docteur en Sciences de l'Environnement

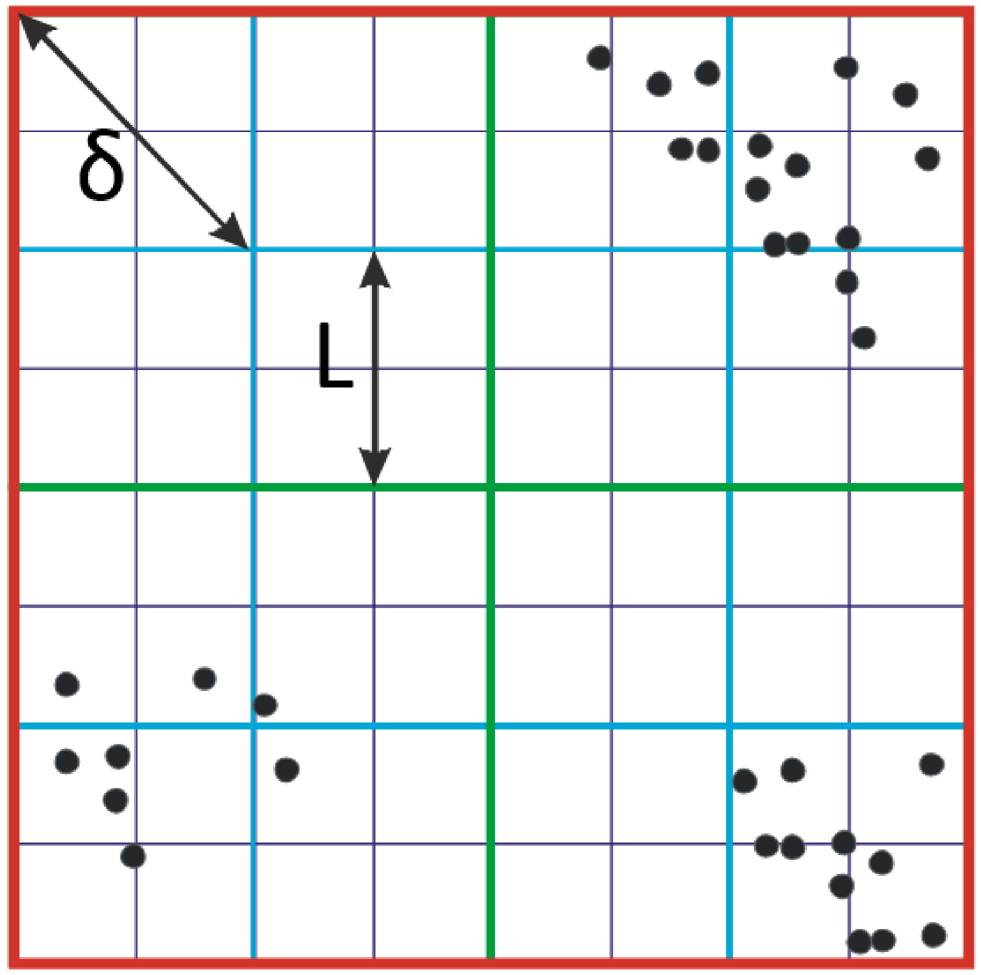
FIGURE 3.1: Computation of the m-Morisita index in two dimensions for $m = 2$ and $m = 3$ and for three benchmark patterns.

Golay, J. & Kanevski, M. 2015, "**A new estimator of intrinsic dimension based on the multipoint Morisita index**", Pattern Recognition, vol. 48, no. 12, pp. 4070-4081

$$I_{m,\delta} = Q^{m-1}\,\frac{\sum_{i=1}^{Q} n_i(n_i-1)(n_i-2)\cdots(n_i-m+1)}{N(N-1)(N-2)\cdots(N-m+1)}$$

$$M_m := E - \left(\frac{S_m}{m-1}\right)$$



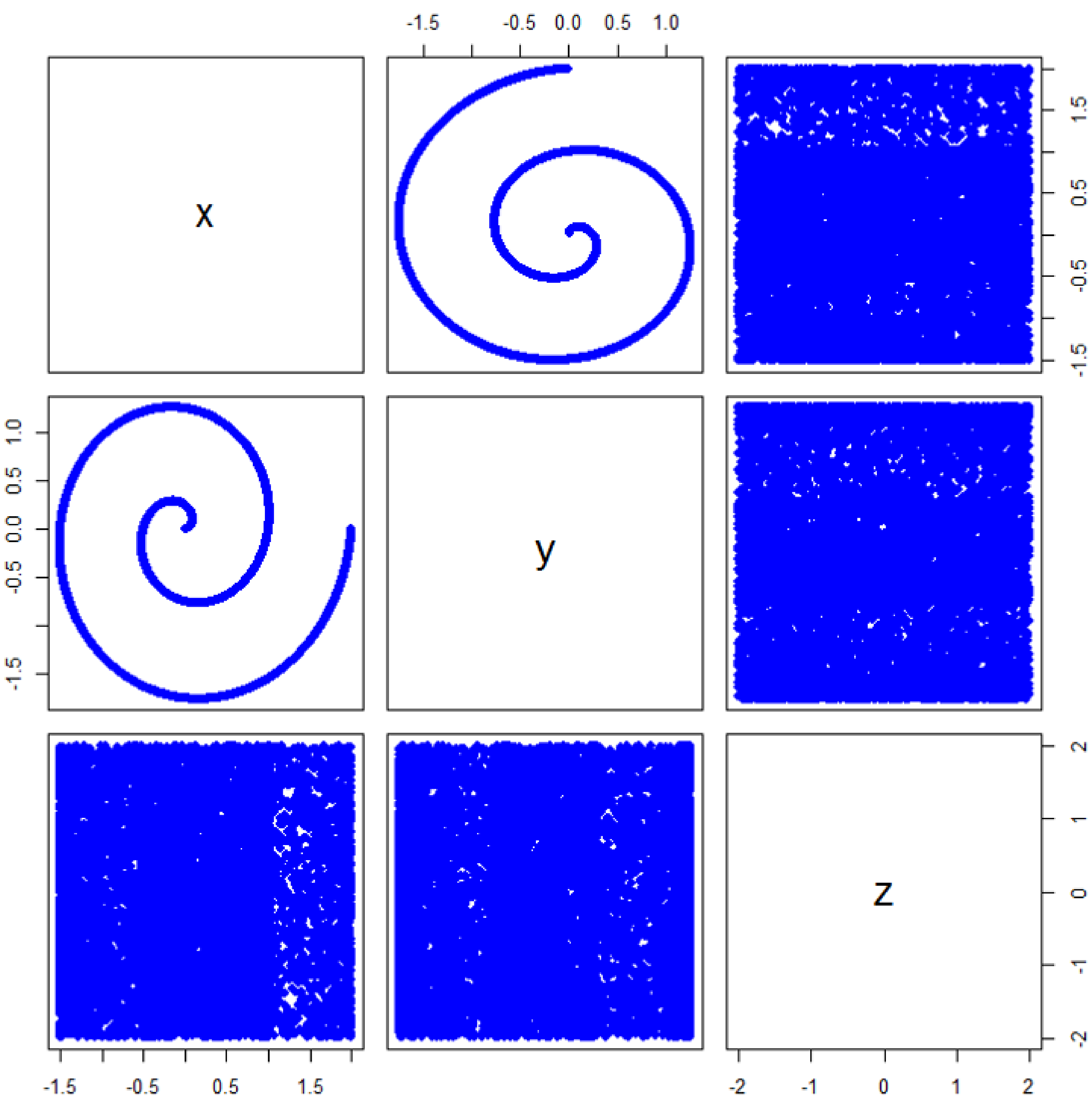


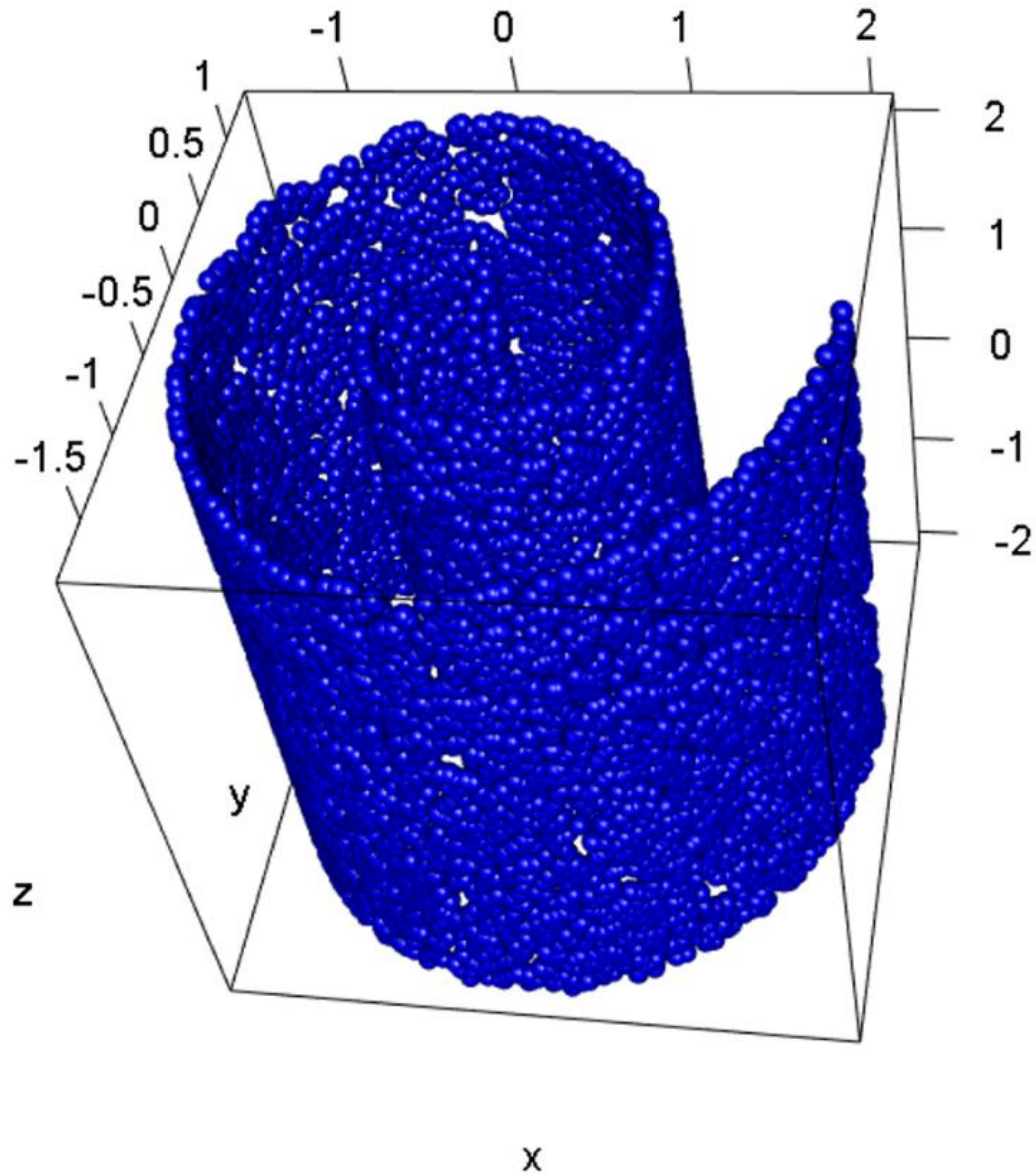
Euclidean space dimension -> E = 3

Intrinsic dimension -> ID = 2

The "**Swiss roll**" (from **R** package "**IDmining**"):

```
x1 <- runif(N, min = -1, max = 1)
x2 <- runif(N, min = -1, max = 1)
x <- sqrt(2 + 2 * x1) * cos(2 * pi * sqrt(2 + 2 * x1))
y <- sqrt(2 + 2 * x1) * sin(2 * pi * sqrt(2 + 2 * x1))
z <- 2 * x2
```





The measures of **ID** can be used for defining **non redundant** and **relevant** features. For example, in an unsupervised setting, redundant features do not contribute to increase the ID. In synthesis, the key idea of ID-based features selection algorithms relies on evaluating the impact of single (or combination of) features on the **ID** of the dataset.

IDmining: Intrinsic Dimension for Data Mining

Contains techniques for mining large and high-dimensional data sets by using the concept of Intrinsic Dimension (ID). Here the ID is not necessarily an integer. It is extended to fractal dimensions. And the Morisita estimator is used for the ID estimation, but other tools are included as well.

Version:	1.0.7
Imports:	data.table, doParallel, parallel, foreach, stats, utils
Published:	2021-05-03
Author:	Jean Golay [aut, cre], Mohamed Laib [aut]
Maintainer:	Jean Golay <jeangolay at gmail.com>
License:	CC BY-NC-SA 4.0
URL:	https://www.sites.google.com/site/jeangolayresearch/
NeedsCompilation:	no
CRAN checks:	IDmining results

Downloads:

The authors of the new ID estimator developed a set of **ID-based algorithms** for **feature selection** both in **unsupervised** as well as in **supervised** learning settings. The tools are implemented in R programming environment (package Idmining).

The algorithms are designed for the analysis of **continuous variables**.

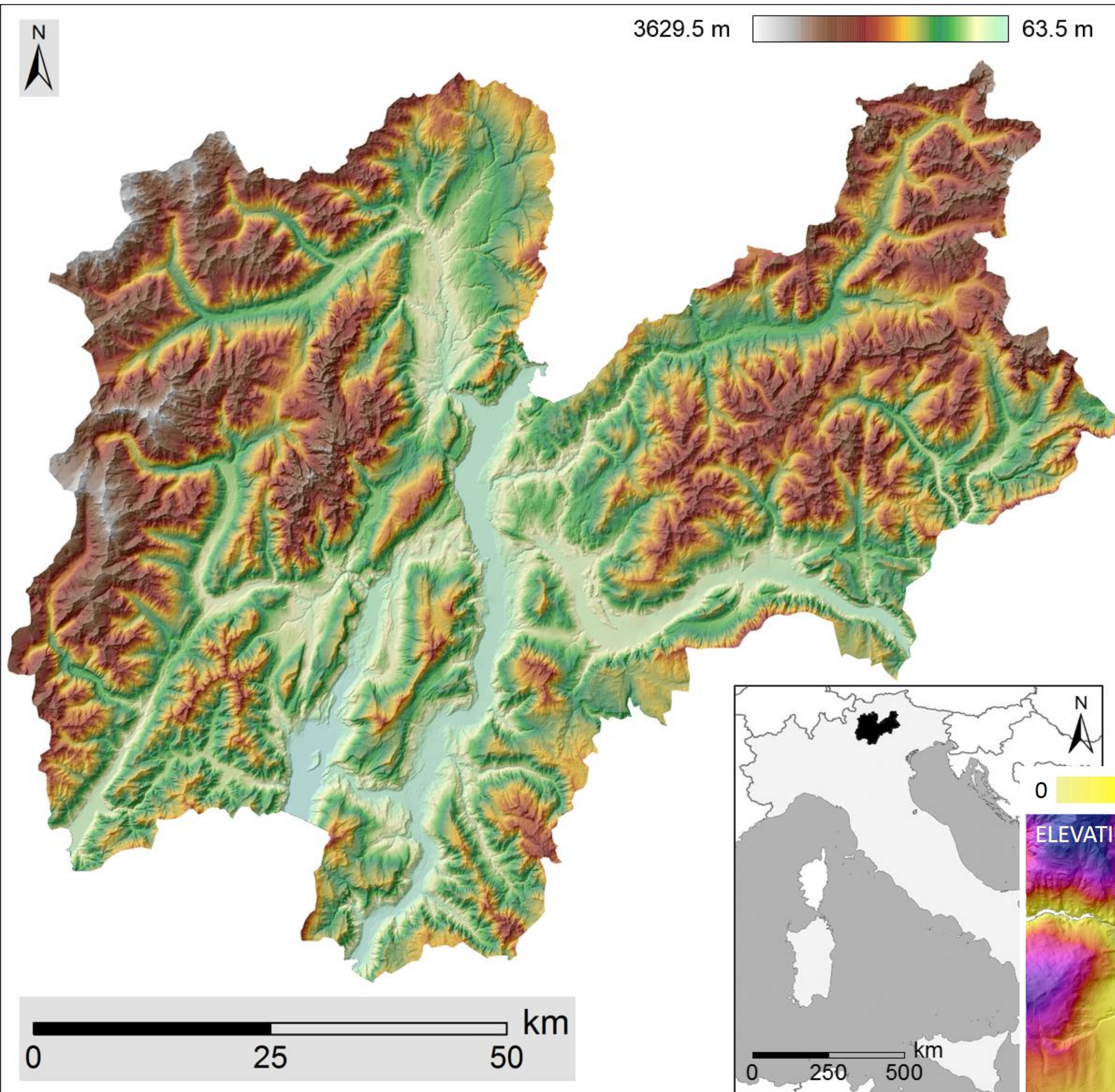
Differently from other approaches, ID-based algorithms do not create new variables for reducing the dimensionality of the data (e.g., as in Principal Component Analysis).

Golay, J. & Kanevski, M. 2017, "**Unsupervised feature selection based on the Morisita estimator of intrinsic dimension**", Knowledge-Based Systems, vol. 135, pp. 125-134.

Golay, J., Leuenberger, M. & Kanevski, M. 2017, "**Feature selection for regression problems based on the Morisita estimator of intrinsic dimension**", Pattern Recognition, vol. 70, pp. 126-138.

Geomorphometric feature selection based on intrinsic dimension estimation: an unsupervised test

Currently, we are exploring various aspects on the application of ID-based algorithms for unsupervised feature selection, considering different geomorphometric features and scale-related issues. A glance of first results is reported from a simple application to basic **morphometric variables** and some **roughness**-related indices (based on Median Absolute directional differences, **MAD**).



Coverage: 6500 km^2

DTM:

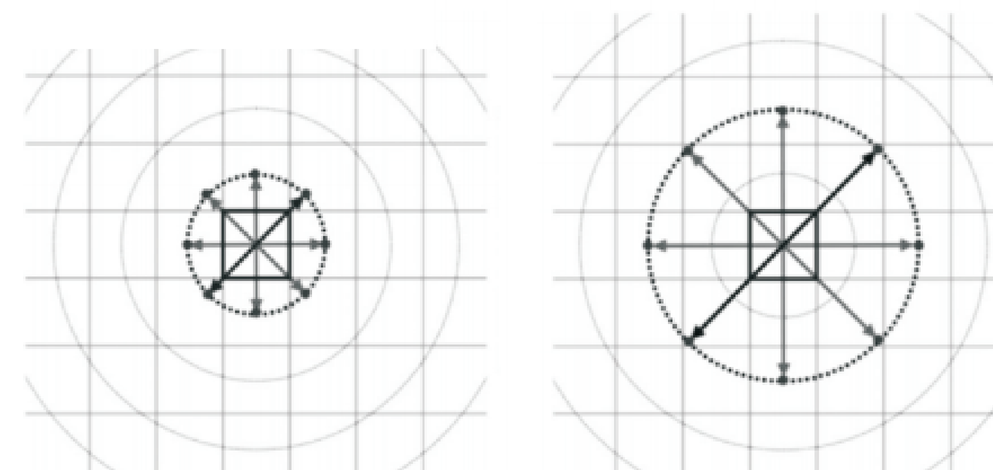
- airborne LiDAR
- resolution considered 25 m (upscaled from 2 m)
- grid nodes: 9971642 (without NAs)

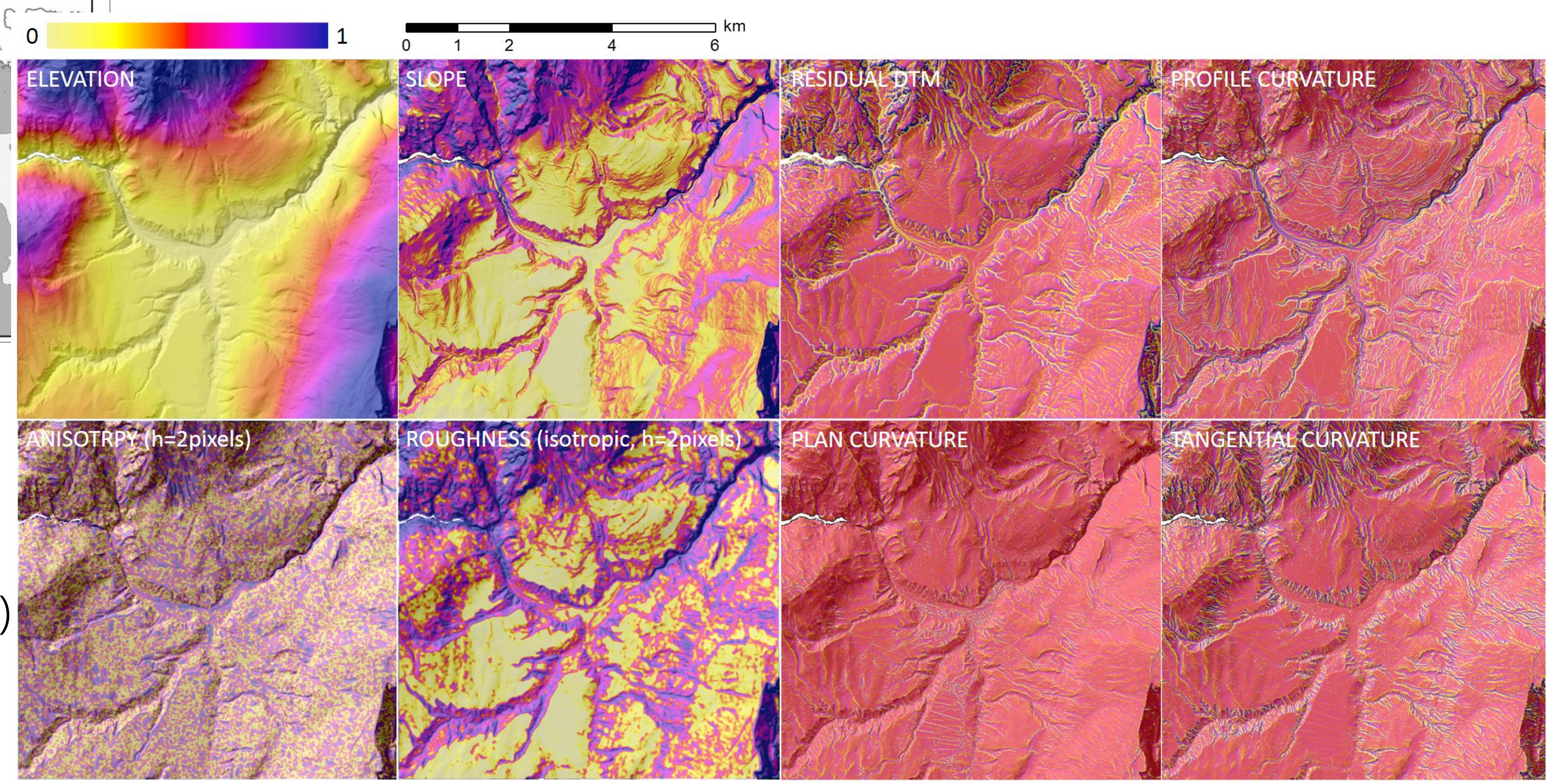
$$\mathrm{MAD}(\mathbf{h}) = Median(|\Delta(\mathbf{h})|) = \begin{cases} |\Delta(\mathbf{h})_{\alpha=(N(\mathbf{h})+1)/2}| & \text{with } N(\mathbf{h}) \text{ odd} \\ 1/2(|\Delta(\mathbf{h})_{\alpha=N(\mathbf{h})/2}| + |\Delta(\mathbf{h})_{\alpha=N(\mathbf{h})/2+1}|) & \text{with } N(\mathbf{h}) \text{ even} \end{cases} \quad (1)$$

where $|\Delta(\mathbf{h})_\alpha| = |z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})|$, with the values $|\Delta(\mathbf{h})_\alpha|$, $\alpha=1,\ldots,N(\mathbf{h})$ sorted into ascending order and $|\Delta(\mathbf{h})|$ the set of the $N(\mathbf{h})$ absolute differences with a separation vector $\mathbf{h}$, i.e., $\{|\Delta(\mathbf{h})_\alpha| \,\|\, \alpha = 1,\ldots,N(\mathbf{h})\}$.

MAD estimator for surface roughness (texture)

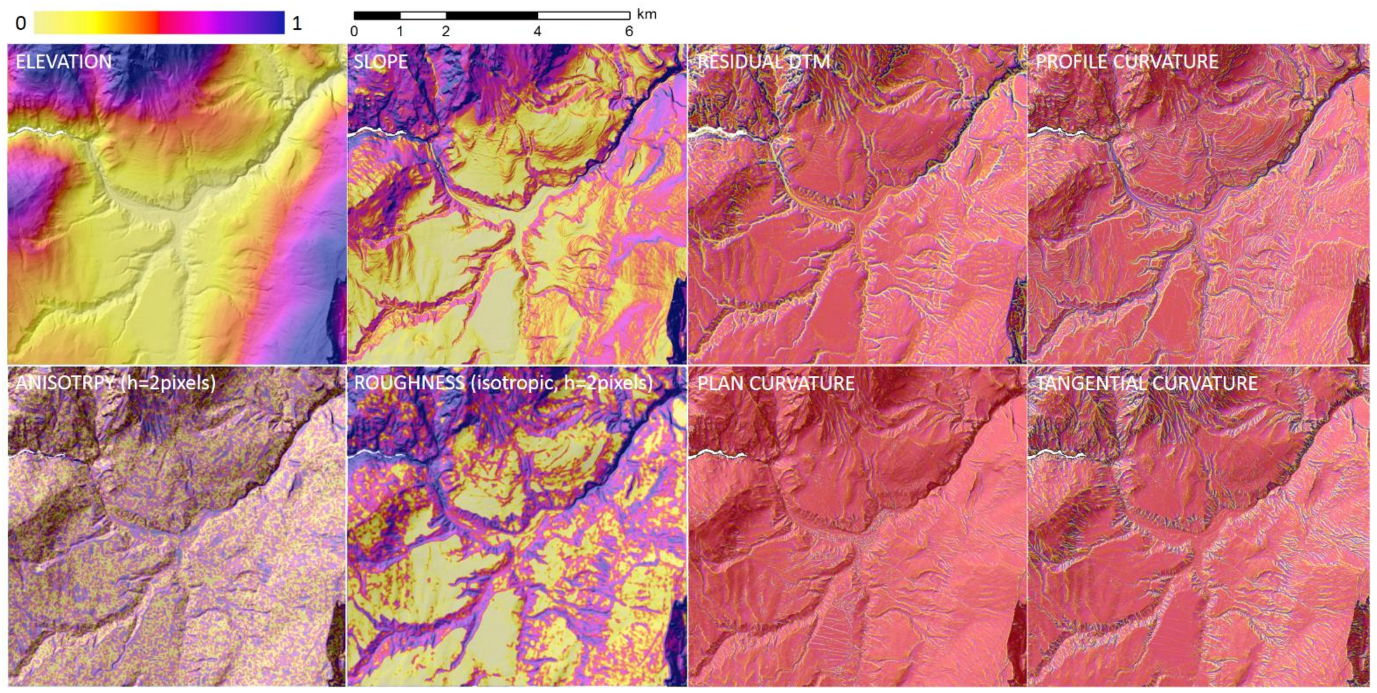
Example of basic kernels used for MAD calculations (see references for details)

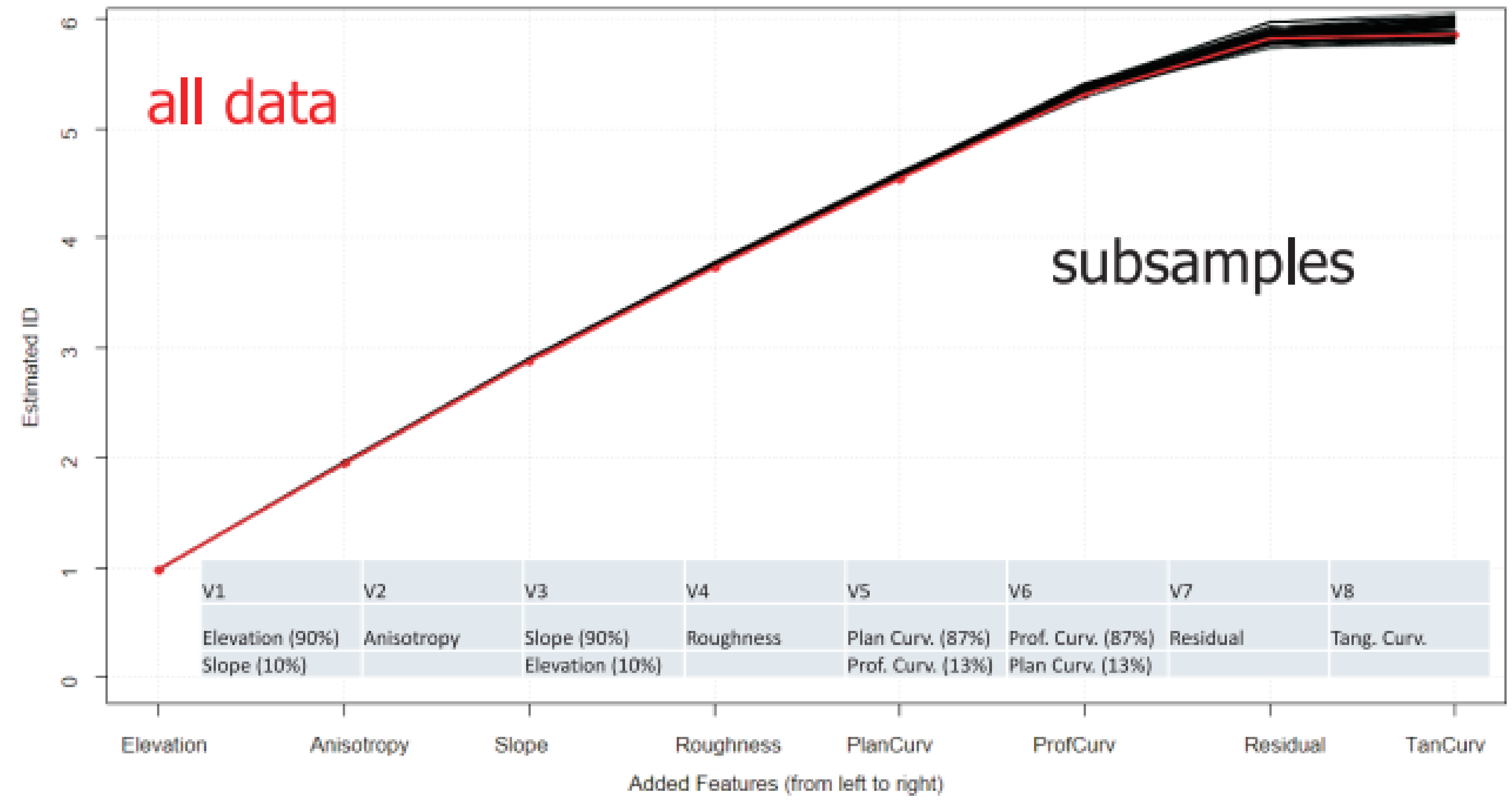




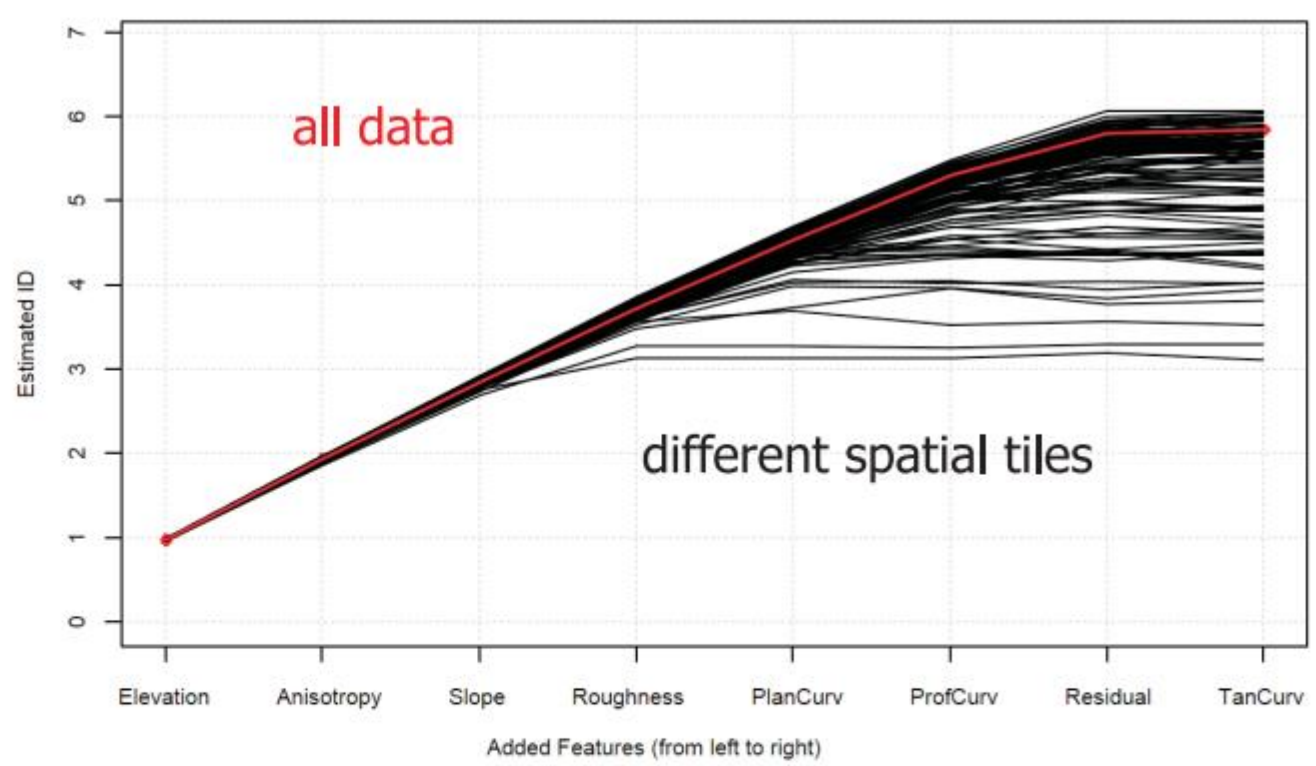
Sebastiano Trevisani - University IUAV of Venice – mail: strevisani@iuav.it

The sensitiveness of subsampling is an aspect to be explored given the high number of nodes, both for computational as well as theoretical reasons (e.g., spatial correlation of features). Another relevant factor is related to the shape of the statistical distribution of variables (skewness, heavy tails, etc.).





Another aspect is related to spatial heterogeneity: an experiment considering tiles of 8 X 8 km^2



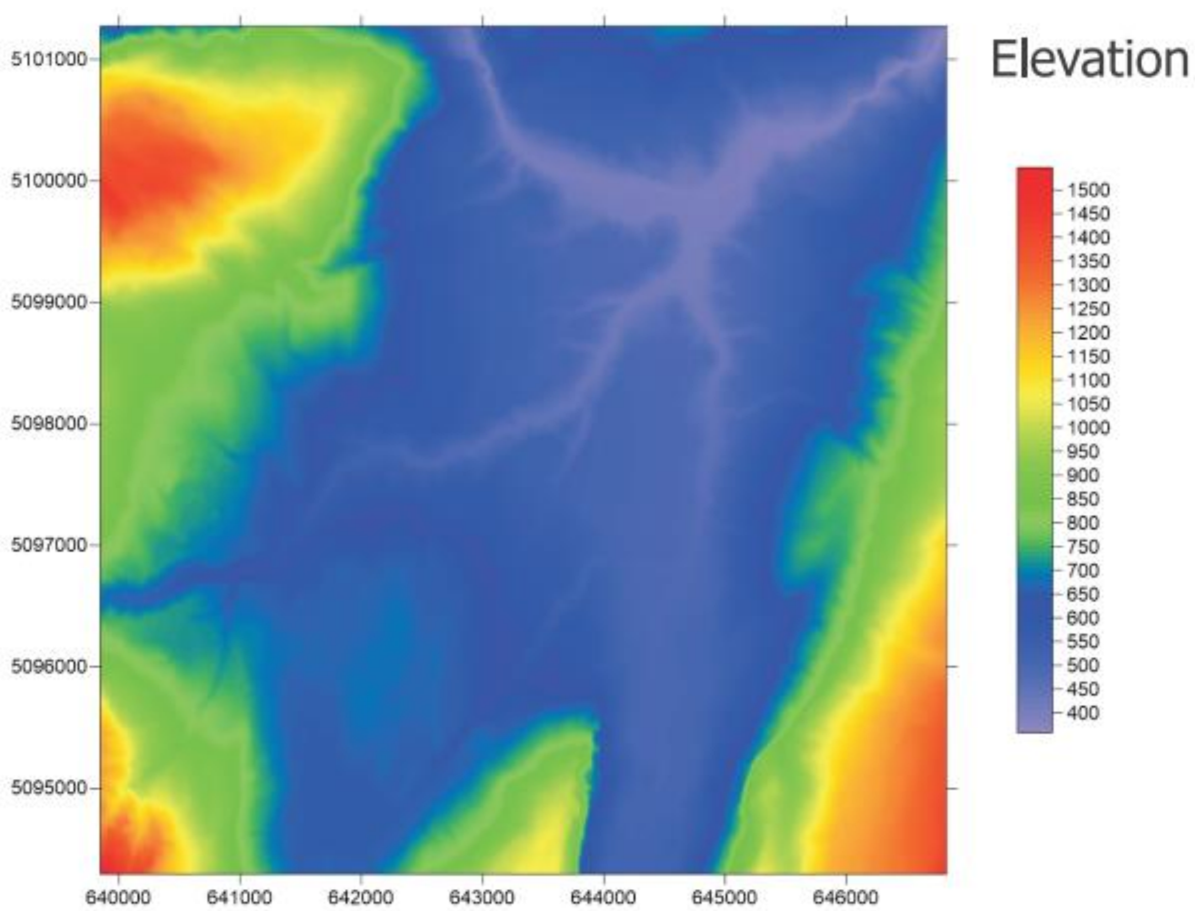
Statistics of selected features obtained applying the approach to the different tiles

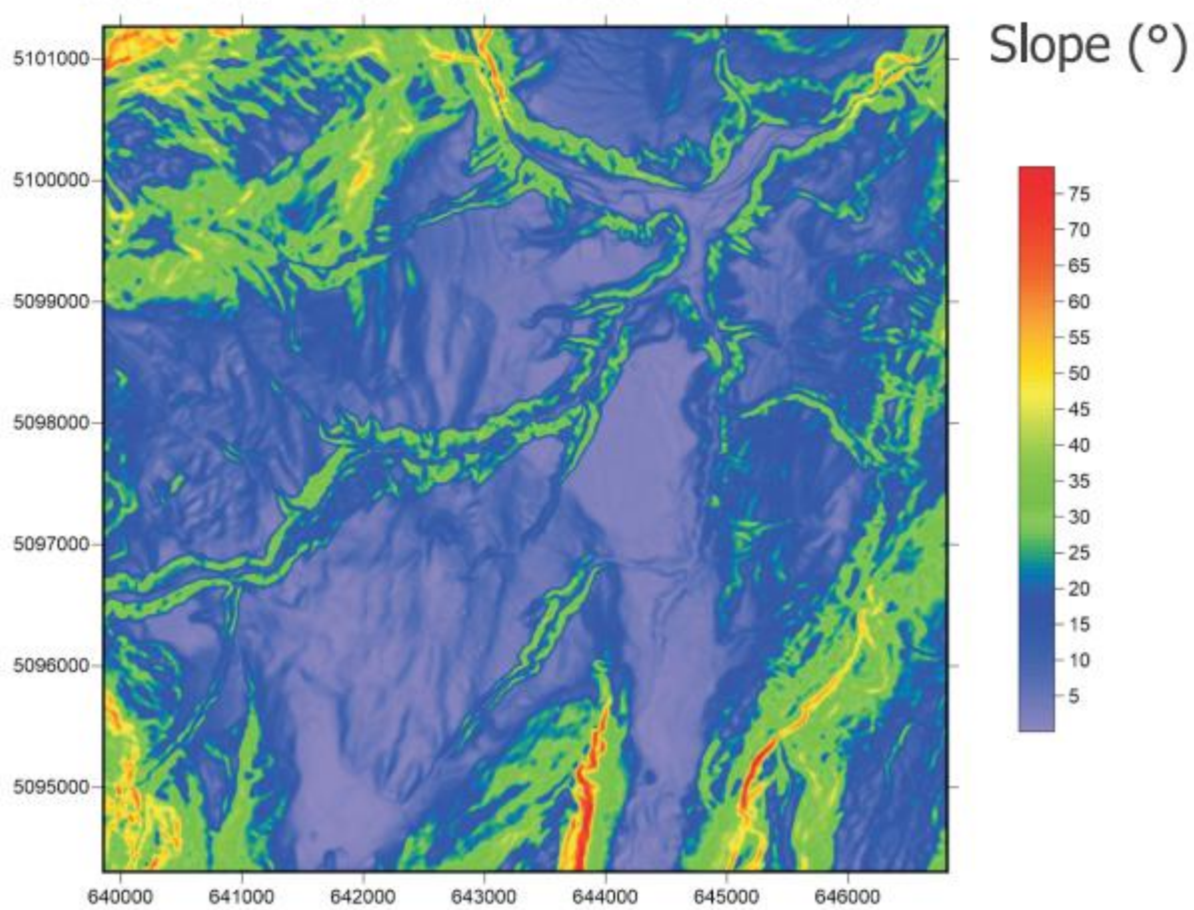
V1	V2	V3	V4	V5	V6	V7	V8
Elevation (44%)	Anisotropy (61%)	Slope (30%)	Roughness (32%)	Roughness (30%)	Plan Curv. (22%)	Residual (54%)	Tang. Curv.(46%)
Anisotropy (38%)	Slope (16%)	Prof. Curv. (26%)	Prof. Curv. (32%)	Prof. Curv. (24%)	Tang. Curv. (19%)	Tang. Curv.(22%)	Plan Curv. (35%)
Slope (15%)	Elevation (13%)	Roughness (20%)	Plan Curv. (14%)	Plan Curv. (18%)	Slope (16%)	Elevation (14%)	Residual (12%)
...	Roughness (9%)	Residual (10%)	Slope (9%)	Tang. Curv.(12%)	Elevation (15%)	...	...
	...	Plan Curv. (8%)	Residual (9%)	Slope (6%)	Prof. Curv. (13%)		
		...	...	...	...		

DECREASING RELEVANCE

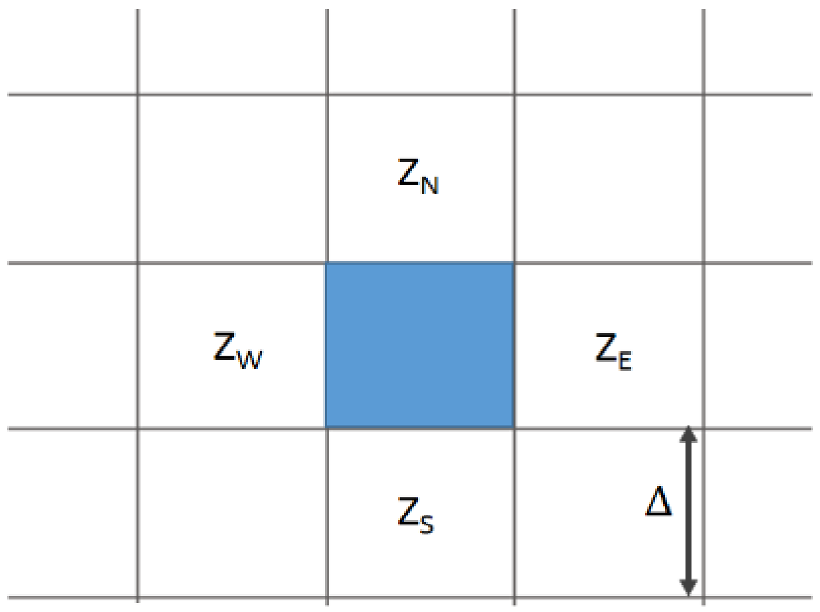
Sebastiano Trevisani - University IUAV of Venice – mail: strevisani@iuav.it

For the supervised test, a **synthetic data set** built from "**real" topography** has been considered; the idea is to generate input features **not linearly** correlated with the output features and **apparently redundant**





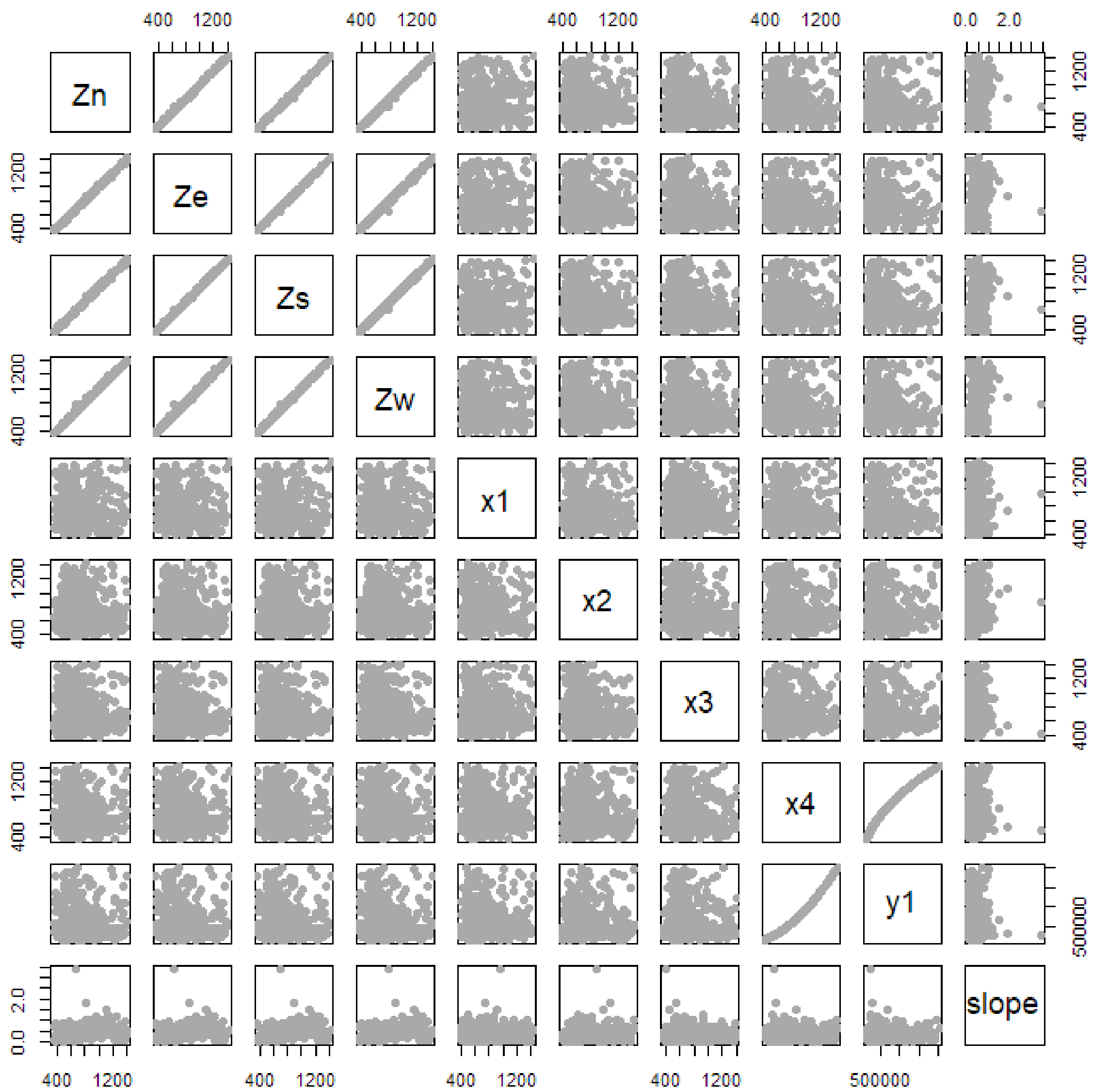
The DTM considered, derived from airborne Lidar technology, is representative of an alpine area with complex morphology and has a grid of 350x350 pixels, with a resolution of 20 m.



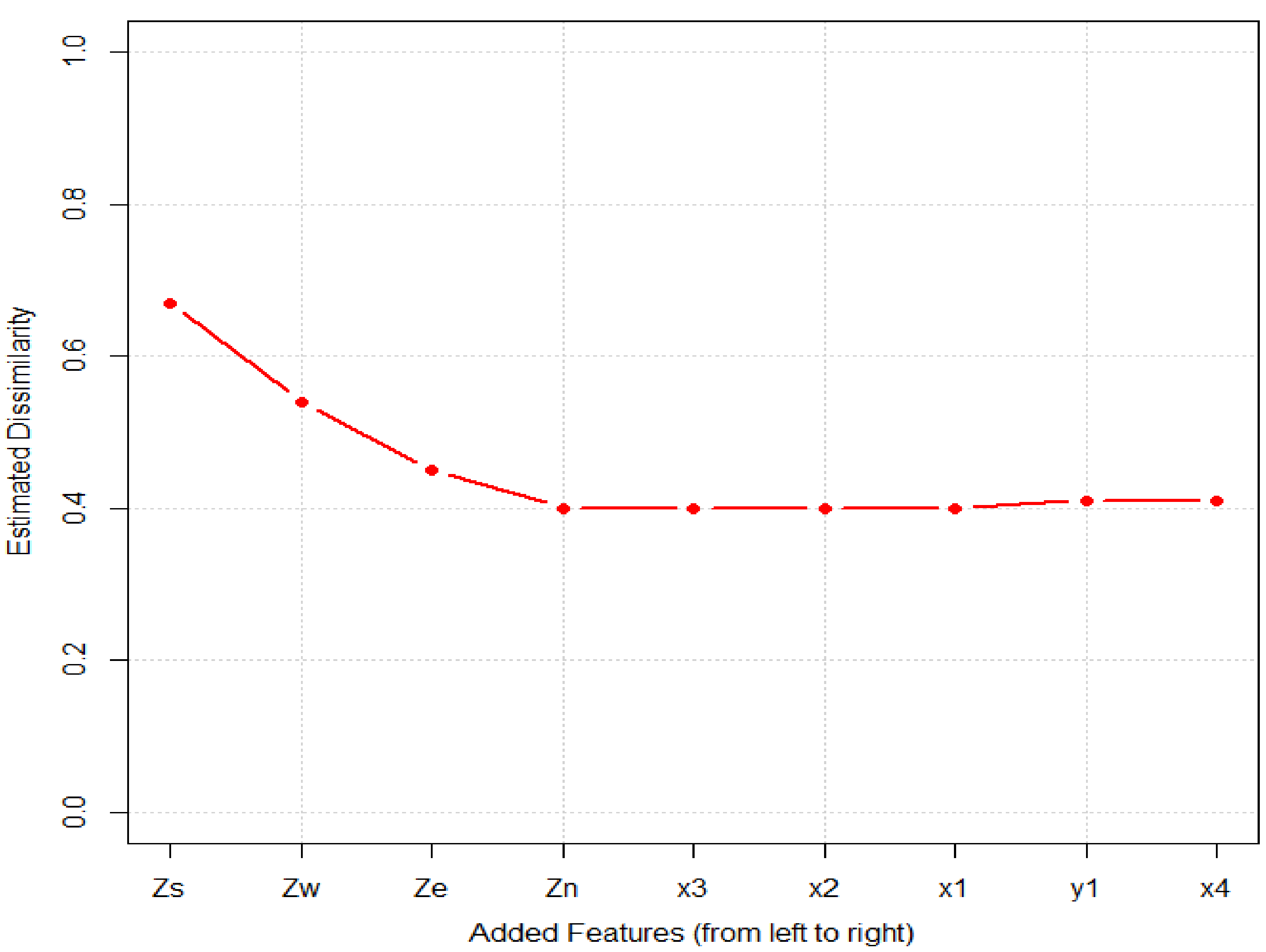
$$Slope \approx \sqrt[2]{\left(\frac{Z_E - Z_W}{2\Delta}\right)^2 + \left(\frac{Z_N - Z_S}{2\Delta}\right)^2}$$

The set of relevant features have a **predictive** power only if used **jointly** and, conversely, the predictive capability of the single relevant feature is marginal. From this viewpoint, topographic slope is a simple and convenient geomorphometric feature for building a synthetic data set for testing purposes.

Sebastiano Trevisani - University IUAV of Venice – mail: strevisani@iuav.it



In addition to the elevations, four **irrelevant** and **non-redundant** features (x1, x2, x3 and x4) have been generated via random shuffling of the elevation and consequently are characterized by the same statistical distribution of relevant features. Finally, a **redundant** (with x4) and irrelevant feature, named y1, has been generated considering the square of x4 plus a Gaussian random noise of zero mean and a standard deviation of 0.1 m. For the dataset, **the ID is 5.62**; excluding the output feature the ID is **5.2**.



Analyzing the index of "dissimilarity" the relevant features are correctly individuated (only elevations have an impact on reducing the dissimilarity index; i.e., the difference between the ID of a set of input features and the ID of the same set plus output feature).

Sebastiano Trevisani - University IUAV of Venice – mail: strevisani@iuav.it

The results are promising; however, more tests should be conducted to fully evaluate potentialities and limitations of the approach in geomorphometry. The capability to handle complex non-linear relationships, the robustness to under-sampling and the straightforwardness of the approach are appealing characteristics. A critical point, to be further investigated, is the sensitivity of the algorithm to the L^{-1} parameter. Another one is how to handle features with statistical distributions characterized by high kurtosis and/or skewness. It is worth noting that this kind of approach is particularly interesting also in the context or remote sensing imagery.

- Trevisani, S. & Rocca, M. 2015, "**MAD: Robust image texture analysis for applications in high resolution geomorphometry**", Computers and Geosciences, vol. 81, pp. 78-92.
- Golay, J. & Kanevski, M. 2015, "**A new estimator of intrinsic dimension based on the multipoint Morisita index**", Pattern Recognition, vol. 48, no. 12, pp. 4070-4081.
- Morisita, M. 1962, "**Iσ-Index, a measure of dispersion of individuals**", Researches on Population Ecology, vol. 4, no. 1, pp. 1-7.
- Golay, J. & Kanevski, M. 2017, "Unsupervised feature selection based on the Morisita estimator of intrinsic dimension", Knowledge-Based Systems, vol. 135, pp. 125-134.
- Golay, J., Leuenberger, M. & Kanevski, M. 2017, "**Feature selection for regression problems based on the Morisita estimator of intrinsic dimension**", Pattern Recognition, vol. 70, pp. 126-138.
- R Development Core Team (2009) **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria
- Golay, J., Kanevski, M., Vega Orozco, C.D. & Leuenberger, M. 2014, "**The multipoint Morisita index for the analysis of spatial patterns**", Physica A: Statistical Mechanics and its Applications, vol. 406, pp. 191-202.
- Kanevski, M. & Pereira, M.G. 2017, "**Local fractality: The case of forest fires in Portugal**", Physica A: Statistical Mechanics and its Applications, vol. 479, pp. 400-410.
- Trevisani S., 2019. "**Unsupervised geomorphometric feature selection based on intrinsic dimension estimation**".Geophysical Research Abstracts. Vol. 21, EGU2019-7318, 2019.