GEOMORPHOMETRY 2020 PERUGIA, ITALY JUNE 22 - 26 2020

Geomorphometric feature selection based on intrinsic dimension estimation

Sebastiano Trevisani
Università IUAV di Venezia
(Venice, Italy)
strevisani@iuav.it

Abstract— **The use of geomorphometric variables or, from a machine learning perspective, geomorphometric features, sometimes coupled with other remote sensing derived variables, is often adopted for the spatial prediction of geoenvironmental properties of interest (e.g. soil and geo-engineering mapping). In other circumstances, geomorphometric features are analyzed for unsupervised approaches in the context of landscape classification and pattern recognition. The detection of the relevant features and the distinction between redundant and irrelevant features is crucial both for improving prediction accuracy as well as for reducing computational cost. Moreover, the detection of relevant features improves the interpretation of studied processes. In this short paper, the potentialities of a new feature selection algorithm are evaluated in a supervised learning problem, tested on ad-hoc designed synthetic dataset. The feature selection algorithm adopted is a Sequential Forward Selection filter, based on a fractal measure of Intrinsic Dimension, relying on a generalization of the Morisita index. The synthetic data set, built from real topography, is characterized by challenging characteristics as for example a strong linear correlation between relevant features. The tests performed on the data set show that the algorithm correctly individuates the relevant features and the irrelevant ones. Moreover, the impact of subsampling on the feature selection algorithm has been tested, showing a stable response up to roughly the 10% of the original data set. The results of this preliminary study suggest that the algorithm is promising in the geomorphometric context and that it is worth to investigate further its applicability in geomorphometry.**

I. INTRODUCTION

The use of geomorphometric variables or, from a machine learning perspective, geomorphometric features (the two terms are used interchangeably in the text), sometimes coupled with other remote sensing derived variables, can be adopted for the spatial prediction of geoenvironmental properties of interest, for example in soil and geo-engineering mapping [1-3]. In other circumstances, geomorphometric features are analyzed via unsupervised approaches for landscape classification and pattern recognition (e.g., [4]). The detection of relevant and non-redundant geomorphometric features in these prediction tasks is crucial both for improving prediction accuracy as well as for reducing computational cost. Moreover, the selection of relevant features improves the interpretation of studied processes, shedding light on main influencing factors and/or processes. Feature selection and reduction are crucial when dealing with geomorphometric analysis. In fact, the quantitative analysis of digital elevation models (e.g., [5-6]) can generate high-dimensional datasets, i.e., characterized by a high number of geomorphometric features. This is partially related to the high number of morphometric variables and local statistical metrics (e.g., [7]) that can be computed. Another reason is related to the spatial-scale dependency of geomorphometric variables and of the related calculation parameters. First, the various geomorphometric variables can be computed from different resolutions and smoothing of the input topography. Second, many geomorphometric variables and local statistical metrics have calculation parameters related to the spatial scale (e.g., the radius of a local search window) or to spatial directionality.

Consequently, given the potentially high number of input geomorphometric features in unsupervised and supervised learning tasks, the discrimination between relevant/irrelevant features (in supervised setting) and of the redundant/non-redundant features is of fundamental importance. The nonlinearity and the complexity of the potential interactions between geomorphometric features make difficult the application of standard parametric data reduction approaches, based for example on principal component analysis or on the linear correlation between variables.

In this context, the recently developed fractal-based estimator of Intrinsic Dimension (ID, [8]), relying on a generalization of Morisita Index [9], is particularly promising. The authors of the new ID estimator developed a set of ID-based algorithms for feature selection both in unsupervised [10] as well as in supervised learning settings [11]. These tools are implemented in R programming environment [12] with a specific package; the algorithms have been designed taking into consideration computational efficiency and ease of use.

Sebastiano Trevisani (2020) Geomorphometric features selection based on intrinsic dimension estimation:

in Massimiliano Alvioli, Ivan Marchesini, Laura Melelli & Peter Guth, eds., Proceedings of the Geomorphometry 2020 Conference, doi:10.30437/GEOMORPHOMETRY2020_11.

This short paper is part of a broader ongoing research exploring the application of these algorithms in geomorphometry both in unsupervised as well as supervised learning settings. The focus of this presentation is on the applicability of the approach in a supervised setting; in particular, the capabilities of the algorithm are tested considering an exemplary and demanding (from the predictive viewpoint) data set, built from real topography.

II. INTRINSIC DIMENSION AND FEATURE SELECTION

ID is strictly related to fractal dimension [8] and is an interesting parameter both for the analysis of spatial point patterns [13-14] as well as in the analysis of multidimensional data [10-11]. In the latter context, it is particularly useful because it allows to detect if the data lie on a lower-dimensional manifold in data space; when data lie on manifolds the ID dimension (not necessarily integer) is lower than the data dimension, i.e. the number of features. The example of the "Swiss roll" [10] distribution is emblematic (figure 1); even if the dataset has 3 variables, the true ID is 2, because the data lie on a 2D surface. The Morisita-based ID estimator is capable of estimating the ID from a multidimensional data set very efficiently and has been proven to be applicable in a wide set of settings, considering noise and under-sampling [8-10-11].

The estimation of ID is at the base of feature selection/reduction algorithms both in unsupervised (e.g., unsupervised clustering) as well as supervised (e.g., regression) settings. The key idea of these algorithms is based on the analysis of the impact of the single features on the ID estimation. For example, in an unsupervised setting, redundant features have a slight impact on ID estimated values. In a supervised setting, the input features can have different characteristics from the perspective of ID and in relation to the predictive capability of the output variable. Concerning ID, some of the input features can be redundant (e.g., strongly correlated) and hence leading to a lower ID respect the number of input features. Some other input features can be irrelevant, i.e., bring no information on output feature, and then contribute to the increase of the ID of the dataset. With real datasets based on geomorphometric features, redundant input features can be irrelevant as well as relevant The supervised feature selection algorithm of Golay et al. [11] is a Sequential Forward Selection filter, using the ID measure for discriminating relevant versus irrelevant features. The algorithm evaluates iteratively, for different subsets of input features, the index of dissimilarity (Diss), according to the equation:

$$Diss(F,Y) = ID(F,Y) - ID(F) \quad (1)$$

where F is a subset of the input features and Y is the output feature. When F is composed exclusively by all relevant features, the dissimilarity index should be theoretically zero. Differently, the irrelevant features have no impact in reducing the dissimilarity.

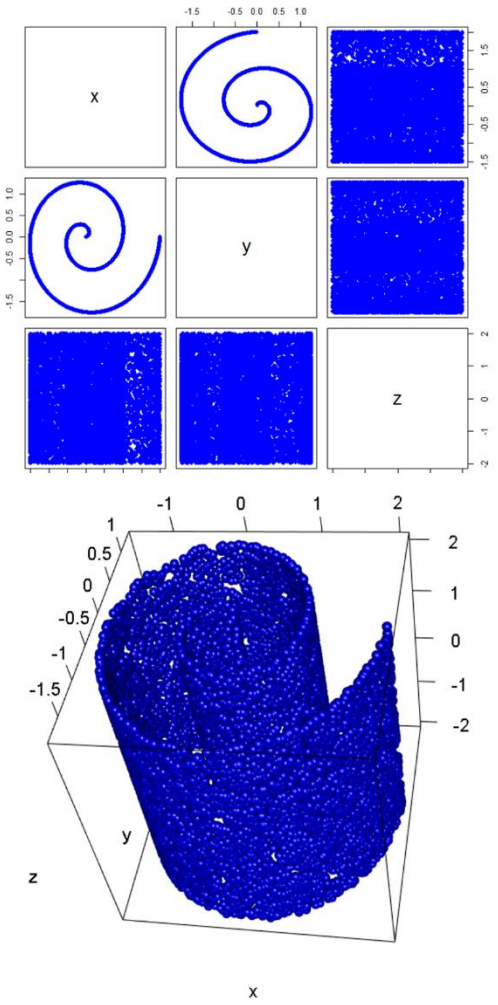

Figure 1. 2D and 3D scatterplots of the "Swiss roll" synthetic dataset. The ID is 2, because data lie on a 2D surface.

III. THE EXPERIMENTAL DESIGN

For testing the supervised feature selection algorithm, a synthetic data set has been built, based on real topography, from which 9 input features and one output feature have been derived. The DTM considered, derived from airborne Lidar technology, is representative of an alpine area with complex morphology (fig. 2, [15]) and has a grid of 350x350 pixels, with a resolution of 20 m. A synthetic dataset has been considered, given the necessity to know exactly the true relationship between input and output features, and test the impact of subsampling on the algorithm performance. The use of real topography and not of a pure

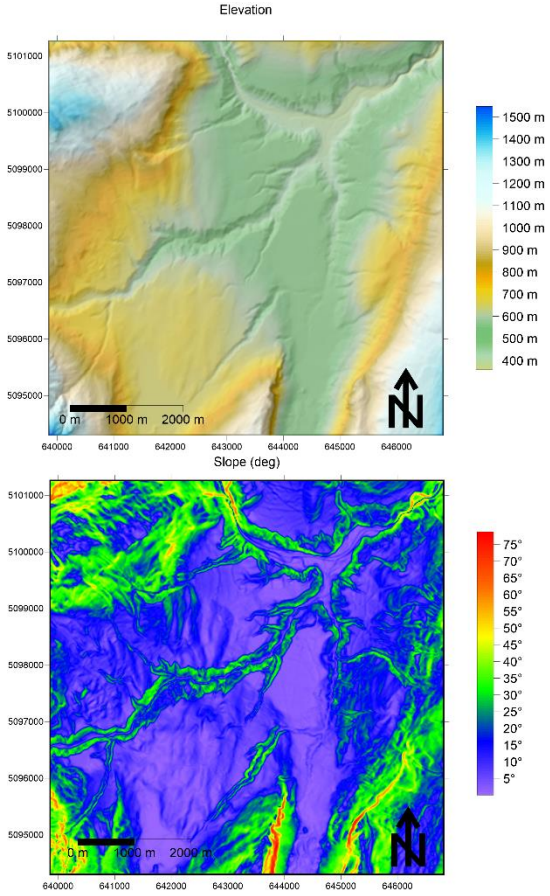

Figure 2. The DTM (Trentino, NE Italy) and the calculated slope. Slope (in the dataset expressed as percent rise) represents the output feature.

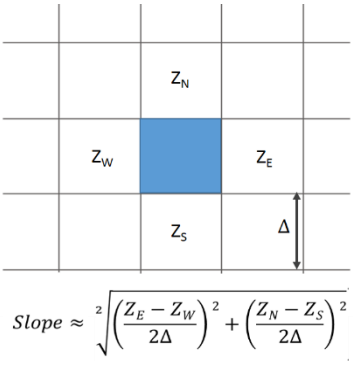

Figure 3. Simplified equation for computing the slope.

synthetic dataset generated from theoretical random distributions (e.g., [11]), is dictated by the need of testing spatial-statistical distributions representing real morphology, even if limited to an alpine setting. Another target of the built dataset is to analyze complex and non-linear relationships between input and output features, including potential redundancy between relevant features. This aspect is particularly relevant in geomorphometry, because of linear correlations between potentially relevant features can be present (e.g., linear correlation between roughness and slope, [4]). Moreover, it may happen that the set of relevant features have a predictive power only if used jointly and, conversely, the predictive capability of a single relevant feature can be marginal. From this viewpoint, topographic slope is a simple and convenient geomorphometric feature for building a synthetic data set for testing purposes. The slope (i.e., output feature) has been computed according to the simple equation reported in figure 3. Consequently, the relevant features are the elevations of the four nodes, here named as features Zn, Ze, Zs and Zw. The formulation of slope permits to model a complex non-linear relationship, with high redundancy between relevant features, and in which the features are relevant only if used in combination. Then, four irrelevant and non-redundant features (named 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, considering all features the ID is 5.62; excluding the output feature the ID is 5.2. It is worth noting, that a feature selection approach based on the analysis of the linear correlation between input features would induce to do not consider some of the relevant features.

IV. Preliminary Results and Future Developments

The tests performed on the synthetic data set are highly positive: the algorithm correctly detects relevant input features and the irrelevant ones. The ID of the set of relevant features with and without the output feature is respectively 1.85 and 1.44; the ID of the set of irrelevant features with and without the output feature is respectively 4.71 and 3.74. The computational parameter to be set in the algorithm is the range of variation of the parameter L^{-1}, controlling the size of the moving windows inherent to Morisita index calculation [8-9]. L is the length of the side of the search windows in the data space, being all features normalized in the 0-1 interval. In this study, after different trials and following the approach suggested in [8], the integer values of L^{-1} were set to {10,11,…, 50}. The output is easily interpretable from the diagnostic curve (figure 4) reporting the impact of the single input features on the variation of the dissimilarity index. Only the features reducing the Dissimilarity index are relevant for the supervised learning.

From the perspective of computational time the algorithm (the non-parallelized version has been tested) is quite fast; the application on the whole dataset required 5.3 minutes with a ten years old Processor Intel® Core™2 Quad Q8300 2.5 GHz and 12 Gb of ram.

A first test on the sensitivity of the algorithm to sampling density has been conducted. The impact of under-sampling has been explored by means of random sub-sampling (100 times) the original distribution with different levels of sub-sampling. Up to an under-sampling of 90 % (only 10% of the values retained), the algorithm is stable in terms of features selection, even if the diagnostic curve is characterized by a high variance.

The results are promising even if more tests should be conducted to fully evaluate potentialities and limitations of the approach in geomorphometry [16]. 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 in presence of 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.

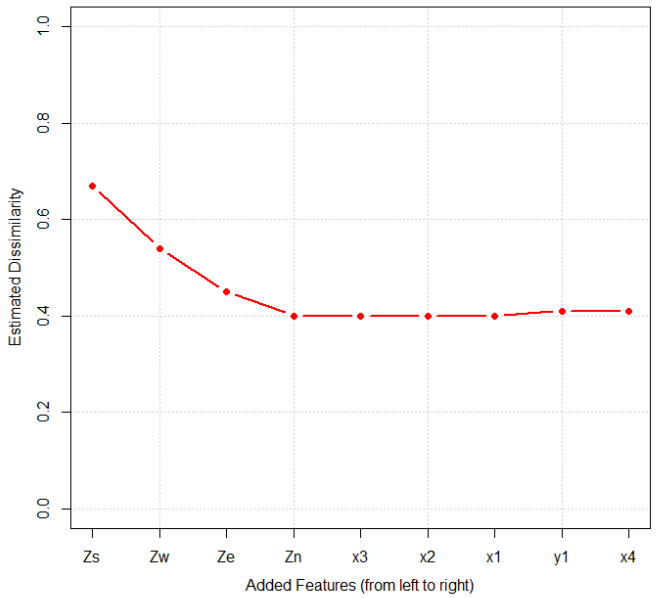

Figure 4. Results of the ID-based supervised feature selection approach applied to the whole data set. The features located on the right of Zn do not reduce the dissimilarity and are considered irrelevant.

REFERENCES

[1] Florinsky, I.V. 2016, "Digital Terrain Analysis in Soil Science and Geology: Second Edition" in Digital Terrain Analysis in Soil Science and Geology: Second Edition, pp. 1-486.

[2] Florinsky, I.V., Eilers, R.G., Manning, G.R. & Fuller, L.G. 2002, "Prediction of soil properties by digital terrain modelling", Environmental Modelling and Software, vol. 17, no. 3, pp. 295-311.

[3] Trevisani, S., Cavalli, M., Golay, J. & Pereira, P. 2019, "Editorial to the topical collection "Learning from spatial data: unveiling the geo-environment through quantitative approaches"", Environmental Earth Sciences, vol. 78, no. 5.

[4] Trevisani, S., Cavalli, M. & Marchi, L. 2012, "Surface texture analysis of a high-resolution DTM: Interpreting an alpine basin", Geomorphology, vol. 161-162, pp. 26-39.

[5] Florinsky, I.V. 2017, "An illustrated introduction to general geomorphometry", Progress in Physical Geography, vol. 41, no. 6, pp. 723-752.

[6] Pike, R.J., Evans, I.S. & Hengl, T. 2009, Geomorphometry: A brief guide.

[7] Trevisani, S. & Rocca, M. 2015, "MAD: Robust image texture analysis for applications in high resolution geomorphometry", Computers and Geosciences, vol. 81, pp. 78-92.

[8] 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.

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

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

[11] 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.

[12] R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

[13] 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.

[14] 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.

[15] Florinsky, I.V., Skrypitsyna, T.N., Trevisani, S. & Romaikin, S.V. 2019, "Statistical and visual quality assessment of nearly-global and continental digital elevation models of Trentino, Italy", Remote Sensing Letters, vol. 10, no. 8, pp. 726-735.

[16] Trevisani S., 2019. "Unsupervised geomorphometric feature selection based on intrinsic dimension estimation".Geophysical Research Abstracts. Vol. 21, EGU2019-7318, 2019.