Finding the Best Combinations of Terrain Attributes and GIS software for Meaningful Terrain Analysis

Vincent Lecours, Alvin Simms, Rodolphe Devillers, Evan Edinger
Department of Geography
Memorial University of Newfoundland
St. John's, Canada
vlecours@mun.ca

Vanessa Lucieer
Institute for Marine and Antarctic Studies
University of Tasmania
Hobart, Australia

Abstract—**Tools that derive terrain attributes from digital elevation models are common in geospatial software. Their accessibility permits applying geomorphometric techniques to a wide range of applications. These tools however, can be considered "black boxes" where the analysis and comparison of the internal workings of the technique are vague and cannot be assessed. Selecting the most effective set of tools for a given task can thus be challenging. This work presents a method for selecting an optimal set of terrain attributes that can help non-expert GIS users make the best use of geomorphometry. The selection of terrain attributes aims to remove redundancy between attributes and maximize the amount of information given on a surface. We derived 230 terrain attributes from an artificial surface using 11 software. This approach is twofold: a pre-selection based on the ranking of attributes was first established using stepwise multicollinearity measures, followed by a final selection of attributes from a principal components analysis (PCA). The results show that using 13 independent terrain attributes can explain up to 83% of the variance for that particular surface: the combination of common attributes that are available in most GIS (i.e. aspect, basic curvatures, slope and a measure of rugosity) can explain 67% of the surface variance. The method proved efficient to reduce a high-dimensional list of terrain attributes to identify combinations of 13 attributes or less that can be used by non-expert GIS users.**

I. INTRODUCTION

Tools allowing geographic information systems (GIS) users to derive terrain attributes from digital elevation models (DEM) are increasingly available in GIS software. These attributes can be used for a wide range of purposes, such as explanatory variables or indicators in biological and ecological studies [1]. The algorithms implemented by the different software are not always specified and can often leave users with little choice of the appropriate algorithm and specific parameters to use. Since different algorithms can produce significantly different results [2], and that few studies outside of the field of geomorphometry report the methods used for the computation of terrain attributes [3], comparisons between studies can be misleading. Users are left with a large number of options and without guidance are often tempted to select a random or sub-optimal set of terrain attributes for their study using the GIS they are familiar with.

Using a random selection of terrain attributes or all the available attributes from a specific software might result in outcomes that are not representative of the observed phenomenon, failing to capture the key properties of a terrain. For instance, being all derivatives from a same surface (the DEM), terrain attributes are likely to show a certain level of covariation [4]. Covariation between variables is known to influence performance of regression analyses [5] and other statistical models [4]. Since multicollinearity makes it difficult to distinguish the influence of individual drivers on a response variable [6], it is important to carefully select the terrain attributes in order to reduce that covariation. Analysts rarely assess multicollinearity between independent variables used in a regression analysis [7].

This study aims to test a method that determines sets of terrain attributes that can (1) minimize multicollinearity between the selected attributes and (2) maximize the variance of the terrain explained by the selected attributes. Such sets of attributes could be used by GIS users to help create more robust models.

II. METHODS

A. *Terrain Attributes*

A 1x1m resolution artificial surface covering an extent of 106x106m was created using the spectral synthesis method in Landserf 2.3 (Figure 1). 230 terrain attributes were derived

In: Geomorphometry for Geosciences, Jasiewicz J., Zwoliński Zb., Mitasova H., Hengl T. (eds), 2015. Adam Mickiewicz University in Poznań - Institute of Geoecology and Geoinformation, International Society for Geomorphometry, Poznań

from this surface using different software packages (Table I). Since some of the software only allow using a 3x3 window of analysis, all analyses were performed using this size of window to allow the inclusion of as many attributes as possbile. The terrain attributes tested are from general geomorphometry, i.e. are computed continuously across a surface, and include both local geometric and statistical attributes. Selected software include both commercial and open source software. To eliminate edge effects, the outer 6m were clipped, resulting in 100x100m surfaces. Each terrain attribute was tested against the others to identify those giving strictly identical results, and thus likely to be using a same algorithm. Since the algorithms should only be accounted for once, only one attribute was kept for each set of duplicates.

A. *Pre-Selection –Dimensionality Reduction*

Several methods exist to detect multicollinearity among variables, and three were selected to examine how all the attributes vary with the others: the Variable Inflation Factor (VIF) [8], the Mutual Information (MI) [9] and the minimum redundancy (W_c) [10]. A known limit of multicollinearity measurements is the lack of meaningful threshold to distinguish values characterized as collinear from values representing the absence of multicollinearity [11]. Some methods, such as the VIF, use arbitrary values as threshold. This makes it difficult to objectively select subsets of terrain attributes based solely on these three measures of multicollinearity. However, variable ranking is often used in machine-learning as a pre-processing step [10] which, even when non optimal, is computationally efficient and statistically robust in preventing over fitting [12]. It was thus possible to rank the terrain attributes based on their level of co-association with the others, without defining any threshold.

Since the levels of co-association vary as soon as one of the variables is removed from the datasets, stepwise measures of VIF, MI and W_c were computed using the statistical software R 3.1.1. The stepwise algorithms (1) calculate the values of the measures for each terrain attribute, (2) rank the terrain attributes based on these values, (3) remove the most collinear or least informative attribute, (4) save it in a list, and (5) repeat the process until all the attributes are ranked in this list. The process is the same for the three measures of multicollinearity. An average of the three rankings was then performed for each terrain attribute, and the 40% top-ranked attributes were kept for further analysis.

B. *Selection and Grouping – Principal Component Analysis (PCA)*

The remaining terrain attributes measurements were imported in the IBM SPSS Statistics software v.22. Principal

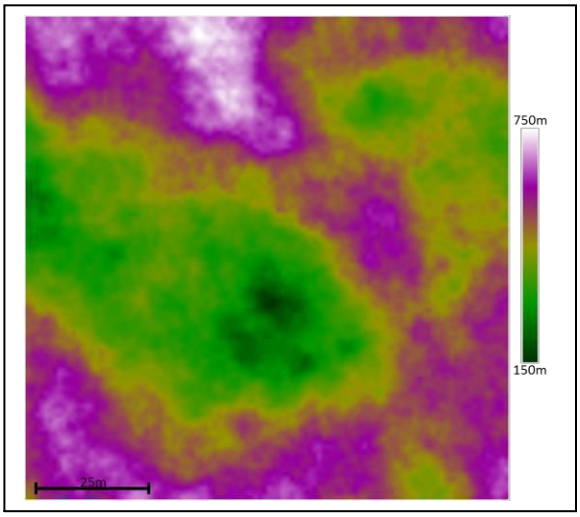

Figure 1. Artificial surface used to derive the 230 terrain attributes

Component Analysis (PCA) is one of the most common techniques used to reduce multicollinearity in a dataset [4]: a stepwise PCA using a Varimax orthogonal rotation was performed. The PCA grouped terrain attributes in independent groups (called components) of highly correlated attributes. An orthogonal rotation allows components to be uncorrelated, hence removing multicollinearity between the groups. Varimax is the most commonly used method for orthogonal rotation and maximizes the variance of component loadings [13].

For each iteration of the PCA, the attributes that loaded equally on two or more components were identified and removed: when a variable was found in more than one group, it was considered redundant, not contributing to the model [13]. Iterations ended once the computation of the PCA ceases to isolate any further redundant attributes. The optimal number of components to be retained was then found in SPSS using a parallel analysis [14] on the remaining attributes. PCA was then performed using the number of components obtained from the tests, and the attributes that did not load on any the components were removed before running a final PCA. Since the first step of the analysis consisted in subjectively removing one attribute over an identical one and that we did not want to favour a software over another, the identical attributes were added back to the final solution under the assumption that if one of the attributes reached the final solution, an identical one would have too.

III. RESULTS AND DISCUSSION

Removal of identical terrain attributes reduced the list of spatial geomorphology derivatives from 230 attributes to 182.

The pre-selection reduced that list further to 73. The iterative PCA removed the redundant attributes in 7 iterations. The 73 attributes of the initial PCA loaded on 17 components, and the seventh iteration left 59 attributes loading on 14 components. The optimal number of components given by three of the four tests was 13. When the PCA was re-run with 13 components, two of the 59 attributes did not load on any of the components and were thus removed. The final solution had therefore 57 terrain attributes loading on 13 components. By adding back the identical attributes, the solution used for interpretation had 67 attributes (Table I).

The percentage of variance explained by each component and the cumulative percentage of the variance are indicated in Table II. The final solution shows a clear association between the type of terrain attributes and the components. The interpretation of the components presented in Table II is based on the terrain attributes that contributed the most to each component. The first four components and the seventh had only one type of terrain attributes in them. For instance, the first component includes only measures of easterness from different software and computed using different algorithms. Other components (5-6, 8-9, 11-13) had a combination of two types of terrain attributes, with one primary and one secondary type. For example, the fifth component had 5 measures of slope and one measure of local fractal dimension. Only the tenth component had three different types of terrain attributes. They were however all from the same software, which might indicate that the algorithms used by Whitebox GAT to measure curvatures are significantly different from those of other software. A general interpretation of the results indicates that a combination of attributes of first (i.e. aspect and slope) and second derivatives (i.e. plan and profile curvatures), which are

TABLE I. SOFTWARE USED, NUMBER OF TERRAIN ATTRIBUTES GENERATED USING EACH OF THEM, AND NUMBER AND PROPORTION OF ATTRIBUTES IN THE FINAL SOLUTION

Software and Versions	Number of Attributes Computed	Number in Final Solution	Proportion Retained
ArcGIS 10.2.2 with Python 2.7.8	22	3	14%
ArcGIS 10.2.2 with DEM Surface Tools (v.2.1.399)	17	4	24%
ArcGIS 10.2.2 with Benthic Terrain Modeler 3.0 rc3	12	0	0%
Diva-GIS 7.5.0	7	1	14%
Idrisi Selva 17.0	7	2	29%
Landserf 2.3	12	3	25%
Quantum GIS 2.4.0 Chugiak	13	1	8%
SAGA GIS 2.0.8	96	26	27%
TNTmips Free 2014 (MicroImages)	25	21	84%
uDig 1.4.0b	9	2	22%
Whitebox GAT 3.2.1 Iguazu	10	4	40%
TOTAL:	**230**	**67**	**29%**

TABLE II. TOTAL OF VARIANCE EXPLAINED BY EACH COMPONENT AND INTERPRETATION

Component	Percentage of Variance	Cumulative Variance	Number of Attributes	Interpretation (Primary / Secondary)
1	15.07%	15.07%	12	Easterness
2	12.92%	28.00%	11	Northerness
3	8.60%	36.60%	5	Plan Curvature
4	8.52%	45.12%	5	Profile Curvature
5	5.88%	51.00%	6	Slope / Local Fractal Dimension
6	5.87%	56.86%	4	Tangential / Plan Curvatures
7	5.25%	62.12%	4	Vector Ruggedness Measure
8	4.57%	66.69%	3	Longitudinal / Profile Curvatures
9	4.33%	71.02%	7	Local Maximum / Slope
10	4.17%	75.19%	3	Curvatures (from Whitebox GAT)
11	3.43%	78.61%	3	Total Curvature / Slope Variability
12	2.46%	81.07%	2	Plan / Profile Curvatures (from Saga GIS)
13	2.42%	83.49%	2	Representativeness / Mean of Residuals

the most common terrain attributes implemented in GIS software, accounts for 51% of the variance of the surface alone (Table II). The addition of other measures of curvature, rugosity and statistical attributes increases the variance explained.

Since all the terrain attributes within each component load almost equally high, they are considered equivalent in importance: PCA regroups highly correlated attributes that also interact with attributes from other components in the same way, therefore making each attribute very similar to all the others within the same component. An optimal combination of terrain attributes would thus consist of one terrain attribute from each component. This indicates that an optimal combination of terrain attributes would have between 5 and 13 attributes.

The algorithm used to derive some of the terrain attributes does not seem to matter much for that particular surface: some algorithms loaded higher than others on each component, but not in a significant manner. TNTmips was the software that had the highest percentage of its surface attributes kept in the analysis (Table I). In addition, at least one attribute from TNTmips is included in each of the five first components, indicating that 51% of the variance can be explained using this software alone (Table II).

The method proved to be efficient in reducing the number of terrain attributes to measure and to provide combinations of terrain attributes to capture the most variance on the surface. More tests will however be necessary to refine the method and test it on natural surfaces, at different scales, and on surfaces of different complexity levels.

IV. Conclusion

This contribution presented a new method to reduce a high-dimensional list of terrain attributes in order to select optimal combinations of terrain attributes to be used by non-expert GIS users. The method reduces multicollinearity and maximizes the variance of the surface that is explained. The proposed method proved to be efficient for the surface on which it was tested and reduced a list of 230 terrain attributes to a list of 67. From these 67 attributes, only 5 were needed to explain 51% of the variance and 13 to explain 83% of it. PCA allowed a meaningful statistical grouping of terrain attributes presenting similar characteristics. Using the results from the final PCA, one can use only one attribute per component and be sure that multicollinearity is removed and that a significant amount of variance is explained. Since principal components will reflect the surface used, future analyses will be conducted using a range of natural and artificial surfaces.

Acknowledgment

Thanks to Dr. Arnaud Vandecasteele for helping with programing, and to the Canadian Natural Sciences and Engineering Research Council (NSERC) for their financial support.

References

[1] Rengstorf, A.M., Grehan, A., Yesson, C., and C. Brown, 2012. Towards high-resolution habitat suitability modeling of vulnerable marine ecosystems in the deep-sea: resolving terrain attribute dependencies. *Marine Geodesy* 35, 343-361.

[2] Dunn, M., and R. Hickey, 1998. The effect of slope algorithms on slope estimates within a GIS. *Cartography* 27, 9-15.

[3] Dolan, M. F. J., and V. L. Lucieer, 2014. Variation and uncertainty in bathymetric slope calculations using geographic information systems. *Marine Geodesy*, 37 (2), 187-219.

[4] Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Gracía Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and S. Lautenbach, 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27-46.

[5] Hijmans, R. J., 2012. Cross-validation of species distribution models: removing spatial bias and calibration with a null model. *Ecology* 93, 679-688.

[6] MacNally, R., 2000. Regression and model-building in conservation biology, biogeography and ecology: the distinction between – and reconciliation of – "predictive" and "explanatory" models. *Biodiversity and Conservation* 9, 655-671.

[7] Graham, M. H., 2003. Confronting multicollinearity in ecological multiple regression. *Ecology* 84, 2809-2815.

[8] Belsley, D. A., Kuh, E., and R. E. Welsch, 2004. Regression diagnostics: Identifying influential data and sources of multicollinearity. *Wiley*, 292 p.

[9] Song, L., Langfelder, P., and S. Horvath, 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* 13, 328

[10] Ding, C., and H. Peng, 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology* 3, 185-205.

[11] Kohavi, R., and G. John, 1997. Wrappers for feature selection. *Artificial Intelligence* 97, 273-324.

[12] Guyon, I., and A. Elisseeff, 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 1157-1182.

[13] Tabachnick, B. G., and L. S. Fidell, 2014. Using multivariate statistics, Sixth edition. *Pearson Education Limited*, 1056 p.

[14] O'Connor, B.P., 2000. SPPS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers* 32, 396-402.