Machine learning for cross-scale geomorphometric classification of landforms: a day at the beach

Ashton Shortridge*, Clayton Queen, & Alan Arbogast
Department of Geography, Environment, & Spatial Sciences
Michigan State University
East Lansing, USA
ashton@msu.edu

Abstract—**This paper investigates the use of random forests and spatial random forests (RFsp) for the classification of coastal dune areas along 41km of Lake Michigan's shoreline using a lidar-derived DEM. Terrain variables across a range of spatial neighborhood scales are utilized, and for two different cell resolutions. Distance is explicitly incorporated into the RFsp models through the calculation of buffer distances around small numbers (6-13) of gridded points in the study area. While classification accuracy is high generally, RFsp produced much more accurate results. At the fine scale, topographic variables and their neighborhood ranges were not predictive of dune areas, perhaps because large (> 0.1 hectare) neighborhoods were not tested at that scale. At the coarse scale these variables were much more important. The use of small numbers of gridded (non-sample) points to improve spatial prediction warrants further investigation.**

I. Introduction

The largest body of freshwater dunes in the world occurs along the eastern shore of Lake Michigan in Michigan. These dunes consist largely of parabolic landforms that, in many cases, are over 30 m high. They line the shore for long (> ~ 1.5 km) stretches along the southwestern shore of Lower Michigan and occur in more isolated embayments and bluff-top locations in the northwestern part of the peninsula [1]. These ecologically important dunes are also heavily utilized for industrial purposes, recreation, and home construction, and are thus one of the most contested landscapes in the Great Lakes region. Due to these pressures, the State of Michigan has established so-called critical dune areas along the coast to facilitate land management of this region. Critical dune areas have been revised several times since the 1980's when they were first enumerated. The latest assessment was conducted in 2017-2018 by two of the authors of this paper for the State of Michigan. This study used submeter lidar-based DEMs and aerial imagery to manually produce a detailed set of polygons along the entire coastline of lower Michigan.

Coastal dune areas are ontologically complex: while these areas are geomorphologically determined, the dune fields they cover may be comprised of an assemblage of distinctive landscape features with varying degrees of crisp spatial boundary definition. Vegetation cover can range from dense forest to bare sand. Moreover, they are intended for management purposes, which means that these areas must be compact and without holes. This complex of physical and management factors makes their delineation challenging.

Machine learning methods have been applied to physical classification applications like this, most notably in landslide hazard mapping and soil classification (e.g., [2], [3], [4]), but such work also comes with significant caveats for geospatial applications [5]. The present paper concentrates on the use of random forests [6], [7] for machine learning classification. Random forests (RF) are an extension of decision tree classification and regression methods which partition data into hierarchical groupings. As implied by its name, an RF is an ensemble of regression trees, each built using subsets of the training data and of the predictor variables at each split. The RF estimate for any set of predictor variables is obtained by averaging across all trees [8].

Given the demonstrated potential of random forests in complex classification environments, it is interesting and important to evaluate its use in a range of geomorphometric applications. The present research investigates the following questions: 1) how effective are RF for classifying coastal dune area complexes using terrain covariates?; 2) How does scale and spatial heterogeneity affect the relative importance of geographical variables used for classification, as well as

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27112v1 | CC BY 4.0 Open Access | rec: 13 Aug 2018, publ: 13 Aug 2018

classification accuracy?; 3) How does the use of distance mesh affect prediction accuracy?

II. Data and methods

The study site spans the western boundary of Allegan County in western lower Michigan (Fig.1). The shoreline extends for 41km roughly south to north, and all locations within 4km of the shore were included, a distance designed to include the extents of the most eastward-extending dune areas. Manually derived polygons defining the reference dune areas were obtained for this region. Dune areas were geographically heterogeneous, ranging from large parabolic dune fields spanning many kilometers to narrow barrier dunes along the beach. Large stretches of coastline had no dunes. A lidar-derived DEM of the county with three-foot (0.914 m.) spacing was obtained from the State of Michigan. The Michigan State Plane South reference system was used, as this was the original format of the data. All analysis was conducted in R, and relied on procedures in libraries rgdal, sp, raster, GSIF, and ranger.

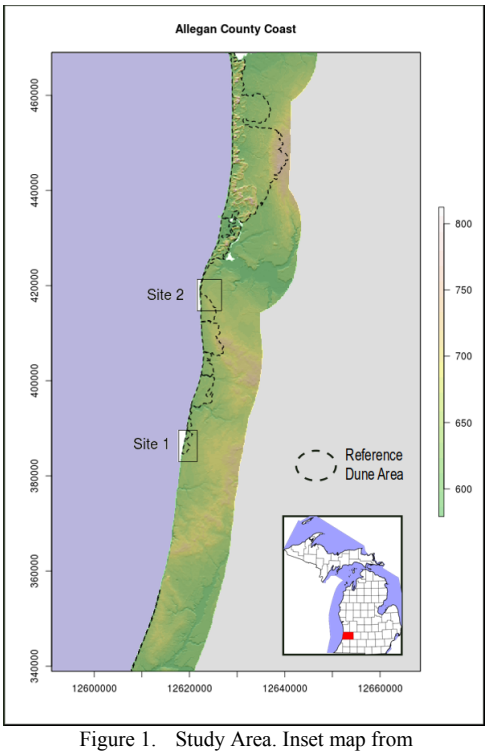

Figure 1. Study Area. Inset map from https://en.wikipedia.org/wiki/Allegan_County,_Michigan

Two distinct geographic scales were used: fine (3 ft cell resolution, neighborhoods of up to 87 ft) and coarse (99 ft cell resolution, neighborhoods of up to 2,871 ft). Due to computational constraints two subareas, Site 1 and Site 2, were used at the fine scale (Figure 1). Distance from lakeshore was the only non-terrain-related variable. Due to the computational expense of calculation, this variable was processed on a raster with 99 foot foot horizontal resolution, then interpolated in two stages to the 3-foot resolution of the fine-scale project. Derivative products were calculated from the DEMs: slope, aspect, topographic position index, and roughness. Focal range of each derivative was calculated for 3x3, 5x5, 9x9, 15x15, and 29x29 cells for cross-scale information on topographic variation. The coarse scale study was conducted region-wide using the DEM resampled to 99ft, and with the same variables calculated at this resolution and further processed with the same focal ranges. As a final variable for experiments at both scales, extending the RFsp concept introduced by [9], distance buffers were calculated for six points spaced in regular grids across each subarea and for 13 points over the entire coastal area.

For training and validation, stratified random point samples were taken from each subarea site and the entire region (n=1,000, evenly split between dunes/non-dunes). At least 20 samples were taken from each of the eight separate dune areas. Each sample was randomly divided into a training set (800 points) and a validation data set (200 points). RF models were run on the training set points using all variables to predict ‘dunes’ or ‘not dunes’. Setting mtry close to the number of variables used in the model improved performance, while reasonable values of other tuning parameters had negligible impacts on accuracy. The effect of sample size on accuracy was evaluated using multiple sampling with varying size on the training set.

III. Results

Excluding distance buffer variables, variable importance for Site 1 and Site 2 was quite similar. Distance from lakeshore and elevation were by far the most important variables. However, model predictive capability was different. Site 1 validation error was 89.5%, while that for Site 2 was 96%. Errors of commission for the ‘dunes’ class were the main source of confusion at Site 1. Accuracy (over 20 replications of RF for each size) was dependent on sample size: accuracy improves as sample size increases (Fig. 2). Finally, these models were used to classify all pixels within each site. These site-wide classifications had somewhat lower overall accuracy, at 88% and 94%, respectively. Errors of commission were a significant issue for Site 1 (Fig. 3, upper row).

When distances from the six gridded points are added to the RF model, validation prediction accuracy increased to 98% and 99% for each site. Distance from lakeshore and elevation remained the most important variables for Site 1, followed by five of the six buffer distances. For Site 2, the two most important variables were buffer distances, followed by lakeshore

distance and elevation, followed by two of the other buffer distances. In both cases terrain derivative variables were not particularly important. Site-wide classifications enjoyed overall accuracies of 97% and 98% (Fig. 3, lower row).

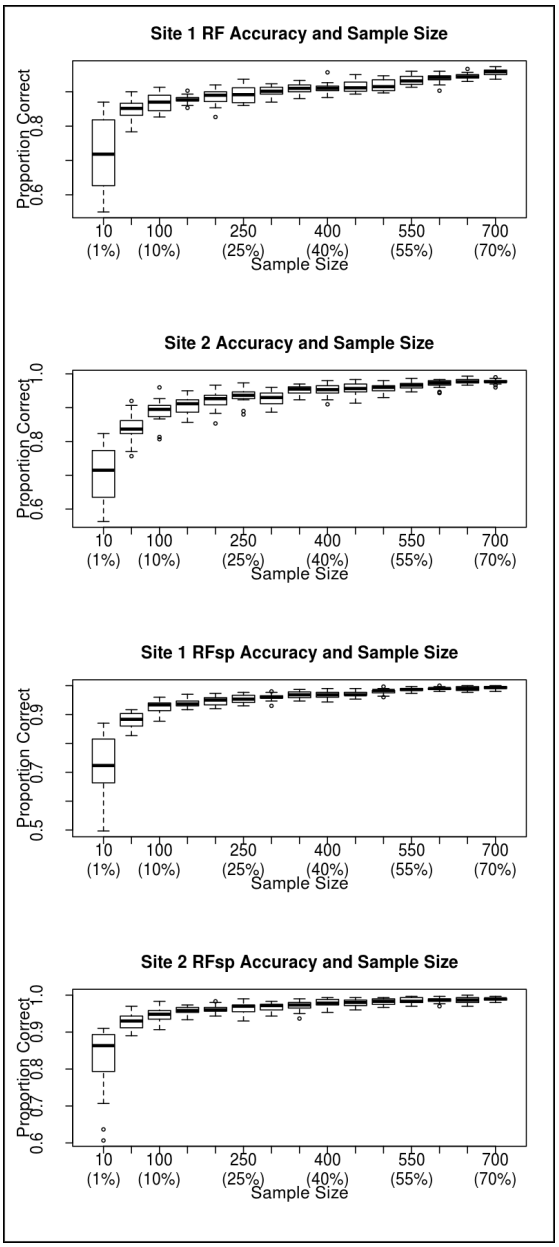

Figure 2. Random Forest accuracy and sample size. 20 samples at each size were drawn to build boxplots.

The coarser-scale, county-wide model had roughly comparable overall accuracy of 90%. The most important variables were slope range over a 29x29 window (a square area 2,971 ft on a side, or about 80 hectares), distance to the lakeshore, the pixel elevation, and TPI range over the same large window. Accuracy increased with sample size, reaching 80% at a size of ~120 and slowly increasing to 90% at a sample size of 400. Study-wide classification accuracy dropped to 86%, with producers accuracy of just 59% for the 'dune' class (Fig. 4).

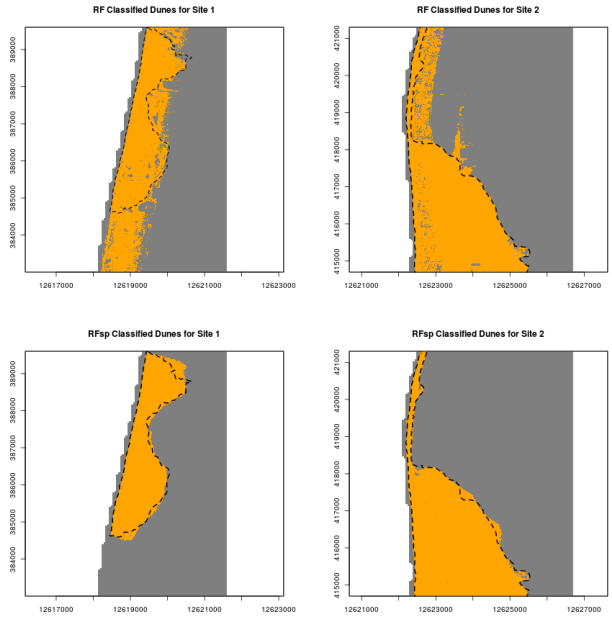

Figure 3. RF (first row) and RFsp (second row) classifications for Site 1 and Site 2. Reference dune area polygons marked with dotted lines.

Incorporating the 13 buffer distance variables improved the model to an overall accuracy of 96% using the validation dataset. Distance to the lakeshore and slope range over the 29x29 window remained important, but the most important variable was one of the buffer distances. Several buffer distances, as well as TPI and roughness range over the 29x29 window were also important. Study-wide classification accuracy was 95%, with errors of commission (producers accuracy) of the 'dune' class largely contributing to the error. Fig. 4 shows the result – errors are particularly noticeable in the northern portion of the county, with good fidelity to the reference data in other areas.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27112v1 | CC BY 4.0 Open Access | rec: 13 Aug 2018, publ: 13 Aug 2018

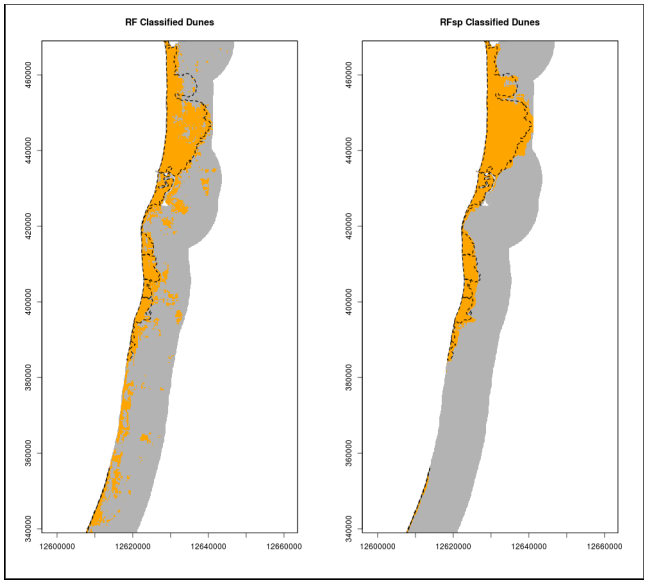

Figure 4. RF (first col) and RFsp (second col) classifications for the entire coast. Reference dune area polygons marked with dotted lines.

IV. Discussion

Over fairly homogenous areas with well-defined coastal dunes, RF performed well at both fine (3 foot) and coarse (99 foot) scales. However, at the fine-scale the topographic variables used were unable to effectively partition the landscape, and the models relied on raw height and distance from lakeshore – a reliance that led to the egregious errors of commission visible in Fig. 3. We suspect that the window sizes for the neighborhood operations may have simply been too small to distinguish dune landscapes effectively using these terrain variables. At the coarse scale, terrain variables were much more important, especially at the larger neighborhood sizes. Dunes were characterized at rates of 80-90% using independent validation samples in all cases. However, errors of commission for dune areas were a substantial problem.

The use of spatial random forests (RFsp) with buffer distances around gridded points (not sampled locations) was highly effective in improving classification accuracy, and reducing errors of commission. We were surprised by the effectiveness of distances from just a few gridded points to improve the classification. We suspect that this improvement is due to two factors: 1) the geographic context that even this small number of points provide to the model; 2) they enable better handling of spatial heterogeneity. Increasing the number of buffer points to 52 did not greatly improve model performance.

Samples were stratified by class but not by space, leading to concerns that spatial autocorrelation may have affected the classifiers. However, spatial autocorrelation was also a positive force in improving accuracy using RFsp with the buffer distances. Further analysis on RF standard errors is warranted. Finally, we note that the reference dune area polygons, while professionally interpreted, are subject to uncertainty themselves, and some classification error in the RF models may in fact identify places with particularly fuzzy dune area properties for computer and human interpreters alike.

Acknowledgment

The forbearance of the organizing committee is gratefully recognized.

References

[1] Arbogast, A. F., 2009. “Sand dunes”. In Michigan Geography and Geology, Pearson, 274-287.

[2] Li, J., A. Heap, A. Potter, and J. Daniell, 2011. “Application of machine learning methods to spatial interpolation of environmental variables”. Environmental Modelling & Software, 26(12), 1647-1659.

[3] Goetz, J.N., A. Brenning, H. Petschko, and L. Leopold, 2015. “Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling”. Computers & Geosciences, 81, 1-11.

[4] Hengl, T., J. M. de Jesus, G. B. Heuvelink, et al., 2017. “SoilGrids250m: Global gridded soil information based on machine learning”. PLoS one, 12(2), p.e0169748.

[5] Millard, K. and M. Richardson, 2015. “On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping”. Remote Sensing, 7(7), 8489-8515.

[6] Breiman, L., 2001. “Random forests”. Machine Learning, 45, 5-32.

[7] Wright, M.N., and A. Ziegler, 2017. “ranger: A fast implementation of random forests for high dimensional data in C++ and R”. Journal of Statistical Software, 77(1), 1-17.

[8] Meyer, H., C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss, 2018. “Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation”. Environmental Modelling & Software, 101(9), 1-9.

[9] Hengl, T., M. Nussbaum, M. Wright, and G.B.M. Heuvelink, 2018. “Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables”, PeerJ Preprints 6:e26693v2 (in review).