RESEARCH PROJECT

Using geomorphometric variables to predict and support benthic and pelagic biodiversity

MSc. Cuvuliuc Alex-Andrei

Table of Contents

1. Introduction

The morphology of the seabed is a good predictor of marine species biodiversity, as proven by previous studies which employed geomorphometric variables and achieved great performance [1], [2], [3], [4]. Such studies often derive land surface parameters (also called geomorphometric variables) [5] like slope, rougness and curvatures, from a DDM (Digital Depth Model) [4].

2. Objectives

In the context of the PhD programme, I propose the following research objectives:

Objective 1. Assess the predictive power of geomorphometric variables in machine learning or deep learning models for estimating the benthic and pelagic biodiversity;

Objective 2. Tune and compare the performance of various regressors, such as Random Forest, XGBoost, Support Vector Regression or Multilayer Perceptrons, for predicting benthic and pelagic biodiversity;

Objective 3. Build an open-source tool (e.g. a QGIS Plugin, Python library, or a web application) that can be used to estimate or view the benthic and pelagic biodiversity.

3. Methodology

In this chapter I will present a method to predict the benthic biodiversity using bathymetry and geomorphometric variables. I have performed the analysis using Python and I have open-sourced the code on GitHub. The project can be accessed using the following link: https://github.com/alecsandrei/biodiv.

3.1. Materials

I have used a dataset of benthic biodiversity (Figure 4) collected near the coast of Tuscany, in the NW Mediterranean Sea [6], [7], which allowed to compute the Margalef's richness index.

$$D = \frac{S - 1}{\ln(N)}$$

where

D Margalef's index
S Number of species
N Number of individuals

The dataset has 209 samples collected from 134 stations. However, to reduce the bias and the spatial autocorrelation, only the first sample from each station was selected. A random subset of 80% of the data were used for training, and 20% for testing.

For the bathymetry, I used an ensemble DTM offered by EMODnet through their Web Coverage Service [8], [9], [10], [11], [12], which I resampled to 500 meters spatial resolution. From the DTM I derived 35 geomorphometric variables using PySAGA-cmd [13], a Python library that works as a wrapper for SAGA GIS [14].

3.2. Training

In this subchapter I will introduce the training approach of the machine learning model employed to predict the Margalef's richness index for the study area.

For this regression problem I chose a machine learning model called XGBoost [15], [16] which is known for it's very high efficiency and accuracy. Despite it working well out of the box, in order to optimize the XGBoost model, hyperparameter tuning is required [17]. A popular hyperparameter tuning method is grid search, which randomly samples hyperparameters. However, it is not as efficient as hyperopt, a Python library which uses Bayesian optimization to sample hyperparameters [18].

Due to the issue of having high-dimensional data with a reduced amount of samples, principal component analysis was used to reduce the dimensionality. The number of principal components were tuned alongside the XGBoost hyperparameters (Table 3). This is a small drawback as we will be unable to explain which predictor has high contribution in the regression task. Before deriving the principal components, the input features were scaled.

To evaluate the hyperparameter tuning iterations, spatial cross-validation was performed with the spatial-kfold Python library [19]. In spatial cross-validation, instead of building the folds randomly, they are created using a spatial clustering algorithm such as KMeans. Randomly building the folds would result in spatial autocorrelation. In order to select the hyperparameters that minimize the spatial cross-validated root mean squared error, 100 tuning iterations were computed. The tuned hyperaprameter values are presented in Table 3. The model which minimized the root mean squared error was evaluated on the testing dataset.

4. Results

The preliminary results show that the model is able to explain 22% of the variance in the Margalef's richness index for the testing area, as reported by the R2 score (Table 4). Considering the

model uses exclusively geomorphometric variables with a single spatial scale, the results are promising. To improve the model, a multiscale geomorphometric analysis could be performed with the WhiteboxTools [20] software which offers a great set of tools for multiscale analysis. The spatial results can be visualized in Figure 5.

5. Conclusions

Considering the preliminary results, the proposed research objectives are attainable as geomorphometry was able to explain some variance of benthic biodiversity. A similar methodology can be employed to predict the biodiversity for the pelagic domain. Perhaps one of the severe limitations of the proposed method was the high-dimensionality of the data when compared to the number of samples. For the future, instead of principal component analysis, a method like recursive feature elimination could be tested in order to reduce the dimensionality of the data, which will help explain which geomorphometric variables have high predictive power.

6. Figures and tables

Table 1: Descriptive statistics of Margalef's richness index for the study area

Descriptive statistics	**Margalef's richness index**
mean	5.43
standard deviation	3.59
minimum	0.0
Q1	2.9
Q2	4.64
Q3	6.63
maximum	18.39

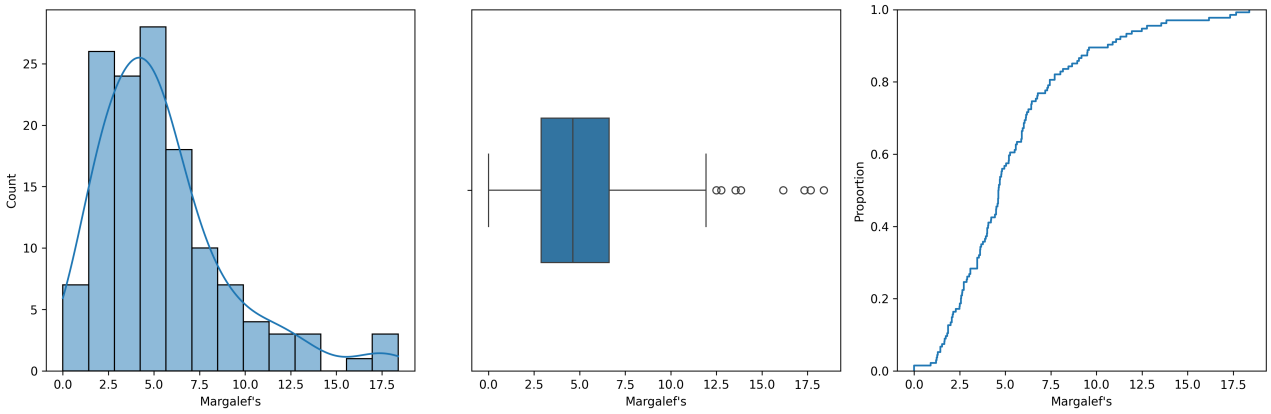

Figure 1: Explanatory data analysis for the predicted variable (Margalef's richness index)

Table 2: The geomorphometric variables computed in SAGA GIS using the DTM

Abbreviation	**Geomorphometric variable**
ioc	Index of convergence
shade	Hillshade
conv	Terrain surface convexity
poso	Positive topographic openness
nego	Negative topographic openness
aspect	Aspect
slope	Slope
northness	Northness
eastness	Eastness
cprof	Profile curvature
cplan	Plan curvature
cgene	General curvature
croto	Flow line curvature
ctang	Tangential curvature
clong	Longitudinal curvature
ccros	Cross-sectional curvature
cmini	Minimal curvature
cmaxi	Maximal curvature
ctota	Total curvature
dem	Digital elevation model
area	Real surface area
tpi	Topographic position index
vld	Valley depth
tri	Terrain ruggedness index
vrm	Vector ruggedness measure
clo	Local curvature
cup	Upslope curvature
clu	Local upslope curvature
cdo	Downslope curvature
cdl	Local downslope curvature
flow	Flow accumulation
fpl	Flow path length
spl	Slope length
cbl	Cell balance
twi	Topographic wetness index
wind	Wind exposition index

Table 3: The tuned hyperparameters in the pipeline (principal component analysis and XGBoost)

Hyperparameter	**Range**	**Sampling**	**Quantized (hyperopt)**	**Tuned value**
xgboost_eta	1e-6, 1.0	loguniform	No	3e-3
xgboost_reg_alpha	1e-6, 2.0	loguniform	No	8e-2
xgboost_reg_lambda	1e-6, 2.0	loguniform	No	1.81
xgboost_gamma	1e-6, 64.0	loguniform	No	1.01e-6
xgboost_subsample	5e-1, 1.0	uniform	Yes	0.55
xgboost_colsample_bytree	3e-1, 1.0	uniform	Yes	0.85
xgboost_max_depth	2, 8	loguniform	Yes	4
xgboost_n_estimators	10, 1000	loguniform	Yes	961
pca_n_components	1, 36	uniform	Yes	13

Table 4: Metrics for the training and testing datasets using the tuned XGBoost regression model

Metric	**Train**	**Test**
Mean squared error	2.62	5.39
Mean absolute error	1.15	1.78
Root mean squared error	1.62	2.32
R2 score	0.81	0.22

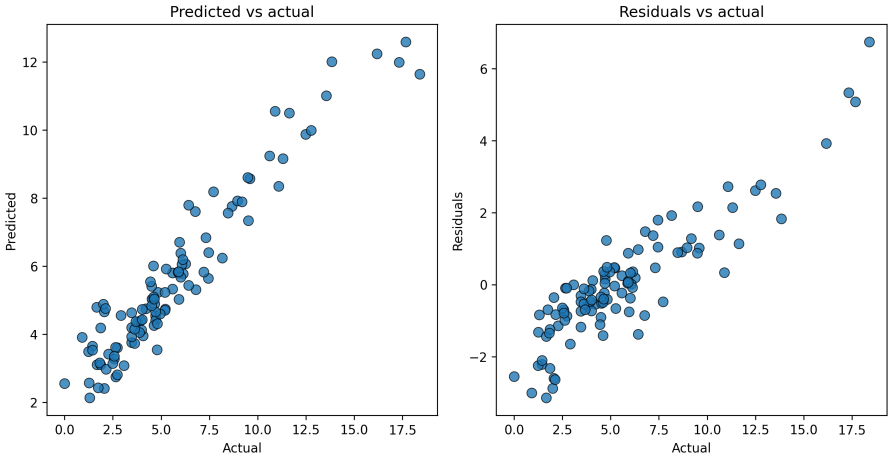

Figure 2: Actual vs predicted and residuals for the training set

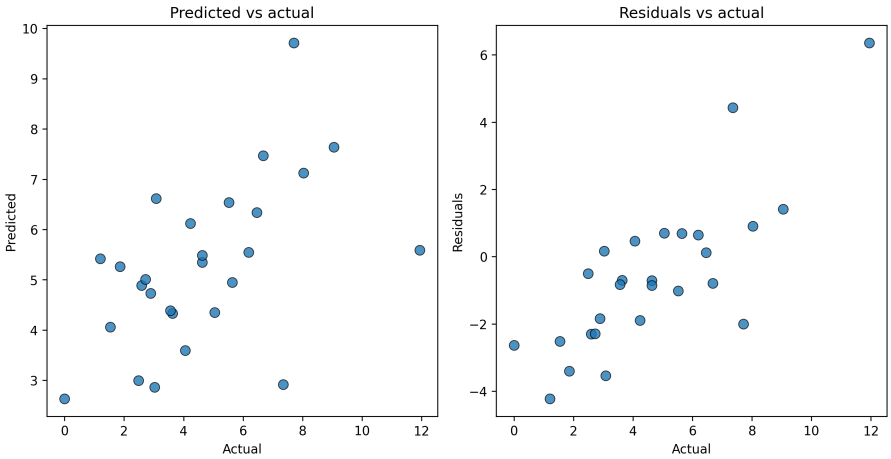

Figure 3: Actual vs predicted and residuals for the testing set

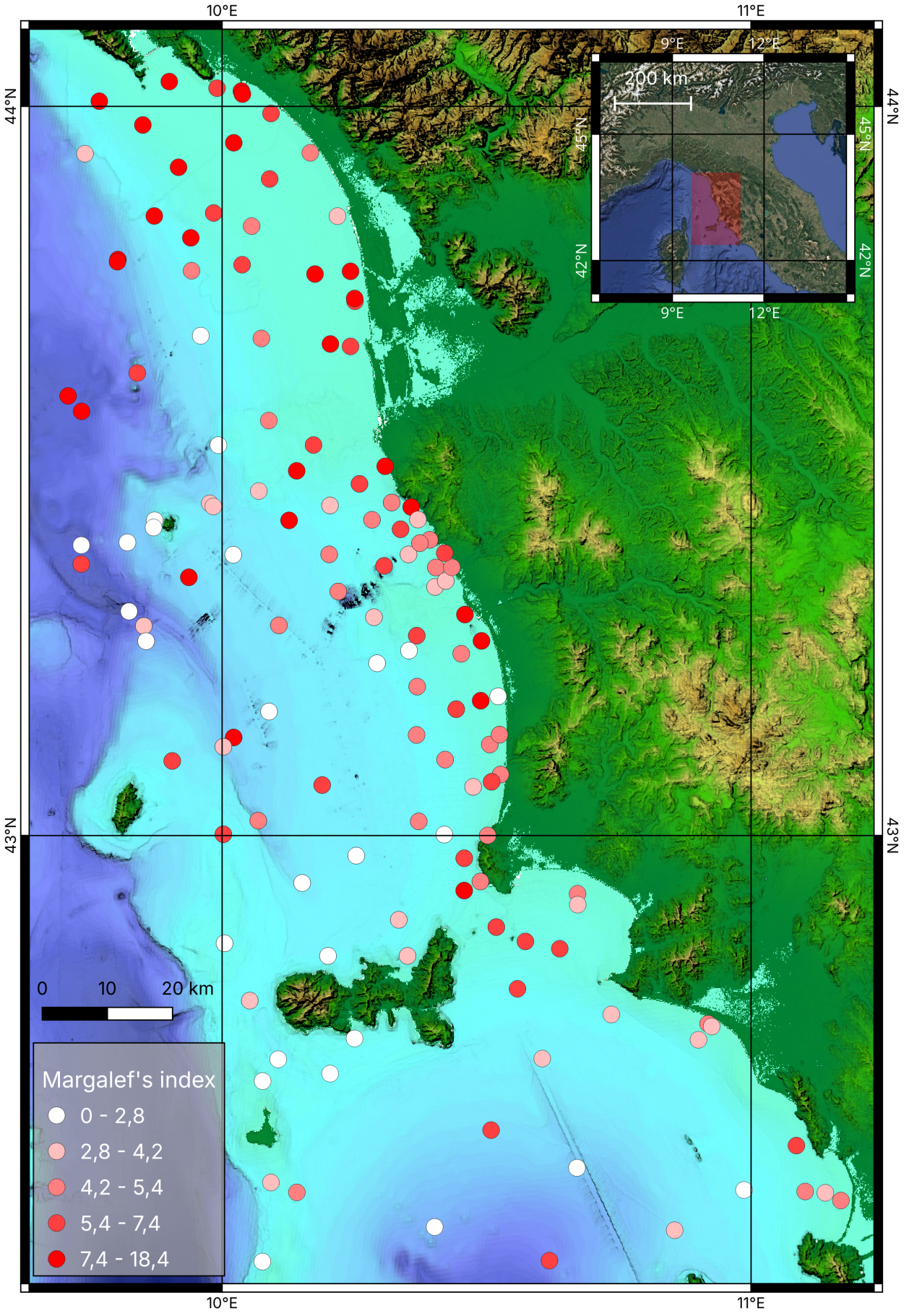

Figure 4: Distribution of the data samples in the study area

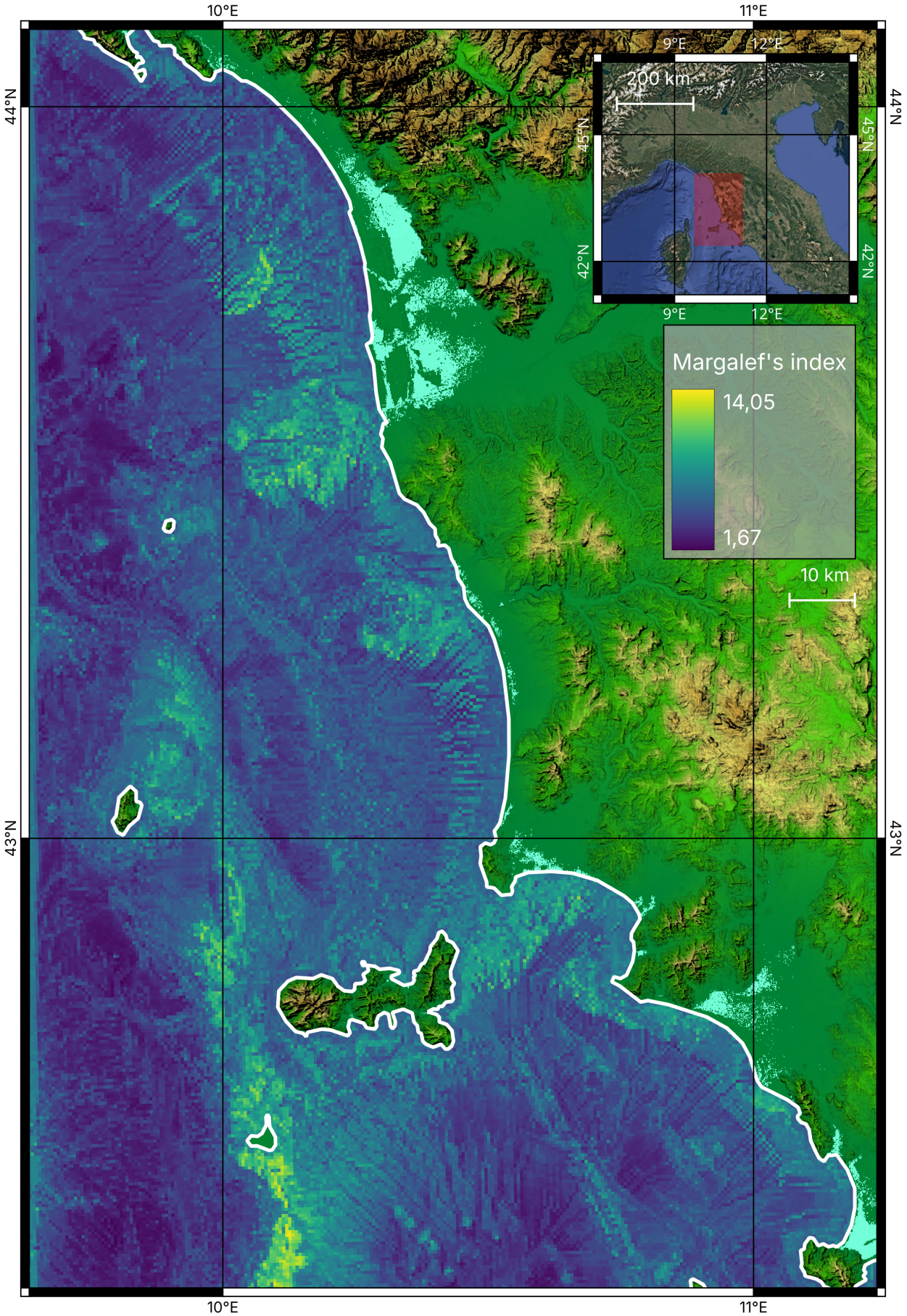

Figure 5: Predictions of Margalef's richness index for the study area using the XGBoost model

7. References

[1] K. L. Yates, C. Mellin, M. J. Caley, B. T. Radford, and J. J. Meeuwig, ‘Models of Marine Fish Biodiversity: Assessing Predictors from Three Habitat Classification Schemes’, *PLOS ONE*, vol. 11, no. 6, p. e0155634, June 2016, doi: 10.1371/journal.pone.0155634.

[2] L. Wedding, A. Friedlander, M. McGranaghan, R. Yost, and M. Monaco, ‘Using bathymetric lidar to define nearshore benthic habitat complexity: Implications for management of reef fish assemblages in Hawaii’, *Remote Sens. Environ.*, vol. 112, pp. 4159–4165, Nov. 2008, doi: 10.1016/j.rse.2008.01.025.

[3] S. Pittman, B. Costa, and T. Battista, ‘Using Lidar Bathymetry and Boosted Regression Trees to Predict the Diversity and Abundance of Fish and Corals’, *J. Coast. Res. - J Coast. RES*, vol. 53, pp. 27–38, Nov. 2009, doi: 10.2112/SI53-004.1.

[4] A. Collin, P. Archambault, and B. Long, ‘Predicting Species Diversity of Benthic Communities within Turbid Nearshore Using Full-Waveform Bathymetric LiDAR and Machine Learners’, *PLOS ONE*, vol. 6, no. 6, p. e21265, June 2011, doi: 10.1371/journal.pone.0021265.

[5] R. J. Pike, I. S. Evans, and T. Hengl, ‘Chapter 1 Geomorphometry: A Brief Guide’, in *Developments in Soil Science*, vol. 33, T. Hengl and H. I. Reuter, Eds, in Geomorphometry, vol. 33. , Elsevier, 2009, pp. 3–30. doi: 10.1016/S0166-2481(08)00001-9.

[6] P. Vassallo, C. Paoli, S. Aliani, S. Cocito, C. Morri, and C. N. Bianchi, ‘Benthic diversity patterns and predictors: A study case with inferences for conservation’, *Mar. Pollut. Bull.*, vol. 150, p. 110748, Jan. 2020, doi: 10.1016/j.marpolbul.2019.110748.

[7] S. Aliani, C. Bianchi, and C. Morri, ‘Lineamenti del benthos dei mari toscani’, *Atti Soc Tosc Sc Nat Mem*, pp. 77–92, 1995.

[8] EMODnet Bathymetry Consortium, ‘EMODnet Digital Bathymetry (DTM 2024).’ doi: https://doi.org/10.12770/cf51df64-56f9-4a99-b1aa-36b8d7b743a1.

[9] EMODnet Bathymetry Consortium, ‘EMODnet Digital Bathymetry (DTM 2022).’ doi: https://doi.org/10.12770/ff3aff8a-cff1-44a3-a2c8-1910bf109f85.

[10] EMODnet Bathymetry Consortium, ‘EMODnet Digital Bathymetry (DTM 2020).’ doi: https://doi.org/10.12770/bb6a87dd-e579-4036-abe1-e649cea9881a.

[11] EMODnet Bathymetry Consortium, ‘EMODnet Digital Bathymetry (DTM 2018).’ doi: https://doi.org/10.12770/18ff0d48-b203-4a65-94a9-5fd8b0ec35f6.

[12] EMODnet Bathymetry Consortium, ‘EMODnet Digital Bathymetry (DTM 2016).’ doi: https://doi.org/10.12770/c7b53704-999d-4721-b1a3-04ec60c87238.

[13] A.-A. Cuvuliuc, *PySAGA-cmd*. (Mar. 19, 2024). Zenodo. doi: 10.5281/zenodo.10839988.

[14] O. Conrad *et al.*, ‘System for automated geoscientific analyses (SAGA) v. 2.1. 4’, *Geosci. Model Dev.*, vol. 8, no. 7, pp. 1991–2007, 2015.

[15] *dmlc/xgboost*. (Aug. 03, 2025). C++. Distributed (Deep) Machine Learning Community. Accessed: Aug. 03, 2025. [Online]. Available: https://github.com/dmlc/xgboost

[16] T. Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System’, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[17] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, ‘Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data’, *Ecol. Model.*, vol. 406, pp. 109–120, Aug. 2019, doi: 10.1016/j.ecolmodel.2019.06.002.

[18] J. Bergstra, D. Yamins, and D. Cox, ‘Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures’, in *Proceedings of the 30th International Conference on Machine Learning*, PMLR, Feb. 2013, pp. 115–123. Accessed: Aug. 03, 2025. [Online]. Available: https://proceedings.mlr.press/v28/bergstra13.html

[19] W. Ghariani, *WalidGharianiEAGLE/spatial-kfold*. (Aug. 03, 2025). Python. Accessed: Aug. 03, 2025. [Online]. Available: https://github.com/WalidGharianiEAGLE/spatial-kfold

[20] J. Lindsay, ‘The whitebox geospatial analysis tools project and open-access GIS’, presented at the Proceedings of the GIS Research UK 22nd Annual Conference, The University of Glasgow, 2014, pp. 16–18.