Regression of biodiversity indicators from satellite observations and environmental parameters

MALIS Project Update Report, Fall Semester 2020

Vincenzo Madaghiele, Cosimo Chetta

I. Motivation

Ecosystems provide the basic support to human life. They ensure that the climatic conditions on Earth are stable, provide indispensable nutrients to the soil and regulate the air we breathe. The survival of ecosystems depends upon the survival of the biodiversity that they host. The more an ecosystem is biodiverse, the more it is stable and able to cope with changes. Therefore, it is fundamental to protect and monitor the biodiversity on Earth, which is decreasing at a fast pace due to human activities. Biodiversity is difficult to measure because it is defined by a series of very complex relations among species, so many variables have to be taken into account in order to explain biodiversity changes. Our experiment aims at exploiting the data generated by the EU satellite technology and sensor networks to provide a reliable estimation of the conditions of biodiversity in a given habitat.

II. Method

Our process was inspired by the Essential Biodiversity Variables (EBVs), a set of variables currently in stage of standardization ([1]). We have therefore followed the method described in [2], which defines the stages for data collection and production regarding the Biodiversity Distribution and Abundance data. In this first section of the experiment we have focused on building an appropriate database by collecting and aggregating different environmental variables, and implementing the first regression models.

III. Preliminary experiments

A. Dataset builduing

The first thing we have done is to build an appropriate training dataset for our model. We have built our dataset using publicly available geospatial data offered by the Copernicus EU space program and the European Environment Agency (EEA). This dataset is composed of features from four main sources:

- **Copernicus Land** [3]
- **Copernicus Climate Change** [4]
- **EEA Richness of forest-related species indicator** [5]

As pointed out in [6], time and space scale are fundamental in biodiversity monitoring, and scales have to be chosen according to the specific phenomena to be measured. The aim of our experiment is to provide a reliable estimation of the overall status of biodiversity in a given area, so we have opted for a community level diversity measurement, and we choose time and space scales accordingly. The Land, Climate and Atmosphere datasets were downloaded taking into account data of the same year and month. When the data was not available for the selected year, we have dowloaded it for the same month of the nearest year. We have chosen to do so because environmental variables do not change much year by year, but it is fundamental to roughly maintain the same month because of the high seasonal dependence of the data. We have chosen a spatial sampling rate of 0.01 degree of latitude/longitude, so each sample in our dataset contains the data of a cell of 0.01 x 0.01 degrees of latitude/longitude.

B. Selection of biodiversity indicator

Measuring biodiversity is a complex task, and many approaches have been proposed in the literature ([7]), depending on the specific need and application of the measure. It is possible to measure biodiversity on all levels of life, from ecosystem diversity to genetic diversity among the same population. In the case of our experiment, we needed a global indicator which could state the overall status of the biodiversity in an area inside the same habitat, and could also be publicly available in a geospatial dataset for download, so we choose to measure species diversity inside the same habitat. The indicator we used is the **Richness of forest-related species and habitats indicator**. This index was developed by the European Environment Agency in order to asses the conservation status of European forest habitats. It expresses the richness of species related to forest habitat as a number between 0 and 1, calculated using a series of sub-indicators. It is publicly available just for European territory [5], and just for one time period in 2012.

C. Area selection

Building a training set for this experiment requires a prior knowledge of the basic ecology of these areas. We have constructed four basic datasets, which represent different biodiversity hotspots in Europe. We have selected the areas among the Sites of Community Importance (SCI), defined in the European Commission Habitats Directive (92/43/EEC) and in the European Commission Bird Directive (79/409/EEC), protected under the Natura 2000 network. We have considered the different types of forest areas and the different biogeographycal regions. We have also considered the internal ecological

similarity of all these data, in order to obtain homogeneous datasets. The chosen areas are shown in table I.

TABLE I: Selected areas

State	*Latitude*	*Longitude*	*Biogeographical region(s)*
France	3.6-4.5	44.4-44.8	Mediterranean, Alpine
Finland	25.0-28.0	67.0-69.0	Boreal, Alpine
Italy	12.5-15.5	40.5-43.0	Continental, Mediterranean, Alpine
Bulgaria	22.0-27.0	41.0-44.0	Continental, Alpine

D. Data cleaning

The Copernicus Climate database values are well defined, there are no missing values and from the exploration of the data we have not found any particular issue therefore we have not performed any computation. It has a precision of 0.1 degree of latitude/longitude so we have joined this dataset with the others by finding the closest point. Since the two remaining datasets have a precision of 0.01 degree, one row of Copernicus Climate has been joined with approximately other ten rows of Land/Richness. The EEA Richness of forest-related species indicator is supposed to have values in range (0,1]. Values less than 0 signal a non-forest areas therefore we have removed those coordinates. For the Copernicus Land database, we have tried different approach to handle the non valid data:

- Remove the rows with invalid features
- Fill with mean of the feature
- Fill with the mean of the four closest coordinates
- Fill with a KNN-imputer

The Copernicus Land documentation reports information about errors in the data acquisition. We have considered the errors in data acquisition as invalid, and we have set the values which were too high or too low respectively to the maximum and the minimum values of the feature. We have considered values that are outside the feature boundaries reported in the documentation as invalid. Finally, by plotting the data of each feature we found some values of the Albedo (ALBH/ALDH) and Top-Of-Canopy Reflectance (TOCR) features that could be considered as outliers. For those two group of features, we set as invalid the values with a z-score greater than 3. After having merged the three dataset over the latitude-longitude coordinates, we tested regional data using cross validation on a random forest regressor.

TABLE II: Outlier handling methods performances

Region	Remove	Mean	Closest mean	knn imputer
Bulgaria	0.918	0.925	0.925	0.923
Finland	0.713	0.717	0.717	0.717
France	0.732	0.742	0.742	0.735
Italy	0.803	0.832	0.827	0.835

As can be seen from the results in Table II, the only outliers handler that caused a meaningful score reduction is the removal of rows with outliers. The other techniques have similar scores, so we chose to continue with the mean of closest coordinates since it performed slightly better than the other two.

E. Comparison of regression models

After obtaining the dataset, we have performed various experiments to select the best model. We have compared four different regression techniques: decision tree, random forest SVM and neural network. We have made a grid search for each model in order to obtain the best performing parameters. In order to measure and compare performances we have used the coefficient of determination R^2 ([8]) as our score. The results of this experiment can be seen in Figure 1.

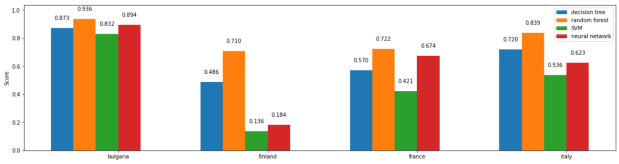


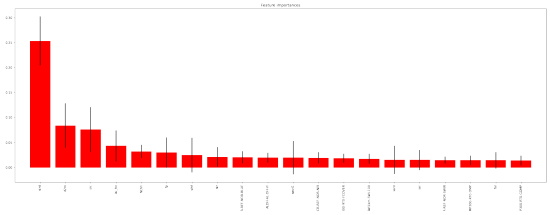
Fig. 1: Score of the model for each region

As can be seen from the figure, the performances are different depending on the model and on the dataset. The best restults have been obtained using random forest on the Bulgaria dataset, while the worst results regard SVM on the Finland dataset. These result are probably due to the differences in the internal composition of the datasets. The Bulgaria dataset better represents the regression index because of the higher internal variability, while the Finland data is more homogeneous and it is not able to correctly describe some of the more different points. Overall, it is possible to observe that the Random Forest model has given the best result with respect to all the other methods for all of the areas, reaching satisfying result in all of them. This is an interesting result, because it shows that, even if the areas are different, there is a coherence among the type of features and how they react to different models.

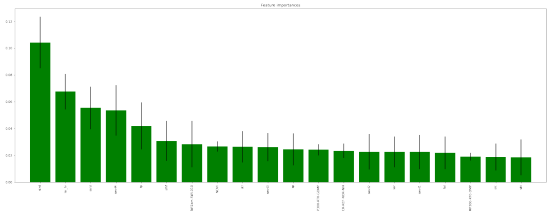
F. Feature importance

We have used random forest to highlight the contribution of each feature to the biodiversity of each area. In order to do so we have trained a different random forest on each dataset and printed the feature importances as a histogram. In figure 2 it is possible to see an example of the importance histogram of France and Finland. The features are displayed on the x axis in order of importance, so it is possible to see how much each feature contributes to the result.

As can be seen from the figure, even though the most important features are similar for the two areas, the importances are distributed in a different way. For example in the graph of Finland many features have a high contribution to the result, while in the case of France the importance is concentrated into fewer features. This could be due to the differences in environmental conditions of these areas, but also to the degree of internal homogeneity of the datasets. More homogeneous areas with fewer, more defined habitats are better described



(a) Most important features for the France dataset (first 20 features)



(b) Most important features for the Finland dataset (first 20 features)

Fig. 2: Feature importances

by a smaller set of features, while bigger areas with higher internal variability are more difficult to describe and predict.

G. Visualization of the regression tree

Random forest and regression trees have been proven useful in many ecological monitoring and forecasting application ([9], [10]), because they provide a good alternative to traditional mathematical models. Being able to effectively visualize the results of the regression is crucial for the practical applications of this model, because it allows to understand the model's decisions. As said in section III-E, the best result in the regression task has been obtained by the random forest. However, visualizing the result of random forest in its entirety would not give many significant information, because the model is too large to be interpreted. In order to obtain a coherent representation of the model's result we have obtained the importance of each feature and then we have trained multiple regression trees on a subset of the most significant features. The best performing trees have then been plotted and compared. An example of the results is shown in Figure 3, which portrays the path followed by the model in order to predict the habitat richness of a sample area in the test set.

The figure portrays the feature space of each node variable chosen by the tree, and how the data was split to obtain the binary division. This tree was trained used only the 10 most important features of the Bulgaria dataset, the maximum depth was set to 4 branches and its overall regression R score on test set is equal to 0.76. Of course, a single tree is less accurate than the correspondent random forest model, but it is possible to inspect the tree by visualizing it. In this case the prediction for this test area was of 0.61 habitat richness. Three main variables were considered by the model for the regression: strd (Surface thermal radiation downwards), skt (Skin Temperature) and ssrd (Surface Solar Radiation Downwards). This kind of analysis is an example of a practical application of our model: first the model will make a more

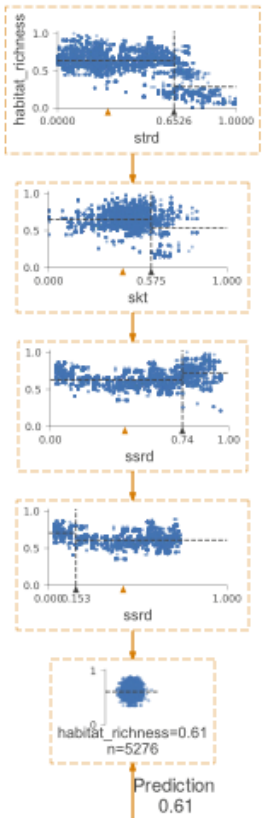


Fig. 3: Path followed by the Bulgaria Decision tree to regress a test sample

accurate prediction using random forest, then it is possible to investigate the reasons of the regression result reducing the number of features and plotting the regression tree, and intervene on those features if necessary.

IV. Next Steps

A. Include satellite images in the training set

Using satellite ortophotos is a common practice in biodiversity and habitat monitoring among ecologists. We want to include satellite images in our estimator models training sets, in order to achieve a higher accuracy in the prediction.

B. Test the model with different biodiversity metrics

The biodiversity indicator we have chosen provides a good general estimation of European forest biodiversity, however it is not a widely recognized indicator in the scientific community, so we will search and experiment with different, more popular indicators.

C. Test the generalization capabilities of the model

So far we have constructed regional models, which have been tested on ecologically homogeneous areas with similar climate and mostly protected by nature reserves. We will test how the different models perform on unseen data of areas with different degrees of similarity from the ones used for training.

V. Contributions

Cosimo Chetta: Copernicus Climate Change data retrieval, Dataset cleaning, Regression models comparison, Grid Search

Vincenzo Madaghiele: Copernicus Land data retrieval, Biodiversity dataset research and retrieval, Areas research, Feature Importances, Tree visualizations

References

[1] G. BON. (2015) What are ebvs? [Online]. Available: https://geobon.org/ebvs/what-are-ebvs/

[2] W. D. K. et al., "Building essential biodiversity variables (ebvs) of species distribution and abundance at a global scale," *Biological Review*, August 2017.

[3] C. EU. (2014) Copernicus global land service. [Online]. Available: https://land.copernicus.eu/global/

[4] ——. (2014) Copernicus global climate change service. [Online]. Available: https://www.copernicus.eu/en/services/climate-change

[5] E. E. Agency. (2012) Richness of forest-related species and habitats indicator 2012 dataset, nov. 2018. [Online]. Available: https://sdi.eea.europa.eu/catalogue/srv/eng/catalog.search/metadata/81754d01-8bc3-49aa-a52c-86b3d212f94e

[6] C. B. Anderson, "Biodiversity monitoring, earth observations and the ecology of scale," *Ecology Letters*, vol. 21, p. 1572 – 1585, 2018.

[7] E. K. M. et al., "Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories," *Ecology and Evolution*, Sept 2014.

[8] Scikit-Learn. sklearn.metrics.r2$_score$.$[Online]$.$Available$: $https$: $//scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html$

[9] M. S. et al., "An applied statistical method to identify desertification indicators in northeastern iran," *Geoenvironmental Disasters*, 2018.

[10] "Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem," *Ecological Indicators*, vol. 60, pp. 870 – 878, 2016.