Assessing ML Models for Species Data Modelling

YU-CHI CHU

Abstract

The study compares the performance of machine learning models and different data preparation techniques for predicting sightings using species data from iNaturalist and world climatic variables from WorldClim. Our aim is to produce a model that cane be generalised and does not have to be tuned and trained for each species prediction. After preprocessing, Random Forest and XGBoost performed best, with Random Forest achieving 0.786 PR AUC. MaxEnt was also found to be effective but not efficient due to its high computational demand. Overall, Random Forest with Principal Component Analysis (PCA) offered the best balance of efficiency and a high evaluation score.

1 Introduction

The aim of this work is to compare the performance of different machine learning (ML) models and pre-processing steps in predicting the sightings of a species at a location. We use data on the sightings of the species, the sightings of other species plus climate and elevation data.

The complex relationships between species and their environment, and the big increase in data available from sources such as remote sensing and citizen scientists have driven interest in ML for species data modelling (SDM)[2] over the last 20 years. The data used in SDM falls into two main categories: Presence Only and Presence-Absence data[2] (Beery et al, 2021). The former, which includes the data collected from iNaturalist[7] used in our analysis, is Presence Only and is based on observations of a species at a location. It does not specifically say the species is absent at a location but rather a sighting has not been recorded. Challenges with such data include training a model without definitive information on where the species is absent[6] (Hastie and Fithian, 2013) and sighting locations being heavily biased to areas accessible to observers[9] (Reddy and Davalos. 2003).

A range of different ML models have been used for SDM. The non-linear relationships given the complexity of ecosystems, imbalanced and correlated data, non i.i.d data, the need for transparency on the most important features and the afore mentioned presence only data have all impacted choices. Decision Tree structures able to cope with the non-linear, non-i.i.d and imbalanced data have been used by multiple authors including Random Forests[3] (Cutler et al., 2007) and XGBoost[5] (Elith et al, 2008). SVM has also been used [4](Drake et al., 2006). Pseudo-absence data has been introduced to these models to handle presence only data [6](Barbet-Massin et al., 2012). The MaxEnt algorithm [8] (Phillips et al., 2006) was specifically introduced to SDM to cope with presence only data and has been one of the most popular and often cited ML models for datasets of this type.

A greater understanding of the capabilities of the best ML approaches to predict the location of species is ever more critical given the pressures on biodiversity of climate change and wider human impact.

2 Data preparation

The data for our modelling is taken from two sources. The species sighting data is from the train_extra dataset provided from iNaturalist, and climate and elevation data is from WorldClim[10]. The species data consists of 1.3 million sightings of 2,418 species spread across all continents. We have also used the 19 bioclimatic variables and elevation data (see appendix) from WorldClim. The number of sightings per species in the dataset range from 50 to 2,000. Whilst the distribution is heavily skewed to the lower sightings per species (Figure 1), over 10% of the species have 2,000 sightings indicating there are sufficient sightings per individual species to train our ML models.

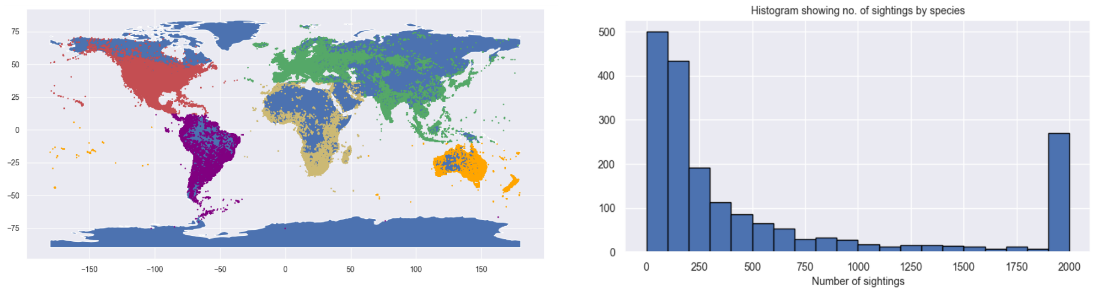


Figure 1: (LHS) Map of sightings by geographic area used in our analysis
. (RHS) Histogram showing the number of sightings for each species

Geographic Region	N America	S America	Europe & Asia	Africa & Middle East	Oceana	Global
Species	684	537	773	705	387	2418
All sightings	609841	109397	272012	161534	186845	1339629

Table 1: Geographic distribution of dataset used in our analysis

We divided the data in to regions for computational efficiency and to reduce the imbalance of sightings verses no sightings. The splits were informed by analysis of the continuity of species sightings. It effectively replicated sea boundaries between land masses so we approximated longitude and latitude rectangles for these areas .

Binning Data: To enable models to run effectively with the computing power available we aggregated data to spatial bins (averaging bioclimate and counting species sightings). The increased efficiency of a smaller dataset was weighed against the loss of detail from aggregation. To examine the trade off we used a section of North America (latitudes 15°- 53° & longitudes -130°- -60°). We assessed bin sizes of 0.25°, 0.5°, 1° and 2°. We compared the count of bins with a species present verses all bins in the area (a proxy for the data sparseness) . We also looked at the distribution of the count of bins for each species to ensure sufficient breadth of geographic sampling to train our ML model. Our analysis indicates 1° bins balance the trade off best with 65% of possible bins filled and 10% of the species in the sample area being present in more than 370 bins.

2.1 Data Quality:

To ensure the reliability of our analysis, we first examine the spatial distribution of sightings for each species to identify irregularities.

Convex Hull: The core of our approach to classifying species as global/local. It constructs a convex polygon encompassing all data points of a species. By analysing its area and perimeter, we can determine the extent of the species' distribution. Table 2 shows a comparison and why we chose this algorithm. To further improve the accuracy of our analysis, we employ two outlier detection techniques:

LOF: Applied on local species based on their local density deviation, suitable for detecting local anomalies and sensitive to parameter selection.

Feature	Convex Hull	DBSCAN	K-Means
Data Requirements	Only need 3 data/species	Determine epsilon, minPts	Determine K value
Cluster Shape	Arbitrary shapes	Arbitrary shapes	Sphere/Elliptic
Complexity	Efficient	Expensive	Sensitive to initialization

Table 2: Comparison to different algorithms

Isolation Forest: Applied on global species, isolating anomalies by partitioning the data space, well-suited for large datasets and global outlier detection.

From the total number of 1339248 data points 61682 were classified as outliers bringing down the number of data points to 1277566.

Some species initially classified as global were reclassified as local after the removal of outliers. This highlights the importance of outlier detection in obtaining reliable results (see Figure 2).

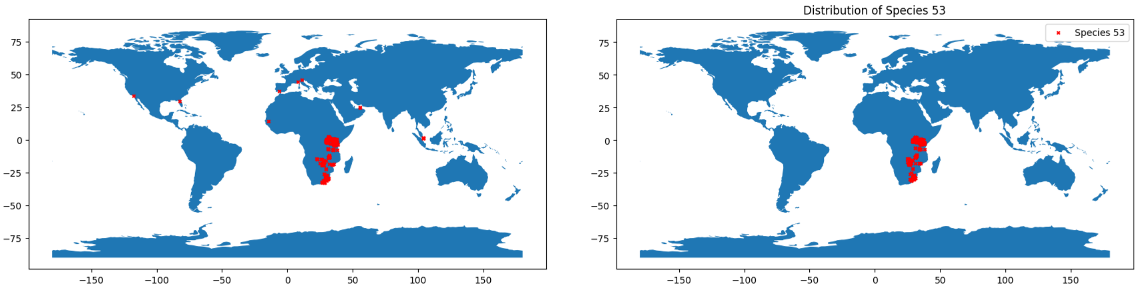


Figure 2: Species ID = 53 Distribution Before/After Outlier Removal (Global->Local)

3 Exploratory data analysis

High correlation of features can lead some ML models to over weight on particular features. In our dataset a significant number of sightings of different species are highly correlated and feature engineering may help model performance. Over 4,836 pairs of species have a Pearson correlation co-efficient greater than 0.8 (there were no coefficients < -0.8). This reflects that species do co-exist in common locations but it is also likely to be impacted by bias in the collection of data from common survey areas.

Bioclimate variables (features for our modelling) also demonstrate high levels of correlation with each other. Removing those with a Pearson correlation coefficient of more than 0.8 reduces the number of bioclimate features by half (see Figure 3) .

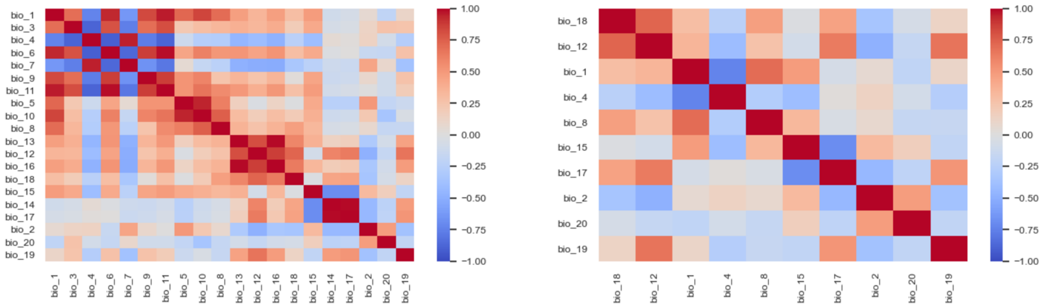
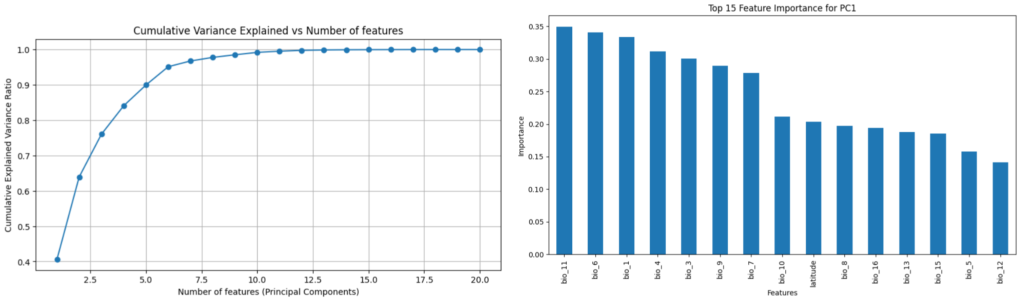


Figure 3: (LHS) Pearson correlation co-efficient for bioclimate data
. (RHS) Bioclimate variables removed so none coefficients > 0.8

To assess in more detail the importance and principal components of the spatial and bioclimatic features we have undertaken Principal Component Analysis (PCA). First, we standardize features and reduce the impact of outliers by limiting extreme values to a predefined range, thus enhancing the robustness of the PCA results. The eigenvectors (principal components) represent the directions of max variance in the data, and feature importance provided insight into the dominant factors. The elbow plot shows the first 6 principal components (PCs) account for 95% of the variance and hence we have used these 6 PCs in the model runs where PCA is applied.

This analysis demonstrates the core variables that require emphasis in modelling species' spatial and environmental interactions. By understanding these key factors, researchers can focus on high-impact variables for further ecological studies or policy-making.



(a)

(b)

Figure 4: (a) Elbow plot of cumulative variance vs Principal Components used. (b) Feature importance

4 Learning methods

Model	Why Chosen
Random Forest	Less prone to overfitting, handles imbalanced data well Can handle both numerical and categorical features Performs well with large numbers of features Provides insights into the relative importance of features Can handle missing values and perform some feature selection
XGBoost	Efficient and accurate Regularization to prevent overfitting Handles imbalanced and correlated data
SVM	Effective in high-dimensional spaces Strong theoretical foundation and good generalization performance Handles imbalanced data through class weighting Kernel trick for non-linear decision boundaries
MaxEnt	Specifically designed for species distribution modeling Accounts for presence-only data Provides insights into the importance of environmental variables Flexible to incorporate different types of environmental data

Table 3: Reasons for Choosing Model

Random Forest (RF): RF is an ensemble learning algorithm used for classification and regression tasks. It combines multiple decision tress. RF is less prone to overfitting and works well with high dimensional data. Moreover, it does not require major preprocessing like normalising and removal of correlated features. However, training the model can be computationally expensive and resource intensive. In previous SDM work RFs have performed well even with imbalanced or correlated data.
XGBoost (XGB): XGB is an advanced implementation of gradient boosting, a technique that builds a model incrementally by training each new model to correct the errors of the previous model. XGBoost uses a series of decision trees where each new tree aims to reduce the errors of the previous tree. A major improvement of XGB over other gradient boosters is its inclusion of regularization to prevent overfitting. The algorithm also works well with imbalanced and correlated data. PCA is not

necessary for XGB however it can be usefule to compare results.

Unlike random forest where the tress are built in parallel the trees are built sequentially. While this might be computationally easier it does increase training times as the size of the dataset and the number of trees increases. Another drawback of XGBoost is that it can be less interpretable

Support Vector Machine (SVM): SVM is a powerful supervised machine learning algorithm used primarily for classification tasks. SVM aims to find the best possible boundary (called a hyperplane) that separates data points into different classes. This boundary is chosen to maximize the margin between the classes. SVM can handle correlated features and high-dimensional spaces like our data. It can also handle imbalanced data through the use of class weights, allowing for penalization of misclassification of the minority class. However it has high training times forcing the hyperparameter space to be smaller to improve efficiency.

MaxEnt: MaxEnt is a machine learning method often applied to SDM and is considered one of the benchmarks for modelling presence only data. The purpose is to maximise the distribution of possible data (entropy), making the model as uniform and unbiased as it can while still being able to fit to the input data. MaxEnt measures the probability distribution of different species across a geographical area using environmental variables and presence only occurrences, which is useful for our task.

MaxEnt models species distribution based on environmental variables and background points. The background points are introduced to represent pseudo-absence data[1]. MaxEnt has only a few hyperparameters, we found the optimum number of background points to be 10,000, This ensured we captured the environmental conditions, but limited the risk of overfitting. We set iterations to 500 so the model converged but not to computational inefficiency. We used 5 fold cross validation consistent with the other models.

MaxEnt has a defined structure for input and output data so there were differences in data prep and outputs to our other models. We used binned occurrence data for the target species. MaxEnt uses continuous environmental data for features No other species data was used. We removed the highly correlated bioclimate data (shown in the analysis above) as MaxEnt is poor at handling correlated input data. We did not use PCA as we wanted to maintain interpretability of the climate feature importance. We used the area under an ROC curve to compare to other models as P-R data was not available from our version of MaxEnt.

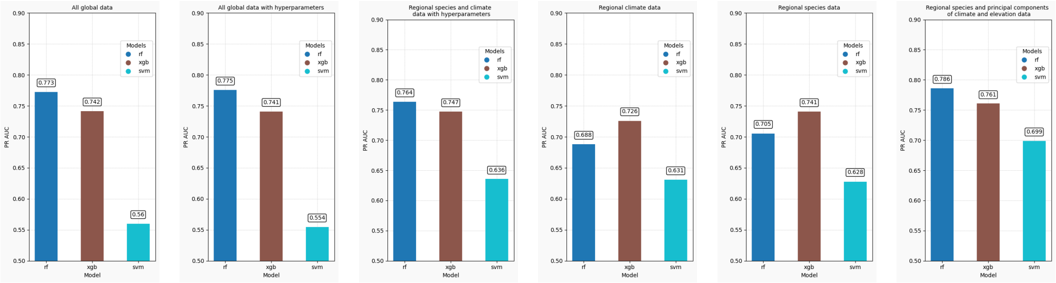
5 Results

Models were evaluated on normalised data using 5 fold cross validation to compare performance. This ensured that each model gave consistent results. The 3 models (RF, XGBoost and SVM) were run on the different approaches to data preparation. PR AUC values were used to select the best model and data preparation process . PR AUC was used over ROC AUC or other metrics because the data available is highly imbalanced towards the negative side and, in such cases, PR AUC provides a better measure of model performance.

15 target species were randomly selected from species with the highest spatial distributions (top decile number of bins) to optimize the models. The target species was set as the variable to be predicted while all the other species served as features. This was then repeated over 15 different target species along with hyperparameter tuning to get the optimal set of hyperparameters. These were then generalised to all the other species in the data. The newly obtained hyperparameters were tested over a set of 10 different species to evaluate whether they perform accordingly.

The models were initially run over the training data in its raw form (post data quality and binning) to baseline the performance of each model (fig 5.a). Random Forest and XGBoost gave PR AUC values of 0.773 and 0.742 while SVM performed poorly. Better results were achieved by fine tuning hyperparameters and by training models on regional data. The maximum PR AUC score of 0.786 was achieved by running Random Forest over a variation of the data containing the species in a region and the principal components of the environmental data accounting for 95% of the variation(fig 5.f). The next best score of 0.775 was from the Random Forest model using all global data with fine tuned hyperparameters(fig 5.b). Even though this is a good score training with the global data over all the species becomes computationally expensive and time consuming.

The MaxEnt algorithm achieved an average ROC AUC score of 0.9503 by running over the different target species. This score though good is still outperformed by Random Forest and XGBoost running on the global data with tuned hyperparameters.



(a) (b) (c) (d) (e) (f)

Figure 5: Comparing the PR AUC of the models over different variations of data.

Scenario No.	Input data (All data has outliers removed)
Scenario 1	Global species and climate data without hyperparameter tuning
Scenario 2	Global species and climate data
Scenario 3	Regional species and climate data
Scenario 4	Regional climate data
Scenario 5	Regional species data
Scenario 6	Regional species data and principal components of climate data

Table 4: Different scenarios of data preprocessing

6 Inferences

Both Random Forest and XGBoost performed well over all variations, in most cases with Random Forest coming up on top. SVM however performs poorly, this could be due to the imbalance of the data as the algorithm performs noticeable better on regional datasets with the less sparse values. The best values were with Random Forest running on the principal components of the environmental species and the species data, but this brings up the problem of feature interpretability being harder.

Though the models seem to work well with the global data this might not be the most optimal solution as trying to make predictions for each species on a global scale would make modelling computationally expensive and time consuming. Moreover the additional data from regions remote to the species taken as the target, logically should not be of importance to the model.

The MaxEnt model, though it run almost as well the best performing model, is not viable here as it is time consuming to run and it requires more complex tuning and preprocessing.

The MaxEnt model though it was mentioned extensively in the literature did not perform as well as the Random Forest algorithm. This could be because of the data being extremely imbalanced or because of the non linear relationships between the features which could not be addressed properly

7 Conclusions

Overall the best combination after assessing computational availability, different ML models and different approaches data preparation was to use Random Forest on the principal components of the bio climate data plus the other species data as features .

The models choosen were based on their ability to handle the challenges of the dataset provided and RF and XGBoost demonstrated good performance, SVM to a lesser extent. Notably, RF and XGB performed well without us specifically adding pseudo-absence data.

There are many areas we'd like to build on this work including introducing pseudo-absence data, a deeper analysis on bin sizes, and additional environmental features.

8 Citations and References

References

[1] Morgane Barbet-Massin, Frédéric Jiguet, Cécile H. Albert, and Wilfried Thuiller. Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, 2012.

[2] Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review. *COMPASS '21: Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, 2021.

[3] D. Richard Cutler, Jr Thomas C. Edwards, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, and Joshue J. Lawler. Random forests for classification in ecology. *Ecology, 88(11), 2007 pp. 2783–2792*, 2007.

[4] John M Drake, Christophe Randin, and Antoine Guisan. Modelling ecological niches with support vector machines. *Applied Ecology 43, 424-432*, 2006.

[5] J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology 2008, 77, 802–813*, 2008.

[6] Trevor Hastie and Will Fithian. Inference from presence-only data; the ongoing controversy. *Ecography 36: 864–867, 2013*, 2013.

[7] iNaturalist. https://www.inaturalist.org/, 2020.

[8] Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. Maximum entropy modeling of species geographic distributions, 2005.

[9] Sushma Reddy and Liliana M. Davalos. Geographic sampling bias and its implications for conservation priorities in africa. *Journal of Biogeography 30(11)*, 2003.

[10] WorldClim. https://www.worldclim.org/data/worldclim21.html, 2020.

9 Appendix

9.1 PR AUC vs ROC AUC

PR AUC: PR AUC refers to the area under the PR curve of a model. It is a comprehensive view of the model's ability to balance precision and recall over an entire range of thresholds. It focuses more on the positive class and evaluates the trade-off between the precision and the recall. It is more sensitive to class imbalance and is often used when the cost of false positives and false negatives are significant. It is heavily influenced by how the model performs on the minority class.

ROC AUC: ROC (Receiver Operating Characteristic) Area Under the Curve is the are under an ROC curve. It provides an overall performance of a model across all classification thresholds. The curve is a plot between the true positive rate against the false positive rate at various threshold settings. A higher ROC AUC indicates that the model is able to distinguish between positive and negative instances with a higher chance of success.

PR AUC and ROC AUC are evaluation metrics both suited for evaluating a model but the heavy imbalance of the data makes PR AUC a better option. ROC AUC will give a higher score when the data is heavily imbalanced towards one side, even if the model is not performing up to the mark. This is why PR AUC has been used as the evaluation metric during hyperparameter tuning and model comparison.

9.2 Hyper-parameters tuned for each ML model

ML Model	Tuned Hyper-parameters
Random Forest	**n_estimators**: the number of trees **max_depth**: maximum depth of each tree **min_samples_split**: minimum samples required to split a node **min_samples_leaf**: minimum samples required to become a leaf node **max_features**: number of features considered when looking for the best split
XGBoost	**n_estimators**: number of boosting rounds or trees to build **learning_rate**: controls the contribution of each tree to the final prediction **max_depth**: maximum depth of each individual decision tree **min_child_weight**: minimum size of each leaf node **subsample**: fraction of the training data to randomly sample for each tree **gamma**: controls whether to split a node **scale_pos_weight**: controls the balance of positive and negative weights (useful for imbalanced classes)
SVM	**kernel**: function to map the original data to a higher-dimensional space for separation **C**: controls the trade-off between maximizing the margin and minimizing classification error **Coef0**: controls the influence of higher-degree polynomials and the sigmoid function **gamma**: defines the influence of a single training example

Table 5: Tuned Hyper-parameters for ML Models

9.3 Pearson's Cross Correlation of Species Data (500 species)

Clustered heatmap of Pearson's cross correlation between 500 species showing significant number of groups of highly correlation.

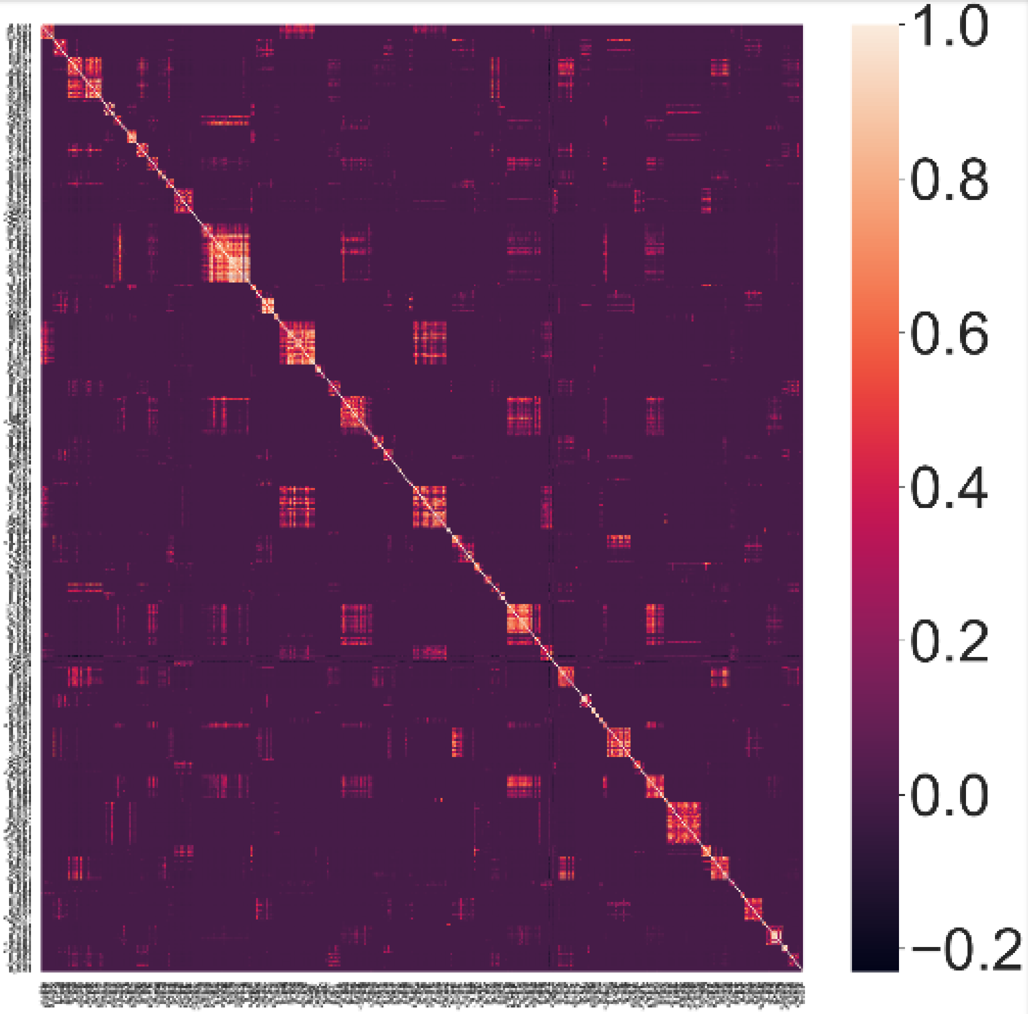


Figure 6: Clustered Pearson’s Cross Correlation of Species Data (500 species)