Biodiversity mapping

s2704516 **s2742860** **s2748897** **s2750316**

Abstract

This project aims to map and study global biodiversity, exploring key questions such as where particular species are located, what is their variety across different geographic areas, and how they interact with environmental factors and other species. Additional data (i.e., environmental and biological features) in combination with various ML techniques (e.g., K-means, Random Forest, K-Nearest Neighbors, LightGBM, XGBoost) and network analysis methods are leveraged to address these pattern recognition tasks, as the initial dataset — composed of online platforms' user-logged sighting records — is both noisy and low-dimensional compared to the underlying problem's complexity.

1 Introduction

Biodiversity mapping is a fundamental science that provides essential data for informed decision-making about species conservation and effective ecosystem management. Recent advancements in citizen science, where the public participates in gathering information on identifying taxa [11], combined with increasingly available and organized environmental and biological data [27], have made such analyses possible. Many methods in this field exist [10], with studies focusing on identifying regions with similar bioclimatic characteristics to assess species' distribution [26] and effectively map ecological regions. The main objective of this analysis is indeed to uncover these factors influencing animals' positioning worldwide, examining environmental conditions and inter-species relations, possibly providing functional data-centric tools for biodiversity protection.

2 Data

The positive-only data is composed of localized sightings for 500 species, provided as training and test sets consisting of $272\,037$ and $1\,706\,646$ records respectively. The former reflects user-reported observations from iNaturalist [19], while the latter represents expert-verified animal distributions from IUCN [17]. Since only one dimension of information was available, these datasets have been enriched with locations' bioclimatic information and species' biological characteristics, to better handle ecology-related questions and investigate models' performances with data increments.

The bioclimatic variables were extracted from WorldClim [16], using TIFF files georeferenced with EPSG:4326 coordinates: six key temperature and precipitation variables were selected, and their values queried for each unique sighting location (appx.1). By a preliminary exploration, some erroneous values were identified, mainly for sightings recorded in the middle of the ocean; these instances were thus considered as outliers and removed to ensure a reliable analysis.

Species' biological information was assembled with taxonomy identifications from GBIF [2] and more composite traits data from Encyclopedia of Life [4] — an USA Federal Government-founded organization aiming to provide "free, open, multilingual, digital access to trusted information on all known species" [6]. These publicly available data were extracted through APIs, respecting the relevant usage policies [5] [3]. Over 100 traits were gathered, including both species-specific characteristics (e.g., *locomotion*, *mass*) and relational features (e.g., *prey on*, *are eaten by*). However, most traits were unreported for a given species, so the sparse dataset required massive cleaning through regular expressions and research on the biological characteristics.

3 Exploratory Data Analysis

The training set's sightings per-species distribution (fig.1) resembles a long tail, with 70 species ($14\,\%$ of the population) accounting for more than $50\,\%$ of the worldwide instances (fig.2). This imbalance can be attributed to the true species' frequencies, but also to data collection factors (e.g., ease of observing particular animals, available technology to capture them, distribution of relevant platform's user-base). The latter might especially be of evident influence for frequently sighted species like the mule deer (*Odocoileus hemionus*) in North America and the Clark's nutcracker (*Nucifraga columbiana*) in Western North America: both are abundant in the wild [1], and captivating to encounter too. Moreover, most of the datasets' records are located in North America and Australia, developed countries placing great importance on conservation efforts [15].

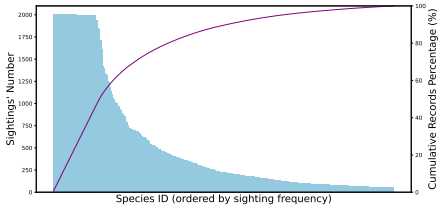


Figure 1: Distribution of sightings per species

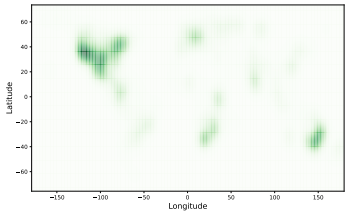


Figure 2: Sightings' geographic distribution

The location scattering of an animal's sightings is then considered, mapping each coordinates pair to a tridimensional Cartesian system — to account for Earth's curvature — and computing the variance of the records with respect to the identified centroid (fig.3). This more accurately pinpointed that most of the species are occurring globally, and just few others are isolated to narrow spots (fig.4). In particular, more than a half of the records in the training set concern animals with significantly high variances (i.e., in the order of 10^5–10^7 Km), reflecting the intrinsic spread of disparate species, such as the Common blackbird (*Turdus merula*), found in Eurasia, North Africa and Australia [8].

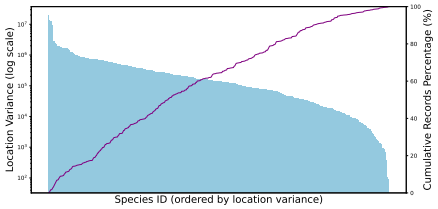


Figure 3: Location variance values per species

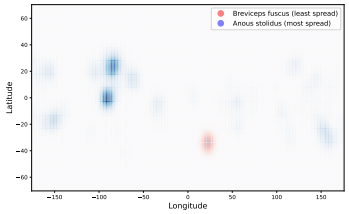


Figure 4: Two species' geographic distribution

It is noticeable that the respective bioclimatic values show similar variances within individual species too (fig.5), revealing the fundamental nature of many animals to fit in vastly dispersed habitats often of different bioclimatic type, and tempering the expectations on exact-location predicting models. Furthermore, plausibly due to the uncommon type (citizens or scientists) and diffusion of loggers animal records' distribution disparities between the location sets are clear: more than 1 in 4 species — accounting for $29.6\,\%$ of sightings — present geographic instances significantly deviating from training to test data (fig.6) (i.e., the relevant centroids are more than 2 standard deviations apart).

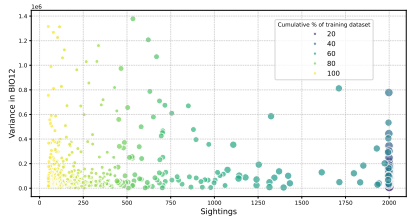


Figure 5: Per-species *annual precipitation* variance vs number of records

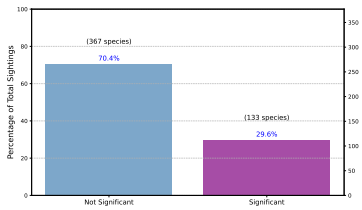


Figure 6: Test-set sightings' locations deviation from training data

These examinations led to the modeling of each species' presence by Kernel Density Estimation (KDE) [31], which provided an unbiased representation of the true population scattering, making no a priori assumptions on the overall data distribution [30]. In particular, a Gaussian kernel function with fine-tuned bandwidth was applied consistently across the species, allowing to sensibly compare their spread (e.g. the above geographic distribution plots). The Jensen-Shannon divergence [22] — a common symmetric distance metric between probability distributions — was used to compute the log-likelihood of the same sighting coordinates under each species' KDE, so that measures closer to 0 implied stronger co-occurrences patterns. The latter could be caused by either actual inter-species relations (e.g., symbiosis) or simply by a comparably fit bioclimatic enviroment, and are more deeply investigated with this additional causal acceptation in the supplementary task.

3.1 Biodiversity-based Locations Clustering

To address geographic noise, latitude and longitude were used as input features in a Decision Tree classifier (DT) [29] to predict species IDs; their entropy was used as the splitting criterion, ensuring the formation of more homogeneous groups. A hierarchical division of the records was performed, limiting the depth to 8 levels to balance granularity and representation (appx.4). This resulted in 256 geographic coordinates' groups, each minimizing the number of distinct species within. The underlying classification algorithm achieved a low validation accuracy of $22\,\%$ and an even lower testing accuracy of $4\,\%$, due to not consistently pure leaf-clusters, confirming that each (even narrow) area is populated by a range of many animals, and outlining the need for alternative location-based prediction methods focused on group-level patterns rather than individual species.

4 Modeling

4.1 Species Prediction

Species prediction given a sightings' location was initially attempted by using geographic coordinates only. In particular, the K-Nearest Neighbors algorithm (KNN) [13] — a simple and interpretable method widely used for local patterns capturing in spatial data — and a Random Forest classifier (RF) [14] — capable to handle high-dimensional data and model complex relationships — were leveraged to predict species' IDs based on latitude-longitude pairs. The original data was split into training and validation subsets (i.e., $20\,\%$ of the instances were randomly selected and kept for models' evaluation). Additionally, the models' hyper-parameters were fine-tuned based on their validation performance through Grid Search, resulting in a neighbors' number of 10 and a maximum trees' depth of 15, both considering the high species' diversity and mitigating the overfitting (appx.4). However, non-satisfactory accuracy levels were reached: $35\,\%$ with KNN and $32\,\%$ with RF, while the testing accuracy was even lower for both models (i.e., $5\,\%$ and $5\,\%$ respectively). A hypothesis t-test was thus performed on each species' predicted and actual values — assuming paired sample independence — to determine if the KNN model significantly outperformed RF; the null hypothesis was not rejected, indicating no statistically meaningful difference in accuracies.

Successively, the bioclimatic data for the records' locations was scaled to ensure uniform feature contribution and prevent dominance by features with larger ranges, and subsequently incorporated, which considerably improved both KNN's and RF's performances: a validation accuracy of $66\,\%$ and $64\,\%$ respectively was achieved, while the testing scores were of $4\,\%$ and $2\,\%$. Moreover, hypothesis testing indicated that KNN's superior performance was statistically significant in this case, likely due to its ability to effectively leverage the spatial relationships and local structure present in the combined location and bioclimatic data. These results pinpoint that environmental information can provide a richer context for each geographic coordinate, enabling models to better capture ecological factors' effects on animals' distribution; incorporating this type of data might also be virtually exploitable for deducing insights on regions where human observers are scarce. Also considering the above, in the next sections the focus would be shifted from plain geographic coordinates to their embedded bioclimatic features.

4.2 Environmental Study

With the aim of grouping areas by environment type, K-means was firstly applied directly on the bioclimatic variables' instances, but it produced unevenly scattered groups, poorly performing in high-biodiversity regions. Therefore, a grid-based approach [25] was followed instead: the world map was divided into 0.5-degree squares. The rationale was that, while individual latitude-longitude pairs were noisy the uniformly sized cells enabled a clearer analysis. The bioclimatic features of the involved coordinates were mean-aggregated for each grid-square, to represent the overall environment of these areas; the latter were then K-means-clustered [9] to obtain 20 groups of cells with

comparable bioclimatic characteristics. In the process, data was scaled, and the number of clusters was selected based on inertia and Davies-Bouldin Index (DBI) scores [18] through the elbow method. Each group actually reflected distinct environmental conditions and ecological composition (appx.2), confirmed by statistically significant differences in bioclimatic variables via Kruskal-Wallis H test [21]. Interestingly, while agglomerative clustering performed worse overall, it revealed an insight: nearby locations were grouped together initially before merging with distant regions, informing our supplementary task. An example of grid clustering in Europe is reported (fig.7): Central Europe primarily falls into *pink* cluster, characterized by a mild, stable climate with consistent rainfall; coastal areas belong to *red* one instead, featuring warmer temperatures, hot summers, and moderate precipitation seasonality. Interestingly the algorithm also captured the Alps as pertaining to the *brown* group, marked by colder climates and significant temperature fluctuations.

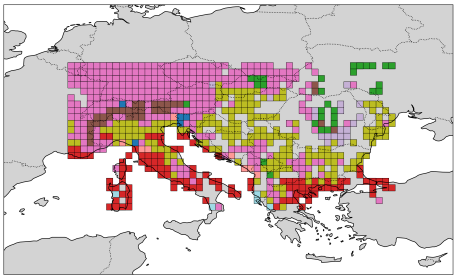


Figure 7: Central Europe environmental clusters

4.3 Locations Prediction

Using species' traits, predicting potential locations and suitable bioclimatic conditions could enhance understanding of the distribution patterns. Using both categorical and continuous features (e.g., *taxonomy class*, *mass*, *locomotion*) a set of models were built to firstly predict the environmental cluster and secondly the grid-cell a species most probably belong to. Due to data sparsity and no means of imputation for specific features, this multi-label classification was carried through LightGBM [20] and XGBoost [12] — tree-based methods that can handle NaN values, as they treat them as a separate category and progressively learn optimal splits for missing data. Since these algorithms are computationally expensive, six regions were selected based on variations in unique animal counts, environmental diversity, and area coverage, to efficiently build one model for these observations and evaluate the two algorithms (appx.3). $20\,\%$ of the training data was kept for validation, consistently for both models. The pre-processing involved binarizing multi-label columns, and one-hot encoding categorical features; labels' coherence between the training, validation, and test sets was ensured. Model parameters (e.g., regularization and learning rates) were tuned using the validation set by testing multiple values (appx.4). Notably, XGBoost required higher run time.

The accuracy of both models was evaluated for cluster and grid predictions using the six predefined regions, with grid predictions considered accurate if they fell within a 4-block radius to accommodate for noise and the large number of cells. Despite careful tuning, both the models achieved approximately $14\,\%$ cluster accuracy and $10\,\%$ grid-ID accuracy on the validation set, suggesting limited effectiveness given the data nature. The results on the test set were also mixed, with a $40\,\%$ cluster accuracy and $1\,\%$ grid accuracy, which can be explained by the already discussed variations (fig.5,fig.6). Interestingly though, it was found that running and tuning a separate model for each region improved performance, achieving for instance $20\,\%$ cluster accuracy and $18\,\%$ grid-ID accuracy for the UK region. This indicates that the overall data may be too noisy, with many species making feature interpretation challenging; focusing on specific areas and grouping observations significantly enhances prediction accuracy. Moreover, assuming normality and paired sample independence for the paired t-test, a hypothesis test showed no difference in cluster accuracy ($T=1.51$, $p=0.192$) or grid accuracy ($T=2.41$, $p=0.061$) at a predefined significance level of 0.05 for the two models; suggesting no evidence to favor either LightGBM of XGBoost.

Below is a feature importance map (fig.8) highlighting the most discriminative features for predictions using the LightGBM model. As expected, the ID is the most critical feature for distinguishing between regions, followed by the number of records and various species characteristics. To better understand the reasons behind misclassifications on the validation set, the predictions from LightGBM were analyzed (fig.9), highlighting the relation between misclassification rate, the grid count bins in which each animal appears (i.e., spread), and the unique species ID counts (i.e., diversity) per grid bin — normalized to address imbalances. A positive relationship is observed between misclassification rate and the number of grid counts an animal appears in, suggesting as expected that more-spread animals are harder to predict. Additionally, misclassification rates are higher in exceptionally biodiverse regions with over 140 species (21+ bin), likely due to the increased com-

plexity and overlap of animals in such areas. Notably, most of the top-20 misclassified animals in terms of location are mostly marked as *Least Concern* by IUCN [17], with some exceptions like the *Near Threatened* Red-headed Woodpecker, highlighting the need for better mapping of such species. Also, the most misclassified areas include the Eastern United States, Southern Africa, and Canada, suggesting prioritizing getting users to note animals there.

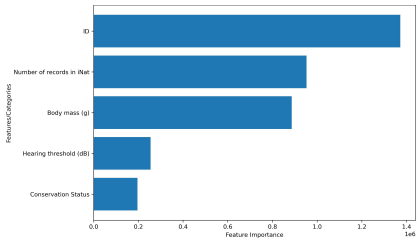


Figure 8: Features' importance for LightGBM

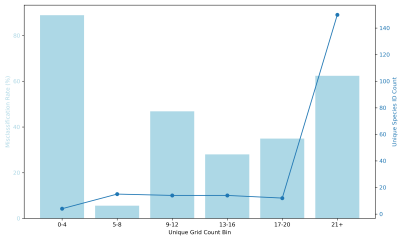


Figure 9: LightGBM misclassification analysis

5 Ecological Analysis (supplementary task)

To enhance biological insight, an ecosystem visualization method was developed, especially useful for researchers to sensibly select and verify targeted conservation ventures for a given species. In particular, the animals' interaction data from EOL [7] — covering 18 types of relations (e.g. *are eaten by*, *prey on*, *are parasitized by*) for the 500 species analyzed — was used to build networks of biological entities, sometimes not even included but connected with animals in the initial dataset (e.g., other species, viruses, plants). A framework of all beings was initially created, then individual ecosystems were plotted: sub-graphs representing data within a specific habitat, conceived as a set of grid-cells of the same environmental type grouped in a geographically cohesive area via DBSCAN [28] — able of handling noise and detect clusters of arbitrary shapes (appx.5). The most crucial interaction is displayed between each entities-pair; nodes are color-coded by taxonomy class and sized by record count in the area. Species among the original 500 with no sightings in the habitat are outlined, the other biological elements are labeled by scientific name only.

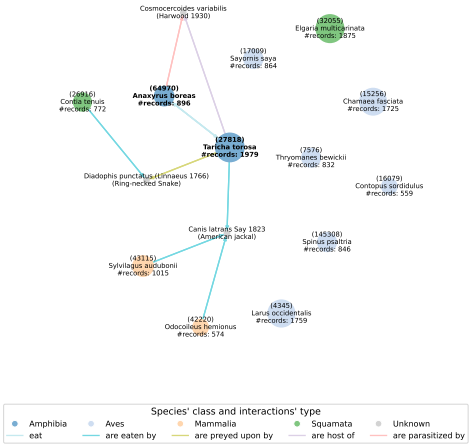


Figure 10: Example of ecosystem core shared by a couple of highly co-occurring species

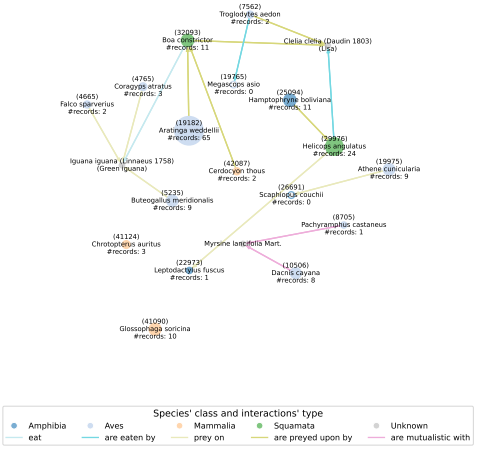


Figure 11: Most populated Amazon rainforest ecosystem's connected core (in Bolivia)

These networks are more sparse and disconnected compared to actual natural ecosystems, since relational data for the 500 studied species was considered only: further entities not among the main animals' interaction targets are not included, thus only weak paths of a single additional species are displayable. However, interesting patterns are highlighted, and a broader perspective on our previous results granted. Causal meaning was given to some species' locations correlation too, considering their roles in shared habitats. For instance, based on the above network (fig.10), it is attributable to dietary reasons that the Western toad (*Anaxyrus boreas*) and the California newt (*Taricha torosa*) ranked in the top 5 most co-occurring species — divergence score of 0.35 as defined in EDA.

Moreover, highly connected nodes representing species with no sightings or not among the main 500 could suggest that these animals are present in the area even if undetected (e.g., Green iguana in fig.11) or that they would fit there for preservation purpose, depending on their connections' quantity and type. Especially if drawn through a more complete framework, these hypotheses could also be tested by platforms like iNaturalist by incentivizing data collection for the potentially unobserved species in its suitable habitat (e.g., via gamification). This graph-based approach is in general of use in modern ecology, as simplified population models alone have demonstrated fallacies in capturing real ecosystems' complexity, particularly when active sustainability interventions are at stake [24].

6 Conclusions

In this paper, we examined biodiversity mapping based on species' occurrence data as recorded by sightseeing platforms; our main task was to develop models to predict a set of possible species given a location on the globe, and to define the areas occupied by each species. After an exploratory analysis of the geographic distribution information, we represented the documented presence of species through KDE, visualizing the probability distribution for the various species to be observed worldwide. This technique was used to identify patterns of species' co-occurrence, and prompted Decision-Tree-based clustering by frequented areas. Early results highlighted the need for further information in order to reach a satisfactory level of species' mapping accuracy.

To better understand biodiversity patterns, we incorporated environmental data, dividing the worldwide coordinates into regions based on bioclimatic characteristics; grid cells were first clustered using K-means, then grouped into geographically cohesive habitats with DBSCAN. K-Nearest Neighbors and Random Forest models were fine-tuned to predict species presence, showing significant performance improvements when bioclimatic features were included. For the more complex task of assigning species to geographic areas, LightGBM and XGBoost models were trained on six selected regions, leveraging the incorporated sparse species' traits data. As discussed above, the demonstrably broad species distribution made this task inherently challenging. However, it was shown that building region-specific models proved beneficial for enhancing predictions.

Our findings highlight the uneven distribution of biodiversity data, shaped by species' natural occurrences and their likelihood of being recorded by platform users. While ordinary methods achieved moderate success in predicting the most observed species near a location, plain species-to-location models showed low accuracy when not evaluated in specific regions. This aligns with the observation that animals are not strictly tied to narrow areas but are instead associated with favorable environmental conditions and co-occurring species. Predicting a species' exact location reduces biodiversity mapping to estimating where sightings were recorded, which is insufficient for understanding species' distribution. A more effective approach involves embedding their biological traits and comparing them to habitats, testing the fit of environmental and ecological factors.

Finally, considering the complex systems that the original dataset qualitatively represents, we delved into an ecosystem-oriented study, leveraging the obtained location- and environment-based habitats determination together with the interaction properties of each species, to address shortages and enhance a causal understanding of our other results. This network-based ecological analysis provided an additional layer of insight, linking species' co-occurrences to biological relations and suggesting potential methods to obtain better ecological knowledge, to be used as citizen science platforms' strategies to gamify collection efforts targeted at verifying predictions and filling information gaps.

6.1 Future Research

Future work could focus on developing more reliable models based on our framework, such as incorporating negative data (i.e., generated data where a species was not recorded) and leveraging MaxEnt modeling for improved results [23]. This approach could help refine species' distribution predictions by capturing the full spectrum of environmental conditions which influence each species' presence/absence. Additionally, collaborating with experts to expand the species' dataset with characteristics — and even actual animal populations' counts for specific regions from higher-quality scientific sources (e.g., natural parks) — could enable a deeper understanding and more accurate mapping of animals to their respective habitats, helping to capture the ecosystems' complexity which was a limitation in our case for both the locations prediction section and the supplementary task. Advanced machine learning techniques (e.g., deep learning and ensemble methods) could be tested as powerful tools for informing conservation policies, identifying priority areas, and assessing the impacts of ecological interventions and socio-political factors.

References

[1] URL https://www.birdguides.com/species-guide/ioc/nucifraga-columbiana/.

[2] . URL https://www.gbif.org.

[3] . URL https://techdocs.gbif.org/.

[4] 2018. URL https://www.eol.org/docs/what-is-eol.

[5] 2018. URL https://eol.org/docs/what-is-eol/terms-of-use-for-eol-application-programming-interfaces.

[6] 2024. URL https://www.citizenscience.gov/catalog/174/#.

[7] 2024. URL https://www.eol.org/docs/discover/food-webs.

[8] 2024. URL https://ebird.org/species/eurbla?siteLanguage=en_GB.

[9] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

[10] Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 329–348, 2021.

[11] Corey T Callaghan, Alistair GB Poore, Thomas Mesaglio, Angela T Moles, Shinichi Nakagawa, Christopher Roberts, Jodi JL Rowley, Adriana VergÉs, John H Wilshire, and William K Cornwell. Three frontiers for the future of biodiversity research using citizen science data. *BioScience*, 71(1):55–63, 2021.

[12] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[13] Gherardo Chirici, Matteo Mura, Daniel McInerney, Nicolas Py, Erkki O. Tomppo, Lars T. Waser, Davide Travaglini, and Ronald E. McRoberts. A meta-analysis and review of the literature on the k-nearest neighbors technique for forestry applications that use remotely sensed data. *Remote Sensing of Environment*, 176:282–294, 2016. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2016.02.001. URL https://www.sciencedirect.com/science/article/pii/S0034425716300293.

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

[15] Hideyuki Doi and Teruhiko Takahara. Global patterns of conservation research importance in different countries of the world. *PeerJ*, 4, Jul 2016. doi: 10.7717/peerj.2173.

[16] Stephen E Fick and Robert J Hijmans. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315, 2017. doi: 10.1002/joc.5086.

[17] International Union for Conservation of Nature. Red list of threatened species 2021: Summary statistics, 2021. URL https://www.iucnredlist.org/resources/summary-statistics. Accessed: 2024-11-17.

[18] GeeksforGeeks. Clustering metrics. https://www.geeksforgeeks.org/clustering-metrics/, 2024. Accessed: 2024-11-16.

[19] iNaturalist Contributors. inaturalist: An online social network for sharing biodiversity information, 2024. URL https://www.inaturalist.org. Accessed: 2024-11-18.

[20] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[21] Thomas W MacFarland, Jan M Yates, Thomas W MacFarland, and Jan M Yates. Kruskal–wallis h-test for oneway analysis of variance (anova) by ranks. *Introduction to nonparametric statistics for the biological sciences using R*, pages 177–211, 2016.

[22] M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997. ISSN 0016-0032. doi: https://doi.org/10.1016/S0016-0032(96)00063-4. URL `https://www.sciencedirect.com/science/article/pii/S0016003296000634`.

[23] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006.

[24] Pierre Quévreux, Ulrich Brose, Nuria Galiana, Anton Potapov, Élisa Thébault, Morgane Travers-Trolet, Sabine Wollrab, and Franck Jabot. Perspectives in modelling ecological interaction networks for sustainable ecosystem management. *Journal of Applied Ecology*, 61, 01 2024. doi: 10.1111/1365-2664.14584.

[25] J Andrew Royle, Richard B Chandler, Charles Yackulic, and James D Nichols. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, 3(3):545–554, 2012.

[26] Peterson AT Soberón J Sánchez-Cordero. V 1999 conservatism of ecological niches in evolutionary time. *Science*, 285:12651267.

[27] Florian D Schneider, David Fichtmueller, Martin M Gossner, Anton Güntsch, Malte Jochum, Birgitta König-Ries, Gaëtane Le Provost, Peter Manning, Andreas Ostrowski, Caterina Penone, et al. Towards an ecological trait-data standard. *Methods in Ecology and Evolution*, 10(12):2006–2019, 2019.

[28] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[29] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[30] Stanislaw Weglarczyk. Kernel density estimation and its application. *ITM Web of Conferences*, 23:00037, 11 2018. doi: 10.1051/itmconf/20182300037.

[31] Stanisław Węglarczyk. Kernel density estimation and its application. In *ITM web of conferences*, volume 23, page 00037. EDP Sciences, 2018.

Statement of contribution

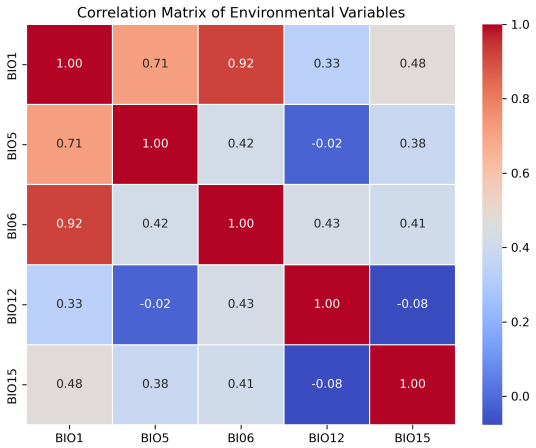
The assignment was approached as a team effort, with the successive analysis steps being outlined and the obtained results shared via weekly meetings and continuous communication. The workload was distributed equally, with new tasks being assigned to either a single person or a couple of members. A specification of each individual's focus is provided below (* indicates works in pair).

s2704516: bioclimatic data extraction, environmental study, grid methodology and location predictions*, exploring misclassifications, report writing and editing;
s2742860: cleaning EOL data, Biodiversity-based locations clustering, Species prediction, grid methodology and location predictions*, report writing and editing;
s2748897: KDE visualizations*, EOL data extraction*, DBSCAN habitat clustering, ecosystem networks building and visualization, ecological analysis, conclusions draft, code and data maintenance in a functional unified repository;
s2750316: KDE visualizations*, EOL data extraction*, species level EDA, variance based analysis, co-occuring species analysis, report writing and editing.

Generative AI usage

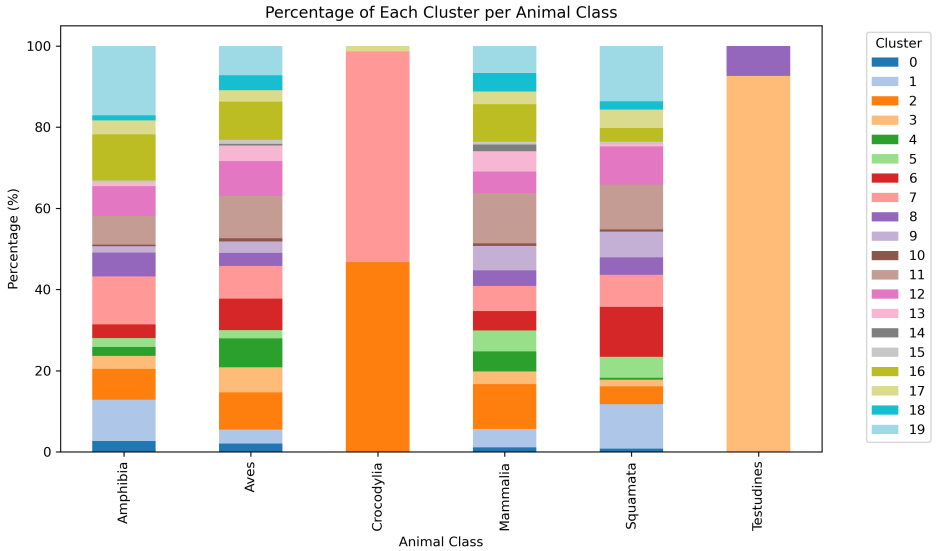
The authors confirm that generative AI was not used in this project. All analyses, models, and written content were independently developed to ensure originality and accuracy.

Appendix 1



Code	Description
BIO1	Annual Mean Temp
BIO5	Max Temp Warmest Month
BIO6	Min Temp Coldest Month
BIO12	Annual Precipitation
BIO15	Precipitation Seasonality

Appendix 2



Appendix 3

Region	Unique Species IDs	Unique Cluster IDs	Area (km^2)
Amazon Rainforest	58	12	4,245,750.00
Eastern United States	60	7	3,774,000.00
United Kingdom	20	3	1,132,200.00
Southern Africa	93	7	2,122,875.00
Central Europe	29	8	1,018,980.00
Northern Canada and Alaska	34	8	15,096,000.00

Table 1: Summary of Regions and Their Characteristics

Appendix 4

Model	Parameters

K-Nearest Neighbors	`n_neighbors=10`
	`weights='uniform'`
	`metric='euclidean'`

Random Forest	`n_estimators=100`
	`criterion='criterion'`
	`max_depth=14`
	`random_state=42`

Decision Tree	`max_depth=8`
	`random_state=42`
	`criterion='entropy'`

LightGBM	`objective='multiclass'`
	`metric='multi_logloss'`
	`boosting_type='gbdt'`
	`n_estimators=500`
	`learning_rate=0.05`
	`max_depth=5`
	`min_data_in_leaf=20`
	`lambda_l1=0.1`
	`lambda_l2=0.1`
	`subsample=0.8`
	`colsample_bytree=0.8`
	`verbose=-1`

XGBoost	`objective='multi:softprob'`
	`eval_metric='mlogloss'`
	`use_label_encoder=False`
	`n_estimators=1500`
	`learning_rate=0.01`
	`max_depth=8`
	`min_child_weight=4`
	`gamma=1`
	`subsample=0.8`
	`colsample_bytree=0.6`
	`colsample_bylevel=0.7`
	`reg_alpha=0.8`
	`reg_lambda=2.5`
	`verbosity=0`

Table 2: Parameters of Models

Appendix 5

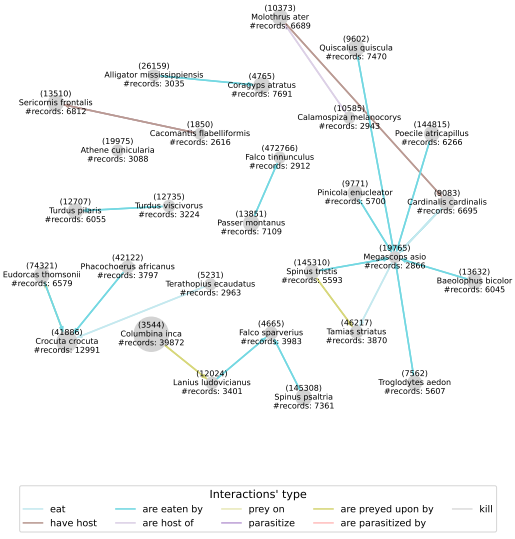


Figure 12: Most recorded species' interactions

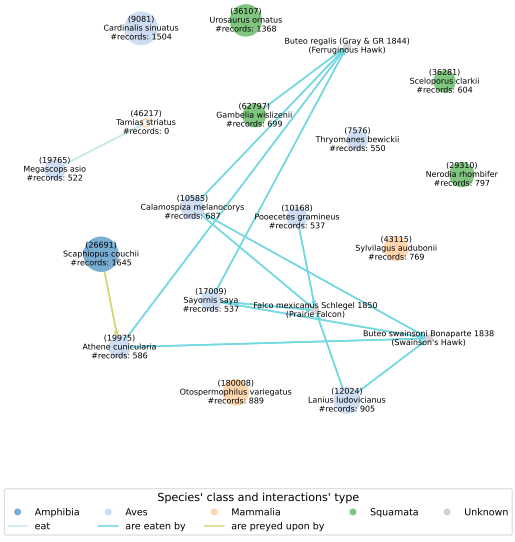


Figure 13: Largest habitat's ecosystem core