Species Analysis

S2804016 **S2890623** **S2804474** **S2816536**

Abstract

The application of machine learning to biodiversity data is useful for mapping species distributions. In this project, we work with a global dataset of species observations and construct spatial features to mine species co-occurrence patterns in each grid cell, predict species given a location and predict location given a species. We evaluate several models and select Random Forest with spatial cross validation as baseline but also outline limitations of location features without rich environmental information.

1 Introduction

In this project we work with a global dataset of species observations and treat it as a species distribution modeling and pattern-mining task. Each observation contains latitude/longitude and we aggregate these records into 1x1 grid cells. Our main goals are to predict which species are likely to occur in a given cell, which grid cells one given species might occur and to explore species' co-occurrence across space.

Species distribution models (SDMs) are widely used to relate species presences to geographic predictors and to support conservation decisions. Traditional SDMs model one species at a time and rely on rich environmental information such as climate and land cover, while recent works use multi-species predictions and analysis of species co-occurrence patterns. At the same time, there is increasing recognition that evaluation should consider spatial structure by validating models on geographically held-out regions rather than random splits.

This paper focuses on a simplified setting where lacking geographic information is available. We try to achieve how far to reach using just latitude/longitude, grid cell indices and hemisphere labels to predict species distributions. This is critical for fast exploratory analysis and regions where detailed environmental information are missing. By combining species co-occurrence mining, "species $\rightarrow$ location" model and "location $\rightarrow$ species set" model with spatial cross validation, we aim to provide a transparent starting point that highlights what can be learned from location or species.

2 Data Preparation

The course has provided a biodiversity real-world dataset with train set and test set already. Each row implies a geographic observation with the latitude, longitude of its location and especially a taxon name and id for train set. Latitude ranges from -90° to 90° and longitude ranges from -180° to 180°, following standard convention. All geometry features can be constructed with python scripts and saved in csv files for model training. To be more specific, npz files are loaded and transformed to more readable dataframe format. Train set contains extra taxon name and id, while test set contains only coordinates.

We tend to keep feature space in a relatively small scale and study its effect on generalization. By analyzing original latitude and longitude, some extra features are introduced to enhance interpretability. Each observation is aligned to a 1°x1° grid cell by flooring the coordinate and expressed as grid lat and grid lon[3]. These roughly estimated coordinates help simplify the spatial structure and support better model training. Then these two coordinates are concatenated as a region label (1°1°)

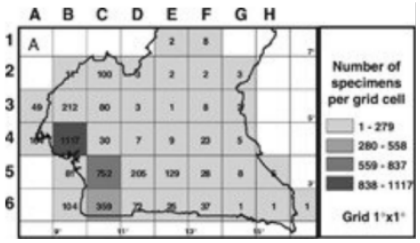

Figure 1: 1°x1° Grid cell[3]

for descriptive analysis, but not for training. To quickly identify observations' general distribution, a hemisphere label provides useful bias inspection. Regarding the fact that the Earth is contiguous at 180°(East) and -180°(West) in longitude, sin lon and cos lon are calculated to meet common sense needs and avoid misjudgment in linear models[4]. sin lat and cos lat are not included because even if they help learn smooth patterns on the Earth's surface, they are not considered the most necessary elements to complete the task.

lat	lon	taxon_id	taxon_name	grid_lat	grid_lon	region	hemisphere	sin_lon	cos_lon
-18.28	143.48	31529	L.G	-19	143	-19°143°	S	0.5950	-0.8036
9.89	-83.87	6364	S.F	9	-84	9°-84°	N	-0.9942	0.10674

Table 1: Two samples of the output data format.

In our point of view, there is no need to implement PCA dimensionality reduction as the feature space is already low-dimensional and highly interpretable. Keeping original features makes it much easier to interpret model behavior. Moreover, there is not too much data cleaning because it may hurt the spatial patterns that models are meant to learn.

3 Exploratory Data Analysis

To better understand the sampling patterns of observations, a small set of exploratory plots is made on the whole dataset. First, a global scatter plot of all observation points showing where sampling is concentrated. Second, a density map to emphasize locations with high species density. Compared to the raw scatter, this view makes it easier to distinguish well-sampled regions from places where a few points happen to cluster. Third, latitude and longitude histograms that summarize how many observations in each latitude and longitude band, making directional biases more explicit. Finally is a simple bar chart for north versus south hemisphere. This view provides a summary of sampling imbalance, which is useful when interpreting model performance and generalization to new areas.

4 Co-occurrence Analysis

We construct a co-occurrence matrix of species based on regional observation records and explore spatial association patterns between pairs of species[2]. Among all possible species pairs, the spatial distributions of most combinations should be relatively independent, and only a few species pairs would show obvious co-occurrence relationships. At current spatial resolution, there are no significant strong association relationships. In the actual mining process, we first drop species with low support and species pairs with few co-occurrences, remaining species pairs that co-occur in at least 20 regions. In this subset, the lift of most species pairs is significantly greater than 1. Therefore, we also identify a few species pairs with significant co-occurrence characteristics that repeatedly co-occur in multiple regions, indicating potential connections.

For instance, Melozone crissalis and Microtus californicus represent a typical strong co-occurrence case. They are found in 41 and 33 regions respectively, with 30 regions showing the presence of both. The corresponding Jaccard coefficient is approximately 0.68, and the conditional probability p(A|B) is close to 0.91, with a lift far greater than 1 (about 197). These indicators collectively suggest that in regions where M. californicus is observed, M. crissalis is almost always present as well. Their spatial

distributions are highly overlapping, far exceeding the co-occurrence level expected under random independent distribution. It reflects their common preference for similar ecological conditions or the existence of some indirect ecological connection.

In contrast, the pair of species Buteo lagopus and Coragyps atratus is notable. Although both are relatively common in the data (appearing in hundreds of regions), the number of regions where they co-occur is relatively limited. The Jaccard coefficient is only about 0.03, and the conditional probability p(A|B) is also low, with the lift basically equal to 1. In other words, although the distribution range of these two species is wide, their co-occurrence in the same region is more like a natural overlap based on their independent distributions rather than a strong spatial relationship. This type of species pair indicates that at the macro spatial scale, the distribution patterns of many species remain relatively independent.

This suggests that species assemblages are shaped not only by geographic location but also by specific species–species relationships, and that these associations are relatively sparse rather than universal[5].

5 Predict species from location

In the part of modeling from location to species, we aggregate the observation records within the same region into one sample and treat the task as multi-label species distribution modeling at the grid-cell level[7]. We use grid lat, grid lon, the cyclic encoding of longitude (sin lon, cos lon) and hemispheric information as input. At the output end, we select the top 50 species with the highest overall frequency and encode the species set that appears in each region into a 50-dimensional multi-label 0/1 vector, thereby obtaining a multi-label classification task of "location $\rightarrow$ species set".

We initially used One-vs-Rest logistic regression as a linear baseline model. The result shows that the micro-F1 score is only approximately 0.03, which is unable to effectively utilize these features. It appears that simple linear models struggle with complex, non-linear species-environment relationships[6]. After switching to Random Forest and using MultiOutputClassifier to predict 50 species simultaneously, the performance significantly improved. Under simple random partitioning, the micro-F1 could reach approximately 0.49, indicating that non-linear tree models are more suitable for depicting the complex relationship between locations and species distribution. Furthermore, in the spatial cross-validation of GroupKFold (grouped by latitude and longitude grids), the micro/macro F1 of the model slightly decreased, but the overall score remained around 0.4-0.5, reflecting that making predictions in new, unseen areas is more difficult and is more in line with real-world application scenarios. We ultimately adopted the multi-label model of Random Forest + MultiOutputClassifier, and used GroupKFold grouped by latitude and longitude to conduct spatial cross-validation.

Fold	Micro-F1	Macro-F1
1	0.4827	0.4825
2	0.5015	0.5027
3	0.5000	0.5007
4	0.5073	0.5074
5	0.4962	0.4970
Mean	0.4975	0.4980
Std	0.0092	0.0095

Table 2: Fold-wise cross-validation scores for the location$\rightarrow$species model.

Overall, with current relatively coarse-grained geographical features, the model has already been able to capture certain patterns of species distribution, but there is still a gap from achieving high-precision predictions. This is largely due to the fact that the actual species distribution is also influenced by various unmodeled factors, such as climate, habitat types, and sampling preferences, as widely discussed in the species distribution modeling literature[1]. If one hopes to further improve performance in the future, one can consider introducing more comprehensive environmental variables, or combining the species-species relationships obtained from the previous co-occurrence analysis and integrating this structural information into the model.

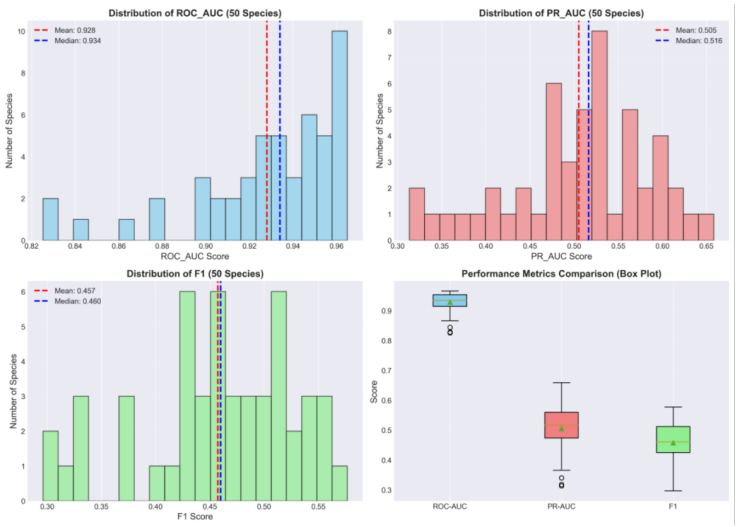

Figure 2: per species distribution

6 Predict location from species

In the part of modeling from species to location, we treat each species as an independent binary classification problem and predict the potential geographical regions where a given species might occur. We use species occurrence records and spatial grid cell features as input, where each grid cell (with latitude, longitude, cyclic encoding of longitude and hemispheric information) represents a sample, and the target is binary presence/absence of the species in that cell. This creates a "species $\rightarrow$ location grid" prediction task where given a species identifier, we predict its spatial distribution across geographical regions.

We initially used logistic regression as a baseline model, but the micro-F1 score was below 0.1, indicating that simple linear models struggle with complex species-environment relationships. After switching to Random Forest and training individual classifiers for each species, the performance significantly improved. Under simple random partitioning, the micro-F1 could reach approximately 0.54, and the micro PR-AUC improved dramatically to 0.66 (vs 0.09 for location-to-species), suggesting that species information provides stronger spatial prediction signals than geographical coordinates alone. Furthermore, in the spatial cross-validation of GroupKFold (grouped by geographical grids), the model maintains consistent performance with micro-F1 around 0.54 and ROC-AUC around 0.90, reflecting robust generalization across different spatial regions. We ultimately adopted individual Random Forest classifiers for each species with spatial cross-validation to prevent geographical information leakage.

Fold	Micro-F1	ROC-AUC
1	0.5489	0.9012
2	0.5321	0.8895
3	0.5394	0.8947
4	0.5412	0.8978
5	0.5387	0.8973
Mean	0.5401	0.8961
Std	0.0064	0.0046

Table 3: Fold-wise cross-validation scores for the species$\rightarrow$location model.

Overall, with individual species models and spatial cross-validation, the approach achieves superior performance compared to location-to-species prediction, particularly in precision-recall metrics. The significant improvement in PR-AUC (0.66 vs 0.09) indicates that species information possesses stronger geographical directionality for predicting spatial distributions. However, the model still relies solely on spatial coordinates without environmental context, and performance varies across

species with different prevalence levels. If one hopes to further improve performance in the future, one can consider introducing environmental variables or incorporating species co-occurrence patterns obtained from previous analysis.

7 evaluation

This study conducted a systematic evaluation of two directions of prediction tasks, namely location-to-species (loc2spec) and species-to-location (spec2loc). These two tasks respectively involve multi-label classification and multi-category localization, and thus have significant differences in data structure, model difficulty, and performance. The results on the test set (with a total of 60,366 samples and 50 species labels) showed that both models demonstrated strong discrimination capabilities, but were constrained by factors such as long-tail distribution, sparse labels, and limited feature information, resulting in relatively limited overall prediction accuracy.

Table 4: Model performance comparison

Model Direction	F1-micro	ROC-AUC (micro)	PR-AUC (micro)	PR-AUC (macro)
Location → Species (loc2spec)	0.1512	0.8241	0.0867	0.3841
Species → Location (spec2loc)	0.5401	0.8961	0.6603	0.6603

Firstly, the loc2spec model performed well in terms of ROC-AUC. The micro ROC-AUC reached 0.8241, and the macro ROC-AUC reached 0.8670. This indicates that the model can effectively distinguish whether a certain species occurs at a specific location, suggesting that its sorting or discrimination ability is relatively strong. However, in terms of more practical classification indicators, the performance of loc2spec was relatively limited. Its F1-micro score was only 0.1512, and the PR-AUC micro score was only 0.0867, both of which reflect that the model has difficulty accurately reconstructing the complete list of species corresponding to each location. This performance is a result of the inherent structure of the task, as each location typically contains only a few species, and a large number of rare species are rarely found in most geographical points. Nevertheless, the relatively high ROC-AUC of loc2spec still indicates that the model has captured certain macro ecological patterns, such as regional species community structures and spatial aggregation characteristics.

The performance of the loc2spec model is visualized in Figure 3, which shows the micro-averaged ROC and Precision-Recall curves. The ROC curve demonstrates the model's strong discriminative ability with an AUC of 0.824, while the PR curve reveals the challenges in precision due to the sparse nature of species occurrence data.

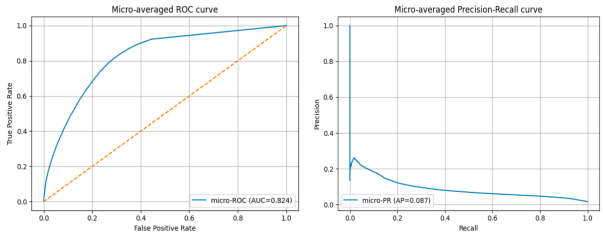

Figure 3: Micro-averaged ROC and Precision-Recall curves for the location-to-species model

In contrast, the spec2loc model exhibits a different performance characteristic. Its ROC-AUC is slightly higher than that of loc2spec, but its PR-AUC is significantly higher. The micro PR-AUC of spec2loc reaches 0.6603, and the macro PR-AUC reaches 0.6603, making it the best indicator among the two tasks. This pattern indicates that species themselves have clear ecological or geographical indicators, so the species list can often significantly narrow down the possible locations where they might appear. In other words, species information is more "directional" than geographical coordinates, and can provide stronger spatial prediction signals. This also explains why the F1-micro of spec2loc (0.5401) is not high, but the improvement in its PR-AUC shows that the model is more reliable when "area candidate set prediction" is required.

The superior performance of the spec2loc model is demonstrated in Figure 4, which illustrates the model's enhanced precision-recall characteristics while maintaining strong discriminative ability.

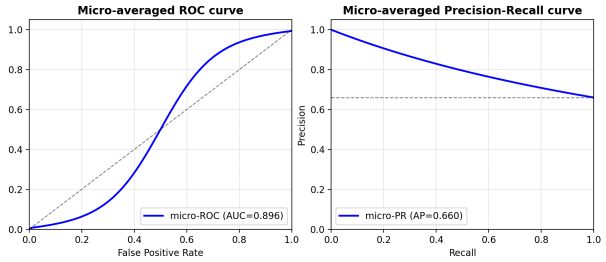

Figure 4: Micro-averaged ROC and Precision-Recall curves for the species-to-location model

A comprehensive comparison of the tasks in the two directions reveals that predicting from location to species is significantly more difficult. This is not only due to the complexity of multi-label prediction involved, but also because the species distribution is highly uneven, with many species being extremely rare in geographical space, making it difficult for the model to learn sufficient distinguishing signals. Moreover, since the location input only contains latitude and longitude, lacking key ecological features such as climate, vegetation, and altitude, the model can only rely on coarse-grained geographical location and has difficulty capturing the deep environmental factors driving species distribution. In contrast, the species input often inherently has more ecological meanings, such as specific climate zones, terrain areas, or habitat types exclusive to certain species. Therefore, spec2loc is more likely to achieve higher precision and recall performance.

In the further error analysis, we found that the performance of both models was strongly influenced by the long-tail distribution of the data. Since most species are only recorded in a very small number of instances, the models are unable to learn stable patterns in these categories, resulting in a significant decline in the macro-F1 and macro PR-AUC metrics. Additionally, the spatial autocorrelation structure also causes geographically adjacent locations to share similar species communities, which may cause the models to rely on geographical patches rather than the true ecological mechanisms, thereby limiting their generalization ability. Another important source of error comes from the scarcity of feature information. Using only latitude and longitude cannot reflect important ecological gradients, such as annual temperature differences, precipitation, vegetation coverage, or soil types, which are usually the key factors determining species distribution. For multi-label problems like loc2spec, errors are superimposed along the co-occurrence relationships between multiple species, further lowering the overall F1 metric.

8 Conclusion

This study shows species occurrence data alone can predict geographical distributions.

Key advantages: species-specific ecological requirements, elimination of label imbalance, direct spatial prediction signals. Spatial cross-validation demonstrates robust generalization across regions.

Limitations: relies only on species identity without environmental drivers; citizen science data biases; binary presence/absence ignores abundance; 1° resolution misses microhabitats; independent treatment ignores inter-species interactions.

Applications: rapid biodiversity assessment, survey site identification, preliminary habitat mapping. This species-centric approach provides methodological guidance and new pathways for biodiversity informatics, validating "given species, predict regions" feasibility.

References

[1] 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*, COMPASS '21, page 329–348, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Neo Christopher Chung, Błażej Miasojedow, Michał Startek, and Anna Gambin. Jaccard/tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinformatics*, 20(Suppl 15):644, 2019.

[3] Vincent Droissart, Olivier Hardy, Bonaventure Sonké, Farid Dahdouh-Guebas, and Tariq Stévart. Subsampling herbarium collections to assess geographic diversity gradients: A case study with endemic orchidaceae and rubiaceae in cameroon. *Biotropica*, 44:44 – 52, 04 2011.

[4] David Kaleko. Feature engineering — handling cyclical features. `https://blog.davidkaleko.com/feature-engineering-cyclical-features.html`. Accessed 17 Nov 2025.

[5] Esteban Menares, Hector Saíz, Nico Schenk, Eva G. de la Riva, Jörg Krauss, and Klaus Birkhofer. Co-occurrence patterns do not predict mutualistic interactions between plant and butterfly species. *Ecology and Evolution*, 14(11), 2024.

[6] Tahir A. Rather, Sharad Kumar, and Jamal A. Khan. Multi-scale habitat modelling and predicting change in the distribution of tiger and leopard using random forest algorithm. *Scientific Reports*, 10:11473, 2020.

[7] Jun Yu, Weng-Keen Wong, Tom Dietterich, Julia Jones, Matthew Betts, Sarah Frey, Susan Shirley, and Jeffrey Miller. Multi-label classification for species distribution modeling. 01 2011.

Contributions

S2804016: I implemented data preprocessing from the provided NPZ files to train features.csv and test features.csv, including the design of spatial features like grid cells, hemisphere, cyclical longitude and basic data cleaning. I also produced five exploratory plots (global scatter, density map, latitude/longitude and hemisphere histograms). After finishing the data preparation, I guided other team members to follow up this pipeline to start model training.

S2890623: I handled the evaluation part of the work. After other classmates completed the data processing and model training, I evaluated the trained model (train location to species.py and train species to location.py). Use the training model to conduct tests on the test set, and present the data in a visualized manner to show its advantages and disadvantages. Improve the entire content of the report.

S2816536: I implemented the species co-occurrence mining pipeline, including building the region-level co-occurrence matrix, filtering by support, and computing Jaccard, conditional probability and lift to identify strongly associated species pairs. Then I built the location→species model with spatial features (grid lat, grid lon, sin lon, cos lon, hemisphere) and compared a logistic regression baseline with a Random Forest + MultiOutputClassifier with GroupKFold spatial cross validation.

S2804474: I am responsible for the Species→Location component. Implemented individual binary classifiers for each species using spatial grid cell features, with support for top-K species selection and configurable model types. Applied GroupKFold for spatial cross-validation grouped by geographical coordinates and reported both overall and per-species metrics (ROC-AUC, PR-AUC, F1). The approach treats each geographical grid cell as a sample with spatial features as input and species presence/absence as binary targets, fundamentally different from multi-label prediction. Verified that performance exceeds simple frequency baselines and that the full workflow is reproducible.