SIADS 696 Milestone II - Final Report

Project Name: Predicting and Clustering Bird Migration Patterns Across Americas
Team: Yangkang Chen, Dean Lawrence
GitHub repo: https://github.com/chenyangkang/MADS_Milestone2_Bird_Migration

Introduction

Background

In this project, our goal is to predict, cluster, and compare different bird migration patterns.

Throughout the years, millions of birds were observed and recorded across the world by birders at eBird citizen science project. However, our knowledge is limited regarding the detailed migration pattern of birds. On the one hand, the global climate and land-use change is threatening birds, and on the other hand the spatial-temporal nature of bird migration makes it challenging to predict movement. Therefore, there is an urgency to model fine-scale bird migration and distribution patterns across the continents, both for understanding and conservation purposes. In this project, we focused on fitting the current migration pattern, so that we could predict birds' presence or not in regions where there is no observation data available.

Goal and purpose

Through this project, we could:

1) Predict the presence or absence of a species at a location at a time, even in a region where there is no observation data available. This will provide a comprehensive graph of the whole-year-round bird distribution pattern.

2) By taking spatial-temporal connectivity into account, we could depict the migration route of certain species, which is significant in understanding the biological process and conservation.

3) The fitted model could further be used to project potential occurrence change under future land-use and climate scenarios to better quantify the impact of global change on migratory birds.

Methods & results summary

For supervised learning, we compared the metrics for 56 different model selection, including five baseline models and one advanced ensemble model. We showed that the gridding-ensemble method (here after, AdaSTEM model) is generally much better than single baseline models. Prediction difficulty for different species is different, with House Wren having the highest AUC score of 0.8646 in AdaSTEM model, and Mallard with the lowest of 0.8615.

In the unsupervised learning part, we first calculated the geographical center of distribution for each calendar week/month, and generated a "migration route" for the species. Then we applied different clustering strategies to depict similarity and dissimilarity of species migration pattern. We found that species in closer evolutionary relationships or with similar tropical niches migrate together. For example, cluster 1 mainly consists of carnivorous predators, while cluster 2 mainly consists of omnivorous and vegetarian preys.

Related work

1) Our modeling framework will largely come from a published academic paper (Fink et al., 2020). In short, this paper use:

- Adaptive gridding method that takes data abundance into consideration. More data-abundant location allows finer gridding.
- Two-step zero inflated model to model count data.

2) Since the nature of our supervised learning task is species distribution modeling, (SDM) there is an example of SDM provided by sklearn (reference [2]): They considered the problem as density estimation and applied a one-class SVM model.

In our case, we treat the problem as a binary classification problem because we have labels for both present and absent.

3) Hubalek (2005) investigated the co-fluctuation among bird species in their migration timing. This work philosophically resembles our unsupervised learning task. Their findings are:

- All short distance migrants with the European winger range clustered together.
- long-distance migrants, who winter in the African range, formed six other smaller clusters.
- They calculated the Pearson correlation of spring migration and made a UPGMA cluster analysis.

<u>Data Source</u>

- eBird citizen science data (reference [5])
    - Time range: year 2018.
    - Originally 4,300,429 observations. 487,293 after subsampling.
    - Important variables:
        - Time of the day when observation started.
        - Date
        - Number of observers
        - Observation protocol type (stationary or traveling)
        - Location (location name and longitude, latitude)
        - Traveling distance
        - The name of each species observed and their count.
    - eBird data were pre-filtered based on following rules:
        - Observation type should be traveling or stationary.
        - Only checklists with more than 5 species observed are included.
        - The travel distance of observer should be less than 3 km (to make sure the high spatial precision and land-use continuity)
        - The observation time duration should be more than 5 minutes.
- ESA CCI global land use data (reference [7])
    - Time range: year 2018.
    - Important variables include: Fraction of land use, landscape index (maximum patch size, patch density, etc.) of different land use (urban, cropland, shrubland, forest, water, etc.)
- WorldClim monthly climate data (reference [8])
    - Time range: year 2018.
    - 19 Bioclimatic variables. For example, precipitation of the wettest quarter, temperature of the warmest month.
    - Monthly raw climate data: maximum temperature (tmax), minimum temperature (tmin), precipitation (prec).
- Elevation and slope data (reference [6])
    - Mean and standard deviation of elevation, slope, eastness, northness.

Feature engineering

1. Data extraction
- We first extracted longitude and latitude data for each eBird checklist.
- We load environmental data based on their format. For example, using the gdal package (GDAL/OGR contributors, 2022) for geo-tif format, and netCDF4 package (reference [10])for NetCDF format data.
- Based on the grid length of the environmental dataset, extract the data at corresponding locations using NumPy operation.

2. Data manipulation
- For missing values in sampling effort parameters, we use -1 to fill the parameters. This is because the most important model in our supervised learning part is tree-based, therefore robust to missing values.
- For supervised learning, we down sampled the data volume so that each 1*1*1 longitude-latitude-day_of_year grid has a maximum of 500 records. This is for releasing the burden of computational power, and also reducing the biased spatial distribution of checklists.
- For the unsupervised learning part, we calculated the geographical center of distribution each day of each species, and used these features for down-stream clustering.

3. Final features selected:
- Features consisted of five parts: 1) climate variables; 2) sampling effort variables (observation duration, number of observers, etc); 3) Temporal variables (day of year, etc); 4) topography; 5) landscape variables. Please refer to the appendix for a complete variables list.

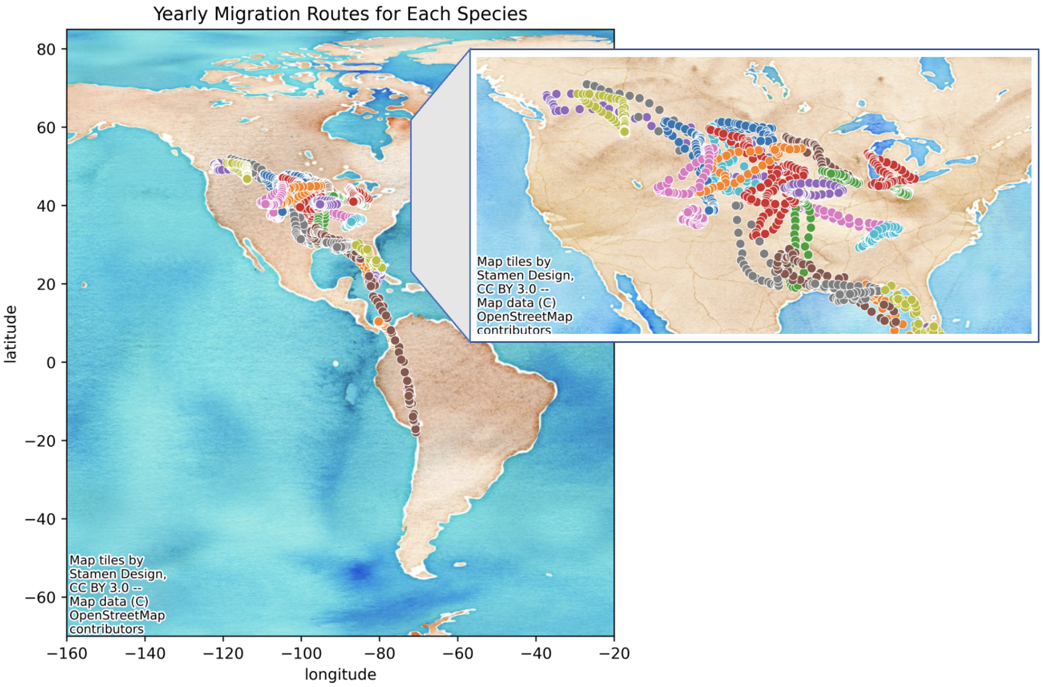
Yearly Migration Routes for Each Species
latitude
80
60
40
20
0
−20
−40
−60
longitude
−160
−140
−120
−100
−80
−60
−40
−20
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors
Map tiles by Stamen Design, CC BY 3.0 -- Map data (C) OpenStreetMap contributors

Figure 1. Visualization of species-level year-round migration route for 27 randomly chosen species. The center of geographical distribution is calculated for each week, and a 10-week-window moving average is applied to smooth the time-series.

Supervised learning

Methods description

The extracted data are ready to be fed into the machine learning workflow. We first did a 80-20 train-test-split to partition the data. While exploring the model selections, we realized that predicting bird occurrence is more than a linear problem, therefore we include only one linear model to ensure that the rest of the baseline model can be competitive to AdaSTEM model. Five basic models were then constructed as baseline models, namely LogisticRegression, GradientBoostingCliassfier, DecisionTreeClassifier, XGBClassifier and RandomForestClassifier. We chose these five models because they represent a wide range of model categories, from linear to non-liner tree-based models, and from single tree model to ensembles and optimized loss functions. Finally, we also include the AdaSTEM model, which uses a grid-train-combine strategy (Figure 2), and we choose XGBClassifier as the "base model" for each grid.

We fine-tuned the grid size hyperparameters by testing the scores for five grid size choices: (2, 5), (4, 10), (8, 20), (16, 40), (24, 60), and 10 ensemble fold choices, ranging from 1 to 10. Taking four evaluation metrics into account - average precision, roc-auc score, f1 score and cohen kappa score - we discovered that the best parameters (that is, at least have 2 metrics ranked 1st in all parameter sets), and therefore the grid size is set to 8 (minimum grid size) and 20 (maximum grid size) (Figure 5). The ensemble fold was also fine-tuned and the value 8 was chosen for it considering that all the metrics did not considerably increase with higher ensemble fold.

All variables used in the supervised learning are listed in the appendix. These include: 1) climate variables; 2) sampling effort variables (observation duration, number of observers, etc); 3) Temporal variables (day of year, etc); 4) topography; 5) landscape variables.

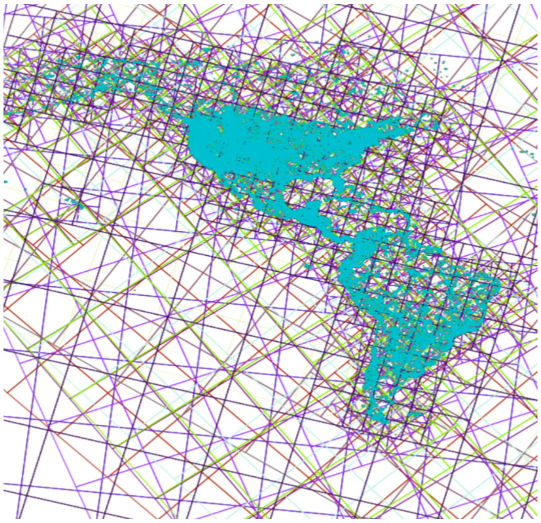

Figure 2. QuadTree implementation
The figure illustrates how the QuadTree gridding algorithm is implemented. There are 10 ensembles in this chart, presented by different line colors. Regions where data are more abundant (shown in blue dots), the grid can be more densely structured. Final result takes the average of all ensembles.

Supervised Evaluation

overall results

We chose four evaluation metrics for the hyperparameters tuning and sensitivity analysis: average precision, roc-auc score, f1 score and cohen kappa score. For comparing baseline model and AdaSTEM model, we chose to use ROC-AUC score for the simplicity and its authenticity and popularity in classifier evaluation.

As a result, the linear LogisticRegression model performed the worst and AdaSTEM performed the best in all six species (Figure 3, Table 1). The average AUC score for AdaSTEM model reached 0.8633 (std 0.001) across all six species, showing a superior and stable performance. Meanwhile the performance for Randomforest and XGBClassifier is also valuable in species like Black-capped Chickadee.

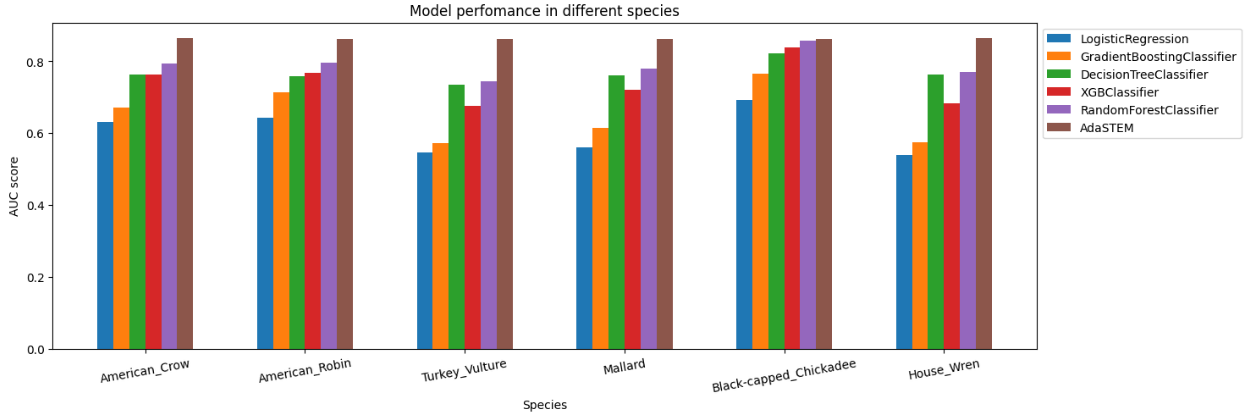

Figure 3. ROC-AUC score for 6 models in 6 different species. All models were run under 5-fold cross validation except for the AdaSTEM model. The standard deviations are too small to enable a visible error bar (see Table 1).

Table 1. ROC-AUC score for 6 models in 6 different species. Values in quotes represent the standard deviation of 5-fold cross validation value.

Species	AdaSTEM	DecisionTree	GradientBoosting	LogisticRegression	RandomForest	XGBoost
American Crow	0.864	0.764 (0.001)	0.672 (0.001)	0.63 (0.002)	0.794 (0.002)	0.764 (0.002)
American Robin	0.863	0.759 (0.001)	0.714 (0.002)	0.643 (0.002)	0.795 (0.01)	0.768 (0.002)
Black-capped Chickadee	0.862	0.821 (0.002)	0.767 (0.002)	0.693 (0.002)	0.858 (0.001)	0.838 (0.002)
House Wren	0.865	0.764 (0.002)	0.574 (0.002)	0.539 (0.002)	0.770 (0.002)	0.682 (0.002)
Mallard	0.862	0.762 (0.003)	0.614 (0.002)	0.561 (0.003)	0.780 (0.002)	0.720 (0.002)
Turkey Vulture	0.863	0.736 (0.002)	0.572 (0.002)	0.547 (0.001)	0.743 (0.002)	0.675 (0.002)

Feature importance analysis

Because of the self-defined format of AdaSTEM model, we conduct feature importance analysis by manually permutating each feature. We use AUC score as the evaluation metric, and feature importance is defined as

$$I_k = 1 - \frac{1}{5}\sum_{1}^{5} \frac{AUC_{k-permutated}}{AUC_{complete}}$$

where I_k is the importance score for the selected feature k. $AUC_{complete}$ is the AUC score on the test set with no features permutated, and $AUC_{k\text{-permutated}}$ is the score where only feature k is permutated. Permutation-evaluations were operated five times, and the average is taken.

Result shows that DOY (day of year) is significantly more important than the second-ranked feature (Figure 4). A randomized DOY feature will reduce the model performance for ~3.3% (measured in AUC). This impact is not as significant as anticipated. However, it makes sense because DOY is not the only feature to indicate seasonal change. Other variables like precipitation, min temperature, max temperature of the month, could also indicate seasons.

Besides DOY, elevation, slope, climates (those "bios") and other environmental factors play a major role in the model. The result makes sense because birds were indicated by climate and other environments to decide their occurrence and movement. Duration minutes, one of the sampling effort parameters also shows significance in the model.

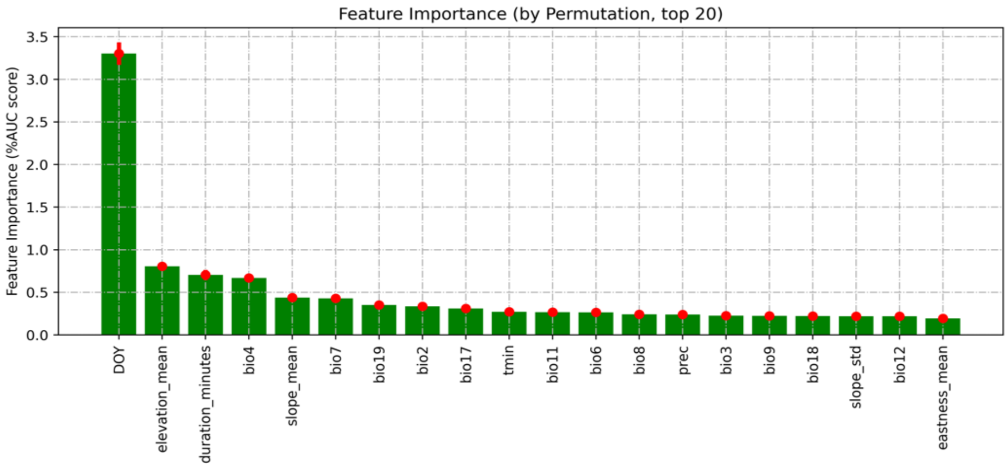

Figure 4. Feature importance calculated by permutation analysis with 5 repeats. Only top 20 features are shown. Red bars represent the standard deviation of importance across repeats.

Sensitivity & learning curve analysis

Sensitivity analysis was conducted in the same manner as hyperparameter tunings. We explore the influence of changing data volume (learning curve), ensemble fold and grid size (in the AdaSTEM model) to the model performance (Figure 5). Overall, our AdaSTEM model is less sensitive to hyperparameters or data volume for training. In the worst case of the combination, the model will still reach an AUC score of ~ 0.8, which shows the robustness of our model to training data volume or parameters.

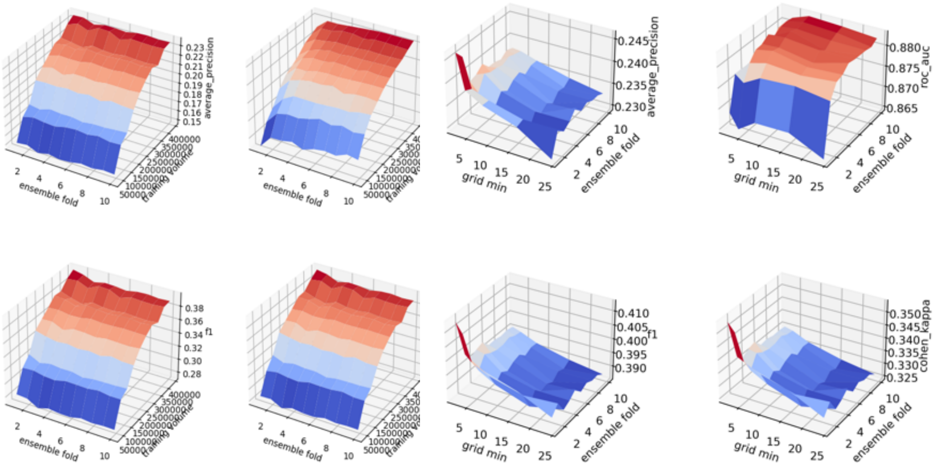

Figure 5. Sensitivity, learning curve and fine-tuning. The four plots on the left show the relationship between evaluation metrics, ensemble fold and data volume. The four plots on the right show the relationship between grid size, ensemble fold and evaluation metrics. The redder the higher scores.

Tradeoffs demonstration

There are multiple tradeoffs in the model training and evaluation stage. One example is that while increasing ensemble fold will almost definitely increase the model performance, the computational cost is also higher. We found a balance between training ensemble fold and computational power, that an 8-ensemble model performs good enough. On the other hand, higher data volume will also increase the training cost, but we suggest using the full dataset because we conclude that the score did not reach a plateau through learning curve analysis.

Failure Analysis

Because of the format of the AdaSTEM model, we conducted wrong-prediction visualization and t-test for features that belong to wrong-right prediction, instead of passing the model to the SHAP package (Lundberg & Lee, 2017). Figure 6 shows the kernel density map for correctly predicted records (green) and wrong prediction records (red). While correct predictions are distributed across the Americas continent and showing highest density around the Caribbean Sea, the wrong predictions are mainly in the US, with some high-density regions. This result is reasonable because the US has the highest data volume, and it's likely to mis-catch some records by using a single model. We further looked into the feature distribution of these two classes to see if the wrongly predicted records are outliers (Supplementary Table 2). We found statistically significant differences between these two classes, with the wrong class being lower in Bio3 (Isothermality), Bio11 (Mean Temperature of Coldest Quarter), slope, Bio9 (Mean Temperature of Driest Quarter), etc. It indicates that those wrongly predicted records are outliers in temperatures, with lower temperature overall. Besides the environment, the "wrong" class has also significantly lower "duration minutes", which means the sampling duration is lower in these records (for example, 5 min observation vs. 100 min). This is also reasonable considering the variability and occasionality in short-term observation.

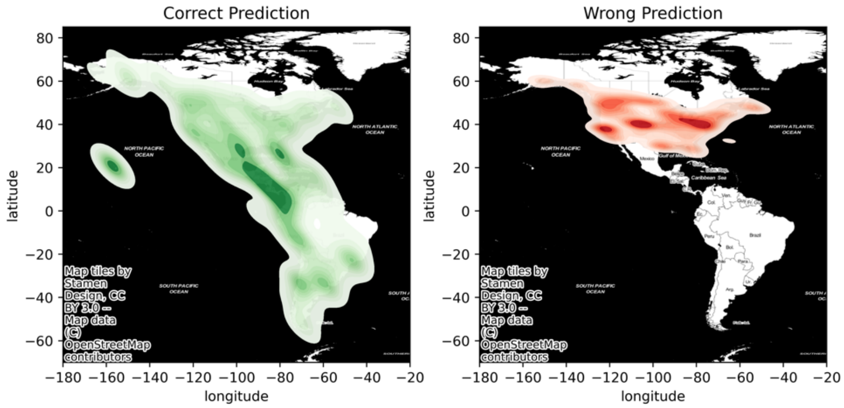

Figure 6. Kernel density map for correct and wrong prediction The green (left) density plot shows the distribution of correctly predicted records, while the red (right) one shows wrong predictions.

For specific examples, we randomly select three records that were mis-classified.

- For the first one, the failure is likely due to the missing of all climate variables. This location is probably on the sea or in a remote area where the climate data is not available. In the feature importance analysis, we showed that climate is important indicators, and thus the failure is reasonable. Further trimming could remove records where the climate is missing.
- For the second failure, the duration minutes is 12 min, which is probably quite short, resulting in undetectability of the species. This situation is expected. When doing formal prediction, we will generate prediction sets with all duration minutes standardized as 60 min, to measure "the likelihood of one expert traveling one km in 60 min and observing this bird".
- For the third case, all variables seem to be falling in reasonable range, except that there is only one land use type: cropland. One explanation for the failure is that the resolution of land use data is not enough (when zooming in, there could be other patches like shrubs and trees). This could be improved by using higher resolution data. The second explanation is that there is occasionality in observation,

especially in citizen science data. This kind of error grows from the data nature and will be hard to avoid.

Unsupervised learning

Methods description

Unsupervised learning methods were used to attempt to cluster bird species into comparable groupings. Because the eBird dataset consists of a set of individual observations of species at locations and times, the observations must be aggregated into individual species features that adequately capture both the spatial and temporal features of the distribution of observations. To do this, two different feature representations were evaluated for spatial aggregation. For both methods, a week-long time interval was used for temporal aggregation. The representations were also evaluated using a month-long time interval for temporal aggregation, but it was found that the more precise week-long interval provided a better result. There were a total of 101 species clustered.

Two clustering methods were compared to determine their effectiveness at grouping similar species. The first used was the K-means clustering method. This was selected for baselining our primary chosen method against. The second method chosen was an agglomerative method. This was chosen as the main method of grouping species due to its potential likeness to genetic species differences. The hypothesis was that species with close genetic similarity may be within close proximity within the clustering. Evaluation of the genetic difference between species was not within the scope of this project, but that was the driving interest in selecting this clustering method and may be one area of future work. The only hyperparameter varied for each of the two methods applied was the number of clusters produced as the output when comparing against the ground truth. This hyperparameter was set equal to the number of clusters in the underlying ground truth being compared against.

The first feature representation used a simpler min-mean-max geographic aggregation to determine the spatial boundaries of each species distribution at different time windows. By aggregating on a weekly temporal basis, this produced three columns for the species latitude bounds and three columns for the species longitude bounds for each week, producing a total of 312 features per species in the final representation.

The second feature representation applied an initial stage of spatial clustering to pull out regions-of-interest (ROIs) from the observation data. The DBSCAN spatial clustering algorithm was used to produce these ROIs due to its ability to differentiate clusters of variable size based on density. This output tended to produce distinct clusters around regions that have high density, such as cities. In the case of small islands, it tends to pull out the entire island as a distinct cluster. The set of clusters found are shown imposed over a section of north America for visualization purposes in Figure 7 below. The polygons were generated by calculating the convex hull over the set of observations within each cluster.

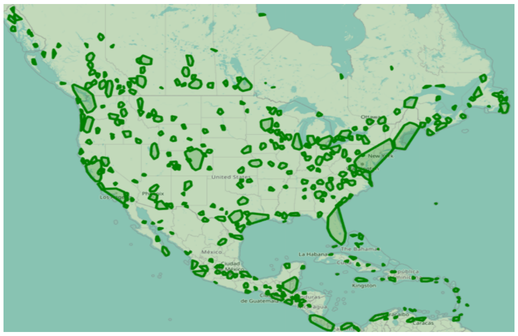

Figure 7. Regions-of-Interest

The parameters used for DBSCAN were an *epsilon* of 0.25 and a *min_samples* of 100. This clustering produced a total of 531 ROIs covering both North and South America. These parameters were hand selected to produce clusters that visually achieved a decent level of spatial granularity, while still not being overly numerous. The final feature vector consisted of zeroes and ones per species where a column represented whether a given species was present in one of the ROIs during a given week of the year. These methods produced a total of 27,612 features per species.

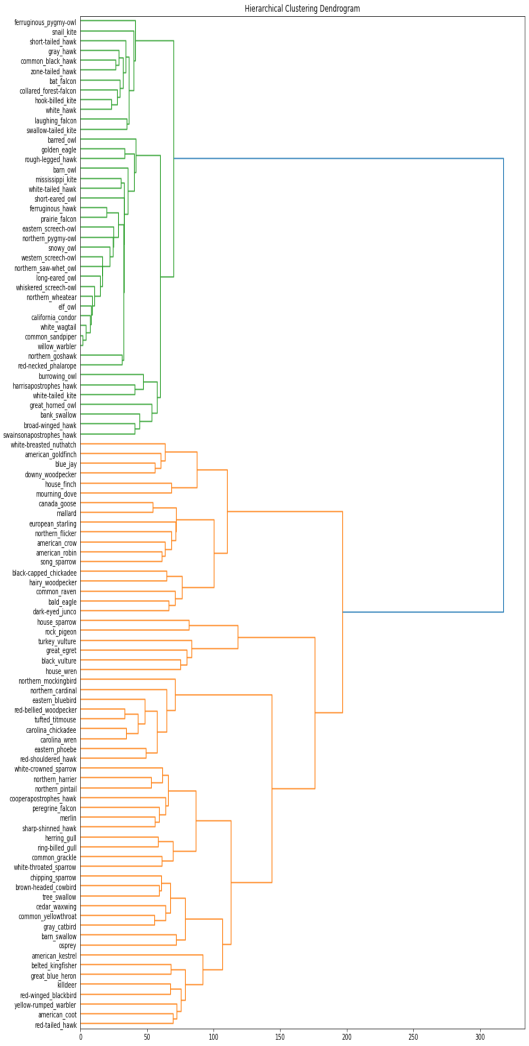

Figure 8. Hierarchical clustering based on ROI feature representation.

Figure 8 shows a dendrogram of the output of the agglomerative clustering approach based on the ROI feature representation. It can generally be found to group the species into two large clusters, one of which mostly contains raptors and other predators, while the other contains mostly herbivores and omnivores.

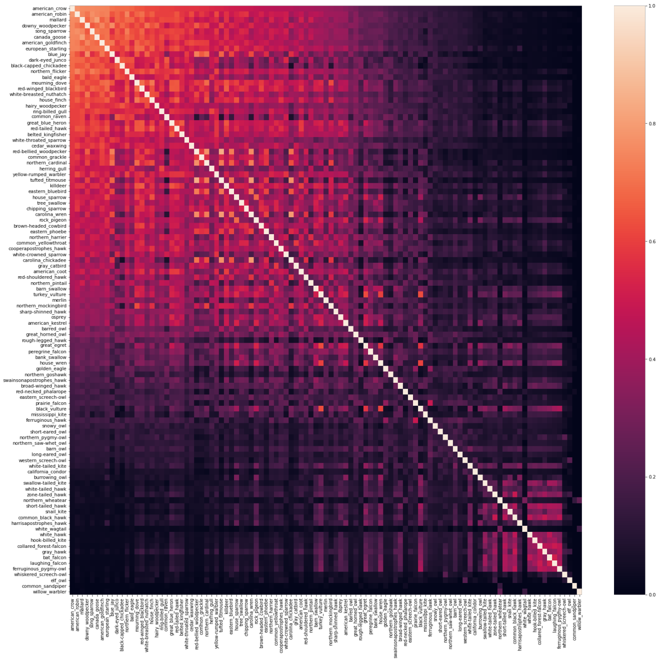

Figure 9. Species similarity based on ROI feature representation.

Figure 9 provides one other mechanism to visualize the species similarity. The heatmap shows the cosine similarity between the ROI-based features for each species. The two large clusters found above can generally be seen in this heatmap, with the upper-left corner containing a large set of similar species, and the lower-right corner containing a smaller but still similar species.

Both of these visualizations were also computed for the min-mean-max feature representation, but were omitted for brevity.

Unsupervised evaluation

Several combinations of feature representations, clustering methods, and ground truth clusterings were evaluated to determine which was the most effective set to describe similarity between species using this dataset.

Three sets of ground truth clusterings were used to evaluate the clusters. These were obtained from species taxonomy data. The first ground truth used was a simplified trophic level, which grouped birds into two groups: herbivores and omnivores against all other predators. The second ground truth used was the species order, which provided a total of 14 clusters. The third ground truth used was the species family, which provided a total of 39 clusters. Number of clusters output from each method was chosen based on the true number of clusters contained in the taxonomic category being compared to.

The adjusted mutual information score was used to compare clustering methods, feature representations, and ground truth classifications. This metric captures the percentage of information found in the clustering when compared to a ground truth that cannot be accounted for by random chance. It is invariant to the number of clusters being output by a method. This is in contrast to the normalized mutual information score, which tends to naturally increase as the number of clusters output increases. Because we were using a variable number of clusters to compare against each ground truth, we opted for the adjusted mutual information score to provide a better comparison.

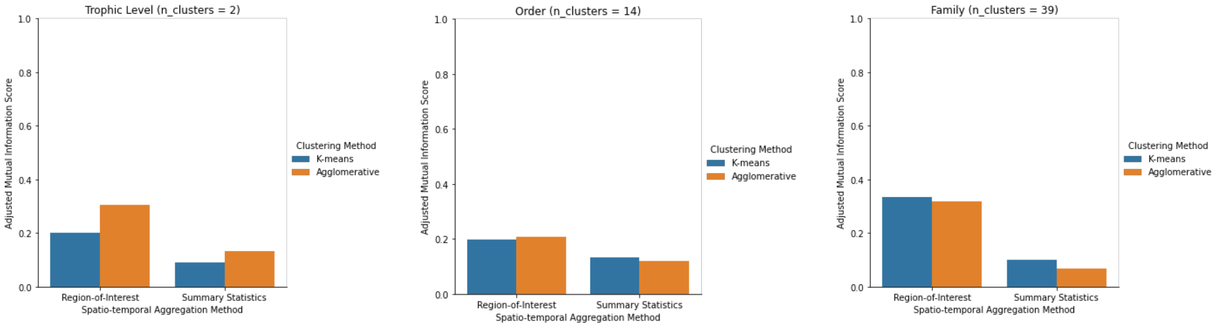

Figure 10. Clustering adjusted mutual information score visualization

Table 2. Adjusted mutual information scores

Taxonomy Category	Feature Representation	Agglomerative	K-means
Trophic Level (*n_clusters* = 2)	Min-mean-max	0.132713	0.090471
	ROI	0.304079	0.200867
Order (*n_clusters* = 14)	Min-mean-max	0.119847	0.131803
	ROI	0.209000	0.197381
Family (*n_clusters* = 39)	Min-mean-max	0.066469	0.099980
	ROI	0.319260	0.333166

Figure 10 visually shows the results of comparison between each feature representation, clustering method, and ground truth clustering. Table 2 contains the numeric values for the adjusted mutual information scores at each level. These results find that the ROI-based feature representation universally fares better than the min-mean-max geographic aggregation representation. When compared at ground truth with a lower number of clusters, the agglomerative clustering outperforms the K-means clustering. When the number of clusters contained in the ground truth increases, the K-means clustering tends to edge out the agglomerative clustering. The best performance among the combinations compared is the K-means clustering, using the ROI-based feature representation, when compared against species family as the ground truth. Although the difference between the hierarchical and K-means clustering with *n_clusters* equal to 39 and the ROI-based feature representation was marginal.

A sensitivity analysis can be done to find the effect of choices made in calculating the ROI-based feature representation. For temporal aggregation, the time period was varied between using week of the year and month of the year. For spatial aggregation, the *min_samples* parameter of DBSCAN was varied, where a smaller number produces a higher number of clusters, and a higher number filters out smaller ROIs. In other words, it raises the threshold of points necessary for something to be considered an ROI.

Table 3. Spatial-temporal aggregation sensitivity analysis

Agglomerative clustering Family ground truth (*n_clusters*=39)		Spatial aggregation (*min_samples*)		Effect
		25 (947 ROIs)	100 (531 ROIs)	
Temporal aggregation	Week	0.291706	0.319260	0.027554
	Month	0.242013	0.241964	-0.000049
Effect		-0.049693	-0.077296	-0.027603

Table 3 contains the adjusted mutual information score produced by the agglomerative clustering method and the species family ground truth when varying the time period for temporal aggregation between a week and a month and the *min_samples* parameter for spatial aggregation between 25 and 100. It was found that varying the parameter for temporal aggregation has a much more significant effect on the output of the model, whereas varying the spatial aggregation parameter has a near-zero effect, but is slightly more pronounced when using weekly temporal aggregation. A more fine grained temporal aggregation while prioritizing fewer ROIs in spatial aggregation produced the highest scoring result.

Discussion

Supervised methods

Understanding the driver of bird migration and accurately predicting migration patterns has been an essential goal in ecology. In this project, with the help of a citizen science database, remote sensing data, and data science methodologies, we successfully explored the important features in bird migration and built a delicate model to predict the whole-year-round spectacular movement. By comparing 6 models with different underlying mechanisms and philosophy, we verified that the AdaSTEM (grid-train-combine) model, which uses spatial-temporal gridding algorithm performed the best, followed by RandomForest and XGBClassifier. With an AUC score >0.8 in all the six examined species, our model shows robustness to the grid size and ensemble fold hyperparameters.

Our feature importance analysis revealed the importance of DOY (day of year), as a proxy for photoperiod, to bird migration. Besides, climate also shows a critical impact on the pattern. These findings are aligned with our prior knowledge that migration is genetically controlled by photoperiod and fine-tuned by the climate of the year. For sampling effort parameters, duration minutes is the most important one, which deserves additional attention when building a citizen science project.

One challenge we faced is the computational cost of the AdaSTEM model. It's difficult to run a 100-ensemble AdaSTEM model because it will cost 100 times more than usual, let alone doing sensitivity analysis on it. While the AdaSTEM model could be computationally costly, we figured out the best grid-size and ensemble fold hyperparameters to strike a balance between computational cost and model performance, making the model more scalable and applicable. Luckily, the AUC score of the model reached a plateau at 8-ensemble, which means we have saved 90% of the expected power compared to the 100-ensemble AdaSTEM model in the original paper. Meanwhile, by conducting failure analysis, we identified four types of potential error: 1) missing value in significant features (e.g., climate); 2) deficient sampling effort (e.g., short-term observation); 3) lack of spatial resolution in some features (e.g., 500m land use data) and 4) randomness in observational data, especially in citizen science program. Future research could head on fine-filtering missing value and increasing the quality of environmental data.

However, we also recognized that our 8-ensemble AdaSTEM model cost 8 times more computational time than other simple models both in training and prediction, which should be considered carefully before adapting them. Future work could include parallel implementation of ensembles, since they are completely independent. Currently, the 8-ensemble model is still efficient enough, but not guaranteed in the ongoing data-booming world.

Unsupervised methods

The initial method chosen was the ROI-based feature representation with the agglomerative clustering method. Both the min-mean-max feature representation and K-means clustering method were later introduced as baselines to compare the primary methods against. Most of our effort and tuning went into refining the ROI-based feature representation. This included a clustering method all its own, which required substantial effort to produce a feature representation that adequately captured the spatial-temporal qualities of the observation data in as concise a way as possible. It was believed based on past personal experience in other geospatial analytic tasks that DBSCAN would be an appropriate choice and the ROI-based feature representation would deliver good results. We were correct in thinking that the ROI-based spatial aggregation method provided a better representation than the baseline min-mean-max method, but the agglomerative clustering method did not always outperform K-means. This was a relatively late finding in the course of the project, and further work would focus more on K-means clustering method to determine its strengths and weaknesses when working with a larger number of clusters. The other opportunity that did not fall in the scope of the work but would be available in the future would be using genetic species similarity to compare our clustering results against.

Ethical Considerations

In the supervised learning part, the data and training have less possibility to involve ethical consideration, but there are some concerns in implementing the data: 1) The model could be used to predict patterns under current climate situations. However, if our model overestimates the range of species, it could provide misleading information that this species is under healthy circumstances and face no threats, which could be wrong. As a result, NGO could pull attention away from this species, resulting in species population decline. 2) This model could also be used for projecting migration patterns under future climate, and like the case in (1), could result in excessive attention or deficient attention toward species.

In the unsupervised learning part, similar concern is raised when the pattern clustering result is used for any conservation action. For example, some conservation biology studies species interaction and how that is important in conservation. Conditions and simplified measures should be stated before diving into any conservation strategies.

Statement of Work

Yangkang Chen contributed the draft of the background, all analysis, visualizations, and results related to supervised learning (except for the baseline models) and ethical consideration.

Dean Lawrence contributed all content, visualizations, and results within this report related to unsupervised methods.

Yangkang Chen and Dean Lawrence both contributed to the editing of the report.

Reference

1. Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W.M. and Kelling, S., 2020. Modeling avian full annual cycle distribution and population trends with citizen science data. Ecological Applications, 30(3), p.e02056.
2. Sklearn example: https://scikit-learn.org/stable/auto_examples/applications/plot_species_distribution_modeling.html
3. Hubalek, Z., 2005. Co-fluctuation among bird species in their migration timing. FOLIA ZOOLOGICA-PRAHA-, 54(1/2), p.159.
4. GDAL/OGR contributors, 2022. GDAL/OGR Geospatial Data Abstraction software Library. Open Source Geospatial Foundation. URL https://gdal.org DOI: 10.5281/zenodo.5884351.
5. eBird data: https://science.ebird.org/en/use-ebird-data
6. EarthEnv topography dataset: https://www.earthenv.org/topography
7. ESA CCI land cover dataset: https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-land-cover?tab=form
8. WorldClim climate dataset: https://www.worldclim.org
9. Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
10. NetCDF4 python API: https://github.com/Unidata/netcdf4-python

Appendix

Supplementary Table 1. Feature importance

Feature	mean_by_permutation	std_by_permutation	mean_by_model_output
DOY	0.03302	0.00132	0.01226
elevation_mean	0.00805	0.00036	0.02481
duration_minutes	0.00705	0.00060	0.01539
bio4	0.00669	0.00023	0.04408
slope_mean	0.00438	0.00030	0.01844
bio7	0.00430	0.00013	0.02880
bio19	0.00351	0.00052	0.02170
bio2	0.00337	0.00012	0.02519
bio17	0.00312	0.00037	0.02058
tmin	0.00273	0.00021	0.01731
bio11	0.00268	0.00026	0.03342
bio6	0.00265	0.00020	0.02876
bio8	0.00242	0.00021	0.02235
prec	0.00239	0.00015	0.01790
bio3	0.00227	0.00038	0.02197
bio9	0.00223	0.00021	0.01761
bio18	0.00222	0.00021	0.02284
slope_std	0.00219	0.00034	0.01602
bio12	0.00219	0.00020	0.02064
eastness_mean	0.00194	0.00025	0.01577
bio5	0.00192	0.00029	0.02021
northness_std	0.00191	0.00025	0.01633
eastness_std	0.00186	0.00028	0.01617
bio1	0.00177	0.00014	0.02066
bio10	0.00173	0.00020	0.01839
bio15	0.00167	0.00024	0.01673

bio14	0.00141	0.00010	0.01772
tmax	0.00123	0.00013	0.01707
northness_mean	0.00119	0.00022	0.01514
bio13	0.00102	0.00015	0.01543
elevation_std	0.00085	0.00028	0.01614
bio16	0.00080	0.00022	0.01214
Traveling	0.00058	0.00008	0.01471
Stationary	0.00047	0.00021	0.01191
urban_and_built_up_lands	0.00045	0.00008	0.00961
open_shrublands	0.00033	0.00003	0.00611
evergreen_needleleaf_forests_ed	0.00030	0.00004	0.00596
woody_savannas_lpi	0.00030	0.00005	0.00997
mixed_forests	0.00030	0.00009	0.01161
deciduous_broadleaf_forests	0.00029	0.00004	0.00833
urban_and_built_up_lands_ed	0.00027	0.00004	0.00613
woody_savannas	0.00026	0.00006	0.01360
grasslands	0.00022	0.00014	0.01335
grasslands_lpi	0.00022	0.00005	0.00810
savannas	0.00020	0.00014	0.01131
permanent_wetlands	0.00019	0.00010	0.01390
mixed_forests_lpi	0.00018	0.00004	0.00790
water_bodies	0.00016	0.00010	0.01232
evergreen_needleleaf_forests	0.00014	0.00008	0.01010
open_shrublands_ed	0.00013	0.00003	0.00185
cropland_or_natural_vegetation_mosaics_ed	0.00012	0.00002	0.00328
evergreen_broadleaf_forests_ed	0.00010	0.00004	0.00155
week	0.00010	0.00001	0.00094

woody_savannas_ed	0.00008	0.00010	0.01029
urban_and_built_up_lands_lpi	0.00006	0.00004	0.00530
evergreen_broadleaf_forests	0.00005	0.00004	0.00281
savannas_lpi	0.00005	0.00009	0.00836
permanent_wetlands_lpi	0.00005	0.00007	0.00548
entropy	0.00004	0.00022	0.01190
mixed_forests_ed	0.00004	0.00006	0.00894
deciduous_broadleaf_forests_ed	0.00004	0.00005	0.00484
croplands_ed	0.00003	0.00004	0.00584
croplands	0.00002	0.00011	0.00846
grasslands_ed	0.00001	0.00011	0.00926
cropland_or_natural_vegetation_mosaics	0.00001	0.00003	0.00552
evergreen_broadleaf_forests_lpi	0.00001	0.00002	0.00141
closed_shrublands_lpi	0.00000	0.00000	0.00041
savannas_pd	0.00000	0.00000	0.00000
open_shrublands_pd	0.00000	0.00000	0.00000
croplands_pd	0.00000	0.00000	0.00000
permanent_wetlands_pd	0.00000	0.00000	0.00000
deciduous_needleleaf_forests_pd	0.00000	0.00000	0.00000
obsvr_species_count	0.00000	0.00000	0.01316
time_observation_started_minute_of_day	0.00000	0.00000	0.01412
cropland_or_natural_vegetation_mosaics_pd	0.00000	0.00000	0.00000
urban_and_built_up_lands_pd	0.00000	0.00000	0.00000
water_bodies_pd	0.00000	0.00000	0.00000
closed_shrublands_pd	0.00000	0.00000	0.00000

woody_savannas_pd	0.00000	0.00000	0.00000
deciduous_needleleaf_forests_lpi	0.00000	0.00000	0.00000
open_shrublands_lpi	0.00000	0.00000	0.00039
month	0.00000	0.00000	0.00000
grasslands_pd	0.00000	0.00000	0.00000
deciduous_needleleaf_forests_ed	0.00000	0.00000	0.00000
evergreen_broadleaf_forests_pd	0.00000	0.00000	0.00000
deciduous_needleleaf_forests	0.00000	0.00000	0.00018
year	0.00000	0.00000	0.00000
evergreen_needleleaf_forests_pd	0.00000	0.00000	0.00000
non_vegetated_lands_pd	0.00000	0.00000	0.00000
deciduous_broadleaf_forests_pd	0.00000	0.00000	0.00000
mixed_forests_pd	0.00000	0.00000	0.00000
Area	0.00000	0.00001	0.00100
closed_shrublands_ed	0.00000	0.00000	0.00045
non_vegetated_lands_lpi	0.00000	0.00001	0.00107
cropland_or_natural_vegetation_mosaics_lpi	-0.00001	0.00004	0.00364
permanent_wetlands_ed	-0.00001	0.00004	0.00561
deciduous_broadleaf_forests_lpi	-0.00001	0.00003	0.00475
non_vegetated_lands_ed	-0.00001	0.00002	0.00093
closed_shrublands	-0.00002	0.00002	0.00171
water_bodies_ed	-0.00002	0.00005	0.00556
water_bodies_lpi	-0.00002	0.00001	0.00345
non_vegetated_lands	-0.00002	0.00001	0.00435
croplands_lpi	-0.00003	0.00006	0.00556

evergreen_needleleaf_forests_lpi	-0.00004	0.00008	0.00629
savannas_ed	-0.00010	0.00004	0.00847

Supplementary Table 2. Failure Analysis (top 20 features)

variable	wrong_pred_mean	wrong_pred_std	right_pred_mean	right_pred_std	t	p	wrong_minus_right
bio3	31.1417	11.4271	47.5257	24.0638	-92.7101	0.00E+00	-16.3840
bio11	-0.1381	7.7520	9.1306	12.0095	-102.5191	0.00E+00	-9.2688
slope_mean	2.0819	3.0772	3.5263	5.0595	-38.1491	0.00E+00	-1.4444
bio9	7.0308	12.3219	13.2751	11.7741	-65.0833	0.00E+00	-6.2443
bio8	11.7343	8.8659	16.2740	9.7189	-59.0228	0.00E+00	-4.5397
bio7	34.6039	11.5767	23.1526	13.3608	109.3235	0.00E+00	11.4513
tmin	5.6592	8.9814	10.1831	9.5991	-59.2592	0.00E+00	-4.5239
bio18	218.9118	161.3034	287.4842	231.3290	-39.0338	0.00E+00	-68.5724
bio4	800.8296	314.5152	443.7489	381.6154	120.3609	0.00E+00	357.0807
bio6	-7.5834	8.5316	3.0836	12.7044	-111.0509	0.00E+00	-10.6670
bio12	888.5442	549.9419	1121.6829	856.7325	-36.1686	2.17E-284	-233.1388

bio2	11.0361	3.9753	9.9140	4.0402	34.5595	5.37E-260	1.1221
duration_minutes	64.3628	78.3682	92.9012	114.0323	-33.0122	1.47E-237	-28.5384
slope_std	0.2960	0.4790	0.4031	0.5578	-24.5268	2.10E-132	-0.1071
prec	73.4752	59.6332	90.3304	97.7459	-23.0356	4.50E-117	-16.8552
elevation_mean	462.8242	604.8264	553.8674	787.3542	-15.0336	5.11E-51	-91.0432
bio19	207.7454	164.3089	234.9254	257.4110	-14.0423	9.58E-45	-27.1800
eastness_mean	0.0196	0.1601	0.0045	0.1701	11.1348	8.87E-29	0.0151
bio17	105.7573	95.4730	100.7385	110.5010	5.7963	6.80E-09	5.0189
DOY	170.5579	95.4176	167.0156	99.2696	4.4624	8.12E-06	3.5423

Supplementary Table 3. Sensitivity & learning curve analysis

training_size	ensemble_fold	recall	precision	average_precision	roc_auc	cohen_kappa	f1
38983	1	0.8264	0.1702	0.1507	0.7892	0.2059	0.2822
	2	0.8577	0.1680	0.1520	0.8029	0.2066	0.2809
	3	0.8702	0.1610	0.1472	0.8033	0.1973	0.2717
	4	0.8673	0.1643	0.1498	0.8060	0.2030	0.2763
	5	0.8574	0.1633	0.1478	0.8011	0.2006	0.2744
	6	0.8714	0.1627	0.1488	0.8058	0.2004	0.2742
	7	0.8714	0.1652	0.1511	0.8078	0.2043	0.2778
	8	0.8714	0.1640	0.1500	0.8066	0.2023	0.2761
	9	0.8855	0.1645	0.1519	0.8120	0.2036	0.2774

	10	0.8793	0.1645	0.1513	0.8099	0.2035	0.2772
77967	1	0.8506	0.1921	0.1724	0.8106	0.2386	0.3135
	2	0.8773	0.1916	0.1753	0.8235	0.2416	0.3145
	3	0.8864	0.1887	0.1738	0.8267	0.2388	0.3111
	4	0.8866	0.1877	0.1730	0.8260	0.2373	0.3099
	5	0.8892	0.1905	0.1758	0.8291	0.2419	0.3138
	6	0.9016	0.1898	0.1768	0.8331	0.2413	0.3136
	7	0.8910	0.1899	0.1754	0.8300	0.2414	0.3131
	8	0.8969	0.1912	0.1774	0.8332	0.2438	0.3153
	9	0.8918	0.1895	0.1752	0.8298	0.2407	0.3126
	10	0.8977	0.1893	0.1758	0.8318	0.2407	0.3127
116950	1	0.8498	0.2011	0.1797	0.8208	0.2553	0.3253
	2	0.8885	0.1930	0.1777	0.8341	0.2479	0.3171
	3	0.9020	0.1950	0.1814	0.8407	0.2519	0.3207
	4	0.9033	0.1936	0.1802	0.8407	0.2500	0.3188
	5	0.9029	0.1937	0.1803	0.8405	0.2502	0.3190
	6	0.9096	0.1908	0.1785	0.8414	0.2462	0.3154
	7	0.9085	0.1926	0.1801	0.8425	0.2492	0.3178
	8	0.9065	0.1930	0.1801	0.8420	0.2497	0.3182
	9	0.9150	0.1929	0.1812	0.8451	0.2499	0.3186
	10	0.9164	0.1936	0.1821	0.8462	0.2511	0.3197
155933	1	0.8659	0.2212	0.1996	0.8355	0.2838	0.3524
	2	0.8996	0.2159	0.2001	0.8485	0.2804	0.3482
	3	0.9138	0.2129	0.1996	0.8525	0.2771	0.3453
	4	0.9167	0.2138	0.2009	0.8546	0.2789	0.3468
	5	0.9190	0.2118	0.1994	0.8543	0.2761	0.3443
	6	0.9241	0.2129	0.2011	0.8569	0.2780	0.3461
	7	0.9129	0.2117	0.1983	0.8517	0.2754	0.3437
	8	0.9158	0.2148	0.2016	0.8548	0.2804	0.3480

	9	0.9214	0.2123	0.2002	0.8554	0.2769	0.3451
	10	0.9180	0.2123	0.1996	0.8540	0.2766	0.3448
194916	1	0.8791	0.2276	0.2073	0.8449	0.2947	0.3616
	2	0.9055	0.2262	0.2103	0.8561	0.2958	0.3619
	3	0.9145	0.2198	0.2061	0.8561	0.2870	0.3545
	4	0.9163	0.2242	0.2104	0.8595	0.2938	0.3603
	5	0.9195	0.2237	0.2104	0.8605	0.2932	0.3598
	6	0.9204	0.2221	0.2091	0.8600	0.2910	0.3578
	7	0.9190	0.2241	0.2107	0.8608	0.2940	0.3604
	8	0.9220	0.2234	0.2106	0.8616	0.2932	0.3597
	9	0.9241	0.2229	0.2104	0.8621	0.2926	0.3591
	10	0.9222	0.2241	0.2112	0.8621	0.2942	0.3605
233900	1	0.8824	0.2345	0.2139	0.8495	0.3047	0.3705
	2	0.9105	0.2333	0.2177	0.8615	0.3063	0.3714
	3	0.9224	0.2304	0.2171	0.8648	0.3030	0.3687
	4	0.9207	0.2309	0.2172	0.8645	0.3036	0.3692
	5	0.9251	0.2329	0.2199	0.8671	0.3068	0.3721
	6	0.9258	0.2317	0.2189	0.8668	0.3052	0.3707
	7	0.9219	0.2321	0.2186	0.8654	0.3054	0.3708
	8	0.9285	0.2323	0.2199	0.8682	0.3063	0.3717
	9	0.9258	0.2321	0.2192	0.8669	0.3056	0.3711
	10	0.9274	0.2317	0.2191	0.8674	0.3052	0.3707
272884	1	0.8867	0.2422	0.2215	0.8550	0.3161	0.3804
	2	0.9097	0.2396	0.2233	0.8641	0.3152	0.3793
	3	0.9264	0.2343	0.2214	0.8683	0.3089	0.3740
	4	0.9257	0.2388	0.2254	0.8706	0.3156	0.3796
	5	0.9251	0.2375	0.2241	0.8697	0.3137	0.3780
	6	0.9312	0.2358	0.2236	0.8713	0.3117	0.3763
	7	0.9283	0.2374	0.2246	0.8710	0.3139	0.3781

	8	0.9291	0.2377	0.2251	0.8715	0.3144	0.3786
	9	0.9330	0.2357	0.2239	0.8720	0.3118	0.3764
	10	0.9308	0.2375	0.2252	0.8721	0.3142	0.3785
311867	1	0.9005	0.2403	0.2222	0.8627	0.3169	0.3794
	2	0.9240	0.2340	0.2206	0.8707	0.3108	0.3735
	3	0.9319	0.2303	0.2184	0.8722	0.3061	0.3693
	4	0.9304	0.2356	0.2231	0.8743	0.3138	0.3760
	5	0.9327	0.2332	0.2213	0.8740	0.3105	0.3731
	6	0.9335	0.2325	0.2208	0.8740	0.3094	0.3722
	7	0.9346	0.2341	0.2225	0.8752	0.3118	0.3744
	8	0.9346	0.2330	0.2215	0.8746	0.3103	0.3730
	9	0.9357	0.2323	0.2211	0.8747	0.3094	0.3723
	10	0.9354	0.2335	0.2221	0.8753	0.3111	0.3738
350851	1	0.9044	0.2491	0.2309	0.8666	0.3283	0.3906
	2	0.9292	0.2485	0.2350	0.8775	0.3304	0.3921
	3	0.9365	0.2424	0.2307	0.8779	0.3226	0.3852
	4	0.9370	0.2484	0.2364	0.8810	0.3312	0.3927
	5	0.9385	0.2476	0.2359	0.8813	0.3303	0.3918
	6	0.9402	0.2412	0.2303	0.8790	0.3212	0.3839
	7	0.9409	0.2454	0.2343	0.8814	0.3274	0.3893
	8	0.9422	0.2444	0.2336	0.8814	0.3260	0.3881
	9	0.9409	0.2441	0.2331	0.8809	0.3256	0.3877
	10	0.9381	0.2452	0.2336	0.8803	0.3270	0.3888
389834	1	0.9056	0.2542	0.2357	0.8704	0.3364	0.3969
	2	0.9308	0.2493	0.2360	0.8802	0.3330	0.3932
	3	0.9370	0.2434	0.2317	0.8802	0.3252	0.3865
	4	0.9360	0.2463	0.2342	0.8813	0.3293	0.3900
	5	0.9394	0.2454	0.2340	0.8823	0.3284	0.3891
	6	0.9443	0.2434	0.2330	0.8834	0.3259	0.3870

	7	0.9440	0.2457	0.2351	0.8844	0.3292	0.3899
	8	0.9436	0.2462	0.2355	0.8844	0.3298	0.3905
	9	0.9451	0.2447	0.2344	0.8843	0.3279	0.3887
	10	0.9447	0.2454	0.2349	0.8845	0.3288	0.3895

Supplementary Table 4. Complete feature list

duration_minutes
protocol_type
effort_distance_km
number_observers
time_observation_started
observation_date
country
obsvr_species_count
elevation_mean
slope_mean
eastness_mean
northness_mean
elevation_std
slope_std
eastness_std
northness_std
prec
tmax
tmin

bio1
bio2
bio3
bio4
bio5
bio6
bio7
bio8
bio9
bio10
bio11
bio12
bio13
bio14
bio15
bio16
bio17
bio18
bio19

closed_shrublands
closed_shrublands_ed
closed_shrublands_lpi
closed_shrublands_pd
cropland_or_natural_vegetation_mosaics
cropland_or_natural_vegetation_mosaics_ed
cropland_or_natural_vegetation_mosaics_lpi
cropland_or_natural_vegetation_mosaics_pd
croplands
croplands_ed
croplands_lpi
croplands_pd
deciduous_broadleaf_forests
deciduous_broadleaf_forests_ed
deciduous_broadleaf_forests_lpi
deciduous_broadleaf_forests_pd
deciduous_needleleaf_forests
deciduous_needleleaf_forests_ed
deciduous_needleleaf_forests_lpi
deciduous_needleleaf_forests_pd
evergreen_broadleaf_forests
evergreen_broadleaf_forests_ed
evergreen_broadleaf_forests_lpi
evergreen_broadleaf_forests_pd
evergreen_needleleaf_forests
evergreen_needleleaf_forests_ed

evergreen_needleleaf_forests_lpi
evergreen_needleleaf_forests_pd
grasslands
grasslands_ed
grasslands_lpi
grasslands_pd
mixed_forests
mixed_forests_ed
mixed_forests_lpi
mixed_forests_pd
non_vegetated_lands
non_vegetated_lands_ed
non_vegetated_lands_lpi
non_vegetated_lands_pd
open_shrublands
open_shrublands_ed
open_shrublands_lpi
open_shrublands_pd
permanent_wetlands
permanent_wetlands_ed
permanent_wetlands_lpi
permanent_wetlands_pd
savannas
savannas_ed
savannas_lpi
savannas_pd

urban_and_built_up_lands
urban_and_built_up_lands_ed
urban_and_built_up_lands_lpi
urban_and_built_up_lands_pd
water_bodies
water_bodies_ed
water_bodies_lpi
water_bodies_pd
woody_savannas
woody_savannas_ed
woody_savannas_lpi
woody_savannas_pd
entropy
doy