Machine Learning for Species Analysis

Student number s1910360

Student number s2572047

Student number s1922260

Abstract

This report tackles the problem of species identification using location data, and a resultant analysis of the effect of climate change on specific species. We explore different types of population distribution and use these findings to accurately train and test several machine learning algorithms (**Logistic Regression**, **K-Nearest Neighbours**, **Random Forest** and **Feed-Forward Neural Networks**). We then compare models using AUC-ROC, AUC-PR, F2 scores, and Cohen's kappa, and further investigate implementing new bio-climatic variables to predict species affected by climate change and analyze the resultant F2 metric.

1 Introduction

A key problem in ecology is understanding the many ecosystems of the world and how they respond to climate change, conservation efforts, and habitat destruction. Central to this work is the monitoring of species present in a fixed area and collection of data from a variety of sources. Ecologists can then perform species distribution modelling (SDM) using this data and determine the species present in an ecosystem, their respective populations, and how these populations change over time. However, the processing and analysis of data can take years due to the large number of samples taken. Machine learning offers a promising method to speed up SDM. In this report, we use various machine learning models to analyse data from iNaturalist (`www.inaturalist.org`), a "crowd-sourced species identification system". We evaluate our machine learning methods to determine which has the best classification accuracy for this data-set, then attempt to improve the model capabilities by training on eight new bio-climatic variables from WorldClim [1], visually explored in Appendix F. We use the most improved model in conjunction with projected temperature data from the Met Office Hadley Centre HadGEM3-GC31-LL model [2] to subsequently determine which species are most affected by climate change, which could then be used to focus conservation efforts.

2 Background

Machine learning methods have been used extensively in the context of SDM; see [3], [4], [5], [6] for examples. Beery et. al. [7] provide a comprehensive review of SDM aimed at computer scientists, highlighting the common methods, terminology, and challenges associated with this area. A common practice in previous work is to combine geospatial data for a particular species with bioclimatic variables (e.g. temperature, precipitation) and topographical attributes (e.g. elevation, slope, flow water direction). Lorena et. al. [8] compare the use of nine different machine learning models in predicting the distribution of thirty five Latin American plant species. Geospatial data for each species was combined with nine environmental layers, comprising of four climatic variables and five topographical attributes. Lee et. al. [9] used a similar approach to model the potential distribution of invasive ant species under climate change. This approach seems viable for our purposes.

3 Exploratory data analysis

A brief overview of the available data structure is included in appendix A. Notably, the mean number of locations attributed to each species is much higher in the test data-set vs. the train data-set. The training data, sourced from citizen scientists, is noisy due to non-regularized collection methods, and important information relevant to conservation efforts such as the observation date of the species is not provided. Additionally, the format of the two data-sets is different. The train data-set provides locations and a singular associated species, whereas the test data-set provides locations and a corresponding list of multiple species. It is important to note the difficulty that this may present in creating models that can accurately predict multiple species per location[1]. The distribution of these two data-sets is best visualized through the plots given in Figure 1.

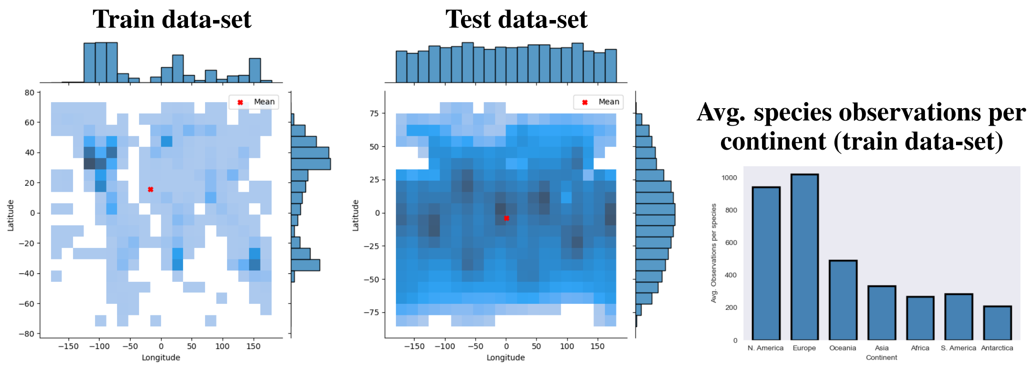

Figure 1: Distribution of data from the train (left) and test (middle) data-sets at locations where there are species observation ≥ 1, note the offset mean location in the train data-set, and the centralized mean location in the test data-set. (Right) is a comparison of the number of observations per species in the train data-set for each continent.

Data from the training data-set is not evenly spread, with larger amounts of data associated with the longitude and latitude ranges (-120, -90) and (20, 50) for example, roughly corresponding to North America. This can be attributed to the fact that there may be more iNaturalist users in this region. An analysis on the number of observations per species for each continent is subsequently carried out to further show this data imbalance. A random sample of twenty locations (converted to a continent name) was taken for every species, and the most common continent among those twenty was taken to be the species continent. As shown in Figure 1, Europe has the highest number of observations (data-points) per species. This shows a clear bias in the data; species from Europe and North America have, on average, a larger number of observations. A more comprehensive table of this data is included in Appendix B. The test data-set on the other hand, considers a (roughly) even spread of location data points. Additionally, the species in the data-set have widely varying population distributions. Figure 2 shows the extremes in species distribution span and density in the test data-set.

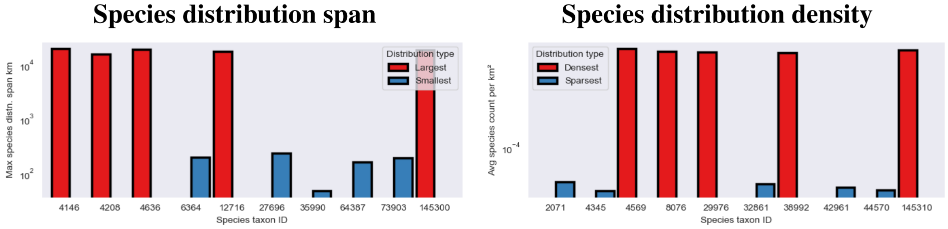

Figure 2: (Left) The variation on population distribution of 10 different species (top 5 largest vs top 5 smallest). (Right) The variation on population density of 10 different species (top 5 densest vs top 5 sparsest), note the logarithmic scaling of both plots.

Notably, population spans vary widely. An initial hypothesis is that this will affect a machine learning model's ability to correctly predict species distributions[2], and so this is something that we would like to investigate. Not only do the species population distributions span different distances, but they also have a wide range of densities. Some species have a much larger count in the train data-set attributed to a smaller area, whereas other species have small counts over larger areas. A second hypothesis

[1] This is because the training data-set provides us with data suited to a singular label classification problem, whereas the test data-set is suited toward a multi-label classification problem.

[2] Largely spread populations may provide less generalization for the ML models to learn about species population.

to consider is our machine learning methods' capability to predict sparse and dense population distributions[3].

4 Data preparation

Through the exploratory analysis, it was evident that a significant imbalance existed within the data-set that needed to be considered before model training could begin. Imbalances in class distributions introduce bias during the training of the models, leading to a loss of performance. To avoid this, individual species weights were introduced via the equation[4]

```python
species_weight = len(train_locs)/(species_count * no.species)
```

These weights ensure that the models place more emphasis on species that may be underrepresented in the training data-set [10]. This method then allowed for the whole data-set to be used in training the algorithms, unlike other methods such as sub-sampling which would sacrifice information about the data-set. Due to our data having only two features, latitude and longitude, dimensionality reduction is unneeded, as our data is already in a low-dimensional space (2D).

5 A Species Distribution analysis using different Machine Learning methods

The central goal of this task is to predict the different species found at a given set of specified coordinates (latitude, longitude) using a variety of machine learning methods. In our investigation, we analyze the multi-label, classification capabilities of 1. **Logistic Regression**, 2. **Random Forest**, 3. **K-Nearest Neighbour** & 4. **Feed Forward Neural Network** models. With these four implemented methods we perform subsequent analysis to evaluate and compare the accuracy of each model. Having done so, we aim to answer questions such as the prevalence of specific species in given regions.

Logistic Regression is a simple but powerful linear model for classification which uses a logistic function to model the probability of a binary outcome. This was implemented using the built-in LogisticRegression implementation in scikit-learn. As we wish to study the per-class probabilities at each location, we use the multinomial method, which seeks to minimise the cross-entropy loss during training, rather than a one-versus-rest scheme. The softmax function is then used to calculate the per-class probabilities. The main disadvantage of logistic regression in this context is that the data is unlikely to be linearly separable; for example, if a particular species is distributed across two different continents. Despite this, the ease of implementation for logistic regression makes it an appealing model to study. **Random Forests** are a suitable choice for our data-set due to their inherent ability to handle complex data and capture non-linear patterns [11]. The ensemble nature of Random Forests is useful to avoid over-fitting and allow for better generalization. Implementation-wise, the model was realized using the scikit-learn library, the number of estimators was set at 100 and the maximum depth was optimized by calculating the F1 scores for different depth values (see Appendix D). At *max_depth=10*, the performance of the model plateaus and so it was chosen to avoid over-fitting. **K-Nearest Neighbours (KNN)** calculates the k-nearest neighbours to a data-point based on some chosen distance metric, then assigns a class based on whichever class makes up the majority of these neighbours. Using sci-kit learn's implementation, we chose $k = 75$ as the number of neighbours, determined by calculating the mean F1 score across all species for different values of k (see Appendix E). In addition, we use the standard Euclidean metric to determine the distance between points and neighbours. The per-class probabilities for a particular location are calculated as the number of neighbours belonging to each class divided by the total number of neighbours. **Feed-Forward Neural Networks (FFNN's)** are able to implicitly detect non-linear relationships through the use of a connected layer architecture. They are therefore a potentially suitable model to use to train population distributions. They were implemented using PyTorch. A visual of the implementation is provided in Appendix C. Specifically, the model takes in longitude and latitude as features and passes them through a set of hidden layers using an activation function. The activation function used between hidden layers is a Rectified Linear unit (ReLu), and the output used a sigmoid activation to scale outputs between 1 & 0 for a k-hot vector, i.e. species present or not. Binary cross entropy was used to evaluate the loss of the model as it is best suited for a multi-label classification, and an Adam optimizer was implemented with a learning rate of 0.001 to best adjust the weights accordingly without over-fitting.

[3] Sparsely spread populations provide less correlation for the ML models to learn, whereas dense populations have fewer outliers.

[4] Note; this is the built-in weight equation from sk-learn.

5.1 Methods & evaluation

To comprehensively assess model performance, we utilized several metrics, including AUC-ROC & AUC-PR scores. To generalize our assessment, we calculated averages for four distinct groups as detailed in Section 3: the top 5 most dense, most sparse, smallest span, and largest span, along with an average across all 500 species. The advantage of employing AUC-ROC and AUC-PR is that they allow for easier comparison between models as they don't depend on the different decision thresholds. Considering the inherent class imbalances in our data-set, where the absence of a species class always dominates, precision and recall emerge as important metrics to evaluate [12].

It is important to note that, given the nature of our project, our primary concern centers around positive class predictions. Consequently, AUC-PR scores assume greater relevance in our model assessment than AUC-ROC scores. Within the positive class, the emphasis is placed on minimizing false negatives for accurate species identification. To address this, we opted to evaluate F2 scores, providing a balanced measure of accuracy that prioritizes recall over precision.[5][13] We also evaluated Cohen's Kappa (CK) to check whether agreement between predictions and reality is achieved by chance.[14]

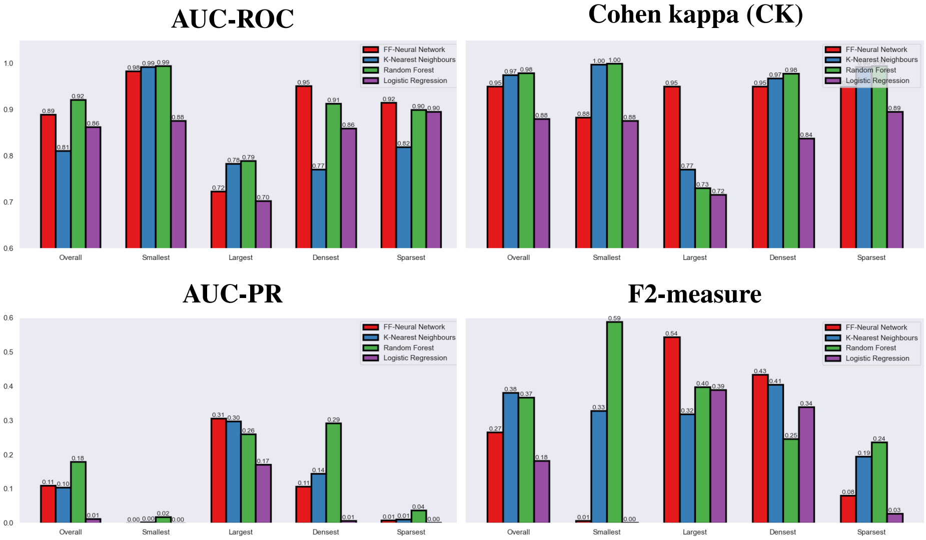

Figure 3: Model performance according to different population distribution types. For each population distribution type the top-5 were evaluated. Note the y axis range for AUC-ROC and Cohen's Kappa (0.6 - 1.0) as opposed to the range for AUC-PR and F-measure (0.0 - 0.6).

5.2 Analysis & results

From Figure 3, we see that models tend to achieve better AUC-ROC values with data that has a smaller distribution compared to larger distributions. However, the opposite is found with the AUC-PR scores. This is because of the inherent imbalance in the test data; the smallest distributions were attributed to species in the data-set which had very low counts, for example, *Gallotia Stehlini* (35990) had only 2 data-points. This skews the Precision and Recall to 0, as well as the False Positive Rate (increasing the ROC score). A variation is not observed between densest and sparsest for the ROC curves, but the densest distribution outperforms the PR score of the sparsest distribution. This result is expected as denser distributions are generally easier to model due to the data being highly correlated, kin to having fewer 'outliers' [15]. The values for CK are all high, confirming the reliability of our models. The F2-score shows variability in performance between models and between the different distribution types, for example RF and KNN perform consistently well across all types yet FFNN and LR perform comparatively worse on the smallest and sparsest distributions. A visual exploration of results provides a more intuitive evaluation of our models' capabilities to predict species distributions (Figure 4).

[5] F2 comes from the generalized F-Beta measure, $F(\beta) = (1 + \beta) * Pr * Re/(\beta^2 * Pr + Re)$; an increase in β emphasizes the importance of Recall.

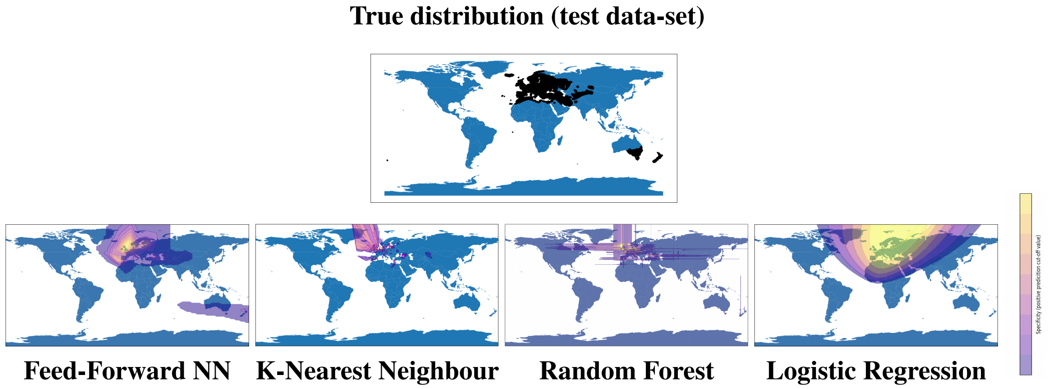
True distribution (test data-set)
Feed-Forward NN
K-Nearest Neighbour
Random Forest
Logistic Regression
Specificity (positive prediction cut-off value)

Figure 4: Predicted Turdus Merula distributions using various trained models (bottom left-right) vs. true distribution (top). Note: The heat-plot nature of the maps is distribution predictions at different model parameters i.e. varying prediction ‘cut-off’ values.

5.3 Conclusion

Analyzing our models it is clear that some models are more adept at predicting different population distributions. RF and KNN performed better on average with small and sparse distributions, whereas FFNN performed better with larger distributions. The Random Forest algorithm consistently did best averaged over all species (0.92, 0.18, 0.99, 0.37 for AUC-ROC, AUC-PR, CK, and F2-score respectively), importantly, however, all models returned very low AUC-PR values. To mitigate our models collective, consistently low AUC-PR results over all population distributions, we further investigate the addition of eight bio-climatic variables into the KNN, FFNN and RF algorithms and their response to the F2 metric.

6 Extended analysis of the effect of climate change on different species

6.1 Introduction of six new features

Having been able to classify species at specific locations we now seek to implement new bio-climatic features to improve the prediction capabilities of our models. Annual temperature range, mean temperature of coldest quarter, mean temperature of warmest quarter, annual precipitation, precipitation of driest month, and precipitation seasonality data were sourced for all the locations in the train and test data-sets using WorldClim bio-climatic variables [1], available at `https://www.worldclim.org/data/worldclim21.html`. Figures of these variables can be found in Appendix (F). Specifically, these new features allow models to produce more accurate results by isolating species predictions within ranges the species can tolerate, for example features like precipitation of driest month can help identify population distributions of desert-based species. The data sourced is restricted to land based, positive-only measurements. As a result, we reduce the task to only consider training for land based species, reducing the train data-set size to 271270. Future work could incorporate marine-based variables such as sea temperature, salinity, and ice content to build a more robust classifier capable of predicting the distribution of marine species in addition to terrestrial species. The Bio-ORACLE data-set [16] [17] would be suitable for this task. Each location data point within the training and test data-sets was combined with the six bio-climatic variables mentioned above, increasing the number of input features to eight. Upon training the models with the newly sourced features, an analysis of the models' F2 scores - specifically chosen due to sensitivity to precision and recall values - was evaluated, shown in Figure 5.

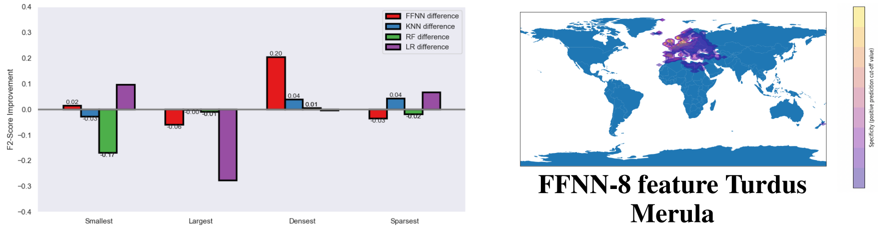

Figure 5: (Left) The metric improvements/losses when trained with new 8-feature data (land-based species only). (Right) Newly predicted Turdus Merula distribution from the FFNN (most improved).

As can be seen from Figure 5, the FFNN model experienced the biggest improvements. Notably, the models performed worse on some types of population distribution; this is because the new features help identify similar regions to that input, smaller distributed populations are then over-identified in a broader range of regions that would be ‘climatically suitable’, as shown in appendix G. Additionally, introducing 6 new features may have led to overfitting of the data, explaining the overall lower than expected improvement in F2 score.

6.2 Effect of climate change on newly trained models

We now use a FFNN, combined with the new bio-climatic features, to predict species most at risk due to climate change. To quantify the effects of climate change, we examine the difference in yearly average temperatures between a baseline value, calculated as the mean over 30 years from 1984-2014, and a projected value for the year 2050, computed under the high emissions scenario (SSP5-8.5). The temperature data is taken from the Met Office Hadley Centre HadGEM3-GC31-LL model [2], prepared for the Coupled Model Intercomparison Project Phase 6 (CMIP6). A visualization of temperature anomaly is shown in appendix H. Based on the 2050 temperature anomaly, we now compute a "score" that quantifies how extreme the predicted temperature increase of a particular location will be due to climate change, which is obtained by normalising the temperature anomaly by the largest increase (in this case 17°C). We note that a few localised patches of ocean saw a decrease in yearly temperatures based on the model data, however, since these areas were small and the decrease was less than a degree, we set the temperature anomaly for these locations to 0°C for the purposes of calculation. Combining this analysis of climate data at different locations with our models’ improved prediction capability we can now provide accurate population distribution predictions and a ‘vulnerability’ score for specific species as shown in Table 1. We note that our vulnerability score is quite simplistic in the sense that a larger increase in temperature does not necessarily translate to a larger negative effect on a species, for example, certain species might be more resilient to changes in temperature than others. For future work, we might define this score differently by considering multiple climate metrics in addition to the yearly average temperature anomaly, for example, the anomalies in maximum temperature of the warmest month, lowest temperature of the coldest month, precipitation rates, sea ice content, etc. This would give a more complete evaluation of how climate change affects habitats beyond an increase in yearly average temperatures. Notably, species with high vulnerability were located near the poles (latitudes of ≈ 80 and ≈ -80 for the Arctic Ground Squirrel, shown in appendix I and Chinstrap Penguin respectively), whereas species with low vulnerability were located in temperate regions between ≈ -50 and ≈ -60 latitude.

Climate change vulnerability analysis					
Most Vulnerable	**Vulnerability**	**Location (largest pop.)**	**Least Vulnerable**	**Vulnerability**	**Location (largest pop.)**
Arctic G. Squirrel	1.0 (298.15)	N.America	*Black Rain Frog*	0.000002	Africa
Chinstrap Penguin	0.70	Antarctica	*Whistling Tree Frog*	0.000012	Australia
Pine Grosbeak	0.38	N.America	*Golden C. Snake*	0.000012	Australia
Common Redpoll	0.20	Europe	*Cape Mole Rat*	0.000015	Africa
Ruffed Grouse	0.18	N.America	*Tusked Frog*	0.000016	Australia

Table 1: An analysis of the most and least vulnerable species to climate change. Scores were scaled according to that achieved by the most vulnerable species.

7 Conclusion

We trained four different machine learning models on a data-set from iNaturalist containing coordinates and names of species identified at those coordinates. The four models used were a Feed-Forward Neural Network (FFNN), Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbours (KNN). We evaluated the performance of these models on the test-data by calculating the PRAUC, ROCAUC, F2-score, and Cohen’s kappa for different subsets of species, as well as an all-species mean. From this, it was determined that the random forest performed the best under our chosen metrics. To improve the accuracy of our models we introduced six bio-climatic features and trained the KNN, FFNN and RF models on these features in addition to the given geospatial data. Overall, this led to some increases in F2 score, particularly for dense populations in the FFNN model. Using past and projected future temperature values, we determined the species that are most likely to be affected by climate change. We defined a "vulnerability score" and in conjunction with the improved eight-feature FFNN, was used to determine that the species in Table 1 are the most likely to be affected by climate change.

References

[1] Stephen E. Fick and Robert J. Hijmans. Worldclim 2: new 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 2017.

[2] Jeff Ridley, Matthew Menary, Till Kuhlbrodt, Martin Andrews, and Tim Andrews. Mohc hadgem3-gc31-ll model output prepared for cmip6 cmip, 2018.

[3] Simone Vincenzi, Matteo Zucchetta, Piero Franzoi, Michele Pellizzato, Fabio Pranovi, Giulio A De Leo, and Patrizia Torricelli. Application of a random forest algorithm to predict spatial distribution of the potential yield of ruditapes philippinarum in the venice lagoon, italy. *Ecological Modelling*, 222(8):1471–1478, 2011.

[4] Jiaxin Jin, Hong Jiang, Jianhui Xu, Wei Peng, Linjing Zhang, Xiuying Zhang, and Ying Wang. Predicting the potential distribution of bamboo with species distribution models. In *2012 20th International Conference on Geoinformatics*, pages 1–4. IEEE, 2012.

[5] Michelle M Jackson, Sarah E Gergel, and Kathy Martin. Citizen science and field survey observations provide comparable results for mapping vancouver island white-tailed ptarmigan (lagopus leucura saxatilis) distributions. *Biological Conservation*, 181:162–172, 2015.

[6] Wim Aertsen, Vincent Kint, Jos Van Orshoven, Kürşad Özkan, and Bart Muys. Comparison and ranking of different modelling techniques for prediction of site index in mediterranean mountain forests. *Ecological modelling*, 221(8):1119–1130, 2010.

[7] Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review. In *ACM SIGCAS conference on computing and sustainable societies*, pages 329–348, 2021.

[8] Ana C Lorena, Luis FO Jacintho, Marinez F Siqueira, Renato De Giovanni, Lúcia G Lohmann, André CPLF De Carvalho, and Missae Yamamoto. Comparing machine learning classifiers in potential distribution modelling. *Expert Systems with Applications*, 38(5):5268–5275, 2011.

[9] Wang-Hee Lee, Jae-Woo Song, Sun-Hee Yoon, and Jae-Min Jung. Spatial evaluation of machine learning-based species distribution models for prediction of invasive ant species distribution. *Applied Sciences*, 12(20):10260, 2022.

[10] Mahdi Hashemi and Hassan Karimi. Weighted machine learning. *Statistics, Optimization and Information Computing*, 6(4):497–525, 2018.

[11] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[12] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[13] Yutaka Sasaki et al. The truth of the f-measure. *Teach tutor mater*, 1(5):1–5, 2007.

[14] Matthijs J Warrens. Five ways to look at cohen's kappa. *Journal of Psychology & Psychotherapy*, 5, 2015.

[15] Samuel Ackerman, Eitan Farchi, Orna Raz, Marcel Zalmanovici, and Parijat Dube. Detection of data drift and outliers affecting machine learning model performance over time. *arXiv preprint arXiv:2012.09258*, 2020.

[16] Lennert Tyberghein, Heroen Verbruggen, Klaas Pauly, Charles Troupin, Frederic Mineur, and Olivier De Clerck. Bio-oracle: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, 21(2):272–281, 2012.

[17] Jorge Assis, Lennert Tyberghein, Samuel Bosch, Heroen Verbruggen, Ester A. Serrão, and Olivier De Clerck. Bio-oracle v2.0: Extending marine data layers for bioclimatic modelling. *Global Ecology and Biogeography*, 27(3):277–284, 2018.

Statement of Contribution

s1910360: Species distribution types (spans and densities), implementation and testing of FFNN (2-feature, 8-feature and vulnerability implementation), 8-feature data-set creation.

s2572047: Background reading and write-up. Implementation and testing of KNN and LR. Obtaining temperature anomaly data.

s1922260: Continent data exploration and data preparation. Implementation and testing of Random Forest. Overall analysis of results.

A Data structure summary of the train and test data-sets

Data structure			
Data-set	**#Data-points**	**#Unique species**	**Mean #locs per species**
Train	272037	500	544
Test	288122	500	3413

Table 2: Data structure of the provided data-sets ‘train’ and ‘test’.

B Species observation count per. continent

Continents	**#Observations**	**#Species**	**Average #Observations per Species**
N. America	126579	134	944.6
Europe	34774	34	1022.8
Oceania	35763	73	489.9
Asia	22317	67	333.1
Africa	32737	122	268.3
S. America	19659	69	284.9
Antarctica	208	1	208

Table 3: Number of Observations, Number of Species ann Average Number of Observations per Species for each Continent.

C Feed-Forward Neural Network Architecture

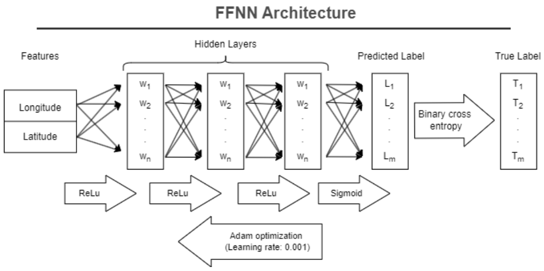

Figure 6: Feed-forward neural network architecture used to predict species on location data.

D Random Forest Optimisation

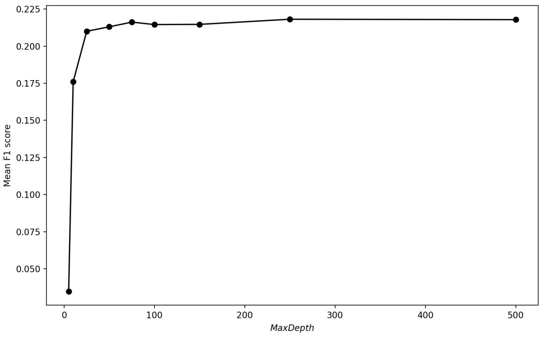

Figure 7: Mean F1 score across all species for different values of maximum depth. At *max_depth=10* the F-score plateaus.

E KNN Optimisation

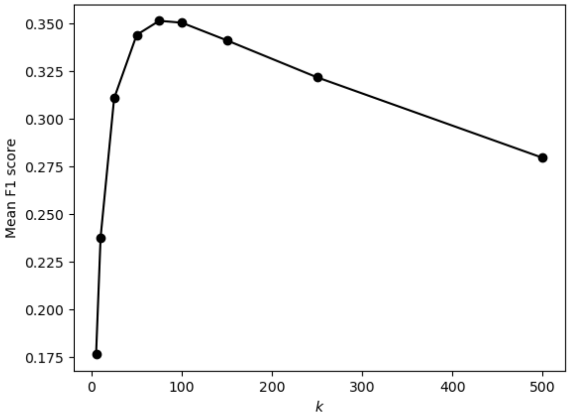

Figure 8: Mean F1 score across all species for different values of k, the number of nearest neighbours used in KNN classification. The maximum value corresponds to $k = 75$.

F Bioclimatic variables

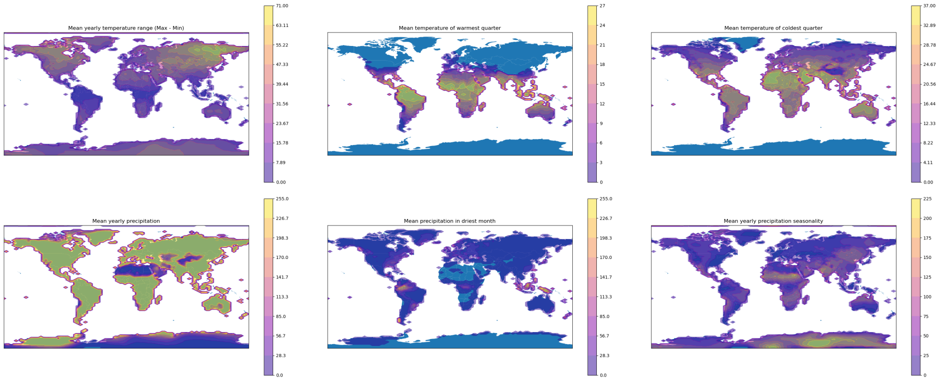

Figure 9: Bio-climatic variables introduced to further train models. (Top) from right to left is the mean yearly temperature range, mean temperature of the warmest quarter, and mean temperature of the coldest quarter. (Bottom) from right to left is the mean yearly precipitation, mean precipitation in the driest month, and mean precipitation seasonality.

G Over estimation of small-distribution species with 8-feature model

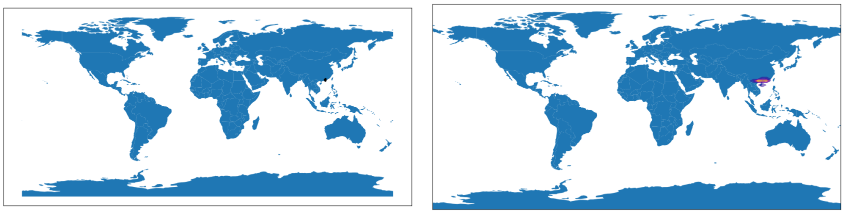

Figure 10: Overestimation of species from the small-population distribution sample (taxon ID 64387) with the use of FFNN trained with bio-climatic variables.

H Temperature anomaly

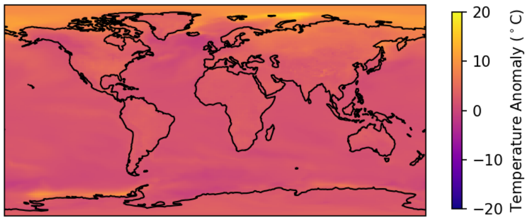

Figure 11: Projected yearly average temperature anomaly for 2050 compared to 1984-2014 baseline

I Arctic Ground Squirrel predicted distribution 8-feature FFNN

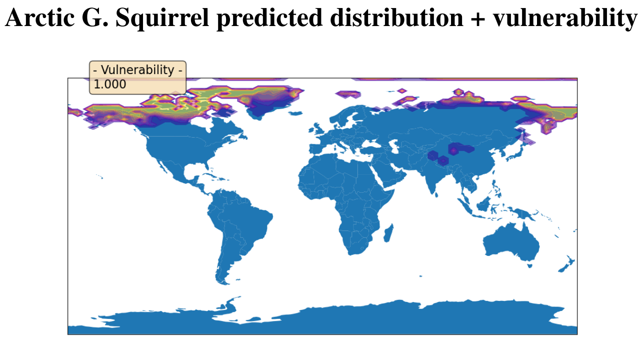
Arctic G. Squirrel predicted distribution + vulnerability

Figure 12: 8-feature FFNN population distribution prediction for the most vulnerable species in the data-set (*Arctic Ground Squirrel*) alongside its scaled vulnerability score, 1.0, indicating the highest net-affect as a result of temperature anomaly.