IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Predicting Animal Population Trends Using Random Forest Models to Enhance Biodiversity Conservation

Aiden Chee

Taipei American School, 800 Zhongshan North Road, Section 6, Taipei, Taiwan, ROC 11152; aidenchee123@gmail.com
Mentor: Jimin Choi

ABSTRACT: Biodiversity is vital for ecological balance, as every species serves a specific function within its ecosystem. The rapid decline in certain animal populations highlights the urgent need for conservation efforts. This study employs a Random Forest model with an integrated data pipeline to predict animal population changes based on physical traits over time. Using the Living Planet Index (LPI) and cross-referenced Wikipedia data, the study examines features such as thermoregulation, habitat, diet, reproductive strategy, and flight capability. Missing data was addressed through forward filling, ensuring continuous and reliable datasets. Results show a minimal correlation between physical traits like habitat and thermoregulation and population trends, indicating that while physical traits offer insights, incorporating environmental or behavioral data is essential for accurate predictions. Future research can build on this framework by integrating advanced modeling techniques and broader datasets to improve biodiversity conservation strategies.

KEYWORDS: Animal Sciences, Ecology, Population Dynamics, Species Extinction, Ecological Modeling.

Introduction

This research aims to predict future animal population trends by analyzing physical traits such as habitat, thermoregulation, diet, and reproductive strategies. These traits influence species' adaptability to environmental changes. For example, climate shifts may impact endothermic animals differently than ectothermic ones, and species with specialized habitat needs may struggle as environments deteriorate. By identifying species at higher risk of population declines, this study seeks to guide conservation efforts toward proactive intervention.

Biodiversity is essential, as every species contributes to its ecosystem. Secondary and tertiary consumers regulate prey populations, pollinators like bees and hummingbirds support plant reproduction, and decomposers such as flies and isopods recycle nutrients. In 2023, 21 animal species were declared extinct in the United States, with estimates predicting up to 150 species will vanish globally each day, equivalent to one extinction every 10 minutes.[1] The United Nations warns that nearly 1 million species are at risk of extinction due to human activities.[2] Freshwater species populations have declined by 83% on average since 1970,[3] and the Amazon rainforest lost approximately 12,000 square kilometers of forest in 2022, an area comparable to Qatar.[3] Population declines disrupt ecosystems, as species' ecological roles go unfulfilled. For instance, gray wolves' extinction in Yellowstone National Park caused an unchecked rise in wild elk populations, leading to overgrazing and vegetation decline.[4] This example underscores how losing a single species can destabilize an entire ecosystem. Preserving current animal populations is, therefore, critical.

By examining the relationship between physical traits and population trends, this research identifies species more vulnerable to environmental changes based on their traits. It seeks to determine how physical traits correlate with population trends and whether these traits can predict species most at risk of population decline. This approach not only enhances understanding of how traits influence survival but also prioritizes conservation for species at higher risk, increasing the chances of preserving biodiversity before irreversible damage occurs.

Researchers argue whether traits alone can reliably predict extinction risk without considering environmental context.[5] This study contributes to that discussion by evaluating how well physical traits predict animal population trends when modeled with statistical population features. The use of a Random Forest regressor aligns with applications of machine learning in conservation, where similar models have been used to assess extinction risk, forecast species distributions, and identify vulnerability patterns across taxonomic groups.[6] By focusing on interpretable models and measurable traits, this study helps solidify the role of trait-based prediction in biodiversity risk.

While prior studies done by Qi have used Random Forest algorithms in bioinformatic settings to evaluate the importance of input features, they often do so in statistical contexts without direct usage in ecological population modelling.[7] Similarly, Moretti and Legg used plant and animal traits to assess responses to ecological disturbances, but did not use historical data.[8] This study extends both studies by using long-term population trends with ecological and physical attributes to better assess the predictive value of traits. Using feature importance values has been common when conducting Random Forest-based studies. However, our usage of the ecological setting could bring challenges in trait interpretation due to environmental variability. These studies provide a strong foundation for applying similar methods to predict species populations and assess vulnerabilities.

We have considered using other models, such as individual-based models (IBMs), which focus on behavior and

DOI: 10.36838/v7i8.15

physiology to predict species responses to environmental changes, making them useful for complex ecosystems. We also considered using dynamic range models (DRMs), which account for population movements and growth, and offer superior predictions in dynamic climates. While these approaches address specific aspects of species responses or general prediction techniques, they do not focus on analyzing physical traits to predict population trends across a broad range of species. This study builds on existing work by integrating physical traits and environmental factors to enhance prediction accuracy and inform targeted conservation strategies.

An efficient data pipeline is essential for this study, as it ensures smooth data flow from collection to model training and performance evaluation. Unlike typical conventional studies, this research deals with high-dimensional, scattered data, incorporating diverse physical and environmental factors across many species. Managing such complexity requires a robust pipeline to handle missing values, select features, and optimize models. This ensures data integrity and enhances prediction accuracy in trait-based population modeling. The pipeline supports data collection, preprocessing, model training, and evaluation.

One major challenge addressed was the prevalence of missing data in the population records. While the Living Planet Index (LPI) provides comprehensive data on over 32,000 populations and 5,200 species, its records for some species over 50 years are incomplete.[9] The pipeline resolved this issue by preprocessing the data, including imputing missing values to ensure continuity and completeness. This step was critical for preparing reliable datasets for modeling. This study utilized the LPI as a primary source due to its extensive biodiversity data. Preprocessing involved cleaning the data and handling missing values to establish a high-quality input for the model. The study sought to uncover correlations between the two by examining species populations alongside their physical traits. The Random Forest model was chosen for its ability to identify correlations and generate accurate predictions. Preprocessed data was fed into the model, which employs multiple decision trees. Each tree acts as an individual model, learning correlations between features to improve predictive accuracy. The model's performance was rigorously evaluated to identify areas for improvement and ensure reliable outcomes.

Methods

Data Collection and Sources:

We analyzed population trends using datasets from the Living Planet Index (LPI), developed by the Zoological Society of London (ZSL) and the World Wildlife Fund (WWF) (WWF/ZSL, 2022). This index tracks population changes across more than 32,000 populations and over 5,200 species, including mammals, birds, amphibians, reptiles, and fish. It gathers data from peer-reviewed studies, government reports, and wildlife surveys, offering a comprehensive view of how environmental changes affect species populations. Integrating the LPI dataset allowed us to examine historical population trends, identify correlations, and improve predictions for future patterns.

Table 1: Formatting of information in the Living Planet Index (LPI) dataset, including data on species populations, geographic locations, and time-series information. The table presents the structure of the dataset used for analysis.

Binomial	Biome	1950	…	1991	…	2020
Falco_punctatus	Tropical and subtropical…	NULL	…	22	…	NULL
Falco_punctatus	Atlantic north temperate	NULL	…	776000	…	1103000

Table 1 summarizes the LPI dataset, which provides taxonomic groupings and population data but lacks information on the physical traits of animals. We developed a data pipeline to fill this gap by sourcing additional details from external resources. Specifically, we used Wikipedia to extract and analyze the physical traits of various species. Recognizing the potential limitations of Wikipedia's credibility, we implemented a validation process. To ensure accuracy, this involved cross-referencing data with reliable scientific databases, such as the Global Biodiversity Information Facility. Using taxonomic classifications, we categorized animals based on shared physical traits, minimizing errors from relying on a single source.

We utilized the Wikipedia-API Python library to retrieve page content and identify relevant details through targeted keywords. Although Wikipedia served as the primary data source, our pipeline is adaptable for incorporating information from other credible databases in future research. This process enabled us to organize species into five sub-datasets, focusing on key traits: thermoregulation, habitat, dietary habits, flight capability, and reproductive strategies. These features were chosen for their influence on population trends. Thermoregulation affects metabolic rates and survival strategies, while habitat provides insight into environmental pressures on species. Dietary habits (e.g., carnivorous, omnivorous, herbivorous) reflect resource availability and feeding behavior. Reproductive strategies (e.g., live birth, egg-laying) influence growth rates and survival. Flight capability impacts species mobility and adaptation to environmental changes. These traits offer a comprehensive understanding of the ecological and biological factors shaping population dynamics.

Data Pipeline:

The effective management and processing of data is of great importance, as the analysis involves complex datasets with diverse features. The data we are working with spans many species and features, resulting in high-dimensional data that may complicate analysis. Additionally, combining data from various sources, such as the Living Planet Dataset (LPD_2022.csv) and the Wikipedia API, requires careful handling to ensure the integrity of the sources while being consistent and accurate. Furthermore, the presence of missing population data in the time-series form creates a challenge for creating continuity and reliability between data points.

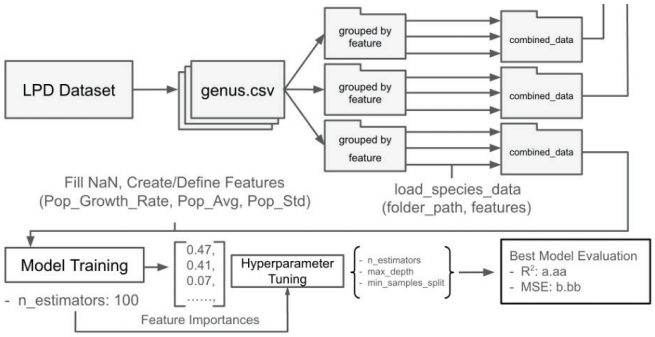

Figure 1: Overview of our data pipeline, illustrating the steps in data collection, preprocessing, and analysis. This pipeline was designed to ensure efficient integration of data into the predictive model, enabling reliable population trend analysis.

As shown in Figure 1, the initial phase of the data pipeline involves data collection, using the LPD dataset and the Wikipedia API. This stage establishes the foundation for subsequent analysis. Therefore, careful consideration must be given to selecting appropriate data sources and methods to ensure reliability.

To prepare the information for use in the Random Forest, the thousands of species needed to be grouped by features. This was addressed by first loading species data, which involved creating a function that would read CSV files from a specified directory. Each file contained data relating to a specific genus. Using the organized formatting of the files, important information such as species names, population counts over the years, and other relevant data was extracted.

The function "load_species_data" aggregated this information, where each dataset was transformed to a melted format, which would simplify analysis by merging data into columns for species, year, and population. Additionally, each entry was supplemented with details about its 'Blood_type' (thermoregulation), habitat, genus, and more, which were used in later stages of analysis.

This data preparation step was also supported by integrating information from the Wikipedia API. Augmenting the datasets with features such as thermoregulation and habitat enhanced the data further. The cross-referencing provided context for the numerical data and allowed the comprehensive analysis of the present ecological patterns. After loading and processing the data from the categorized species, these datasets were merged with species with similar groupings, and the resulting DataFrame was termed 'combined_data.' The merging process enabled the conduct of analyses spanning multiple species.

Data Cleaning and Imputation:

To ensure the quality of the data, missing values were addressed through strategies such as filling in missing data. The forward fill imputation was used to handle missing values in the population. This method propagates the most recent observed value forward to replace subsequent missing values, which is especially suitable for time-series data. By carrying forward the last known population value, the forward fill imputation is able to maintain continuity in the dataset without introducing completely unrealistic values, making it effective for the context.

We chose forward filling for this task to maintain continuity for species with many gaps in time-series data, however, the method may create bias if earlier values are not representative of later trends. Alternative techniques such as linear interpolation or KNN-based imputation were considered, but were not implemented due to data sparsity and computational cost. A comparative study of imputation methods could be a valuable direction for refinement in the future.

This preprocessing was essential as it prepared the data for subsequent statistical analyses and applications in machine learning. Furthermore, additional features, such as population growth rates, averages, and standard deviation, were added. These values were calculated for each species in the dataset, with $G = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100$ represented population growth rate, where P_t and P_{t-1} represent the population in year t and the previous year t-1, respectively. The average population across all available years was calculated by $P_{avg} = \frac{1}{n}\Sigma_{i=1}^{n} P_i$, where n is the total number of years for which population data is available, and Pi is the population in year i. Finally, the standard deviation was calculated with the equation:

$$\sigma = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(P_i - P_{avg}\right)^2}$$

Feature Engineering and Final Dataset:

Additionally, the data pipeline used feature engineering to create new variables based on existing data. For instance, calculating derived metrics such as population growth rates, averages, and standard deviations from the raw data counts enriches the dataset and provides additional context that may improve model performance. This was added in the hopes that the model's ability to capture patterns and relationships in the data would be enhanced. This approach helps the Random Forest model use a more comprehensive set of features, which leads to more accurate predictions and insights about species populations and ecological relationships.

Ultimately, the data collection and preparation approach, which spanned from the initial loading of species data to the handling of missing values, created a foundation for further analysis using the Random Forest model. At the same time, feature engineering enabled us to make more accurate predictions and insights.

Methods

Random Forest Model:

The Random Forest model was selected for its ability to manage complex, high-dimensional data commonly encountered in ecological research. Unlike linear regression models, which assume linear trends, Random Forest captures non-linear associations with traits and trends, which is essential for ecological data. For example, relationships between environmental factors and species populations are rarely linear and can involve intricate interactions that linear models would miss. While linear methods may be sufficient in other contexts, the complexity of ecological data makes Random Forest better suited for predicting population trends.

DOI: 10.36838/v7i8.15

Random Forest achieves this by using an ensemble of decision trees, each trained on random subsets of the data. This ensemble approach minimizes overfitting, a common issue with single decision trees, and enhances the model's robustness to variations in the data. Moreover, Random Forest is resistant to outliers, as errors from individual trees tend to offset one another when aggregated. These attributes make it highly effective for analyzing real-world ecological data.

Another key advantage is its interpretability. The feature importance metric highlights the traits most influential in shaping population trends, enabling targeted conservation strategies, a central goal of this study. Although techniques like neural networks and support vector machines (SVMs) are viable alternatives, they require extensive tuning and often lack the interpretability that Random Forest provides. Although techniques like neural networks and support vector machines (SVMs) are viable alternatives, they require extensive tuning and often lack the interpretability that Random Forest provides. Neural networks are well-suited for identifying patterns in large, unstructured datasets, and SVMs excel in high-dimensional spaces, but their limitations in transparency make them less ideal for this project.

In contrast, Random Forest ideally balances predictive power and interpretability for understanding the drivers of population trends based on ecological factors. For an analysis where the relationships between variables and species populations are complex and non-linear, Random Forest is the best fit. While other techniques may offer benefits in certain contexts, the strengths of Random Forest align most closely with the goals of this study.

Model Inputs and Feature Selection:

This process enabled the extraction of meaningful insights from the data. In this study, the initial step involved selecting features and defining the target variable. The features included physical characteristics, habitat, population growth rate, standard deviation, and population averages, while the target variable was population counts. These inputs allowed the model to learn the relationships required for accurate predictions. An 80-20 train-test split ensured sufficient data for training and evaluation.

Data Preprocessing and Standardization:

Before fitting the model, features were standardized using the StandardScaler from scikit-learn, which scales them to have a mean of 0 and a unit variance. It transforms each feature x according to the using $z = \frac{x-\mu}{\sigma}$, where z is the standardized value, x is the original value of the feature, is the mean of the feature, and σ is the standard deviation of the feature. Using the StandardScaler reduced the impact of any discrepancies in scale among features.

Hyperparameter Tuning:

The first iteration of the Random Forest Regressor used 100 trees and was trained on the scaled training dataset. To enhance model performance even further, hyperparameter tuning was executed using GridSearchCV, an approach that evaluates a range of different combinations of hyperparameters.

GridSearchCV is especially beneficial when optimizing model performance, as it allows for thorough exploration of hyperparameters. By specifying a grid of hyperparameters, it is able to automate the tuning process, which ensures that all potential configurations are considered. It begins with defining the parameter grid and continues to employ cross-validation to evaluate the model's performance for each combination of parameters. This involves separating the dataset into K subsets, where the model is trained on K-1 folds. This process is repeated for every combination of parameters, allowing for an accurate assessment of each configuration's performance. Finally, it aggregates the results, calculating the mean performance metrics and identifying the best set of hyperparameters.

The tuned hyperparameters included the number of trees, maximum tree depth, and minimum samples required to split a node. For optimization, the parameter grid specified ranges of [100, 200, 300] for tree count, [10, 20, 30] for maximum depth, and [2, 5, 10] for minimum samples to split a node. This process identified a configuration that maximized predictive accuracy while maintaining resistance to overfitting.

Table 2: 3D table displaying the combinations of hyperparameters used in the analysis. The optimal combination found by tuning is bolded for clarity, ensuring reproducibility of future results.

Number of samples	Number of trees	Tree depth limit = 10	Tree depth limit = 20	Tree depth limit = 30
2	100	100, 10, 2	100, 20, 2	100, 30, 2
	200	200, 10, 2	200, 20, 2	200, 30, 2
	300	300, 10, 2	300, 20, 2	300, 30, 2
5	100	100, 10, 5	100, 20, 5	100, 30, 5
	200	200, 10, 5	200, 20, 5	200, 30, 5
	300	300, 10, 5	**300, 20, 5**	300, 30, 5
10	100	100, 10, 10	100, 20, 10	100, 30, 10
	200	200, 10, 10	200, 20, 10	200, 30, 10
	300	300, 10, 10	300, 20, 10	300, 30, 10

Model Evaluation and Metrics:

Following the identification of optimal parameters depicted in Table 2, the model with the highest R^2, a measure of how well the model accounts for differences in observed data, was evaluated on the test set from the train-test split to evaluate its predictive capabilities. Performance metrics such as R^2 and Mean Squared Error (MSE) were calculated. R^2 is measured with $R^2 = \frac{SS_{res}}{SS_{tot}} = \frac{\Sigma(y_i - y_{pred})^2}{\Sigma(y_i - y_{mean})^2}$, where SS_{res} is the sum of squares of the differences between the real and predicted values, and SS_{tot} is the total sum of squares, or the difference between the actual values and the mean of the values.

MSE is measured with $\Sigma(y_i - y_{pred})^2$, where n is the number of observations, y_i are the actual values, and y_{pred} are the predicted values. Lower MSE values illustrate better model performance.

In addition to numerical metrics, a visualization of prediction errors was conducted through the generation of a confusion matrix, shown in Figure 2. The confusion matrix summarizes predictions across multiple categories, with rows representing actual categories and columns representing the predicted categories. The diagonal cells show correct predictions for each category, while the cells outside represent misclassifications.

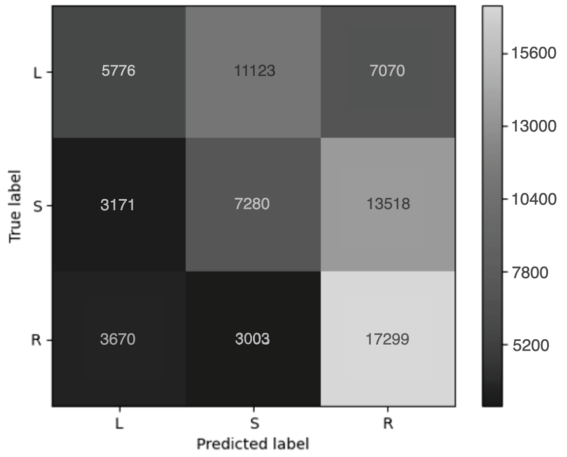

Figure 2: Confusion matrix showing the comparison of predicted and actual population values, categorized into three groups: Low (L), Medium (S), and High (R). This classification provides an alternative assessment of the model's performance in predicting population trends. Although correct predictions are present, the overall predictive power of the model is quite weak.

Results

The comparison of thermoregulation type, habitat, diet, reproductive strategy, and flight capability against statistical data such as species growth rate, average population across years, and standard deviation revealed minimal correlation with animal population trends.

The random forest regressor generates a Relative Importance Value (RIV). Calculated $RIV = \frac{\text{Importance}_{feature}}{\text{Importance}_{all}} \times 100$, the RIV value represents the importance of each feature to the predictions made by the model. This model calculates the RIV by measuring Gini impurity, which calculates feature importance and determines the final value. Gini impurity measures the likelihood of misclassifying a randomly chosen data point if labeled according to the class distribution. A feature's importance score is calculated based on its ability to decrease misclassification and improve decision-making in individual trees. In Random Forest, the RIV represents the average contribution of a feature to the reduction in Mean Squared Error (MSE) across all decision trees. Each split in the forest evaluates how much it reduces the MSE, and the features that result in greater reductions are assigned higher RIVs. The final RIV is the average reduction in error attributed to a feature across all trees, providing a robust measure of its importance in the model.

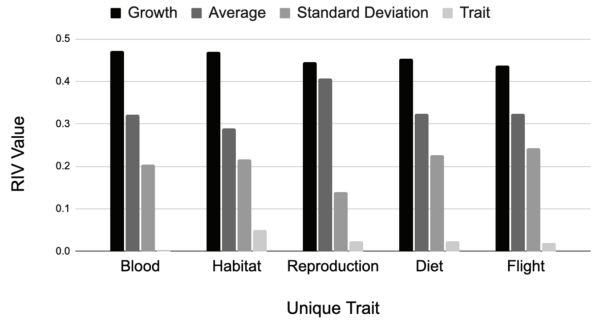

Figure 3: Relative Importance Values (RIV) for assessing the significance of five physical traits in predicting population trends. The RIV values are compared with statistical data to determine which traits most strongly correlate with population decline. The figure depicts low values for traits when compared to statistical data.

From the Random Forest regressor output, shown in Figure 3, the species' overall growth rate emerged as the most important feature in predicting future populations. It consistently held the highest RIV value compared to the other five features, indicating that species with stable or changing population trends are more likely to maintain these trajectories. Conservation efforts should prioritize species with declining populations, as they are at a higher risk of extinction. The standard deviation of populations over the years was then consistently the second most important feature. A higher standard deviation suggests greater fluctuations in population numbers, potentially signaling vulnerability to environmental changes or pressures.

The stronger performance of statistical features is likely since they encompass cumulative ecological effects over time. Growth rate and standard deviation reflect actual demographic responses to diverse pressures, such as habitat degradation, climate variability, or competition, without needing to explicitly model those factors. Physical traits, on the other hand, are more general, which means they encompass characteristics that may differently influence population trends under specific ecological conditions. As a result, while traits like thermoregulation and habitat are biologically meaningful, their predictive power is limited without environmental context, explaining why statistics outperformed trait-based variables.

In contrast, population growth rate, habitat, and thermoregulation features showed negligible RIV values, ranging from 0 to 0.05. This suggests that habitat changes may affect individual species rather than causing consistent changes across multiple species. The minimal RIV for thermoregulation indicates no significant correlation with future population trends.

The relatively low RIV value for habitat indicates that habitat changes do not consistently correlate with population changes across multiple species. Instead, these changes often alter species dynamics. This variability complicates the use of habitat as a reliable predictor for individual species populations, as seen in the application of Random Forest Regressor models. A relevant example is the extinction of gray wolves in Yellowstone National Park. The loss of this apex predator caused an unchecked increase in the elk population, which led to severe overgrazing. This vegetation loss degraded habitats, negatively impacting other species reliant on those ecosystems, illustrating the cascading effects of predator-prey dynamics on broader ecological systems.

The negligible RIV for thermoregulation indicates that population trends are likely shaped by combinations of factors rather than broad classifications alone. Simplistic classification risks overlooking significant variations in behavioral patterns, ecological adaptations, and physiological responses within these categories. For example, reptiles and fish may be classified together based on thermoregulation, but their population trends diverge due to significant differences in behavior and environmental sensitivities. Reptiles, such as lizards and snakes, often regulate body temperature by basking, influencing their activity and habitat use. Conversely, fish adjust their depth to manage temperature but are more vulnerable to water quality and temperature fluctuations, which directly affect

DOI: 10.36838/v7i8.15

breeding cycles and health. While reptiles are particularly susceptible to deforestation and its impact on food sources, fish face greater risks from aquatic environmental changes. These distinctions illustrate how population trends within thermoregulation-based groups can vary significantly. Relying solely on thermoregulation categories to assess populations may overlook critical factors such as competition for resources, predation pressures, and other environmental stressors. Species-specific traits and their interactions with complex ecological variables often provide more accurate insights into population dynamics than broad classifications.

Further analysis, however, reveals insights beyond just feature importance. First, the RIV values for certain features alone, such as habitat, do not exclude the possibility of interactions between variables. Although one feature alone may show minimal predictive influence, it could be more significant when combined with other features. For example, certain species in specific habitats may experience population changes due to environmental or ecological pressures that are not apparent when examining these features individually. Random Forest models can capture such interactions through decision tree splits. These splits divide data based on specific features or values, helping the model detect patterns. For example, a split could divide the data depending on whether the habitat type is "Marine." However, understanding these interactions requires more dedicated analysis techniques, including interaction effect plots or pairwise feature importance metrics, which assess the combined effect of two features. Using these tools could help clarify how features work together to influence predictions, which would create in-depth insights and likely inform conservation strategies that could address multiple factors simultaneously. This would provide a more comprehensive understanding of population patterns, especially for species that may be in complex environments.

While feature importance values provide interpretability in model behavior, we acknowledge their limitations in capturing causal relationships. An ablation study, where models are trained by removing features incrementally, was considered but not conducted due to the sparsity in our dataset and computational constraints. Future studies could expand on this study by the usage of ablation methods to evaluate feature combinations more accurately.

Regarding model accuracy, the stability of feature rankings across multiple Random Forest runs suggests consistent RIVs. This credibility further emphasizes features such as growth rate and average populations as important predictors. However, it is important to recognize any limitations in the application of RIVs.

Random Forests may favor features with more unique values, which could inflate their importance. To mitigate this, techniques like permutation importance can be used, which shuffle feature values to assess their true impact on prediction accuracy. Shuffling feature values could help in avoiding bias for abnormal values in the dataset, which would otherwise impact the importance values of features. From a conservation perspective, the findings suggest focusing on species with historically high population fluctuations, as the standard deviation indicates. These species may be more susceptible to minor environmental pressures. The high importance of population growth suggests that conservation efforts should target populations with declining trends, as these are likely to continue without intervention.

Discussion

The interconnectedness of species creates a complex web of interactions, making it challenging to discern overall population trends. Each species occupies a unique ecological role and responds differently to environmental pressures. For example, while large herbivores may thrive in the absence of predators, competing species or those dependent on vegetation face adverse effects, such as habitat degradation or reduced food availability. Understanding whether a species' population is increasing or declining requires a broader ecological context, as factors like competition, mutualism, and environmental changes significantly influence responses to habitat shifts.

The inclusion of historical data aimed to establish a baseline for understanding population trends, but it should not dominate the analysis when predicting future changes. The primary focus is on analyzing how physical traits, such as thermoregulation and habitat, influence species' adaptability and survival in dynamic environments. While historical population sizes provide a reliable foundation for estimating future sizes, they are less informative for identifying changes in population trends—the core objective of this study. The results suggest that excluding historical data in future iterations of the model may enhance its alignment with the study's goals. By focusing on the impacts of physical traits and emphasizing small deviations as early indicators for conservation efforts, the model can more effectively predict population trends and support targeted interventions.

Limitations of Physical Traits:

Although physical traits certainly influence a species' ability to survive in different conditions, they do not alone account for the shaping of population dynamics. For instance, while a group of species may have a consistent diet or reproductive strategy, such as being herbivorous or laying eggs, other aspects of the animals may be extremely varied due to other physical traits of the animal or different habitats. In this sense, the reasoning behind the low predictive scores for thermoregulation is also reflected when measuring the scores of flight capability, diet, and ways of birthing offspring. Each of these traits plays a role within a broader web of factors that influence population trends. For instance, while flight capability is crucial for certain species adapting to environmental changes, it does not address how populations are impacted by predator-prey dynamics or resource availability. Similarly, other traits often interact with environmental and ecological factors in ways that minimize or alter their individual influence on population trends. Thermoregulation affects sensitivity to climate variability, where ectothermic species are more susceptible to extreme temperatures. These examples illustrate how the predictive power of individual traits can be limited without considering their combined effects. Exploring such combinations could reveal

hidden vulnerabilities that are not apparent when examining traits in isolation, where methods like pairwise interaction analysis within Random Forest models could provide deeper insight into how physical traits shape population resilience or decline. This highlights the need to analyze a more comprehensive set of variables, emphasizing interactions among traits and environmental factors. Considering these combinations can provide a more accurate understanding of population dynamics and their broader ecological implications.

Implications:

The findings of this research highlight the importance of refining analyses to better understand how physical traits interact with environmental factors and population dynamics. Given the complexity of population trends, future studies should explore how traits like habitat interact with historical population data to improve the accuracy of predictions. Monitoring changes in population trends can help identify species at risk due to climate change, insights that may not be apparent from historical data alone. While historical population data is useful for predicting trends in stable conditions, physical traits become critical for understanding changes under increasing environmental pressures.

Future Research and Model Expansion:

This study also introduced methodological advancements. The data processing pipeline effectively managed extensive missing values, and trait data were sourced using the Wikipedia API and taxonomic groupings. These developments create an infrastructure for further exploration of similar topics, potentially supporting more effective and stronger analyses in future studies. Additionally, the flexibility of the Random Forest model allows for the integration of new features, traits, or population data to expand the model's range. Future research could incorporate environmental indicators, such as climate change or pollution data, to capture ecological interactions influencing population trends within specific groupings. The use of more detailed subcategories based on alternative or mixed physical and behavioral traits could further enhance the precision of population analyses.

Enhancing the pipeline with automated processes for feature selection and hyperparameter tuning by using grid search or other evolutionary algorithms could also optimize model performance as well as prevent the model from overfitting. Another potential improvement involves exploring ensemble methods, combining Random Forests with models like Dynamic Range Models (DRMs) or Individual-Based Models (IBMs). These hybrid approaches could better capture overall population trends while accounting for specific behavioral differences at the individual level. Continued development of this framework holds promise for creating more effective predictive tools. By incorporating advanced methodologies and diverse modeling approaches, future research can support more accurate analyses, guiding conservation efforts and species protection initiatives with greater precision.

Real Time Monitoring Applications:

The predictive framework developed in this study could be used to support real-time monitoring systems for conservation decision making. By using continuously updated population data from monitoring programs to feed directly into this model, population counts processed through the Random Forest algorithm would pre-emptively detect population decline based on species-specific traits. This would enable conservation managers to identify species with abnormal declines and prioritize interventions before substantial loss occurs. By automating this process, the model could support a real-time alert system for regions or species where consistent population tracking is available.

■ Conclusion

This study assessed the role of physical traits in predicting changes in animal populations, focusing on thermoregulation, habitat, diet, reproductive strategy, and flight capability. The results show that historical population data, including growth rate, standard deviation, and average population size, were crucial predictors of population trends. However, physical traits provided valuable insight into population changes, where small deviations in population trajectories can lead to significant ecological shifts. The Random Forest model demonstrated that the historical population data effectively predicted trends based on past patterns. While physical traits showed lower Relative Importance Values (RIVs) individually, their interactions with other environmental and biological factors may reveal more complex relationships. For instance, habitat may exert a stronger influence on population trends when analyzed alongside additional ecological variables, emphasizing the importance of assessing these interactions.

Although historical data is valuable for estimating population levels, the physical traits studied hold promise for identifying deviations from ongoing trends. Such deviations, even minor ones, can serve as early indicators of larger ecological or environmental changes, offering insights into species' long-term survival. Predicting changes in population trends requires a broader framework that integrates historical data with various physical and ecological factors for a more comprehensive understanding.

■ Acknowledgments

I would like to thank my mentor, as well as Dr. Matthew Caesar, a Computer Science professor at the University of Illinois Urbana-Champaign, for providing insights during the writing process and planning stages. Their guidance allowed me to shape my ideas into a more cohesive, well-structured paper.

■ References

1. Kaufman, M. These animals went extinct in 2023. *Mashable*, 2023, December 27. https://mashable.com/article/extinct-species-animals-2023 (accessed Dec 27, 2023).
2. Bongaarts, J. IPBES, 2019. Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of

DOI: 10.36838/v7i8.15

the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. *Population and Development Review*. 2019, *45* (3), 680–681. https://doi.org/10.1111/padr.12283.
3. An 83% decline in freshwater animals underscores the need to keep rivers connected and flowing. *World Wildlife Fund*. https://www.worldwildlife.org/stories/an-83-decline-of-freshwater-animals-underscores-the-need-to-keep-rivers-connected-and-flowing (accessed Dec 27, 2023).
4. White, P. J.; Garrott, R. A. Northern Yellowstone elk after wolf restoration. *Wildlife Society Bulletin* 2005, *33* (3), 942–955. https://doi.org/10.2193/0091-7648(2005)33.
5. Pacifici, Michela, et al. "Assessing Species Vulnerability to Climate Change." Nature Climate Change, vol. 5, 2015, pp. 215–224. https://doi.org/10.1038/nclimate2448.
6. Cardillo, Marcel, et al. "Multiple Causes of High Extinction Risk in Large Mammal Species." Science, vol. 309, no. 5738, 2005, pp. 1239–1241.
7. Qi, Y. Random Forest for bioinformatics. In *Springer eBooks*; Springer: New York, 2012; pp 307–323. https://doi.org/10.1007/978-1-4419-9326-7_11.
8. Moretti, M.; Legg, C. Combining plant and animal traits to assess community functional responses to disturbance. *Ecography* 2009, *32* (2), 299–309. https://doi.org/10.1111/j.1600-0587.2008.05524.x.
9. Almond, R. E. A.; Grooten, M.; Juffe Bignoli, D.; Petersen, T. (Eds.) *Living Planet Report 2022 – Building a nature-positive society*; WWF: 2022. https://wwf.panda.org (accessed Dec 27, 2023).
10. Johnston, A. S. A.; Boyd, R. J.; Watson, J. W.; Paul, A.; Evans, L. C.; Gardner, E. L.; Boult, V. L. Predicting population responses to environmental change from individual-level mechanisms: towards a standardized mechanistic approach. *Proceedings of the Royal Society B: Biological Sciences* 2019, *286* (1913), 20191916. https://doi.org/10.1098/rspb.2019.1916.
11. Pagel, J.; Schurr, F. M. Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography* 2011, *21* (2), 293–304. https://doi.org/10.1111/j.1466-8238.2011.00663.x.

Authors

Aiden Chee:

Aiden Chee is a tenth-grade student at Taipei American School. He is interested in environmental sciences as well as engineering and is passionate about preventing further damage within ecosystems due to a lack of awareness.

Jimin Choi:

University of Michigan, Ann Arbor, USA