Using Publicly Available Data to Model Six Species of Lepidoptera in Kruger National Park

Jennifer Sailor **May 10, 2023** **MSSC 6975** **Yu / Lemoine**

Climate change has and will continue to affect ecosystems and biodiversity. As our planet begins to change it is important to model how species environmental niches will shape future habitat availability. The environmental niche models will aim to assist conservation efforts of rare and common lepidoptera species. With this in mind, the goal of the Summer Practicum is to use publicly available datasets to model six (three common and three rare) species of lepidoptera (butterflies). I will focus on Kruger National Park (KNP), a 7,500 square mile national park located in eastern South Africa.

The project is a part of the MSSC 6975 Practicum in Statistics and Data Science course that will run from May 22nd to August 19th, 2023. Upon completion of the course, a 5-page summary report will be submitted that includes an outline of the work completed during the 12 weeks, a description of milestones, and an explanation of how the work contributed. In the Fall of 2023, I will give a 15-20 minute presentation over the same criteria to students and faculty of the Mathematical and Statistical Sciences (MSSC) Department at Marquette University. Over the 12 weeks, I will be mentored by Dr. Nathan Lemoine an Assistant Professor in the Biological Sciences department. With a focus on Trophic Ecology and Quantitative Ecology, he has published articles on Lepidoptera and machine learning techniques for ecological sciences. He currently contributes to KNP research. I have attended lab meetings for over a semester, started working in his lab in April, and will complete the work required for the practicum virtually from Kansas City. We will meet a minimum of once a week via Microsoft Teams for mentorship and project updates and then will be in contact via email for questions and concerns outside of our weekly meetings.

The data are obtained from GBIF (Global Biodiversity Information Facility) using an API (Application Programming Interface), which will include occurrence records of the six species of interest. To select the target species of interest, I referenced the Johan Kloppers 1978 book, "Butterflies of the Kruger National Park" to determine the rarity or commonness of each species within KNP, and then cross validated that assessment with GBIF occurrence reports for the continent of Africa and KNP. The top three highest occurrences in KNP were chosen for the three common species and the first three rare species with a high number of KNP occurrences and high difference in African occurrence in KNP occurrences were chosen for the three rare species. The rare species were chosen in this manor for the goal of seeing if there is a particular biological reason for the rarity in the park and what conversational efforts can be implemented to help increase their abundance. Therefore I will be focusing on *Bicyclus safitza*, *Eronia cleodora*, and *Cacyreues marshalli* as rare species and *Danaus chrysippus*, *Cynthia cardui*, and *Belenois aurota* which are all classified as common. All coding will be in Python due to its capabilities of working with GIS (Geographic information system) data as well as my desire to improve my experience in that particular language. The code will majority of the time be run through Dr. Lemoine's lab server due to the scale.

The first goal of the project is to quantify the species distribution of the six species both across Africa and within KNP. Based off those distributions I hypothesize that the habitat requirements of the species will be similar but not perfectly aligned with their distributions due to the potential observation bias near roads. Additionally, I will be interested in finding if there are any additional environmental predictors, such as soil characteristics or vegetation, that may contribute to their occurrences. By gaining a better understanding of the habitat requirements and distribution patterns of these butterfly species, we can then inform conservation strategist that will make a plan to protect their populations.

To predict how the distribution of butterfly species may change in the future due to climate change, we will use a combination of machine learning models such as: Logistic Regression, K-neighbors classifier, Gaussian Process Classifier, Decision Tree Classifier, Random Forest Classifier, Artificial Neural Network, Ada Boost Classifier, Naïve Bayesian Classifier. This approach would be similar to Dr. Lemoine's 2021 paper which used all nine models above to predict grasshopper species distributions in future climates. Likewise, my project will use the machine learning models to predict habitat suitability for the six species of butterflies in future climates based off of different climate scenarios. We know that climate change will affect the suitability of their habitat, but my hypothesis is that by 2050, the majority of the butterfly species will not be pushed out of KNP but rather shifted south further away from the equator to a different suitable habitat. However, it is important to note that rare species located near the southern border of the park will be at a greater risk of no longer being a part of the ecosystem of the park.

In conclusion, the 12 week practicum aims to develop species distributions and apply models to predict changed on habitat suitability on six species of butterflies. By using statistical approaches, I will gain information on how the species will be affected by the changing environmental conditions. This work hopes to inform conservation efforts in Kruger National Park, while giving MSSC faculty and students an example of how Statistical and Data Science methodologies can be applied to ecological issues.