Predicting Monarch Butterfly Sex

Stuart Sims

# Overview

The general goal of this project is to identify what features most reliably predict the sex class of the monarch butterfly, danaus plexippus. The aim is to narrow down the number of features to find those that are most relevant to classifying the sex given a wide spectrum of reflectivity data. There are approximately 1000 features in the dataset and a labeled feature for sex class.

## The Problem

There are two main problems given this data:

First, the spectral properties outside the visible spectrum of light is not known for Monarchs. Do they reflect the same, more, or less light in those areas as they do in the visible areas and do they display any sexual dimorphism?

Secondly, suppose that you wanted to reliably determine the sex class of a monarch butterfly from an image collected in the lab. Students may not be able to do so reliably and there are thousands of samples in cryogenic storage to process. Similarly, what if the most important features are not those that require extremely expensive and delicate imaging hardware and could be discerned from regular photographs? Identifying key features that may hold predictive power could help in feature selection among more commonly available and public data. A predictive model could aid in determining these types of characteristics. Similarly, it may be able to assist with wing fragments or partial specimens found in the field.

## Client Profile

Researchers studying Monarch populations would benefit from these findings, as they may have samples waiting to be processed for this very issue. The Altizer Lab, the Xerces Society for Invertebrate Conservation, North American Butterfly Association, USFS, University of Minnesota Monarch Lab, and several other entomological research labs come to mind as clients invested in similar research.

## Source Data

In remote sensing, reflectance represents the proportion of light that is returned from a target (i.e. reflects off of a surface). Certain wavelengths of light are associated with specific natural structures (like chlorophyll a and b concentrations in a leaf, leaf structure, amount of water, and so on). The reflectance at those wavelengths can quantify the presence, absence, or amount of a substance in an image. Remote sensing is often used after satellite, plane, or UAV collection methods. In this work, I set up an experiment to gauge reflectance of the dorsal view of monarch butterfly wings (outstretched) at a very close distance, comparatively. The intensity of the orangeness in monarch butterflies does vary between sex classes, on infectious status (whether the individual is parasitized by OE), and on life cycle stage (a summer-breeding individual vs a migrating individual). This data was all collected within the Deepak Mishra Lab at the University of Georgia School of Geography. It contains about 700 wavelength reflectance features and labeled sex data. The data is complete, so no imputation is needed for this specific data.

# Approach

## Data Collection

Sample specimens were provided by the Altizer Lab from The University of Georgia Odum School of Ecology. All Monarchs were deceased and some from cryogenic storage were provided as well. Sampling occurred in multiple sessions using a hyperspectral imaging camera known as the “SuperGIR.” The superGIR can collect reflectance of light at wavelengths ranging from 340 to 2500nm. Before each session, a calibration was performed against an industry-standard white test media. This calibration ensures reflectance values’ accuracy across sampling sessions. Due to the small imaging area, halogen lighting was utilized to boost the available light off the sample. Due to the warm-up time, the imaging area was left for 15 minutes for the lights to reach standard operating characteristics.

A small brass c shaped clip was used to ensure wings were fully outstretched for imaging. A baseline sample was also taken with the background and brass clip alone to ensure no additional noise was introduced to the reflectance data.

Each specimen was spread with a rear, dorsal view with the central axis of the specimen parallel and centered the central axis of the camera. Wings were kept outstretched using the small brass C shaped for each sample.

Each specimen was scanned three times. The scan number, specimen number, sex, and other categorical characteristics were recorded. Additionally, the camera captures a traditional image of the imaged area for each specimen. If any scan was disturbed or the individual was not centered properly in the imaging area, it was noted and discarded later. An additional sample was captured to ensure that each specimen had at least 3 samples.

Each set of scans was aggregated per specimen using an unweighted mean at each wavelength reflectance. The result is complete reflectance data across 96 Monarchs.

## Data Cleaning and Transformation

This foundational data contained reflectance values but also categorical data some categorical fields:

* ORIGINAL DATA SCAN #
* RECORDED SCAN #
* Monarch
* Group
* INFO
* INFECTED
* NOTES

These fields are not used in subsequent modeling at this time. The “Group” value “ORIENTATION TEST” was used to determine if the rotation of the specimen heavily influenced reflective characteristics. It contained multiple scans at multiple angles compared to the central axis of the camera and was removed first. Subsequently, these categorical fields were dropped from the dataset.

Some scan aggregations were missing specimen sex. These scan aggregations correspond to background samples of the area. Records with no “Sex” label were therefore dropped.

Finally, outlier detection was used to determine reflectance values outlier positions. Z scored outliers with a threshold of 3 standard deviations was investigated. Valid interpretation does require normality in each feature.

Inter-Quartile range was also used to determine outlier values in the dataset. Outlier thresholds were established per field at 1.5\* IQR on either side of the 25th and 75th percentile values. Ultimately, an ensemble method suited to handle outlers was used, so no steps were added to then remove the values identified as outliers at the cleaning stage.

## Exploratory Analysis and Inferential Statistics

The highest and lowest reflectance were identified among wavelengths. The standard deviation of each feature was explored to better understand where the features differed most. Subsequently, plotting each reflectance value across the spectrum of wavelength values (columns) allowed visual exploration of the reflectance “curves.” The mean reflectance at each feature for both sex classes as well as each individually was plotted to characterize where reflectance differed from the overall mean among male and female specimens. When coloring the line plots based on Sex, reflectance was visibly diverging starting at the orange/red wavelengths with males having higher reflectance values. Statistical testing was used to discern whether the mean reflectance in the blue, green, and red wavelengths differed significantly between sex classes. In all three tests, there was evidence to reject the null hypothesis at alpha = 0.1. Among visual portions of the spectrum, there are significant differences in reflectance. In this portion, dimensionality reduction was intentionally postponed. Because the primary goal is to identify features of interest, potentially reducing the feature space before modeling was counterproductive at this point.

## Other Potential Datasets

There are no other datasets that have the same experimental approach that could be found at this point. There was limited research into butterfly reflectance spectra, but they are not focused in Monarchs and the data was not available in any public-facing form. Unfortunately, it appears this is the only data available for this specific problem at the time.

However, I would hope that after sharing any findings with the lab and those labs whose equipment I was able to utilize, there may be some curious undergraduate or graduate students interested in capturing additional data for analysis!

## Current Findings

The reflectance in red, greed, and blue portions of the visible spectrum are significantly different and exhibit sexual dimorphism at a high likelihood (alpha = 0.01). Most features in the dataset are also non-normal which may be important in future testing.