Prevalence of *Haemoproteus plataleae* in American White Ibis from Southern Florida

Ryan Grunert

2021-11-12

#Loading required packages #for data processing  
library(here) #to set paths

## here() starts at /Users/ryangrunert/Desktop/Fall2021\_Classes/MADA/RyanGrunert-MADA-project

library(tidyverse) #all required data manipulation packages

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.5 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.4 ✓ stringr 1.4.0  
## ✓ readr 2.0.2 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(xlsx) #to read in excel spreadsheets  
library(janitor) #To assist in cleaning up dates and tidy data

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(lubridate) #To extract times from raw data

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(tidymodels) #to run the statistical analyses

## Registered S3 method overwritten by 'tune':  
## method from   
## required\_pkgs.model\_spec parsnip

## ── Attaching packages ────────────────────────────────────── tidymodels 0.1.4 ──

## ✓ broom 0.7.9 ✓ rsample 0.1.0   
## ✓ dials 0.0.10 ✓ tune 0.1.6   
## ✓ infer 1.0.0 ✓ workflows 0.2.3   
## ✓ modeldata 0.1.1 ✓ workflowsets 0.1.0   
## ✓ parsnip 0.1.7 ✓ yardstick 0.0.8   
## ✓ recipes 0.1.17

## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## x scales::discard() masks purrr::discard()  
## x dplyr::filter() masks stats::filter()  
## x recipes::fixed() masks stringr::fixed()  
## x dplyr::lag() masks stats::lag()  
## x yardstick::spec() masks readr::spec()  
## x recipes::step() masks stats::step()  
## • Search for functions across packages at https://www.tidymodels.org/find/

library(ranger) #implementation of random forests  
library(vip) #For plotting decision tree model

##   
## Attaching package: 'vip'

## The following object is masked from 'package:utils':  
##   
## vi

library(rpart) #For plotting machine learning models

##   
## Attaching package: 'rpart'

## The following object is masked from 'package:dials':  
##   
## prune

library(rpart.plot) #For plotting machine learning models  
library(png)  
set.seed(144)  
#Loading required materials - Figures  
data\_location5 <- here::here("processed\_data","CompleteIbisDataset.rds")  
figure\_loc\_4 <- here::here("results", "Figures", "DTTree.Rda")  
figure\_loc\_5 <- here::here("results", "Figures", "ImpPlotDT.Rda")  
figure\_loc\_6 <- here::here("results", "Figures", "LASSOFig.Rda")  
figure\_loc\_7 <- here::here("results", "Figures", "ImpPlotRF.Rda")  
figure\_loc\_8 <- here::here("results", "Figures", "RFImpPlotTest.Rda")  
figure\_loc\_9 <- here::here("results", "Figures", "RF\_Test\_ROCPlot.Rda")  
  
IbisC <- readRDS(data\_location5)  
DTTree <- readRDS(figure\_loc\_4)  
ImpPlotDT <- readRDS(figure\_loc\_5)  
LASSOFig <- readRDS(figure\_loc\_6)  
ImpPlotRF <- readRDS(figure\_loc\_7)  
RFImpPlotTest <- readRDS(figure\_loc\_8)  
RF\_Test\_ROCPlot <- readRDS(figure\_loc\_9)  
  
#Loading required materials - Tables  
mod\_loc\_1 <- here::here("results", "Models", "Model1PA.Rda")  
mod\_loc\_2 <- here::here("results", "Models", "Model2PA.Rda")  
mod\_loc\_3 <- here::here("results", "Models", "Model3PA.Rda")  
mod\_loc\_4 <- here::here("results", "Models", "Model4PA.Rda")  
mod\_loc\_5 <- here::here("results", "Models", "Model5PA.Rda")  
mod\_loc\_6 <- here::here("results", "Models", "Model6PA.Rda")  
mod\_loc\_7 <- here::here("results", "Models", "Model7PA.Rda")  
mod\_loc\_8 <- here::here("results", "Models", "Model8PA.Rda")  
mod\_loc\_9 <- here::here("results", "Models", "Model9PA.Rda")

# 1 Introduction

## 1.1 General Background Information

The data this analysis concerns is American White Ibis (*Eudocimus albus*) blood parasite data that was collected from 2010-2017(Coker et al., 2017). The parasite the data focuses on is the protozoan *Haemoproteus*, which causes a chronic blood infection in White Ibises and many other birds. *Haemoproteus* is a vector-borne disease, it is transmitted by *Culicoides* biting midges when they feed on the birds and causes long-lasting parasitemia. Adult healthy birds tolerate haemosporidian infections well, but young birds and birds outside their normal range that are naive to the pathogen are more likely to develop clinical disease. Haemosporidian infections usually cause long-lasting chronic infections in natural hosts, and can cause reduced reproductive success, host fitness, and increased stress and disease susceptibility.

The purpose of this analysis is to determine whether the presence of *Haemoproteus* was increasing over time in Ibis in southern Florida, and if so, determine any demographic trends among the Ibis or temporal or spatial trends among the system at large. Here we investigate whether age, sex, sampling site, capture season, date of capture, body condition, and/or mass of captured Ibis make significant differences on the overall presence of *Haemoproteus* in the birds.

# 2 Methods

## 2.1 Data Collection

The data was collected by various undergraduate and graduate students of Dr. Sonia Hernandez and Dr. Michael Yabsley at the University of Georgia. Wild American White Ibis were captured during multiple sampling seasons between 2010-2017 in various urban and natural habitat sites from southern Florida. Data was recorded in excel sheets that included information about the sampling sites (habitat type, location, etc.) and various characteristics about the captured Ibis (weight, sex, age, etc.). Ibis were bled and lab diagnostics were performed at the Southeastern Cooperative Wildlife Disease Study (SCWDS) at UGA in order to determine the presence of *Haemoproteus* parasitemia, and if possible, amount of parasitemia in the blood.

## 2.2 Data Cleaning

The data was organized in 2 excel sheets; one containing all the lab diagnostic data concerning the sampled Ibis and the other containing all the descriptive, demographic, and geographic data about the sampled Ibis. Both excel files were combined while removing uninformative and duplicate variables in both. Variables of importance were standardized once both datasheets were combined. If sampling season was not recorded but sampling date was for an Ibis, season was determined from the date and added. I dichotomized the recorded amount of *Haemoproteus* parasitemia to either “1” (presence) or “0” (absence) by whether or not the Ibis had a parasitemia value of above zero. Locations that did not have presence of *Haemoproteus* and had less than 5 observations with little demographic information were removed from analysis.

## 2.3 Data Analysis

Univariate relationships between the presence of *Haemoproteus* parasitemia and each predictor were assessed using logistic regression treating *Haemoproteus* parasitemia presence as a categorical outcome. Multivariate classification was also performed, determination of variables to include was performed by sequential forward floating selection. In addition, a Least Absolute Shrinkage and Selection Operator (LASSO) model, a decision tree, and a random forest model were all conducted with *Haemoproteus* parasitemia presence as the outcome variable. The goal of all models was maximization of accuracy and roc\_auc on test data using a 10 fold cross-validation (80/20 split, 5 times repeated).

## 2.4 Study Population

During all of the sampling seasons, 668 Ibis were sampled. Among those sampled, 355 Ibis had a positive *Haemoproteus* parasitemia value. 20 Ibis were removed from the analysis because they did not have a positive *Haemoproteus* parasitemia value and the site from which they were captured had less than or equal to 5 sampled Ibis and had no presence of infected Ibis. Ibis were either classified as having parasitemia (1 - presence) or not having parasitemia (0 - absence).

## 2.5 Exploratory Analysis

### 2.5.1 Tables

### 2.5.2 Figures

Now we can move onto creating some figures for the data to see any patterns visually.

Figure 1: Haemoproteus parasitemia from sampled infected Ibises from 2010-2017 by habitat type

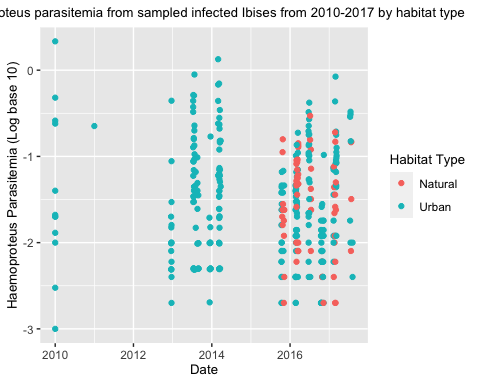
ExpFigure1 <-ggplot(IbisC) +  
 geom\_point(aes(x = Date, y = HaeParasitLog10, color = HabType)) +  
 ggtitle("Haemoproteus parasitemia from sampled infected Ibises from 2010-2017 by habitat type") +  
 xlab("Date") +  
 ylab("Haemoproteus Parasitemia (Log base 10)") +  
 scale\_color\_discrete("Habitat Type", na.translate = F) +  
 theme(plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10))  
  
Figure\_file1 <- here("results", "Figures", "ExpFigure1.png")  
ggsave(filename = Figure\_file1,plot = ExpFigure1)

## Saving 5 x 4 in image

## Warning: Removed 293 rows containing missing values (geom\_point).

ExpFigure1

## Warning: Removed 293 rows containing missing values (geom\_point).



This scatterplot graphs the sample dates chronologically against the parasitemia measured from the sampled ibis on a transformed log base 10 scale. Habitat type is depicted by color of the dots. Overall, there is a general narrowing of the data over time with a slight negative correlation. There isn’t a clear difference between natural and urban habitat types when it comes to amount of parasitemia.

Figure 2: Haemoproteus parasitemia from adult and juvenile sampled Ibises

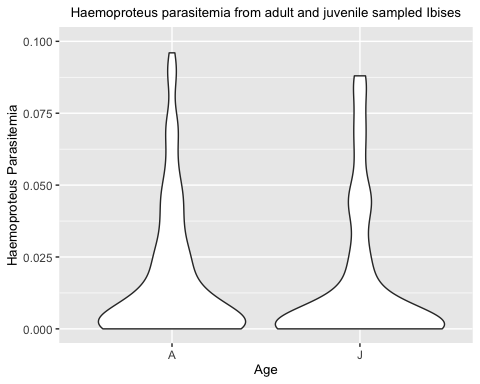
ExpFigure2 <- IbisC %>%  
 drop\_na(Age) %>%  
 ggplot() +  
 geom\_violin(aes(x = Age, y = HaeParasit)) +  
 ggtitle("Haemoproteus parasitemia from adult and juvenile sampled Ibises") +  
 xlab("Age") +  
 ylab("Haemoproteus Parasitemia") +  
 scale\_y\_continuous(limits = c(0, 0.1)) +  
 theme(plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10))  
  
Figure\_file2 <- here("results", "Figures", "ExpFigure2.png")  
ggsave(filename = Figure\_file2, plot = ExpFigure2)

## Saving 5 x 4 in image

## Warning: Removed 142 rows containing non-finite values (stat\_ydensity).

ExpFigure2

## Warning: Removed 142 rows containing non-finite values (stat\_ydensity).



This violin plot graphs the haemoproteus parasitemia distribution for the adult and juvenile Ibises sampled. The majority of the parasitemia values are on the lower end for both adults and juveniles. The distribution for adults reaches higher than juveniles. The overall structure for the adult distribution is higher and skinnier, while juveniles have more of their parasitemias lower in value.

Figure 3: Haemoproteus parasitemia from sampled Ibises by sex

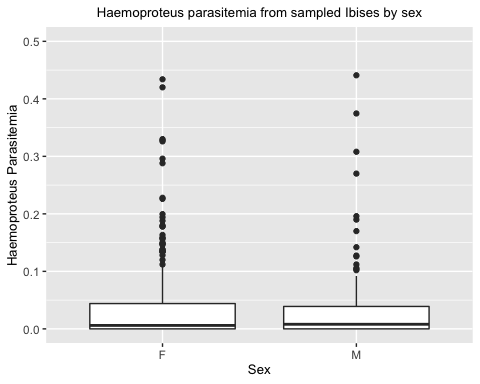
ExpFigure3 <- IbisC %>%  
 drop\_na(Sex) %>%  
 ggplot() +  
 geom\_boxplot(aes(x = Sex, y = HaeParasit)) +  
 scale\_y\_continuous(limits = c(0, 0.5)) +  
 ggtitle("Haemoproteus parasitemia from sampled Ibises by sex") +  
 xlab("Sex") +  
 ylab("Haemoproteus Parasitemia") +  
 theme(plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10))  
  
Figure\_file3 <- here("results", "Figures", "ExpFigure3.png")  
ggsave(filename = Figure\_file3, plot = ExpFigure3)

## Saving 5 x 4 in image

## Warning: Removed 21 rows containing non-finite values (stat\_boxplot).

ExpFigure3

## Warning: Removed 21 rows containing non-finite values (stat\_boxplot).



The haemoproteus parasitemia distributions between sexes is extremely similar. The median value for males is slightly higher than females. The majority of the measured parasitemias are between 0.0 and 0.1.

Figure 4: Haemoproteus parasitemia by total female and male Ibis mass

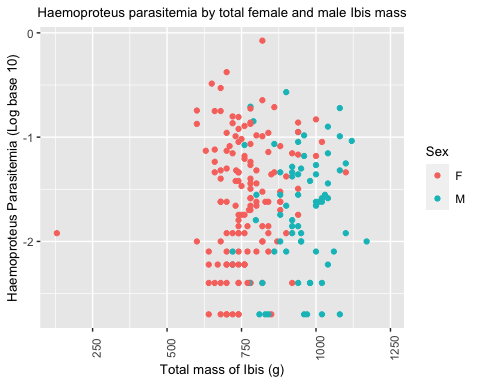
ExpFigure4 <- IbisC %>%  
 drop\_na(BirdMassG, Sex) %>%  
 ggplot() +  
 geom\_point(aes(x = BirdMassG, y = HaeParasitLog10, color = Sex)) +  
 theme(axis.text.x = element\_text(angle = 90)) +  
 scale\_x\_continuous(limits = c(130, 1240)) +  
 ggtitle("Haemoproteus parasitemia by total female and male Ibis mass") +  
 xlab("Total mass of Ibis (g)") +  
 ylab("Haemoproteus Parasitemia (Log base 10)") +  
 theme(plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10))  
  
Figure\_file4 <- here("results", "Figures", "ExpFigure4.png")  
ggsave(filename = Figure\_file4, plot = ExpFigure4)

## Saving 5 x 4 in image

## Warning: Removed 107 rows containing missing values (geom\_point).

ExpFigure4

## Warning: Removed 107 rows containing missing values (geom\_point).



Looking at the plot, male birds seem to weigh more than females. However, when comparing the parasitemia between females and males, there isn’t a clear difference between the two.

Figure 5: Haemoproteus parasitemia by female and male Ibis mass and body condition score

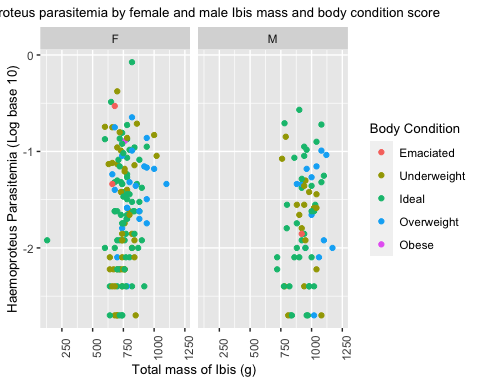
ExpFigure5 <- IbisC %>%  
 drop\_na(BirdMassG, BodyCondScore, Sex) %>%  
 ggplot() +  
 geom\_point(aes(x = BirdMassG, y = HaeParasitLog10, color = BodyCondScore)) +  
 facet\_wrap(~ Sex) +  
 ggtitle("Haemoproteus parasitemia by female and male Ibis mass and body condition score") +  
 xlab("Total mass of Ibis (g)") +  
 ylab("Haemoproteus Parasitemia (Log base 10)") +  
 theme(axis.text.x = element\_text(angle = 90),  
 plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10)) +  
 scale\_x\_continuous(limits = c(130, 1240)) +  
 scale\_color\_discrete("Body Condition", na.translate = F)  
  
Figure\_file5 <- here("results", "Figures", "ExpFigure5.png")  
ggsave(filename = Figure\_file5, plot = ExpFigure5)

## Saving 5 x 4 in image

## Warning: Removed 107 rows containing missing values (geom\_point).

ExpFigure5

## Warning: Removed 107 rows containing missing values (geom\_point).



Looking at this plot, there are more underweight female ibises with higher parasitemias compared to males. This could be due to the higher number of sampled females though.

Figure 6: Haemoproteus parasitemia by sample site and habitat type

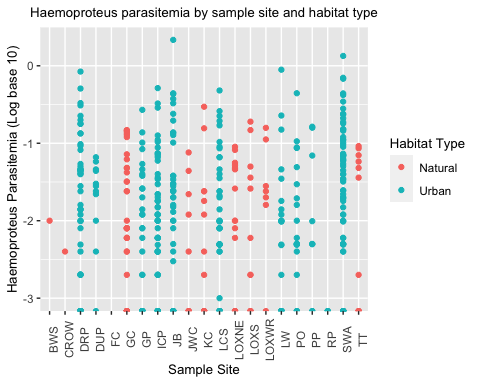
ExpFigure6 <- IbisC %>%  
 drop\_na(HabType, Site) %>%  
 ggplot() +  
 geom\_point(aes(x = Site, y = log10(HaeParasit), color = HabType)) +  
 theme(axis.text.x = element\_text(angle = 90),  
 plot.title = element\_text(hjust = 0.5, size = 10),  
 axis.title.x = element\_text(size = 10),  
 axis.title.y = element\_text(size = 10),  
 legend.title = element\_text(size = 10)) +  
 ggtitle("Haemoproteus parasitemia by sample site and habitat type") +  
 xlab("Sample Site") +  
 ylab("Haemoproteus Parasitemia (Log base 10)") +  
 scale\_color\_discrete("Habitat Type", na.translate = F)   
  
Figure\_file6 <- here("results", "Figures", "ExpFigure6.png")  
ggsave(filename = Figure\_file6, plot = ExpFigure6)

## Saving 5 x 4 in image

## Warning: Removed 115 rows containing missing values (geom\_point).

ExpFigure6

## Warning: Removed 115 rows containing missing values (geom\_point).



Looking at this plot broadly, it seems that the sites with a natural habitat have a slightly lower average parasitemia than the sites with an urban habitat.

Overall, there are no huge patterns or clear differences between covariates and variates so far. Statistical analyses may pick up on any significant differences later on, but I will return to the exploratory analysis as I continue with the project. I will also be meeting with those involved with the project to discuss as well.

## 2.6 Statistical Analyses

### 2.6.1 Univariate and Multivariate Logistic Models

This section details the results of a preliminary statistical analysis on the Haemoproteus parasitemia data just looking at the presence and absence data, as opposed to the amount of parasitemia per sampled bird.

The following models were setup and tested through a 10-fold cross-validation. The training/testing data proportions were 80/20. 10 folds were taken on 80% of the observations, and the results are depicted here. I am currently struggling with applying the models generated by CV to the testing data.

data\_locationPA1 <- here::here("results", "Tables", "Model1MetricsPA.Rda")  
table1 <- readRDS(data\_locationPA1)  
table1

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.569 50 0.00985 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.445 50 0.00734 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of habitat type on the presence of Haemoproteus in infected birds. The results were unsatisfactory.

data\_locationPA2 <- here::here("results", "Tables", "Model2MetricsPA.Rda")  
table2 <- readRDS(data\_locationPA2)  
table2

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.606 50 0.0112 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.534 50 0.0123 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of sex of the Ibis on the presence of Haemoproteus in infected birds. The results were unsatisfactory.

data\_locationPA3 <- here::here("results", "Tables", "Model3MetricsPA.Rda")  
table3 <- readRDS(data\_locationPA3)  
table3

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.584 40 0.0102 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.594 40 0.0122 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of age of the Ibis on the presence of Haemoproteus in infected birds. The results were unsatisfactory.

data\_locationPA4 <- here::here("results", "Tables", "Model4MetricsPA.Rda")  
table4 <- readRDS(data\_locationPA4)  
table4

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.639 50 0.0108 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.656 50 0.0125 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of sampling site on the presence of Haemoproteus in infected birds. The results were unsatisfactory.

data\_locationPA5 <- here::here("results", "Tables", "Model5MetricsPA.Rda")  
table5 <- readRDS(data\_locationPA5)  
table5

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.539 50 0.00927 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.549 50 0.00943 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of sampling date on the presence of Haemoproteus in infected birds. The results were unsatisfactory.

data\_locationPA6 <- here::here("results", "Tables", "Model6MetricsPA.Rda")  
table6 <- readRDS(data\_locationPA6)  
table6

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.535 50 0.00880 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.476 50 0.00754 Preprocessor1\_Model1

This univariate classification GLM looked at the effect of sampling season on the presence of Haemoproteus in infected birds. The results were unsatisfactory, however, these results were better than the other univariate models so far.

data\_locationPA7 <- here::here("results", "Tables", "Model7MetricsPA.Rda")  
table7 <- readRDS(data\_locationPA7)  
table7

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.627 50 0.0129 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.511 50 0.0178 Preprocessor1\_Model1

This multivariate classification GLM looked at the effect of age and sex on the presence of Haemoproteus in infected birds. The results were unsatisfactory, this model scored about the same as the previous univariate models for age and sex.

data\_locationPA8 <- here::here("results", "Tables", "Model8MetricsPA.Rda")  
table8 <- readRDS(data\_locationPA8)  
table8

## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.624 50 0.0107 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.668 50 0.0123 Preprocessor1\_Model1

ImpLoc <- here::here("results", "Figures", "ImpPlotDT.jpeg")

This multivariate classification GLM looked at the effect of sampling date and sampling season on the presence of Haemoproteus in infected birds. The results here had the highest performance of all the other models so far. One way to interpret this model would be that as the sampling effort continued from 2010-2017, the passage of time and season had a significant effect on the presence of Haemoproteus in Ibis. However, the performance of this model isn’t high enough to successfully conclude with that explanation.

#### 2.6.1.1 Decision Tree Model

The decision tree model was run on all the predictors against the *Haemoproteus* parasitemia presence and absence data. The model was run on all 10 folds of the training data. The roc\_auc of the best tree was 0.681 and the accuracy of the model was 0.631. The two predictors that had the most significant effect on the accuracy of the model were the sample date and mass of the sampled birds in grams. The mos tree with the best parameters had a maximum depth of 8 nodes.

#### 2.6.1.2 LASSO Model

#### 2.6.1.3 Random Forest Model

#### 2.6.1.4 Conclusions so far

This concludes the statistical analysis for part 3. Overall, the models did not perform well, especially those looking at the amount of Haemoproteus in infected birds. The presence/absence data has more potential for conclusions to be made based on generated models. The next steps would be to fine-tune the classification models for the presence/absence data and create a random forest model.

Current problems that have arose in this process include imputing missing values to run models (as there is a large amount of missing data in this dataset), and visualizing the tidymodels results on a graph.

Coker, S. M., Hernandez, S. M., Kistler, W. M., Curry, S. E., Welch, C. N., Barron, H. W., … Yabsley, M. J. (2017). Diversity and prevalence of hemoparasites of wading birds in southern Florida, USA. *International Journal for Parasitology: Parasites and Wildlife*, *6*(3), 220–225. <https://doi.org/10.1016/j.ijppaw.2017.08.003>