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

Ryan Grunert

2021-11-28

## 0.1 Abstract

Blood parasites of the *Haemoproteus* genus can infect various bird species and cause chronic infections that last an entire lifetime. A general increase in the presence of parasitic pathogens of birds led to the collection of blood samples from American White Ibis in Southern Florida from 2010-2017. A data analysis on the demographic, temporal, spatial, and parasitemia data collected from over 300 sampled Ibis was conducted. Various univariate and multivariate logistic regression models, along with a random forest model, concluded that the sampling date was the most important predictor in determining whether a sampled Ibis had a *Haemoproteus* infection (ROC-AUC 0.725, Accuracy 0.715). This led to the observation that more Ibis have *Haemoproteus* infections in Southern Florida in 2017 compared to 2010, however, there are numerous other epidemiologic and pathogenic factors about the parasite that makes any further conclusions difficult. An increased sampling effort and further studies are needed.

## 0.2 Loading Required Data and Packages

#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 3.1.6 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.4 ✓ stringr 1.4.0  
## ✓ readr 2.1.0 ✓ forcats 0.5.1

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

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 0.7.10 ✓ rsample 0.1.1   
## ✓ dials 0.0.10 ✓ tune 0.1.6   
## ✓ infer 1.0.0 ✓ workflows 0.2.4   
## ✓ 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()  
## • Dig deeper into tidy modeling with R at https://www.tmwr.org

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  
set.seed(144)  
#Loading in the complete Ibis Dataset  
data\_location1 <- here::here("processed\_data","CompleteIbisDataset.rds")  
IbisC <- readRDS(data\_location1)  
#Loading required materials - Figures  
#Exploratory Figures  
figure\_Exp\_1 <- here::here("results", "Figures", "ExpFigure1.png")  
figure\_Exp\_4 <- here::here("results", "Figures", "ExpFigure4.png")  
figure\_Exp\_6 <- here::here("results", "Figures", "ExpFigure6.png")  
#Machine Learning Model Diagnostic Figures  
figure\_ML\_4 <- here::here("results", "Figures", "ImpPlotRF.jpeg")  
figure\_ML\_5 <- here::here("results", "Figures", "ROC\_Test\_RF.jpeg")  
#Loading and preparing the Exploratory Analysis Tables  
IbisCount1 <- IbisC %>% count(Sex, Age, wt = HaeParasitPA)  
IbisCount2 <- IbisC %>% count(Season, wt = HaeParasitPA)  
IbisCount3 <- IbisC %>% count(Site, wt = HaeParasitPA)  
#Loading Model Results  
data\_locationPA4 <- here::here("results", "Tables", "Model4MetricsPA.Rda")  
model4 <- readRDS(data\_locationPA4)  
data\_locationPA5 <- here::here("results", "Tables", "Model5MetricsPA.Rda")  
model5 <- readRDS(data\_locationPA5)  
data\_locationPA8 <- here::here("results", "Tables", "Model8MetricsPA.Rda")  
model8 <- readRDS(data\_locationPA8)

## 0.3 Introduction

Avian haemosporidians are vector-borne parasites of the blood that have a global distribution (Clark, 2014). *Plasmodium* and *Haemoproteus* are two of the most common genera, however, *Haemoproteus* is much more common among avian hosts (Clark, 2014). They are believed to be the least pathogenic and have a narrow host range (Atkinson, 1991). Previous studies have shown that *Haemoproteus* parasitemia has been detected in Green Herons (*Butorides virescens*), Roseate Spoonbills (*Platalea ajaja*), Glossy Ibis (*Plegadis falcinellus*), and American White Ibis (*Eudocimus albus*) (Coker et al., 2017).

### 0.3.1 General Background Information

The data this analysis concerns is American White Ibis (*Eudocimus albus*) blood parasite data that was collected from 2010-2017. 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 (Coker et al., 2017). 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 (Coker et al., 2017). 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 (Coker et al., 2017). Lower parasitemia values suggest that the infections in the Ibis are chronic (Coker et al., 2017).

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 *Haemoproteus* ecology. Here we investigate whether age, sex, sampling site, capture season, date of capture, body condition, and mass of captured Ibis make significant differences on the overall presence of *Haemoproteus* in the birds.

## 0.4 Methods

### 0.4.1 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.

### 0.4.2 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.

### 0.4.3 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).

### 0.4.4 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).

### 0.4.5 Exploratory Analysis

Here we looked at various variate and covariate relationships between the predictor variables and the main outcome variable of presence of *Haemoproteus* in the sampled Ibis. The following tables and figures show those relationships prior to the statistical analysis. Tables and figures not included here are located in the supplemental file.

#### 0.4.5.1 Tables

IbisCount1

## Sex Age n  
## 1 F A 109  
## 2 F J 48  
## 3 F <NA> 29  
## 4 M A 51  
## 5 M J 32  
## 6 M <NA> 11  
## 7 <NA> A 60  
## 8 <NA> J 2  
## 9 <NA> <NA> 13

Overall, looking at the distribution of Ibis sex and age, the are more females than males and adults than juveniles. Female adults have the highest number with 109. Another clear observation from this table is that there are a large number of these missing variables in the dataset. This will most likely have to be completed with and without imputation depending on the model used. The results could be somewhat biased due to this distribution.

IbisCount2

## Season n  
## 1 Fall 2012 16  
## 2 Fall 2013 15  
## 3 Fall 2015 35  
## 4 Fall 2016 39  
## 5 Spring 2014 40  
## 6 Spring 2016 65  
## 7 Spring 2017 41  
## 8 Summer 2013 44  
## 9 Summer 2016 35  
## 10 Summer 2017 12  
## 11 <NA> 13

This table shows the number of Ibis sampled during each season. The majority of the Ibis were sampled in the later years (2015-2017) than in the earlier years (2010-2014). 13 samples did not have a recorded season. The highest number of Ibis sampled were in Spring, followed by Summer in the earlier years and Fall in the later years.

IbisCount3

## Site n  
## 1 BWS 1  
## 2 CROW 1  
## 3 DRP 40  
## 4 DUP 10  
## 5 FC 0  
## 6 GC 27  
## 7 GP 20  
## 8 ICP 48  
## 9 JB 31  
## 10 JWC 5  
## 11 KC 8  
## 12 LCS 34  
## 13 LOXNE 11  
## 14 LOXS 8  
## 15 LOXWR 6  
## 16 LW 15  
## 17 PO 17  
## 18 PP 7  
## 19 RP 0  
## 20 SWA 59  
## 21 TT 7

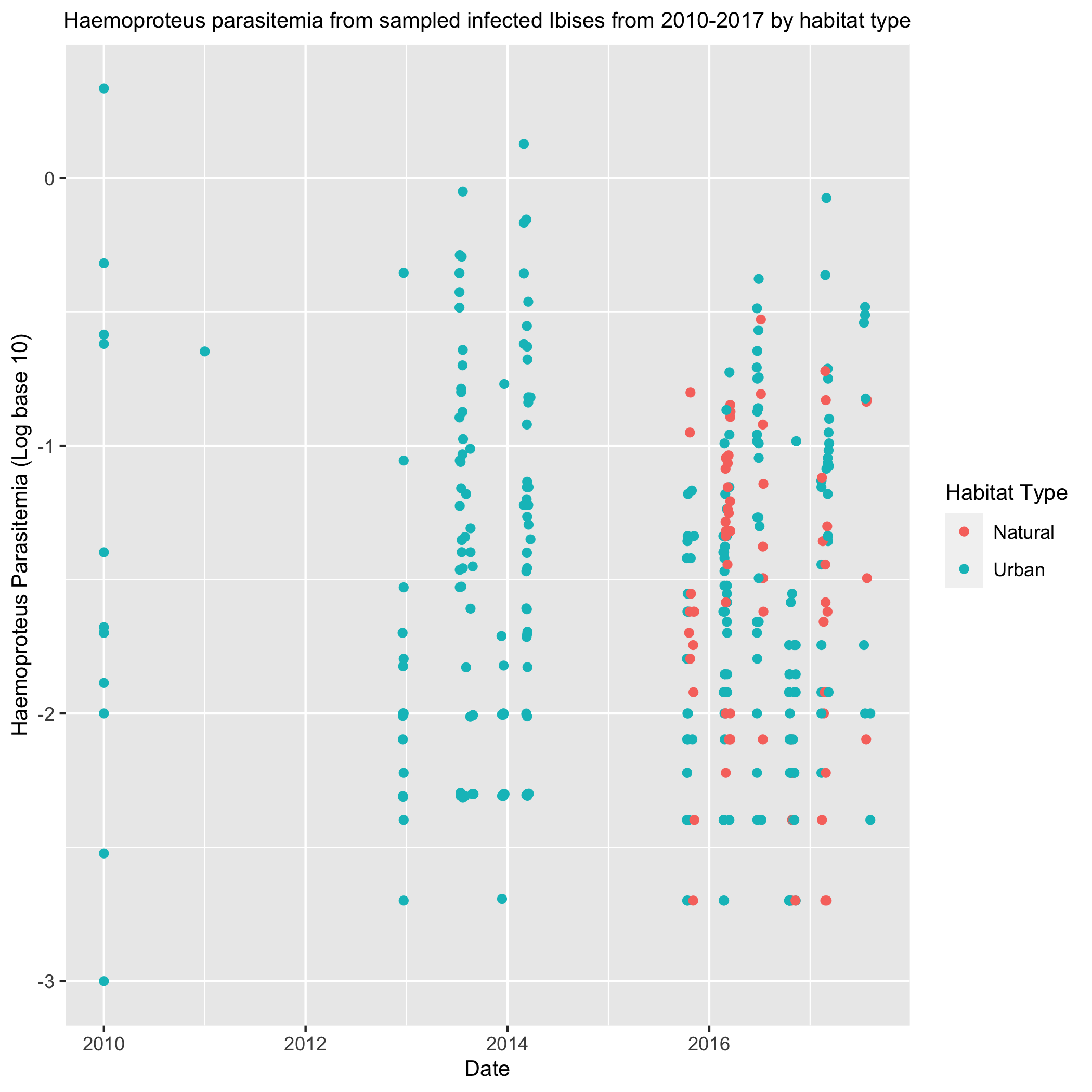
The majority of the Ibis were sampled in SWA, DRP, and ICP. Some sites had very low sampling numbers. Most of the Ibis were sampled in urban sites compared to natural sites.

#### 0.4.5.2 Figures

Now we can move onto creating some figures for the data to see any patterns visually. Haemoproteus parasitemia was transformed on a log base-10 scale for the exploratory figures here. This was conducted in order to see if there was any clear patterns in the amount of parasitemia among Ibis with *Haemoproteus* infection.

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

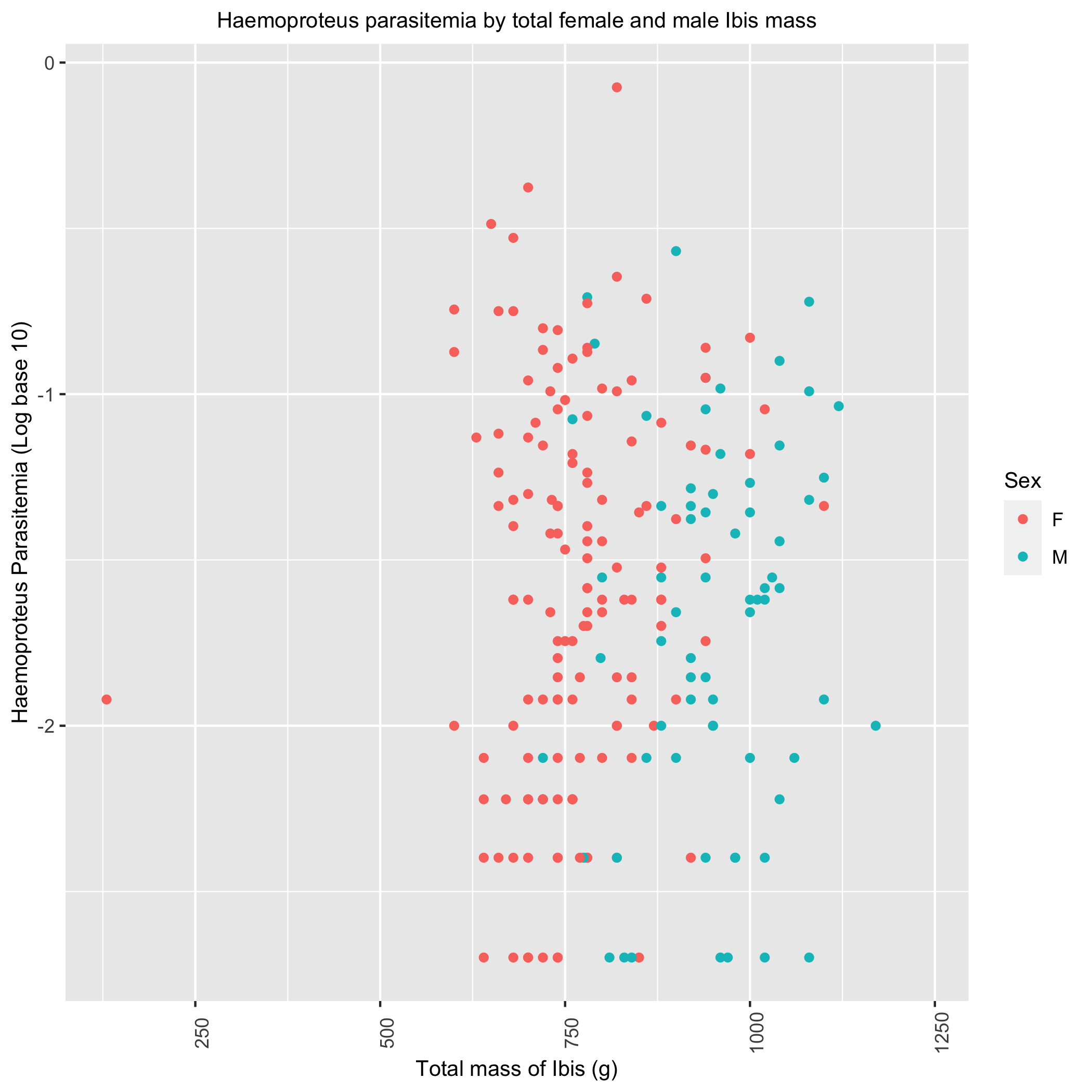
knitr::include\_graphics(figure\_Exp\_1, error = FALSE)



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 in the later years.

Figure 2: Haemoproteus parasitemia by total female and male Ibis mass

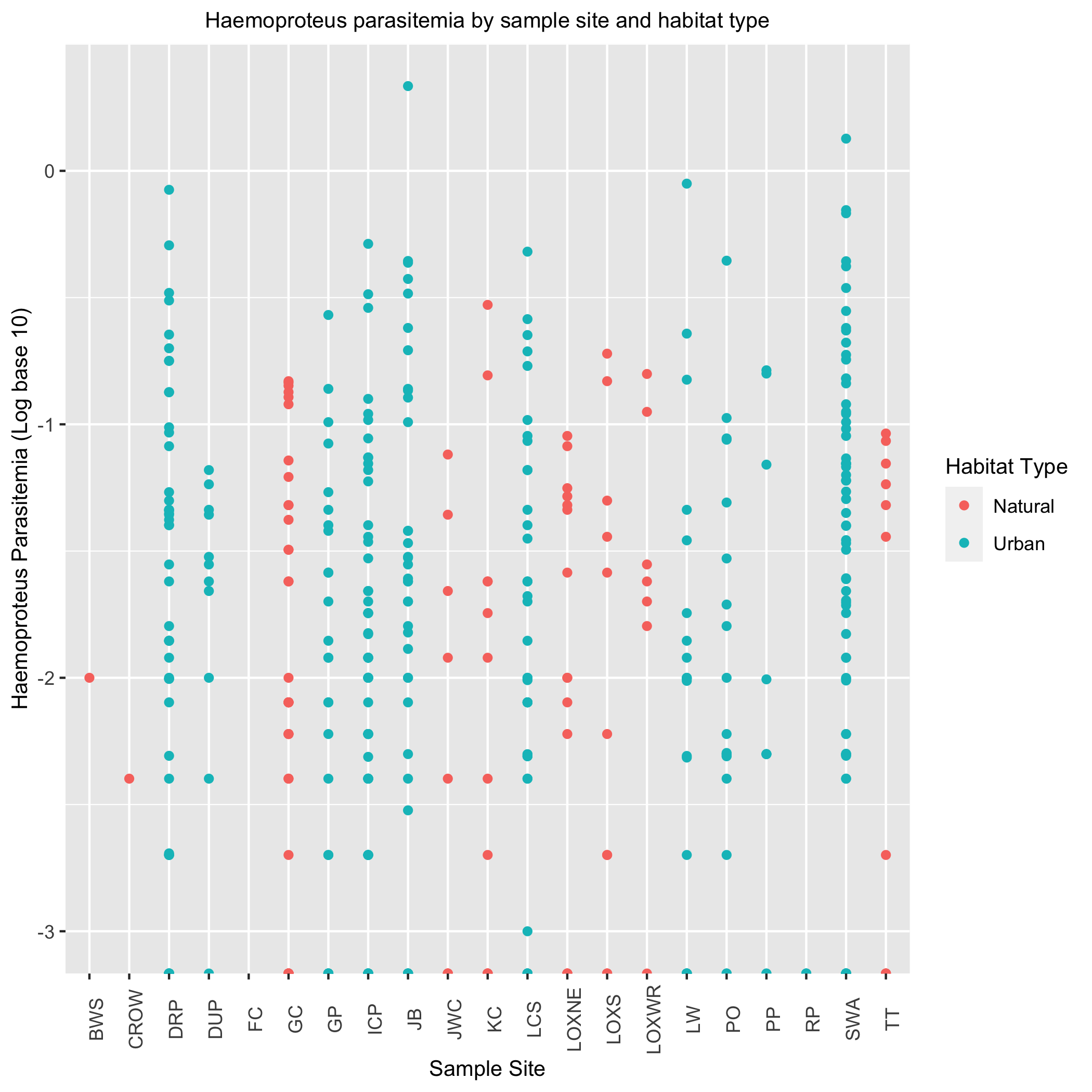
knitr::include\_graphics(figure\_Exp\_4, error = FALSE)



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. There seems to be an hourglass shape in the graph however, with majority of the parasitemia values either being a higher or lower value.

Figure 3: Haemoproteus parasitemia by sample site and habitat type

knitr::include\_graphics(figure\_Exp\_6, error = FALSE)



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.

### 0.4.6 Statistical Analyses

#### 0.4.6.1 Univariate and Multivariate Logistic Models

This section details the results of the 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 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. Models with variables not included here are included in the supplemental information.

model4

## # 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 logistic regression looked at the effect of sampling season on the presence of *Haemoproteus* in infected birds. The results were somewhat satisfactory with an accuracy of 0.639 and roc\_auc of 0.655.

model5

## # 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 logistic regression looked at the effect of sampling date on the presence of *Haemoproteus* in infected birds. The results were somewhat unsatisfactory with an accuracy of 0.539 and roc\_auc of 0.548.

model8

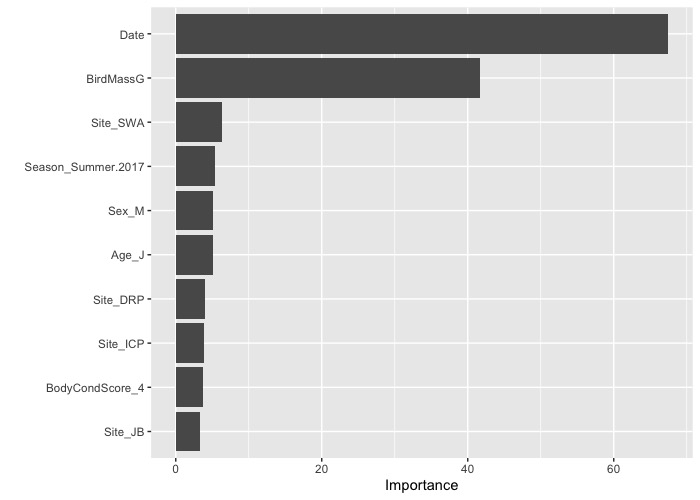
## # A tibble: 2 × 6  
## .metric .estimator mean n std\_err .config   
## <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 accuracy binary 0.624 50 0.0107 Preprocessor1\_Model1  
## 2 roc\_auc binary 0.668 50 0.0123 Preprocessor1\_Model1

This multivariate logistic regression model 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. 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 justify successfully concluding with that explanation.

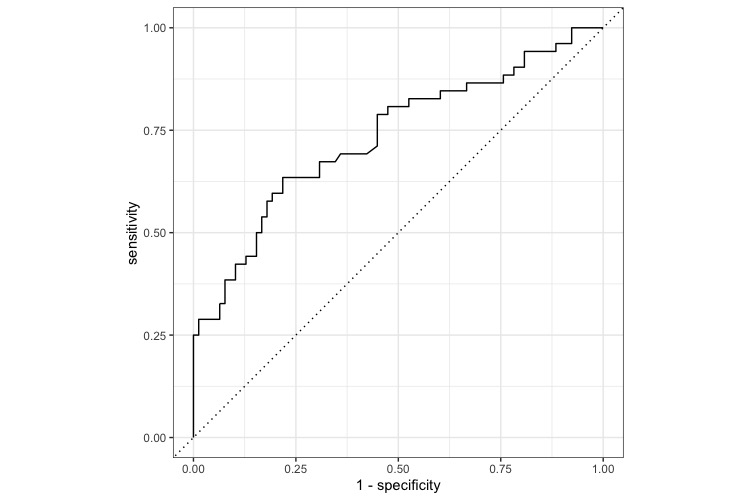
### 0.4.7 Random Forest Model

A random forest model was run on all the predictors against the *Haemoproteus* parasitemia presence and absence data. The number of features to consider at each split and the minimum number of data points required for the node to be split were both tuned with a tuning grid. The random forest model consisted of 500 different decision trees. The best model resulted in an roc\_auc of 0.712 on the training data.

knitr::include\_graphics(figure\_ML\_4, error = FALSE)

 The variable importance plot suggests that the sampling date and mass of the sampled bird in grams are the two most important predictors of *Haemoproteus* presence in the sampled Ibis.

The most accurate random forest model was run on the testing dataset, which resulted in an roc\_auc of 0.725 and accuracy of 0.715. The ROC curve for that random forest model is shown below.



### 0.4.8 Discussion

Based on the analysis conducted on the *Haemoproteus* dataset, sampling date and mass of the sampled Ibis are the most important predictors in determining whether the sampled Ibis will develop parasitemia. This study has shown that the number of *Haemoproteus* infections in Ibis are increasing in southern Florida from 2010 to 2017. The mass of the sampled Ibis is also an important predictor in determining whether or not the Ibis will develop an infection.

However, there are several other determinants in the *Haemoproteus* epidemiology and pathogenesis that can influence the development of infection. The results from this analysis aren’t powerful enough to successfully conclude that the mass of the birds is a good predictor. The life cycle of the *Culicoides* midges, pathogenesis of the *Haemoproteus* parasite itself in both the vector and host organism, and *Culicoides* biting rate could all have a more significant effect but can’t be directly applied to this dataset.

It may also be the case that the results of the analysis suggesting sampling date as the most important predictor is simply due to the higher number of Ibis being sampled in the later years of the study. The other individual predictors were all used in univariate logistic regression models, and none of them had highly significant results.

### 0.4.9 Conclusions

This concludes the statistical analysis for the *Haemoproteus* presence and absence data. Overall, there isn’t a clear predictor variable that determines whether or not sampled Ibis will have *Haemoproteus* parasitemia. The random forest model performed the best and had increased performance on the testing dataset. It suggested that sampling date was the most important predictor in determining whether the sampled Ibis would develop *Haemoproteus* parasitemia or not, but this may just be the case of their being a higher number of sampled Ibis in the later years. However, there are numerous other factors regarding the *Haemoproteus* epidemiology and pathogenesis that make any further conclusions difficult.

Atkinson, C., Carter T.; Van Riper. (1991). Pathogenicity and epizootiology of avian haematozoa: Plasmodium, leucocytozoon, and haemoproteus. *Oxford Ornithology Series*, 19–48.

Clark, S. M. ;. L., Nicholas J.; Clegg. (2014). A review of global diversity in avian haemosporidians (plasmodium and haemoproteus: Haemosporida): New insights from molecular data. *International Journal for Parasitology*, *44*, 329–338.

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>