# MSTISR001
# STA2007H
# Project 1

# Pollinating Fynbos Birds


## Abstract

This project investigates the population density of pollinating fynbos birds and its association with multiple diverse factors. The study aims to understand the influence of time since the last fire, vegetation type, presence of an alien species, altitude, and their interactions on bird density. The dataset includes information on bird density, time since the last fire, vegetation type, presence of alien species, and altitude. Univariate and multivariate exploration techniques were used to analyse the data. Linear regression models were constructed to examine the relationships between bird density and the explanatory variables. Models were then tested to identify which of them best fit the data. The model including vegetation type and interaction between time since the last fire and alien species demonstrated improved performance in explaining bird density.


## Introduction

A fictional study will be examined that investigates the density of pollinating birds. By analysing the provided dataset, the aim is to shed light on the relationships between bird density and variables such as fynbos patch size, distance to the nearest patch, altitude, time since the last fire, presence of alien vegetation, presence of key nectar-providing protea, and vegetation type.

## Statistical Methods

- Descriptive statistics: Provides insights on the variability and distribution of the data as well as helping in understanding the basic features of the variables and formulating the initial hypothesis.
- Correlation analysis: Is used to identify potential predictors for further analysis.
- Linear regression analysis: Allows estimation of the magnitude and significance of the effects of variables on bird density.
- Diagnostic plot analysis: Used to validate the regression models and identify potential violations of assumptions.
- Interaction analysis: The effect of one predictor variable on the response variable depends on the level of another predictor variable.
- Model selection: Adjusted R-squared and AIC will be used to determine which of the models fits the data the best.

## Data Exploration

### Univariate Exploration

| Variables | Minimum | Median | Mean | Maximum | Standard Deviation | Interquartile Range |
|---|---|---|---|---|---|---|
| Size (Hectares) | 482.300 | 4660.3 | 4576.1 | 7304.000 | 1467.603 | 2235.220 |
| Distance (Km) | 8.505 | 26.623 | 25.837 | 40.227 | 7.321 | 11.756 |
| Altitude (m) | 15.650 | 910.020 | 1016.130 | 2009.680 | 646.488 | 1448.173 |
| Time (years) | 2.560 | 9.700 | 9.497 | 17.599 | 3.212 | 4.280 |
| Density (Population per Hectare) | 1.082 | 22.410 | 22.187 | 42.147 | 9.267 | 14.082 |

**Table 1**

- Size:
  The smallest fynbos patch is 482.3 hectares. The median size of fynbos patches is 4660.300 hectares. The mean size of fynbos patches is 4576.100 hectares. It represents the average size of the patches in the dataset. The largest fynbos patch in the dataset has a size of 7304.000 hectares. The standard deviation of the fynbos patch sizes is 1467.603 hectares. The interquartile range is 2235.220 hectares, indicating the spread of the middle 50% of the patch sizes.

- Distance:
  The minimum distance to the nearest patch is 8.505 km. The median distance is 26.623 km. The mean distance is 25.837 km. This represents the average distance to the nearest patch and provides a measure of central tendency for the data. The maximum distance is 40.227 km. The standard deviation of 7.321. A higher standard deviation implies a wider spread of data points. The interquartile range is 11.756 km. A larger IQR indicates a greater dispersion of distances.

- Altitude:
  The minimum altitude recorded is 15.650 meters. The median altitude is 910.020 meters. The mean altitude is 1016.130 meters. The maximum altitude recorded is 2009.680 meters. The standard deviation of 646.488 suggests that the altitudes of the fynbos patches vary around the mean altitude. A higher standard deviation implies a wider spread of data points. The interquartile range is 1448.173 meters. A larger IQR indicates a greater dispersion of altitudes.

- Time:
  The minimum time recorded is 2.560 years. The median time is 9.700 years, indicating a moderate range of time since the last fire. The mean time is 9.497 years. The maximum time recorded is 17.599 years. The standard deviation of 3.212 suggests that the time intervals since the last fire vary around the mean time. The interquartile range is 4.280 years.

- Density:
  The minimum density recorded is 1.082 birds per hectare. The median density is 22.410 birds per hectare. The mean density is 22.187 birds per hectare. The maximum density recorded is 42.147 birds per hectare. This indicates that there are areas with a relatively high density of pollinating fynbos birds. The standard deviation of 9.267 suggests that the bird densities vary around the mean density. A higher standard deviation implies a wider spread of data points. The interquartile range is 14.082 birds per hectare. A larger IQR indicates a greater dispersion of bird densities.

## Relationships

- Density vs Altitude:

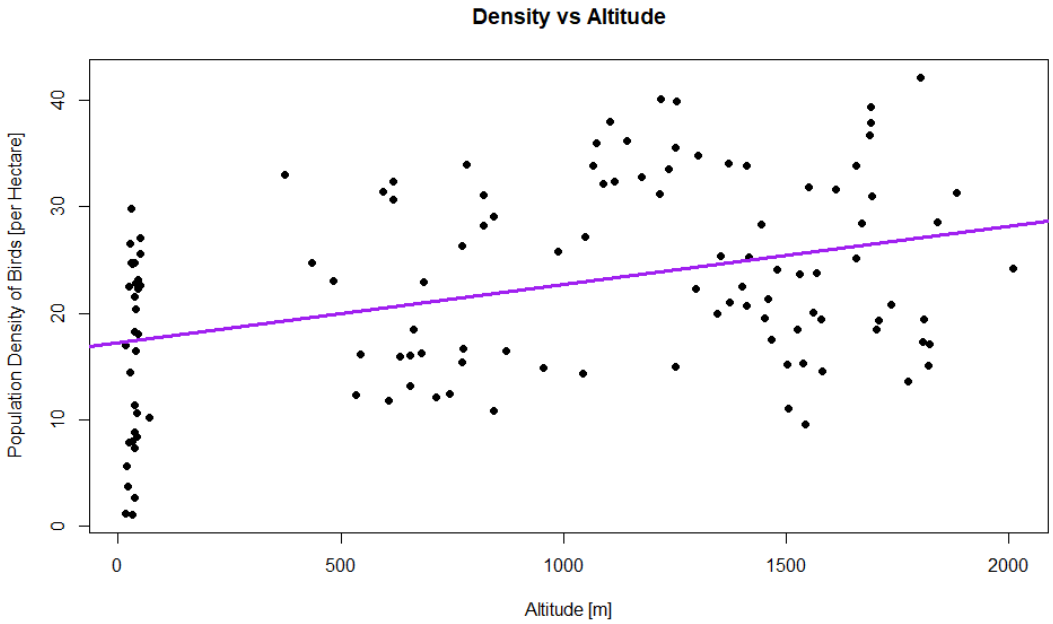


**Fig 1.1**

This is the second highest correlation(0.381), as altitude increases, there tends to be a slight increase in the density of birds, but the relationship is weak. This means that altitude alone may not be the most influential factor in determining the density of pollinating birds, as other variables may have stronger effects.

- Density vs Time:

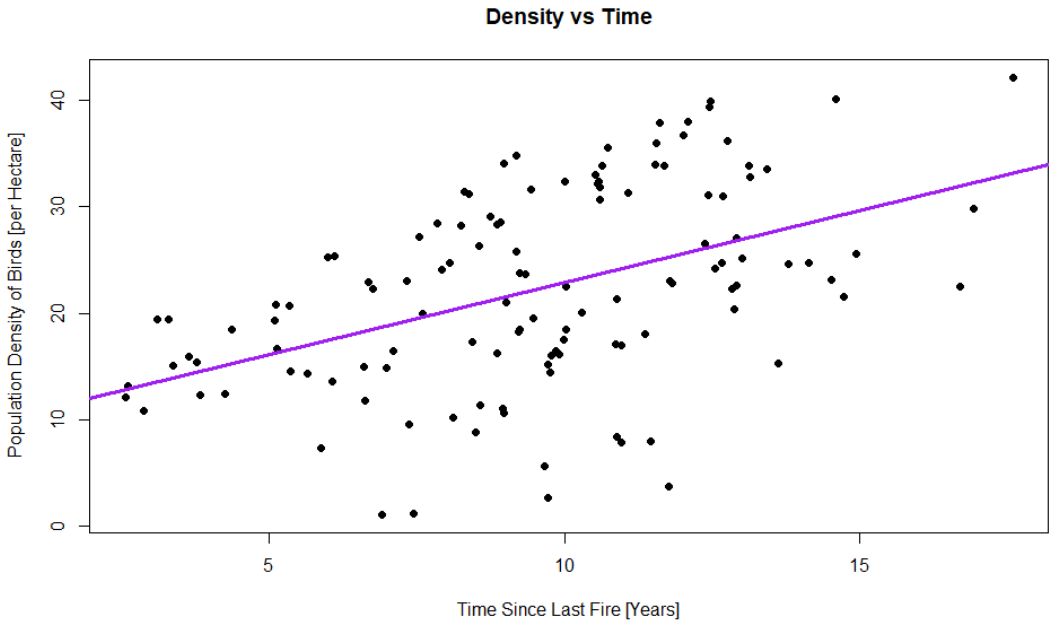


**Fig 1.2**

This has the highest correlation(0.468) out of the variables. As the time since the last fire increases, there is a moderate increase in the density of birds. This indicates that there may be a positive linear relationship. However, it is important to consider other variables in the analysis as well, as they may also contribute to the observed density patterns.

- Other relationships:
  The other relationships have a lot weaker correlations(Density vs Distance: 0.052 and Density vs Size: 0.140), this could be that there is little to no possibility that there are any linear relationships present, there could possibly be relationships other than linear that exist between them. Refer to section 1 of appendix for these graphs.

## Boxplots(Multivariate Exploration)

- Density and Alien Vegetation

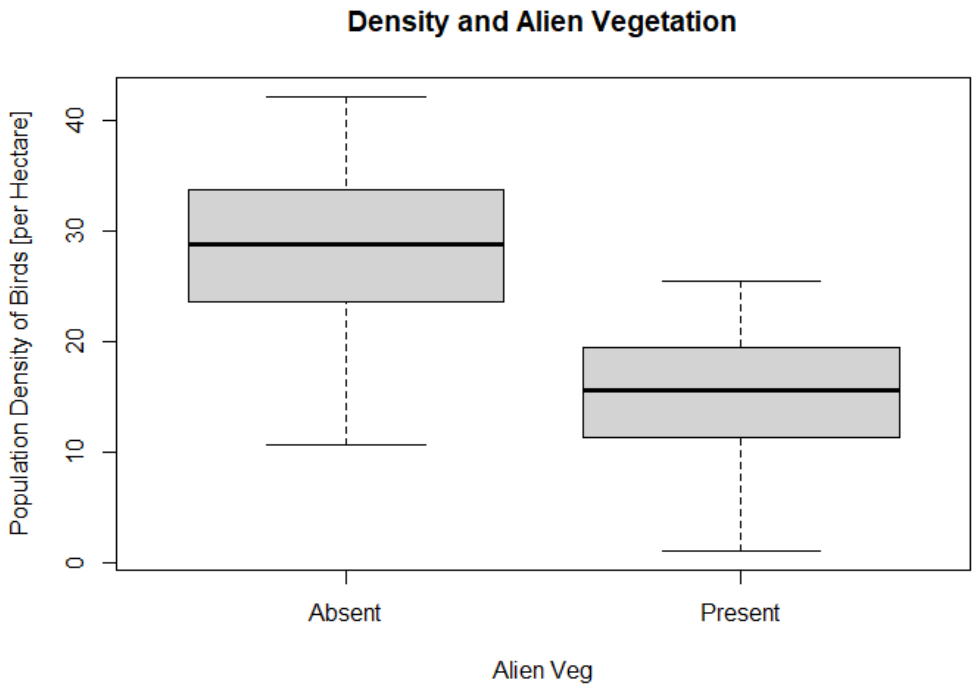


**Fig 1.5**

In areas without alien vegetation, the median bird density is higher compared to areas with alien vegetation. The interquartile range for areas without alien vegetation is wider, indicating a broader range of bird densities compared to areas with alien vegetation. The density range within areas without alien vegetation is larger compared to areas with alien vegetation. The boxes **Fig 1.5** do not overlap, suggesting some difference in bird densities between areas with and without alien vegetation. The lower median and narrower interquartile range in areas with alien vegetation indicate a potential negative influence on bird density.

- Density and Prescence of Protea

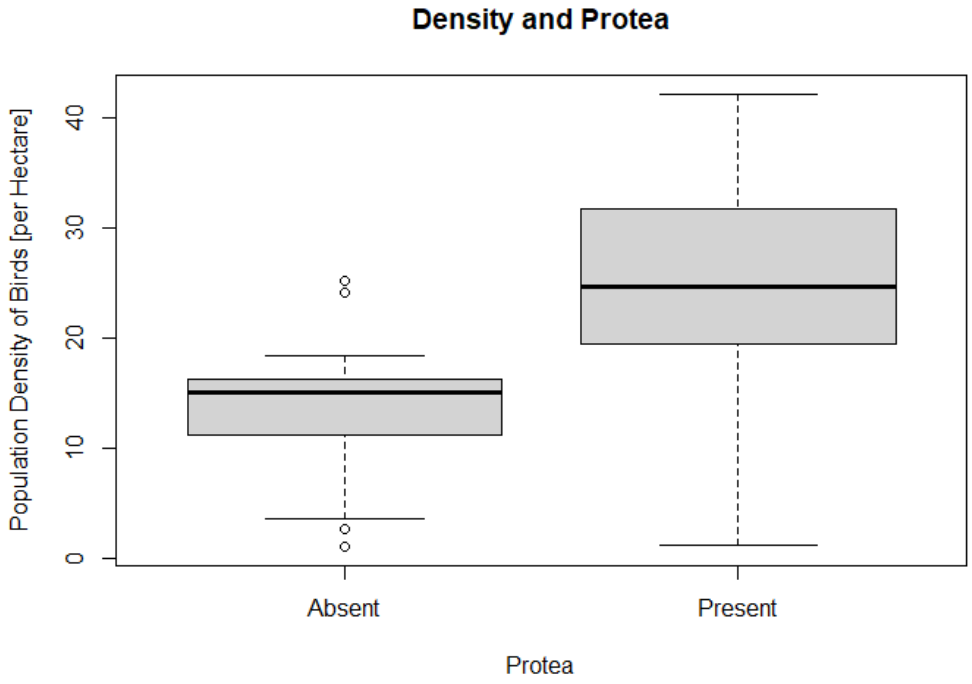


**Fig 1.6**

In areas without proteas, the median bird density is relatively low, compared to areas with proteas. The interquartile range for areas without proteas is narrower, this shows a smaller range of bird densities compared to areas with proteas. The density range within areas without proteas is smaller compared to areas with proteas. The boxes in **Fig 1.6** do not overlap, this suggests a difference in bird densities between areas with and without proteas. The higher median and wider interquartile range in areas with proteas indicate a potential positive influence on bird density.

- Density and Vegetation Type

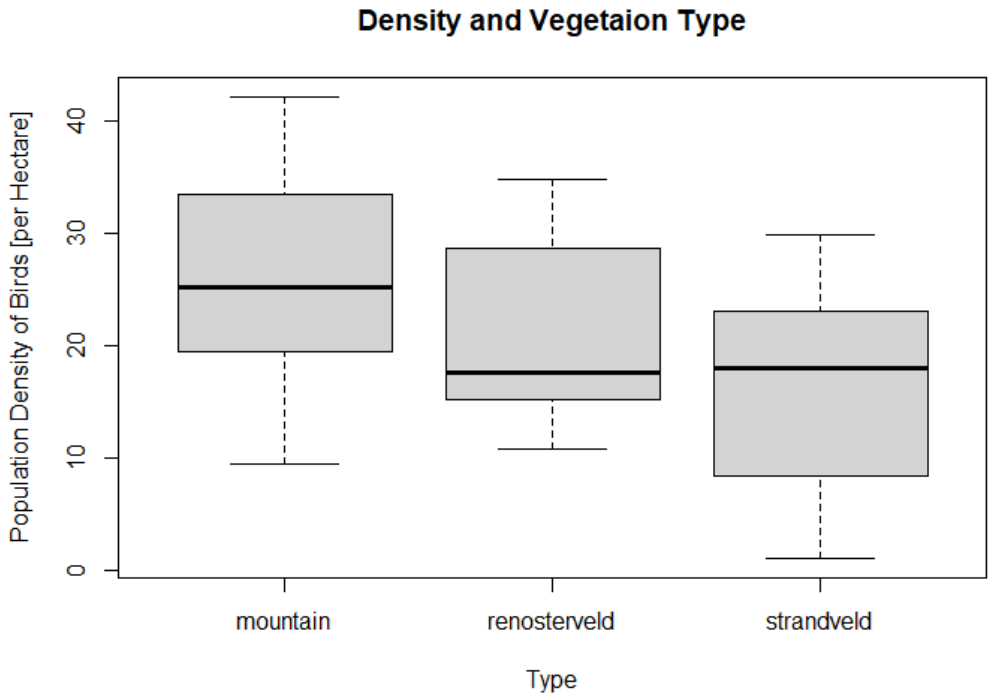


**Fig 1.7**

The median bird density is highest in the Mountain fynbos vegetation type, followed by the Renosterveld vegetation type, and lowest in the Strandveld vegetation type. The interquartile range for bird density is wider in the Mountain fynbos and Renosterveld vegetation types, compared to the Strandveld vegetation type. The density range within each vegetation type varies, with the highest range observed in the Mountain fynbos, followed by Renosterveld, and the lowest range in the Strandveld. The boxes in **Fig 1.7** overlap, indicating overlap in bird densities between different vegetation types. The highest median bird density in the Mountain fynbos suggests a potential preference or suitability of this vegetation type for supporting higher bird populations.

## Model Building

### $H_1$ : Food availability

$H_1$ will explore the possibility of the effect of the presence of key protea species and the type of vegetation. To test the hypothesis, a linear regression model will be constructed as followed:

$H_1$: Density = $\beta_o + \beta_1$Protea + $\beta_2$Vegtype

For $H_1$ it is assumed that Protea = absent and Vegtype = Mountain will be the references and since there are three categories in Vegtype, $H_1$ will look like this:

$H_1$: Density = $\beta_o + \beta_1$Protea + $\beta_2$Vegtype$_1$ + $\beta_2$Vegtype$_2$

| Coefficients | Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| $\beta_o$ | 16.235 | 1.525 | 10.644 | <2.000E-16 |
| $\beta_1$Protea | 11.800 | 1.524 | 7.743 | 3.750E-10 |
| $\beta_2$Vegtype$_1$ | -1.769 | 1.605 | -1.102 | 0.273 |
| $\beta_2$Vegtype$_2$ | -10.106 | 1.478 | -6.839 | 3.750E-10 |

**Table 2**

### $H_2$: Fire's role on the Fynbos

Bird density is driven by the length of time that has elapsed since the last fire, the type of vegetation, and whether there is alien vegetation present. It will be split into two hypotheses, one fitted with an interaction between time passed since last fire and the presence of alien vegetation, the other hypothesis won't have an interaction term. The linear regression models will be constructed as followed:

$H_{2a}$: Density = $\beta_o + \beta_1$Time + $\beta_2$Vegtype + $\beta_3$Alien

For $H_{2a}$ it is assumed that Vegtype = Mountain and that Alien = Absent will be used as reference. So $H_{2a}$ will be:

$H_{2a}$: Density = $\beta_o + \beta_1$Time + $\beta_2$Vegtype$_1$ + $\beta_2$Vegtype$_2$ + $\beta_3$Alien

| Coefficients | Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| $\beta_o$ | 25.689 | 1.784 | 14.400 | <2.000E-16 |
| $\beta_1$Time | 0.676 | 0.148 | 4.558 | 1.280E-05 |
| $\beta_2$Vegtype$_1$ | -2.744 | 0.907 | -3.026 | 0.003 |
| $\beta_2$Vegtype$_2$ | -12.718 | 0.822 | -15.472 | <2.000E-16 |
| $\beta_3$Alien | -12.306 | 0.854 | -14.412 | <2.000E-16 |

**Table 3**

$H_{2b}$: Density = $\beta_o + \beta_1$Time + $\beta_2$Vegtype + $\beta_3$Alien + $\beta_4$(Time × Alien)

The same assumptions are made from $H_{2a}$, so $H_{2b}$ will look like:

Density = $\beta_o + \beta_1$Time + $\beta_2$Vegtype$_1$ + $\beta_2$Vegtype$_2$ + $\beta_3$Alien + $\beta_4$(Time × Alien)

| Coefficients | Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| $\beta_o$ | 19.759 | 2.275 | 8.687 | 2.770E-14 |
| $\beta_1$Time | 1.202 | 0.195 | 6.167 | 1.050E-08 |
| $\beta_2$Vegtype$_1$ | -2.595 | 0.858 | -3.025 | 0.003 |
| $\beta_2$Vegtype$_2$ | -12.885 | 0.778 | -16.566 | <2.000E-16 |
| $\beta_3$Alien | -3.161 | 2.490 | -1.270 | 0.207 |
| $\beta_4$(Time × Alien) | -0.956 | 0.246 | -3.882 | 1.720E-04 |

**Table 4**

The interaction term allows us to examine whether the relationship between time since the last fire and bird density differs depending on the presence or absence of alien vegetation. The interaction term is the combined effect of time and alien species presence on density, providing insights into how these variables interact and influence bird density.

## $H_3$: Fragmentation

$H_3$ proposes that bird density is a linear function of the distance to the nearest patch and the size of the patch. The regression model for this hypothesis is as followed:

$H_3$: Density = $\beta_o + \beta_1$Distance + $\beta_2$Size

| Coefficients | Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| $\beta_o$ | 12.880 | 5.033 | 2.559 | 0.012 |
| $\beta_1$Distance | 0.153 | 0.123 | 1.242 | 0.217 |
| $\beta_2$Size | 0.001 | 6.136E-04 | 1.908 | 0.059 |

**Table 5**

## $H_4$: Environmental gradient

Bird density is driven by the presence of alien species as well as by the altitude of the patch. The hypothesis is split into two models, one fitted with altitude as a linear effect, and one fitted with altitude as a quadratic effect. The regression models are constructed as followed:

$H_{4a}$: Density = $\beta_o + \beta_1$Alien + $\beta_2$Altitude; it is assumed that Alien = absent is used as a reference.

| Coefficients | Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| $\beta_o$ | 22.760 | 0.760 | 29.960 | <2.000E-16 |
| $\beta_1$Alien | -14.810 | 0.814 | -18.210 | <2.000E-16 |
| $\beta_2$Altitude | 0.007 | 6.311E-04 | 11.270 | <2.000E-16 |

**Table 6**

$H_{4b}$: Density = $\beta_o + \beta_1$Alien + $\beta_2$Altitude + $\beta_3$Altitude$^2$

The same assumption is made that Alien = absent is used as a reference

| Coefficients | Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| $\beta_o$ | 20.770 | 0.770 | 26.992 | <2.000E-16 |
| $\beta_1$Alien | -14.940 | 0.729 | -20.493 | <2.000E-16 |
| $\beta_2$Altitude | 0.018 | 0.002 | 8.749 | 1.780E-14 |
| $\beta_3$Altitude$^2$ | -6.391E-06 | 1.158E-06 | -5.519 | 2.060E-07 |

**Table 7**

The use of a quadratic effect of altitude is sensible as it has a positive effect on the adjusted-R value going from 0.7703 to 0.8159. This means that model $H_{4b}$ better fits the data compared to $H_{4a}$ and that it explains a larger proportion of the variance in the dependent variable while accounting for the complexity of the model.

More details for model building can be found in section 2 of the appendix.

## Model Selection

### $H_1$ : Food availability

The adjusted R-Squared value is 0.456, this means that the model $H_1$ only describes around 45.6% of the variability of density. This is low and shows us that it is not a good model to fit the data.

### $H_2$: Fire's role on the Fynbos

$H_{2a}$ has an adjusted R-squared value of 0.8421, which means that about 84.21% of the variability in density is explained by $H_{2a}$. This value is high, this indicates that more tests will need to be conducted to confirm if it is the best model to fit the data.

$H_{2b}$ has a slightly higher adjusted R-squared value of 0.859, so 85.9% of the variability in density is explained by $H_{2b}$, this is the highest in the study and is an indication that this model is the model that best fits the data.

### $H_3$: Fragmentation

$H_3$ has the lowest adjusted R-squared values in the study, 0.016, which is only 1.6% of the variability in density being explained. This shows that $H_3$ is probably not a good model to fit data.

### $H_4$: Environmental gradient

$H_{4a}$ has an adjusted R-squared value of 0.7703, this is not bad but there are models with higher values in the study.

$H_{4b}$ has a slightly higher value than $H_{4a}$, after adding the quadratic effect of altitude, the adjusted R-squared is 0.8159, which is high but not as high as others. Conducting another test will solidify the model that best fits the data.

| Model | AIC | AIC Weight |
|---|---|---|
| $H_1$ | 821.138 | 4.327E-36 |
| $H_{2a}$ | 671.188 | 1.576E-03 |
| $H_{2b}$ | 658.285 | 0.998 |
| $H_3$ | 892.467 | 1.404E-51 |
| $H_{4a}$ | 714.963 | 4.918E-13 |
| $H_{4b}$ | 688.952 | 2.188E-07 |

**Table 8**

In the AIC weighted table(**Table 8**), it is evident that $H_{2b}$ has the lowest AIC value and the highest AIC weight being close to one, this further supports that $H_{2b}$ is the model that best fits the data of the study.

However, when conducting a residual plot for $H_{2b}$ it shows that for the values 4, 44 and 63 are all below the y = 0 line, this means that for these values the model is overestimating the predicted values and hints that in general, the model could overestimate values of density.

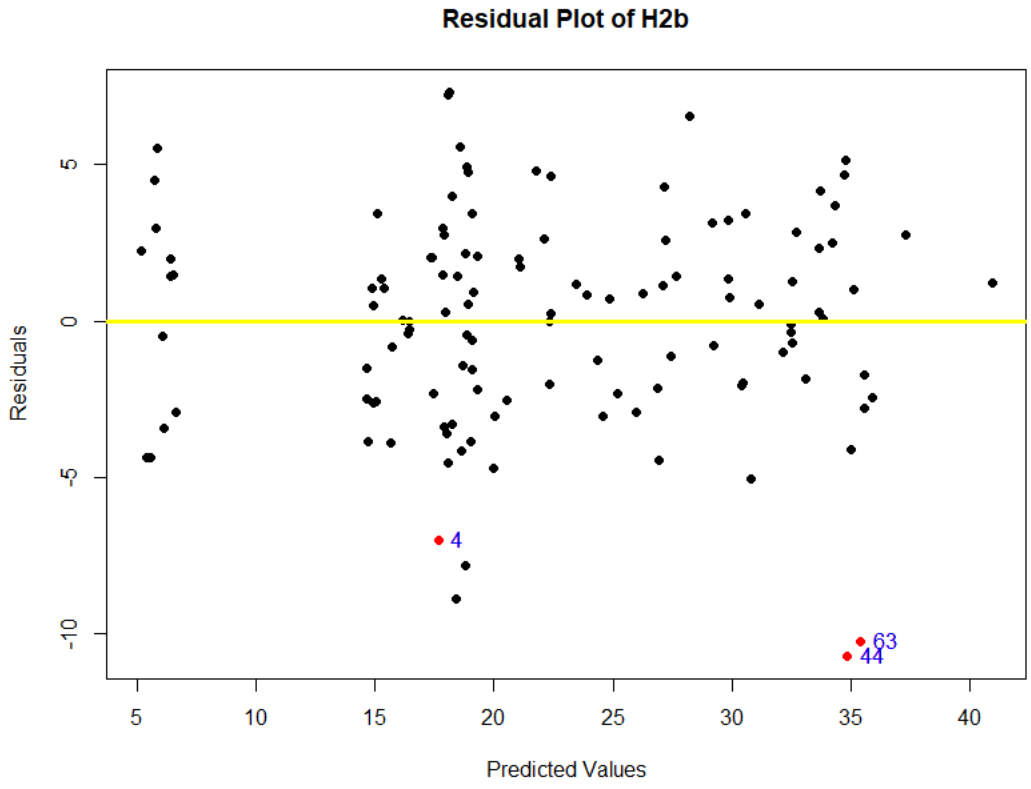


**Fig 3.1**

When testing the second and third best models, $H_{2a}$ and $H_{4b}$ respectively, it is found that there is a similar outcome, so all models will in general overestimate density of birds.(See **section 3** of appendix for other relevant graphs)

## Interpretation of $H_{2b}$

### Coefficients

| Coefficients | Estimate | Std. Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| $\beta_o$ | 19.759 | 2.275 | 8.687 | 2.770E-14 |
| $\beta_1$Time | 1.202 | 0.195 | 6.167 | 1.050E-08 |
| $\beta_2$Vegtype$_1$ | -2.595 | 0.858 | -3.025 | 0.003 |
| $\beta_2$Vegtype$_2$ | -12.885 | 0.778 | -16.566 | <2.000E-16 |
| $\beta_3$Alien | -3.161 | 2.490 | -1.270 | 0.207 |
| $\beta_4$(Time × Alien) | -0.956 | 0.246 | -3.882 | 1.720E-04 |

**Table 4**

- $\beta_o$ = 19.7593(intercept) represents the estimated density when all predictors are zero.
- $\beta_1$Time = 1.2023 indicates that for every year, the density of pollinating birds increases by approximately 1.2023, holding other predictors constant.
- $\beta_2$Vegtype$_1$ = -2.5947 suggests that the presence of renosterveld is associated with a decrease in the density of pollinating birds, compared to other vegetation types, holding other predictors constant.
- $\beta_2$Vegtype$_1$ = -12.8852 indicates that the presence of strandveld is associated with a significant decrease in the density of pollinating birds, compared to other vegetation types, holding other predictors constant.
- $\beta_3$Alien = -3.1610 suggests that the presence of alien species has a negative effect on the density of pollinating birds, although it is not statistically significant at the conventional significance level [p-value: 0.206783].
- $\beta_4$(Time × Alien) = -0.9559 shows that the interaction between time and alien presence has a significant negative effect on the density of pollinating birds.

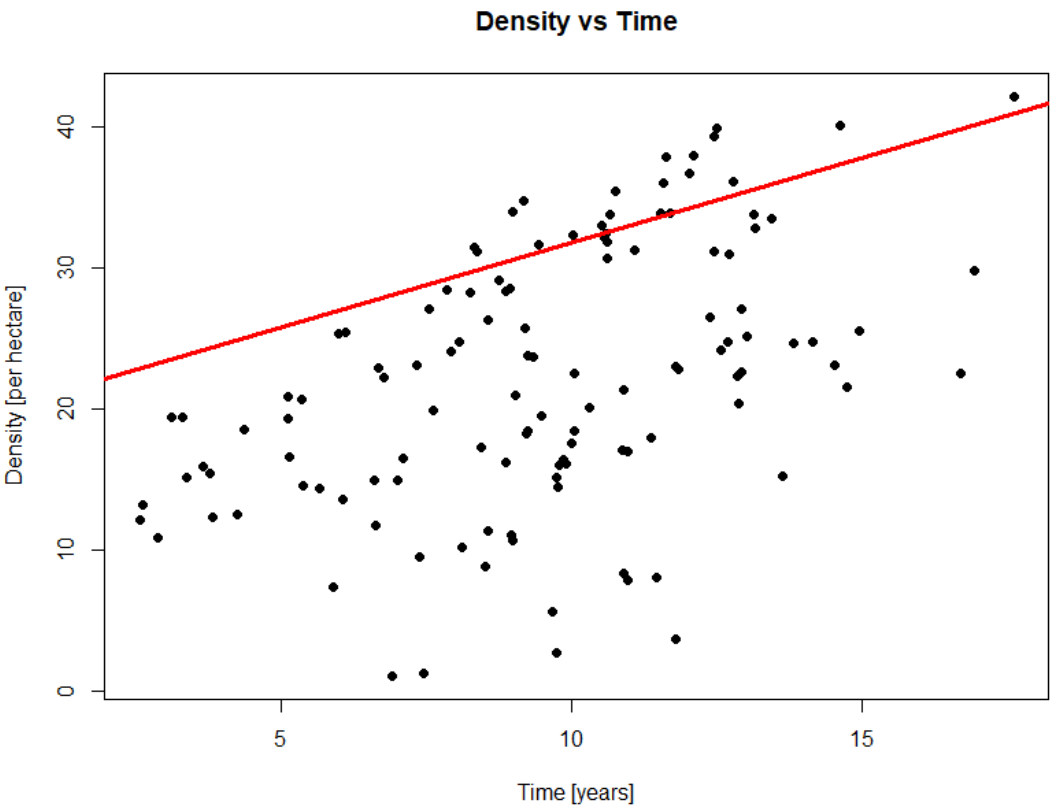


**Fig 4**

In **Fig 4**, the regression line for $H_{2b}$ is applied and it follows the established trend of density increasing as time increases. The line being higher than most of the data points is more evidence showing that the model is overestimating values.

## Discussions & Conclusions

In conclusion, the model that best fits the data in this study is:

$H_{2b}$: Density = $\beta_o + \beta_1$Time + $\beta_2$Vegtype + $\beta_3$Alien + $\beta_4$(Time × Alien)

However, the residuals show that there could be a better model that will more accurately predict the data that will need to be established.

# Appendix

## Section 1:

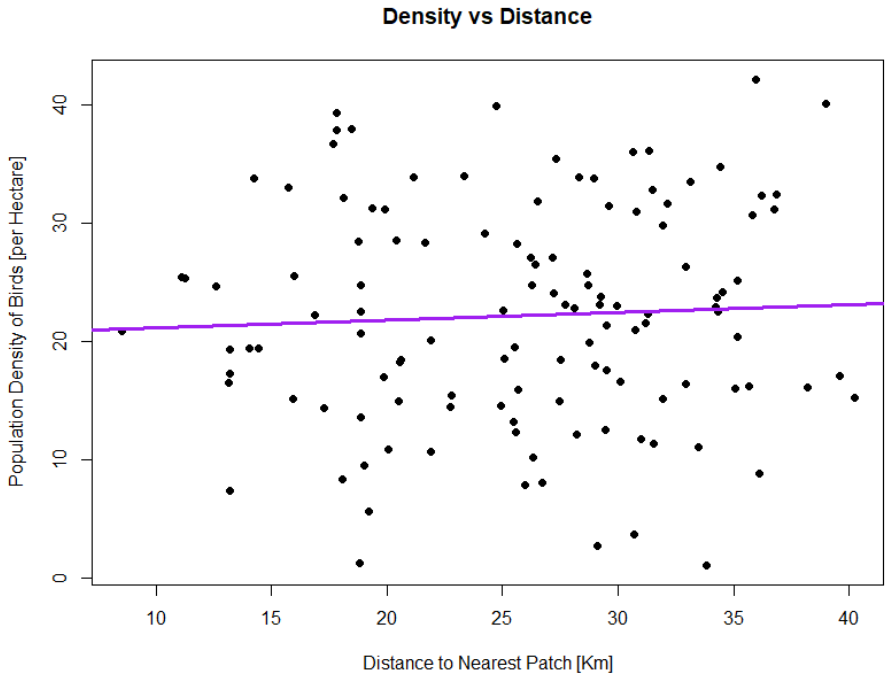


**Fig 1.3**

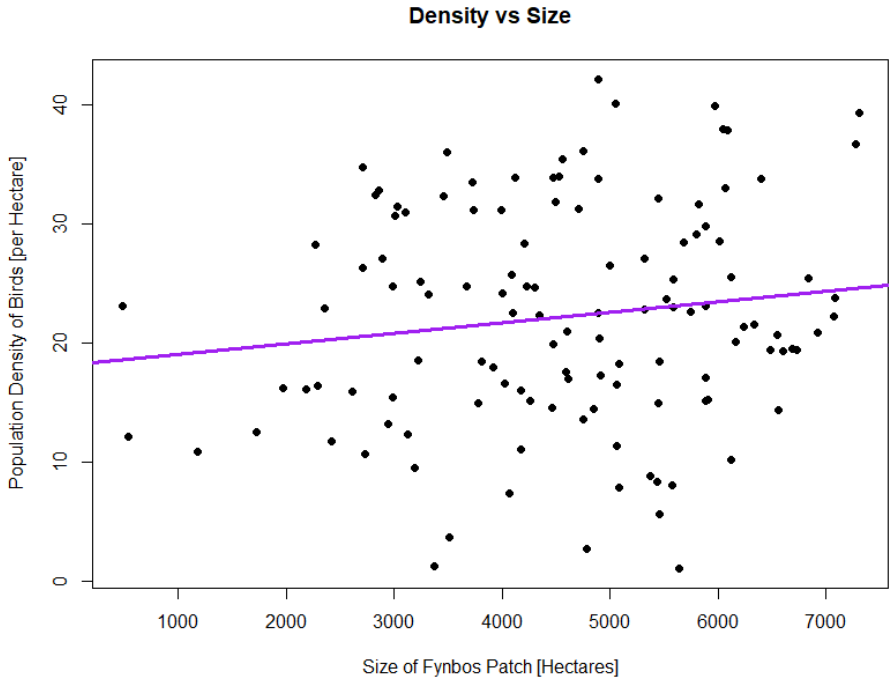


**Fig 1.4**

## Section 2:

```
> H1 <- lm(density ~ protea + vegtype, data = birb)
> summary(H1)

Call:
lm(formula = density ~ protea + vegtype, data = birb)

Residuals:
     Min       1Q   Median       3Q      Max
-16.7178  -4.8788   0.8645   5.1732  14.1122

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          16.235      1.525  10.644  < 2e-16 ***
proteaPresent        11.800      1.524   7.743 3.70e-12 ***
vegtyperenosterveld  -1.769      1.605  -1.102    0.273
vegtypestrandveld   -10.106      1.478  -6.839 3.75e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.835 on 118 degrees of freedom
Multiple R-squared:  0.4695,      Adjusted R-squared:  0.456
F-statistic:  34.8 on 3 and 118 DF,  p-value: 3.459e-16
```

```
> H2a <- lm(density ~ time + vegtype + alien, data = birb)
> summary(H2a)

Call:
lm(formula = density ~ time + vegtype + alien, data = birb)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0121  -1.8938  -0.2416   2.5437   7.9119

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         25.6890     1.7839  14.400  < 2e-16 ***
time                 0.6758     0.1483   4.558 1.28e-05 ***
vegtyperenosterveld -2.7442     0.9069  -3.026  0.00305 **
vegtypestrandveld  -12.7180     0.8220 -15.472  < 2e-16 ***
alienPresent       -12.3057     0.8539 -14.412  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.682 on 117 degrees of freedom
Multiple R-squared:  0.8473,      Adjusted R-squared:  0.8421
F-statistic: 162.3 on 4 and 117 DF,  p-value: < 2.2e-16
```

```
> H2b <- lm(density ~ time + vegtype + alien + time:alien, data = birb)
> summary(H2b)

Call:
lm(formula = density ~ time + vegtype + alien + time:alien, data = birb)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6903  -2.3010   0.2867   2.2172   7.3153

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         19.7593     2.2746   8.687 2.77e-14 ***
time                 1.2023     0.1950   6.167 1.05e-08 ***
vegtyperenosterveld -2.5947     0.8577  -3.025 0.003061 **
vegtypestrandveld  -12.8852     0.7778 -16.566  < 2e-16 ***
alienPresent        -3.1610     2.4898  -1.270 0.206783
time:alienPresent   -0.9559     0.2462  -3.882 0.000172 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.479 on 116 degrees of freedom
Multiple R-squared:  0.8649,      Adjusted R-squared:  0.859
F-statistic: 148.5 on 5 and 116 DF,  p-value: < 2.2e-16
```

```
> H3 <- lm(density ~ distance + size, data = birb)
> summary(H3)

Call:
lm(formula = density ~ distance + size, data = birb)

Residuals:
    Min      1Q  Median      3Q     Max
-23.572  -5.639  -1.277   7.206  18.040

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.288e+01  5.033e+00   2.559   0.0117 *
distance    1.528e-01  1.230e-01   1.242   0.2167
size        1.171e-03  6.136e-04   1.908   0.0588 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.192 on 119 degrees of freedom
Multiple R-squared:  0.03227,     Adjusted R-squared:  0.01601
F-statistic: 1.984 on 2 and 119 DF,  p-value: 0.142
```

```
> H4a <- lm(density ~ alien + altitude, data = birb)
> summary(H4a)

Call:
lm(formula = density ~ alien + altitude, data = birb)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8922  -2.9929   0.1556   3.1620   8.6530

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.276e+01  7.597e-01   29.96   <2e-16 ***
alienPresent -1.481e+01  8.136e-01  -18.21   <2e-16 ***
altitude      7.112e-03  6.311e-04   11.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.441 on 119 degrees of freedom
Multiple R-squared:  0.7741,      Adjusted R-squared:  0.7703
F-statistic: 203.9 on 2 and 119 DF,  p-value: < 2.2e-16

> A <- birb$altitude^2
> H4b <- lm(density ~ alien + altitude + A, data = birb)
> summary(H4b)

Call:
lm(formula = density ~ alien + altitude + A, data = birb)

Residuals:
    Min      1Q  Median      3Q     Max
-10.882  -2.980   0.538   2.665   9.546

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.077e+01  7.695e-01  26.992  < 2e-16 ***
alienPresent -1.494e+01  7.288e-01 -20.493  < 2e-16 ***
altitude      1.809e-02  2.067e-03   8.749 1.78e-14 ***
A            -6.391e-06  1.158e-06  -5.519 2.06e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.976 on 118 degrees of freedom
Multiple R-squared:  0.8205,      Adjusted R-squared:  0.8159
F-statistic: 179.7 on 3 and 118 DF,  p-value: < 2.2e-16
```

## Section 3:

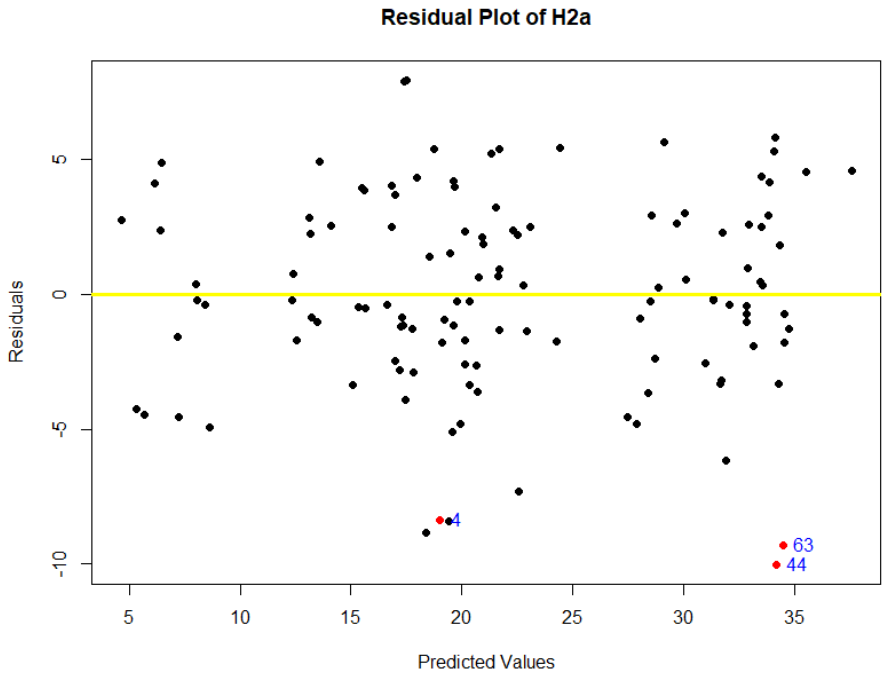


**Fig 3.2**

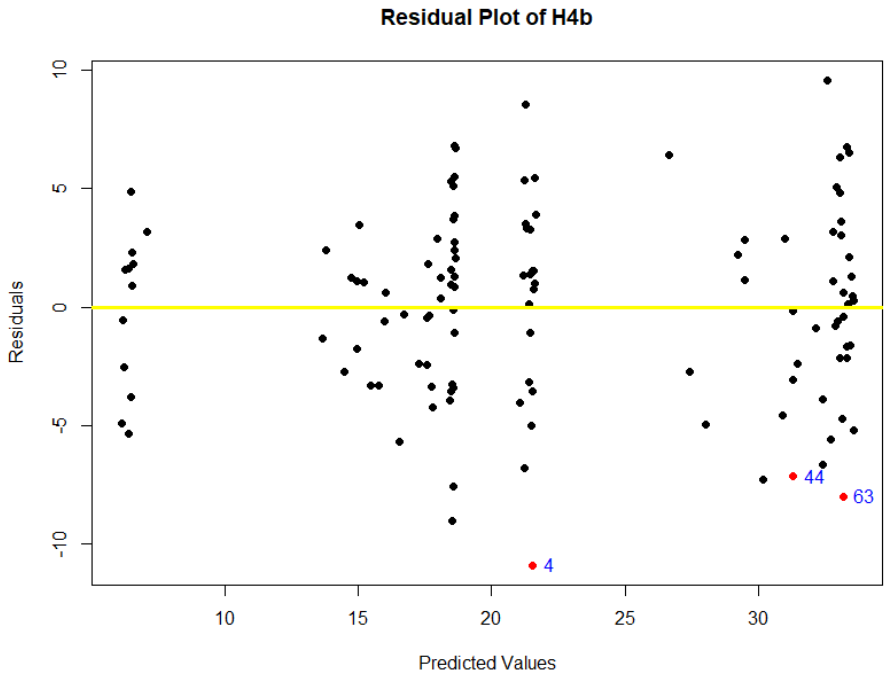


**Fig 3.3**