**Title: Natural Regeneration Capacity of Eucalyptus spp. Overstorey**

**Background**

Low levels of natural tree regeneration have been reported in agricultural and intensively grazed landscapes globally. An understanding of the extent and pattern of natural regeneration is critical to help landowners, conservation agencies and governments understand the impact of this issue. To explore the possibility of Eucalyptus spp. overstorey regeneration, this study utilized the counts of Eucalyptus spp seedlings which spans from 0 to more than 2 meters in height as a regeneration indication. This study is to identify what parts of the different aspects of vegetation cover or plant composition within a specific environmental, tree canopy cover, distance to tree canopy is at greatest threat from the lack of regeneration.

**Methods**

I obtained data from a restoration project in Australia, where livestock grazing was removed from properties. The purpose of this study was to reduce human active planting and in the hopes that the Eucalyptus spp. overstorey would regenerate. Three rounds of surveys took place across 18 sites. Each survey involved the random placement of distinct sets of 15 x 15 m quadrats within 60 m of existing tree canopies across every site. This study focused on the number of Eucalyptus seedlings in each quadrat, Eucalyptus tree canopy cover, distance to tree canopy and ground-layer plant species composition within each quadrat. I had converted several character variables into factors, they were date, season, property, aspect and landscape. It allowed RStudio software (R4.2.3) to recognize these variables as categorical rather than numerical. I grouped five exotic variables related to vegetation cover into a group\_Exotic, five native variables related to vegetation cover into a group\_Native, and three count of eucalypt seedlings that were between 0 cm to > 2 meters in height into a group\_seedlings (sum of counts) (Table 1), percentage computation of group\_Exotic and group\_Native were published in R script (Appendix).

Table 1 Grouping of multiple variables

|  |  |  |  |
| --- | --- | --- | --- |
| group\_Exotic (sum%) | group\_Native (sum%) | group\_seedlings (sum#) | |
| ExoticAnnualGrass\_cover | NativePerennialFern\_cover | euc\_sdlgs0\_50cm | |
| ExoticAnnualHerb\_cover | NativePerennialGrass\_cover | euc\_sdlgs50cm.2m | |
| ExoticPerennialHerb\_cover | NativePerennialHerb\_cover | euc\_sdlgs.2m | |
| ExoticPerennialGrass\_cover | NativePerennialGraminoid\_cover |  |  |
| ExoticShrub\_cover | NativeShrub\_cover |  |  |

The pairs.panels() function was used to create a matrix of scatterplots for visualizing relationships between multiple variables in a dataset. I used pairs.panels() with the argument lm = TRUE, it added fitted regression lines along with scatterplots in the lower triangle of the panel (Appendix: supplementary figure 1 and 2). I removed rows with missing values with this syntax: dat <- na.omit(dat).

**Result**

I explored associations between the dependent variable (presence of Eucalyptus regeneration) and the independent variables (Table 1), distance between the quadrat and the closest edge of the nearest eucalypt canopy in meters (Distance\_to\_Eucalypt\_canopy) and percentage foliage projective cover by eucalypt canopy per quadrat (Euc\_canopy\_cover). The pairs panels (supplementary figure 1 and 2) provided hints about these datasets were skewed. Given that the response variable (the number of eucalypt seedlings) was count integer. I tested for dependence by examining relationships between the probability of Eucalyptus regeneration and the explanatory variables in the candidate models (data not shown but analysis done in R script) using generalized linear model of Poisson regression. The overdispersion was serious (the ratio of the residual deviance to the residual degrees of freedom was 2350.6 on 341 degrees of freedom), and I discarded the model estimates. Therefore, I searched for an alternative link function that allowed the variance to increase more than the mean. To assess the response of Eucalyptus regeneration to variation in explanatory variables, I fitted a generalized linear model with negative binomial error distribution (accounting for severe overdispersion) and a log link function (Table 2). Each candidate model was ranked from best to worst fitting on the basis of ascending Akaike’s Information Criterion (AIC) values. AIC values were transformed into AIC weights (w) and AIC differences (delta). Models with AIC differences less than two (delta < 2) had substantial empirical support and should all be considered (Burnham and Anderson, 2002). Degrees of freedom (df) represented the number of parameters in each model. logLik (The maximized log-likelihood) of the model, indicating how well the model explained the observed data. Higher values suggested a better fit to the data. In my case, I chose only model\_0 to be analyzed (Table 3).

Table 2: Syntax of each model

# Negative Binomial Regression

model\_0 <- glm.nb(group\_seedling ~ group\_Exotic + Distance\_to\_Eucalypt\_canopy.m.,

data = dat)

model\_1 <- glm.nb(group\_seedling ~ group\_Exotic + group\_Native, data = dat)

model\_2 <- glm.nb(group\_seedling ~ group\_Exotic \* group\_Native, data = dat)

model\_3 <- glm.nb(group\_seedling ~ group\_Exotic + group\_Native + Euc\_canopy\_cover,

data = dat)

model\_4 <- glm.nb(group\_seedling ~ group\_Exotic \* Euc\_canopy\_cover,

data = dat)

model\_5 <- glm.nb(group\_seedling ~ group\_Exotic + group\_Native + Euc\_canopy\_cover

+ Distance\_to\_Eucalypt\_canopy.m., data = dat)

model\_6 <- glm.nb(group\_seedling ~ group\_Native \* Distance\_to\_Eucalypt\_canopy.m.,

data = dat)

Model Equation of model\_0

The model equation is in the form of:

group\_seedling = b0 + b1 \* group\_Exotic + b2 \* Distance\_to\_Eucalypt\_canopy.m.

Table 3 Candidate models for natural regeneration that were examined and their associated

AIC, AIC differences (delta) and Akaike weights (w).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | df | AIC | logLik | delta | w |
| model\_0 | 4 | 902.2210 | -447.1105 | 0.00 | 0.74 |
| model\_5 | 6 | 906.0847 | -447.0424 | 3.86 | 0.11 |
| model\_2 | 5 | 907.8274 | -448.9137 | 5.61 | 0.05 |
| model\_1  model\_6  model\_4  model\_3 | 4  5  5  5 | 908.0123  909.0553  909.2846  909.9572 | -450.0061  -449.5276  -449.6423  -449.9786 | 5.79  6.83  7.06  7.74 | 0.04  0.02  0.02  0.02 |

Table 4 Result of fitting a generalized linear model with a negative binomial distribution to model\_0

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Coefficients | Estimate | Std. Error | z value | Pr(>|z|) |  |
| (Intercept) | 1.27177 | 0.25588 | 4.970 | 6.69e-07 | \*\*\* |
| group\_Exotic | -0.09123 | 0.03471 | -2.628 | 0.00859 | \*\* |
| Distance\_to\_Eucalypt\_canopy.m. | -0.03027 | 0.012613 | -2.400 | 0.01639 | \* |

Significant codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

To interpret the parameter estimates, I backtransformed the predicted values for some relevant scenarios to the original count scale.There were 3.57 of Eucalyptus seedlings counted when distance between the quadrat and the closest edge of the nearest eucalypt canopy (m) was zero and group Exotic was set to its mean. For every one-unit (m) increased in the distance from the quadrat to the nearest edge of the closest Eucalypt canopy, there was an approximate expected count decreased of 3.46 times, keeping other variables constant. For every one-unit (%) increased in the exotic vegetation, there was an approximate expected count decreased of 3.26 times, keeping other variables constant (Table 4). To illustrate the response of Eucalyptus seedlings to distance from quadrat to nearest Eucalyptus canopy edge, while also representing the effect of variables representing exotic vegetation cover, I plotted the response curves for three different fixed values of the effect of exotic variables related to vegetation cover (Figure 1).

**A graph of seedlings and distance

Description automatically generated**

Figure 1. Number of Seedlings vs Distance to Eucalypt Canopy

**Conclusion**

Based on the model's coefficients and significance, a negative coefficient (-0.09123) suggests that areas with a higher presence of exotic plants (group\_Exotic) are associated with a slightly lower likelihood of seedling regeneration. This implies that the presence of exotic plants might have a negative impact on seedling regeneration. Distance\_to\_Eucalypt\_canopy (m) is another negative coefficient (-0.03027) indicates that as the distance to Eucalypt tree canopies increases, there is a slightly lower likelihood of seedling regeneration (Table 4). This implies that areas farther away from Eucalypt tree canopies might experience difficulties in seedling regeneration. It suggests that monitoring or managing these factors could be crucial for fostering regeneration.

**Appendix**

**A screenshot of a graph

Description automatically generated**

Supplementary figure 1. Scatterplots of Eucalyptus canopy cover (%), Distance to canopy cover (m) and seedlings (0 - > 2 m)

A grid of squares with different colored squares

Description automatically generated with medium confidence

Supplementary figure 2. Scatterplots of Exotic vegetations (%), Native vegetations (%) and seedlings (50 - 2 m)

**R script of this study**

rm(list = ls())

library(MASS)

library(dplyr)

library(psych)

dat = read.csv("~/git/exploring-data/BIOS14/exam2023\_data.csv")

meta = read.csv("~/git/exploring-data/BIOS14/exam2023\_metadata.csv")

# Convert Date object to factor

dat$Date = as.Date(dat$Date, format = "%d/%m/%Y")

dat$Date = as.factor(dat$Date) # Convert Date object to factor

dat$Season = as.factor(dat$Season) # Convert Season object to factor

dat$Property = as.factor(dat$Property) # Convert Property object to factor

dat$Aspect = as.factor(dat$Aspect) # Convert Aspect object to factor

dat$Landscape.position = as.factor(dat$Landscape.position) # Convert Landscape.

# position object to factor

# Assuming the column numbers 10 to 14 represent group\_Exotic and 15 to 19

# represent group\_Native

# Grouping for pairwise scatterplots

group\_Exotic <- dat[, 10:14]

group\_Native <- dat[, 15:19]

group\_seedling <- dat[, 26:28]

# Use pairs.panels

pairs.panels(dat[, 24:28], lm = TRUE) # seedlings vs distance

# Combine the selected columns

combined\_groups\_euc0\_50 <- cbind(group\_Exotic, group\_Native, dat$euc\_sdlgs0\_50cm)

# Use pairs.panels with the combined groups

# pairs.panels(combined\_groups\_euc0\_50, lm = TRUE)

# Combine the selected columns

combined\_groups\_euc50\_2 <- cbind(group\_Exotic, group\_Native, dat$euc\_sdlgs50cm.2m)

# Use pairs.panels with the combined groups

# pairs.panels(combined\_groups\_euc50\_2, lm = TRUE)

# Combine the selected columns

combined\_groups\_euc\_2m <- cbind(group\_Exotic, group\_Native, dat$euc\_sdlgs.2m)

# Use pairs.panels with the combined groups

# pairs.panels(combined\_groups\_euc\_2m, lm = TRUE)

# Summing columns 26 through 28 to create 'group\_seedling'

# Grouping for data exploration

dat$group\_seedling <- rowSums(dat[, 26:28])

# Assuming columns 10 to 14 contain percentage values

dat <- dat %>%

mutate(

group\_Exotic = rowSums(select(., 10:14) / 100), # Sum percentages and divide by 100

group\_Exotic = group\_Exotic / 5, # Divide the column by 5

group\_Exotic = group\_Exotic \* 100 # Multiply by 100 to get percentage

)

dat <- dat %>%

mutate(

group\_Native = rowSums(select(., 15:19) / 100), # Sum percentages and divide by 100

group\_Native = group\_Native / 5, # Divide the column by 5

group\_Native = group\_Native \* 100 # Multiply by 100 to get percentage

)

ncol(dat) # Number of columns in 'dat'

nrow(dat) # Number of rows in 'dat'

colnames(dat) # Column names in 'dat'

# Removing rows with missing values

dat <- na.omit(dat)

# Running the Poisson regression

model\_P0 <- glm(group\_seedling ~ group\_Exotic + group\_Native, data = dat, family = poisson)

model\_P1 <- glm(group\_seedling ~ group\_Exotic \* group\_Native, data = dat,

family = poisson)

model\_P2 <- glm(group\_seedling ~ group\_Exotic + group\_Native + Euc\_canopy\_cover,

data = dat, family = poisson)

model\_P3 <- glm(group\_seedling ~ group\_Exotic + group\_Native + Euc\_canopy\_cover

+ Distance\_to\_Eucalypt\_canopy.m., data = dat, family = poisson)

model\_P4 <- glm(group\_seedling ~ group\_Native \* Euc\_canopy\_cover,

data = dat, family = poisson)

model\_P5 <- glm(group\_seedling ~ group\_Native \* Distance\_to\_Eucalypt\_canopy.m.,

data = dat, family = poisson)

model\_P6 <- glm(group\_seedling ~ group\_Exotic \* Euc\_canopy\_cover,

data = dat, family = poisson)

mlist = list(model\_P0, model\_P1, model\_P2, model\_P3, model\_P4, model\_P5, model\_P6)

AICTab = AIC(model\_P0, model\_P1, model\_P2, model\_P3, model\_P4, model\_P5, model\_P6)

AICTab$logLik = unlist(lapply(mlist, logLik))

AICTab = AICTab[order(AICTab$AIC, decreasing=F),]

AICTab$delta = round(AICTab$AIC - min(AICTab$AIC), 2)

lh = exp(-0.5\*AICTab$delta)

AICTab$w = round(lh/sum(lh), 2)

AICTab

summary(model\_P3) # Model with the lowest AIC value

# Negative Binomial Regression

model\_0 <- glm.nb(group\_seedling ~ group\_Exotic + Distance\_to\_Eucalypt\_canopy.m.,

data = dat)

model\_1 <- glm.nb(group\_seedling ~ group\_Exotic + group\_Native, data = dat)

model\_2 <- glm.nb(group\_seedling ~ group\_Exotic \* group\_Native, data = dat)

model\_3 <- glm.nb(group\_seedling ~ group\_Exotic + group\_Native + Euc\_canopy\_cover,

data = dat)

model\_4 <- glm.nb(group\_seedling ~ group\_Exotic \* Euc\_canopy\_cover,

data = dat)

model\_5 <- glm.nb(group\_seedling ~ group\_Exotic + group\_Native + Euc\_canopy\_cover

+ Distance\_to\_Eucalypt\_canopy.m., data = dat)

model\_6 <- glm.nb(group\_seedling ~ group\_Native \* Distance\_to\_Eucalypt\_canopy.m.,

data = dat)

mlist = list(model\_0, model\_1, model\_2, model\_3, model\_4, model\_5, model\_6)

AICTab = AIC(model\_0, model\_1, model\_2, model\_3, model\_4, model\_5, model\_6)

AICTab$logLik = unlist(lapply(mlist, logLik))

AICTab = AICTab[order(AICTab$AIC, decreasing=F),]

AICTab$delta = round(AICTab$AIC - min(AICTab$AIC), 2)

lh = exp(-0.5\*AICTab$delta)

AICTab$w = round(lh/sum(lh), 2)

AICTab

# Compare the Summary of Each Model

# summary(model\_0)

# summary(model\_1)

# summary(model\_2)

summary(model\_0)

# summary(model\_4)

# summary(model\_5)

# summary(model\_6)

coefs = summary(model\_0)$coef

exp(coefs[1,1]) #Intercept

exp(coefs[1,1] + coefs[2,1]) # Intercept + Exotic

exp(coefs[1,1] + coefs[3,1]) # Intercept + Distance\_to\_Eucalypt\_canopy.m.

plot(dat$Distance\_to\_Eucalypt\_canopy.m., dat$group\_seedling, xlab="Distance to Eucalypt canopy (m)",

ylab = "Number of seedlings", main = "Number of Seedlings vs Distance",

col = 'lightblue', las = 1)

newDistance = seq(min(dat$Distance\_to\_Eucalypt\_canopy.m.),

max(dat$Distance\_to\_Eucalypt\_canopy.m.), length.out=200)

newExotic = rep(mean(dat$group\_Exotic), length(newDistance)) # Predictor values

y\_hat = predict(model\_0, newdata = list(group\_Exotic = newExotic,

Distance\_to\_Eucalypt\_canopy.m. = newDistance),

type = "response")

# Plot predicted values

lines(newDistance, y\_hat, lwd=2)

newExotic2 = rep(mean(dat$group\_Exotic)+sd(dat$group\_Exotic), length(newDistance))

y\_hat2 = predict(model\_0, newdata=list(group\_Exotic = newExotic2,

Distance\_to\_Eucalypt\_canopy.m. = newDistance),

type = "response")

newExotic3 = rep(mean(dat$group\_Exotic)-sd(dat$group\_Exotic), length(newDistance))

y\_hat3 = predict(model\_0, newdata=list(group\_Exotic = newExotic3,

Distance\_to\_Eucalypt\_canopy.m. = newDistance),

type = "response")

lines(newDistance, y\_hat2, lwd=2, col="blueviolet")

lines(newDistance, y\_hat3, lwd=2, col="darkgoldenrod")

legend("topleft", lty=1, lwd=2, col=c(1, "blueviolet", "darkgoldenrod"), bty="n",

legend=c("Exotic = Mean",

"Exotic = Mean + SD",

"Exotic = Mean - SD"))

abline(h=exp(coefs[1,1]), lty=2) # Intercept

abline(h=exp(coefs[1,1] + coefs[3,1]), lty=2, col="red") # Intercept +

# Distance\_to\_Eucalypt\_canopy.m.