evaluate_SDMs

steppe

3/13/2022

Contents

List of Figures

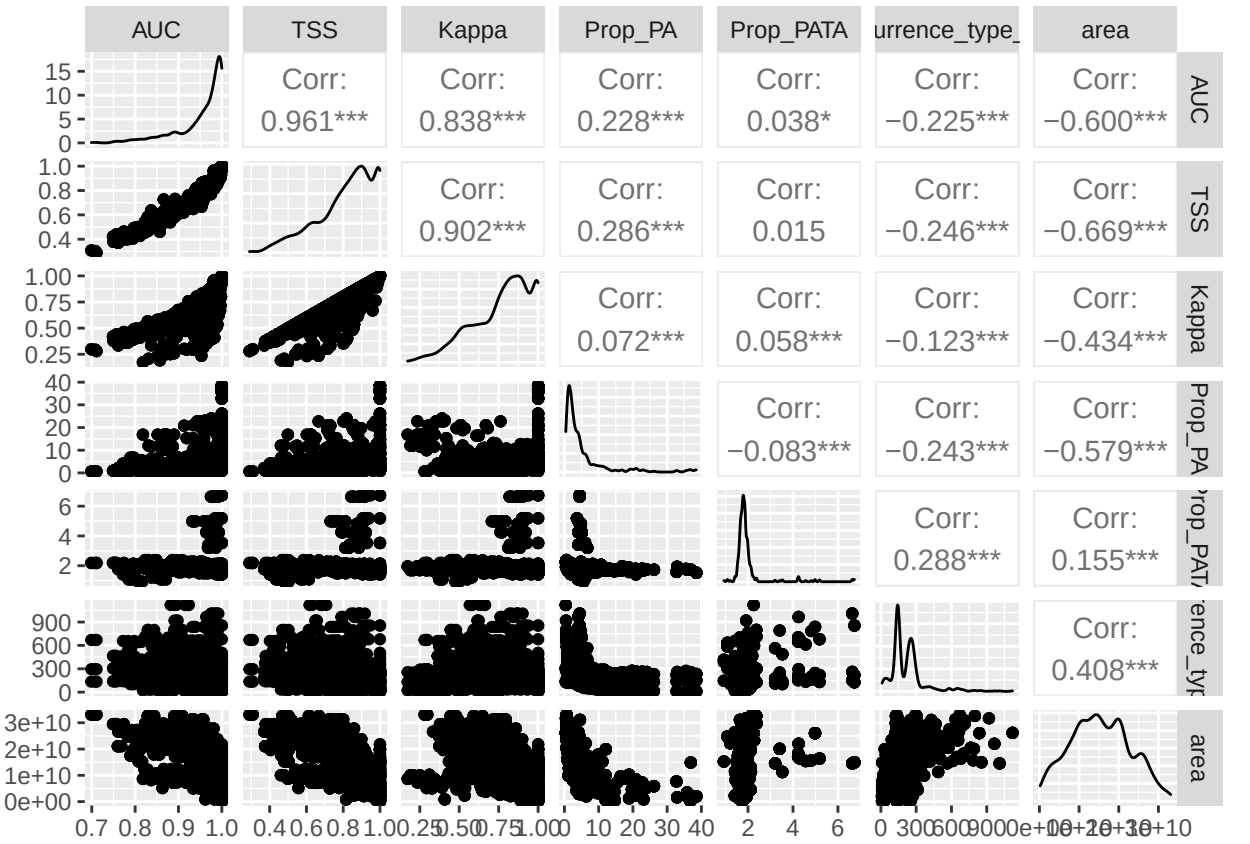
List of Tables

1 Import and Wrangle Data

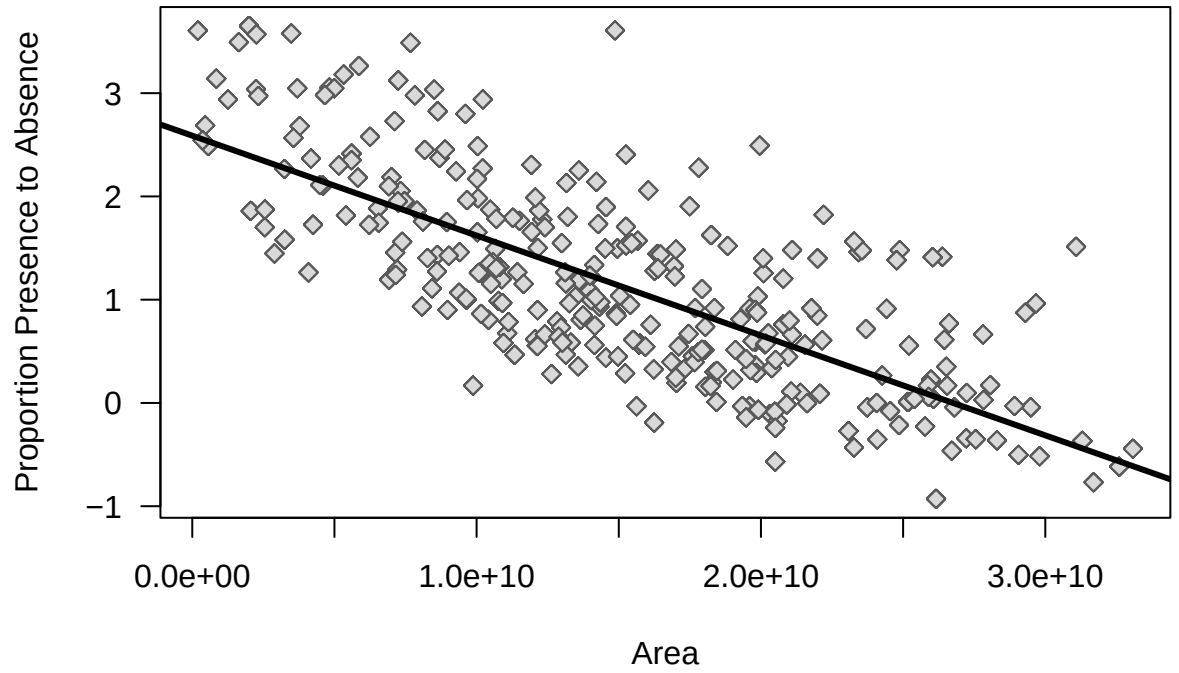
```
## `summarise()` has grouped output by 'binomial'. You can override using the
## `.groups` argument.
```

2 Data Exploration

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

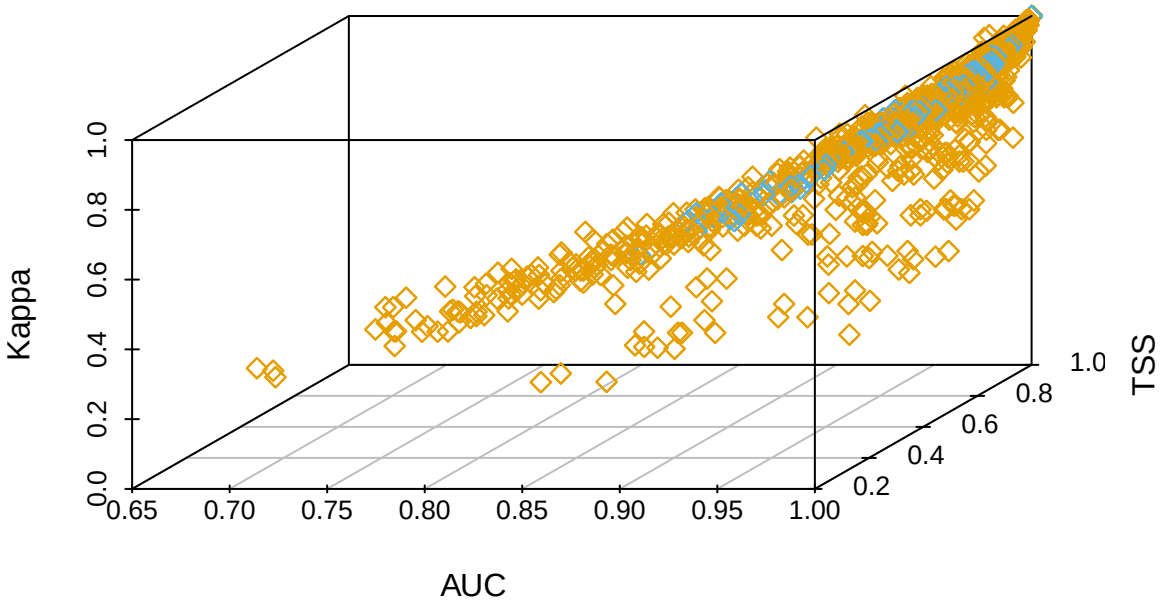


Relationship between Area and Proportion of Presence to Absence

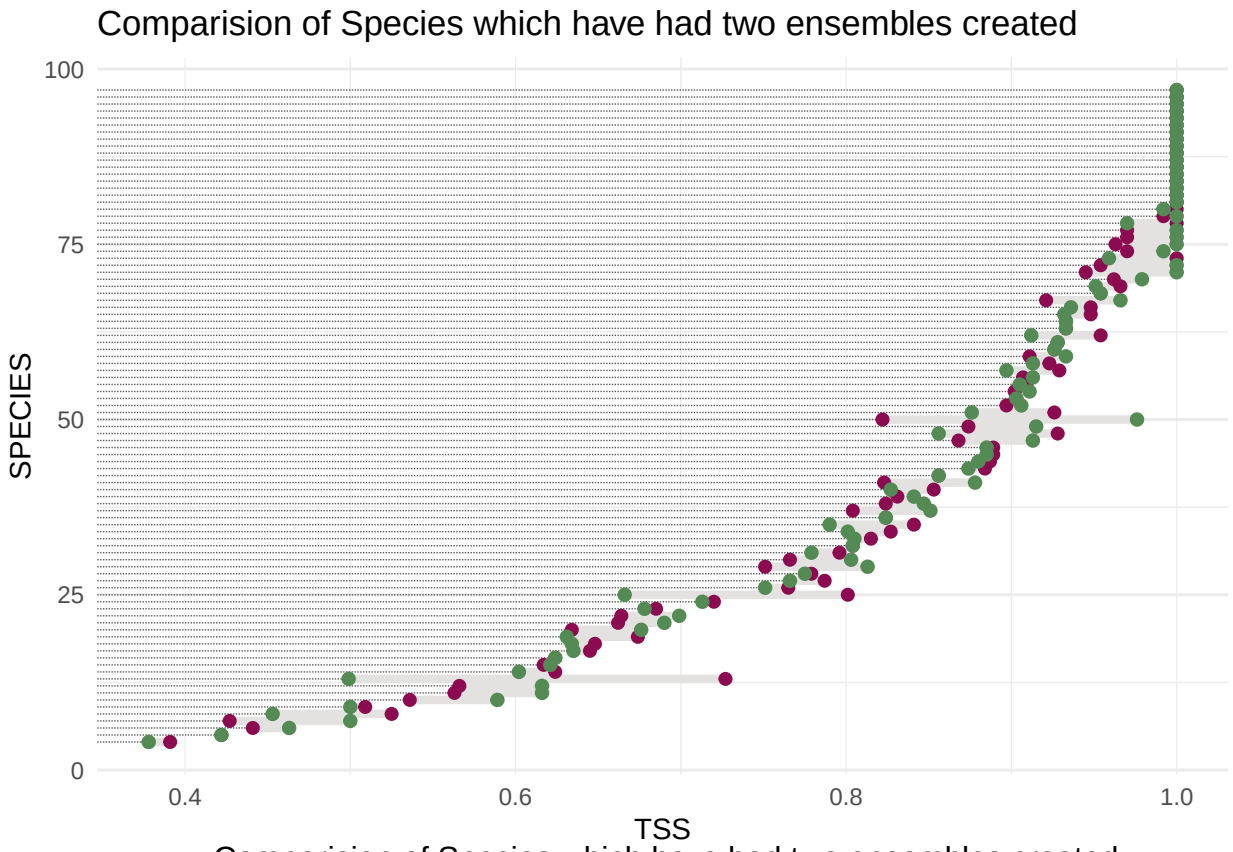


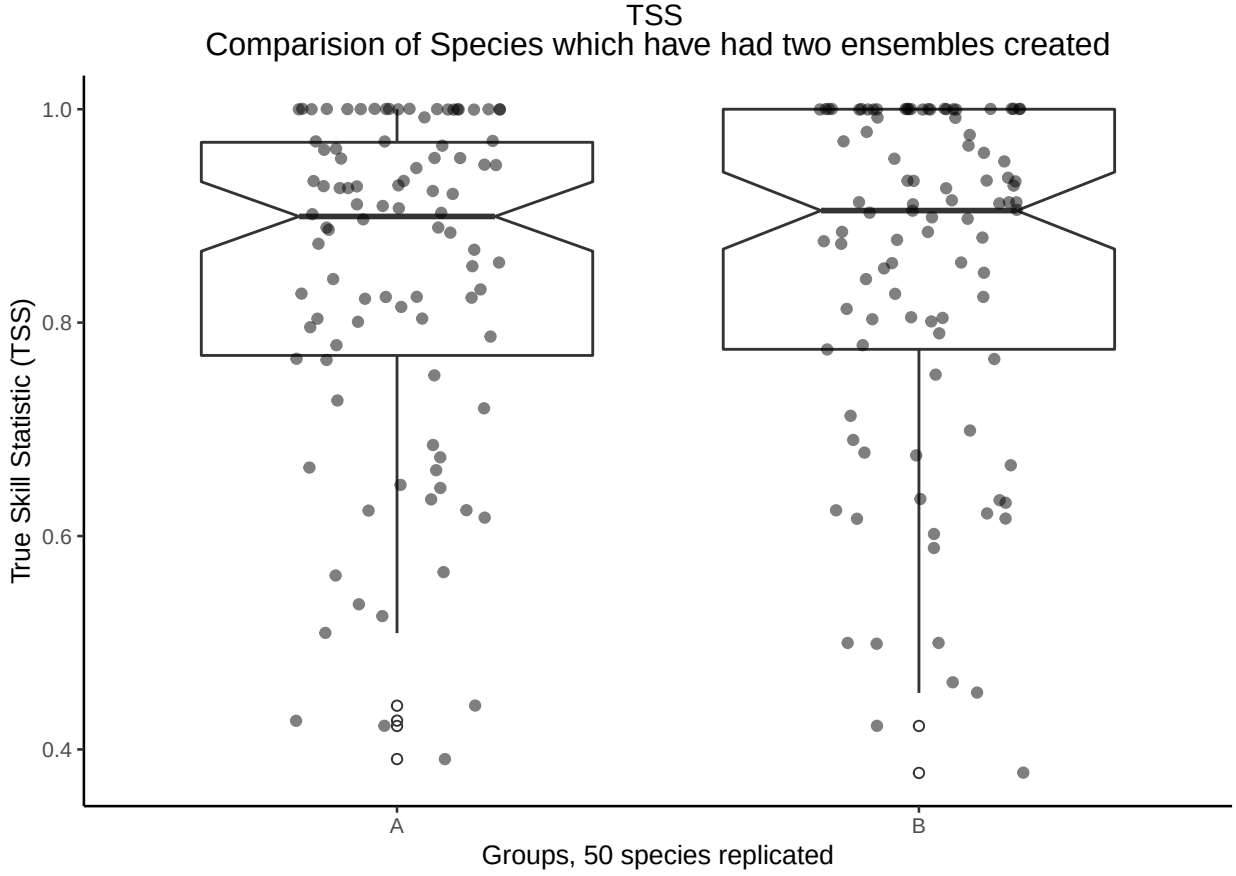
As expected, we see that the three evaluation of model fit criteria, Area Under the Curve (AUC), the True Skill Statistic (TSS), and Kappa are high correlated. Area has significant correlations with all variables examined here, and has strong negative correlations with the evaluation criteria AUC, TSS, and Kappa, as

well as with the Proportion of records which are Presences to Absences.




```
Registered S3 methods overwritten by 'ggalt':
  method                  from
  grid.draw.absoluteGrob  ggplot2
  grobHeight.absoluteGrob ggplot2
  grobWidth.absoluteGrob  ggplot2
  grobX.absoluteGrob      ggplot2
  grobY.absoluteGrob      ggplot2
```






```
Wilcoxon signed rank test with continuity correction
```

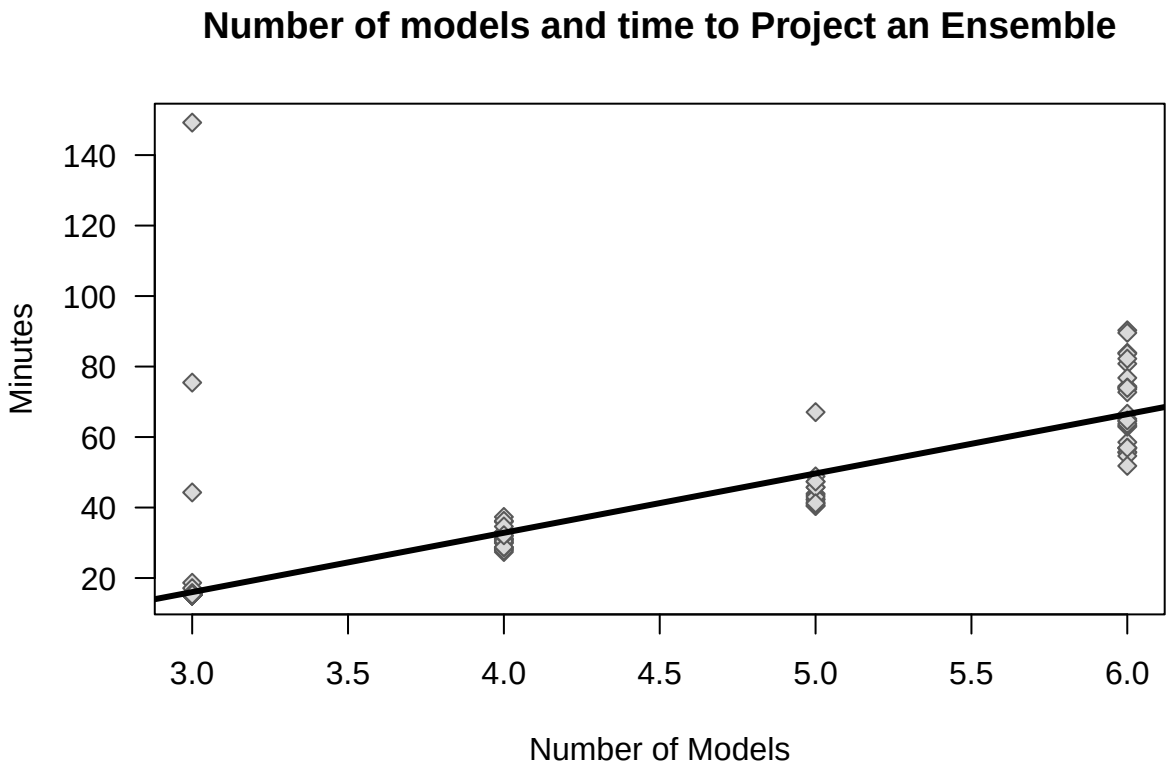
```
data:  grpA and grpB
V = 1988.5, p-value = 0.8139
alternative hypothesis: true location shift is not equal to 0
```

Using a dependent Wilcoxon signed rank test different modelling runs appear unlikely to give substantially different results for the models which were run.

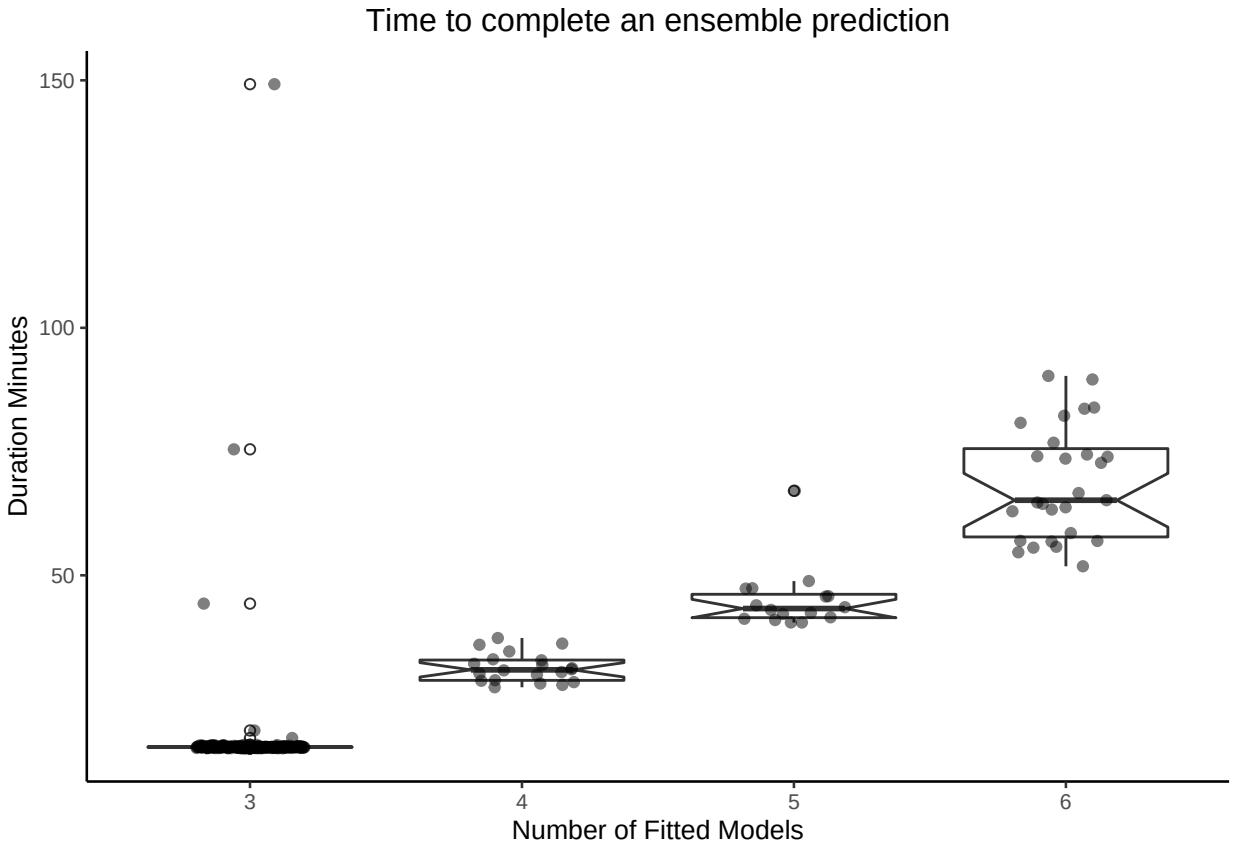
3 Analyses

How long does it take to run an ensemble forecast? # THIS DOES NOT MAKE TOTAL SENSE AS TIME(GLM) =? TIME(GAM)

```
Call:
lm(formula = Duration ~ Number_Models, data = how_long)

Residuals:
    Min      1Q  Median      3Q     Max
-14.676  -0.816  -0.733  -0.650 133.200

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -34.4599     1.9633  -17.55   <2e-16 ***
Number_Models  16.8254     0.5554   30.29   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.162 on 320 degrees of freedom
Multiple R-squared:  0.7414,    Adjusted R-squared:  0.7406
F-statistic: 917.6 on 1 and 320 DF,  p-value: < 2.2e-16
```



```
notch went outside hinges. Try setting notch=FALSE.
```




```
    Fligner-Killeen test of homogeneity of variances

data:  Duration by as.factor(Number_Models)
Fligner-Killeen:med chi-squared = 159.94, df = 3, p-value < 2.2e-16

Call:
glm(formula = Duration ~ Number_Models, family = poisson, data = how_long)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3298  -0.3145  -0.3145  -0.3145  19.8696

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.318529   0.037951   34.74   <2e-16 ***
Number_Models 0.489884   0.008937   54.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3164.1  on 321  degrees of freedom
Residual deviance:  646.8  on 320  degrees of freedom
AIC: 2197.3

Number of Fisher Scoring iterations: 4
```

```
    Overdispersion test

data:  glm_poiss
z = 1.0342, p-value = 0.1505
alternative hypothesis: true alpha is greater than 0
sample estimates:
  alpha
3.52037
```

```
Call:
MASS::glm.nb(formula = Duration ~ Number_Models, data = how_long,
    init.theta = 23.47679262, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2679  -0.2294  -0.2294  -0.2294  12.4316

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.29389    0.05878   22.01   <2e-16 ***
Number_Models  0.49659    0.01550   32.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(23.4768) family taken to be 1)

    Null deviance: 1381.9  on 321  degrees of freedom
Residual deviance:  261.0  on 320  degrees of freedom
AIC: 2016.9

Number of Fisher Scoring iterations: 1

              Theta:  23.48
          Std. Err.:  3.30

 2 x log-likelihood:  -2010.946
```

```
'log Lik.' 1 (df=3)
```

```
Waiting for profiling to be done...
```

```
               Estimate     2.5 %    97.5 %
(Intercept)   1.2938868 1.1774651 1.4096704
Number_Models 0.4965948 0.4660526 0.5274022
```

```
              Estimate    2.5 %   97.5 %
(Intercept)   3.646934 3.246135 4.094605
Number_Models 1.643117 1.593691 1.694525
```

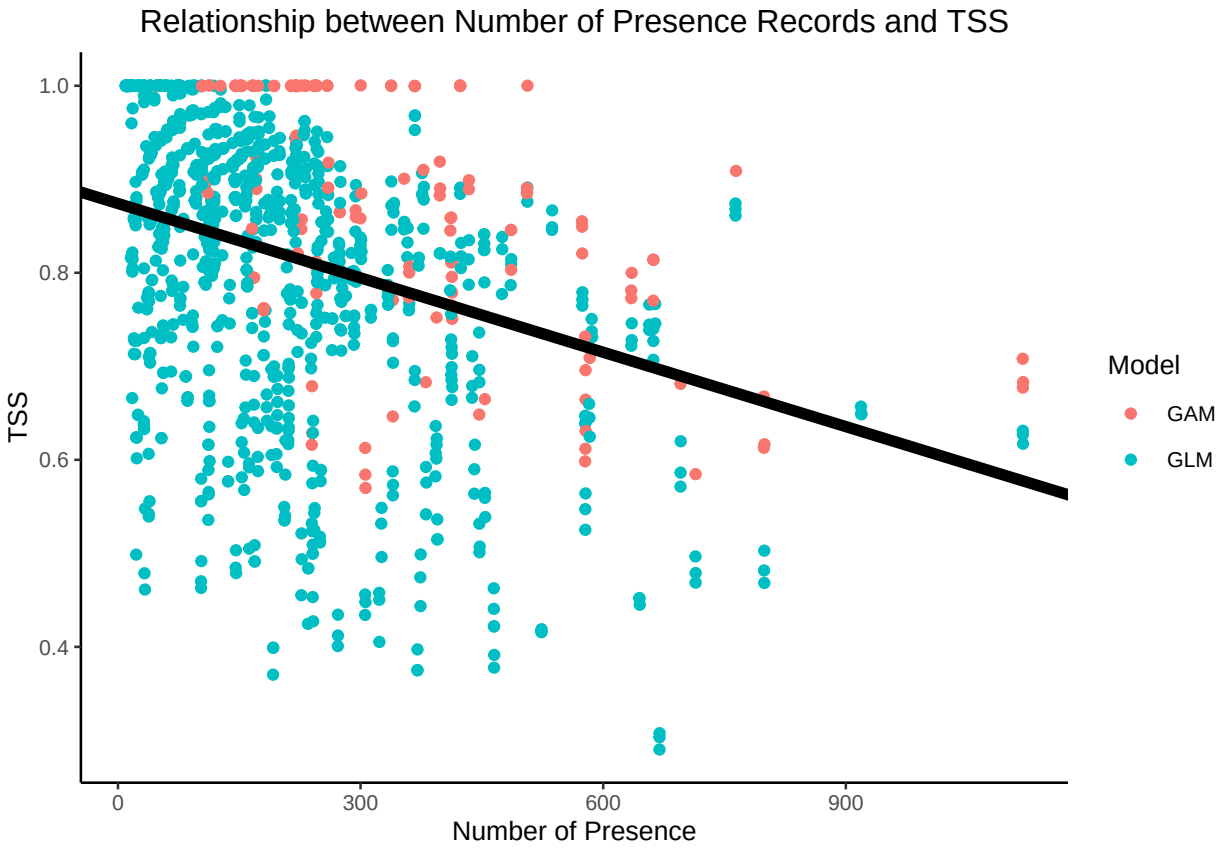
Fligner-Killeen test null hypothesis: the sample variances have equal variance, this is soundly rejected for the alternative.

On average it looks like it may take 65% more time to ensemble a prediction as each additional model is added.

Do the Number of Presence/Absences predict the accuracy of models ?

```
Call:
```

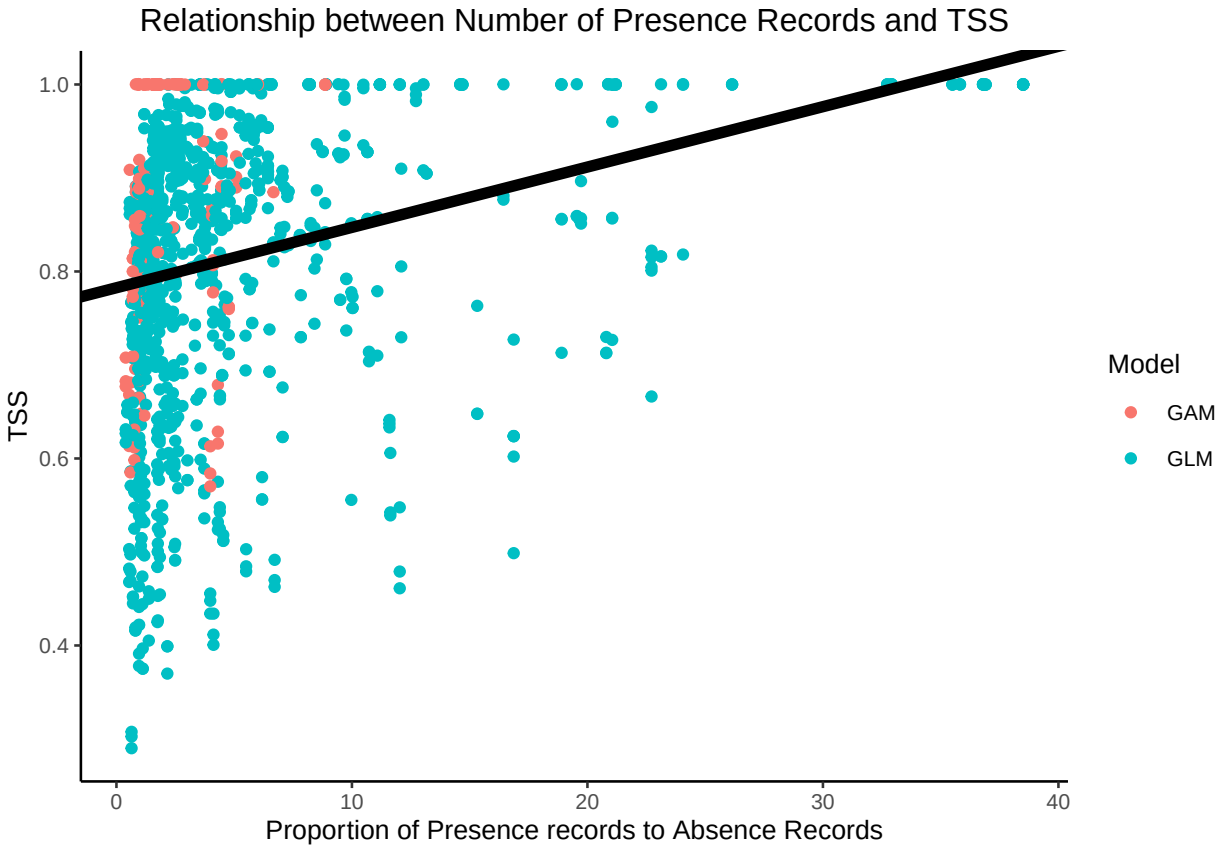
```
lm(formula = TSS ~ occurrence_type_cnt, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54829 -0.08565  0.03445  0.11533  0.35257

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.8740799  0.0046487  188.03   <2e-16 ***
occurrence_type_cnt -0.0002651  0.0000177  -14.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1502 on 3472 degrees of freedom
Multiple R-squared:  0.06069,   Adjusted R-squared:  0.06042
F-statistic: 224.3 on 1 and 3472 DF,  p-value: < 2.2e-16
```



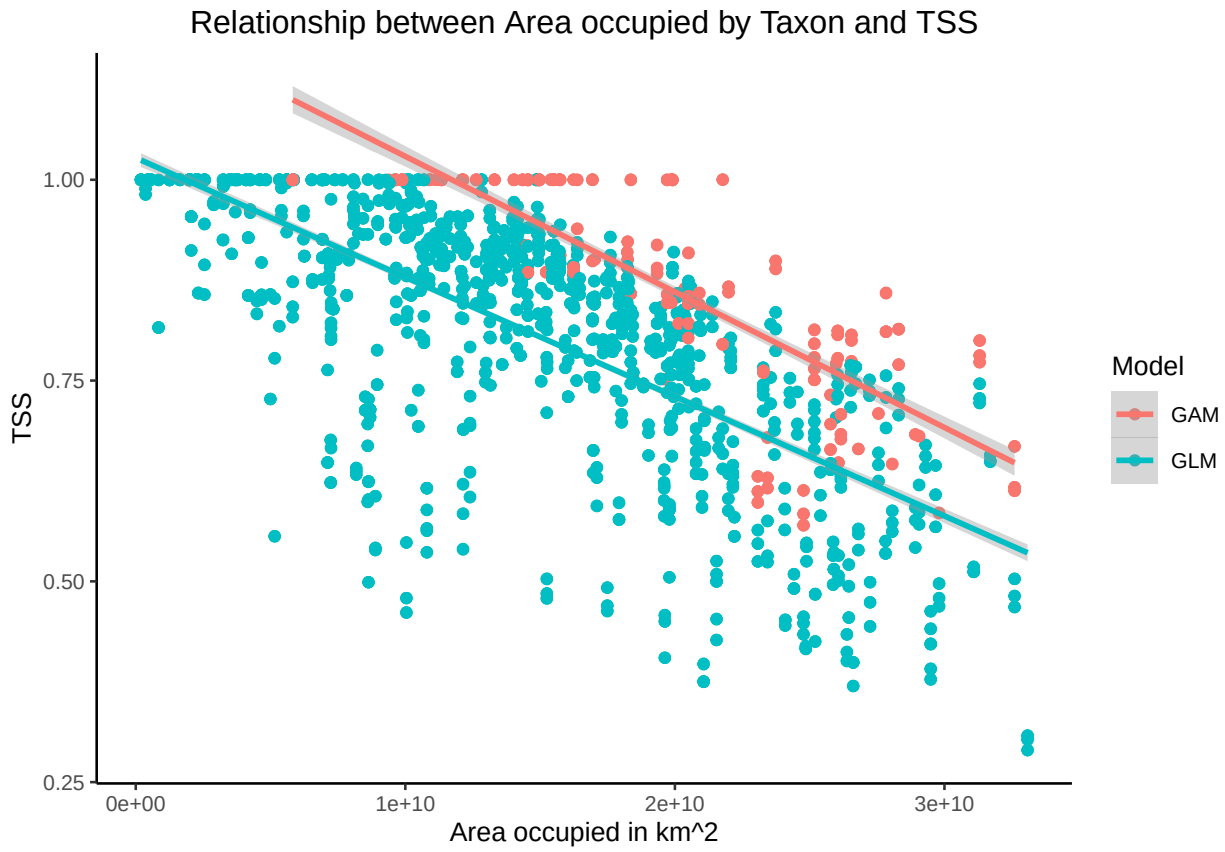
Obviously more data does not make prediction worse, rather the root cause is likely to be the correlation between the number of records and the number of habitat types the species grows in. The more generalized a species, the more likely that a presence is to occur in an area deemed as predicted as absence.

```
Call:
lm(formula = TSS ~ Prop_PA, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49673 -0.08139  0.03134  0.11482  0.21208
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7825878  0.0031495  248.48   <2e-16 ***
Prop_PA     0.0064551  0.0003667   17.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1485 on 3472 degrees of freedom
Multiple R-squared:  0.08195,   Adjusted R-squared:  0.08169
F-statistic: 309.9 on 1 and 3472 DF,  p-value: < 2.2e-16
```



Do the accuracy of models decrease with increases in the geographic extent which the species ranger covers in



the study area?

```
## Warning in rm(areas_occ): object 'areas_occ' not found
```

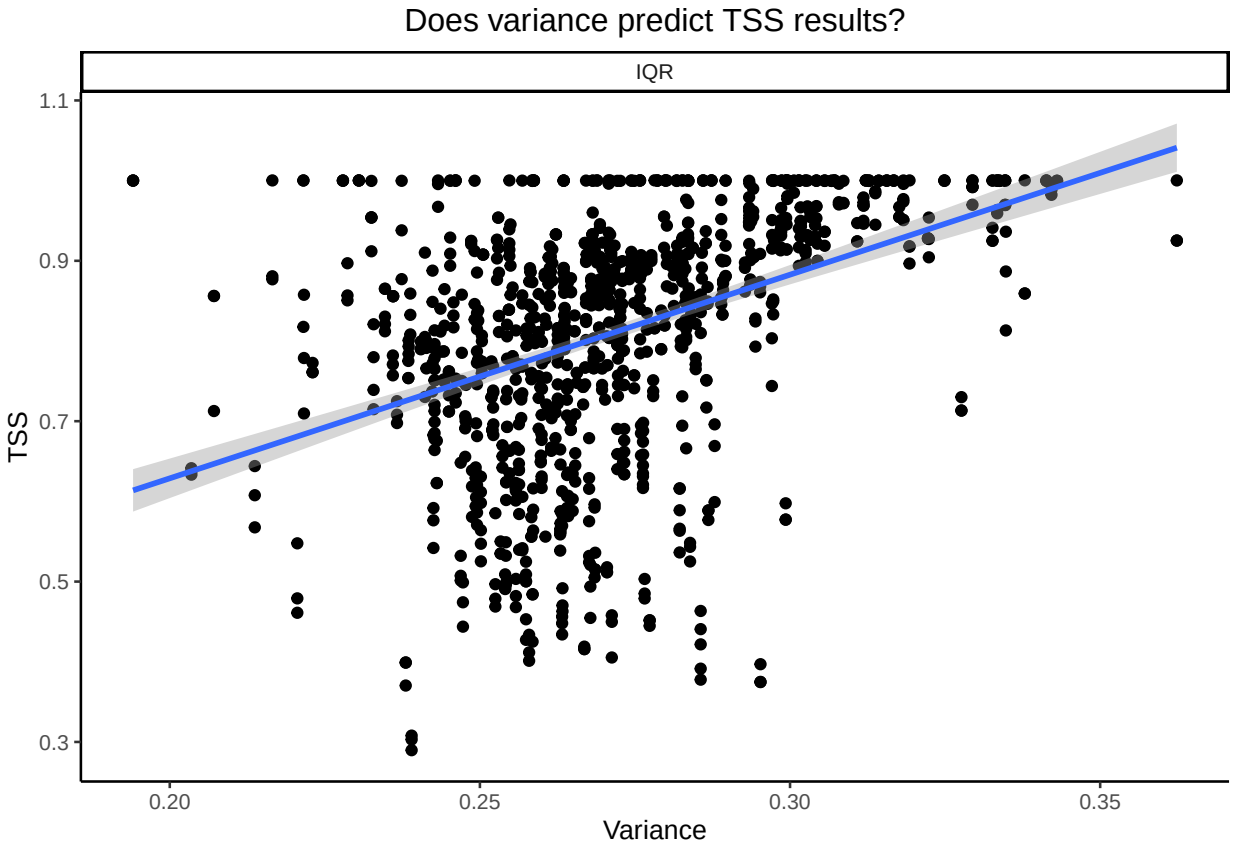
How does variance affect model fit?

```
variance_df <- data %>%
  dplyr::select(Taxon, modelID, TSS, Model) %>%
  distinct(.keep_all = T) %>%
  left_join(., shout, by = c('Taxon' = 'binomial'))

occurrence_pred_tss <- lm(TSS ~ Dispersion, data = variance_df)
occ_pr_tss_sum <- summary(occurrence_pred_tss)
occ_pr_tss_sum
```

```
Call:
lm(formula = TSS ~ Dispersion, data = variance_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52247 -0.08547  0.03453  0.11553  0.18753

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.125e-01  4.599e-03   176.7   <2e-16 ***
DispersionSE  -3.243e-19  6.504e-03     0.0        1
DispersionVAR -1.370e-19  6.504e-03     0.0        1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1544 on 3378 degrees of freedom
Multiple R-squared:  1.326e-31, Adjusted R-squared:  -0.0005921
F-statistic: 2.239e-28 on 2 and 3378 DF,  p-value: 1
```

```
variance_df %>%
  filter(Dispersion == 'IQR') %>%
  ggplot( aes(x = Value, y = TSS)) +
  facet_wrap(~Dispersion) +
  geom_jitter() +
  geom_smooth(method = "lm", formula = 'y ~ x') +
  theme_classic(base_size = 10) +
  labs(title = 'Does variance predict TSS results?',
       y = 'TSS', x = 'Variance') +
  theme(plot.title = element_text(hjust = 0.5))
```




```
rm(variance_df, occurrence_pred_tss, occ_pr_tss_sum)
```

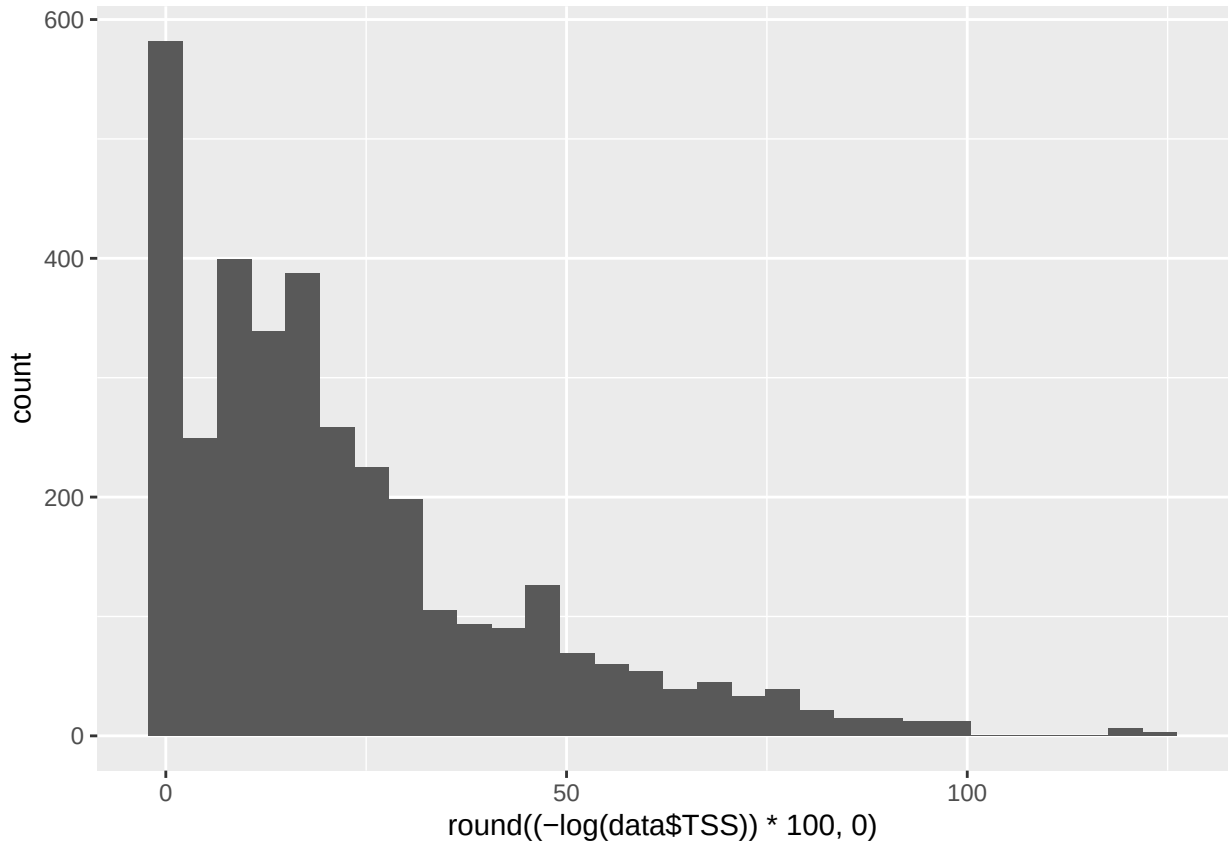
hmm not as expected, bigger sample size and more is explained,

Let's try and fit a model

```
ggplot(data, aes(round((-log(data$TSS))*100,0))) +
  geom_histogram()
```

```
Warning: Use of `data$TSS` is discouraged. Use `TSS` instead.
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
glm_poiss <- glm(round((-log(data$TSS))*100,0) ~ data$area + data$occurrence_type_cnt, family = poisson
summary(glm_poiss)
```

```
Call:
glm(formula = round((-log(data$TSS)) * 100, 0) ~ data$area +
    data$occurrence_type_cnt, family = poisson, data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-8.046  -2.883  -1.156   1.212  12.639

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.650e+00  1.074e-02 153.698  < 2e-16 ***
data$area                 8.519e-11  5.308e-13 160.505  < 2e-16 ***
data$occurrence_type_cnt -1.887e-04  2.341e-05  -8.062 7.53e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 70602  on 3473  degrees of freedom
Residual deviance: 40578  on 3471  degrees of freedom
AIC: 54708

Number of Fisher Scoring iterations: 5
```

```
AER::dispersiontest(glm_poiss, trafo=1)
```

```
    Overdispersion test

data:  glm_poiss
z = 22.434, p-value < 2.2e-16
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
12.29108
```

```
summary(glm_area <- glm(round((-log(TSS))*100,0) ~ area, family = poisson, data = data))
```

```
Call:
glm(formula = round((-log(TSS)) * 100, 0) ~ area, family = poisson,
    data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-7.954  -2.907  -1.169   1.235  12.563

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.633e+00  1.052e-02   155.2   <2e-16 ***
area        8.362e-11  4.950e-13   168.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 70602  on 3473  degrees of freedom
Residual deviance: 40645  on 3472  degrees of freedom
AIC: 54772

Number of Fisher Scoring iterations: 5
```

```
summary(glm_records <- glm(round((-log(TSS))*100,0) ~ occurrence_type_cnt, family = poisson, data = data
```

```
Call:
glm(formula = round((-log(TSS)) * 100, 0) ~ occurrence_type_cnt,
    family = poisson, data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-9.954  -3.804  -1.403   1.869  15.701

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.814e+00  6.137e-03  458.54   <2e-16 ***
occurrence_type_cnt 1.273e-03  1.971e-05   64.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 70602  on 3473  degrees of freedom
Residual deviance: 67022  on 3472  degrees of freedom
AIC: 81150

Number of Fisher Scoring iterations: 5
```

```
AER::dispersiontest(glm_area, trafo=1)
```

```
    Overdispersion test

data:  glm_area
z = 22.471, p-value < 2.2e-16
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
12.30735
```

```
AER::dispersiontest(glm_records, trafo=1)
```

```
    Overdispersion test

data:  glm_records
z = 29.256, p-value < 2.2e-16
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
18.88928
```

```
modelTABLE <- MuMIn::model.sel(glm_area,glm_records)
modelTABLE
```

```
Model selection table
            (Int)       are occ_typ_cnt df    logLik    AICc    delta weight
glm_area    1.633 8.362e-11              2 -27384.22 54772.4     0.00      1
glm_records 2.814             0.001273  2 -40573.03 81150.1 26377.62      0
Models ranked by AICc(x)
```

```
rm(glm_poiss, glm_area, glm_records)
```