First approach to models

Daniela Linero

2/6/2020

1. Load the diversity metrics that have been calculated from a table that:

- Had corrected abundance measures using sampling effort
- Had meged sites

```
diversity3 <- readRDS("./output/intermediate_files/02_Analysis_walkthrough_Site_metrics.rds")
```

2. I am going to merge the land-use classes "Primary forest" and "Primary non-forest". I'm also going to merge the use intensities "Intense" and "light". Finally, I will set the land-use or use intensity as NA if they are:

- Secondary vegetation (indeterminate age)
- Cannot decide
- Urban

```
diversity4 <- diversity3 %>%

  mutate(

    # collapse primary forest and non-forest together into primary vegetation as
    # these aren't well distinguished
    Predominant_land_use = recode_factor(Predominant_land_use,
                                         "Primary forest" = "Primary",
                                         "Primary non-forest" = "Primary"),

    # indeterminate secondary veg and cannot decide get NA, urban too because it
    # has only 40 sites
    Predominant_land_use = na_if(Predominant_land_use, "Secondary vegetation (indeterminate age)"),
    Predominant_land_use = na_if(Predominant_land_use, "Cannot decide"),
    Predominant_land_use = na_if(Predominant_land_use, "Urban"),


    # set reference levels
    Predominant_land_use = factor(Predominant_land_use),
    Predominant_land_use = relevel(Predominant_land_use, ref = "Primary"),
    Use_intensity = factor(Use_intensity),
    Use_intensity = relevel(Use_intensity, ref = "Minimal use")
  )
```

```r
diversity5 <- diversity4 %>%
  mutate(Use_intensity = str_replace_all(Use_intensity,
                                         pattern = c("Intense use" = "Intense light use",
                                                     "Light use" = "Intense light use")),
         Use_intensity = na_if(Use_intensity, "Cannot decide"),
         Use_intensity = factor(Use_intensity),
         Use_intensity = relevel(Use_intensity, ref = "Minimal use"))
```

3. Test for collinearity

```r
source("https://highstat.com/Books/Book2/HighstatLibV10.R")

corvif(diversity5[ , c("Predominant_land_use", "Use_intensity")])
```

```
##
##
## Variance inflation factors
##
##                          GVIF Df GVIF^(1/2Df)
## Predominant_land_use 1.119901  6     1.009481
## Use_intensity        1.119901  1     1.058254
```

4. Get complete cases, that means dropping the rows that have NA in the columns of total abundance, predominant land use and use intensity

```r
model_data <- drop_na(diversity5,
                      Total_abundance, Predominant_land_use, Use_intensity) %>%
  droplevels()
```

```r
table(model_data$Predominant_land_use, model_data$Use_intensity)
```

```
##
##                                      Minimal use Intense light use
##   Primary                                    739               448
##   Young secondary vegetation                 123                35
##   Intermediate secondary vegetation          138               152
##   Mature secondary vegetation                 69               152
##   Plantation forest                          109               616
##   Pasture                                     21                53
##   Cropland                                    97               146
```

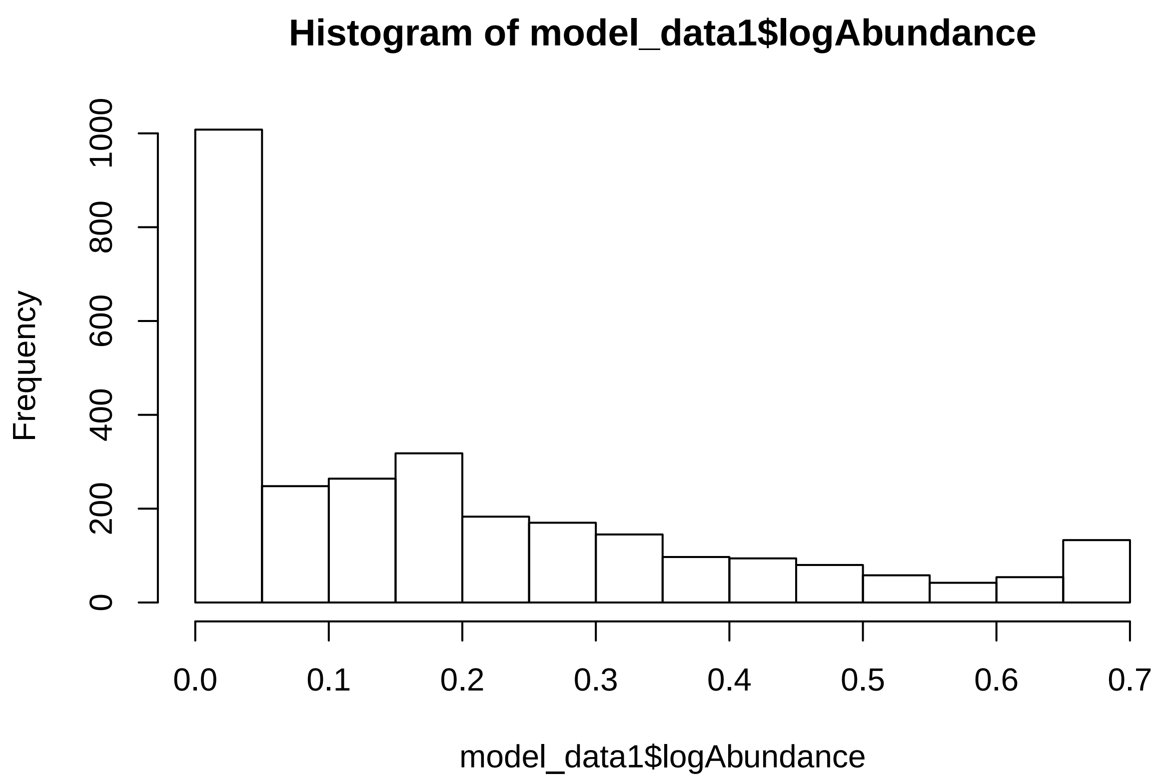
5. Transform abundance measures

Abundance data usually display a nonnormal error distribution because they have a positive mean-variance relationship and are zero-inflated (Purvis et al., 2018). Given that some abudance measures are not integers (some are relative abudance or densities), I am not going to model the abudance with a Poisson distribution, but I'm going to transform it in order to meet the assumptions of linear mixed models.

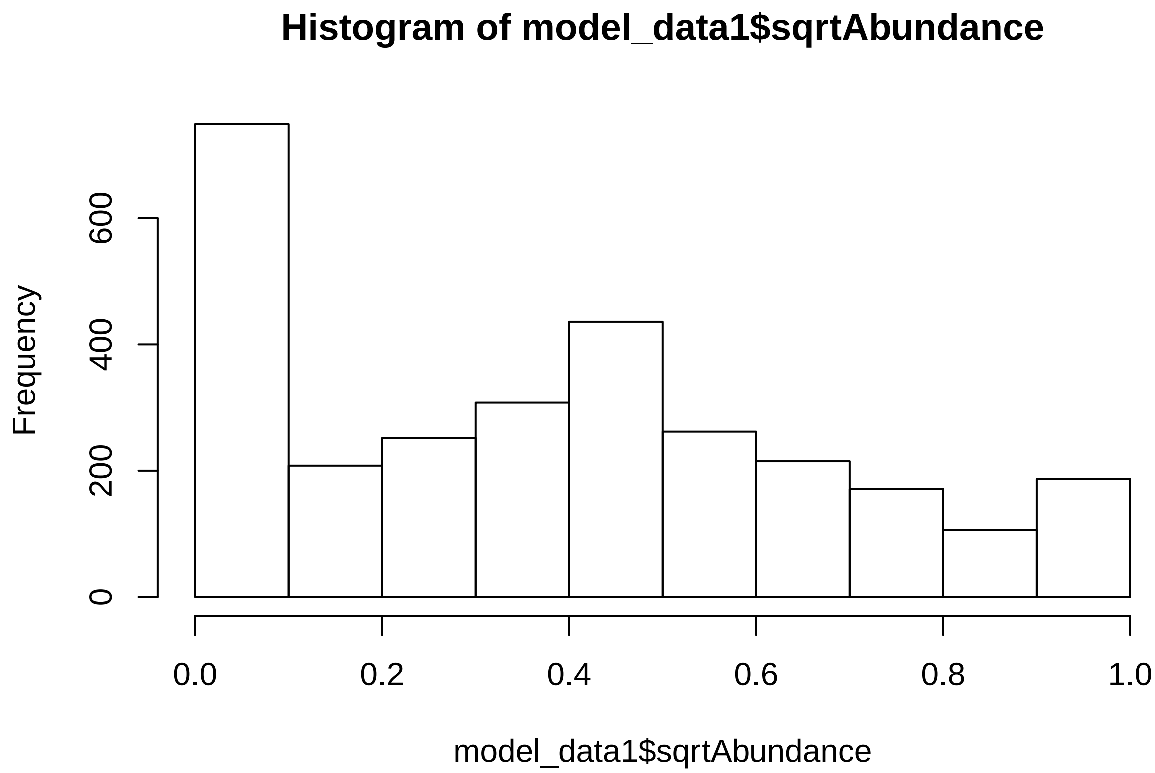
```r
model_data1 <- model_data %>% mutate(logAbundance = log(RescaledAbundance + 1),
                     sqrtAbundance = sqrt(RescaledAbundance)
)
```

6. Data explorations

```
# Distribution of the log rescaled total abundance
hist(model_data1$logAbundance)
```

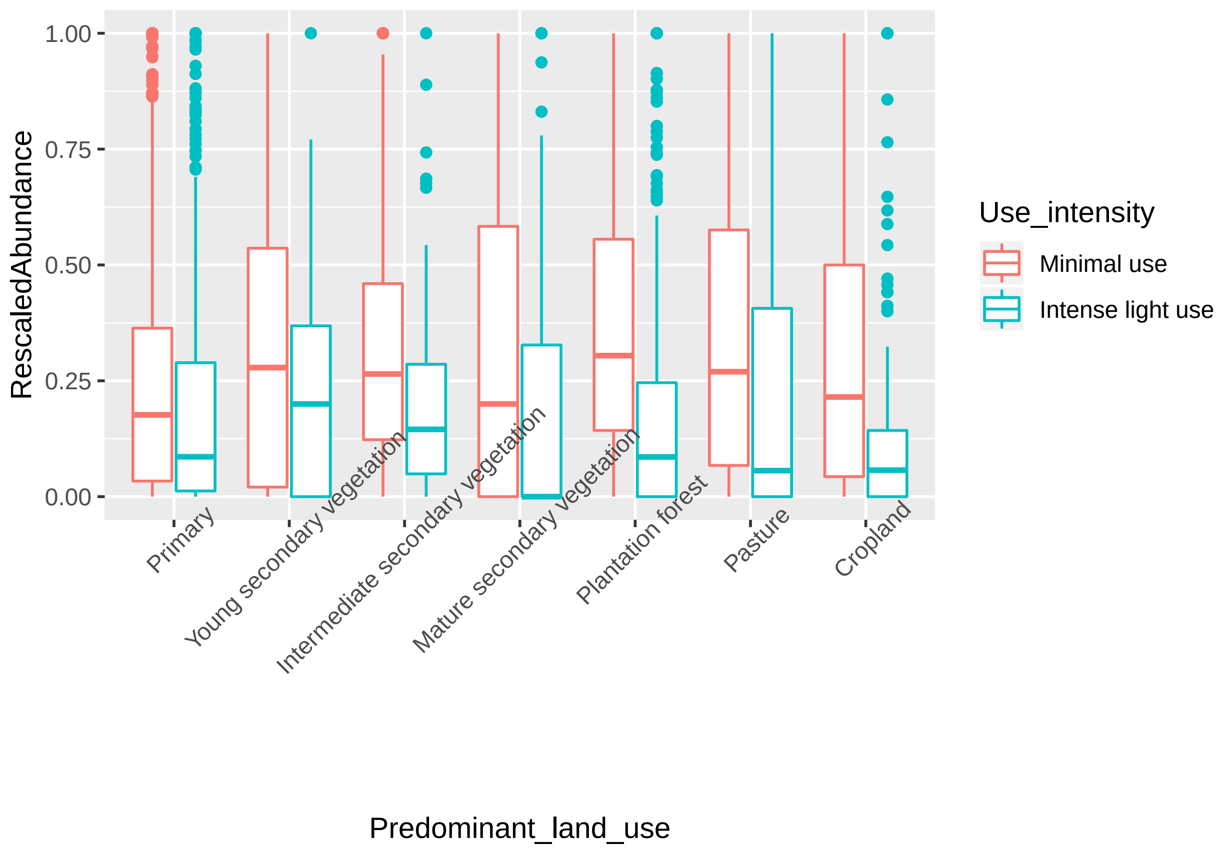



```
# Distribution of the square root of the rescaled total abundance
hist(model_data1$sqrtAbundance)
```




```
# Boxplot showing rescaled total abundance differences between land-use types and intensities
ggplot(model_data1, aes(x=Predominant_land_use, y= RescaledAbundance, color= Use_intensity)) +
  geom_boxplot() + theme(axis.text.x = element_text(size=9, angle=45))
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



Since the distribution looks a little bit more normal using square root, I am going to start with that as response variable

7. Select random effects structure

```
First_model <- lmer(sqrtAbundance ~ Predominant_land_use + Use_intensity +
             Predominant_land_use:Use_intensity +
             (1|SS) + (1|SSB), data = model_data1)
```

```
Second_model <- lmer(sqrtAbundance ~ Predominant_land_use + Use_intensity +
             Predominant_land_use:Use_intensity +
             (1+Predominant_land_use|SS) + (1|SSB), data = model_data1)
```

```
## boundary (singular) fit: see ?isSingular
```

```
Third_model <- lmer(sqrtAbundance ~ Predominant_land_use + Use_intensity +
             Predominant_land_use:Use_intensity +
             (1+Use_intensity|SS) + (1|SSB), data = model_data1)
```

Compare the model using the Akaike's Information Criterion

```
AIC(First_model, Second_model, Third_model)
```

```
##              df       AIC
## First_model  17 116.71920
```

```
## Second_model 44  28.24005
## Third_model  19  57.73412
```

I am going to select the last model as it didn't have a warning and has a lower AIC than the first one.

8. Select fixed effects structure

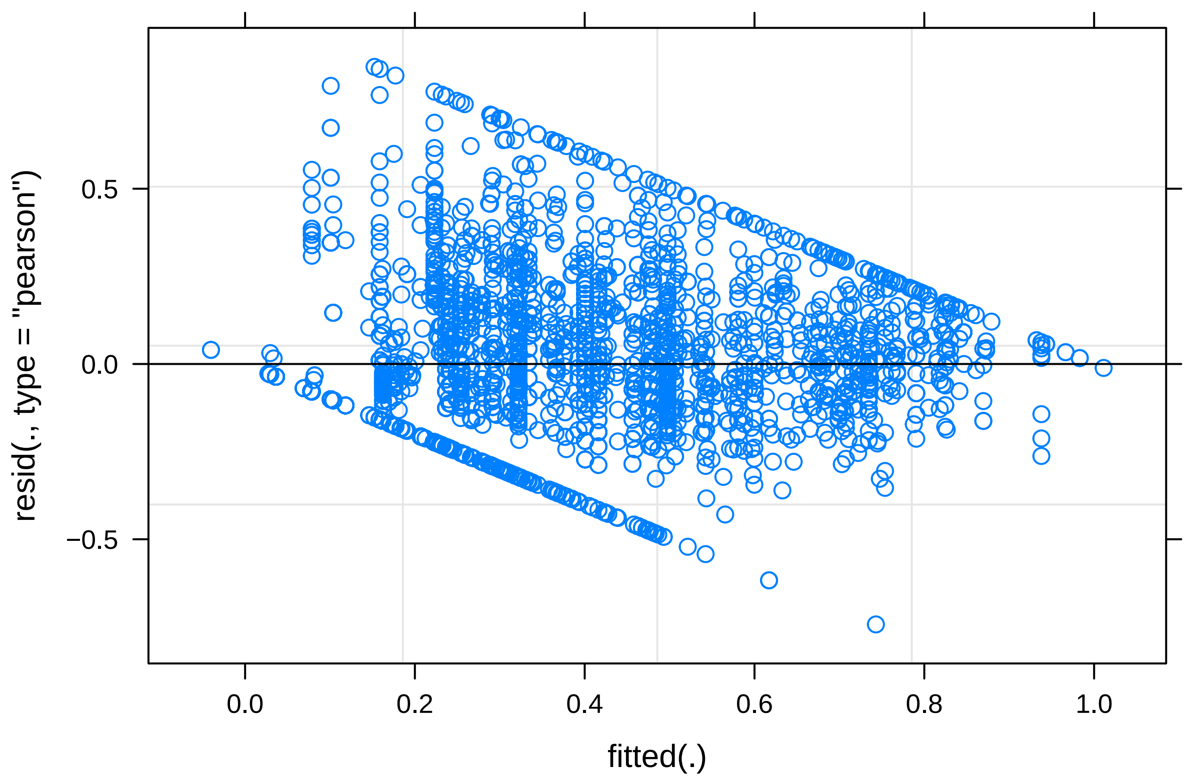
```
Anova(Third_model)
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: sqrtAbundance
##                                      Chisq Df Pr(>Chisq)
## Predominant_land_use               57.7001  6  1.318e-10 ***
## Use_intensity                       2.2688  1      0.132
## Predominant_land_use:Use_intensity 34.2906  6  5.912e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

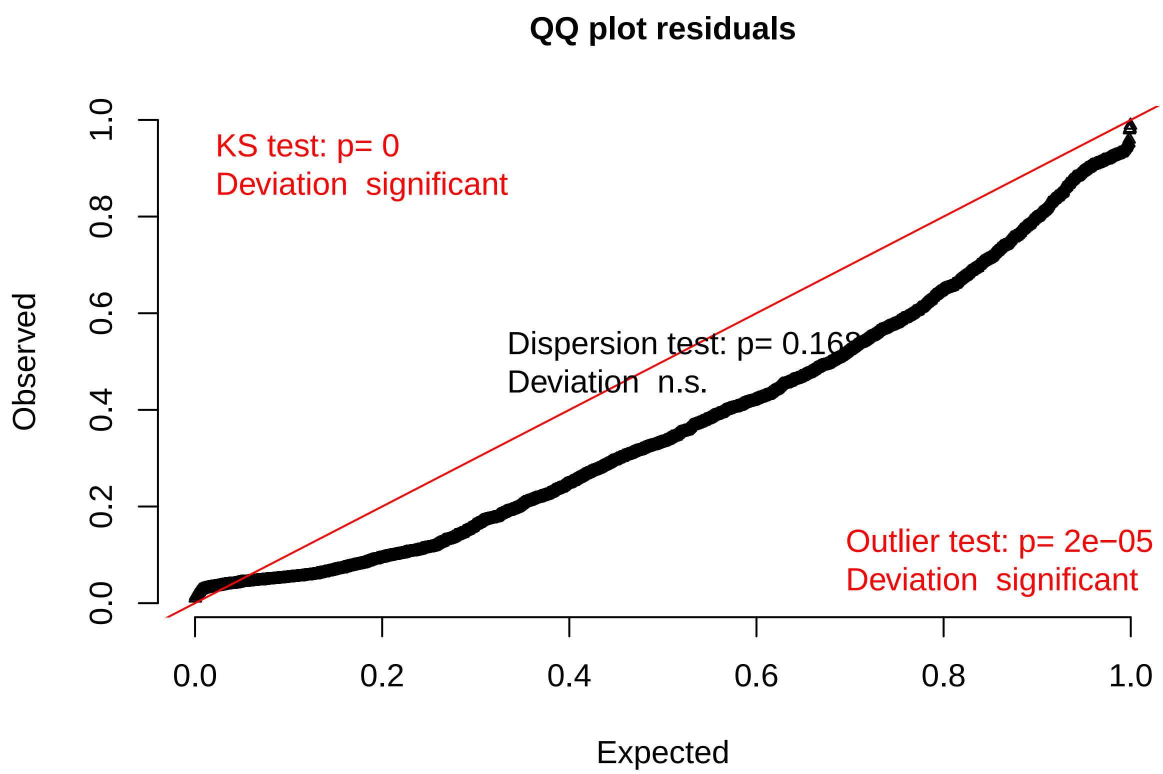
Since the interaction is significant, I am going to leave the two explanatory variables.

9. Plot residuals of the model

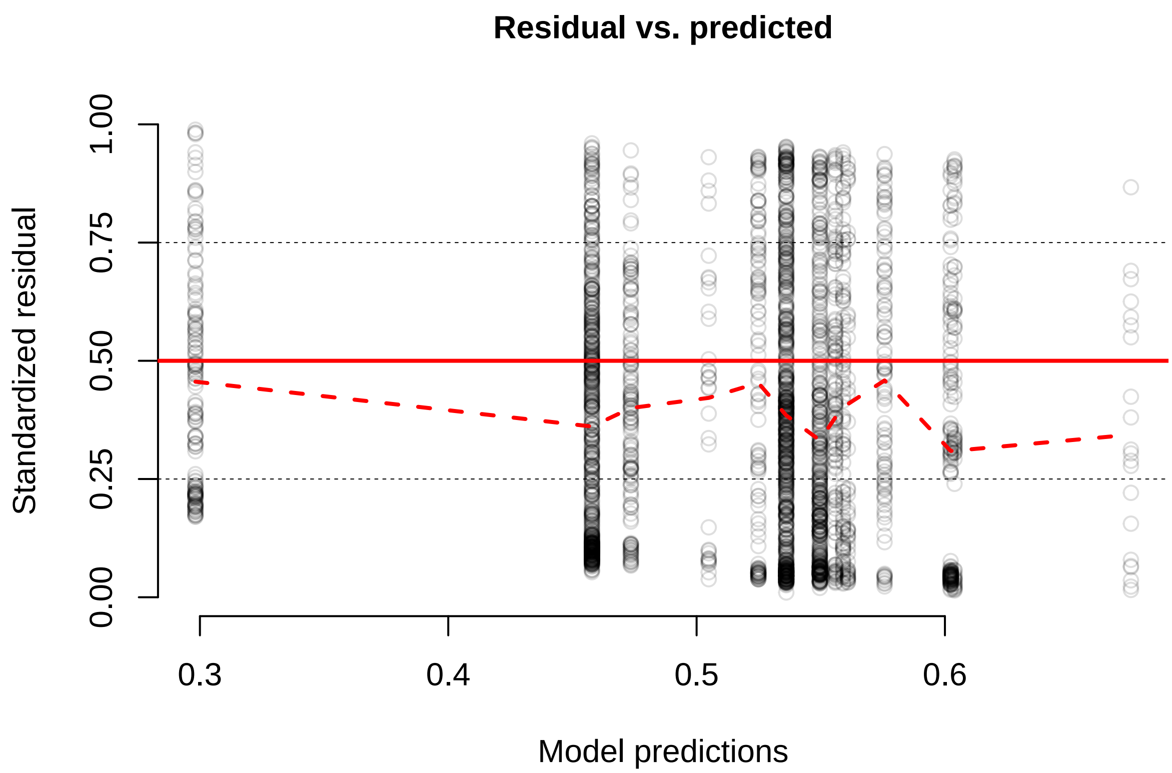
```
plot(Third_model)
```



```r
simulationOutput <- simulateResiduals(fittedModel = Third_model)
# Acces the qq plot
plotQQunif(simulationOutput)
```




```r
# Plot the residuals against the predicted value
plotResiduals(simulationOutput)
```



10. Model estimates

```r
summary(Third_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## sqrtAbundance ~ Predominant_land_use + Use_intensity + Predominant_land_use:Use_intensity +
##     (1 + Use_intensity | SS) + (1 | SSB)
##    Data: model_data1
##
## REML criterion at convergence: 19.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2869 -0.6743 -0.1226  0.5939  3.7496
##
## Random effects:
##  Groups   Name                            Variance Std.Dev. Corr
##  SSB      (Intercept)                     0.004681 0.06842
##  SS       (Intercept)                     0.048039 0.21918
##           Use_intensityIntense light use 0.033339 0.18259  -0.27
##  Residual                                 0.051095 0.22604
## Number of obs: 2894, groups:  SSB, 198; SS, 99
##
## Fixed effects:
##                                                                                  Estimate
```

```
## (Intercept)                                                                         0.53611
## Predominant_land_useYoung secondary vegetation                                     -0.01124
## Predominant_land_useIntermediate secondary vegetation                               0.01976
## Predominant_land_useMature secondary vegetation                                     0.06774
## Predominant_land_usePlantation forest                                               0.03957
## Predominant_land_usePasture                                                         0.13879
## Predominant_land_useCropland                                                        0.02290
## Use_intensityIntense light use                                                      0.01351
## Predominant_land_useYoung secondary vegetation:Use_intensityIntense light use      -0.03350
## Predominant_land_useIntermediate secondary vegetation:Use_intensityIntense light use -0.09585
## Predominant_land_useMature secondary vegetation:Use_intensityIntense light use     -0.01506
## Predominant_land_usePlantation forest:Use_intensityIntense light use               -0.13135
## Predominant_land_usePasture:Use_intensityIntense light use                         -0.12749
## Predominant_land_useCropland:Use_intensityIntense light use                        -0.27434
##                                                                                    Std. Error
## (Intercept)                                                                           0.02707
## Predominant_land_useYoung secondary vegetation                                        0.02887
## Predominant_land_useIntermediate secondary vegetation                                 0.02646
## Predominant_land_useMature secondary vegetation                                       0.04884
## Predominant_land_usePlantation forest                                                 0.03349
## Predominant_land_usePasture                                                           0.06532
## Predominant_land_useCropland                                                          0.03560
## Use_intensityIntense light use                                                        0.03395
## Predominant_land_useYoung secondary vegetation:Use_intensityIntense light use         0.05897
## Predominant_land_useIntermediate secondary vegetation:Use_intensityIntense light use   0.04387
## Predominant_land_useMature secondary vegetation:Use_intensityIntense light use        0.06220
## Predominant_land_usePlantation forest:Use_intensityIntense light use                  0.04516
## Predominant_land_usePasture:Use_intensityIntense light use                            0.07737
## Predominant_land_useCropland:Use_intensityIntense light use                           0.05165
##                                                                                    t value
## (Intercept)                                                                         19.805
## Predominant_land_useYoung secondary vegetation                                      -0.389
## Predominant_land_useIntermediate secondary vegetation                                0.747
## Predominant_land_useMature secondary vegetation                                      1.387
## Predominant_land_usePlantation forest                                                1.182
## Predominant_land_usePasture                                                          2.125
## Predominant_land_useCropland                                                         0.643
## Use_intensityIntense light use                                                       0.398
## Predominant_land_useYoung secondary vegetation:Use_intensityIntense light use       -0.568
## Predominant_land_useIntermediate secondary vegetation:Use_intensityIntense light use  -2.185
## Predominant_land_useMature secondary vegetation:Use_intensityIntense light use      -0.242
## Predominant_land_usePlantation forest:Use_intensityIntense light use                -2.909
## Predominant_land_usePasture:Use_intensityIntense light use                          -1.648
## Predominant_land_useCropland:Use_intensityIntense light use                         -5.311
```

```
##
## Correlation matrix not shown by default, as p = 14 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

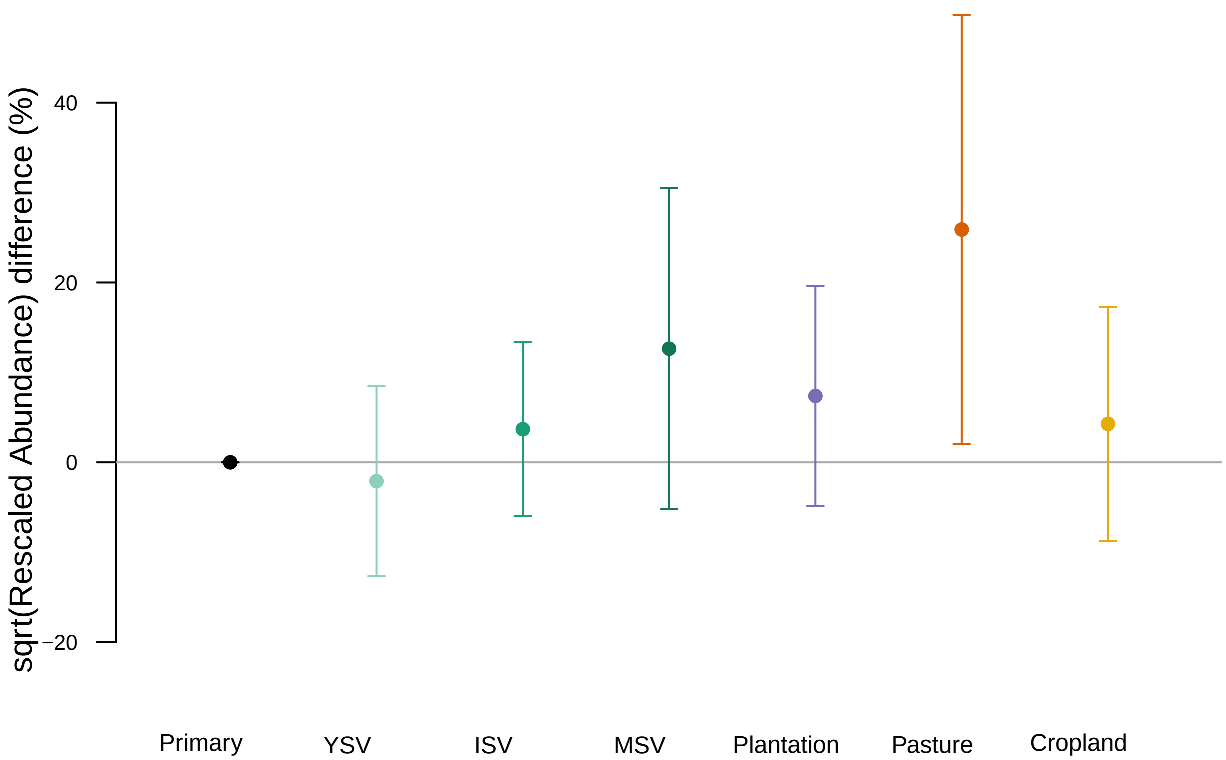
According to the model estimates:

- The average square root of the rescaled abundance in Primary forest with minimal use is 0.536

- The average square root of the rescaled abundance in minimally-used Pasture is 0.14 higher. Meaning that the abundance in minimally-used Pasture is 0.54+0.14=0.68
- The average square root of the rescaled abundance in Intermediate secondary vegetation with an intense-light use is 0.09 lower. Meaning that the abundance in Intermediate secondary vegetation with an intense-light use is 0.54-0.09 = 0.45
- The average square root of the rescaled abundance in Plantation forest with an intense-light use is 0.13 lower. Meaning that the abundance in Plantation forest with an intense-light use is 0.54-0.13 = 0.41
- The average square root of the rescaled abundance in Crops with an intense-light use is 0.27 lower. Meaning that the abundance in Crops with an intense-light use is 0.54-0.13 = 0.27

11. Plot the results

```
roquefort::PlotErrBar(model = Third_model,
                      data = model_data5,
                      responseVar = "sqrt(Rescaled Abundance)",
                      seMultiplier = 1.96,
                      secdAge = TRUE,
                      logLink = "n",
                      catEffects = c("Predominant_land_use"),
                      forPaper = TRUE,
                      plotLabels = FALSE)
```



12. Run the models with the log of Abundance

```r
First_model2 <- lmer(logAbundance ~ Predominant_land_use + Use_intensity +
               Predominant_land_use:Use_intensity +
               (1|SS) + (1|SSB), data = model_data1)
```

```r
Second_model2 <- lmer(logAbundance ~ Predominant_land_use + Use_intensity +
               Predominant_land_use:Use_intensity +
               (1+Predominant_land_use|SS) + (1|SSB), data = model_data1)
```

```
## boundary (singular) fit: see ?isSingular
```

```r
Third_model2 <- lmer(logAbundance ~ Predominant_land_use + Use_intensity +
               Predominant_land_use:Use_intensity +
               (1+Use_intensity|SS) + (1|SSB), data = model_data1)
```

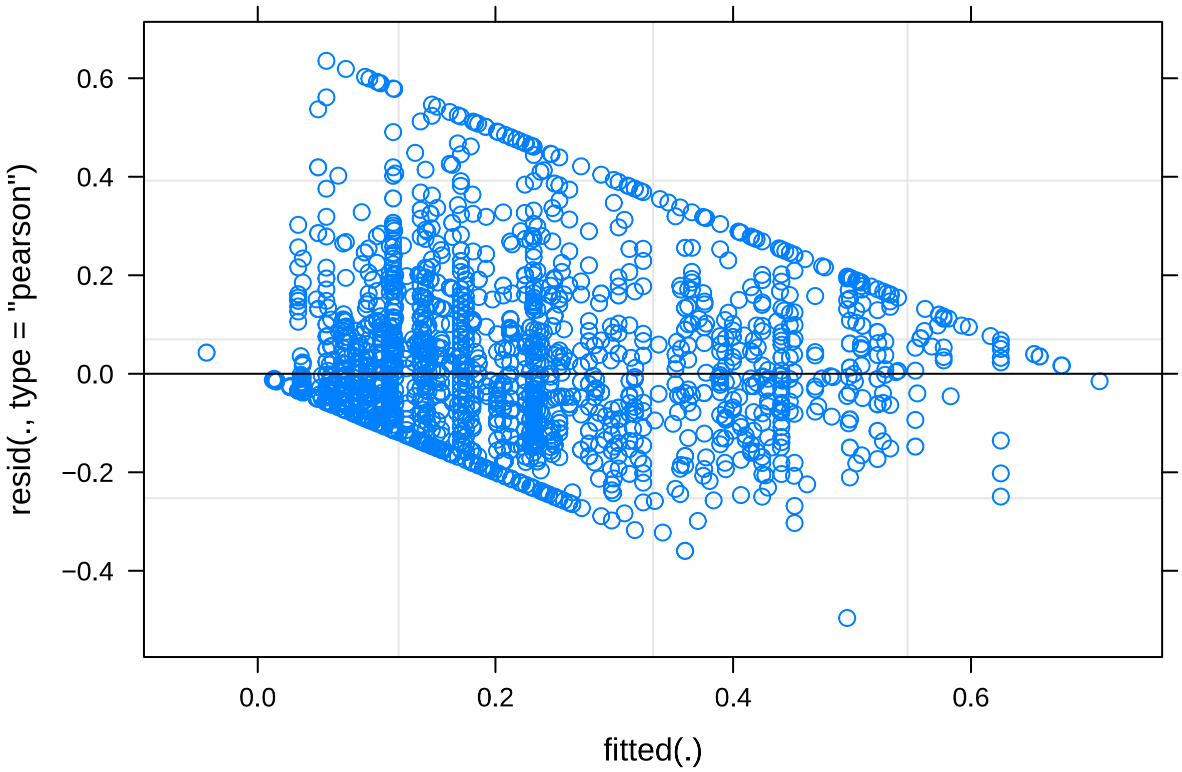
```r
# Choose best random effects structure
AIC(First_model2, Second_model2, Third_model2)
```

```
##               df       AIC
## First_model2  17 -2333.459
## Second_model2 44 -2449.155
## Third_model2  19 -2402.798
```

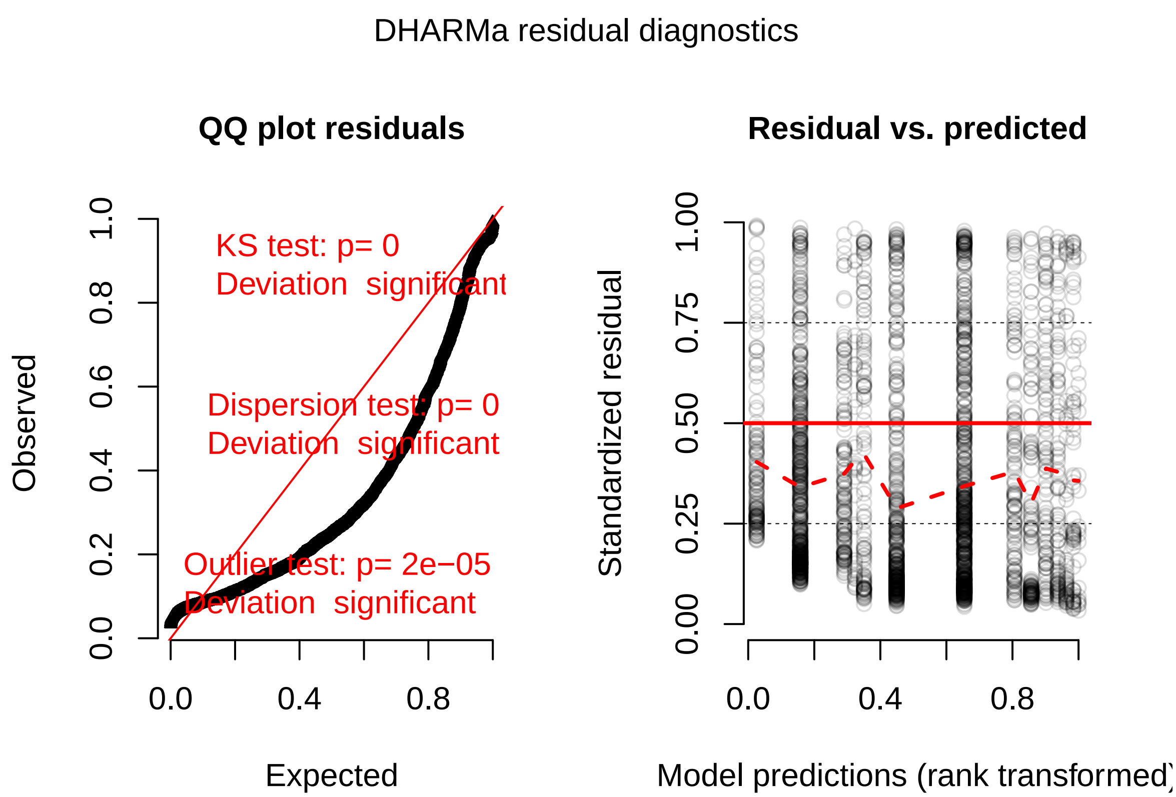
```r
# See the significance of fixed variables
Anova(Third_model2)
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: logAbundance
##                                      Chisq Df Pr(>Chisq)
## Predominant_land_use               38.9000  6  7.488e-07 ***
## Use_intensity                       3.2822  1   0.070034 .
## Predominant_land_use:Use_intensity 26.0602  6   0.000217 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

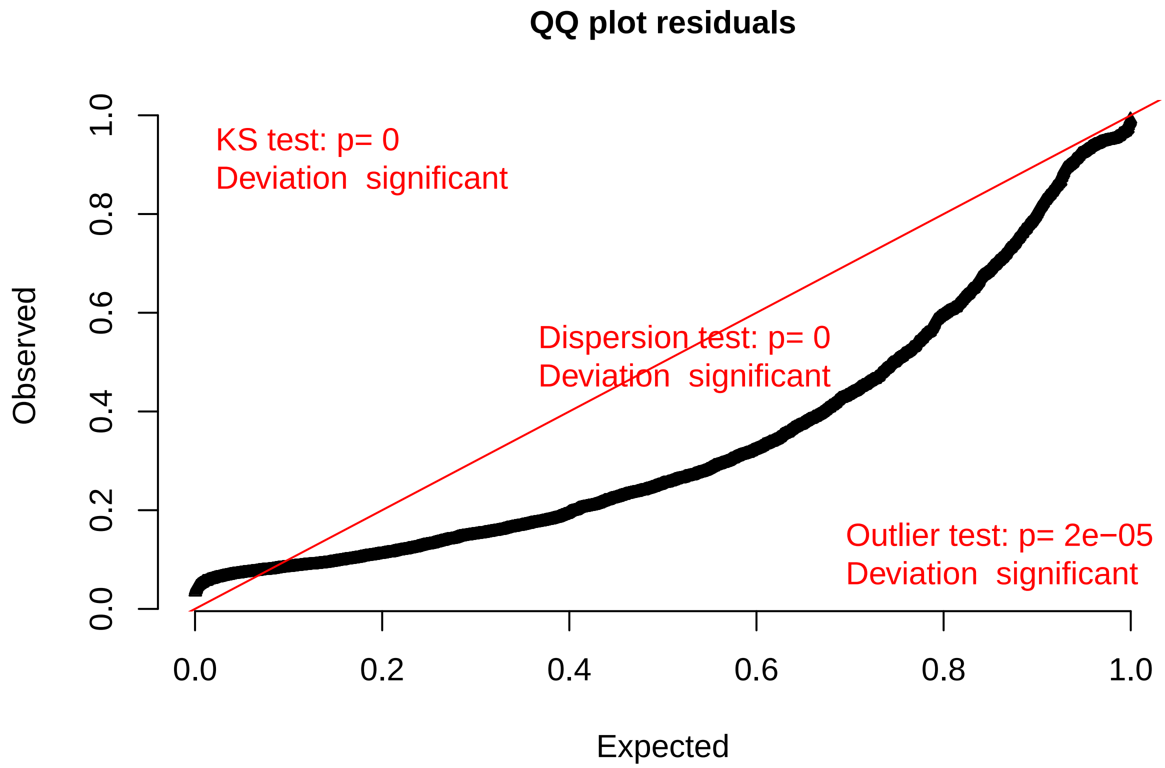
```r
plot(Third_model2)
```



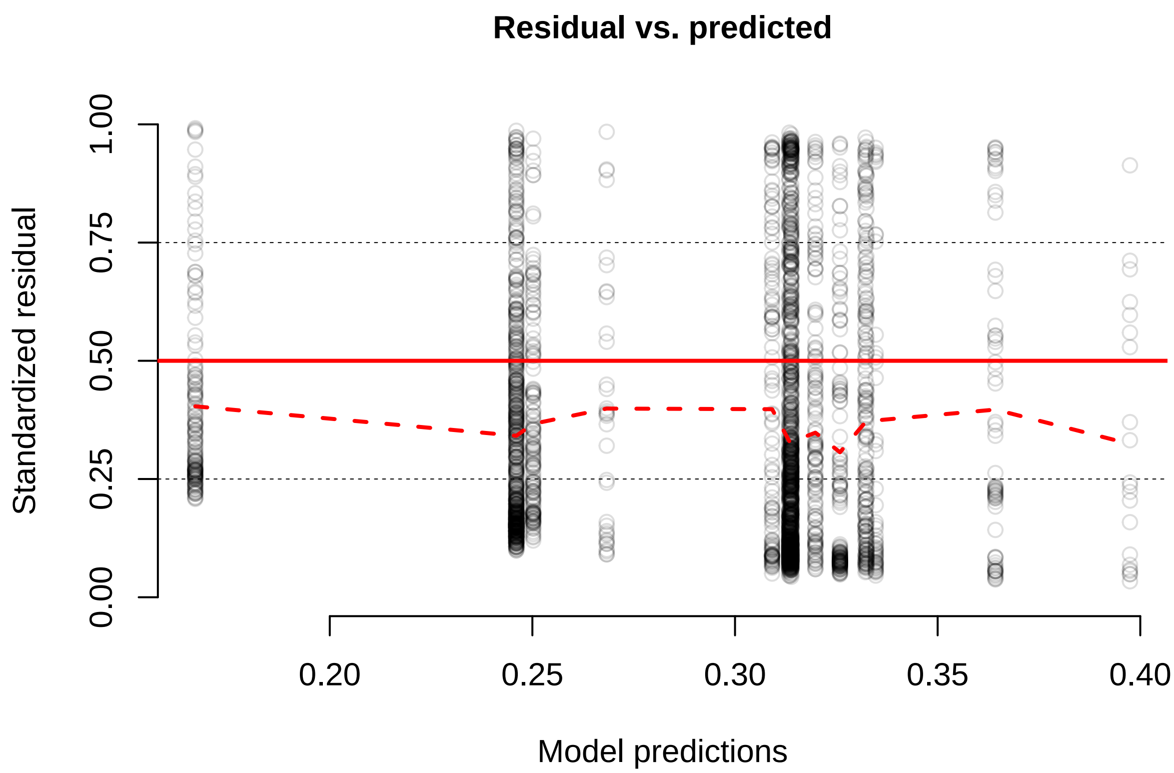

```r
# Simulate the residuals and plot them
simulationOutput1 <- simulateResiduals(fittedModel = Third_model2, plot = T)
```




```
# Acces the qq plot
plotQQunif(simulationOutput1)
```




```
# Plot the residuals against the predicted value
plotResiduals(simulationOutput1)
```




```r
# model estimates
summary(Third_model2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## logAbundance ~ Predominant_land_use + Use_intensity + Predominant_land_use:Use_intensity +
##     (1 + Use_intensity | SS) + (1 | SSB)
##    Data: model_data1
##
## REML criterion at convergence: -2440.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3821 -0.6297 -0.2246  0.4384  4.3335
##
## Random effects:
##  Groups   Name                          Variance Std.Dev. Corr
##  SSB      (Intercept)                   0.001686 0.04106
##  SS       (Intercept)                   0.029343 0.17130
##           Use_intensityIntense light use 0.021546 0.14679  -0.45
##  Residual                               0.021496 0.14662
## Number of obs: 2894, groups:  SSB, 198; SS, 99
##
## Fixed effects:
##                                                                                                 Estimate
## (Intercept)                                                                                    0.3138800
```

```
## Predominant_land_useYoung secondary vegetation                                  -0.0047272
## Predominant_land_useIntermediate secondary vegetation                            0.0059535
## Predominant_land_useMature secondary vegetation                                  0.0503693
## Predominant_land_usePlantation forest                                            0.0183064
## Predominant_land_usePasture                                                      0.0835796
## Predominant_land_useCropland                                                     0.0185618
## Use_intensityIntense light use                                                  -0.0004478
## Predominant_land_useYoung secondary vegetation:Use_intensityIntense light use   -0.0403715
## Predominant_land_useIntermediate secondary vegetation:Use_intensityIntense light use -0.0691870
## Predominant_land_useMature secondary vegetation:Use_intensityIntense light use  -0.0378970
## Predominant_land_usePlantation forest:Use_intensityIntense light use            -0.0856998
## Predominant_land_usePasture:Use_intensityIntense light use                      -0.0622889
## Predominant_land_useCropland:Use_intensityIntense light use                     -0.1651837
##                                                                                  Std. Error
## (Intercept)                                                                       0.0202547
## Predominant_land_useYoung secondary vegetation                                    0.0188836
## Predominant_land_useIntermediate secondary vegetation                             0.0173070
## Predominant_land_useMature secondary vegetation                                   0.0324543
## Predominant_land_usePlantation forest                                             0.0220418
## Predominant_land_usePasture                                                       0.0432806
## Predominant_land_useCropland                                                      0.0235288
## Use_intensityIntense light use                                                    0.0246133
## Predominant_land_useYoung secondary vegetation:Use_intensityIntense light use     0.0385663
## Predominant_land_useIntermediate secondary vegetation:Use_intensityIntense light use  0.0288735
## Predominant_land_useMature secondary vegetation:Use_intensityIntense light use    0.0412839
## Predominant_land_usePlantation forest:Use_intensityIntense light use              0.0298384
## Predominant_land_usePasture:Use_intensityIntense light use                        0.0512020
## Predominant_land_useCropland:Use_intensityIntense light use                       0.0340985
##                                                                                  t value
## (Intercept)                                                                       15.497
## Predominant_land_useYoung secondary vegetation                                    -0.250
## Predominant_land_useIntermediate secondary vegetation                              0.344
## Predominant_land_useMature secondary vegetation                                    1.552
## Predominant_land_usePlantation forest                                              0.831
## Predominant_land_usePasture                                                        1.931
## Predominant_land_useCropland                                                       0.789
## Use_intensityIntense light use                                                    -0.018
## Predominant_land_useYoung secondary vegetation:Use_intensityIntense light use     -1.047
## Predominant_land_useIntermediate secondary vegetation:Use_intensityIntense light use  -2.396
## Predominant_land_useMature secondary vegetation:Use_intensityIntense light use    -0.918
## Predominant_land_usePlantation forest:Use_intensityIntense light use              -2.872
## Predominant_land_usePasture:Use_intensityIntense light use                        -1.217
## Predominant_land_useCropland:Use_intensityIntense light use                       -4.844
```

```
##
## Correlation matrix not shown by default, as p = 14 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```r
roquefort::PlotErrBar(model = Third_model2,
                      data = model_data5,
                      responseVar = "log Abundance",
                      seMultiplier = 1.96,
                      secdAge = TRUE,
```

```
                      logLink = "n",
                      catEffects = c("Predominant_land_use", "Use_intensity"),
                      forPaper = TRUE,
                      plotLabels = FALSE)
```