Accounting for spatial variability in forest genetic trials using breedR: a case study with black poplar

BreedR Workshop
June 30 – July 3, 2015
Jaca, Spain

Mesfin N. Gebreselassie
PhD student (1st year)
Research Unit: UAGPF

Supervisors:
- **Dr. Catherine Bastien**
- **Dr. Vincent Segura**

PhD Thesis: **Genetic architecture of lignocellulosic biomass yield and quality in black poplar for its use in biorefinery**

General objective:

- **To elucidate the <u>molecular and genetic basis</u> of <u>biomass yield and quality</u> in European black poplar using large population size, contrasting environments and modern statistical tools.**

Specific objectives:

(1) To quantify the genetic variation and the heritability of components of lignocellulosic biomass yield and quality.

(2) To analyze interactions between the genetic variability of the target traits and environment.

(3) To assess the existence of any trade-offs between biomass quantity and quality and to identify the genetic basis of significant trait correlations.

(4) To identify and locate genetic polymorphisms that control the genetic variability of biomass yield and quality related traits.

Populus nigra association mapping population and a field trial:

- **Population:** Association mapping population of European black poplar: ~1000 *populus nigra* clones /genotypes

- **Experimental sites:** Orléans (France) within the NovelTree project

- **Experimental design:** the trial was established using a randomized complete block design (RCBD) with 6 replications.

- **Traits of interest:** Bioenergy related traits
  - Components of biomass yield,
  - Biomass yield, and
  - Biomass quality

Figure 1. Geographical origin (France, Italy, Germany, Netherlands) of the subpopulations constituting the *Populus nigra* association population and location of a field trial
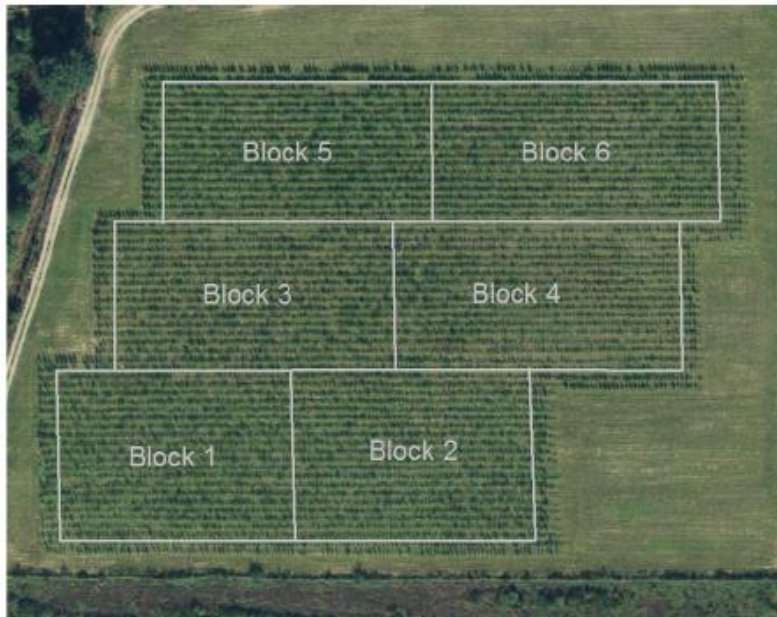
Methodological approaches:
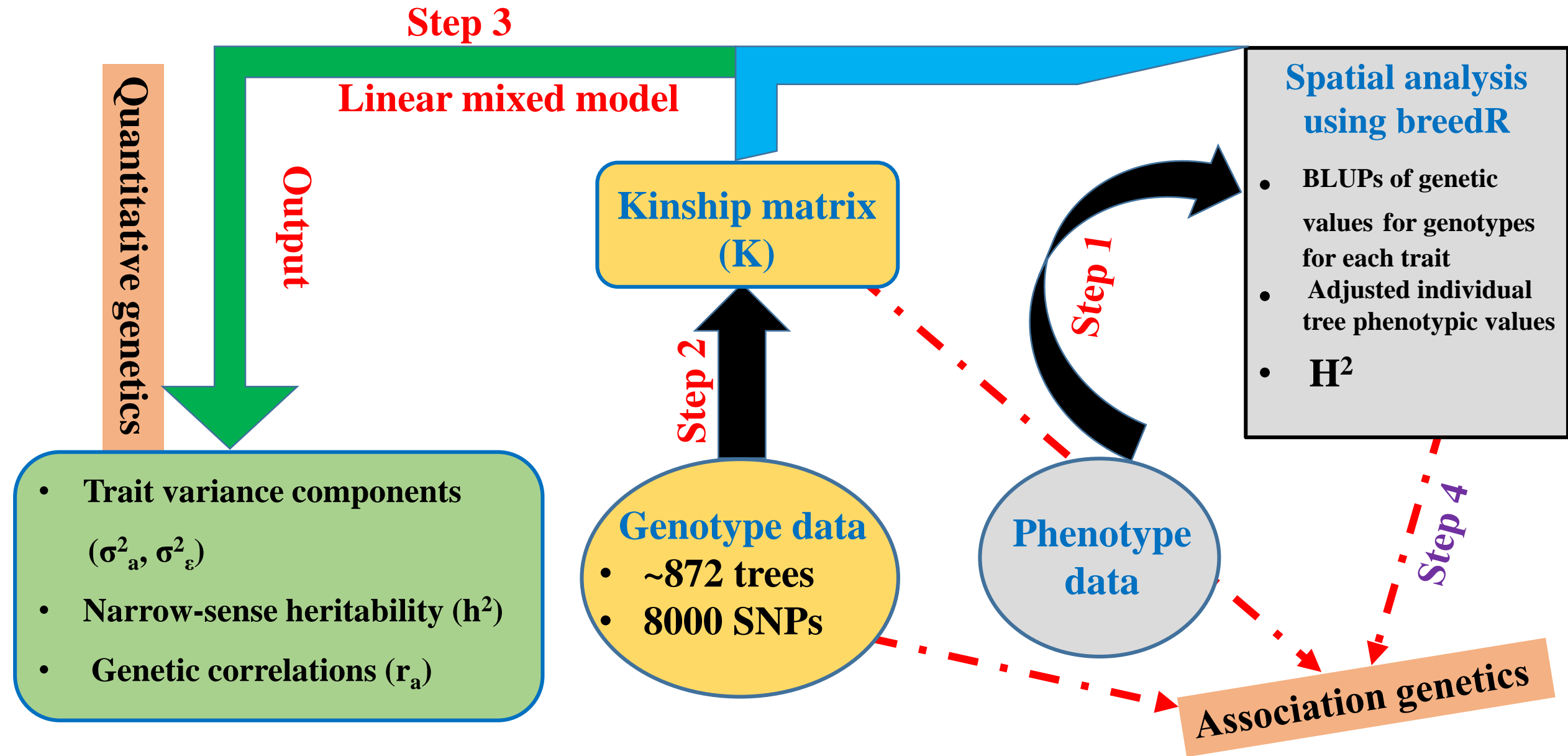
Accounting for spatial variability using breedR

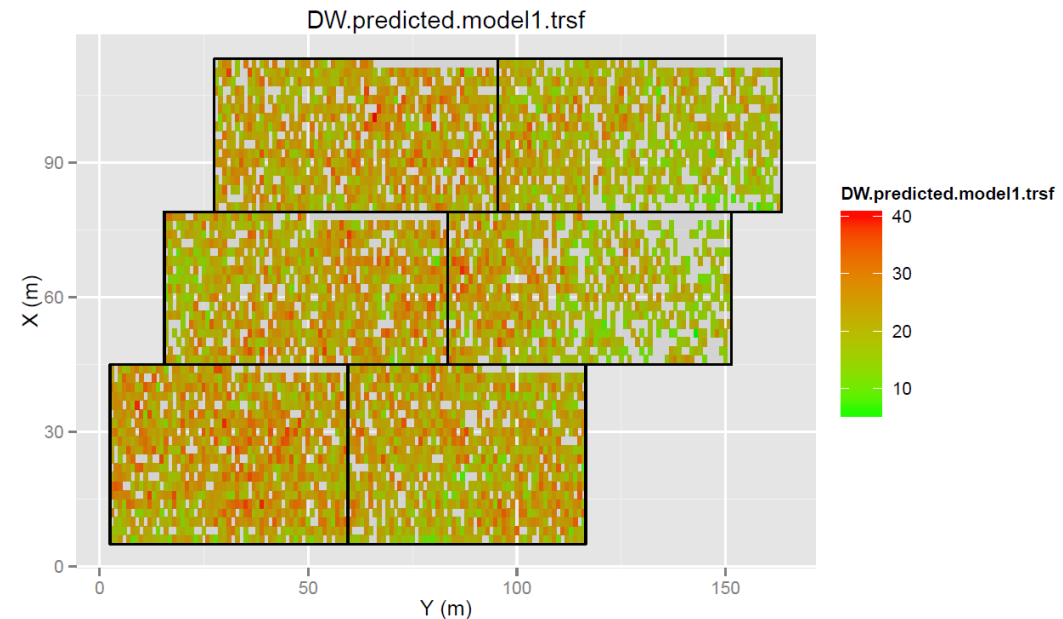## 1.   Introduction

❑ Why do we care about spatial effects?

❑ Is RCBD efficient in capturing variability in the field test when we have large genetic entries of forest trees?

Aerial view of the field trial

2D map of phenotype

2. Statistical method

❑ Individual tree data from Orléans experimental design were analysed by a linear mixed model (Henderson, 1984) with or without a spatial effect using breedR (Muñoz and Sanchez, 2014).

❑ The model fit to the data followed the general linear mixed model of the form:

$$\text{Classical mixed model: } Y = X\beta + Zu + \varepsilon \qquad (1)$$

$$\text{Spatial mixed model: } \quad Y = X\beta + Zu + s + e \quad (2)$$

Where;

Y is a vector of phenotypic values,
β is a vector of fixed effects of blocks,
u is a vector of random effects of genotypes,
X and Z are design matrices relating the observations to the fixed and random effects, respectively.
s is a spatially dependent random error vector, and
e is a spatially independent random residual vector.

Statistical method cont'd . . .

Classical model: $Y = \boxed{X\beta} + \boxed{Zu + \varepsilon}$ (1)

Spatial model: $Y = \boxed{X\beta} + \boxed{Zu + s + e}$ (2)

                 fixed   **random**

$$u \sim N(0, \sigma^2_g I) \qquad s \sim N(0, \sigma^2_s H) \qquad e \sim N(0, \sigma^2_e I)$$

$$H = [\mathbf{AR1}(\rho_{col}) \otimes \mathbf{AR1}(\rho_{row})]$$

- **AR1(ρ)** represents a first-order autoregressive correlation matrix which, for ordered coordinates of size n, has the form:

$$AR1(\rho) = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^n \\ \rho & 1 & \rho & \cdots & \\ \rho^2 & \rho & 1 & \cdots & \\ \vdots & \vdots & \vdots & \ddots & \\ \rho^n & \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

3. Trait to be analysed for illustration of the new statistical package breedR: "predicted dry biomass yield" from Orléans experimental design

Developing prediction model: Mathematical relationships between various morphological descriptors and biomass yield have been assessed.

Morphological descriptors measured on all trees:

- Shoot length 2010, 2011

- Height 2009, 2010, 2011

- Circumference 2009, 2010, 2011

- Predicted date for bud set score

- Predicted date for bud flush score

- Sylleptic ramification score

- Tree architectural traits

Dry biomass yield measured on biomass subsamples:

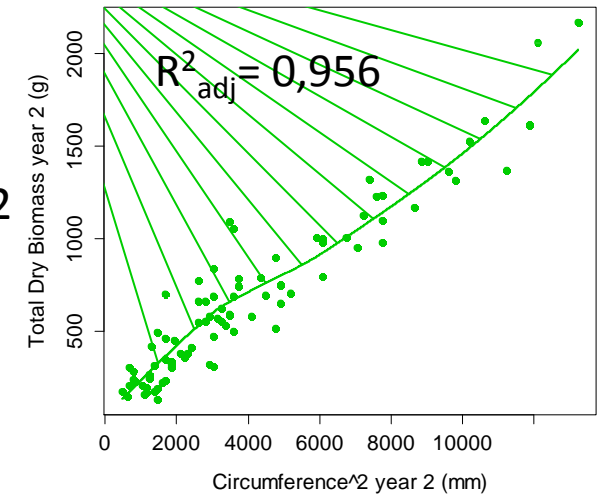- 30 genotypes * 3 replications

Trait to be analysed cont'd . . .

❑ Trait: DW.predicted.model1

- "tree total dry biomass weight" was predicted based on biomass components measured at Orleans in 2011 (2nd year of the 2nd production cycle).

- **model1: DW.total ~ -1 + HT2011+ Circ2011.trsf + date15.fitted.doy**

« Circumference year 2 »
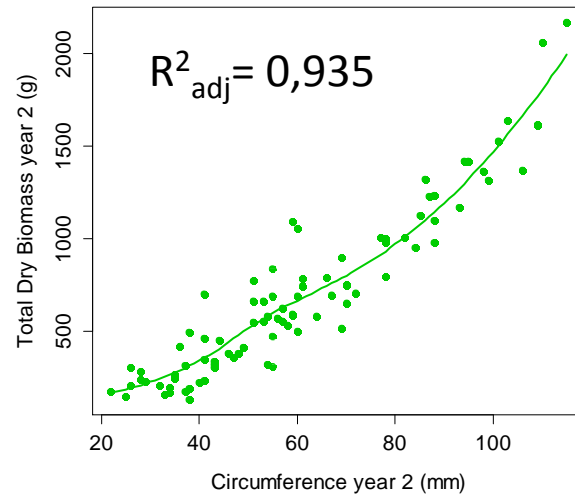
« Circumference year 2 » ^2

4. Spatial and genetic analysis of *Populus nigra* association mapping experiment using breedR

❑ **A linear mixed model approach involving spatial effects was applied**

❑ **The objectives of the spatial analysis were to obtain:**

- an accurate genotypic value, i.e., adjusted for any micro-environmental effect

- an adjusted individual tree phenotypic value

- an accurate estimate of broad-sense heritability (H^2)
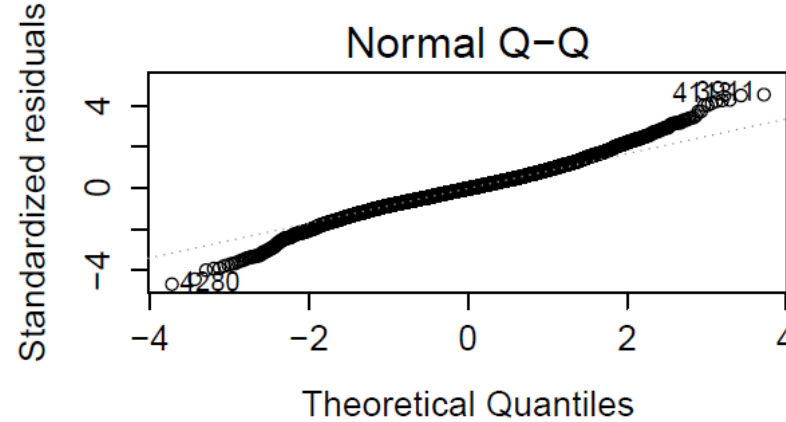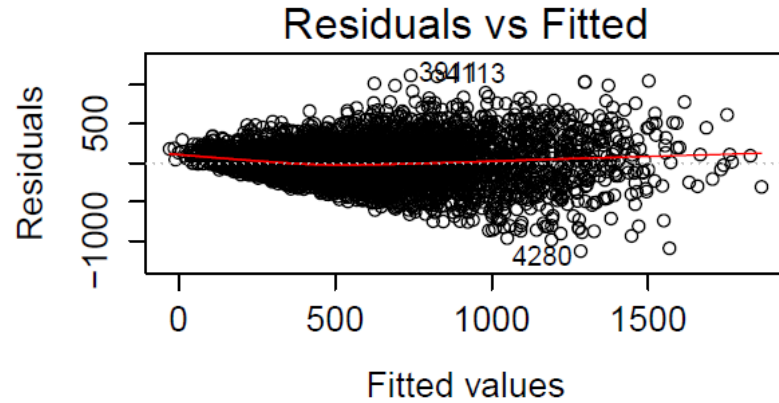
- an accurate estimate of the genetic variation of the trait

Data exploration:

❑ Histogram of raw phenotype data: the normal distribution of the phenotype was evaluated

> hist(data_ok$DW.predicted.model1, col = "grey", xlab = "DW.predicted.model1", main = "")
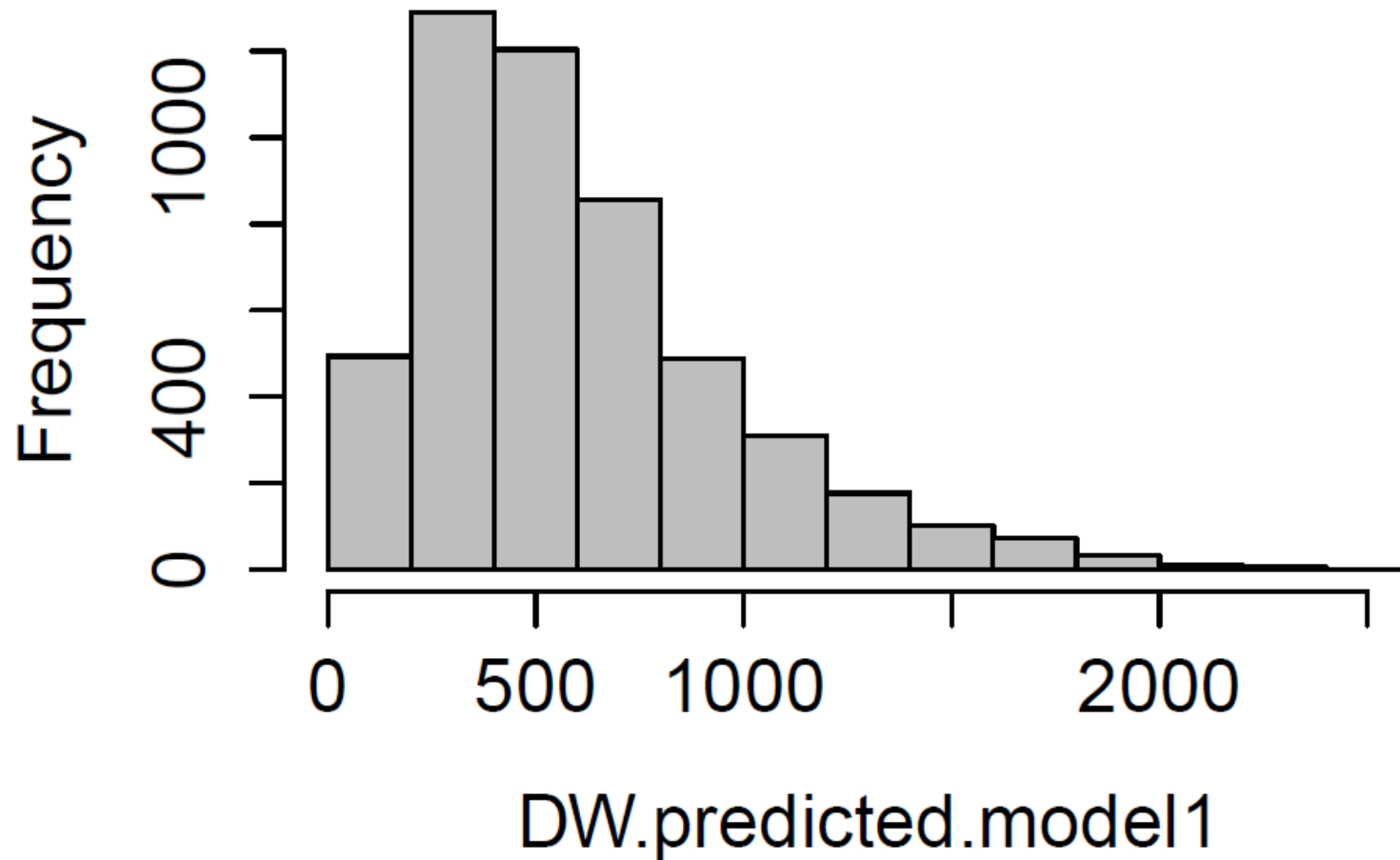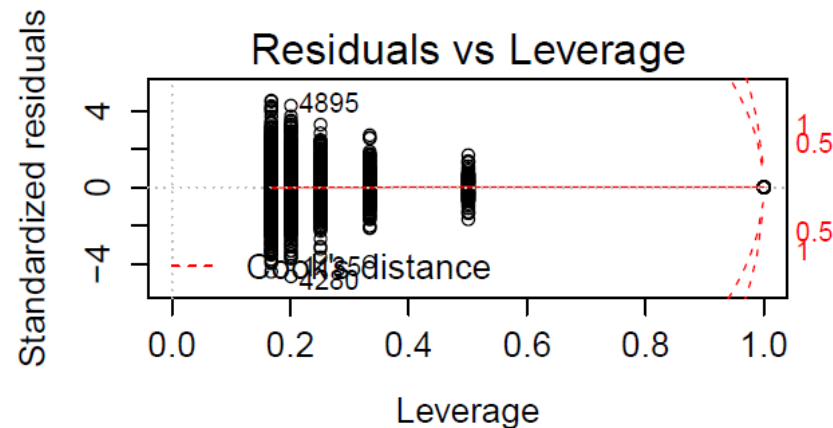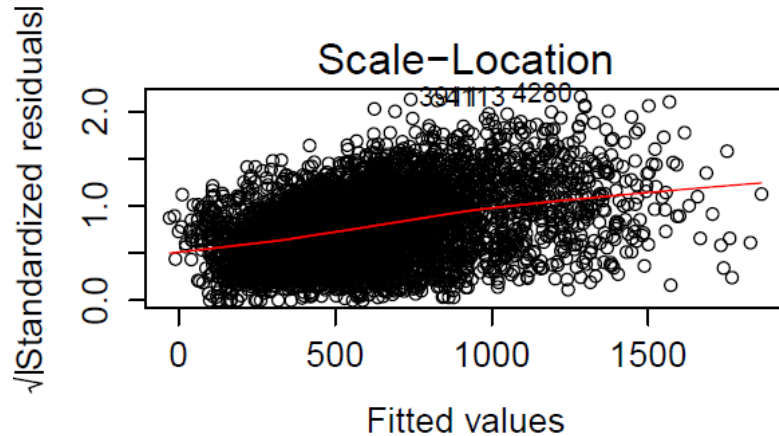
Fixed effect ANOVA model:

Model QC: Assumptions on distribution of residuals of the ANOVA model were checked

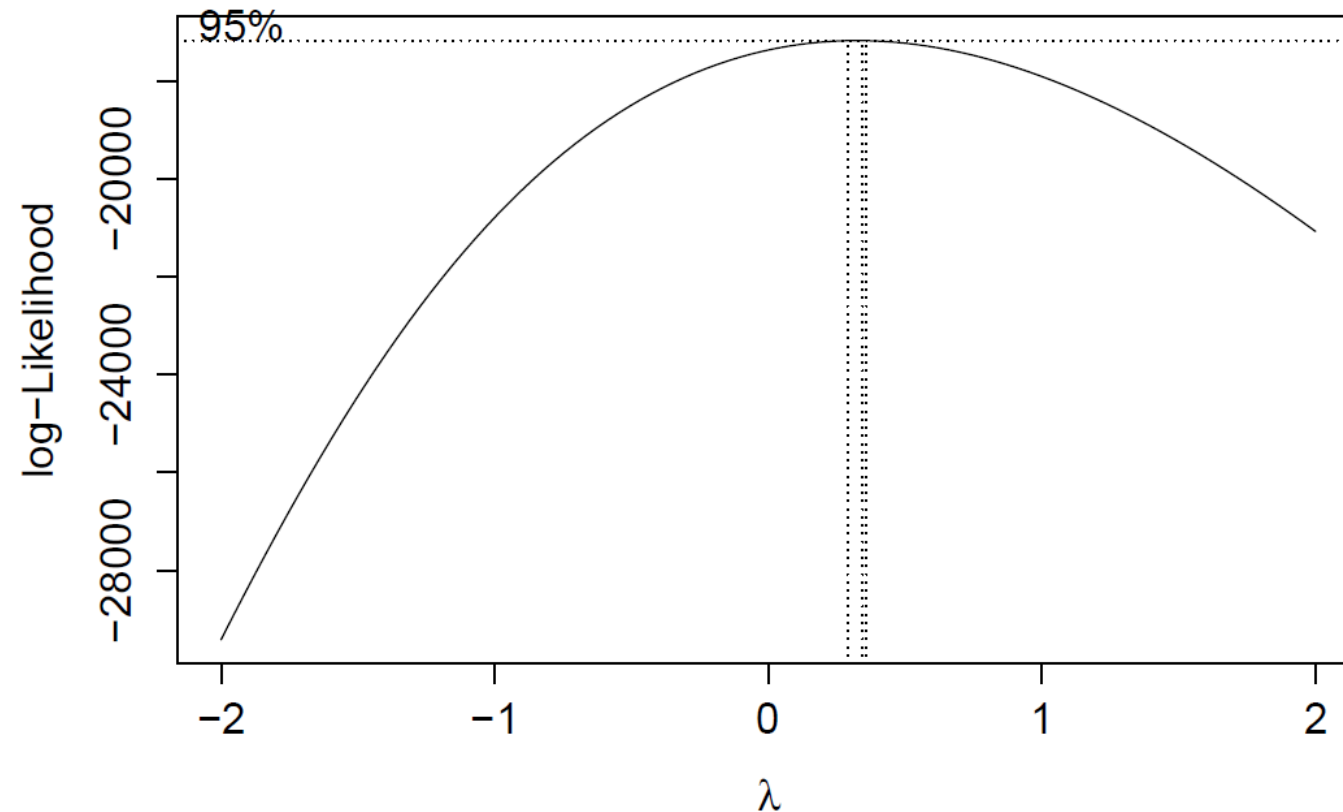
```
>op <- par(mfrow = c(2, 2))
>plot(linmod)
>par(op)
```

ANOVA model cont'd . . .

Boxcox transformation: was used to normalize the predicted total dry biomass yield
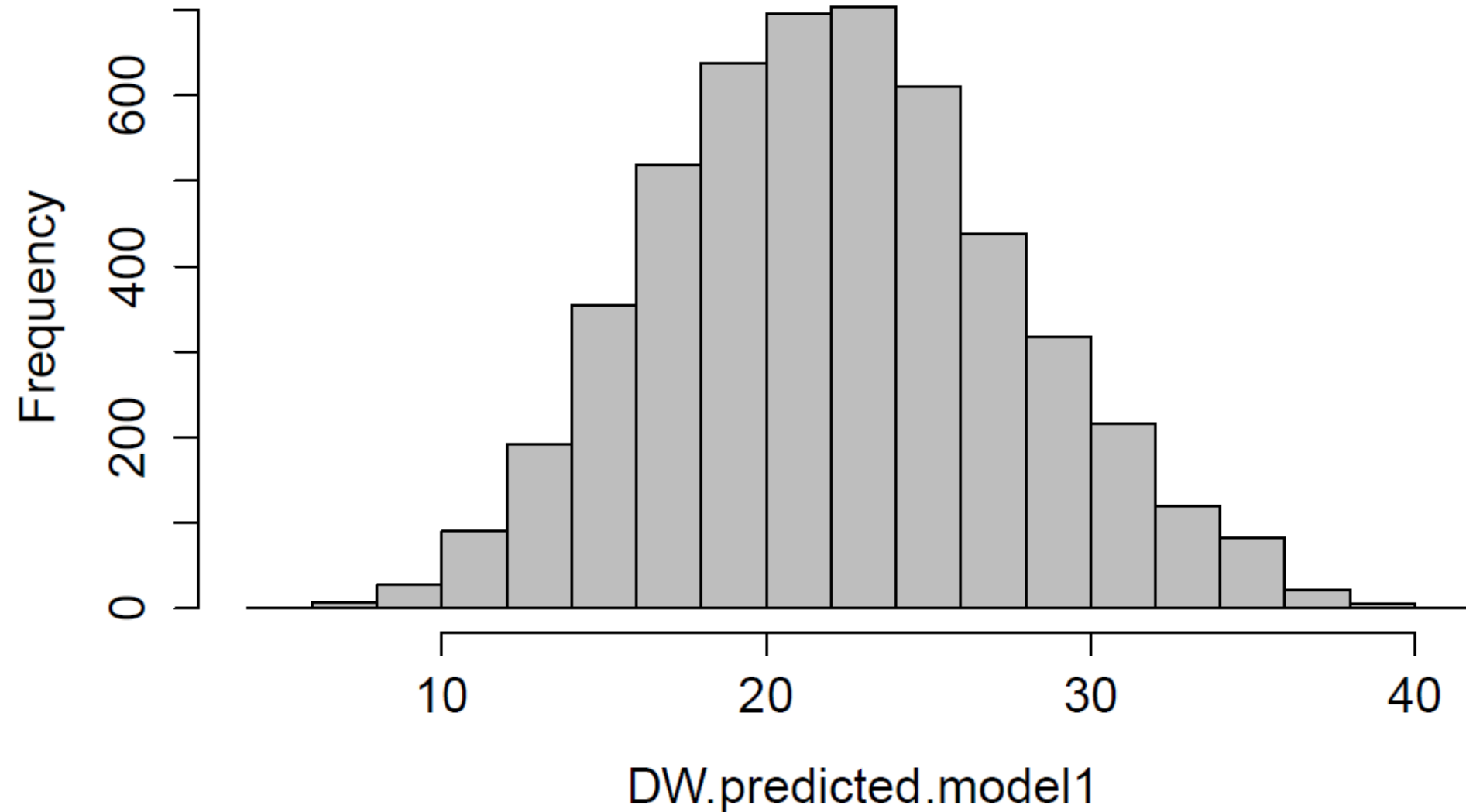
boxcox_transf <- boxcox(linmod)

lambda <- boxcox_transf$x[which.max(boxcox_transf$y)]

lambda = 0.3434343

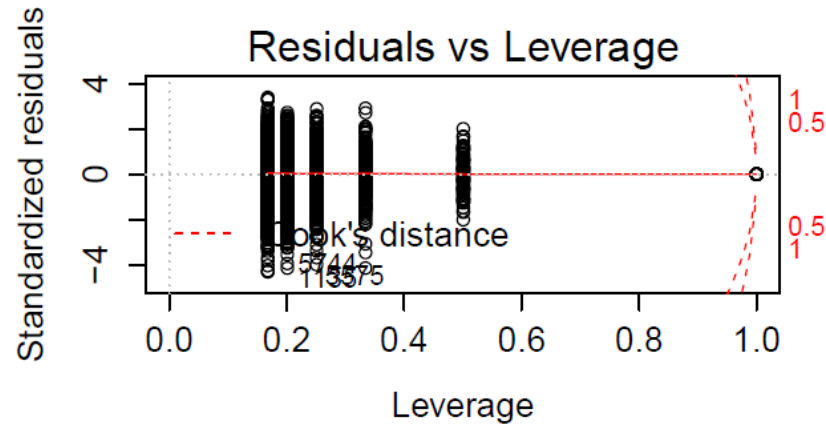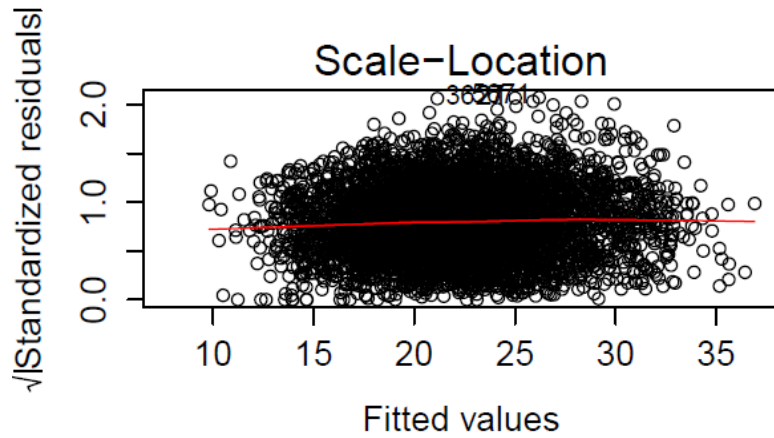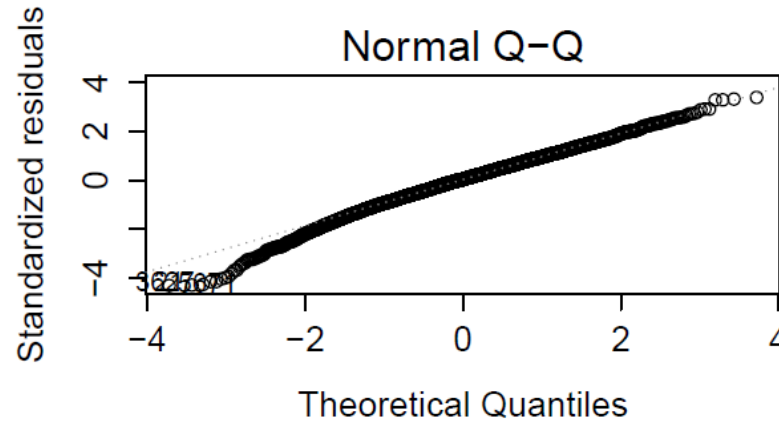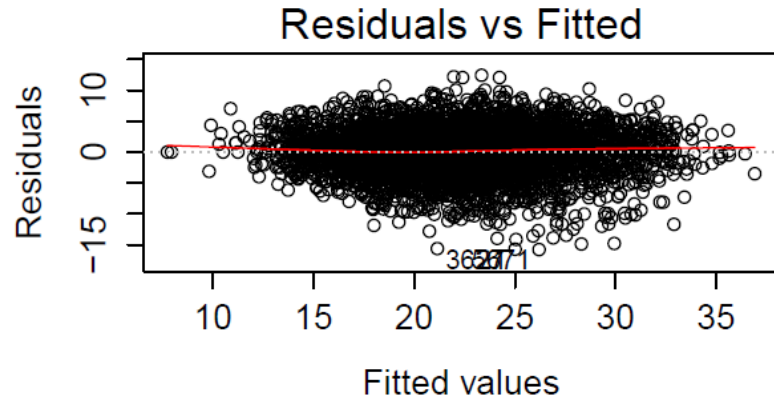# ANOVA model cont'd . . .

Boxcox transformation: Y_trsf = (Y^lambda -1)/lambda

Check if the ANOVA model is improved after data transformation:


```
>op <- par(mfrow = c(2, 2))
>plot(linmod_trsf)
>par(op)
```

Mixed model analysis without a spatial effect using breedR (classical model)

```
> mixmod_breedR <- remlf90(fixed = DW.predicted.model1.trsf ~ 1 + as.factor(Bloc),
+                          random = ~ Ident,
+                          data = data_ok,
+                          method = "ai")


> summary(mixmod_breedR)
Linear Mixed Model with pedigree and spatial effects fit by AI-REMLF90 ver. 1.110
```

Data: data_ok

AIC	BIC	logLik
29940	unknown	-14968

Variance components:

	Estimated variances	S.E.
Ident	12.24	0.6990
Residual	16.21	0.3635

Fixed effects:

|	value	s.e.
as.factor(Bloc).1	23.749	0.1721
as.factor(Bloc).2	22.148	0.1734
as.factor(Bloc).3	22.508	0.1756
as.factor(Bloc).4	19.698	0.1887
as.factor(Bloc).5	22.996	0.1778
as.factor(Bloc).6	19.277	0.1841

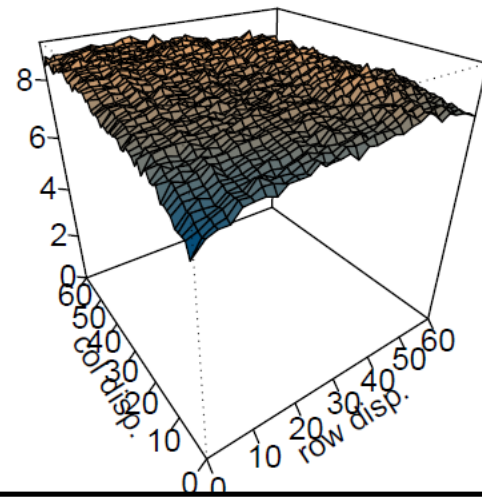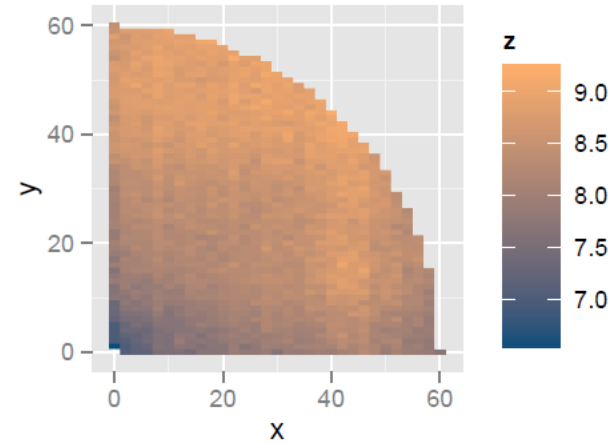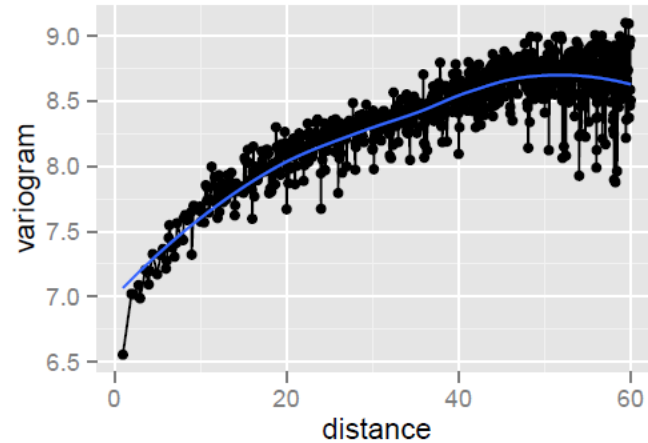# Classical model cont'd . . .

- **Spatial effect diagnosis: 2D plot of residuals from classical model**
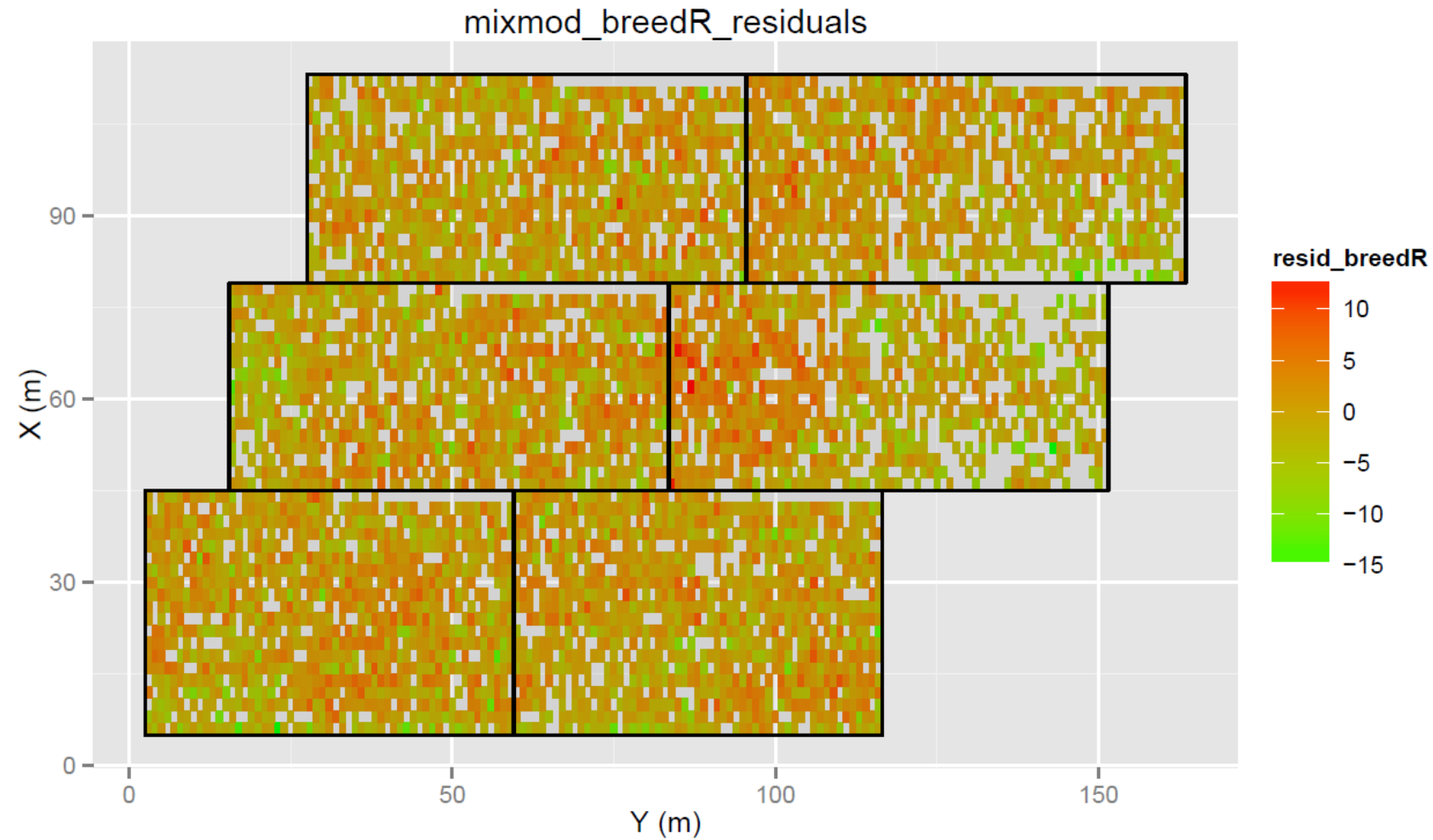
Classical model cont'd . . .

- **Spatial effect diagnosis: Variograms of residuals from classical model**

> variogram(mixmod_breedR, coord = data_ok[, colnames(data_ok) %in% c("X_ok", "Y_ok")],
R = 60)
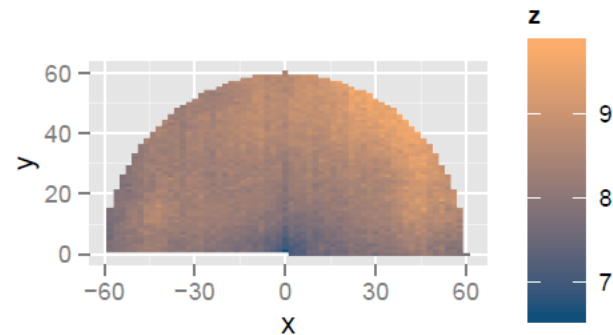
Conclusions based on classical model:

- The analysis of 2 year biomass yield indicated that the RCB design was not adequately accounting for field variation.

- In order to improve the estimation of genotype effects, a spatial analysis was used on 2 year biomass data using the breedR (Muñoz and Sanchez, 2014) statistical package.

Mixed model with a spatial effect using breedR: autoregressive with Block effects (Selected spatial model)

```
#grid5:
rho.grid <- expand.grid(rho_r = seq(0.83, 0.88, length = 4),
             rho_c = seq(0.95, 0.99, length = 4))


mixmod_breedR_AR1_bloc_grid5 <- remlf90(fixed = DW.predicted.model1.trsf ~ 1 + as.factor(Bloc),
                  random = ~ Ident,
                  spatial = list(model = "AR",
                           coordinates = data_ok[, c("X_ok", "Y_ok")],
                           rho = rho.grid),
                 data = data_ok,
                 method = "ai")


 save(mixmod_breedR_AR1_bloc_grid5, file =
"mixmod_breedR_AR1_bloc_grid5_DW.predicted_model1_trsf.Rdata")
```
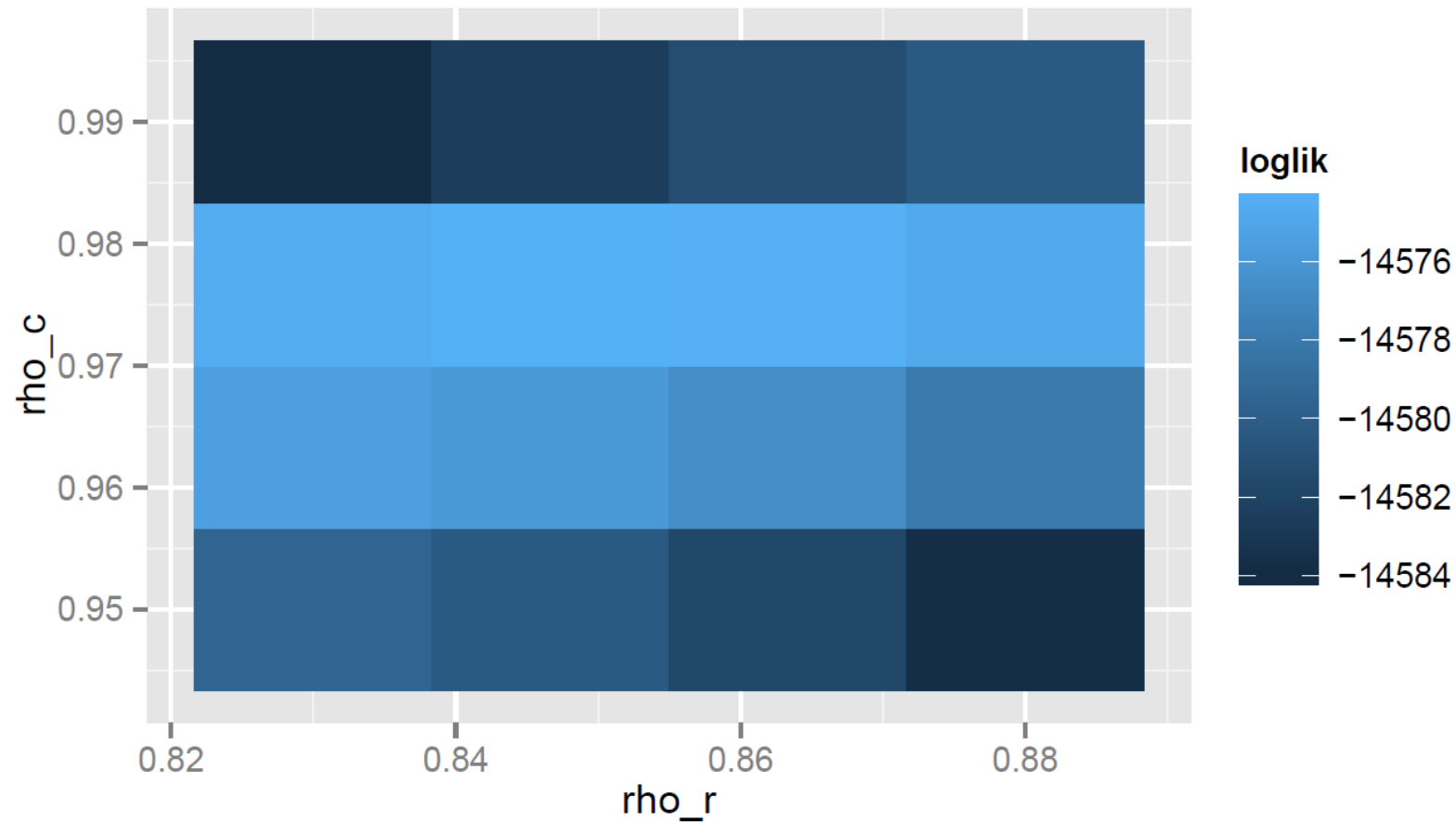
Selected spatial model cont'd . . .

> qplot(rho_r, rho_c, fill = loglik, geom = "tile", data = mixmod_breedR_AR1_bloc_grid5$rho)

Autoregressive parameters **for** rows and columns: (0.846, 0.976)

Selected spatial model cont'd . . .

```
> selmod <- remlf90(fixed = DW.predicted.model1.trsf ~ 1 + as.factor(Bloc),
+                           random = ~ Ident,
+                           spatial = list(model = "AR",
+                                   coordinates = data_ok[, c("X_ok", "Y_ok")],
+                                   rho = c(rho_r= 0.846, rho_c = 0.976)),
+                    data = data_ok,
+                    method = "ai")

> summary(selmod)
```

Variance components:

|	Estimated variances	S.E.
Ident	12.836	0.6921
spatial	6.125	0.7843
Residual	11.908	0.2912

Fixed effects:

	value	s.e.
as.factor(Bloc).1	22.497	0.9883
as.factor(Bloc).2	22.236	0.9777
as.factor(Bloc).3	20.612	0.9237
as.factor(Bloc).4	20.459	0.9396
as.factor(Bloc).5	20.273	0.9890
as.factor(Bloc).6	19.490	1.0039

Selected spatial model cont'd . . .

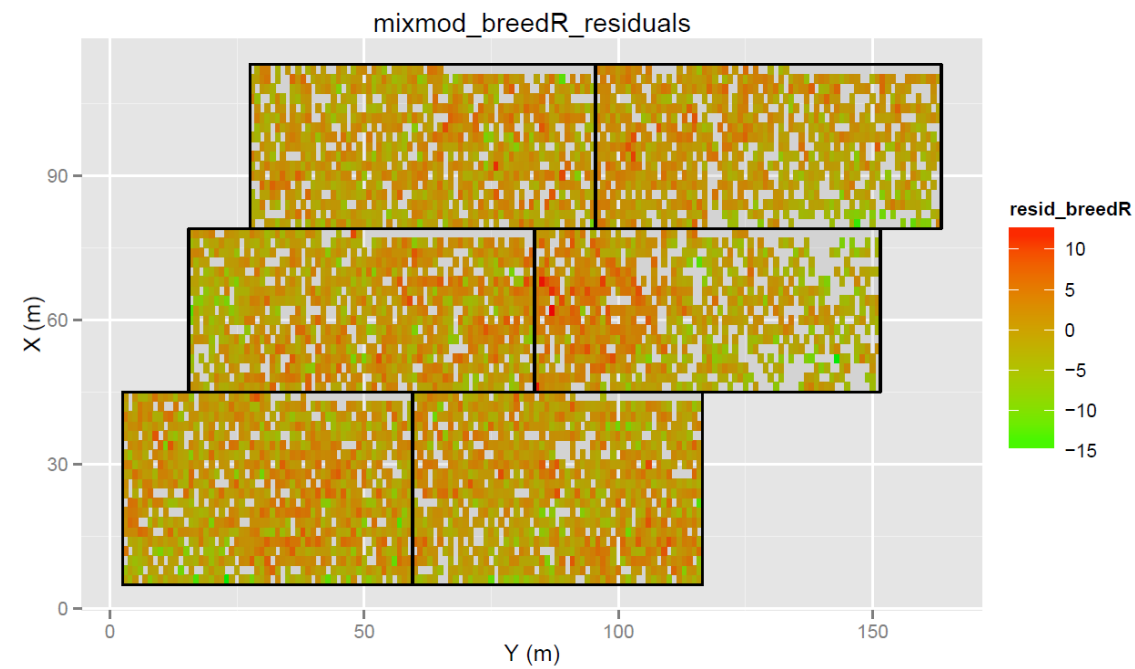## Selected spatial model: 2D map of spatial effects
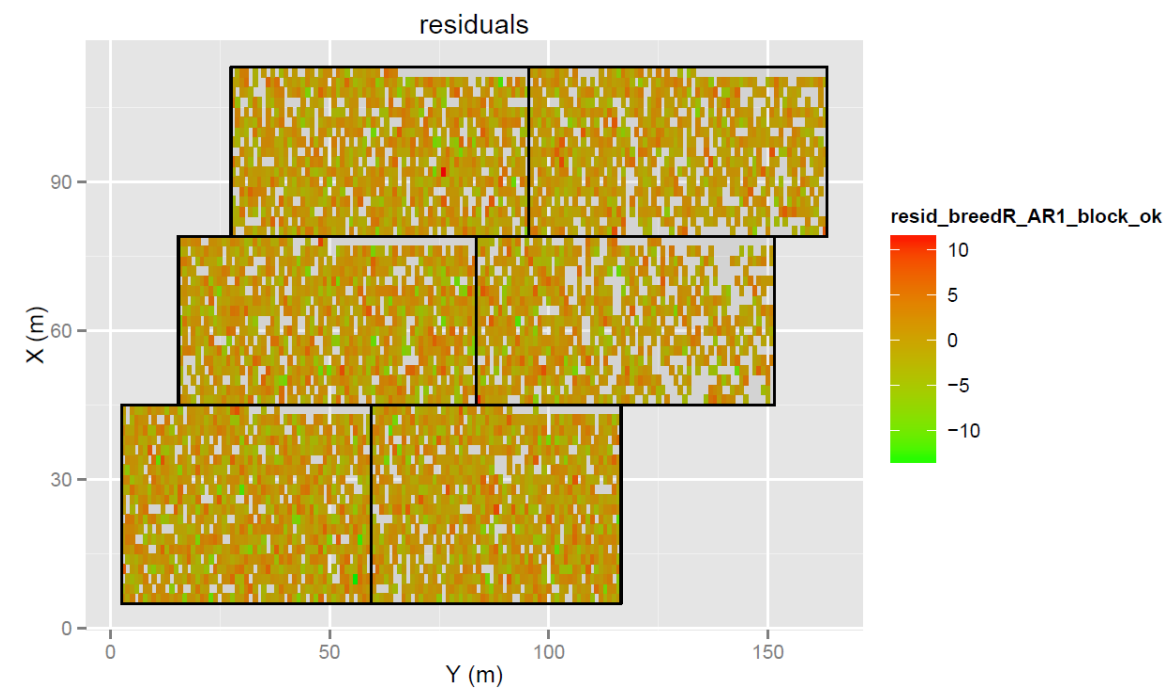
Aerial view of the field trial
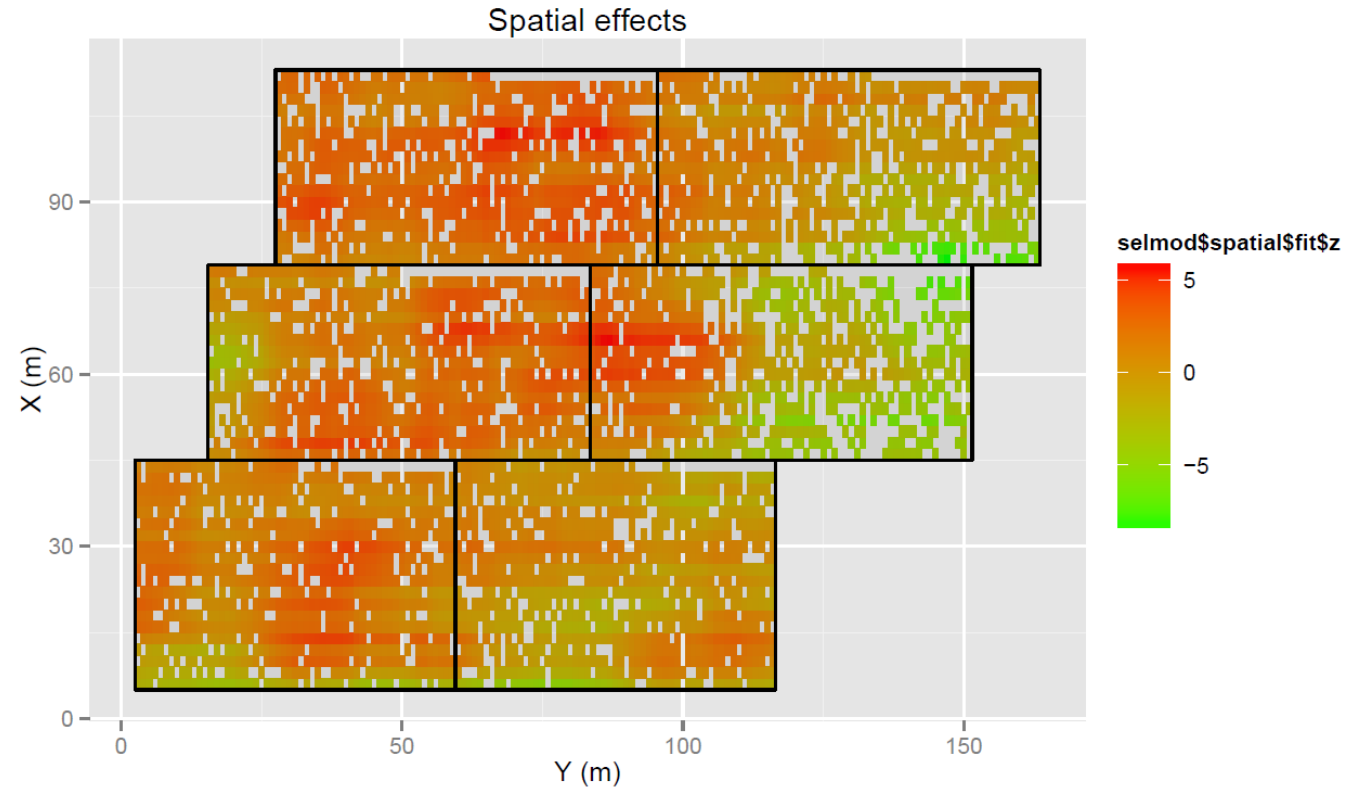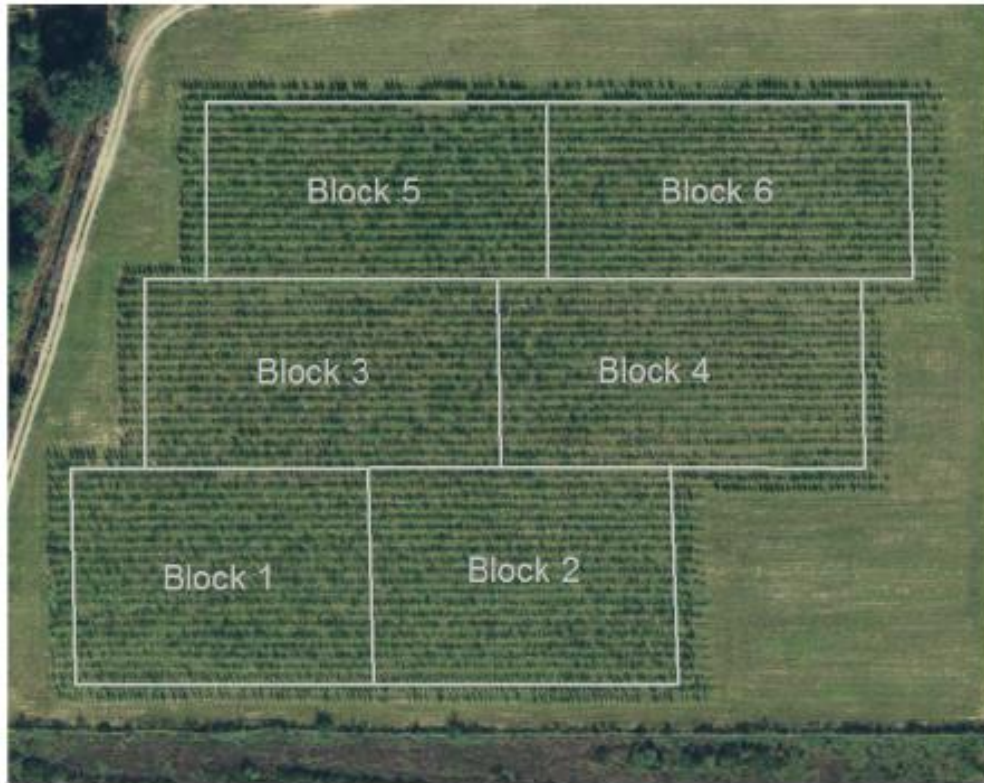
Selected spatial model cont'd . . .
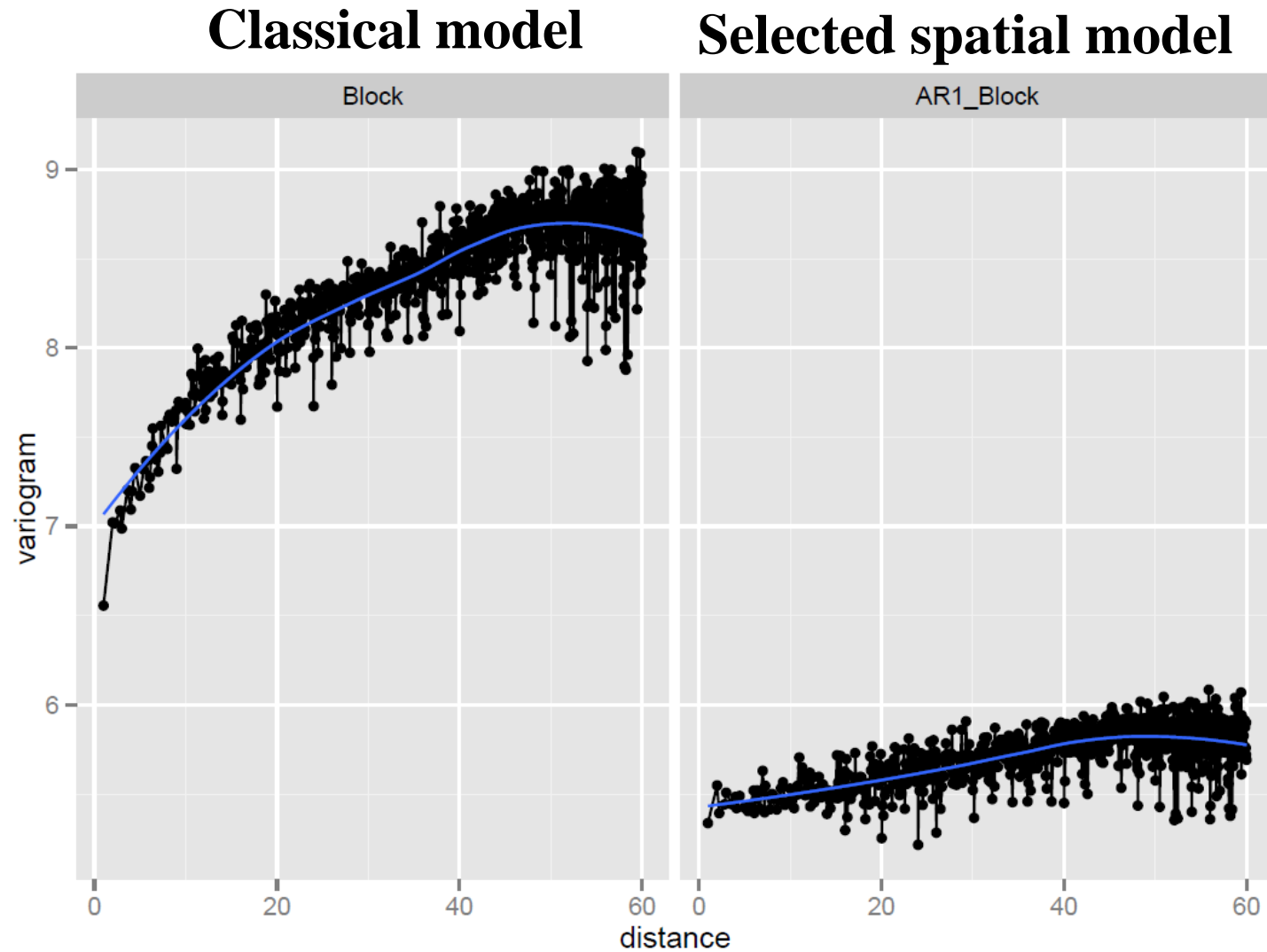
2D map of residuals:
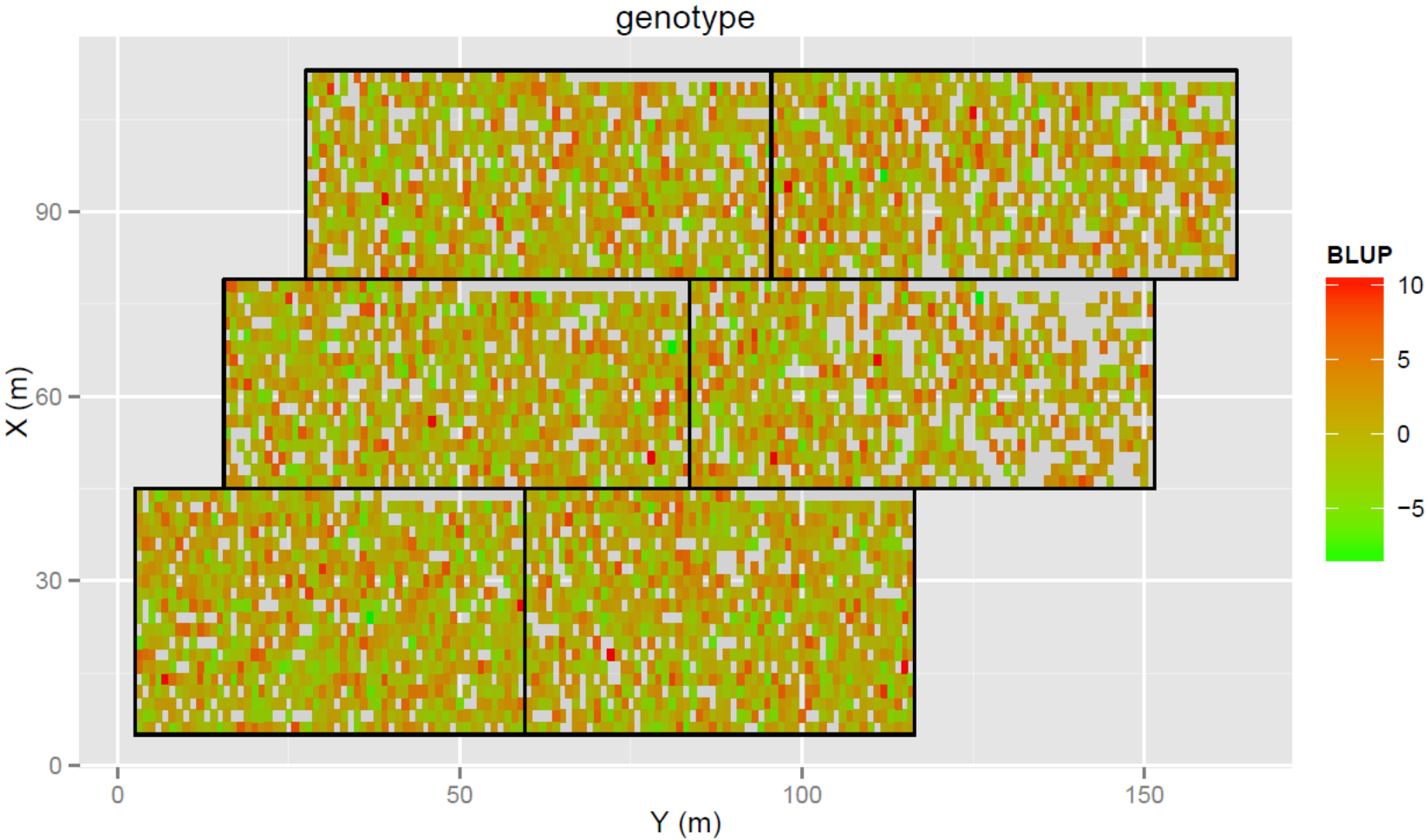
Classical model

Selected spatial model

Selected spatial model cont'd . . .

Variograms of residuals:

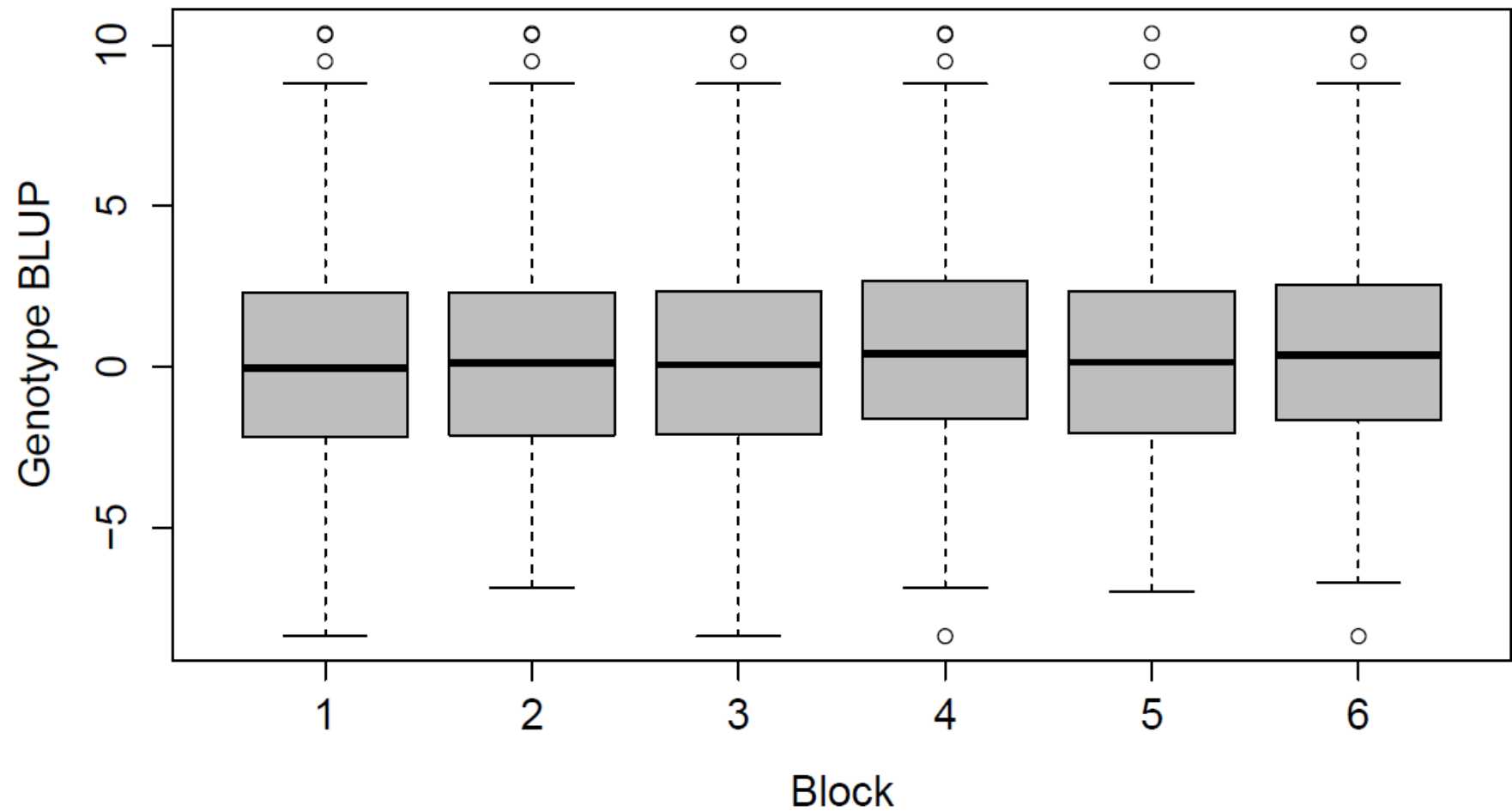**Classical model**   **Selected spatial model**
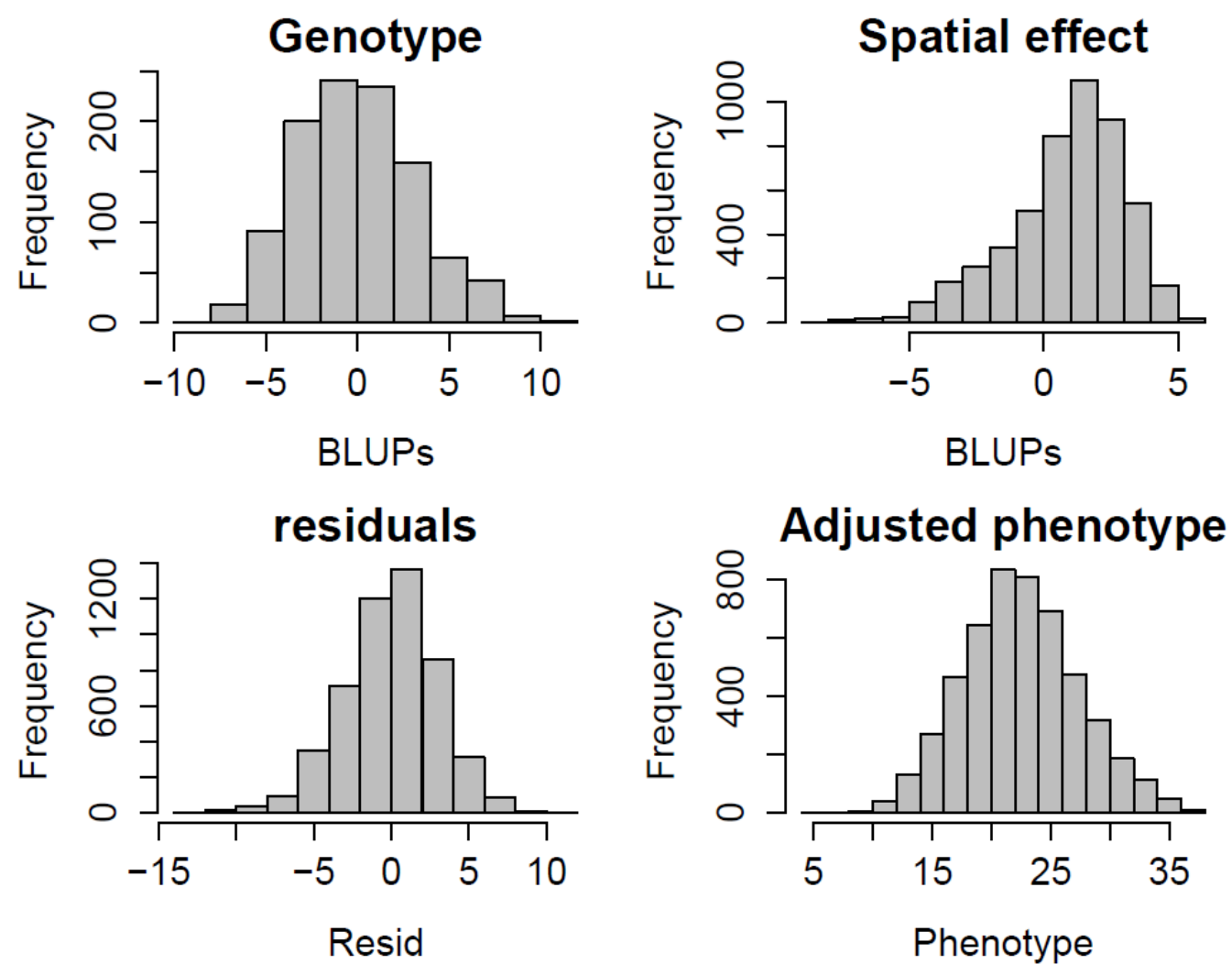
Selected spatial model: 2D map of genotype BLUPs

Selected spatial model cont'd . . .

Selected spatial model: Boxplot of genotype BLUPs per Block

Selected spatial model cont'd . . .

Selected spatial model: histograms of random effect BLUPs

Extraction of AI matrix, estimation of H² together with its standard errors

#Extraction of AI matrix from breedR output:
```r
aimat <- which(selmod$reml$output == " inverse of AI matrix (Sampling Variance)")
varcov_mat_breedR <- matrix(na.omit(as.numeric(unlist(apply(data.frame(selmod$reml$output[
  (aimat + 1):(aimat + 3)]), 1, function(x){strsplit(x, " ")})))), 3, 3)
colnames(varcov_mat_breedR) <- c("Ident", "spatial", "Residual")
rownames(varcov_mat_breedR) <- c("Ident", "spatial", "Residual")
```

#Estimation of heritability (H²) together with its standard errors:
```r
 library(msm)
H2 <- selmod$var["Ident", "Estimated variances"] /
  (selmod$var["Ident", "Estimated variances"] + selmod$var["Residual", "Estimated variances"])
se_H2 <- deltamethod(~ x1 / (x1 + x2),
            + c(selmod$var["Ident", "Estimated variances"],
              + selmod$var["Residual", "Estimated variances"]),
            + varcov_mat_breedR[c("Ident", "Residual"), c("Ident", "Residual")])
round(c(H2, 1.96*se_H2), 2)
```

Extraction of AI matrix, estimation of H² together with its standard errors

❑ **an accurate estimate of broad-sense heritability (H²)**

Classical model: $H^2 = \sigma^2_g / (\sigma^2_g + \sigma^2_\varepsilon)$
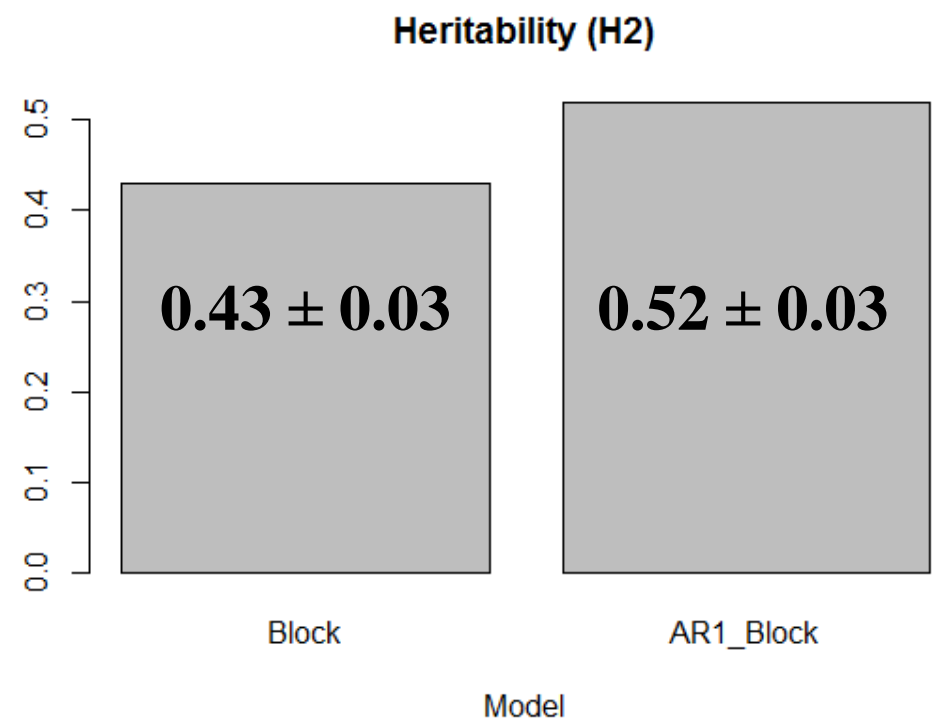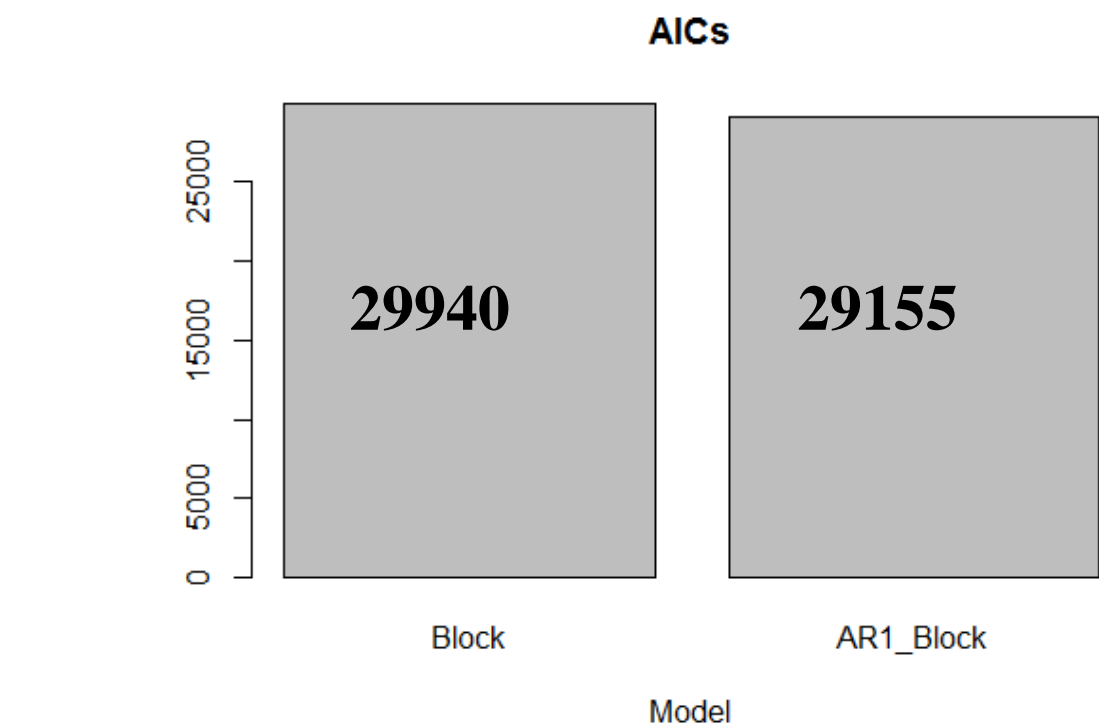
Selected spatial model:

$$H^2 = \sigma^2_g / (\sigma^2_g + \boxed{\sigma^2_s} + \sigma^2_e)$$

X

$$H^2 = \sigma^2_g / (\sigma^2_g + \sigma^2_e)$$

Extraction of AI matrix, estimation of H^2 together with its standard errors

Graphical representation of AICs and broad-sense heritability estimates from classical & selected models:

Conclusions:

❑ The analysis of biomass yield indicated that the RCB design was not adequately accounting for field variation resulting in high error term and low heritability.

❑ Linear mixed model with AR1 yielded best results for all traits.

❑ Including a fixed effect of Block in the spatial model improves the model fit.

❑ Spatial analysis always improves H^2 estimates.

❑ Data transformation do not seem to signifcantly affect H^2 estimates nor spatial effect parameters.

Acknowledgements

Dr. Vincent Segura

Dr. Catherine Bastien

Dr. Leopoldo Sanchez

Dr. Facundo Muñoz