Module 4:End-to-End Analytical Pipeline in R

Fundamentals of Genomic Prediction and Data-Drive Crop Breeding

(November 24-28, 2025)

Waseem Hussain
Senior Scientist-I
International Rice Research Institute
Rice Breeding Innovations Platfrom
waseem.hussain@cgiar.org
whussain2.github.io

Mahender Anumalla
Scientist-I
International Rice Research Institute
South-Asian Hub, Hyderabad
m.anumalla@cgiar.org

Margaret Catolos
Associate Scientist
International Rice Research Institute
South-Asian Hub, Hyderabad
m.catolos@cgiar.org

November 20, 2025

Contents

Load the Libraries

```
> # Load the Required Libraries
>   rm(list=ls()) # Remove previous work
>   library(rrBLUP)
>   library(BGLR)
>   library(AGHmatrix)
>   library(ggplot2)
>   library(DT)
>   #library(cvTools)
>   library(dplyr)
>   library(lme4)
>   library(arm)
>   library(statgenSTA)
```

This section shows the analysis of filtered phenotypic data in lme4 and other open source R packages. The filtered data set was obtained after pre-processing and Quality check of data

Phenotypic Data Analysis in lme4 R Package

- Here in this section phenotypic data analysis is performed in an open source R package called **lme4**. More on this R package can be found here lme4 Tutorial 1, and lme4 Tutorial 2.
- The purpose of this section is to repeat the phenotypic data analysis in lme4 as ASReml R package is commercial package and may not available for all the users.
- Filtered data set will be used, same one used in ASReml R package to perform the analysis in lme4.
- ANOVA, variance components, BLUPS, BLUES and heritability is extracted for the results part.

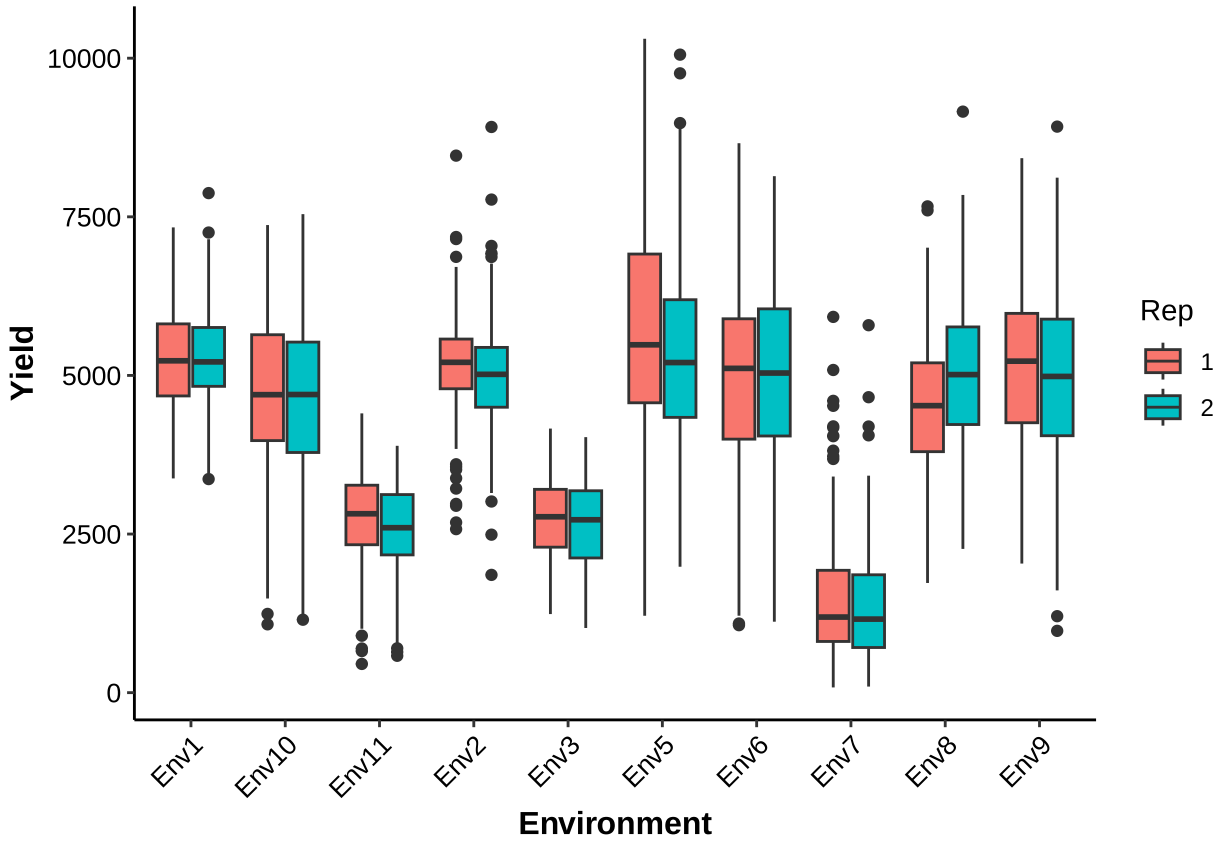
Upload the Filtered Phenotypic Data

```
> demo.data.filtered<-read.csv(file="./Data/demo.data.filtered.csv",
+                              header = TRUE)
> # factor conversion if below are not in factors
> columns<-c("Environment", "Genotype", "Rep", "Block", "Row", "Column", "Line.type")
> demo.data.filtered[, columns]<-lapply(columns, function(x) as.factor(demo.data.filtered[[x]]))
> demo.data.filtered$Yield<-as.numeric(demo.data.filtered$Yield)
> demo.data.filtered$HT<-as.numeric(demo.data.filtered$HT)
> demo.data.filtered$DTF<-as.numeric(demo.data.filtered$DTF)
>
> # Subset the required columns
> demo.data.filtered<-demo.data.filtered[, c("Environment", "Genotype", "Rep",
+                                           "Block", "Row", "Column", "Line.type",
+                                           "Yield", "HT", "DTF")]
> # First we will arrange the rows and columns for spatial analysis.
> # Now we will subset the environments and Yields for analysis
> demo.data.filtered<-data.frame(demo.data.filtered%>% group_by(Environment)%>%arrange(Row, Column)) # 
> demo.data.filtered<-data.frame(demo.data.filtered%>% arrange(Environment)) # Arrange by environment
>
> #demo.data.filtered<-demo.data.filtered[!demo.data.filtered$Environment %in% c("Env2", "Env5","Env8",
> # View as table in file
> head(demo.data.filtered)
```

Environment	Genotype	Rep	Block	Row	Column	Line.type	Yield	HT	DTF
Env1	44	1	1	1	1	Entry	4956.395	115.6	96
Env1	131	1	1	1	2	Entry	5059.207	116.0	89
Env1	17	1	1	1	3	Entry	4948.038	99.0	101
Env1	146	1	1	1	4	Entry	6012.658	102.8	92
Env1	123	1	1	1	5	Entry	4456.759	112.2	94
Env1	116	1	1	1	6	Entry	4473.946	108.0	98

Quick Visualization of Data

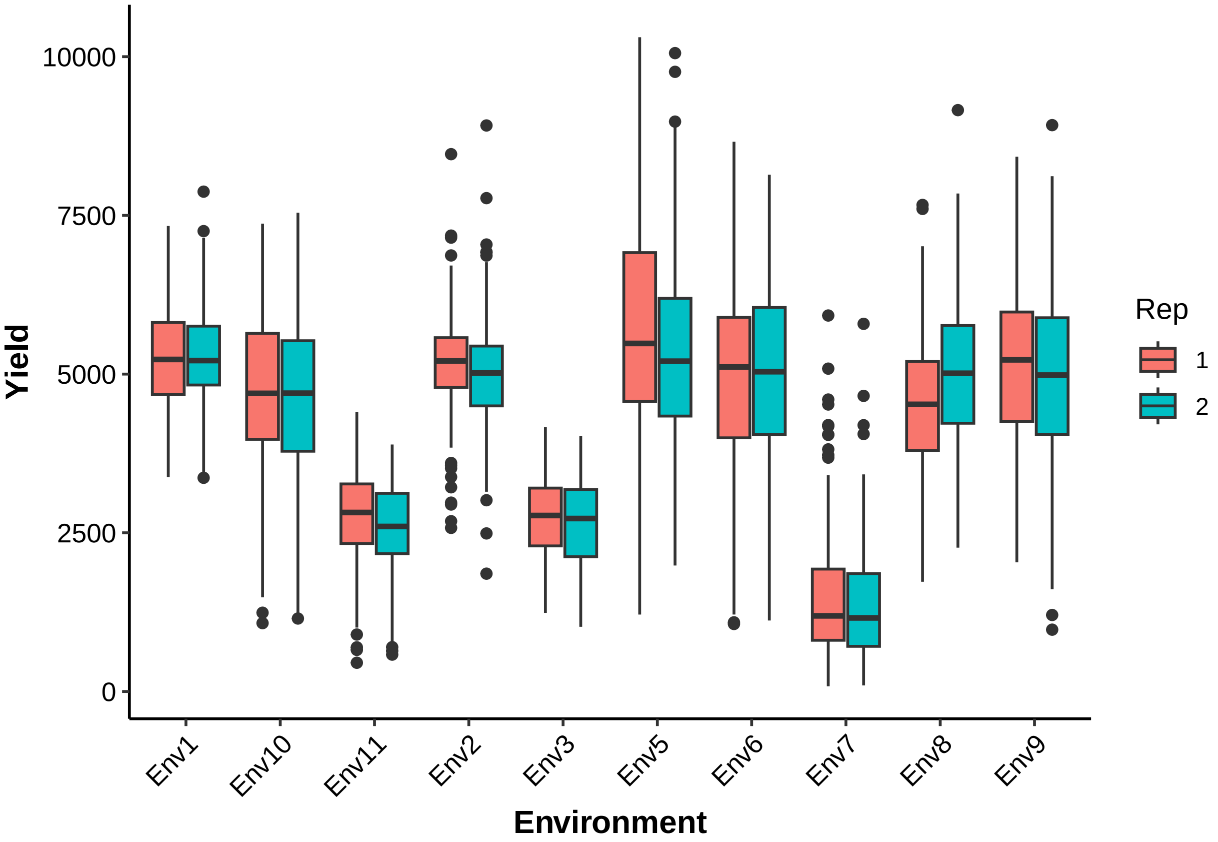
```
> ggplot(data = demo.data.filtered, aes(x = Environment, y = Yield, fill = Rep))+
+   geom_boxplot()+
+   theme_classic()+
+   theme(axis.text.x = element_text(angle = 45, hjust = 1)) +# fill by timepoint to give different col
+   #scale_fill_manual(values = c("", ""))+
+   #scale_color_manual(values = c("", ""))
+   theme (plot.title = element_text(color="black", size=12,hjust=0.5, face = "bold"), # add and modify
+          axis.title.x = element_text(color="black", size=12, face = "bold"), # add and modify title t
+          axis.title.y = element_text(color="black", size=12, face="bold")) + # add and modify title t
+   #scale_y_continuous(limits=c(0,15000), breaks=seq(0,15000,1000), expand = c(0, 0))+
+   theme(axis.text= element_text(color = "black", size = 10)) # modify the axis text
```



##

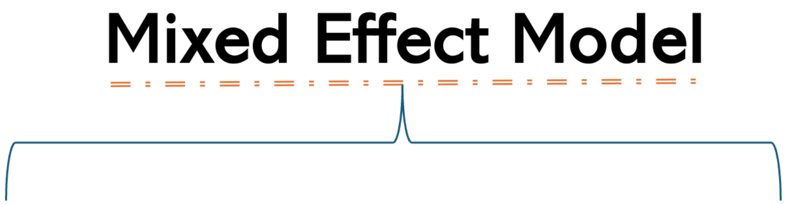
Quick Visualization of Data

```
>   ggplot(data = demo.data.filtered, aes(x = Environment, y = Yield, fill = Rep))+
+   geom_boxplot()+
+   theme_classic()+
+   theme(axis.text.x = element_text(angle = 45, hjust = 1)) +# fill by timepoint to give different col
+   #scale_fill_manual(values = c("", ""))+
+   #scale_color_manual(values = c("", ""))
+   theme (plot.title = element_text(color="black", size=12,hjust=0.5, face = "bold"), # add and modify
+   axis.title.x = element_text(color="black", size=12, face = "bold"), # add and modify title to x axi
+   axis.title.y = element_text(color="black", size=12, face="bold")) + # add and modify title to y axi
+ #scale_y_continuous(limits=c(0,15000), breaks=seq(0,15000,1000), expand = c(0, 0))+
+  theme(axis.text= element_text(color = "black", size = 10)) # modify the axis text
```



Single Stage/Step Wise Analysis

- In this section, data analysis will be shown only for grain yield trait using a **Linear Mixed-Model Approach** in lme4 R Package package, and will be useful to the users who do not have access to the commercial **ASReml-R package**.
- In general analysis pipeline is divided in two parts:
  1. **Separate analysis/step-wise**: In this each environment/trial is analyzed separately.
  2. **Combined analysis or Multi-environment trial (MET) analysis**: In this analysis all the environments will be analyzed jointly.
  - Various mixed models from basic to advanced models will be used will for MET analysis.
  - First let us subset the data for on environment to show how to perform the analysis for one trial or environment in lme4 R package
  - We will run models which are feasible in lme4 R package. Note spatial models are not possible to run in lme4 R package.
  - We will use basic models and show how to extract the results



Mixed Effect Model

Single Stage Analysis

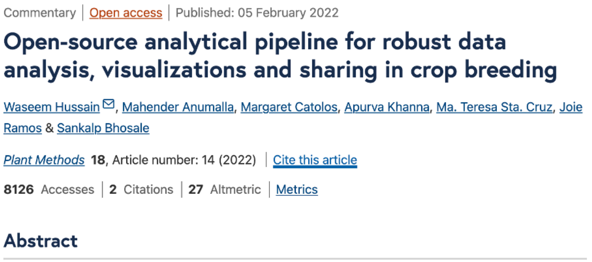
Each trial or environment
is analyzed separately

Two Stage Analysis

All trials combined and
analyzed together

Stage-wise analysis is more appropriate.

- ➢ Trials with unbalanced data sets,
- ➢ Different experimental design factors acros trials,
- ➢ Avoid the computational challenges of analyzing a huge number of trials.



Commentary | Open access | Published: 05 February 2022

Open-source analytical pipeline for robust data analysis, visualizations and sharing in crop breeding

Waseem Hussain, Mahender Anumalla, Margaret Catolos, Apurva Khanna, Ma. Teresa Sta. Cruz, Joie Ramos & Sankalp Bhosale

Plant Methods **18**, Article number: 14 (2022) | Cite this article

8126 Accesses | **2** Citations | **27** Altmetric | Metrics

Abstract

Subset the Data for One Environment

- Subset the data for one environment first.

```
> # Subset the environment 1
>   sub.data<-subset(demo.data.filtered, Environment=="Env1")
>   sub.data<-droplevels.data.frame(sub.data)
```

Run the Mixed model

Model 1.lme4

- The model described below is equivalent to *model 1* described in ASReml R package analysis.

$y_{ijk} = \mu + g_i + r_j + b_{jk} + \epsilon_{ijk}$ Where $y_{ijk} =$ is the effect of ith genotype in jth replication and kth block within the jth replication, μ = overall mean, g_i = random effect of the ith genotype, r_j = fixed effect of the jth replication, $b_{jk}= random effect of kth block nested within j replication,□*{ijk}=residual error,here we assume errors are independent and identically dis*

```
> # Now apply model
>   model1<-lmer(Yield~Rep+(1|Genotype)+ (1|Rep:Block), data =sub.data)
```

Results

- Here we will summarize the results using ***summary()*** function. The first few lines of output indicate that the model was fitted by REML as well as the value of the REML criterion. The second piece of the summary output provides information regarding the random-effects and residual variation. The third piece of the summary output provides information regarding the fixed-effects and the fourth piece of summary output provides information regarding the correlation of fixed effects.

```
> # Summarise the results
>   summary(model1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Yield ~ Rep + (1 | Genotype) + (1 | Rep:Block)
   Data: sub.data

REML criterion at convergence: 6239.3

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.90048 -0.59387  0.03899  0.60311  1.71001

Random effects:
 Groups    Name        Variance Std.Dev.
 Genotype  (Intercept) 431861   657.2
 Rep:Block (Intercept)  28499   168.8
 Residual              193255   439.6
Number of obs: 394, groups:  Genotype, 197; Rep:Block, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)  5233.19      94.20  55.552
Rep2           60.38     115.60   0.522

Correlation of Fixed Effects:
     (Intr)
Rep2 -0.614
```

Extract variance components

- Here we will extract variance components

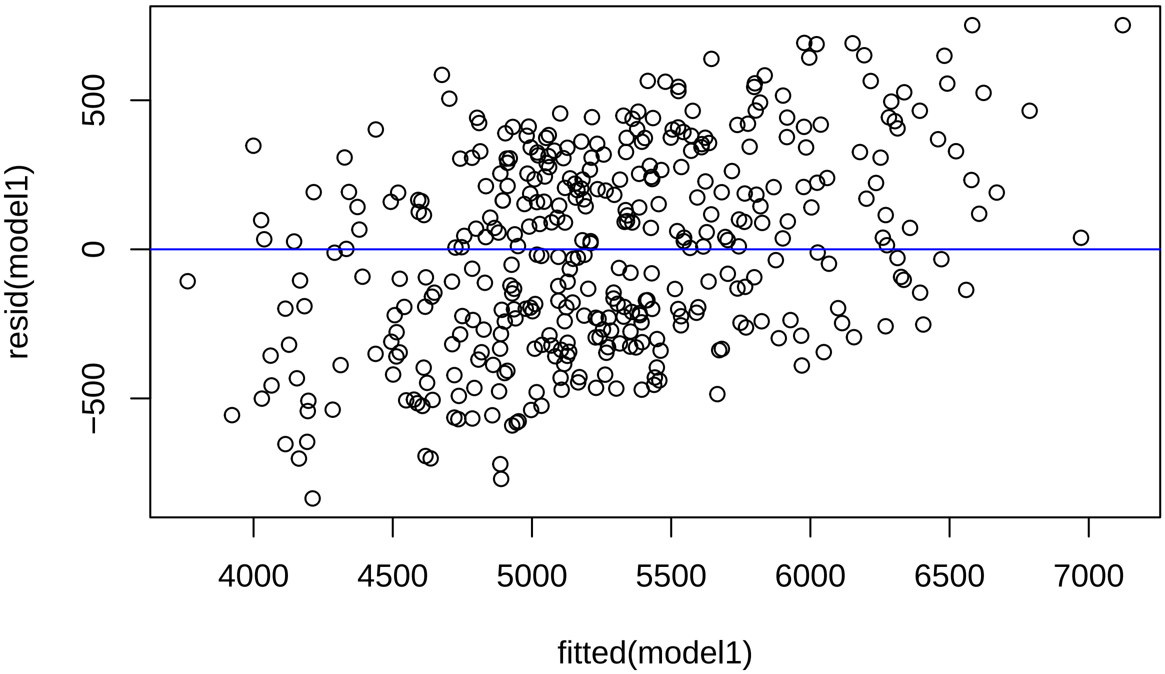
```
>   Ve<- VarCorr(model1)
>   Ve
```

```
 Groups    Name        Std.Dev.
 Genotype  (Intercept) 657.16
 Rep:Block (Intercept) 168.82
 Residual              439.61
```

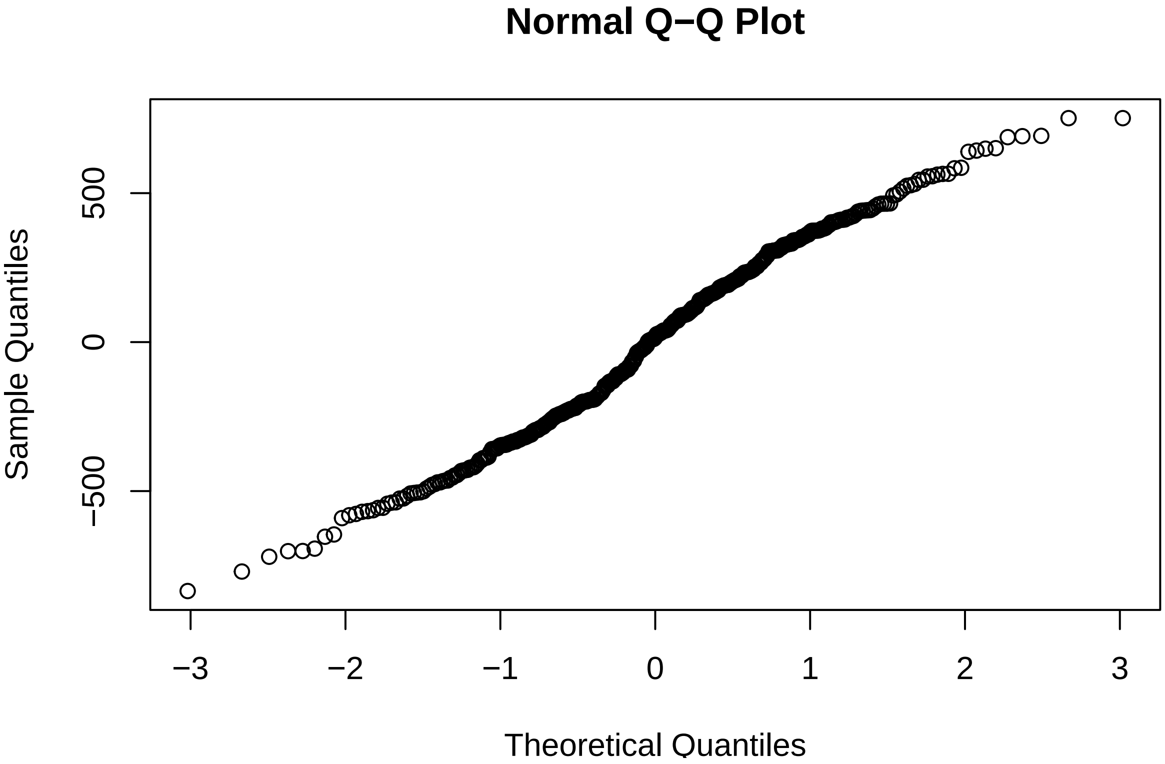
Plot the residual vs fitted plot

- Here will show how to check for check for homoscedasticicty

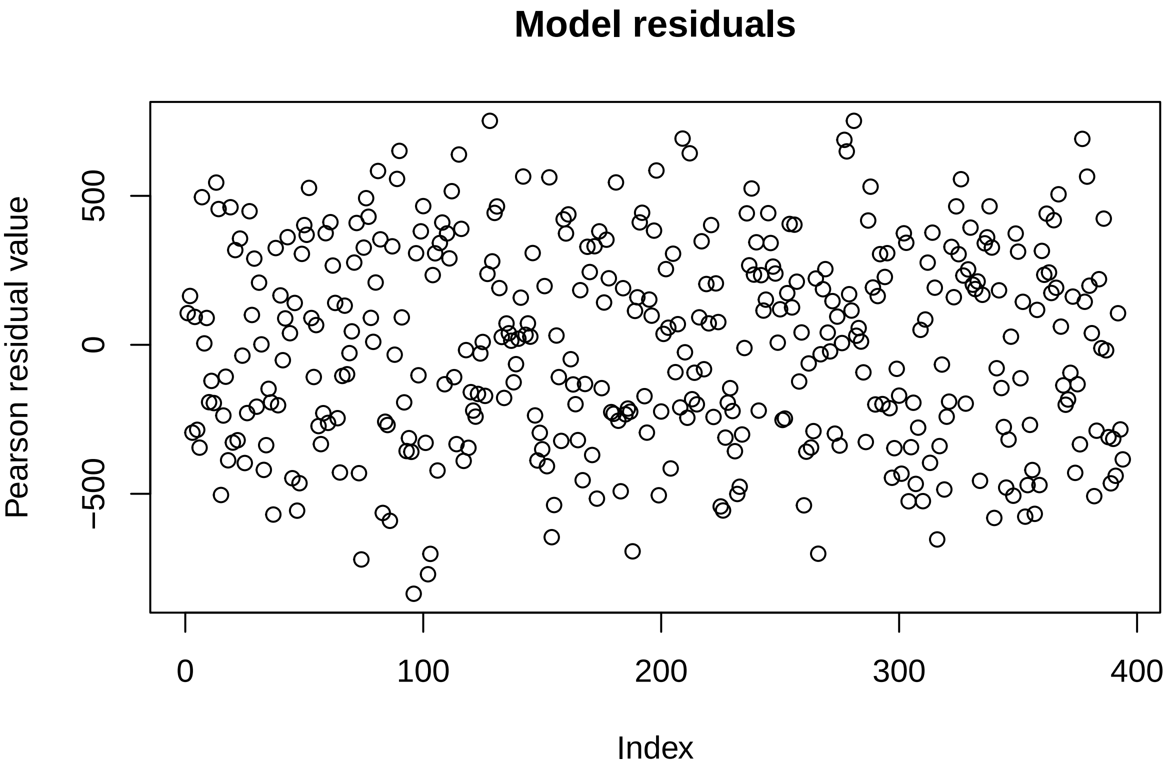
```
> # Plot the residual plot
>   plot(fitted(model1), resid(model1), type="pearson")
>   abline(0,0, col="blue")
```



```
> # Plot QQ plot
>   qqnorm(resid(model1))
```



```
> # Residual plot
>   plot(residuals(model1,type="pearson"), main='Model residuals',
+   ylab='Pearson residual value')
```



ANOVA for fixed effects

```
> # ANOVA
>   anova(model1)
```

	npar	Sum Sq	Mean Sq	F value
Rep	1	52724.26	52724.26	0.2728223

Extract the Fixed effects

- Here will show how to extract the BLUEs.

```
>   BLUEs<-fixef(model1)
>   BLUEs
```

```
(Intercept)        Rep2
  5233.1856     60.3794
```

Extract the Random effects

- Here will show how to extract the BLUPs.

```
> # Extract the Random effects
>   BLUPs<-data.frame(Blups.yield=ranef(model1)$Genotype)
>   GV<-data.frame(BLUps.GY=coef(model1)$Genotype[,1]) #Genotype values (Blups +Intercept)
```

Heritability

- Here will show how to calculate the heritability. Two approaches will be show how to estimate heritability: 1) Based on Variance components and 2) Based on Cullis et al. 2006 is also ....$1 - \frac{\overline{V}_{BLUp}}{2\sigma^2 g}$. Where $\overline{V}_{BLUP}$ is mean variance difference of two genotypes based on BLUPs and $\sigma^2 g$ is variance of genotypes.

```
> # Extract the variance components
>   Ve<- data.frame (VarCorr(model1))
>   Ve
```

grp	var1	var2	vcov	sdcor
Genotype	(Intercept)	NA	431860.94	657.1613
Rep:Block	(Intercept)	NA	28498.55	168.8151
Residual	NA	NA	193254.94	439.6077

```
> # Now calculate heritability using variance components
>   genotype.var=Ve[1,4]
>   error.var=Ve[2,4]
> # Now heritability
>   h2=genotype.var/(genotype.var+error.var)*100
>   h2
```

```
[1] 93.8095
```

```
> # Reliability
>   std.err<-se.ranef(model1)$Genotype
>   v_BLUP<- mean(std.err)
> # Heritability/Reliability
>   h2<- (1-((v_BLUP)^2/(Ve[1,4]*2)))*100
>   h2
```

```
[1] 90.55036
```

Run the Analysis for all Environments

```
> # Run the analysis and check reliability
>   # For Non-Stress Data using DTF as co-variate
>   demo.data.filtered$Environment<- as.character(demo.data.filtered$Environment)
>   un.exp<- unique(demo.data.filtered$Environment)
>   for(i in 1:length(un.exp)){
+     sub<- droplevels.data.frame(demo.data.filtered[which(demo.data.filtered$Environment==un.exp[i]),]
+
+       model<-lmer(Yield~Rep+(1|Genotype)+ (1|Rep:Block), data =sub)
+       #BLUPs<-data.frame(Blups.yield=ranef(model)$Genotype, Environment=un.exp[i])
+         BLUPs<-data.frame(BLUps.GY=coef(model)$Genotype[,1], Environment=un.exp[i])
+     if(i>1){
+       BLUPs.all<-rbind(BLUPs.all, BLUPs)
+     }
+     else{
+       BLUPs.all<- BLUPs
+     }
+   }
> # Save the BLUES out put file for Genomic Predictions
>  #estimates.all$Genotype<-gsub("^.{8}", "",  estimates.all$Genotype)
```

MET Analysis

Model 2.lme4

- Here we will analyze all the environments jointly and extract the single BLUE for each genotype. We will use mixed model analysis in lme4 r package model. We will treat genotypes as fixed and environment as random effect.

Combined ANOVA

- Here ANOVA will be generated for all the factor levels.
- Replications are nested with environments and Blocks are within Replications which are nested within environment.

```
> # Linear model to get ANOVA
>   demo.data.filtered$Environment<-as.factor(demo.data.filtered$Environment)
>   model.anova<-lm(formula = Yield~Genotype+Environment+Genotype*Environment+Environment:Rep+ Environm
+     data=demo.data.filtered)
> # Get ANOVA
>   anova(model.anova)
```

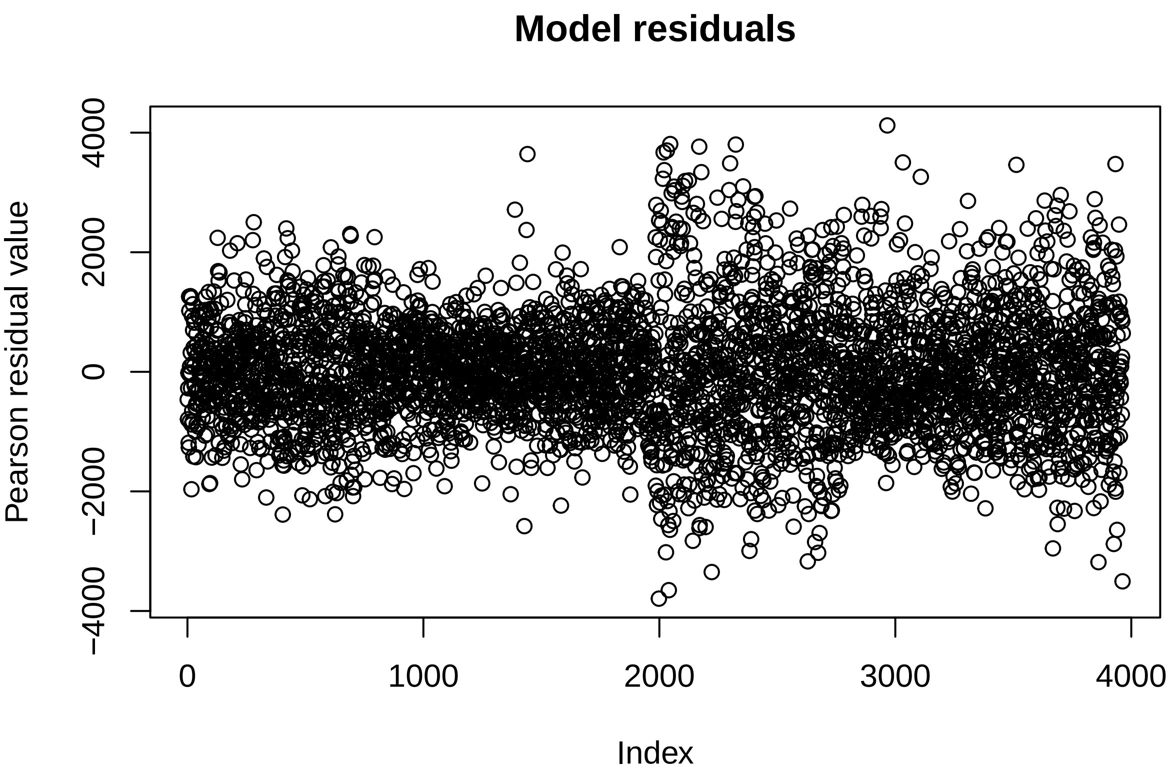
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Genotype	199	1185402954	5956798.8	14.585107	0.00e+00
Environment	9	7105461322	789495702.4	1933.065053	0.00e+00
Genotype:Environment	1779	3073244084	1727512.1	4.229780	0.00e+00
Environment:Rep	10	57746059	5774605.9	14.139012	0.00e+00
Environment:Rep:Block	76	57387089	755093.3	1.848831	1.72e-05
Residuals	1890	771907223	408416.5	NA	NA

Significant differences are observed for all factors and genotype by environment interactions are significant

Check for Homogeneity of Variance

- Some test can be used to check variance between pair of environments as given below:
- More on this can be found on this: Source 1, Source 2
- Here we will check the distribution of residuals and see how they vary as we have more than two environments. For that we will run the mixed model in lme4 and then plot the residuals

```
> #
> model2<- lmer(Yield~Rep+(1|Genotype)+(1|Environment)+
+           (1|Environment:Rep)+(1|Environment:Rep:Block),
+           data=demo.data.filtered)
>
> #plot residuals
> plot(residuals(model2,type="pearson"), main='Model residuals',
+ ylab='Pearson residual value')
```



```
> #var.test(Yield~Environment,data=demo.data.filtered)
```

From the plot it is clear that residuals are not same and highly different

Combined Analysis in lme4

- The model we will use is give below:

$y_{ijkl} = \mu + g_i + e_j + (ge)_{ij} + r_{jk} + b_{jkl} + \epsilon_{ijklm}$

Where, μ = overall mean, g_i = random effect of the ith genotype, e_j = random effect of the jth environment, $(ge)_{ij}$ = is the interaction effect of ith genotypes with the jth environment, r_{jk} = fixed effect of the kth replication nested within jth e
\$b_{jkl}=random effect of lth block nested with j environment and kth replication,□_{ijkl}=residual error,here we assume residuals a

---

- Mixed models are powerful tools to handle assumptions of linear model Read this one
- We will extract variance components and also calculate heritability.

```
> demo.data.filtered$Environment<-as.factor(demo.data.filtered$Environment)
> Model3.lme4<-lmer(Yield~Genotype+(1|Rep)+(1|Environment:Genotype)+
+           +(1|Environment:Rep:Block), data=demo.data.filtered)
```

Summary of MET results

- In summary we will get following summarized results: 1) Description of model we used, 2) Random effects and varainces, 3) Fixed effects, 4) Correlation of fixed efefcts

```
>   summary(Model3.lme4)
```

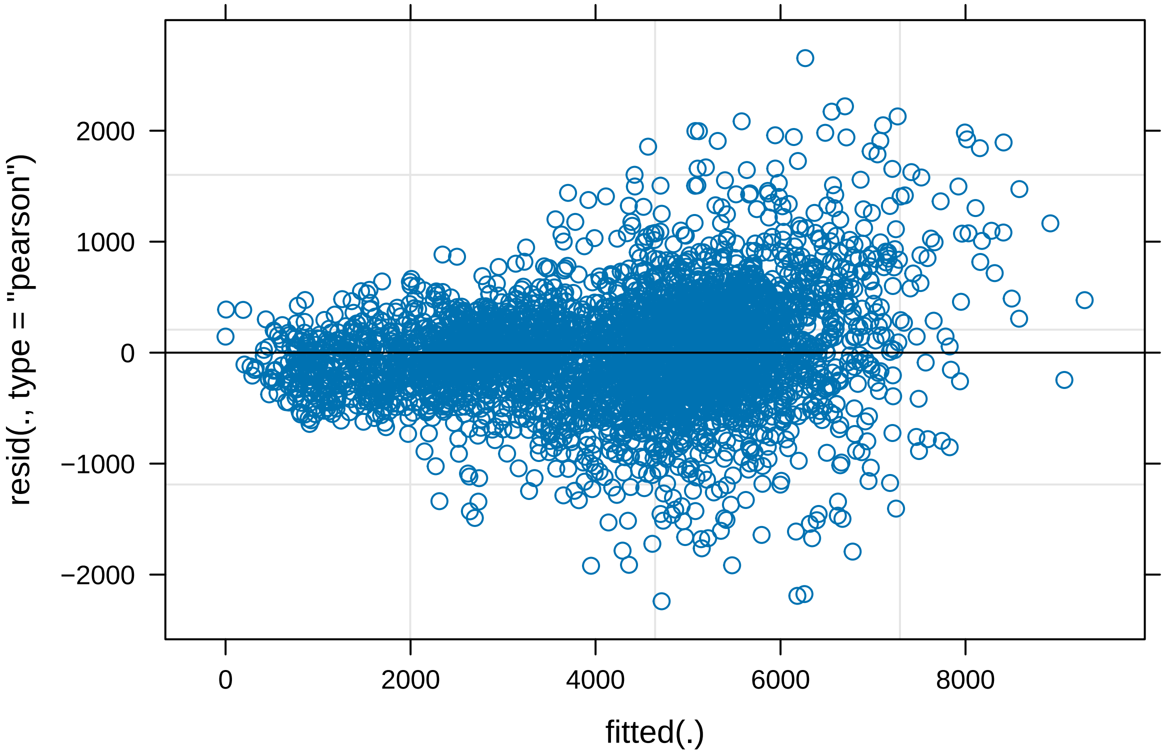
```
Linear mixed model fit by REML ['lmerMod']
Formula: Yield ~ Genotype + (1 | Rep) + (1 | Environment:Genotype) + +(1 |
    Environment:Rep:Block)
   Data: demo.data.filtered

REML criterion at convergence: 62923.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5042 -0.4373 -0.0158  0.4092  4.1520

Random effects:
 Groups                Name        Variance  Std.Dev.
 Environment:Genotype  (Intercept) 6.559e+05 8.099e+02
 Environment:Rep:Block (Intercept) 1.859e+06 1.363e+03
 Rep                   (Intercept) 5.465e-03 7.393e-02
 Residual                          4.086e+05 6.393e+02
Number of obs: 3964, groups:
Environment:Genotype, 1988; Environment:Rep:Block, 96; Rep, 2

Fixed effects:
              Estimate Std. Error t value
(Intercept)  4006.5428   325.7186  12.301
Genotype2    -158.3061   418.7820  -0.378
Genotype3    -100.7883   416.7684  -0.242
Genotype4     685.7349   416.6863   1.646
Genotype5     395.5221   416.9535   0.949
Genotype6     409.2635   415.8521   0.984
Genotype7    -355.5394   416.8936  -0.853
Genotype8     546.5261   416.4542   1.312
Genotype9    -480.3568   416.5919  -1.153
Genotype10   -474.1254   418.4225  -1.133
Genotype11    589.1146   416.2162   1.415
Genotype12    215.6045   416.2547   0.518
Genotype13     49.1649   416.0796   0.118
Genotype14    477.5898   416.8830   1.146
Genotype15   -260.8794   416.2345  -0.627
Genotype16   -223.6248   428.3856  -0.522
Genotype17    241.2194   416.6967   0.579
Genotype18    197.2838   416.2462   0.474
Genotype19   -411.0785   416.8845  -0.986
Genotype20    377.8620   416.5646   0.907
Genotype21    -26.2123   415.8055  -0.063
Genotype22    489.4758   416.5019   1.175
Genotype23      0.6426   416.7124   0.002
Genotype24    642.2857   419.0594   1.533
Genotype25    304.2129   416.6928   0.730
Genotype26     -2.0851   416.2181  -0.005
Genotype27   -309.2393   416.5684  -0.742
Genotype28   -258.6101   416.7560  -0.621
Genotype29    165.2624   417.9146   0.395
Genotype30    104.9037   415.9059   0.252
Genotype31   -670.4483   416.7135  -1.609
Genotype32    583.0878   416.5086   1.400
```

```
Genotype33    -14.6119   416.2335  -0.035
 [ reached getOption("max.print") -- omitted 167 rows ]
```

Plot of model

- With the plot function model we will get the residuals vs fitted values

```
> plot(Model3.lme4)
```



```
> Ve<- data.frame (VarCorr(Model3.lme4))
> Ve
```

Extract the variance components

grp	var1	var2	vcov	sdcor
Environment:Genotype	(Intercept)	NA	6.559377e+05	809.8998098
Environment:Rep:Block	(Intercept)	NA	1.858562e+06	1363.2908155
Rep	(Intercept)	NA	5.465100e-03	0.0739264
Residual	NA	NA	4.086498e+05	639.2572114

Heritability

- Here will estimate the combined heritability based on **Cullis et al.2006**

```
> #std.err<-se.ranef(Model3.lme4)$Genotype
> #v_BLUP<- mean(std.err)
> # Heritability/Reliability
> #h2<- (1-((v_BLUP)^2/(Ve[2,4]*2)))*100
> #h2
```

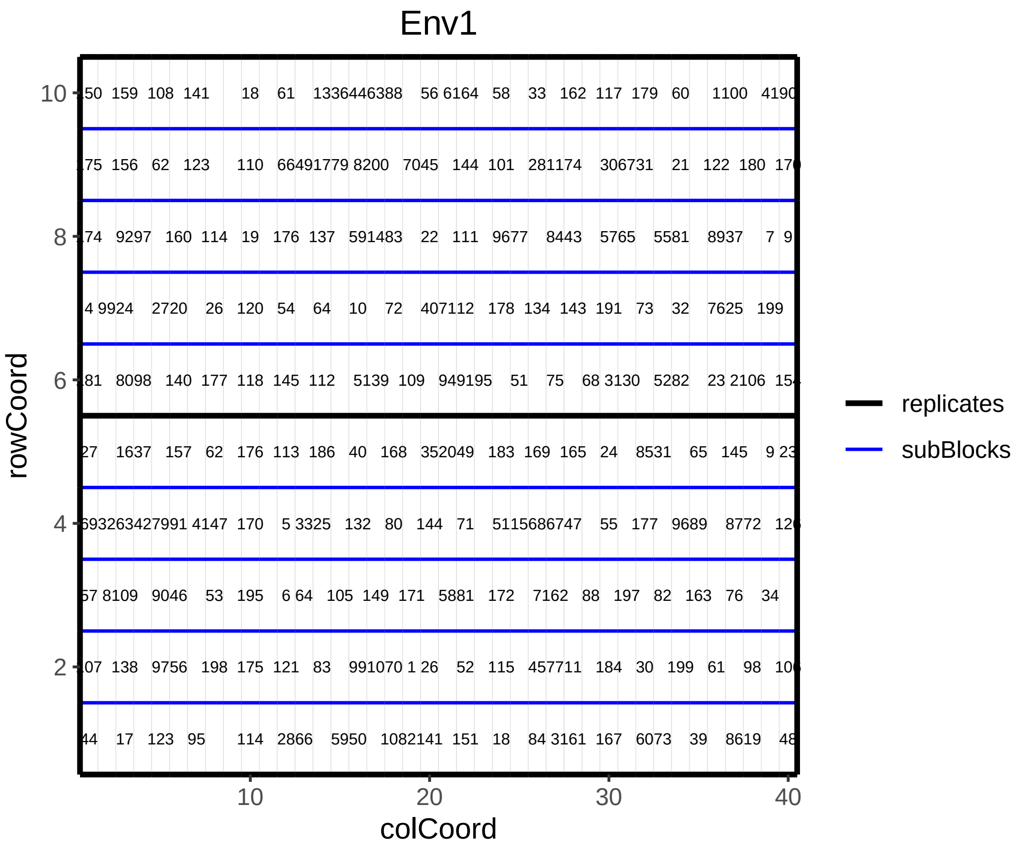
BLUEs for Random Effects

```
> # BLUEs
>   BLUEs.all<-data.frame(BLUEs.Yield=fixef(Model3.lme4))
>   #BLUPs<-data.frame(BLUps.GY=coef(Model3.lme4)$Genotype[,1])
>   head(BLUEs.all)
```

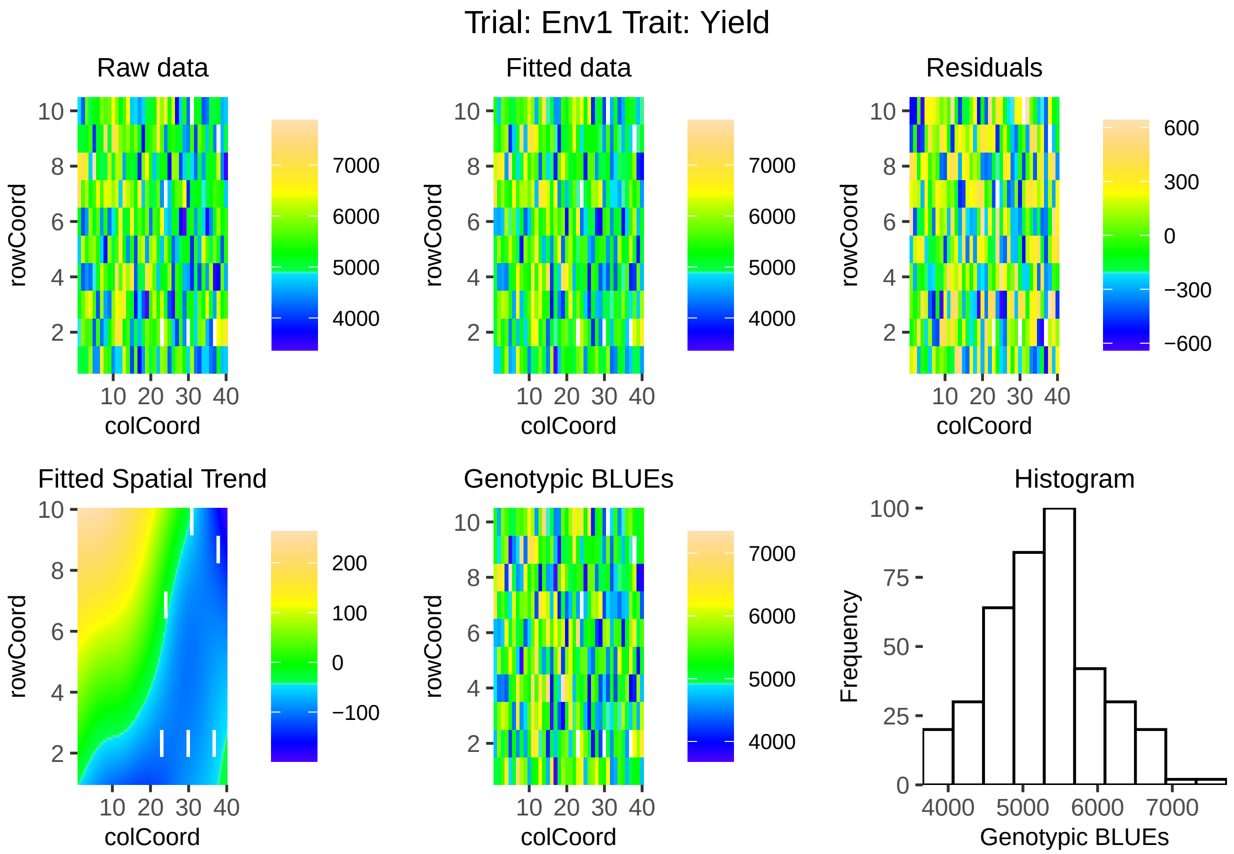
	BLUEs.Yield
(Intercept)	4006.5428
Genotype2	-158.3061
Genotype3	-100.7883
Genotype4	685.7349
Genotype5	395.5221
Genotype6	409.2635

Accounting Spatial Varaibility

```
> library(statgenSTA)
> TD_STA <- createTD(sub.data,
+                    trial = "Environment",
+                    genotype = "Genotype",
+                    rowCoord = "Row",
+                    colCoord = "Column",
+                    repId = "Rep",
+                    subBlock = "Block",
+                    trDesign = "res.ibd")
>
> ## Create layout plot with variety labels
> plot(TD_STA,
+         plotType = "layout",
+         showGeno = TRUE)
```



```
> ## Model specification (using engine = "SpATS")
> sta_model_SpATS <- fitTD(TD = TD_STA,
+                          traits = "Yield",
+                          design = "res.ibd",
+                          what = "fixed",
+                          spatial = TRUE,
+                          engine = "SpATS")
+
>
> plot(sta_model_SpATS,
+      plotType = "spatial",
+      traits = 'Yield')
```



```
> ## Extract all available statistics from the fitted model.
>   extr <- extractSTA(sta_model_SpATS)
> ## Extract only the BLUEs from the fitted model.
>   BLUEs <- extractSTA(sta_model_SpATS,
+                       what = "BLUEs")
```

Additional on MET and Stability Analysis

- Here in this section we are giving some useful *R* resources that can be used for stability and MET analysis.

1. metan-R: Multi-environment Trial Analysis
2. gge-R: Functions for GGE and GGB

Additional Literature

- Screening experimental designs
- Analysis and Handling of G × E in a Practical Breeding Program
- A stage□wise approach for the analysis of multi□environment trials
- Analysis of series of variety trials with perennial crops
- A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels

- Experimental design matters for statistical analysis: how to handle blocking
- Random effects structure for confirmatory hypothesis testing: Keep it maximal
- Generalized linear mixed models: a practical guide for ecology and evolution
- Mixed Models Offer No Freedom from Degrees of Freedom
- Perils and pitfalls of mixed-effects regression models in biology
- A brief introduction to mixed effects modelling and multi-model inference in ecology
- Modeling Spatially Correlated and Heteroscedastic Errors in Ethiopian Maize Trials
- More, Larger, Simpler: How Comparable Are On□Farm and On□Station Trials for Cultivar Evaluation
- Rethinking the Analysis of Non□Normal Data in Plant and Soil Science
- The Design and Analysis of Long□Term Rotation Experiments
- Analysis of Combined Experiments Revisited
- Fundamentals of Experimental Design: Guidelines for Designing Successful Experiments

---

For any suggestions or comments, please feel to reach at waseem.hussain@cgiar.org; m.anumalla@cgiar.org; m.catolos@cgiar.org

---

If your experiment needs a statistician, you need a better experiment - Ernest Rutherford