Foundations of linear mixed models

Luis Fernando Delgado Munoz
luis.delgado@cgiar.org
Faisalabad, November 2023

@BiovIntCIAT_eng
@BiovIntCIAT_esp

#Alliance4Science

The (very) basics about mixed models

Mixed models consist of:

Fixed terms

- Interest in specific levels
- Parameters of interest are individual treatment effects

Random terms

- Treatment levels are chosen to represent a population of possible treatments
- Parameters of interest variance parameters

Analysis of large experiments.

Analysis steps:

1. Fit a model with **genotypes as random**
   - to calculate trial heritability
   - genotype means are called 'BLUPs'

2. Fit a model with **genotypes as fixed**
   - to extract genotype adjusted means (BLUEs) to be used for variety selection, GxE analysis or QTL detection.

Heritability (broad sense)

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_\epsilon^2}{n}}$$

σ_g^2 = genotypic variance

σ_ϵ^2 = error variance (residual)

n= number of replicates

$H^2 = 0$
variation is not explained by the genotypes

$H^2 = 1$
variation is fully explained by the genotypes

Analysis of large experiments

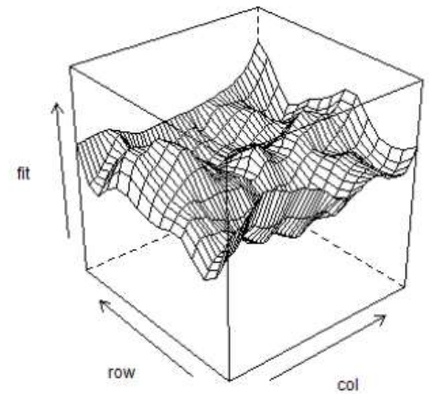
Example: each experiment has:

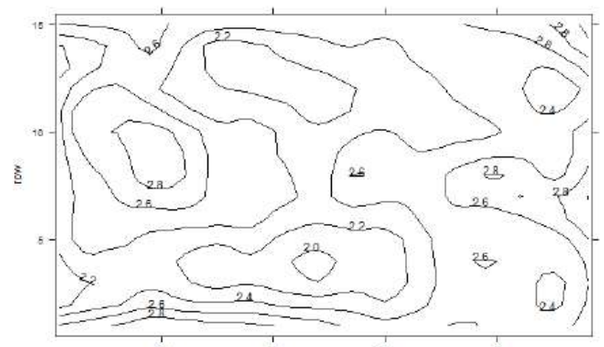
- 150 barley genotypes
- 2 replicates
- The only difference is the design that was used:

a) RCBD
b) Alpha
c) Row-column
d) Resolvable row-column

Uniformity trial (barley)

- **15 rows x 48 columns grid**
- **Variation across experiment due to uncontrolled conditions (="noise")∘Minimum 1.490 ton/ha**
- **Maximum 2.290 ton/ha**
- **Variance = 0.084 (ton/ha)2**
- **Std = 0.289 ton/ha**





The use of uniformity data in the design and analysis of cotton and barley variety trials

ER Williams and DJ Luckett

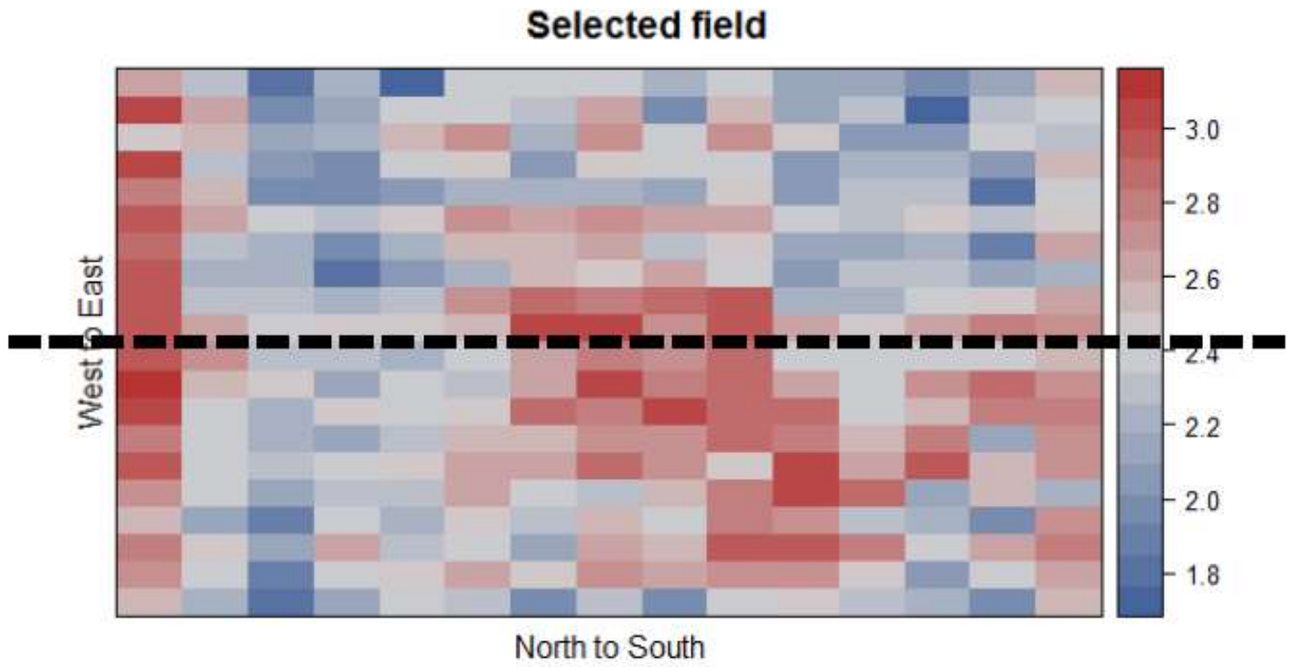
Australian Journal of Agricultural Research 39(3) 339 - 350

Published: 1988

https://doi.org/10.1071/AR9880339

Suppose we plan a trial in this field …

- 150 barley genotypes
- 2 reps each



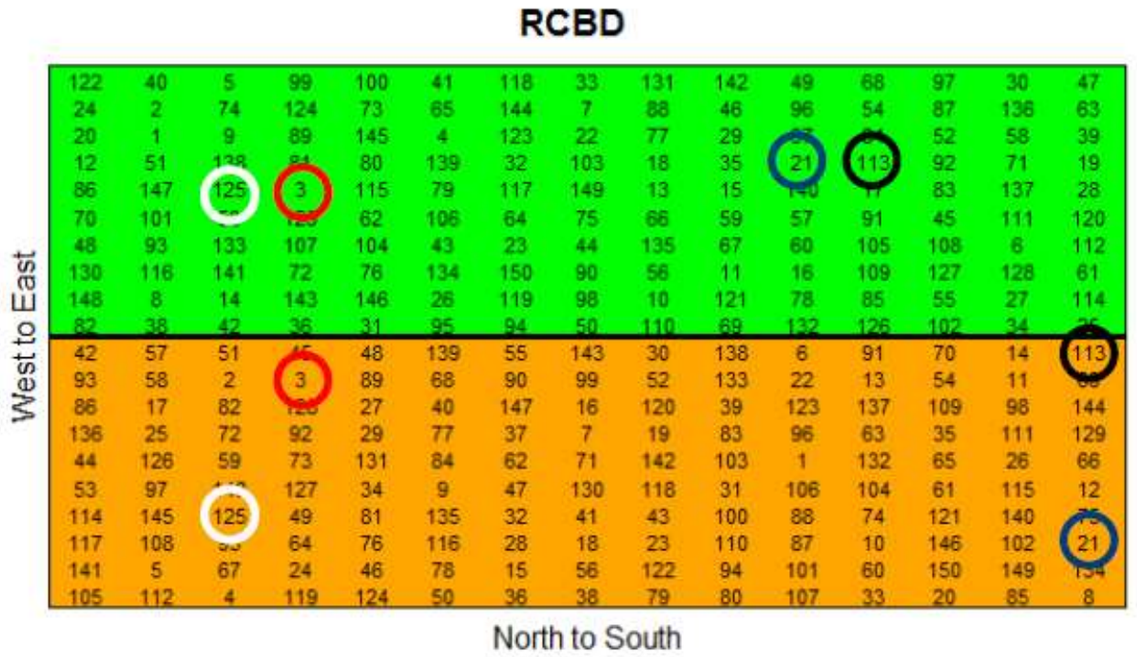
Two complete blocks = RCBD

Estimates: surrogates of true value

Phenotype	Phenotypic mean	BLUE	BLUP	PBLUP	GBLUP
The phenotypic value of a single observation (e.g. plant or plot) is a surrogate of the observation's true genetic value. However, the phenotype deviates from the true genetic value due to non-genetic effects like environment and random chance. How much phenotypes deviate from the true genetic value can be estimated by measuring heritability. If phenotypes have high heritabilities, they are relatively accurate estimates of true value. If not, they are relatively inaccurate, and other surrogates can increase accuracy.	Phenotypic means of multiple observations are generally more accurate than single phenotypic values. This is because the error component of phenotypes is random, so the mean of the phenotypes is closer to the true genetic value than a single phenotype. Phenotypic means derived from phenotypes with high heritabilities still have higher accuracies than phenotypic means derived from phenotypes with low heritabilities. The increase in accuracy from using the phenotypic mean vs. phenotype is higher if the phenotypes have low heritabilities.	Best linear unbiased estimates can be obtained from linear regression of phenotypes on fixed effects, e.g.: Pheno ~ mean + block + genotype + error BLUEs allow more complex evaluation strategies which can increase accuracy (e.g. blocking) and accounting for non-genetic effects. In absence of non-genetic effects besides error and with equal replication, BLUEs of genetic value are equal to phenotypic means. BLUEs can increase accuracy compared to using phenotypic means indirectly by enabling use of experimental designs which increase accuracy.	Best linear unbiased predictions can also be obtained from linear regression of phenotypes, but the regression is on random effects or a mix of fixed and random effects, e.g.: Pheno ~ mean + block + genotype + error The random genotype effects, which estimate genetic value, are shrunk to the mean by an estimate of heritability for the genotype. Genotypes with fewer replications (as in unbalanced data) have lower heritabilities, so estimates of their genetic value are shrunk more to anticipate their regression to the mean. This can increase accuracy.	Pedigree best linear unbiased predictions are similar to BLUPs, but the random genetic effects are not assumed to be independent (since genotypes are related): Pheno ~ mean + block + genotype + error |	A	B
---	---	---
A	1	0.50
B	0.5	1 | The genetic effects are not only shrunk by heritability, but also weighted by the pedigree expectation of their relatedness to other observations. This borrowing of information from relatives can increase accuracy in some scenarios.	Genomic best linear unbiased predictions are quite similar to PBLUPs, except that genotype relatedness is estimated from markers rather than pedigree expectations. Pheno ~ mean + block + genotype + error |	A	B
---	---	---
A	1	0.54
B	0.54	1	 Genotype relatedness deviates from pedigree expectations because of recombination (Mendelian sampling), so accounting for this deviation can increase accuracy compared to PBLUP in some scenarios.

RCBD

150 barley genotypes
2 reps each



- All 150 present once within each replicate (green and orange area).

Analysis RCBD trial (genotypes random = BLUPs)

$$y_{ik} = \mu + b_k + G_i + \epsilon_{ik}$$

$G_i \sim N(0, \sigma_g^2)$
$\epsilon_{ik} \sim N(0, \sigma_\epsilon^2)$

```
m1 <- lmer(yield ~ block + (1|geno), data = mytrial_RCBD$book)
summary(m1)
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 geno     (Intercept) 0.01946  0.1395
 Residual             0.08233  0.2869
Number of obs: 300, groups:  geno, 150

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  2.50360    0.02605 287.49000  96.110  < 2e-16 ***
block2      -0.12587    0.03313 149.00000  -3.799 0.000211 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_\epsilon^2}{n}}$$

$$H^2 = \frac{0.01946}{0.01946 + \frac{0.08233}{2}} = 0.32$$

Analysis RCBD trial (genotypes fixed = BLUEs)

$$y_{ik} = \mu + b_k + G_i + \epsilon_{ik} \qquad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon)$$

```
m1f <- lm(yield ~ block + geno, data = mytrial_RCBD$book)
anova(m1f)
lsmeans(m1f, "geno")
```

Analysis of Variance Table

Response: yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
block	1	1.1882	1.18818	14.4326	0.0002111	***
geno	149	18.0652	0.12124	1.4727	0.0093520	**
Residuals	149	12.2666	0.08233			

geno	lsmean	SE	df	lower.CL	upper.CL
1	2.654493	0.202887	149	2.253585	3.055400
2	2.498493	0.202887	149	2.097585	2.899400
3	2.086625	0.202887	149	1.685717	2.487532
4	1.928961	0.202887	149	1.528053	2.329868
5	2.140785	0.202887	149	1.739877	2.541692
6	2.303412	0.202887	149	1.902504	2.704319
7	2.808384	0.202887	149	2.407476	3.209291
8	2.400185	0.202887	149	1.999278	2.801093
9	2.282892	0.202887	149	1.881985	2.683800
10	2.834490	0.202887	149	2.433583	3.235398

- Adjusted means (= simple average over two blocks).
- Compare means using sed

$$sed = \sqrt{2 \times MSE/n} = \sqrt{2 \times 0.08233/2} = 0.287$$

(k=1...n), n= 2 replicates

RCBD Precision experiment → Least Significant Difference

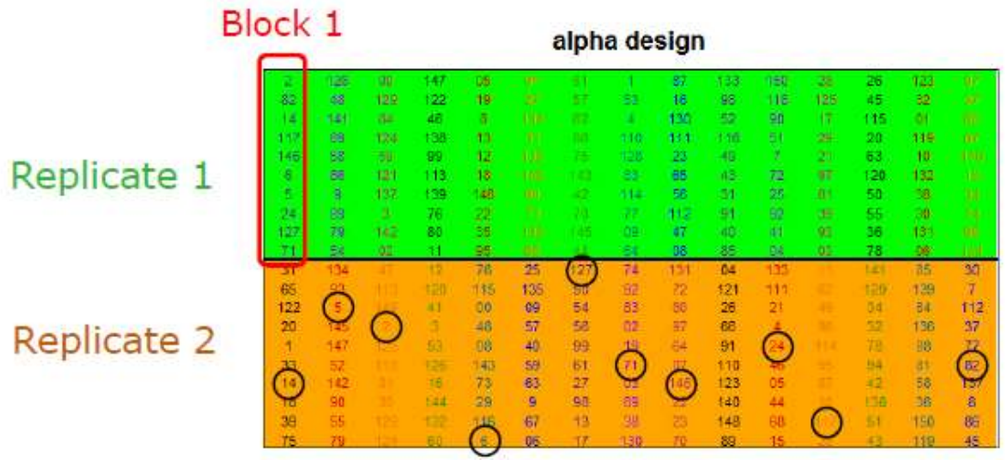
$$LSD = 2 * sed = 2 * 0.287 = 0.574\ ton/ha$$

Standard error of the difference RCBD – the same

Because all genotypes are tested in all blocks, the s.e.is the same for all genotype comparisons

```
diff_rcbd[1:20,1:4]
 contrast     estimate        SE
    1 - 2 -0.466782955 0.2744649
    1 - 3 -0.013491128 0.2744649
    1 - 4 -0.107449108 0.2744649
    1 - 5 -0.150717265 0.2744649
    1 - 6  0.108819636 0.2744649
    1 - 7  0.165882834 0.2744649
    1 - 8  0.001472315 0.2744649
    1 - 9 -0.289928923 0.2744649
   1 - 10 -0.061383326 0.2744649
   1 - 11 -0.157999909 0.2744649
   1 - 12 -0.066398252 0.2744649
   1 - 13  0.067532540 0.2744649
   1 - 14  0.226920294 0.2744649
   1 - 15 -0.450909076 0.2744649
   1 - 16 -0.030708723 0.2744649
   1 - 17 -0.017941970 0.2744649
   1 - 18  0.161718042 0.2744649
   1 - 19  0.126490129 0.2744649
   1 - 20  0.703500385 0.2744649
   1 - 21  0.015052154 0.2744649
```

Alpha design



- Resolvable: 2 full replicates
- 15 Incomplete blocks within each replicate
  - Blocks occur along the columns direction
  - A subset of genotypes within a particular block
  - Partially-balanced: genotypes occur together in a block either once ($\lambda_2 = 1$) or never ($\lambda_1 = 0$) over the whole experiment → $\alpha(0,1)$ design

Analysis α-design trial (genotypes random = BLUPs)

$$y_{ijk} = \mu + R_j + b_{k(j)} + G_i + \epsilon_{ijk}$$

$$G_i \sim N(0, \sigma_g^2)$$
$$b_{k(j)} \sim N(0, \sigma_b^2)$$
$$\epsilon_{ijk} \sim N(0, \boldsymbol{\sigma_\epsilon^2})$$

When analysing an alpha design, blocks MUST be random, because they are incomplete

```
m2 <- lmer(yield ~ replication + (1|block:replication) + (1|geno),
           data = mytrial_alpha$book)
summary(m2)
```

```
Random effects:
 Groups            Name        Variance Std.Dev.
 geno              (Intercept) 0.02508  0.1584
 block:replication (Intercept) 0.03463  0.1861
 Residual                      0.03840  0.1960
Number of obs: 300, groups:  geno, 150; block:replication, 30

Fixed effects:
             Estimate Std. Error       df t value Pr(>|t|)
(Intercept)   2.50360    0.05227 28.41700  47.902   <2e-16 ***
replication2 -0.12587    0.07162 25.38300  -1.758   0.0909 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_\epsilon^2}{n}}$$

$$H^2 \cong \frac{0.02508}{0.02508 + \frac{0.03840}{2}} = 0.56$$

Analysis α-design trial (genotypes fixed = BLUEs)

$$y_{ijk} = \mu + R_j + b_{k(j)} + G_i + \epsilon_{ijk}$$

$$b_{k(j)} \sim N(0, \sigma_b^2)$$
$$\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$$

```
m2f <- lmer(yield ~ replication + (1|block:replication) + geno,
            data = mytrial_alpha$book)
```

```
Least Squares Means table:
          replication geno Estimate Standard Error  DF t-value Lower CI Upper CI
geno  1            NA    1    2.315          0.151 142    15.3     2.02     2.61
geno  2            NA   63    2.289          0.151 142    15.2     1.99     2.59
geno  3            NA   74    2.379          0.151 142    15.8     2.08     2.68
geno  4            NA   85    2.257          0.151 142    14.9     1.96     2.56
geno  5            NA   96    2.507          0.151 142    16.6     2.21     2.81
geno  6            NA  107    2.621          0.151 142    17.4     2.32     2.92
geno  7            NA  118    2.596          0.151 142    17.2     2.30     2.89
geno  8            NA  129    2.686          0.151 142    17.8     2.39     2.98
geno  9            NA  140    2.203          0.151 142    14.6     1.90     2.50
geno 10            NA    2    2.381          0.151 142    15.8     2.08     2.68
```

- Adjusted means (≠ simple average over two reps).
- Lower standard errors! (was 0.202 and now 0.151!)
- Standard error of difference = 0.206

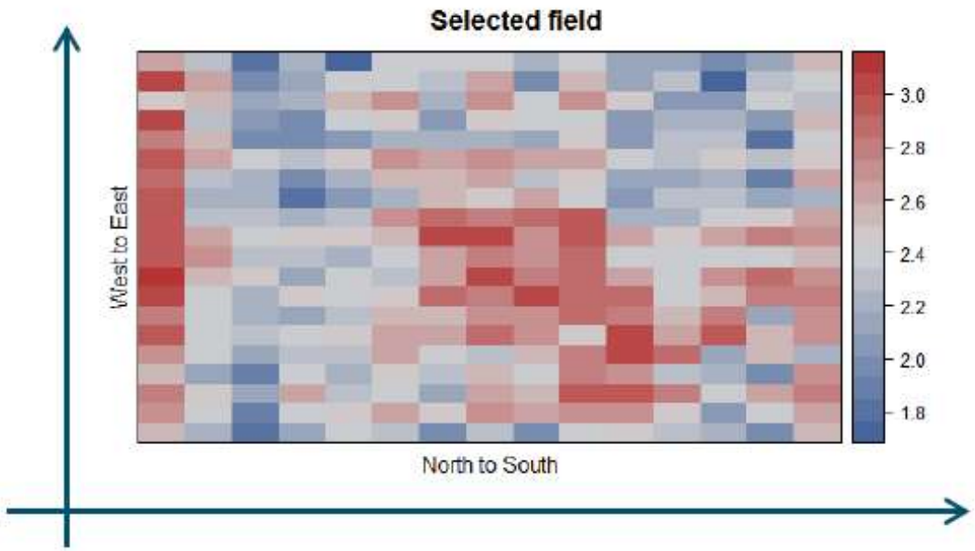
Higher precision than the RCBD → $LSD = 2 \times 0.206 = 0.412\ ton/ha$

Standard error of the difference α-design

- It might differ between pairs
- The SE depends on the incomplete block and error variance and on with whom the genotypes occur together in incomplete blocks

```
> diff_alpha$contrasts
 contrast   estimate    SE
 1 - 2     -7.44e-01 0.235
 1 - 3     -2.69e-01 0.234
 1 - 4     -3.78e-01 0.236
 1 - 5     -4.12e-01 0.235
 1 - 6     -3.49e-01 0.235
 1 - 7     -6.15e-01 0.235
 1 - 8     -1.87e-01 0.235
 1 - 9     -4.43e-01 0.236
 1 - 10    -8.44e-02 0.236
 1 - 11    -1.46e-01 0.234
 1 - 12    -1.89e-01 0.234
 1 - 13    -2.50e-01 0.235
 1 - 14    -3.81e-01 0.226
 1 - 15    -4.66e-01 0.234
 1 - 16    -2.51e-01 0.235
 1 - 17    -3.27e-01 0.234
 1 - 18    -1.77e-01 0.234
 1 - 19    -5.75e-01 0.234
 1 - 20    -2.36e-01 0.226
```

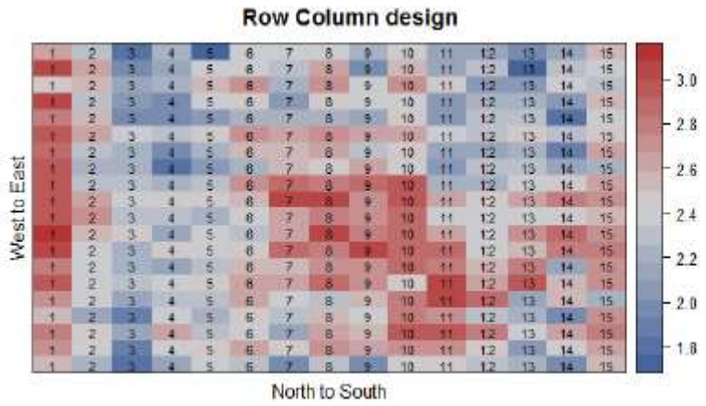
Double blocking: row-column designs

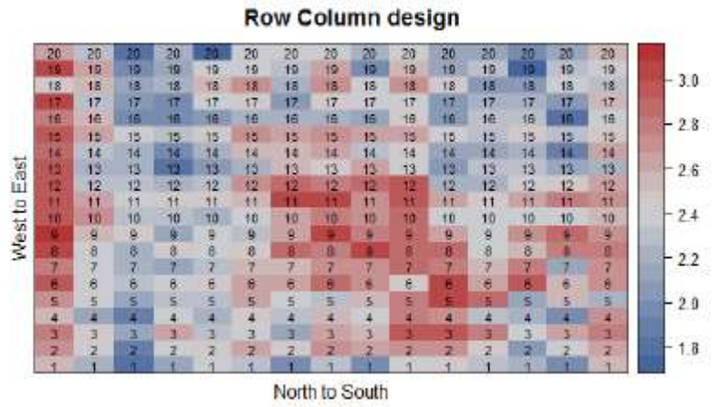


- Regard the field as a full $(r \times c)$ grid.
- Rows and columns → incomplete blocks over both directions.

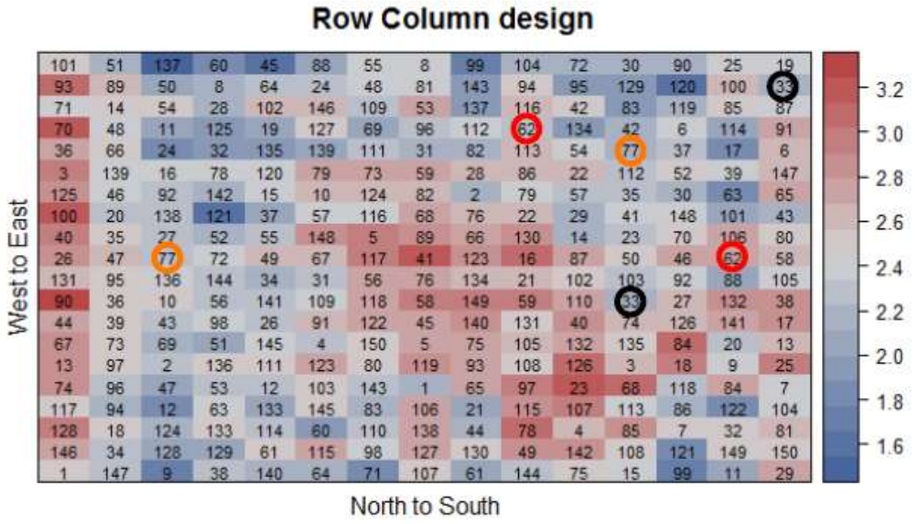
Double direction blocking

Columns





Rows



Analysis of a row-column design (genotypes random = BLUPs)

$$y_{ijk} = \mu + R_j + C_k + G_i + \epsilon_{ijk}$$

$$G_i \sim N(0, \sigma_g^2)$$
$$R_j \sim N(0, \sigma_r^2)$$
$$C_k \sim N(0, \sigma_c^2)$$
$$\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$$

```
m3 <- lmer(yield ~ (1|Row) + (1|Col) + (1|Geno),
           data = rowcol1$Book)
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 Geno     (Intercept) 0.02096  0.1448
 Col      (Intercept) 0.02456  0.1567
 Row      (Intercept) 0.03075  0.1754
 Residual             0.03217  0.1794
Number of obs: 300, groups:  Geno, 150; Col, 20; Row, 15

Fixed effects:
            Estimate Std. Error t value
(Intercept)  2.44067    0.05937   41.11
```

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_\epsilon^2}{n}}$$

$$H^2 \cong \frac{0.02096}{0.02096 + \frac{0.03217}{2}} = 0.57$$

- Note that the model consists of rows and columns
  - Non-resolvable design because subsets of rows/cols do not form complete replicates
- Similar H^2 as with the α-design (and better than RCBD).

Analysis of a row-column design (genotypes fixed>BLUEs)

$$y_{ijk} = \mu + R_j + C_k + G_i + \epsilon_{ijk}$$

$$R_j \sim N(0, \sigma_r^2)$$
$$C_k \sim N(0, \sigma_c^2)$$
$$\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$$

```
m3f <- lmer(yield ~ (1|Row) + (1|Col) + Geno,
            data = rowcol1$Book)
```

```
Least Squares Means table:
          Geno Estimate Standard Error  DF t-value Lower CI Upper CI p-value
Geno  1      1    2.146          0.148 147  14.530     1.85     2.44  <2e-16 ***
Geno  2     63    2.360          0.149 147  15.860     2.07     2.65  <2e-16 ***
Geno  3     74    2.503          0.147 147  16.990     2.21     2.79  <2e-16 ***
Geno  4     85    2.353          0.148 147  15.930     2.06     2.64  <2e-16 ***
Geno  5     96    2.712          0.148 147  18.310     2.42     3.00  <2e-16 ***
Geno  6    107    2.645          0.148 147  17.930     2.35     2.94  <2e-16 ***
Geno  7    118    2.367          0.148 147  16.040     2.08     2.66  <2e-16 ***
Geno  8    129    2.457          0.148 147  16.600     2.16     2.75  <2e-16 ***
Geno  9    140    2.249          0.148 147  15.250     1.96     2.54  <2e-16 ***
Geno  10     2    2.604          0.148 147  17.580     2.31     2.90  <2e-16 ***
```

When analysing a row-column design, rows and columns MUST be random, because they are incomplete

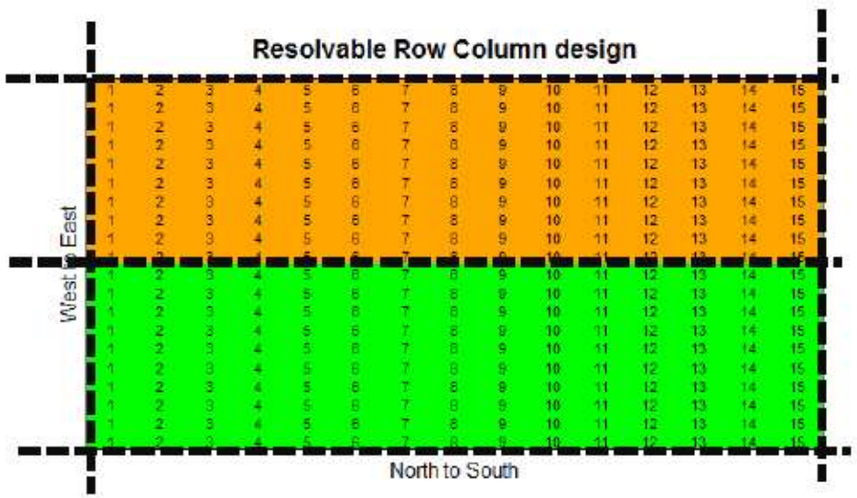
- Adjusted means (≠ simple average over two reps).
- Standard errors slightly lower than with alpha design
- Average standard error of difference = 0.192

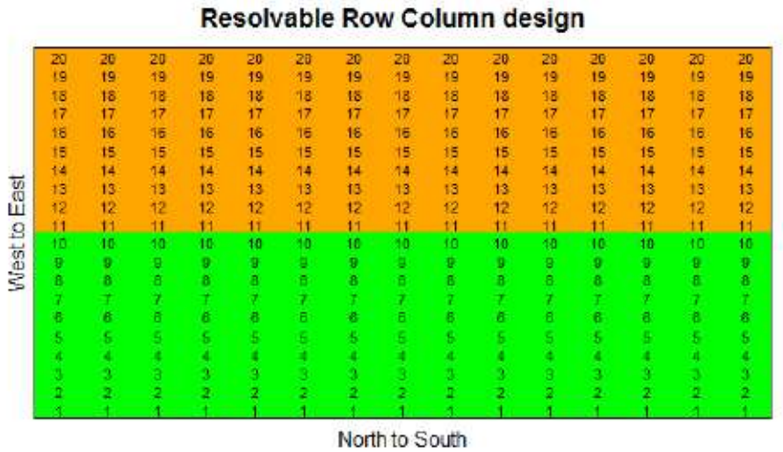
Standard error of the difference row column design

- It might differ between pairs
- SE depends on row, column, and error variance and which genotypes occur together in rows and columns.

```
> diff_rowcol$contrasts
 contrast    estimate    SE
 1 - 2     -2.14e-01 0.195
 1 - 3     -3.57e-01 0.190
 1 - 4     -2.07e-01 0.195
 1 - 5     -5.66e-01 0.191
 1 - 6     -4.99e-01 0.194
 1 - 7     -2.21e-01 0.189
 1 - 8     -3.11e-01 0.190
 1 - 9     -1.04e-01 0.189
 1 - 10    -4.58e-01 0.195
 1 - 11    -2.79e-01 0.189
 1 - 12    -8.41e-02 0.189
 1 - 13    -2.35e-01 0.190
 1 - 14    -1.83e-01 0.195
 1 - 15    -5.42e-01 0.189
 1 - 16    -4.26e-01 0.197
 1 - 17    -1.82e-01 0.195
 1 - 18    -5.85e-01 0.194
 1 - 19    -2.47e-01 0.194
 1 - 20    -2.54e-01 0.194
```

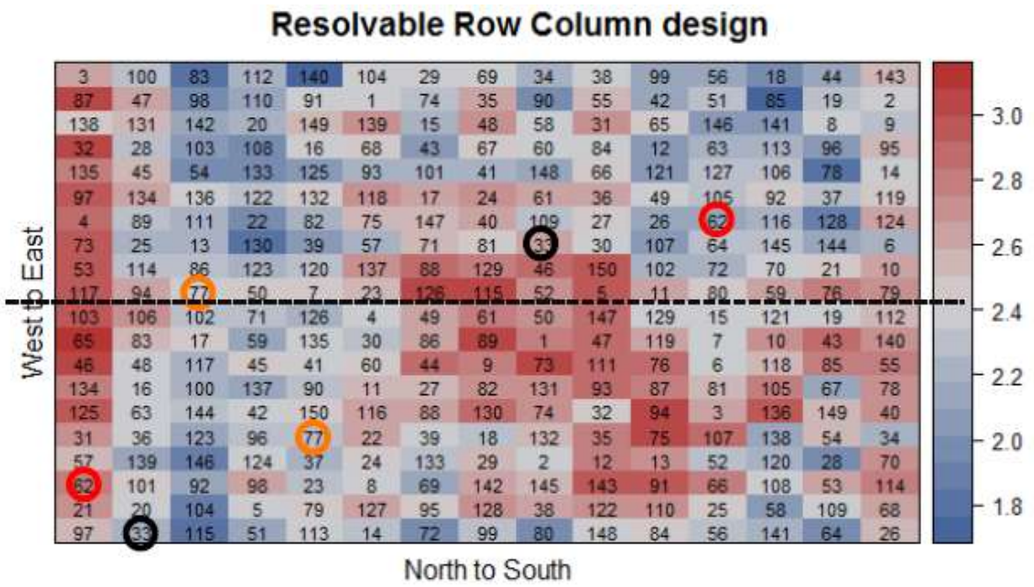
Resolvable Row-Column design





- Define two full replicates → make it resolvable
- Within Replicates incomplete rows and columns (as before).

Randomization: resolvable row-column



- All 150 in each half of the field.
- Genotypes within same row (column) in rep 1 will not be in the same row (column) in rep 2.

Analysis Resolvable Row-Column design (genotypes random = BLUPs)

$$y_{ijkm} = \mu + Rep_m + R_{j(m)} + C_{k(m)} + G_i + \epsilon_{ijkm}$$

$$G_i \sim N(0, \sigma_g^2)$$
$$R_j \sim N(0, \sigma_r^2)$$
$$C_k \sim N(0, \sigma_c^2)$$
$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$$

```
m4 <- lmer(yield ~ Rep + (1|Rep:Row) + (1|Rep:Col) + (1|Geno),
           data = rowcol2$Book)

Random effects:
 Groups   Name        Variance Std.Dev.
 Geno     (Intercept) 0.02652  0.1629
 Rep:Row  (Intercept) 0.03792  0.1947
 Rep:Col  (Intercept) 0.01986  0.1409
 Residual             0.02464  0.1570
Number of obs: 300, groups:  Geno, 150; Rep:Row, 30; Rep:Col, 20

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  2.37774    0.06968 42.04000  34.123   <2e-16 ***
Rep1         0.12587    0.09673 39.15000   1.301    0.201
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_\epsilon^2}{n}}$$

$$H^2 \cong \frac{0.02652}{0.02652 + \frac{0.02464}{2}} = 0.68$$

- Note that rows and columns are **nested** within replicates (the : operator used in R).
- Heritability is the highest among all the designs.

Analysis resolvable row-column (genotypes fixed = BLUEs)

$$y_{ijkm} = \mu + Rep_m + R_{j(m)} + C_{k(m)} + G_i + \epsilon_{ijkm}$$

$$R_j \sim N(0, \sigma_r^2)$$
$$C_k \sim N(0, \sigma_c^2)$$
$$\epsilon_{ijkm} \sim N(0, \sigma_\epsilon^2)$$

```
m4f <- lmer(yield ~ Rep + (1|Rep:Row) + (1|Rep:Col) + Geno,
            data = rowcol2$Book)
```

Least Squares Means table:

		Rep	Geno	Estimate	Standard Error	DF	t-value	Lower CI	Upper CI	p-value	
Geno	1	NA	1	2.379	0.136	140	17.6	2.11	2.65	<2e-16	***
Geno	2	NA	63	2.379	0.135	140	17.6	2.11	2.65	<2e-16	***
Geno	3	NA	74	2.355	0.134	139	17.5	2.09	2.62	<2e-16	***
Geno	4	NA	85	2.241	0.135	139	16.6	1.97	2.51	<2e-16	***
Geno	5	NA	96	2.632	0.135	140	19.5	2.36	2.90	<2e-16	***
Geno	6	NA	107	2.312	0.136	140	17.0	2.04	2.58	<2e-16	***
Geno	7	NA	118	2.347	0.135	140	17.4	2.08	2.61	<2e-16	***
Geno	8	NA	129	2.382	0.133	139	17.9	2.12	2.65	<2e-16	***
Geno	9	NA	140	2.218	0.135	139	16.5	1.95	2.48	<2e-16	***
Geno	10	NA	2	2.484	0.135	139	18.4	2.22	2.75	<2e-16	***

- Adjusted means (≠ simple average over two reps).
- Standard errors lower than with α design and row-column.
- Average standard error of difference = 0.178

→ $LSD = 2 \times 0.178 = 0.356\ ton/ha$

Standard error in a resolvable row column design

- It might differ between pairs
- SE depends on row, column, and error variance and which genotypes occur together in rows and columns.

```
> diff_resrowcol$contrasts
 contrast    estimate    SE
 1 - 2       0.000383 0.165
 1 - 3       0.024864 0.181
 1 - 4       0.138576 0.183
 1 - 5      -0.252565 0.178
 1 - 6       0.067122 0.181
 1 - 7       0.032708 0.175
 1 - 8      -0.002491 0.182
 1 - 9       0.161730 0.181
 1 - 10     -0.104665 0.176
 1 - 11     -0.259703 0.180
 1 - 12     -0.073505 0.180
 1 - 13     -0.215201 0.177
 1 - 14     -0.125324 0.183
 1 - 15     -0.006636 0.182
 1 - 16     -0.148138 0.180
 1 - 17     -0.241101 0.176
 1 - 18      0.178004 0.182
 1 - 19      0.049960 0.175
 1 - 20     -0.098286 0.183
```

Broad sense heritability

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_\epsilon^2}{n}}$$

RCBD $H^2 \cong \frac{0.01946}{0.01946 + \frac{0.08233}{2}} = 0.32$	*Alpha design* $H^2 \cong \frac{0.02508}{0.02508 + \frac{0.03840}{2}} = 0.56$
Row-Col (non-resolvable) $H^2 \cong \frac{0.02096}{0.02096 + \frac{0.03217}{2}} = 0.57$	*Row-Col (resolvable)* $H^2 \cong \frac{0.02652}{0.02652 + \frac{0.02464}{2}} = 0.68$

Variance components estimated for models with genotypes as random

Least Significant Difference (LSD) – precision experiments

RCBD	$\rightarrow LSD = 2 * 0.287 = 0.574\ ton/ha$
Alpha design	$\rightarrow LSD = 2 \times 0.206 = 0.412\ ton/ha$
Row-Col (non-resolvable)	$\rightarrow LSD = 2 \times 0.192 = 0.384\ ton/ha$
Row-Col (resolvable)	$\rightarrow LSD = 2 \times 0.178 = 0.356\ ton/ha$

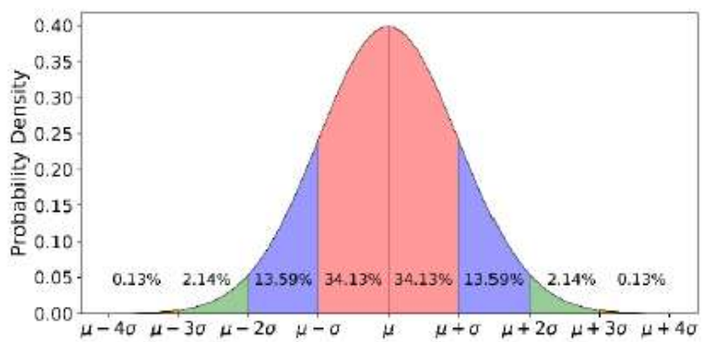
Estimated for models with genotypes fixed

Shrinkage (genotypes fixed or random?)

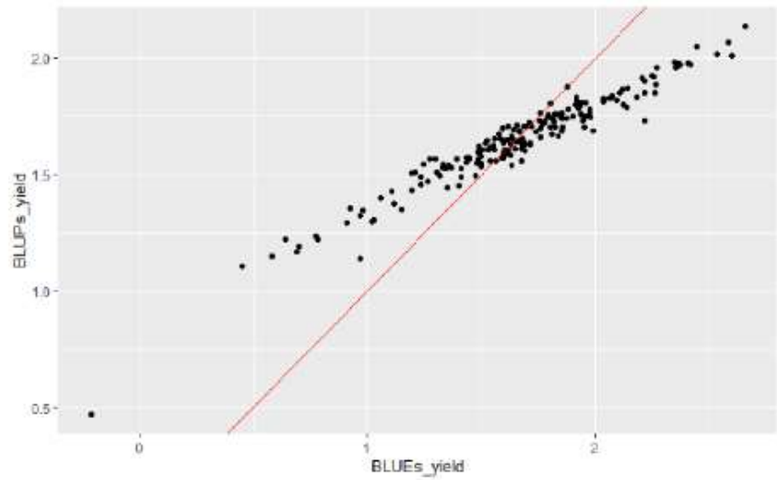
Shrinkage: Random effects are pulled towards the mean.

Extreme values are less likely to occur

$$\hat{g}_i = \frac{\sigma_g^2}{(\sigma_g^2 + {}^{\sigma_\epsilon^2}/_n)}(\bar{y}_{i.} - \bar{y}_{..})$$



Example: wheat data (next practical)




```
> # calculate the range for BLUES
> range(WHO9BB$BLUEs_yield)
[1] -0.2129382  2.6567850
>
>
> # calculate the range for BLUPS
> range(WHO9BB$BLUPs_yield)
[1] 0.4710533 2.1358297
```

Experimental design and analysis

- Good experimental design that helps to control extraneous variation.
  - Soil variability, operator/equipment effects, etc
- Precise data collection/phenotyping.
  - Trait definition (units, measurement procedure, etc)
- Advanced data analysis methods (simple ANOVA models far **too restrictive**)
  - Mixed models
  - Spatial models

Takeways

- With large trials we need to account for local variability
  - Smaller blocks
  - Row / column effects
- Also true in greenhouses and platforms!
- Clear design = clear analysis
- Useful designs are to separate spatial field variation from genotypic variation:
  - Alpha designs (genotypes assigned to incomplete blocks that form full replicates)
  - Row-column designs (blocking in the row and column directions)
- Design + advanced analyses leads to better/more accurate genotypic evaluation (e.g. higher Heritability)

Thank you.