

CH 05. Notions de *blocking*
Plan en blocs aléatoires complets
Randomized complete block design (RCBD)

Contexte

Idéalement, les unités expérimentales (UE) utilisées pour comparer les traitements devraient être homogènes, de manière à ce que la variance entre les UE, en absence de traitement, soit faible par rapport à l'effet des traitements qui seront appliqués.

Pourtant, si l'on songe à notre expérience en absence de traitement, il est assez aisé d'imaginer si des différences naturelles existent entre les UE.

Ex. 1: les pots disposés au côté N ou S d'une serre

Ex. 2: des échantillons qui sont analysés le matin ou l'après-midi

Le *blocking* consiste à regrouper les UE « semblables ». Il s'agit donc d'une restriction à la randomisation. En organisant les UE en blocs, on sépare la variance résiduelle en deux composantes: une variance entre unités d'un même bloc (**intra bloc**) et une variance entre les blocs (**inter bloc**). De cette manière, la comparaison de traitements au sein d'un bloc, se fait sur base d'une variabilité résiduelle plus faible et la précision des estimations des traitements est améliorée.

Nous verrons plus tard une autre motivation au *blocking* (cf *split plot design*).

Plan

1. Présentation du plan en blocs aléatoires complet
2. Exemple
3. Génération du plan
4. Modèle d'analyse
5. Inférence sur les traitements

Le plan en blocs aléatoires complets

Caractéristiques du RCBD:

- Les UE sont regroupées en blocs d'unités homogènes
- Au sein d'un bloc, les traitements sont assignés aux UE de manière **aléatoire**
- Les blocs sont **complets**: chaque bloc contient l'ensemble des traitements
- Il y a **une** UE par traitement au sein d'un bloc
- Le nombre de réplifications de chaque traitement est donc égal au nombre de blocs

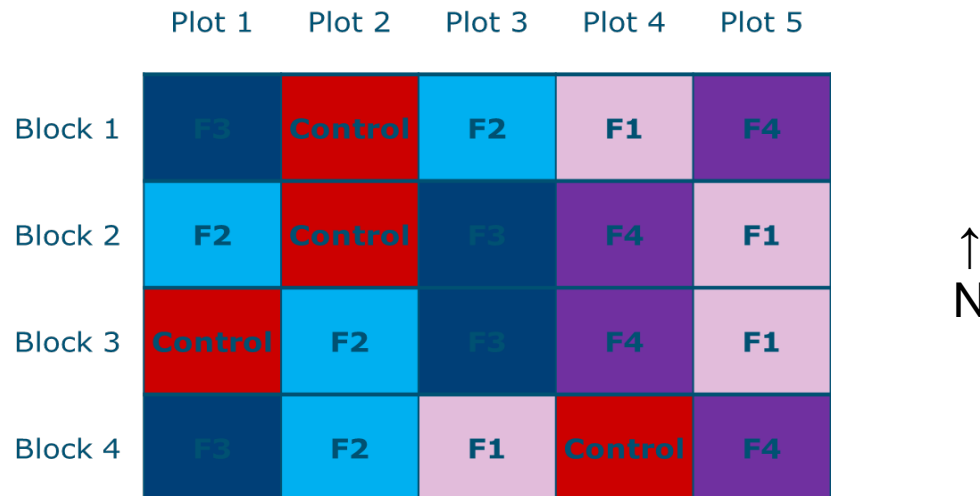
Avantages:

- Simplicité
- Risque: il n'y a pas de réplification au sein d'un bloc. Si on perd une UE (dans un bloc), le plan n'est plus balancé (estimation des moyennes des blocs biaisée).
- Assurance: une bonne partie des problèmes qui arrivent en cours d'expérience affectent un groupe de parcelle contigues – donc vraisemblablement un bloc, que l'on pourra sortir de l'expérience sans compliquer l'analyse.

Exemple

Une expérience au champ investigant les effets de quatre fongicides (F1, F2, F3, F4) sur le rendement de la pomme de terre, en comparaison avec un contrôle (sans traitement).

Le plan utilisé est un RCBD avec quatre blocs de cinq parcelles chacun, avec un total de 20 parcelles. On soupçonne l'existence d'un gradient naturel entre les parcelles, orienté Nord-Sud. Les blocs sont donc orientés Est-Ouest



Le jeu de données

ID	Block	Plot	Type	Fungicide	Yield
1	1	1	Treated	F3	642
2	1	2	Control	Control	377
3	1	3	Treated	F2	633
4	1	4	Treated	F1	527
5	1	5	Treated	F4	623
6	2	1	Treated	F2	600
7	2	2	Control	Control	408
8	2	3	Treated	F3	708
9	2	4	Treated	F4	550
10	2	5	Treated	F1	604
11	3	1	Control	Control	500
12	3	2	Treated	F2	650
13	3	3	Treated	F3	662
14	3	4	Treated	F4	562
15	3	5	Treated	F1	606
16	4	1	Treated	F3	504
17	4	2	Treated	F2	567
18	4	3	Treated	F1	533
19	4	4	Control	Control	333
20	4	5	Treated	F4	667

	Plot 1	Plot 2	Plot 3	Plot 4	Plot 5
Block 1	F3 642	Control 377	F2 633	F1 527	F4 623
Block 2	F2 600	Control 408	F3 708	F4 550	F1 604
Block 3	Control 500	F2 650	F3 662	F4 562	F1 606
Block 4	F3 504	F2 567	F1 533	Control 333	F4 667

Dans le cadre d'un RCBD, on s'intéressera aux effets du traitement et du bloc. On veillera toutefois à ne pas oublier la colonne Plot, sans laquelle on ne pourra pas, a posteriori, retrouver comment les traitements ont été assignés au sein d'un bloc.

Equation du modèle

Le modèle utilisé pour analyser les données de l'expérience est le suivant:

$$y_{ij} = \mu + b_i + \alpha_j + e_{ij}$$

avec :

y_{ij} : le rendement de la parcelle du bloc i qui a reçu le traitement j

μ : la moyenne générale de la population des échantillons

b_i : l'effet du bloc i , exprimé en différence de la moyenne générale

α_j : l'effet du traitement j , exprimé en différence de la moyenne générale

e_{ij} : la déviation (erreur résiduelle) entre la valeur observée y_{ij} et la valeur attendue ($\mu + b_i + \alpha_j$) de la parcelle du bloc i qui a reçu le traitement j

$$i = 1 \dots n$$

$$j = 1 \dots t$$

$$N = n.t$$

$$e_{ij} \sim iN(0, \sigma^2)$$

Si on décide de traiter bloc comme aléatoire, on ajoutera que $b_i \sim iN(0, \sigma_b^2)$

Décomposition de la moyenne générale

```
mean(data$Yield[data$Block == 1])
```

	Plot 1	Plot 2	Plot 3	Plot 4	Plot 5
Block 1	F3 642	Control 377	F2 633	F1 527	F4 623
Block 2	F2 600	Control 408	F3 708	F4 550	F1 604
Block 3	Control 500	F2 650	F3 662	F4 562	F1 606
Block 4	F3 504	F2 567	F1 533	Control 333	F4 667

$$\bar{y}_{1.} = \frac{642 + 377 + 633 + 527 + 623}{5} = 560.4$$

Moyenne d'un bloc

Comme chaque bloc est complet, les différences entre les moyennes de 2 blocs ne représentent que des effets de blocs.

Ce ne serait pas le cas s'il y avait une interaction bloc*trt ou si le plan n'était pas balancé (blocs incomplets)

$$\bar{y}_{.1} = \frac{377 + 408 + 500 + 333}{4} = 404.5$$

Moyenne d'un traitement

Comme chaque traitement est présent dans les 4 blocs, les différences entre les moyennes de 2 traitements ne représentent que les effets des traitements.

Ce ne serait pas le cas s'il y avait une interaction bloc*trt ou si le plan n'était pas balancé (blocs incomplets)

Construction des termes b_i , α_j et e_{ij} du modèle

$$\begin{aligned}
 e_{11} &= y_{11} - \hat{b}_1 - \hat{\alpha}_1 - \bar{\mu} \\
 &= y_{11} - (\bar{y}_{1.} - \bar{y}) - (\bar{y}_{.1} - \bar{y}) - \bar{y} \\
 &= y_{11} - \bar{y}_{1.} - \bar{y}_{.1} + \bar{y} \leftarrow \\
 &= 377 + 2.4 + 158.3 - 562.8 = -25.1
 \end{aligned}$$

Les plus perspicaces reconnaîtront ici le terme d'interaction de l'ANOVA2

	Plot 1	Plot 2	Plot 3	Plot 4	Plot 5
Block 1	F3 642	Control 377	F2 633	F1 527	F4 623
Block 2	F2 600	Control 408	F3 708	F4 550	F1 604
Block 3	Control 500	F2 650	F3 662	F4 562	F1 606
Block 4	F3 504	F2 567	F1 533	Control 333	F4 667

$$\hat{b}_1 = \bar{y}_{1.} - \bar{y} = 560.4 - 562.8 = -2.4$$

$$\hat{\alpha}_1 = \bar{y}_{.1} - \bar{y} = 404.5 - 562.8 = -158.3$$

On a donc bien:

$$\begin{aligned}
 y_{11} &= \bar{\mu} + \hat{b}_1 + \hat{\alpha}_1 + e_{11} \\
 377 &= 562.8 - 2.4 - 158.3 - 25.1
 \end{aligned}$$

Bizarre, ce terme d'erreur en absence de réplication ?

Le modèle d'ANOVA2 vu en Biométrie considère qu'il y a plusieurs réplications pour chaque combinaison des facteurs (ici pour chaque traitement au sein de chaque bloc). Il aurait été construit comme suit:

$$y_{ijk} = \mu + b_i + \alpha_j + \gamma_{ij} + e_{ijk}$$

L'absence de réplication au sein des blocs, propre au RCBD, a pour conséquence que $k = 1$, d'où la suppression de l'indice k et du terme d'erreur.

L'astuce du RCBD consiste à renommer γ_{ij} en e_{ij} . **Keksekseka ?**

Si on ne peut calculer les e_{ijk} , on peut quand même estimer leur variance σ^2 !!!

1. On a tout ce qu'il faut pour estimer γ_{ij}
2. On peut donc estimer $SSAB$ et $CMAB$
3. On se souvient que $E(MSAB) = \sigma^2 + n \sum \sum \gamma_{ij} / (n - 1)(t - 1)$ Biométrie, Ch 03, p. 12
4. Si on fait l'hypothèse qu'il n'y a pas d'interaction entre blocs et traitements, on peut alors en déduire que $E(MSAB) = \sigma^2$

Et le tour est joué, puisqu'on a notre estimateur de σ^2

Particularités de ce modèle

Le modèle du RCBD ne peut donc être appliqué qu'en absence d'interaction entre blocs et traitements!!!

Pour vérifier qu'il n'y a pas d'interaction, retour au cours de biométrie, chapitre 3... Il faudrait plusieurs réplifications au sein de chaque bloc pour estimer σ^2 et en déduire les effets d'interaction.

Dans la réalité, notre connaissance du domaine nous permet de décider si on peut (ou pas) faire l'hypothèse qu'il n'y a pas d'interaction et appliquer le plan du RCBD.

En absence de connaissance, mieux vaut se donner les moyens d'estimer la variance résiduelle de manière propre, au risque de sauter sans filet !

On trouve parfois le modèle RCBD décrit de la manière suivante:

- un facteur explicatif Fongicide, associé aux effets des traitements α_j
- un facteur Bloc, associé aux déviations des blocs b_i
- un facteur Unit.Block, associé aux déviations e_{ij} et qui signifie que les UE sont associées à un bloc

Modélisation GLM avec un facteur bloc fixe

Le tableau de décomposition de la variance (approche GLM) donne:

$$SST = SSB + SSA + SSE$$

$$\sum_{i=1}^n \sum_{j=1}^t (y_{ij} - \bar{y})^2 = \sum_{i=1}^n \sum_{j=1}^t (\bar{y}_{i.} - \bar{y})^2 + \sum_{i=1}^n \sum_{j=1}^t (\bar{y}_{.j} - \bar{y})^2 + \sum_{i=1}^n \sum_{j=1}^t (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2$$

avec les degrés de liberté suivants:

$$N - 1 = (n - 1) + (t - 1) + (n - 1)(t - 1)$$

On reconnaît ici que le terme SSE correspond au terme d'interaction de l'ANOVA2, ainsi que son nombre de degrés de liberté.

Les espérances des carrés moyens sont:

$$E(MSA) = \sigma^2 + \frac{n}{t-1} \sum_{j=1}^t \alpha_j^2 \quad E(MSB) = \sigma^2 + \frac{t}{n-1} \sum_{i=1}^n b_i^2 \quad E(MSE) = \sigma^2$$

Au départ desquelles on peut construire les tests sur les facteurs traitement et bloc, ainsi que les IC aux moyennes de traitement:

$$\left(\hat{\mu}_j - t^{\alpha/2} \sqrt{\frac{s^2}{n}}, \hat{\mu}_j + t^{\alpha/2} \sqrt{\frac{s^2}{n}} \right), \text{ et } SED = \sqrt{\frac{2s^2}{n}}$$

Modélisation R

```
require(lme4)
require(emmeans)
require(lmerTest)

data <- read.table(file="C:/Users/draye/Home/Cours/BRAI2222 - Mixtes et plans/R/POTATO.DAT", header = TRUE)
data$Block <- as.factor(data$Block)
data$Fungicide <- as.factor(data$Fungicide)

## Approche Modele GLM

mod1 <- lm(Yield ~ Fungicide + Block, data)
summary(mod1)

# test F sur les facteurs fixes: traitement et Block
anova(mod1)

# Estimation des moyennes par Traitement (ici SE est l'erreur std sur une moyenne (SEM))
emml <- emmeans(mod1, ~Fungicide)
emml

# Estimation d'une différence entre deux traitements (ici SE est l'erreur std sur une différence (SED))
# Les éléments de c() correspondent aux moyennes des 5 trts, voir sortie de emmeans())
contrast(emml, method = list("Difference bw ctl and F1 " = c(1, -1, 0, 0, 0)))
```

Modélisation R

```
> summary(mod1)
```

Call:

```
lm(formula = Yield ~ Fungicide + Block, data = data)
```

Residuals:

```
    Min       1Q   Median       3Q      Max
-83.00 -26.20   2.05   23.40  108.50
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	402.10	37.33	10.773	1.59e-07	***
FungicideF1	163.00	41.73	3.906	0.002088	**
FungicideF2	208.00	41.73	4.984	0.000318	***
FungicideF3	224.50	41.73	5.380	0.000165	***
FungicideF4	196.00	41.73	4.697	0.000517	***
Block2	13.60	37.33	0.364	0.721930	
Block3	35.60	37.33	0.954	0.359023	
Block4	-39.60	37.33	-1.061	0.309611	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.02 on 12 degrees of freedom

Multiple R-squared: 0.7803, Adjusted R-squared: 0.6521

F-statistic: 6.087 on 7 and 12 DF, p-value: 0.003339

Modélisation R

```
> anova(mod1)
```

Analysis of Variance Table

Response: Yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fungicide	4	133419	33355	9.5763	0.001026 **
Block	3	14987	4996	1.4343	0.281402
Residuals	12	41797	3483		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> emm1 <- emmeans(mod1, ~Fungicide)
```

```
> emm1
```

Fungicide	emmean	SE	df	lower.CL	upper.CL
Control	404	29.5	12	340	469
F1	568	29.5	12	503	632
F2	612	29.5	12	548	677
F3	629	29.5	12	565	693
F4	600	29.5	12	536	665

Results are averaged over the levels of: Block

Confidence level used: 0.95

```
> contrast(emm1, method = list("Difference bw ctl and F1 " = c(1, -1, 0, 0, 0)))
```

contrast	estimate	SE	df	t.ratio	p.value
Difference bw ctl and F1	-163	41.7	12	-3.906	0.0021

Results are averaged over the levels of: Block

Modélisation R avec terme d'interaction

```
> mod3 <- lm(Yield ~ Fungicide*Block, data)
> summary(mod3)
```

```
Residuals:
ALL 20 residuals are 0: no residual degrees of freedom!
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	377	NA	NA	NA
FungicideF1	150	NA	NA	NA
...				
Block2	31	NA	NA	NA
...				
FungicideF1:Block2	46	NA	NA	NA
...				

```
> anova(mod3)
```

Analysis of Variance Table

Response: Yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fungicide	4	133419	33355		
Block	3	14987	4996		
Fungicide:Block	12	41797	3483		
Residuals	0	0			

Warning message:

In anova.lm(mod3) :

les tests F d'ANOVA sur un ajustement pratiquement parfait ne sont pas fiables

Fungicide:Block utilise les degrés de liberté qui étaient affectés à la résiduelle dans le modèle sans interaction. Idem pour les sommes des carrés. Ce modèle ne peut pas calculer s^2 et les sorties n'ont pas de SE si de test F

Modélisation mixte avec un facteur bloc aléatoire

Si on décide de traiter le facteur bloc comme aléatoire, on se retrouve avec un modèle mixte (cf Ch 3), dont on estime les paramètres par la méthode REML.

On peut alors réaliser l'inférence sur les effets des traitements avec la méthode générale vue au chapitre 3.

Modélisation R

```
## Approche Modele Mixte

mod2 <- lmer(Yield ~ Fungicide + (1|Block), data)
summary(mod2)

# test F sur le facteur fixe traitement
anova(mod2)

# test LRT sur le facteur aleatoire block
ranova(mod2)

# Estimation des moyennes par Traitement (ici SE est l'erreur std sur une moyenne (SEM))
emm2 <- emmeans(mod2, ~ Fungicide)
emm2

# Estimation d'une différence entre deux traitements (ici SE est l'erreur std sur une différence (SED))
# Methode 1: lmerTest (ici les éléments de c() correspondent aux elements de b, voir sortie de summary())
lmerTest::contest(mod2, c(0, 1, 0, 0, 0), joint = FALSE)
# Methode 2: contrast (ici les éléments de c() correspondent aux moyennes des 5 trts, voir sortie de
emmeans())
contrast(emm2, method = list("Difference bw ctl and F1 " = c(1, -1, 0, 0, 0)))

# Connaissant SED, on peut trouver LSD et estimer (a posteriori) la puissance du test F
# ou du contraste sur la différence entre les deux moyennes
```

Modélisation R

```
> mod2 <- lmer(Yield ~ Fungicide + (1|Block), data)
> summary(mod2)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: Yield ~ Fungicide + (1 | Block)
Data: data
```

```
REML criterion at convergence: 172.9
```

```
Scaled residuals:
```

```
   Min       1Q   Median       3Q      Max
-1.9025 -0.5851  0.1172  0.4693  1.4478
```

```
Random effects:
```

Groups	Name	Variance	Std.Dev.
Block	(Intercept)	302.5	17.39
	Residual	3483.1	59.02

Number of obs: 20, groups: Block, 4

$\hat{\sigma}_b^2$

$\hat{\sigma}^2$

Estimation de la
moyenne du contrôle
(SE correspond à *SEM*)

```
Fixed effects:
```

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	404.50	30.76	14.63	13.149	1.68e-09 **
FungicideF1	163.00	41.73	12.00	3.906	0.002088 **
FungicideF2	208.00	41.73	12.00	4.984	0.000318 ***
FungicideF3	224.50	41.73	12.00	5.380	0.000165 ***
FungicideF4	196.00	41.73	12.00	4.697	0.000517 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimation de la
différence entre F2 et
le contrôle (SE
correspond à *SED*)

```
Correlation of Fixed Effects:
```

	(Intr)	FngcF1	FngcF2	FngcF3
FungicideF1	-0.678			
FungicideF2	-0.678	0.500		
FungicideF3	-0.678	0.500	0.500	
FungicideF4	-0.678	0.500	0.500	0.500

Modélisation R

```
> anova(mod2)
```

Type III Analysis of Variance Table with Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Fongicide	133419	33355	4	12	9.5763	0.001026 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Test F sur l'effet du
facteur Fongicide

```
> ranova(mod2)
```

ANOVA-like table for random-effects: Single term deletions

Model:

Yield ~ Fongicide + (1 | Block)

	npar	logLik	AIC	LRT	Df	Pr(>Chisq)
<none>	7	-86.458	186.92			
(1 Block)	6	-86.542	185.08	0.16736	1	0.6825

Test LRT sur l'effet du
facteur Bloc

```
> emm2 <- emmeans(mod2, ~ Fongicide)
```

```
> emm2
```

Fongicide	emmean	SE	df	lower.CL	upper.CL
Control	404	30.8	14.6	339	470
F1	568	30.8	14.6	502	633
F2	612	30.8	14.6	547	678
F3	629	30.8	14.6	563	695
F4	600	30.8	14.6	535	666

SEM

Degrees-of-freedom method: kenward-roger

Confidence level used: 0.95

Modélisation R

```
> lmerTest::contest(mod2, c(0, 1, 0, 0, 0), joint = FALSE)
```

	Estimate	Std. Error	df	t value	lower	upper	Pr(> t)
1	163	41.73168	12	3.905905	72.07447	253.9255	0.002087653

```
> contrast(emm2, method = list("Difference bw ctl and F1 " = c(-1, 1, 0, 0, 0)))
```

contrast		estimate	SE	df	t.ratio	p.value
Difference bw ctl and F1		163	41.7	12	3.906	0.0021

SED

Création d'un RCBD avec DiGGer

```
require(DiGGer)
```

```
# Création d'un design
```

```
d <- corDiGGer(  
  numberOfTreatments = 5,  
  rowsInDesign = 4,  
  columnsInDesign = 5,  
  rowsInReplicate = 1,  
  columnsInReplicate = 5,  
  spatial = FALSE,  
  blockSequence = list(c(1,5)),  
  rngSeeds = c(111, 222))
```

```
# Vérification du nombre de répétitions de chaque traitement dans chaque bloc
```

```
# Les arguments 1, 5 précisent la taille du bloc
```

```
desTab(getDesign(d), 1, 5)
```

```
      B1 B2 B3 B4  
freq_1 5  5  5  5
```

← 5 traitements apparaissent avec une fréquence de 1 dans chacun des 4 blocs

```
# Affiche le nombre de fois que chaque paire de traitement apparaît dans un même bloc
```

```
getConcurrence(d, 1, 1, 1)
```

```
B RANDOM BLOCK: 1 Row by 5 Columns, VarianceRatio 1.000000
```

```
Treatment-Block
```

```
Number of Concurrences      4  
Number of Treatment Pairs 10
```

← chacune des 10 paires de traitements apparaît 4 fois dans un même bloc (comme il y a 4 blocs, cela signifie qu'elles apparaissent une fois dans chaque bloc)

Création d'un RCBD avec DiGGer

```
# Représentation graphique du design
plot(d, trts=1, col=2, new=TRUE, label=FALSE, rlab=TRUE, clab=FALSE, rstr="Block", cstr="", chtdiv=3)
plot(d, trts=2, col=3, new=FALSE, label=FALSE, rlab=FALSE, clab=FALSE)
plot(d, trts=3, col=3, new=FALSE, label=FALSE, rlab=FALSE, clab=FALSE)
plot(d, trts=4, col=3, new=FALSE, label=FALSE, rlab=FALSE, clab=FALSE)
plot(d, trts=5, col=3, new=FALSE, label=FALSE, rlab=FALSE, clab=FALSE)
desPlot(getDesign(d), trts = 1:5, new=FALSE, label=TRUE, chtdiv=4, rlab=FALSE, clab=FALSE,
bdef=cbind(1,5),bwd=8))
```

Final Design

1	3	5	2	1	4
2	5	2	4	3	1
3	1	3	5	4	2
4	4	1	3	2	5