



MIXED MODELS

Dr. Vladimir Grosbois
vladimir.grosbois@cirad.fr

CIRAD
UR AGIRs

Examples of fixed and random variables

- Two types of explanatory categorical variables
 - Fixed variables (effects)
 - Random variables (effects)

- Fixed variables
 - Their levels have been specifically selected by the investigator for the purpose of it's study

- Random variables
 - Their levels have been randomly selected among a large population of possible levels. They thus represent a random sampling within a large population of possible levels

Examples of fixed and random variables

➤ Typical fixed effects

- Treatment/Control in an experiment
- Exposed to a risk factor/Not exposed to a risk factor in an epidemiological investigation
- Presence of predators / Absence of predators in an investigation of wild ungulates vigilance behaviour

➤ Typical random effects

- Plot/Block within an experimental field
- Village
- Water point

Parameters of fixed and random variables

➤ The parameters estimated for fixed effects

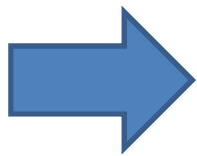
- Are used to depict the mean of the response variable for different combinations of levels of the fixed explanatory variables.
- The estimation of the response variable for each possible combination of levels of the fixed variables are of interest

➤ The parameters estimated for random effects

- Are used to depict the variance of the response variable that remains unexplained by the fixed effects.
- The estimations of the elementary random terms have no interest.
- All we need to know is the extent to which the response variable varies among the levels of the random variables

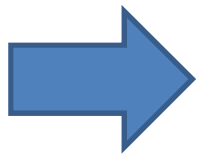
A same variable can be fixed or random depending on the question asked

➤ A national park wants to estimate the density of ungulates at two water points in order to decide where to settle an observation platform



Fixed variable

➤ In an investigation of the vigilance behaviours of ungulates, you select 10 water points where to observe these behaviours



Random variable

Why should we care about random variables: pseudo-replication

○ In statistical models, the individual error terms (departure between the prediction and the observation) need to be independent from each other.

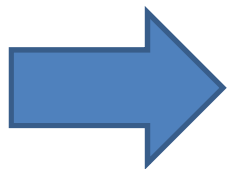
○ You investigate the variation in the productivity of two breeds of cattle in a randomly selected sample of 10 mixed herds. In each herd you select 3 animals of each breed and measure its weekly weight gain in its first year of life.

○ There are a large number of reasons why two animals from a same herds should show more similar weight gains than two animals from different herds:

- ✓ Animals from a same herd graze in the same pastures
- ✓ Animals from a same herd are treated with the same vaccines
- ✓ Animals from a same herd experience similar exposures to the same diseases
- ✓ other reasons which can be unsuspected by the investigator

Why should we care about random variables: pseudo-replication

- Because of these shared, uncontrolled and unmeasured influences, the error terms of the animals from a same herd are expected to be positively correlated. Animals from a same herd are not independent stat. units
- This also means that the number of independent statistical units available to estimate the model parameters (degrees of freedom) are less than the number of individual cattle sampled.



- ✓ Analysing such data without taking into account the herd random effect results in underestimating the uncertainty in the value of the parameters of the fixed effects
- ✓ The standard errors of the coefficients associated with the breed fixed effect will be underestimated

Why should we care about random variables: pseudo-replication

- There are a large number of reasons why two animals from a same herds should show more similar weight gains than two animals from different herds:
 - ✓ Animals from a same herd graze in the same pastures
 - ✓ Animals from a same herd are treated with the same vaccines
 - ✓ Animals from a same herd experience similar exposures to the same diseases
 - ✓ other reasons which can be unsuspected by the investigator
- Note that one way to deal with this pseudo-replication problem is
 - ✓ To measure the variables that underlie the similarity of the response for animals in a same herd (e.g. ask the farmer about vaccines and diseases)
 - ✓ And to incorporate this information in the model in the form of fixed effects.
- However it is likely that there will always be some unmeasured/unsuspected shared conditions within a herd

Why should we care about random variables

Variance components estimations

- An epidemiological investigation of a cattle disease prevalence
 - Random selection of 10 out of 50 districts in the study region
 - Random selection of 10 villages in each selected districts
 - Random selection of 5 herds in each selected village
 - Random selection of 5 animals in each selected village
 - Measure the epidemiological status of each selected animal

- A typical situation where district, village and herd have to be considered as random variables

- A model with these random effects will allow evaluating the level (between districts, between villages, between herds) at which prevalence varies the most

Why not use fixed effects for all the explanatory variables

- Fixed effects are costly: 1 degree of freedom per level
- Random variables have typically many levels
- Random effects are much less costly: less degree of freedom than the number of levels

A linear mixed model for pseudo replication
script: mixed_buffalo.R
data: buffalo faeces.csv

Variation among herds and seasons in the diet quality of buffalos in the W park

```
dat<-read.table("buffalo faeces.csv",sep=" ",header=T)
head(dat,5)
```

Herd	Year	Month	Season	Date	NinMO	ADLinMO	Stress
H1	2007	4	LDS	08/04/2007	2.0	13.4	336.258
H1	2007	4	LDS	08/04/2007	2.0	14.9	312.155
H1	2007	4	LDS	08/04/2007	2.2	15.9	415.852
H1	2007	4	LDS	08/04/2007	2.0	15.0	402.648
H1	2007	4	LDS	08/04/2007	2.1	16.7	591.945

```
summary(dat$Season)      EDS  EWS  LDS  LWS  MDS  MWS
                        158  110  142   88  189  142
```

```
summary(dat$Herd)        H1   H2
                        413  416
```

Sort the season variable in the right order

```
summary(dat$Season)
```

```
EDS EWS LDS LWS MDS MWS  
158 110 142 88 189 142
```

```
dat$Season<-  
factor(dat$Season,levels=c("EWS","MWS","LWS","EDS"  
,"MDS","LDS"),ordered=is.ordered(dat$Season))
```

```
summary(dat$Season)
```

Research question

- We want to depict the variation in diet quality among seasons and herds.
- We want to test whether the seasonal pattern of variation is similar in the two herds

- Dependent variable: $C(\text{Nitrogen})/c(\text{Lignin})$

```
dat$Diet<- (dat$NinMO/dat$ADLinMO)  
attach(dat)
```


- Explanatory variables: Herd and Season

The pseudoreplication problems

- For a given season and herd, many faeces collected on the same day.

	Order	Herd	Year	Month	Season	Date	NinMO	ADLinMO	Stress
1	97	H1	2007	4	LDS	08/04/2007	2.0	13.4	336.258
2	124	H1	2007	4	LDS	08/04/2007	2.0	14.9	312.155
3	125	H1	2007	4	LDS	08/04/2007	2.2	15.9	415.852
4	126	H1	2007	4	LDS	08/04/2007	2.0	15.0	402.648
5	127	H1	2007	4	LDS	08/04/2007	2.1	16.7	591.945

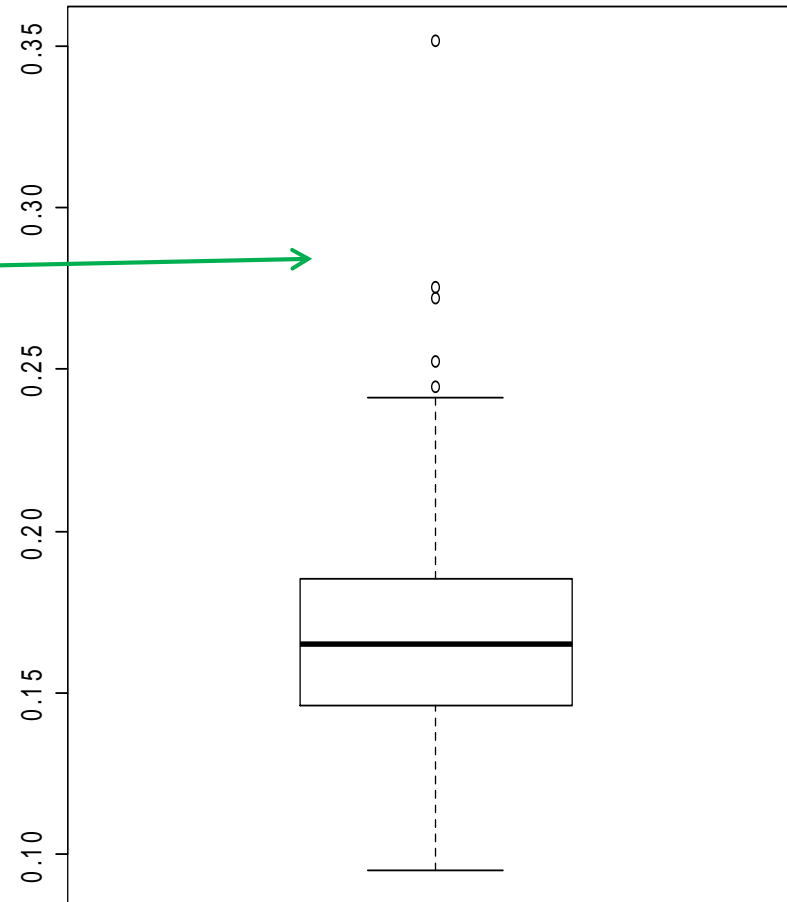
- The faeces collected on a same day are pseudoreplicates
- Same weather conditions
 - Same pastures exploited
 -

 So, a random effect of date has to be incorporated in the model

Models for continuous variables

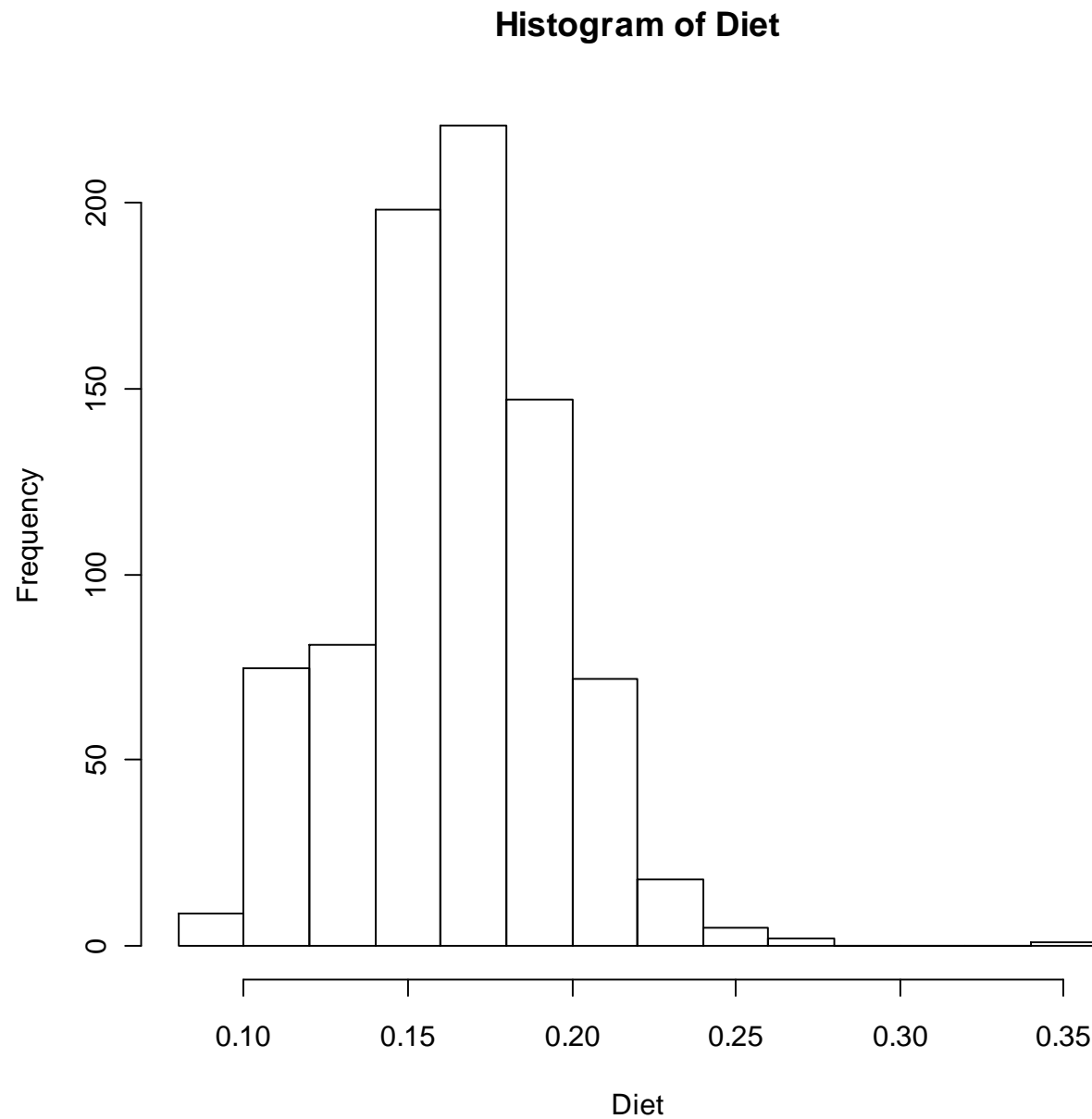
`boxplot(Diet)`

Outliers



Examination of the response variable

`hist(Diet)`



Remove the outliers

➤ An outlier

$$\begin{aligned} &> 3^{\text{rd}}Q + 1.5 * (3^{\text{rd}}Q - 1^{\text{st}}Q) \\ &< 1^{\text{st}}Q - 1.5 * (3^{\text{rd}}Q - 1^{\text{st}}Q). \end{aligned}$$

```
summary(Diet)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.09479 0.14580 0.16500 0.16440 0.18490 0.35190
```

Define the limit above which a data point should be considered as an outlier

```
limout<-0.1849+1.5*(0.1849-0.1458)
```

Remove the outliers

```
datnoout1<-dat[diet<limout,]
```

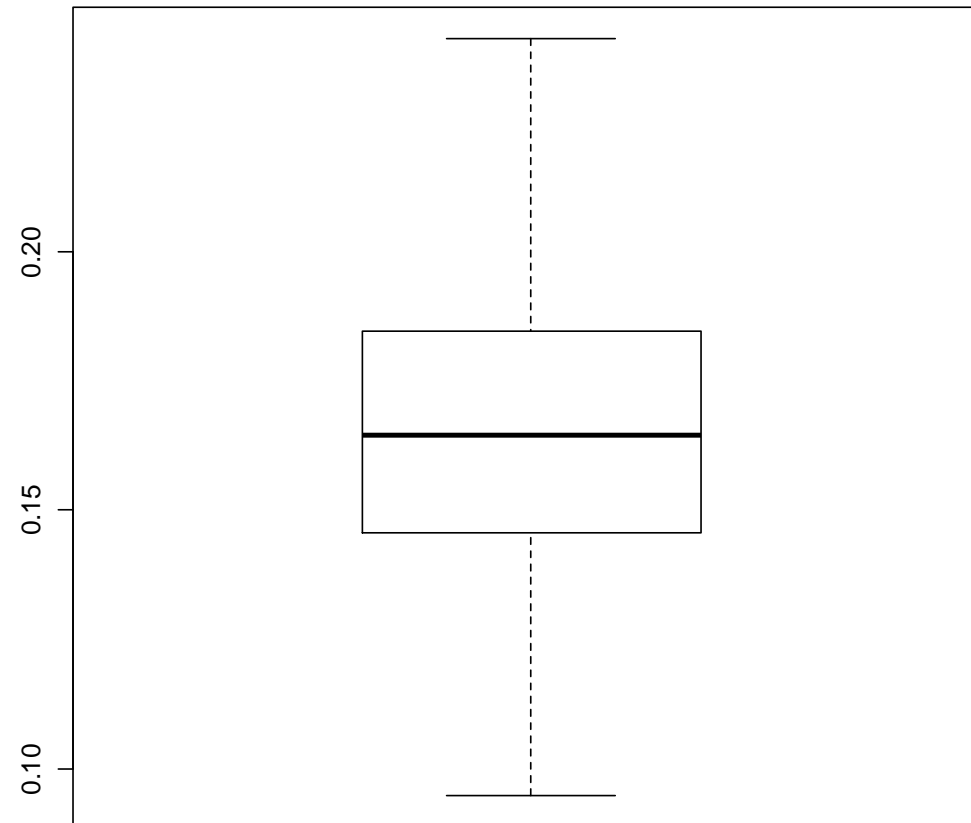
We will now use this new data frame

```
detach(dat)
```

```
attach(datnoout1)
```

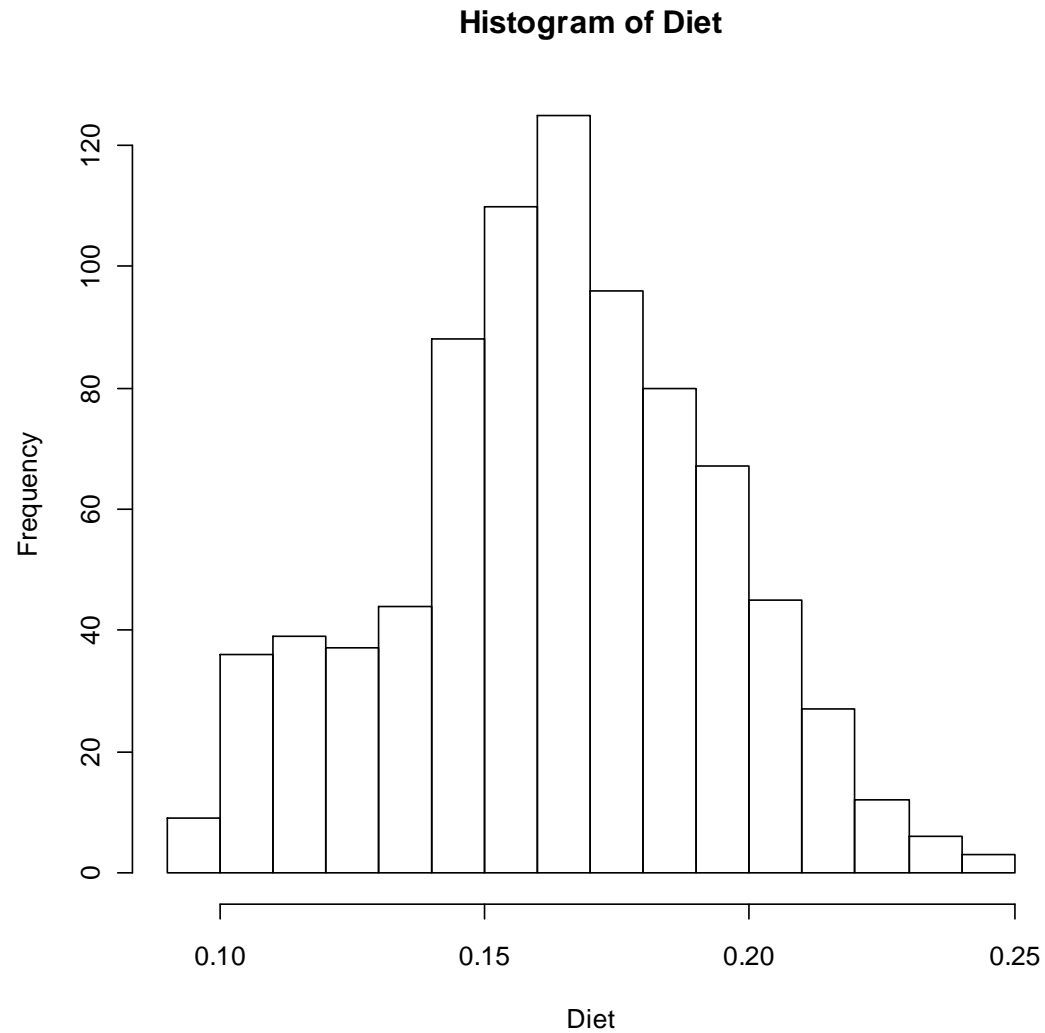
Models for continuous variables

`boxplot(Diet)`



Look at the distribution of Diet without outliers

```
hist(Diet)
```



Distribution of dates across herds and across seasons

table(Date, Herd)

Each herd has been sampled on many different dates
On a few dates, both herds have been sampled

Date	Herd	
	H1	H2
03/11/2008	16	0
04/09/2008	27	0
04/11/2008	0	24
05/02/2008	0	24
05/09/2007	23	0
06/02/2008	23	0
06/06/2007	0	8
06/08/2008	2	21
.....		

table(Date, Season)

Within a season, sampling has occurred
over several distinct dates

Date	EDS	EWS	LDS	LWS	MDS	MWS
03/11/2008	16	0	0	0	0	0
04/09/2008	0	0	0	27	0	0
04/11/2008	24	0	0	0	0	0
05/02/2008	0	0	0	0	24	0
05/09/2007	0	0	0	23	0	0
06/02/2008	0	0	0	0	23	0
06/06/2007	0	8	0	0	0	0
06/08/2008	0	0	0	0	0	23
07/12/2007	19	0	0	0	0	0
.....						

Run the mixed effect model with lme {nlme}

$$Y_{hsdi} = b_0 + b_h + b_s + b_{hs} + \sigma_d + \varepsilon_{hsdi}$$

```
library(nlme)
mixmod1 <- lme(Diet ~ Herd + Season + Herd:Season, random = ~1 | Date)
anova(mixmod1)
```



	numDF	denDF	F-value	p-value
(Intercept)	1	779	4800.368	<.0001
Herd	1	779	9.456	0.0022
Season	5	33	16.582	<.0001
Herd:Season	5	779	0.968	0.4363

The interaction between herd and season is not significant, so the seasonal pattern of variation can be considered as similar in the two herds.

Run the mixed model without the interaction

```
mixmod2<-lme(Diet~Herd+Season,random=~1|Date)  
anova(mixmod2)
```

	numDF	denDF	F-value	p-value
(Intercept)	1	784	4796.032	<.0001
Herd	1	784	9.454	0.0022
Season	5	33	16.566	<.0001

Both Herd and Season have significant effects.
So diet quality varies among seasons and herd.

Parameter estimates

`summary(mixmod2)`

$$Y_{hsdi} = b_0 + b_h + b_s + b_{hs} + \sigma_d + \varepsilon_{hsdi}$$

Linear mixed-effects model fit by REML

Data: NULL

	AIC	BIC	logLik
	-4563.122	-4520.771	2290.561

Random effects:

Formula: ~1 | Date

	(Intercept)	Residual
StdDev:	0.01434703	0.01350327

The output provides only for random effects standard deviation estimates

Fixed effects: Diet ~ Herd + Season

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.17302695	0.005089928	784	33.99399	0.0000
HerdH2	-0.00730380	0.002391195	784	-3.05446	0.0023
SeasonEWS	0.01188493	0.007833630	33	1.51717	0.1387
SeasonLDS	-0.05132769	0.007730441	33	-6.63968	0.0000
SeasonLWS	0.00158589	0.007820213	33	0.20279	0.8405
SeasonMDS	-0.01132728	0.007394936	33	-1.53176	0.1351
SeasonMWS	0.01722919	0.008208952	33	2.09883	0.0436

Look at the random effect

$$Y_{hsdi} = b_0 + b_h + b_s + b_{hs} + \sigma_d + \varepsilon_{hsdi}$$

```
head(random.effects(mixmod2),10)
```

```
03/11/2008 -0.014621850
04/09/2008  0.001512380
04/11/2008 -0.003198053
05/02/2008  0.006011611
05/09/2007  0.027329604
06/02/2008  0.012295166
06/06/2007 -0.001367915
06/08/2008 -0.020618939
07/12/2007 -0.006624506
08/04/2007  0.012464597
```

```
nrow(random.effects(mixmod2)) 39
```

```
nlevels(Date) 39
```

You can get estimates of the individual terms of the random effect

Compare with estimation of a linear model

```
summary(lm(Diet~Herd+Season))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.173459	0.001636	106.005	< 2e-16	***
HerdH2	-0.008369	0.001284	-6.516	1.26e-10	***
SeasonEWS	0.011913	0.002290	5.202	2.50e-07	***
SeasonLDS	-0.051381	0.002128	-24.147	< 2e-16	***
SeasonLWS	0.006527	0.002471	2.642	0.0084	**
SeasonMDS	-0.010313	0.001984	-5.198	2.54e-07	***
SeasonMWS	0.019930	0.002135	9.333	< 2e-16	***

```
summary(mixmod2)
```

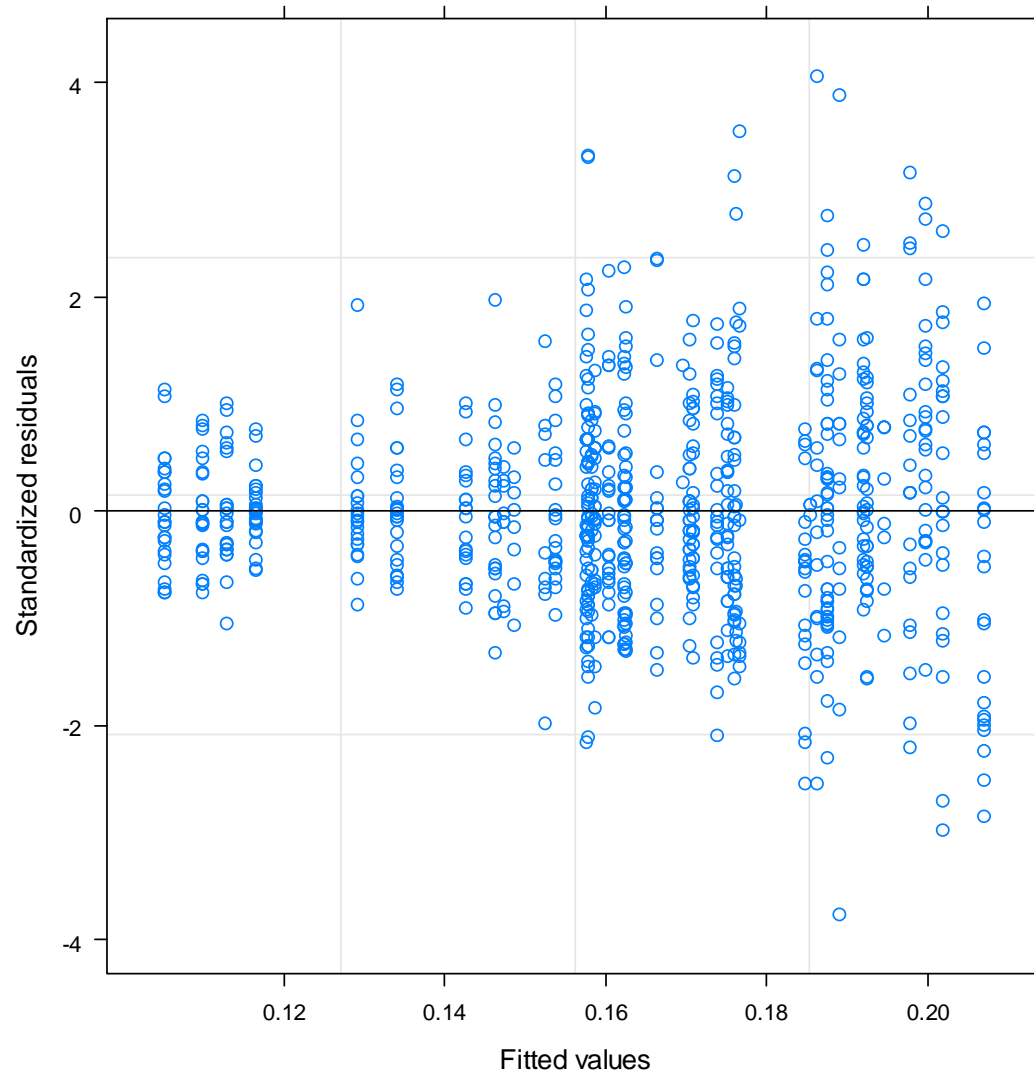
Fixed effects: Diet ~ Herd + Season

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.17302695	0.005089928	784	33.99399	0.0000
HerdH2	-0.00730380	0.002391195	784	-3.05446	0.0023
SeasonEWS	0.01188493	0.007833630	33	1.51717	0.1387
SeasonLDS	-0.05132769	0.007730441	33	-6.63968	0.0000
SeasonLWS	0.00158589	0.007820213	33	0.20279	0.8405
SeasonMDS	-0.01132728	0.007394936	33	-1.53176	0.1351
SeasonMWS	0.01722919	0.008208952	33	2.09883	0.0436

Note the much larger standard errors of the mixed model estimates

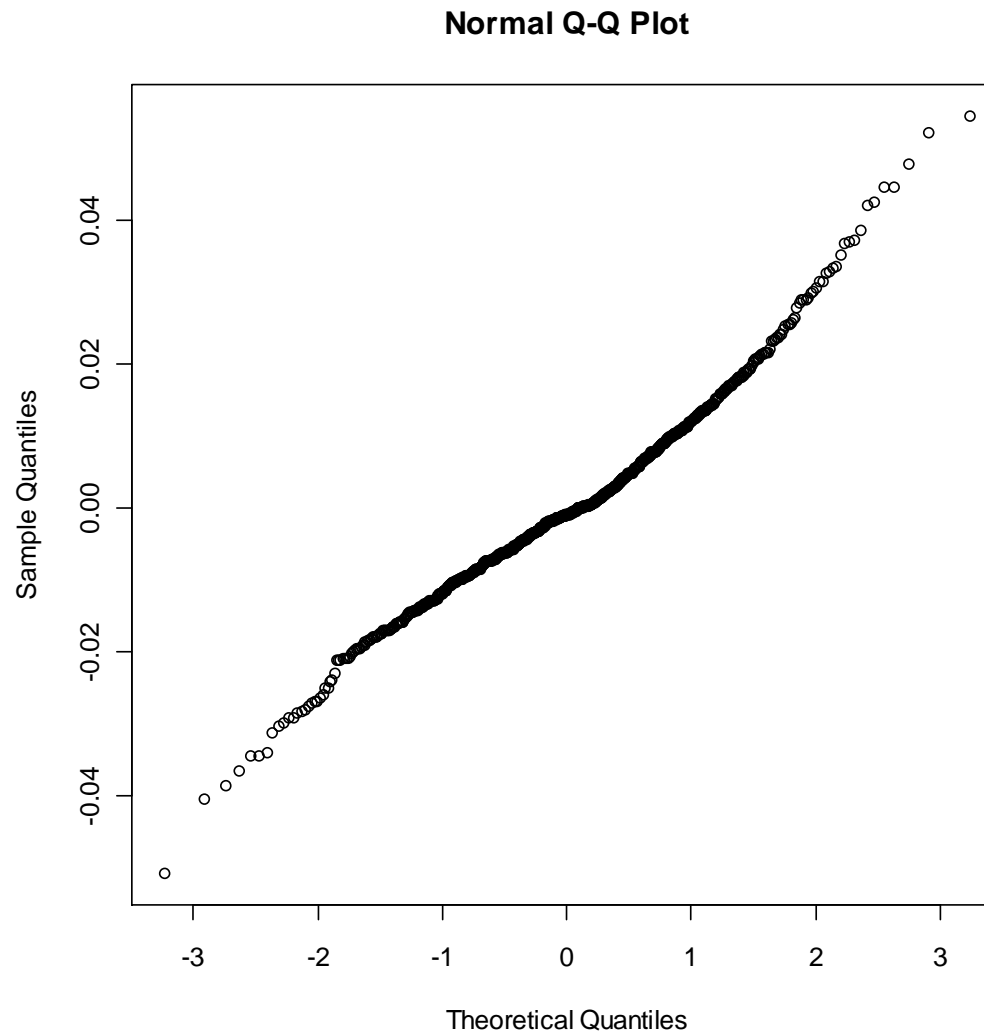
Check conditions of application

`plot(mixmod2)`



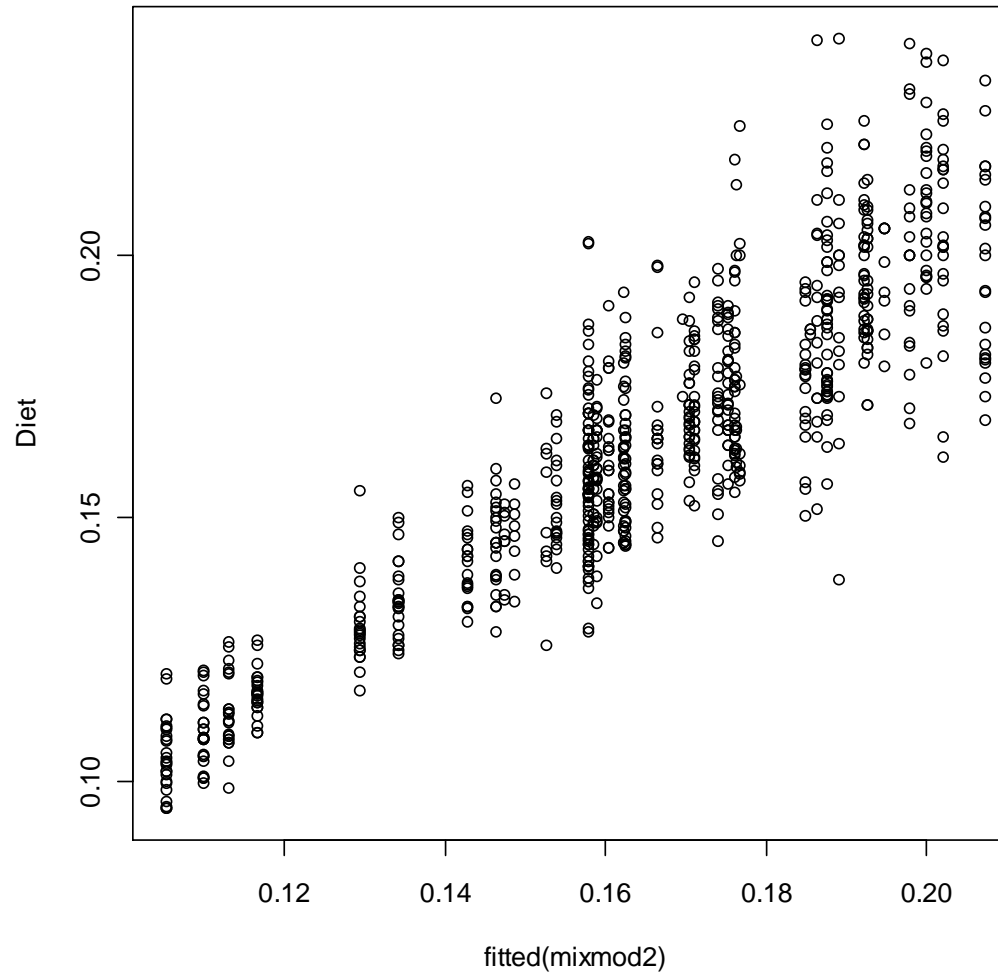
Check conditions of application

```
qqnorm(residuals(mixmod2))
```



Look at how well the model fits the data

```
plot(fitted(mixmod2),Diet)
```



Represent the estimations of the mixed model

Paste the predicted values in a new column of the data frame

```
datnooutl$preddietq<-fitted(mixmod2)
```

Select only the necessary columns

```
predframe<-datnooutl[,c("Date","Herd","Season","preddietq")]
```

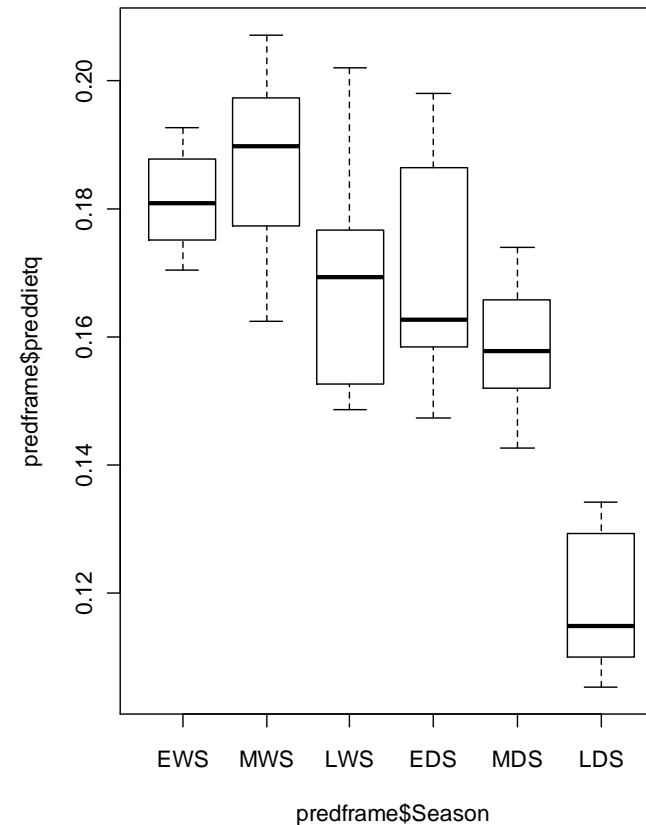
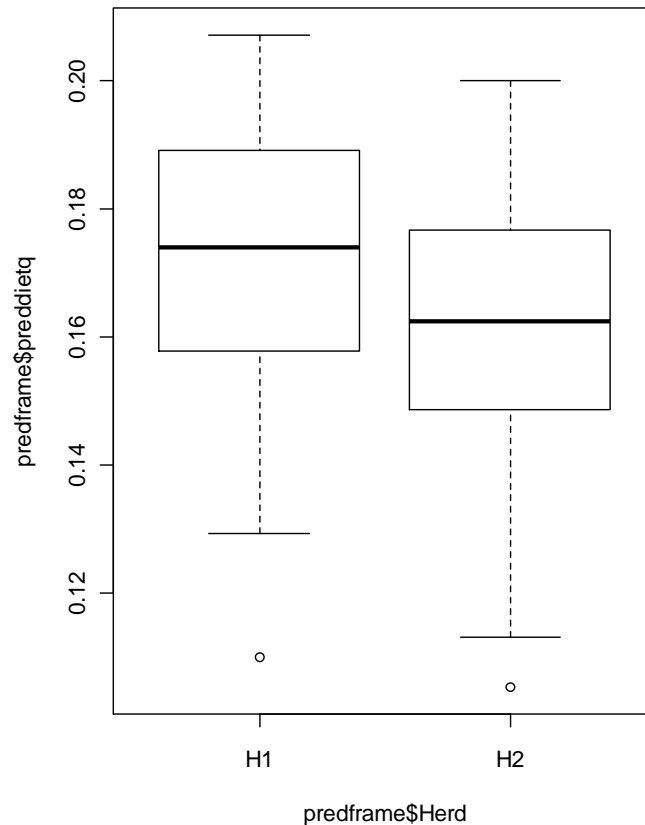
Keep only one line per distinct combination of the fixed and random exp variables

```
predframe<-unique(predframe)
```

Represent the estimations of the mixed model

Paste the predicted values in a new column of the data frame

```
par(mfrow=c(1,3))  
plot(predframe$preddietq~predframe$Herd+predframe$Season)
```



A generalized linear mixed model
for hierarchical random effects
script: hierarchical_random.R
data: prevPPCB.txt

Description of the variation in PPCB prevalence among cattle herds in Mali

```
dat<-read.table("prevPPCB.txt",header=TRUE,sep=";")
summary(dat)
```

Detection of PPCB
antibodies (Y/N)

PO
Min. :0.0000
1st Qu.:0.0000
Median :0.0000
Mean :0.1582
3rd Qu.:0.0000
Max. :1.0000

Successive administrative subdivisions

RE	DI	CO
Mopti:1569	Bandiagara:413	Yeredon Sagnona: 306
Ségou:1421	Douentza :744	Monimpébougou : 278
	Macina :464	Kalasiguida : 168
	Mopti :412	Douentza : 135
	Niono :553	Gandamia : 135
	San :404	Kerena : 134
		(Other) :1834

HE	SI	MA
Dallah3 : 35	Large:2190	Sedentary : 981
Déberé4 : 35	Small: 800	Transhumant:2009
Déberé5 : 35		
Douentza4: 35		
Douentza5: 35		
Douentza6: 35		
(Other) :2780		

Herd Size

Herd management

Herd ID

Preparation/Description of the data

Change the type of PO, it has to be a categorical variable (factor)

```
dat$PO<-as.factor(dat$PO)
```

```
summary(dat$PO)
```

	0	1
	2517	473

```
attach(dat)
```

Look at the potential association between herd size and herd management

		MA	
table(TA,TY)	SI	Sedentary	Transhumant
	Large	319	1871
	Small	662	138

```
chisq.test(table(SI,MA))
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: table(SI, MA)
```

```
X-squared = 1232.625, df = 1, p-value < 2.2e-16
```

Characterization of the association

Change the type of PO, it has to be a categorical variable (factor)

```
chisq.test(table(SI,MA))$observed
```

SI	MA	
	Sedentary	Transhumant
Large	319	1871
Small	662	138

```
chisq.test(table(SI,MA))$expected
```

SI	MA	
	Sedentary	Transhumant
Large	718.5251	1471.4749
Small	262.4749	537.5251

Small herds are more often sedentary

Large herds are more often transhumant



Potential collinearity issues

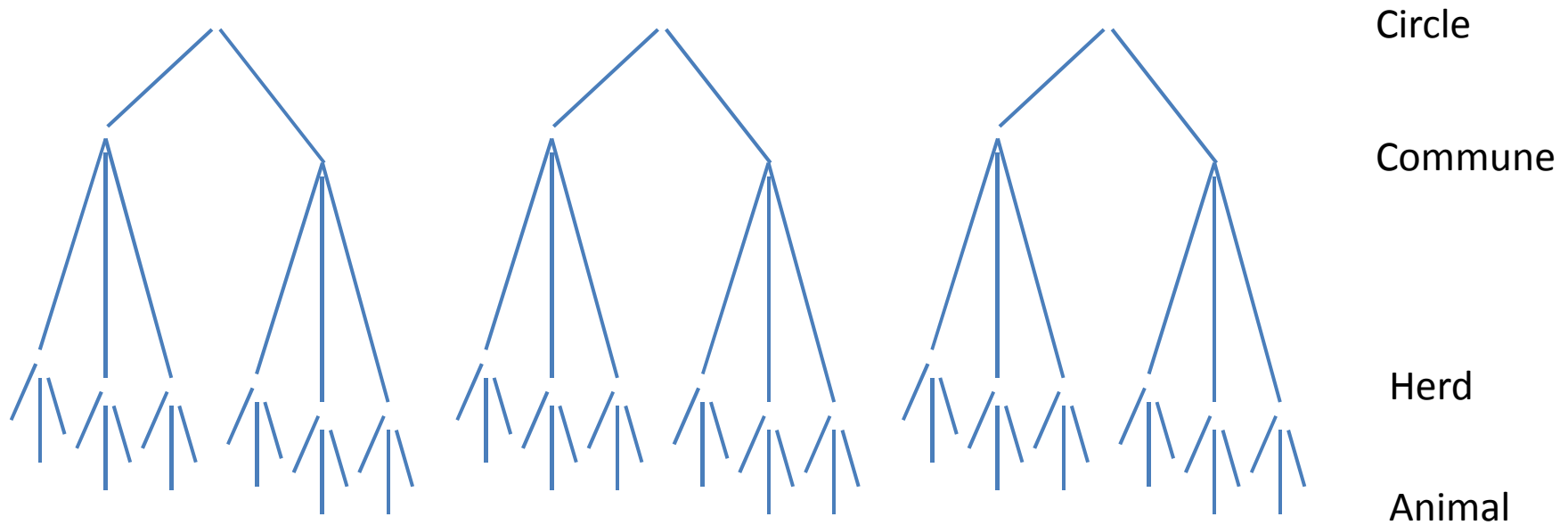
Hierarchical geographic (administrative structure) and structured correlations among individual status

We are analysing the consequences (serological status) of the spatial spread of a contagious disease

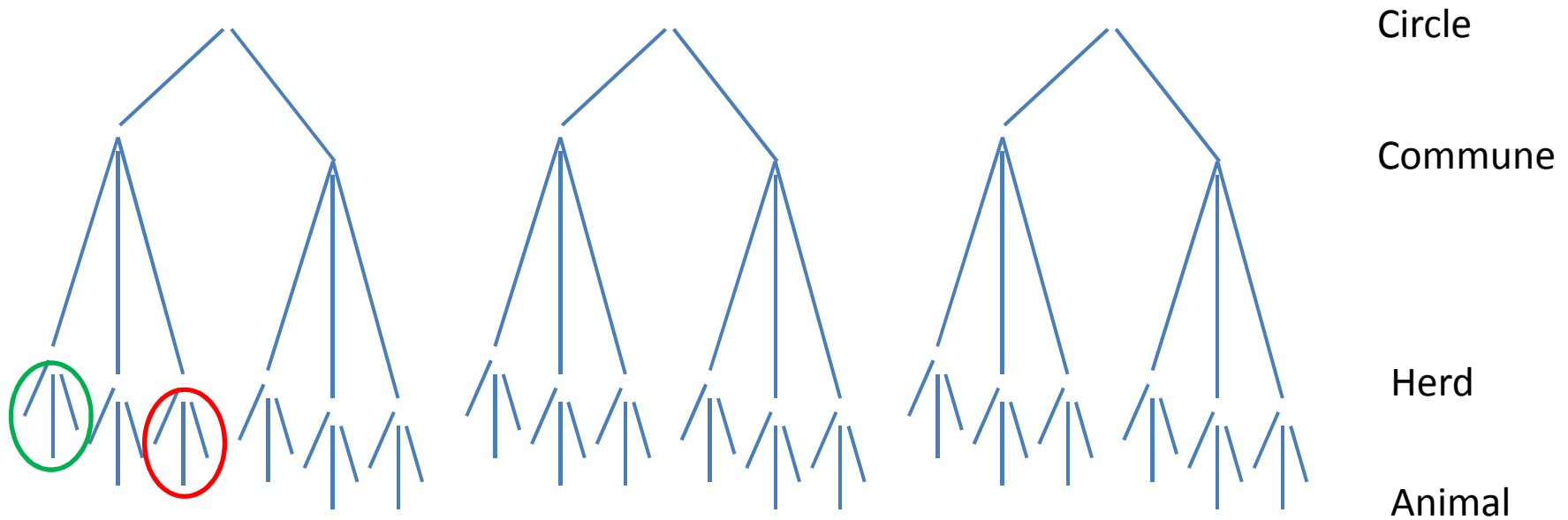
- The individuals that live close together probably do not represent independent information pieces
- The error terms of the model will probably not be independent
- Their interdependencies probably follow a hierarchical structure

Note that the same type of issue arises when with comparative analyses across species when one has to account for the shared evolutionary history (need to account for the phylogenetic structure)

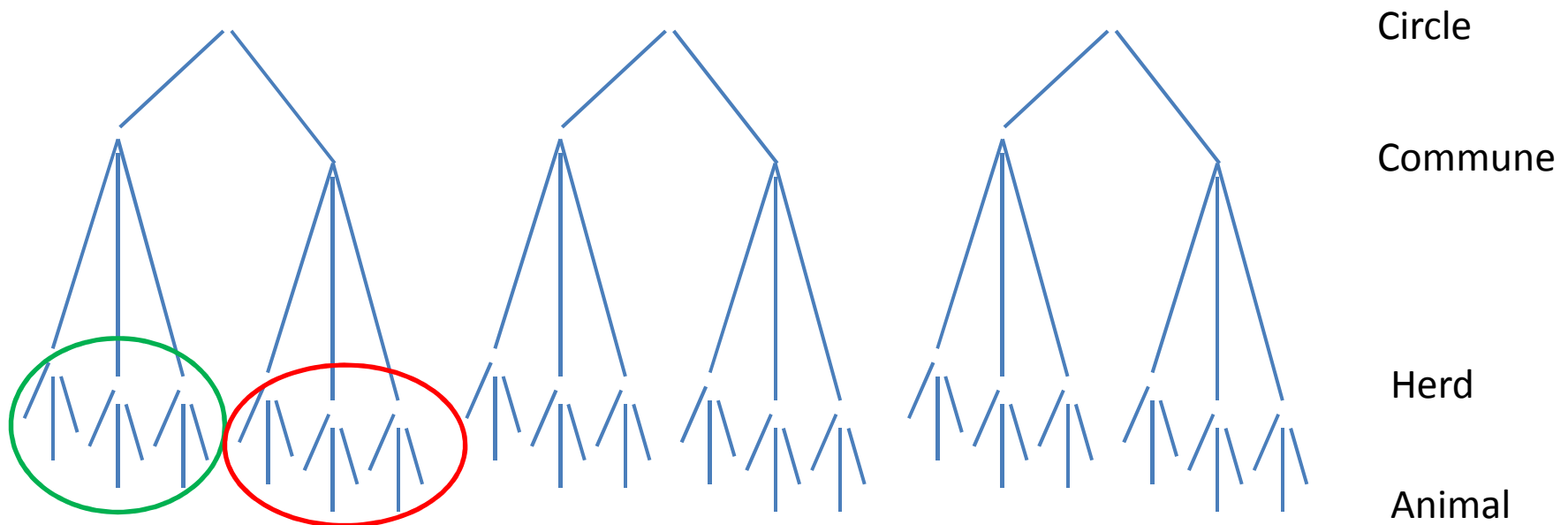
Nested error structures



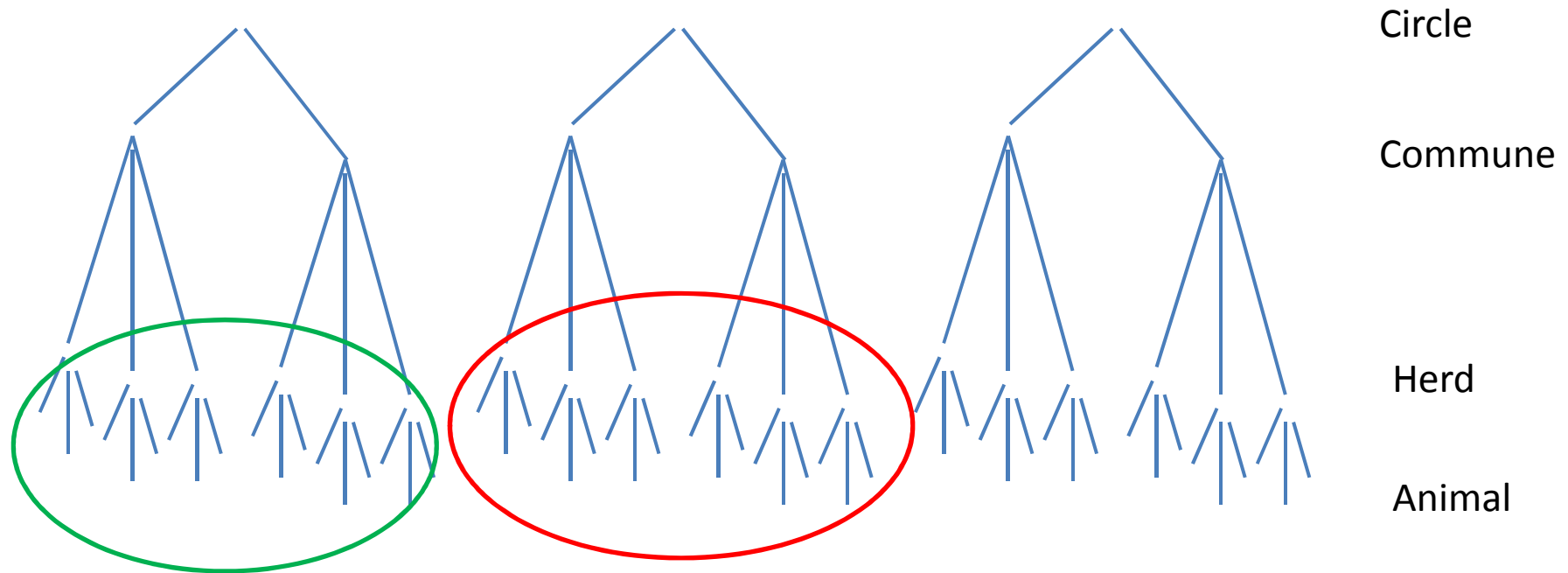
Nested error structures



Nested error structures



Nested error structures



Research question

➤ Assess the influence of herd size and management strategy on the prevalence of CBPP

- Requires to account for the pseudo-replication and nested correlation structure

➤ Assess the scale at which the disease spreads

- Requires to estimate the variance at the different spatial scales

Generalized Mixed Linear Model (GLMM)

Note that here, the dependent variable is 0 or 1, so no need to specify (nbpos,nbneg)

```
library(lme4)
randmod0<-glmer(PO ~ 1 + (1|RE/DI/CO/HE), family = binomial)
```

Generalized linear mixed model fit by the Laplace approximation

Formula: PO ~ 1 + (1 | RE/DI/CO/HE)

AIC BIC logLik deviance

2448 2478 -1219 2438

Random effects:

Groups	Name	Variance	Std.Dev.
HE:(CO:(DI:RE))	(Intercept)	0.46005	0.67827
CO:(DI:RE)	(Intercept)	0.35477	0.59562
DI:RE	(Intercept)	0.00000	0.00000
RE	(Intercept)	0.00000	0.00000

Large inter-herd variance

Large inter-district variance

Number of obs: 2990, groups: HE:(CO:(DI:RE)), 153; CO:(DI:RE), 27; DI:RE, 6; RE, 2

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.878	0.143	-13.13	<2e-16 ***

Estimation of
logit(prevalence)

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

Model selection

```
library(lme4)
randmodmax<-glmer(PO ~ SI +MA + (1|RE/DI/CO/HE), family = binomial)
summary(randmodmax)
```

Generalized linear mixed model fit by the Laplace approximation

Formula: PO ~ SI + MA + (1 | RE/DI/CO/HE)

AIC	BIC	logLik	deviance
2449	2491	-1217	2435

Random effects:

Groups	Name	Variance	Std.Dev.
HE:(CO:(DI:RE))	(Intercept)	0.43908	0.66263
CO:(DI:RE)	(Intercept)	0.34317	0.58581
DI:RE	(Intercept)	0.00000	0.00000
RE	(Intercept)	0.00000	0.00000

Large inter-herd variance

Large inter-district variance

Number of obs: 2990, groups: HE:(CO:(DI:RE)), 153; CO:(DI:RE), 27; DI:RE, 6; RE, 2

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.9181	0.2617	-7.330	2.3e-13 ***
SISmall	0.2477	0.2376	1.042	0.297
MATranshumant	-0.1224	0.2463	-0.497	0.619

Size and Management
seem to have no
influence

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

Correlation of Fixed Effects:

	(Intr)	SISmll
SISmall		-0.753
MATranshmnt	-0.783	0.661

Model selection

```
library(lme4)
randmodmax<-glmer(PO ~ SI +MA + (1|RE/DI/CO/HE), family = binomial)
summary(randmodmax)
```

Generalized linear mixed model fit by the Laplace approximation

Formula: PO ~ SI + MA + (1 | RE/DI/CO/HE)

AIC	BIC	logLik	deviance
2449	2491	-1217	2435

Random effects:

Groups	Name	Variance	Std.Dev.
HE:(CO:(DI:RE))	(Intercept)	0.43908	0.66263
CO:(DI:RE)	(Intercept)	0.34317	0.58581
DI:RE	(Intercept)	0.00000	0.00000
RE	(Intercept)	0.00000	0.00000

Large inter-herd variance

Large inter-district variance

Number of obs: 2990, groups: HE:(CO:(DI:RE)), 153; CO:(DI:RE), 27; DI:RE, 6; RE, 2

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.9181	0.2617	-7.330	2.3e-13 ***
SISmall	0.2477	0.2376	1.042	0.297
MATranshumant	-0.1224	0.2463	-0.497	0.619

Size and Management
seem to have no
influence

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

Correlation of Fixed Effects:

	(Intr)	SISmll
SISmall		-0.753
MATranshmnt	-0.783	0.661

Model selection

```
library(lme4)
randmod1<-glmer(PO ~ MA + (1|RE/DI/CO/HE), family = binomial)
```

Generalized linear mixed model fit by the Laplace approximation

Formula: PO ~ MA + (1 | RE/DI/CO/HE)

AIC	BIC	logLik	deviance
2448	2484	-1218	2436

Random effects:

Groups	Name	Variance	Std.Dev.
HE:(CO:(DI:RE))	(Intercept)	4.4672e-01	6.6837e-01
CO:(DI:RE)	(Intercept)	3.4428e-01	5.8675e-01
DI:RE	(Intercept)	9.3841e-21	9.6871e-11
RE	(Intercept)	0.0000e+00	0.0000e+00

Number of obs: 2990, groups: HE:(CO:(DI:RE)), 153; CO:(DI:RE), 27; DI:RE, 6; RE, 2

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.7155	0.1731	-9.912	<2e-16 ***
MATranshumant	-0.2919	0.1858	-1.571	0.116

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

Correlation of Fixed Effects:

(Intr)

MATranshmnt -0.579

Model selection

```
library(lme4)
randmod1<-glmer(PO ~ MA + (1|RE/DI/CO/HE), family = binomial)
```

Generalized linear mixed model fit by the Laplace approximation

Formula: PO ~ SI + (1 | RE/DI/CO/HE)

AIC	BIC	logLik	deviance
2447	2483	-1217	2435

Random effects:

Groups	Name	Variance	Std.Dev.
HE:(CO:(DI:RE))	(Intercept)	4.4021e-01	6.6348e-01
CO:(DI:RE)	(Intercept)	3.4463e-01	5.8705e-01
DI:RE	(Intercept)	9.0098e-16	3.0016e-08
RE	(Intercept)	0.0000e+00	0.0000e+00

Number of obs: 2990, groups: HE:(CO:(DI:RE)), 153; CO:(DI:RE), 27; DI:RE, 6; RE, 2

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.0203	0.1629	-12.399	<2e-16 ***
SISmall	0.3254	0.1782	1.826	0.0679 .

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

Correlation of Fixed Effects:

	(Intr)
SISmall	-0.503

When Management is not included in the model, size seems to have an influence: colinearity

Random effect examination

$$-\sigma_r + \sigma_d + \sigma_c + \sigma_h$$

```
ranef(randmod2)
```

\$`HE:(CO:(DI:RE))`

	(Intercept)
Bandiagara1:Bandiagara:Bandiagara:Mopti	-0.3651189237
Bandiagara2:Bandiagara:Bandiagara:Mopti	0.1342612757
Bandiagara3:Bandiagara:Bandiagara:Mopti	0.3646181041

..... ● ●

$$\$_{\text{CO}}: (\text{DI}:\text{RE})_{\text{}}$$

	(Intercept)
Bandiagara:Bandiagara:Mopti	0.56340488
Dallah:Douentza:Mopti	-0.16802064
Déberé:Douentza:Mopti	-0.63050851
Djaptodji:Douentza:Mopti	1.03408896

.....●

\$`DI:RE`

	(Intercept)
Bandiagara:Mopti	4.970099e-16
Douentza:Mopti	4.727615e-16
Macina:Ségou	-1.863407e-15
Mopti:Mopti	6.037477e-16

..... ● ●

\$RE

	(Intercept)
Mopti	0
Ségou	0