

Randomization-based analysis of experiments (Tiers over mixed models)

Chris Brien

School of Mathematics and Statistics
University of South Australia



$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\theta}$$

$$\text{var}[\mathbf{Y}] = \sum \sigma_i^2 \mathbf{S}_i + \sigma^2 \mathbf{I}$$



chris.brien@unisa.edu.au

Outline

1. Options for analyzing
2. Randomization-based analysis of two-tiered experiments: why?
3. Randomization-based analysis of multitiered experiments
4. Intertier interactions
5. Lab-phase design: a biodiversity example
6. Some questions
7. Summary

1. Options for analyzing

In the following (model) terms are derived from generalized factors.

- A. Regression or traditional ANOVA involving sequential fitting of terms.
 - Generally obtain SS_q for terms adjusted for previously fitted terms so for nonorthogonal designs:
 - analysis depends on order;
 - lose information as no combination of information.
 - Does not exhibit confounding in experiment.

Options for analyzing (cont'd)

B. Tier-based ANOVA

- Can be efficient, if exploits structure;
- Limited software availability:
 - for two structure formulae: Genstat, R and S-Plus
 - for three structure formulae: Genstat
 - for > 3 structure formulae: none
- For nonorthogonal experiments, must be balanced and, provided orthogonal variance structure, manual combination of information.

C. Mixed model estimation

- Does not exhibit confounding in experiment and models of convenience often required;
- For orthogonal experiments, inefficient and inexact;
- May have convergence problems, e.g. for variance components with small df;
- Often pseudofactors are not required;
- For nonorthogonal experiments, neither OVS nor balance required and automatic combination of information;
- Can fit a broader class of models e.g. unequal correlation and variance.

My preferences

- Use tier-based ANOVA for establishing properties and for analyzing orthogonal experiments.
- Use mixed model estimation for analyzing nonorthogonal experiments and when more complex models required.
- No Genstat means mixed-model estimation for all, but still examine properties using tier-based decomposition table.
- Concentrate on mixed model estimation from here.

Some notation

- The form of the mixed model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\mathbf{u}$ with
 - a random term ($\mathbf{Z}_j\mathbf{u}_j$) for each random, generalized factor;
 - a fixed term ($\mathbf{X}_j\boldsymbol{\theta}_j$) for each fixed, generalized factor.
- Symbolic mixed model (Patterson, 1997, SMfPVE)

Fixed terms | random terms

(A*B | Blocks/Runs)

$A*B = A + B + A\wedge B$

$\text{Blocks/Runs} = \text{Blocks} + \text{Blocks}\wedge\text{Runs}$

- Corresponds to the mixed model:

$$\mathbf{Y} = \mathbf{X}_A\boldsymbol{\theta}_A + \mathbf{X}_B\boldsymbol{\theta}_B + \mathbf{X}_{A\wedge B}\boldsymbol{\theta}_{A\wedge B} + \mathbf{Z}_b\mathbf{u}_b + \mathbf{Z}_{b\wedge R}\mathbf{u}_{b\wedge R}.$$

where the Xs and Zs are indicator variable matrices for the generalized factor in its subscript.

- Could replace $\mathbf{Z}_{b\wedge R}\mathbf{u}_{b\wedge R}$ with a unit error term, \mathbf{e} .

2. Randomization-based analysis of two-tiered experiments: why?

- What is the model for the analysis?
- For RCBD, is it the traditional two-way model:
 - $y_{ik} = \mu + \beta_i + \tau_k + \varepsilon_{ik}$
 - where $\varepsilon_{ik} \sim N(0, \sigma^2)$ is the interaction of the i th block and j th treat?
- or, the model equivalent to a randomization model:
 - $y_{ijk} = \mu + \beta_i + \varepsilon_{ij} + \tau_k$
 - where $\varepsilon_{ij} \sim N(0, \sigma^2)$ is the effect of the j th plot in the i th block?
- Does it matter?
 - I think it does as the two terms are reflecting different mechanisms for generating variability.
 - Samuels, Casella & McCabe (1991, JASA) distinguish between random interactions and inherent contributions, but do not introduce different terms.
 - Randomization model includes term for EU, two-way model does not. (B \wedge T an EU??)
 - The two-way model is a **model of convenience**

Decomposition tables equivalent to models

Traditional

source	df
Blocks	$b - 1$
Treats	$v - 1$
Error	$(b - 1)(v - 1)$

Tier-based

plots tier		treatments tier	
source	df	source	df
Blocks	$b - 1$		
Plots[Blocks]	$b(v - 1)$	Treats	$v - 1$
		Residual	$(b - 1)(v - 1)$

Other reasons for first obtaining a randomization-based model

- Aside from the pertinent-identification-of-sources argument, there are two others?
- Reason 2: ensure all potential sources and no stray sources included in the model, particularly for complex experiments.
 - Formulating tiers encourages the identification of all factors in the experiment (and a set of objective rules produces the model).
- Reason 3: make explicit where the model deviates from the randomization model.
 - e.g. including block-treatment (intertier) interaction

Model varies with randomization

- Randomization depends on sources researcher thinks important:
 - In this sense the randomization depends on the model;
 - Need to restrict randomization to take into account.
- Consequently, the nesting and crossing relations are based on intrinsic relations and those for the design.
 - So, given same intrinsic relations, different relations reflect different restrictions placed on the randomization.
- Sets of Latin squares:
 - $Sets * Rows * Columns$
 - $(Sets / Rows) * Columns$
 - $Sets / (Rows * Columns)$

Obtaining a randomization-based mixed model

(Brien & Bailey, 2006, Fig. 24;
Brien & Demetrio, 2008)

Identify the sets of objects and nominate the observational unit

Stage I:
intratier

Determine the tiers: the factors indexing the sets

(randomization) model

Determine the intratier formulae

Add intertier interactions to form the analysis formulae

Stage II:
intertier

Expand each analysis formula to produce a list of model terms

mixed model

Designate each term as fixed or random and form model

(Augment the model for other terms considered important)

Stage III:
randomization-

Identify totally confounded terms and remove to leave one

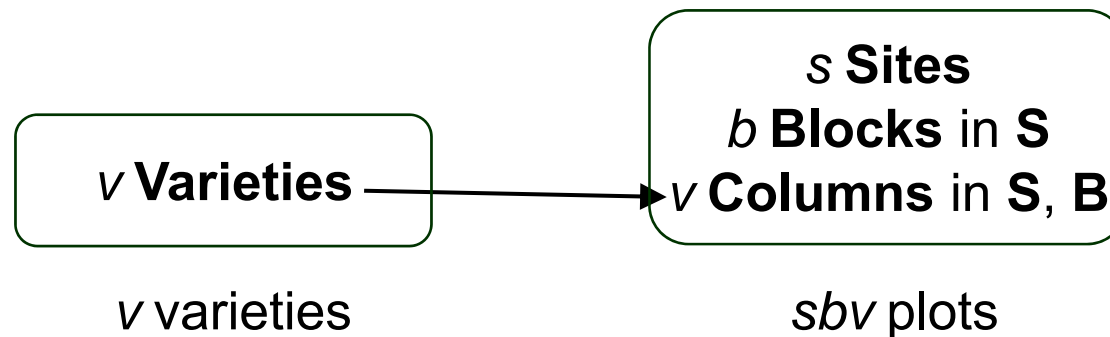
based mixed model

Vary parameterization of terms

- Although focused on normally-distributed data, not restricted to LMM, as could be used to formulate a GLMM (or HGLM?).

Stage I: Randomization for RCBD at several sites

- An experiment to investigate several plant varieties
- At each of s sites
 - RCBD of b Blocks containing v Varieties put on a $b \times v$ rectangle of plots.



■ Structure formulae

- Intratier: s Sites / b Blocks / v Columns
- (Intrinsic: s Sites / (b Blocks * v Columns))
- v Varieties

Terms and model for variety experiment

■ Model terms

- plots: Sites + Sites \wedge Blocks
+ Sites \wedge Blocks \wedge Columns
- varieties: Varieties

■ To obtain a mixed model that corresponds to randomization analysis assume random = unrandomized; fixed = randomized

- random: Sites, Blocks, Columns
- fixed: Varieties

■ From mixed model viewpoint, if Sites are not representative, unrealistic to specify as random

- random: Blocks, Columns
- fixed: Sites, Varieties

■ But, randomization justifies regarding it as random?

Mixed model equivalent to randomization model

- Varieties | Sites + Sites \wedge Blocks + Sites \wedge Blocks \wedge Columns
- Our process has ensured a model including all terms that
 - are intrinsic to the physical set-up **and**
 - have been allowed for in the design's randomization, the minimum set that ought to be included
- Advantage of randomization analysis is robust as only assumption is additivity of randomized (“treatment”) and unrandomized (“block”) factors.
- However, conclusions apply only to observed units.

Stages II & III: Deviating from the randomization model

- Randomization-based mixed model:

- a mixed model that is derived from a randomization model .

- a) *Replacing unrandomized factors with randomized factors
 - b) Incorporating block-treatment (intertier) interactions
 - c) *Removing totally confounded terms
 - d) Augmenting to add intrinsic terms not allowed for in the design
 - e) Varying the parameterization of terms in the randomization model

* Only done for computational convenience and does not reflect the proposed model.

a) Replacing unrandomized with randomized factors

- Piepho et al. (2003, JACR) suggest this in their rule 5:
Varieties | Sites + Sites^Blocks + Sites^Blocks^Columns
⇒ Varieties | Sites + Sites^Blocks + Sites^Blocks^Varieties
- This model of convenience
 - gives correct analysis and saves on factors
 - but obscures identity of type of variation and removes experimental-unit terms:
 - Sites^Blocks^Columns is an EU and reflects inherent variability
 - Sites^Blocks^Varieties required when “block-treatment” interaction
 - How do you randomize Varieties to Sites^Blocks^Varieties?
- Their habitual use leads to loss of randomized layouts as it encourages presentation and reporting in treatment order.
- Indeed, without randomized layout, Columns = Varieties.

b) Incorporating block-treatment (intertier) interactions

■ Tiers

- plots: {Sites, Blocks, Columns}
- varieties: {Varieties}

■ Structure formulae

- 3 Sites / 4 Blocks / 3 Columns
- 4 Varieties * Sites

■ Mixed model (with Sites fixed)

Varieties + Sites + Varieties^Sites

| Sites^Blocks + Sites^Blocks^Columns

- In this case, can estimate the intertier interaction (next slide)

c) Removing totally confounded terms

- Useful to form decomposition tables at this point.

plots tier		varieties tier	
source	df	source	df
Sites	$s - 1$		
Blocks[Sites]	$s(b - 1)$		
Columns[Sites \wedge Blocks]	$sb(v - 1)$	Varieties	$v - 1$
		Varieties#Sites	$(s - 1)(v - 1)$
		Residual	$s(b - 1)(v - 1)$

- As opposed to the RCBD, the intertier interaction can be estimated
- No terms totally confounded by others so model unchanged.

d) Augmenting to include intrinsic terms not allowed for in design

- Perhaps there are intrinsic terms that ought be included in the model, even though not allowed for in randomization.
- Such terms will be unbalanced, but fitting them not a problem with computers
- Difficulty is estimation efficiency determined randomly and could be very inefficient (0)
- Better to determine efficiency by (nonorthogonal) design
- So, provided sufficient df, better to allow for any possibly important intrinsic terms in the design

Augmented model for variety experiment

- In retrospect might suspect Columns is an important source of variability whose overall differences should be isolated.
- In this case, use sources and terms derived from Sites/(Blocks*Columns) instead of Sites/Blocks/Columns
- Will be an unbalanced analysis with information on Varieties and Sites#Varieties lost but a reduced value for the Residual MSq for Blocks#Columns[Sites] — compensates?

e) Varying the parameterization of terms in randomization model

- Designate some unrandomized terms fixed and/or some randomized terms random, as have already seen.
- Specifying trend models for describing the effects of fixed factors.
- Specifying variance matrices that do not conform to compound symmetry pattern (e.g. heterogeneous variances, temporal or spatial correlation).

3. Randomization-based analysis of multitiered experiments (Brien & Bailey, 2006, Sec. 7; Brien & Demetrio, 2008)

Identify the sets of objects and nominate the observational unit

Stage I:
intratier

Determine the tiers: the factors indexing the sets

(randomization) model

Determine the intratier formulae

Add intertier interactions to form the analysis formulae

Stage II:
intertier

Expand each analysis formula to produce a list of model terms

mixed model

Designate each term as fixed or random and form model

(Augment the model for other terms considered important)

Stage III:
randomization-

Identify totally confounded terms and remove to leave one

based mixed model

Vary parameterization of terms

Same method as before

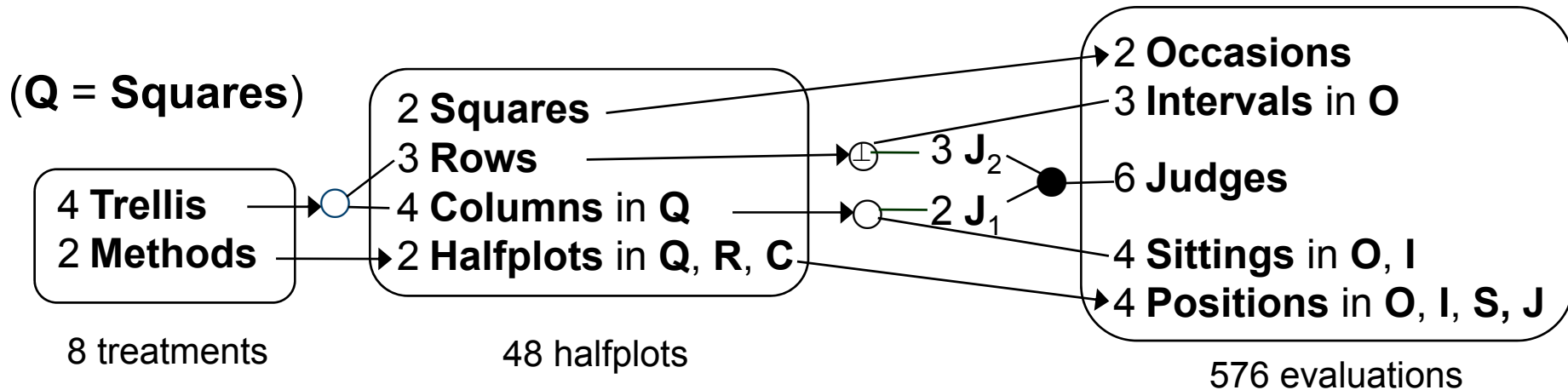
3(i) A Two-Phase Sensory Experiment

(Brien & Bailey, 2006, Example 15)

(Brien & Payne, 1999)

■ Involves two randomizations:

- *Field phase*: 8 treatments to 48 halfplots using split-plot with 2 Youden squares for main plots.
- *Sensory phase*: 48 halfplots randomized 576 evaluations, using Latin squares and an extended Youden square.



■ Analysis formulae

$((2 \text{ Occasions} / 3 \text{ Intervals} / 4 \text{ Sittings}) * 6 \text{ Judges}) / 4 \text{ Positions}$

$(3 \text{ Rows} * (2 \text{ Squares} / 4 \text{ Columns})) / 2 \text{ Halfplots}$

4 Trellis * 2 Methods * Judges (include to generate intertier interactions)

➤ Intratier model: $T * M \mid ((O / I / S) * J) / P + (R * (Q / C)) / H$

A Two-Phase Sensory Experiment (continued)

- Fixed: treatments factors + Judges;
- Random: all halfplots and evaluations factors except Judges.

- Nonorthogonality:

- Judges interactions are unbalanced;
- Nonorthogonal variance structure (Columns);

so need mixed model estimation to get analysis with combination of information.

- Intertier mixed model:

$$T + M + T \wedge M + J + T \wedge J + M \wedge J + T \wedge M \wedge J \\ | \textcircled{O} + O \wedge I + O \wedge I \wedge S + O \wedge J + O \wedge I \wedge J + O \wedge I \wedge S \wedge J + \underline{O \wedge I \wedge S \wedge J \wedge P} \\ + \textcircled{Q} + R + Q \wedge R + Q \wedge C + Q \wedge R \wedge C + Q \wedge R \wedge C \wedge H.$$

(Q = Square)

- Randomization-based mixed model

- Delete one of \textcircled{O} and \textcircled{Q} (see decomposition table on next slide)
⇒ model of convenience.
- Might consider model with unequal variances between judges.

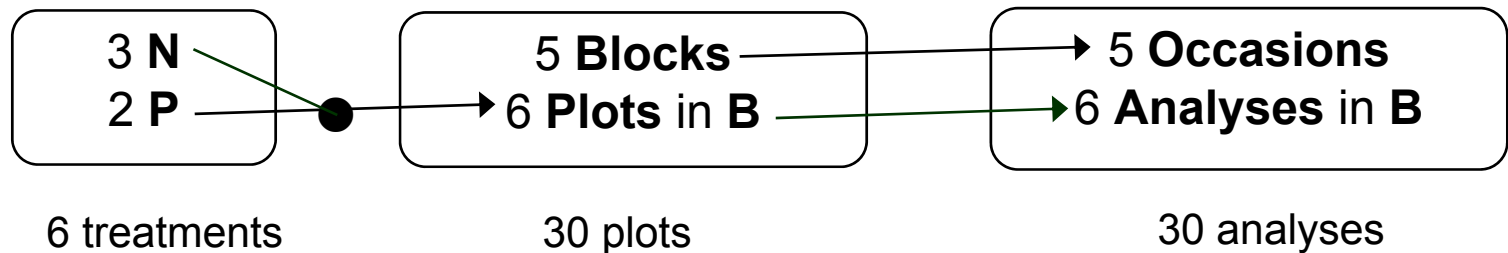
Decomposition table for sensory exp't

positions tier		wines tier			treatments tier			
source	df	eff	source	df	eff	source	df	
Occasions	1		Squares	1				
Judges	5							
O#J	5							
Intervals[O]	4							
I#J[O]	20		Rows	2				
			Q#R	2				
			Residual	16				
Sittings[O^I]	18	1/3	Columns[Q]	6	1/27	Trellis	3	
						Residual	3	
				Residual	12			
S#J[O^I]	90	2/3	Columns[Q]	6	2/27	Trellis	3	
						Residual	3	
				R#C[Q]	12	8/9	Trellis	3
						Residual	9	
				Residual	72			
Positions[O^I^S^J]	432		Halfplots[R^C^Q]	24		Method	1	
						T#M	3	
						Residual	20	
				Residual	408			

3(ii) A simple example

- Field phase
 - RCBD in which 3 levels of N x 2 levels of P assigned to 6 plots in 5 blocks
- Laboratory phase
 - RCBD for 6 analyses in 5 occasions in which blocks assigned to occasions and plots to analyses.
- What are the sets of objects?
- What is the randomization diagram?
- What are the tiers?
- What are the model formulae and sources?
- What is the decomposition table, including the idempotents?

3(ii) A simple example (cont'd)



■ Decomposition table:

analyses tier		plots tier		treatments tier	
source	df	source	df	source	df
Occasions	4	Blocks	4		
Analyses[O]	25	Plots[B]	25	N	2
				P	1
				N#P	2
				Residual	20

- What is the intratier or randomization model?
- What is the model of convenience for a software package?

4. Intertier interactions

(Brien & Bailey, 2006, Section 7.1)

- Interaction between factors from different tiers
 - Includes block-, unit- and experiment-treatment interaction.

- Randomized-randomized
 - e.g. treatments factors randomized in separate randomizations as with Varieties and Regimes in Brien and Bailey (2006, Example 5).

- Unrandomized-randomized
 - Scientists interested in interaction of treatments with an unrandomized factor as for Trellis or Methods with Judges in Brien and Bailey (2006, Example 15).

- Need to take account of intertier interactions in design otherwise may be unbalanced
 - Trellis#Judges interaction in Brien and Bailey (2006, Example 15) is unbalanced.

5. Lab phase design: a biodiversity example

(Harch et al., 1997)

- Effect of tillage treatments on bacterial and fungal diversity
- Two-phase experiment: field and laboratory phase

Field phase:

- 2 tillage treatments assigned to plots using RCBD with 4 blocks
- 2 soil samples taken at each of 2 depths

⇒ $2 \times 4 \times 2 \times 2 = 32$ samples

Laboratory phase:

- Then analysed soil samples in the lab using **Gas Chromatography - Fatty Acid Methyl Ester (GC-FAME)** analysis
- 2 preprocessing methods randomized to 2 samples in each Plot^Depth
- All samples analysed twice — necessary?
 - once on days 1 & 2; again on day 3

	Day			
	1	2	am of 3	pm of 3
Int1	1	1	2	2
Int2	1	2	1	2
Blocks	1 & 2	3 & 4	1 & 2	3 & 4

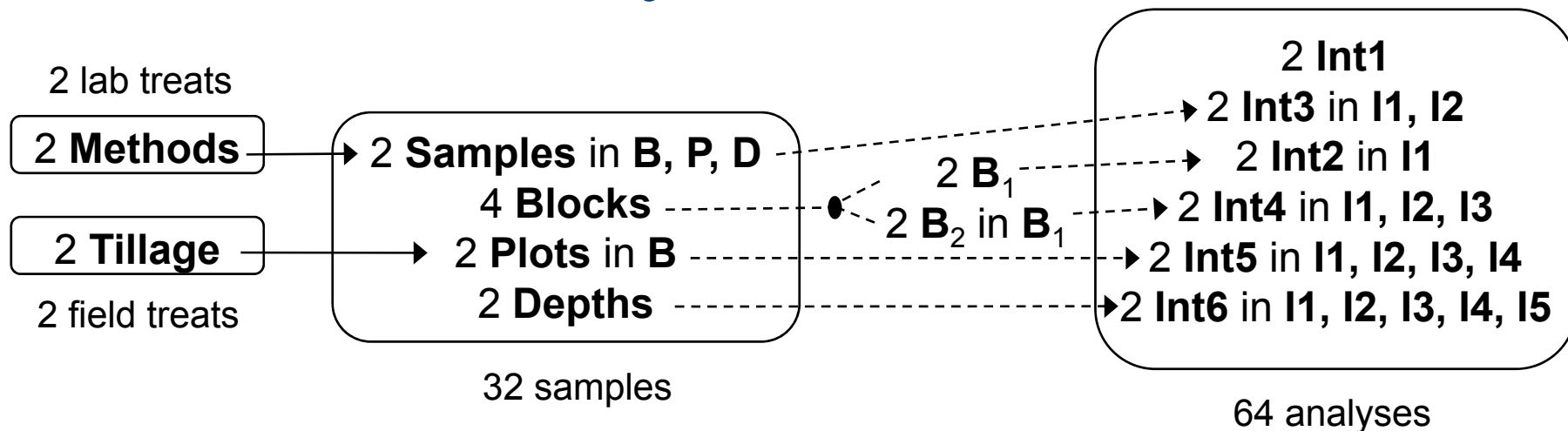
- In each Int2, 16 samples analyzed

Processing order within $\text{Int1} \wedge \text{Int2}$

Analysis	Method	Block	Plot	Depth	Analysis	Method	Block	Plot	Depth
1	ground	A	1	0-5cm	9	sieved	A	1	0-5cm
2	ground	A	1	5-10cm	10	sieved	A	1	5-10cm
3	ground	A	2	0-5cm	11	sieved	A	2	0-5cm
4	ground	A	2	5-10cm	12	sieved	A	2	5-10cm
5	ground	B	1	0-5cm	13	sieved	B	1	0-5cm
6	ground	B	1	5-10cm	14	sieved	B	1	5-10cm
7	ground	B	2	0-5cm	15	sieved	B	2	0-5cm
8	ground	B	2	5-10cm	16	sieved	B	2	5-10cm

- Logical as similar to order obtained from field
- But confounding with systematic laboratory effects:
 - Preprocessing method effects
 - Depth effects
- Depths assigned to lowest level — sensible?

Towards an analysis



- 64 analyses divided up hierarchically by 6 x 2-level factors Int1...Int6.

- Dashed arrows indicate systematic assignment

Int1	Int2	Block
1	1	1&2
1	2	3&4
2	1	1&2
2	2	3&4

Int3	Int4	Int5	Int6	Method	Block	Plot	Depth
2	1	1	1	grievd	A	1	0-5cm
2	1	1	2	grievd	A	1	5-10cm
2	1	2	1	grievd	A	2	0-5cm
2	1	2	2	grievd	A	2	5-10cm
2	2	1	1	grievd	B	1	0-5cm
2	2	1	2	grievd	B	1	5-10cm
2	2	2	1	grievd	B	2	0-5cm
2	2	2	2	grievd	B	2	5-10cm

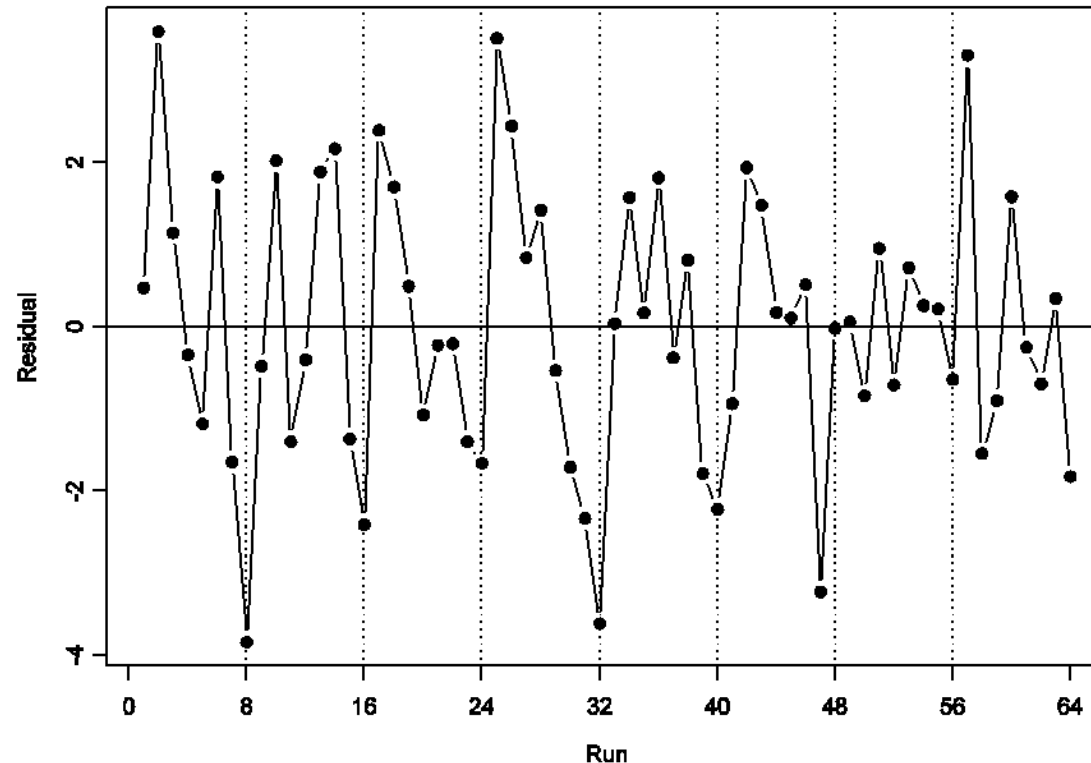
Analysis of example for lab variability

Source	df	SSq	MSq	F	p
Int1	1	0.72	0.72	19.98	0.140
Int2[Int1]	2	58.84			
Block	1	58.80	58.80	1621.46	0.016
Residual	1	0.04	0.04	0.02	
Int3[Int1^Int2]	4				
Sample[Block^Plot^Depth]	2	121.82			
Method	1	121.28	121.28	224.89	0.042
Residual	1	0.54	0.54		
Residual	2	3.19	1.59	0.43	
Int4[Int1^Int2^Int3]	8	64.54			
Block	2	37.68	18.84	5.05	0.080
Sample[Block^Plot^Depth]	2	11.94	5.97	1.60	0.308
Residual	4	14.91	3.73	2.43	0.133
Int5[Int1^Int2^Int3^Int4]	16	56.72			
Plot[Block]	4	43.22			
Tillage	1	1.60	1.60	0.12	0.757
Residual	3	41.62	13.87		
Sample[Block^Plot^Depth]	4	3.20			
Tillage#Method	1	3.04	3.04	56.26	0.005
Residual	3	0.16	0.05		
Residual	8	12.30	1.54	1.58	0.207
Int6[Int1^Int2^Int3^Int4^Int5]	32	209.67			
Depth	1	137.66	137.66	141.75	<0.001
Block#Depth	3	13.74	4.58	4.72	0.015
Plot#Depth[Block]	4	15.87			
Tillage#Depth	1	3.87	3.87	0.97	0.398
Residual	3	12.00	4.00	1.9	
Sample[Block^Plot^Depth]	8	26.86			
Depth#Method	1	13.22	13.22	6.28	0.046
Tillage#Depth#Method	1	1.01	1.01	0.48	0.514
Residual	6	12.63	2.10		
Residual	16	15.54	0.97		

- Variability for:
 - Int4 > Int5 > Int6
 - 8 > 4 > 2
 - Analyses
 - Int1, Int2, Int3 small (< Int4)
 - Recalibration?

Trend in the biodiversity example

- Trend can be a problem in laboratory phase. Is it here?
- Plot of Lab-only residuals in run order for 8 Analyses within Times



- Linear trend that varies evident and is significant.
- Need to design bearing this in mind.

6. Some questions

- How to design the laboratory phase of a two-phase experiment? Trend over time to be allowed for, but with variation from two phases.
- What is the role for multiphase experiments in microarray experiments?
- Are two-phase experiments applicable in clinical trials?
- What is the multitiered perspective on multistage, reprocessing experiments?
- Are there other experiments that employ double randomizations?

7. Summary

- Tier-based decomposition/ANOVA tables useful for:
 - Establishing properties of design;
 - Analyzing orthogonal experiments;
 - Formulating mixed models.
- Mixed model estimation is most general and widely available.
- Randomization-based mixed models have advantages that:
 - pertinent sources are identified;
 - ensures all potential sources included in the model, particularly for complex experiments;
 - makes explicit where the model deviates from the randomization model.
- Intertier interactions are often required
- Often need to drop inextricably confounded terms and so models of convenience often input to packages.

Web addresses

- INI site

<http://www.newton.ac.uk/webseminars>

- Multitiered experiments site

<http://chris.brien.name/multitier>

(see Research > Slides for talks)