Day 4
Spatial variation and genotype by environment interaction

Linear mixed models

Daniel Tolhurst, Jon Bancic, Chris Gaynor, Gregor Gorjanc

Lecture overview

1. **Design of plant breeding field trials**
   - Fundamental concepts of experimental design
   - Classical and model-based designs

2. **Linear mixed models for plants ←**
   - Complex residual variance structures
   - Spatial variation


Linear mixed models for
plant breeding field trials

Outline

- **Background**
- **Formulating a linear mixed model**
- **Spatial variation**
  - Global and local trend
  - Random error
  - Extraneous variation

Randomised complete block (RCB) design

$$y_{ij} = \mu + g_i + b_j + e_{ij}$$

phenotype mean genotype block residual

- Simple to construct
- Balanced, complete and resolvable
- Genotypes and blocks are orthogonal
- **But,** assumes blocks are homogeneous

Block 1		Block 2	
G1	G4	G5	G12
G14	G11	G10	G16
G6	G8	G7	G2
G10	G16	G3	G15
G2	G5	G9	G8
G3	G13	G1	G6
G12	G15	G11	G4
G9	G7	G14	G13

Scalar notation

$$y_{ij} = \mu + g_i + b_j + e_{ij}$$

phenotype mean genotype block residual

- y_{ij} is the phenotype of genotype i in block j (n in total)
- μ is the overall mean
- g_i is the effect of genotype i ($i = 1, \ldots, n_g$)
- b_j is the effect of block j ($j = 1, \ldots, n_b$)
- e_{ij} is the plot residual of genotype i in block j (n in total)

Scalar → vector notation

$$\begin{bmatrix} y_{1;1} \\ \vdots \\ y_{n_b;n_g} \end{bmatrix} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} + \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} g_1 \\ \vdots \\ g_{n_g} \end{bmatrix} + \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_{n_g} \end{bmatrix} + \begin{bmatrix} e_{1;1} \\ \vdots \\ e_{n_b;n_g} \end{bmatrix}$$

phenotype **mean** **design matrix** **genotype** **design matrix** **block** **residual**

- y_{ij} is the phenotype of genotype i in block j (n in total)
- μ is the overall mean
- g_i is the effect of genotype i $(i = 1, \dots, n_g)$
- b_j is the effect of block j $(j = 1, \dots, n_b)$
- e_{ij} is the plot residual of genotype i in block j (n in total)

Vector notation

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{Zg} + \mathbf{Wb} + \mathbf{e}$$

phenotype mean genotype block residual

- $\mathbf{y}$ is the n-vector of phenotypes (ordered as plots within blocks)
- μ is the overall mean, $\mathbf{1}_n$ is a n-vector of ones
- $\mathbf{g}$ is the n_g-vector of genotype effects, with $n \times n_g$ design matrix $\mathbf{Z}$ which links plots to genotypes
- $\mathbf{b}$ is the n_b-vector of block effects, with $n \times n_b$ design matrix $\mathbf{W}$ which links plots to blocks
- $\mathbf{e}$ is the n-vector of residuals

Linear mixed models (LMMs)

- Ordinary linear models comprise fixed effects + random error
- Linear mixed models comprise **fixed + random** effects + random error
  - Analysis of incomplete/unbalanced block designs with recovery of interblock information (Patterson & Thompson, 1971)
  - LMM with fixed genotype effects and random (incomplete) blocks
  - First application of residual/restricted maximum likelihood (REML)
  - Equivalent to ANOVA estimates when blocks are equal size (Nelder, 1968)

Eisenhart (1947)

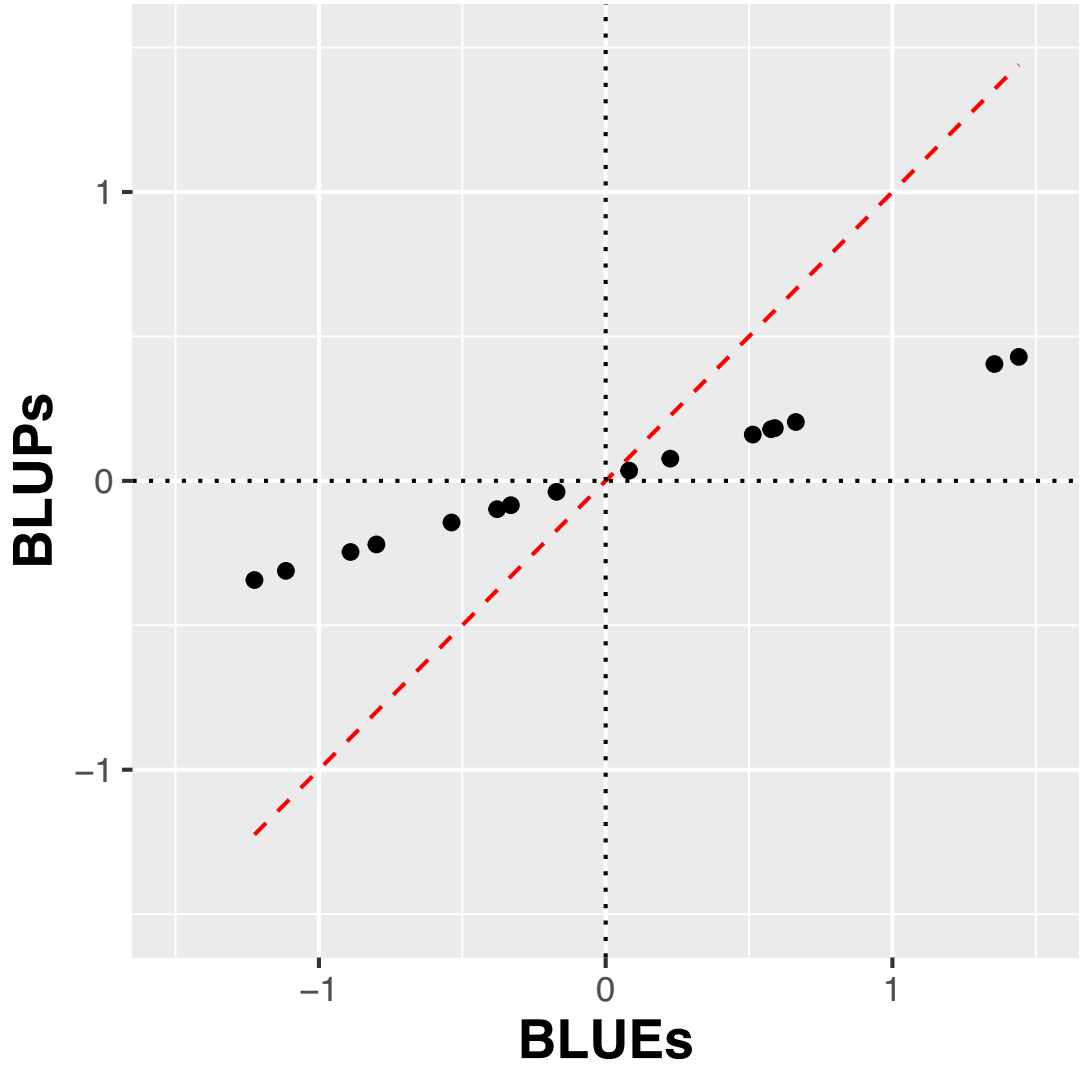
Fixed or random effects

- **The choice of fitting effects as fixed or random is important**
  - Fixed effects
    - Contribute to $\mathrm{E}(\boldsymbol{y})$
    - Best linear unbiased estimates (BLUEs)
  - Random effects
    - Contribute to $\mathrm{Var}(\boldsymbol{y})$
    - Realisations of random variables
    - Best linear unbiased predictions (BLUPs); shrinkage to mean according to amount of information

Henderson (1975)

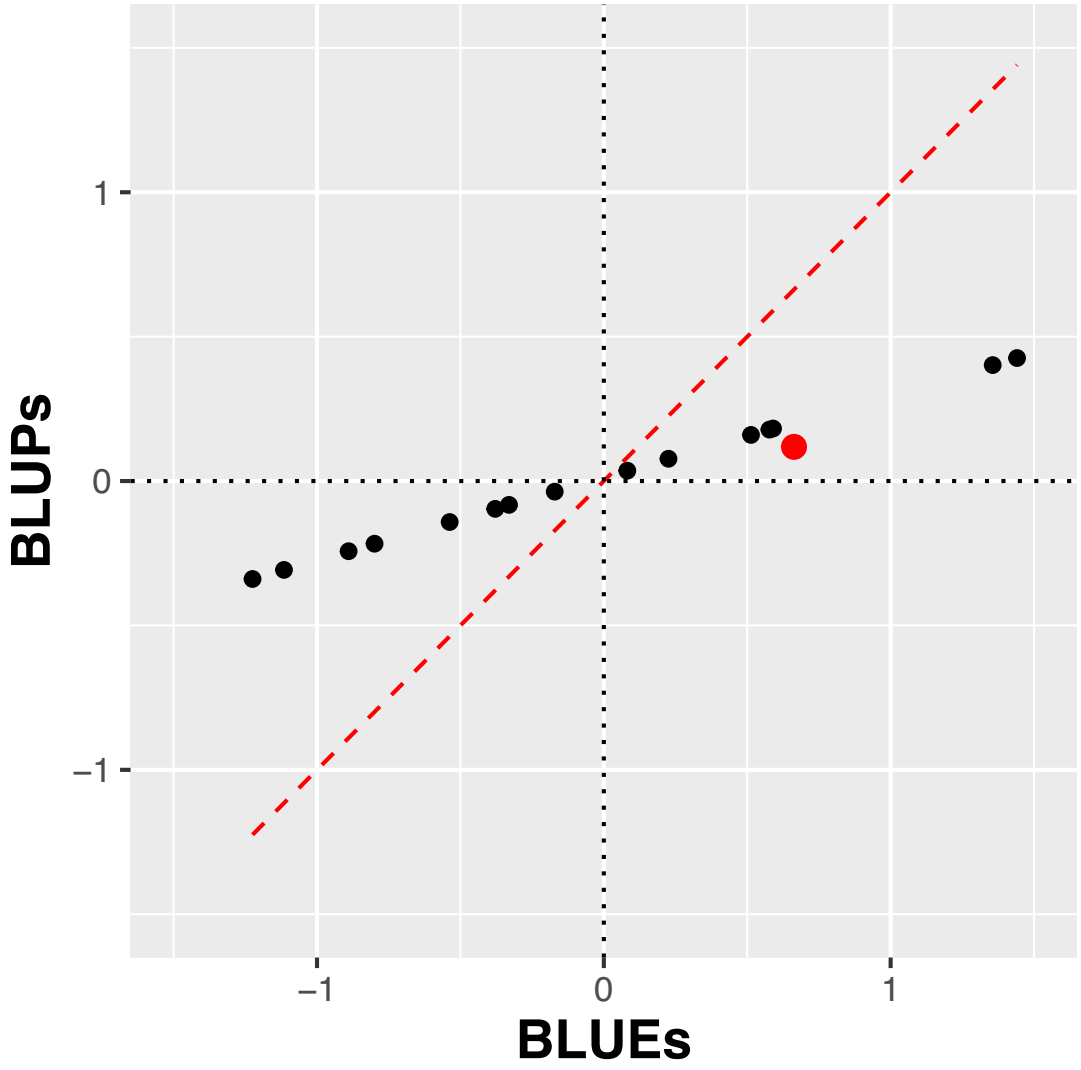
Shrinkage

- BLUP = BLUE x shrinkage
  - **Same number of replicates**
    = same information
    = same shrinkage
  - Different number of replicates
    = different information
    = different shrinkage
  - Spatial variability in the field
    = different information
    = different shrinkage



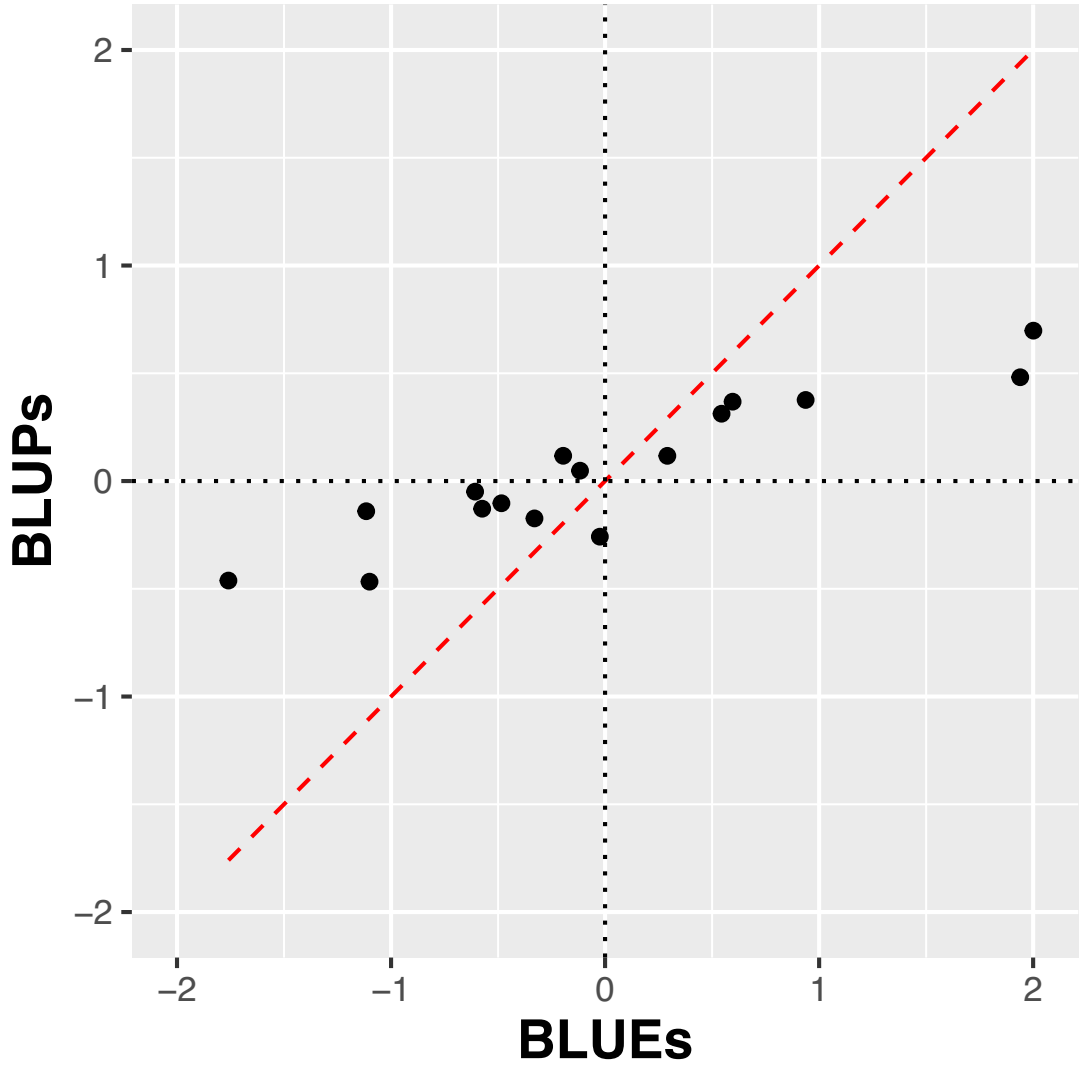
Shrinkage

- BLUP = BLUE x shrinkage
  - Same number of replicates
    = same information
    = same shrinkage
  - **Different number of replicates**
    = different information
    = different shrinkage
  - Spatial variability in the field
    = different information
    = different shrinkage



Shrinkage

- BLUP = BLUE x shrinkage
  - Same number of replicates
    = same information
    = same shrinkage
  - Different number of replicates
    = different information
    = different shrinkage
  - **Spatial variability in the field**
    = different information
    = different shrinkage



Model assumptions

Assume the genotype and block effects are fitted as random

$$\mathrm{E}(\boldsymbol{y}) = \mathbf{1}_n\mu \quad \text{and} \quad \mathrm{Var}(\boldsymbol{y}) = \sigma_g^2\mathbf{Z}\mathbf{Z}' + \sigma_b^2\mathbf{W}\mathbf{W}' + \sigma_e^2\mathbf{I}_n$$

$$\begin{bmatrix}\mathbf{g}\\ \mathbf{b}\\ \mathbf{e}\end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_g^2\mathbf{I}_{n_g} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \sigma_b^2\mathbf{I}_{n_b} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \sigma_e^2\mathbf{I}_n\end{bmatrix}\right)$$

– σ_g^2 is a genetic variance, $\mathbf{I}_{n_g}$ is a $n_g \times n_g$ identity matrix

– σ_b^2 is a block variance, $\mathbf{I}_{n_b}$ is a $n_b \times n_b$ identity matrix

– σ_e^2 is a residual variance, $\mathbf{I}_n$ is a $n \times n$ identity matrix

Updating model assumptions

Complex genetic and residual variance structures:

$$\mathrm{E}(\boldsymbol{y}) = \mathbf{1}_n\mu \quad \text{and} \quad \mathrm{Var}(\boldsymbol{y}) = \sigma_g^2\mathbf{Z}\mathbf{G}\mathbf{Z}' + \sigma_b^2\mathbf{W}\mathbf{W}' + \mathbf{R}$$

$$\begin{bmatrix}\mathbf{g}\\ \mathbf{b}\\ \mathbf{e}\end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_g^2\mathbf{G} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \sigma_b^2\mathbf{I}_{n_b} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{bmatrix}\right)$$

- σ_g^2 is the genetic variance, $\mathbf{G}$ is a $n_g \times n_g$ genotype relationship matrix
- σ_b^2 is the block variance, $\mathbf{I}_{n_b}$ is a $n_b \times n_b$ identity matrix
- $\mathbf{R}$ is a $n \times n$ residual variance matrix

Mixed model equations (MMEs)

- Estimates of fixed effects (BLUEs) and predictions of random effects (BLUPs) obtained from the mixed model equations

$$\begin{bmatrix} \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{1}_n & \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{1}_n & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}/\sigma_g^2 & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{1}_n & \mathbf{W}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \mathbf{I}_{n_b}/\sigma_b^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \tilde{\mathbf{g}} \\ \tilde{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\mu} \\ \tilde{\mathbf{g}} \\ \tilde{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \left(\mathbf{1}_n'\mathbf{H}^{-1}\mathbf{1}_n\right)^{-1}\mathbf{1}_n'\mathbf{H}^{-1}\mathbf{y} \\ \sigma_g^2\mathbf{G}\mathbf{Z}'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{1}_n\hat{\mu}) \\ \sigma_b^2\mathbf{W}'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{1}_n\hat{\mu}) \end{bmatrix}$$

$$\mathbf{H} = \mathbf{Var}(\mathbf{y}) = \sigma_g^2\mathbf{Z}\mathbf{G}\mathbf{Z}' + \sigma_b^2\mathbf{W}\mathbf{W}' + \mathbf{R}$$

Henderson (1950)

Residual variance structure (R)

- Variance component model:

$$\mathbf{R} = \sigma_e^2 \mathbf{I}_n$$

- Assumes plots within a field are **independent**
  - Rarely sensible because plots are known to have some level of correlation
  - Spatial variation is ubiquitous in field trials

Residual variance structure (R)

- Covariance model:

$$\mathbf{R} = \sigma_e^2 \, \boldsymbol{\Sigma}_{\mathbf{c}}(\rho_c) \otimes \boldsymbol{\Sigma}_{\mathbf{r}}(\rho_r)$$

  - σ_e^2 is the residual variance
  - $\boldsymbol{\Sigma}_{\mathbf{c}}$ is a $n_c \times n_c$ matrix with column autocorrelation parameter ρ_c
  - $\boldsymbol{\Sigma}_{\mathbf{r}}$ is a $n_r \times n_r$ matrix with row autocorrelation parameter ρ_r

- Assumes plots within a field are **correlated**
  - Two-dimensional stochastic variance matrix
  - Plots closer together are more correlated than those further apart

Spatial variation

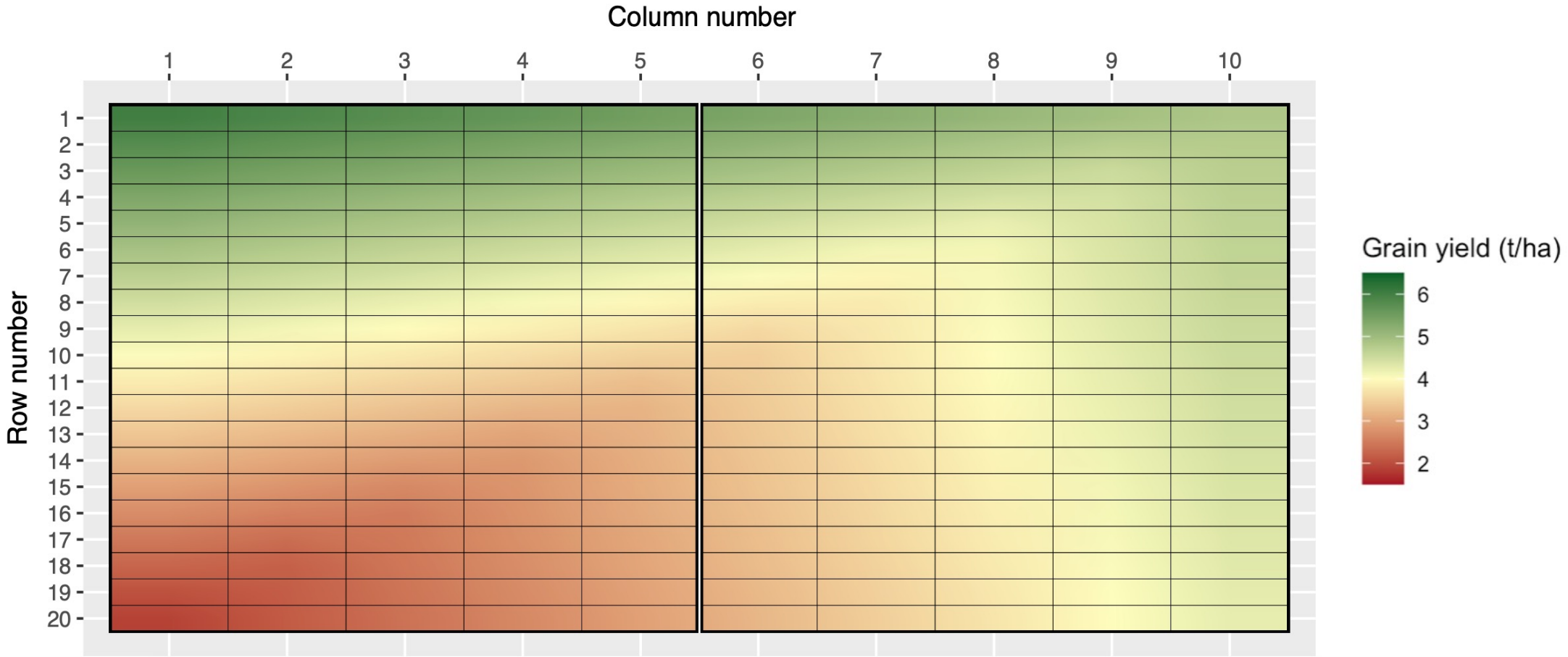
Some concepts

- Spatial variation arises from heterogeneity across the trial area
- **Global and local trend (smooth spatial trend)**
  - Large and small scale changes in fertility/soil composition
- **Random error (noise)**
  - Measurement error, or variability in the plots themselves
- **Extraneous variation (systematic variation)**
  - Induced during the conduct of the trial

Gilmour et al. (1997)

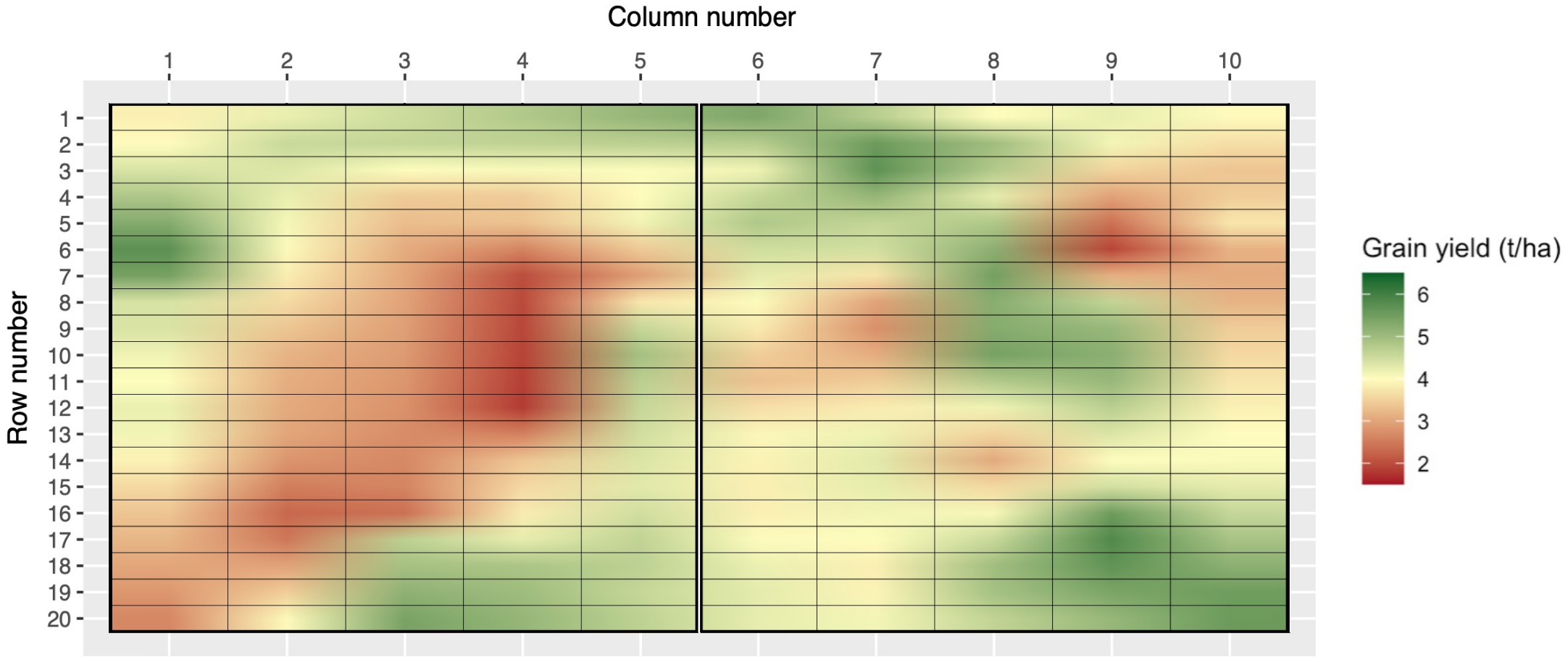
Global and local trend (smooth spatial trend)

- **Large** and small scale changes in fertility/soil composition



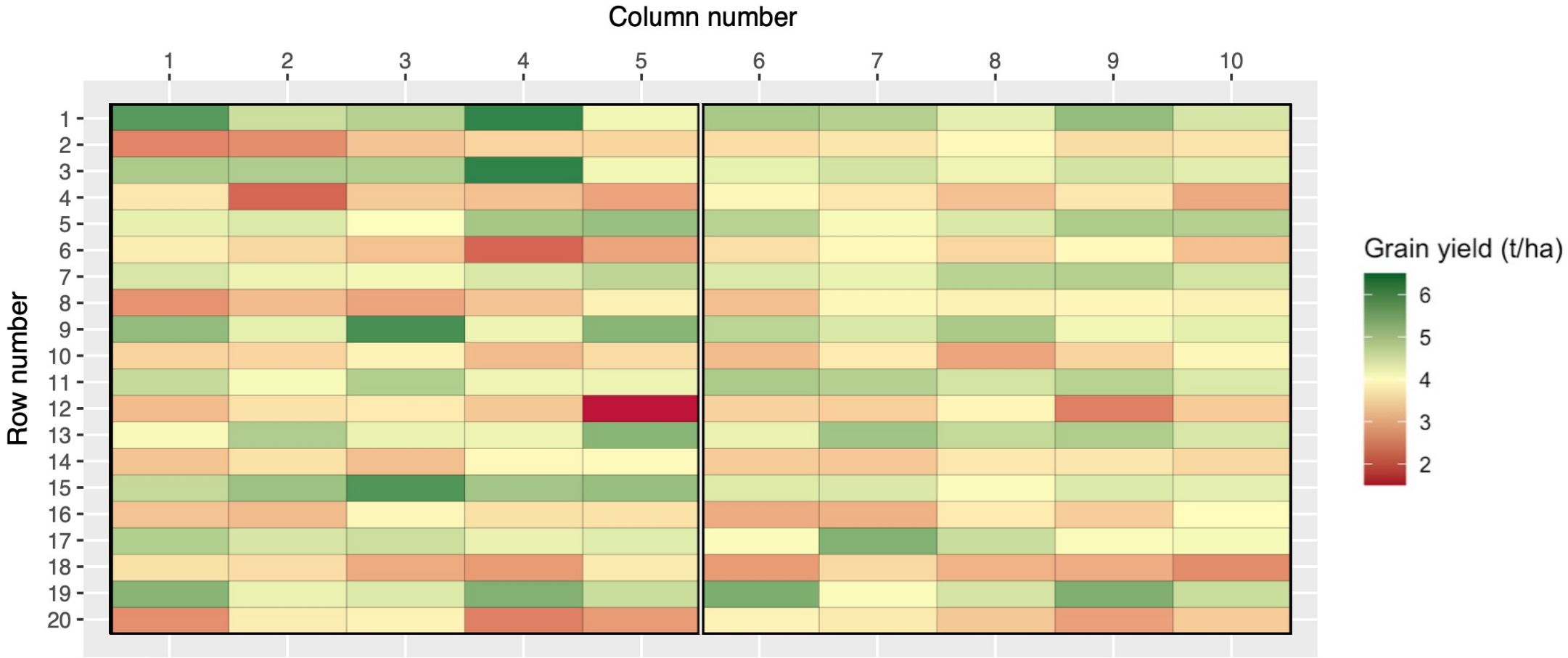
Global and local trend (smooth spatial trend)

- Large and **small** scale changes in fertility/soil composition



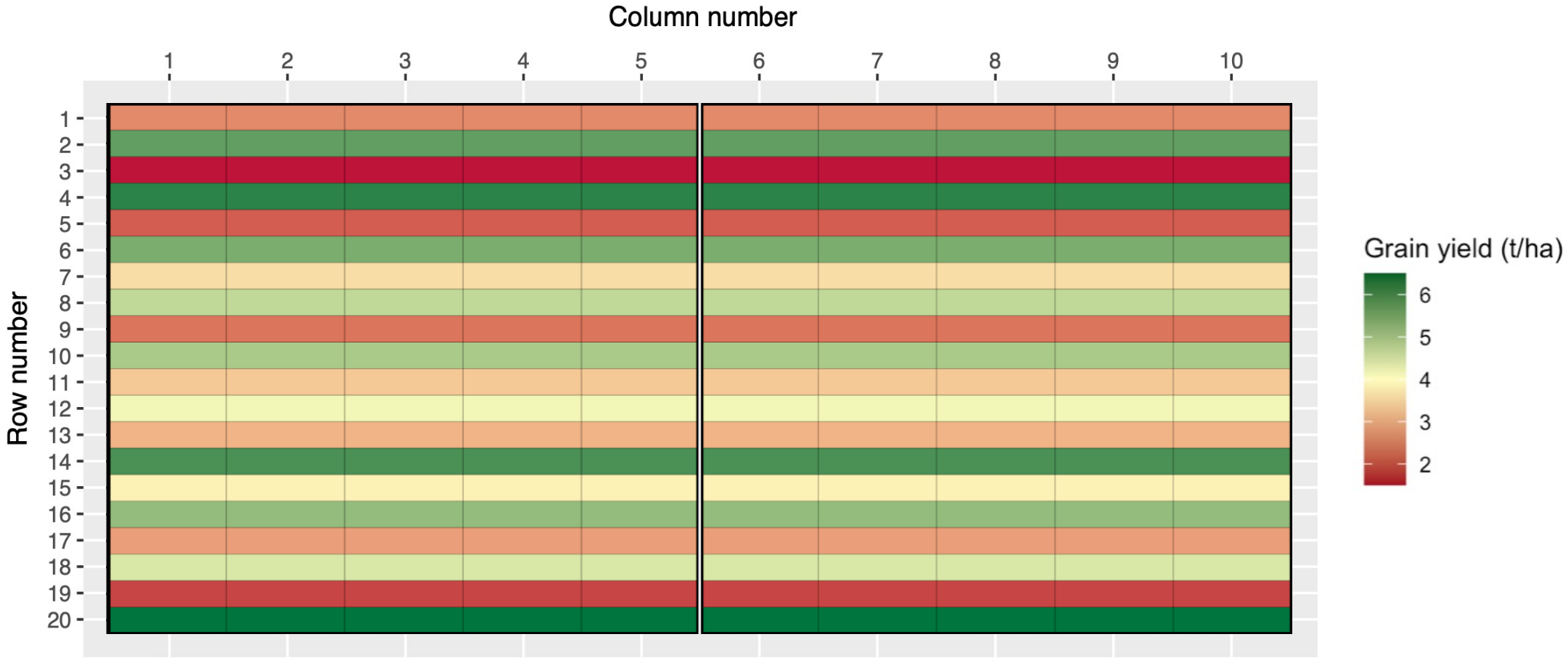
Random error (noise)

Measurement error, or variability in the plots themselves



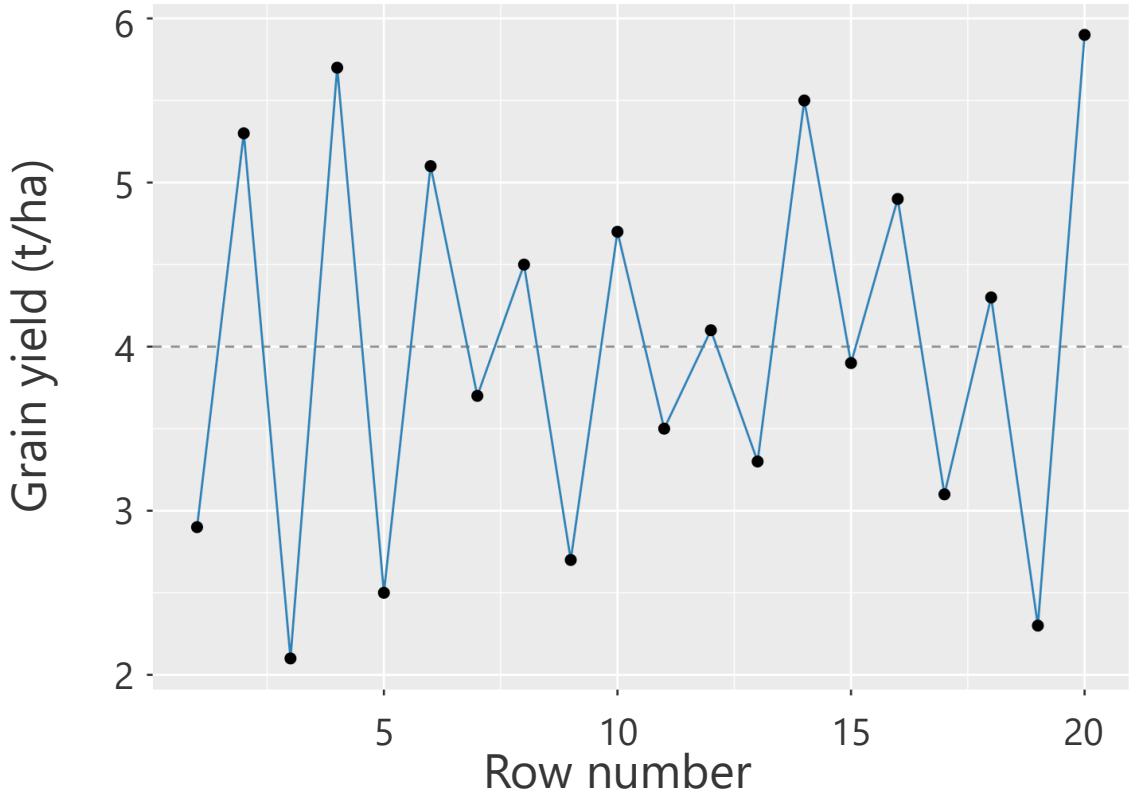
Extraneous variation

- Induced during the conduct of the trial



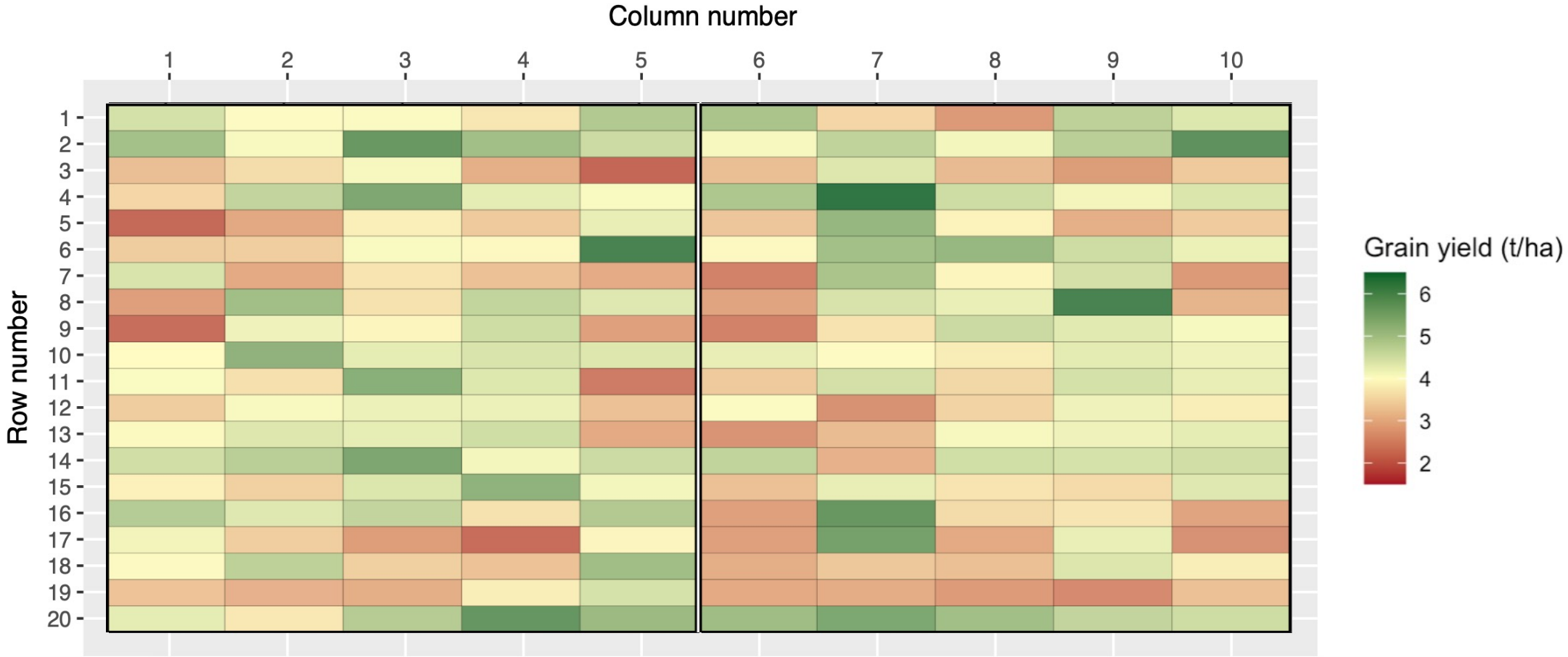
Extraneous variation

- Induced during the conduct of the trial
  - Typically aligned with the field columns and row



Total spatial variation

- Some combination of global and local trend, random error and extraneous variation



Spatial models

Some popular approaches

- **Separable autoregressive process (Cullis & Gleeson, 1991)**
  - **Stochastic variance matrix**
- Tensor product penalised splines (Rodríguez-Álvarez et al., 2018)
  - Smoothing function in two dimensions
- Nearest neighbour adjustments (Papadakis, 1937)
  - Adjust phenotypes based on neighbouring plots

Accounting for trend and noise

- Autoregressive covariance model + random error:

$$\mathbf{R} = \sigma_s^2 \, \boldsymbol{\Sigma}_{\mathbf{c}}(\rho_c) \otimes \boldsymbol{\Sigma}_{\mathbf{r}}(\rho_r) + \sigma_r^2 \mathbf{I}_n$$

trend noise

  - σ_s^2 is the autoregressive scaling component
  - $\boldsymbol{\Sigma}_{\mathbf{c}}$ is a $n_c \times n_c$ matrix with column autocorrelation ρ_c
  - $\boldsymbol{\Sigma}_{\mathbf{r}}$ is a $n_r \times n_r$ matrix with row autocorrelation ρ_r
  - σ_r^2 is the random error variance

Autoregressive covariance matrix

- **Autoregressive covariance model** + random error:

$$\mathbf{R} = \sigma_s^2 \, \mathbf{\Sigma_c}(\rho_c) \otimes \mathbf{\Sigma_r}(\rho_r) + \sigma_r^2 \mathbf{I}_n$$

trend noise

- Assumes the phenotypes are ordered as rows in cols

$$\begin{bmatrix} y_{1;1} \\ \vdots \\ y_{n_b;n_g} \end{bmatrix} \rightarrow \begin{bmatrix} y_{1;1} \\ \vdots \\ y_{n_c;n_r} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} e_{1;1} \\ \vdots \\ e_{n_b;n_g} \end{bmatrix} \rightarrow \begin{bmatrix} e_{1;1} \\ \vdots \\ e_{n_c;n_r} \end{bmatrix}$$

Autoregressive covariance matrix

- **Autoregressive covariance model** + random error:

$$\mathbf{R} = \sigma_s^2 \, \boldsymbol{\Sigma}_{\mathbf{c}}(\rho_c) \otimes \boldsymbol{\Sigma}_{\mathbf{r}}(\rho_r) + \sigma_r^2 \mathbf{I}_n$$

trend noise

- Assumes exponential decay according to a first order process in the column and row directions (AR1 x AR1)

$$\mathbf{R} = \sigma_s^2 \begin{bmatrix} 1 & \rho_c & \rho_c^2 & \cdots & \rho_c^{n_c-1} \\ \rho_c & 1 & \rho_c & \rho_c^2 & \vdots \\ \rho_c^2 & \rho_c & 1 & \rho_c & \rho_c^2 \\ \vdots & \rho_c^2 & \rho_c & 1 & \rho_c \\ \rho_c^{n_c-1} & \cdots & \rho_c^2 & \rho_c & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho_r & \rho_r^2 & \cdots & \rho_r^{n_r-1} \\ \rho_r & 1 & \rho_r & \rho_r^2 & \vdots \\ \rho_r^2 & \rho_r & 1 & \rho_r & \rho_r^2 \\ \vdots & \rho_r^2 & \rho_r & 1 & \rho_r \\ \rho_r^{n_r-1} & \cdots & \rho_r^2 & \rho_r & 1 \end{bmatrix} + \sigma_r^2 \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \vdots \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & 0 & 0 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix}$$

Autoregressive covariance matrix

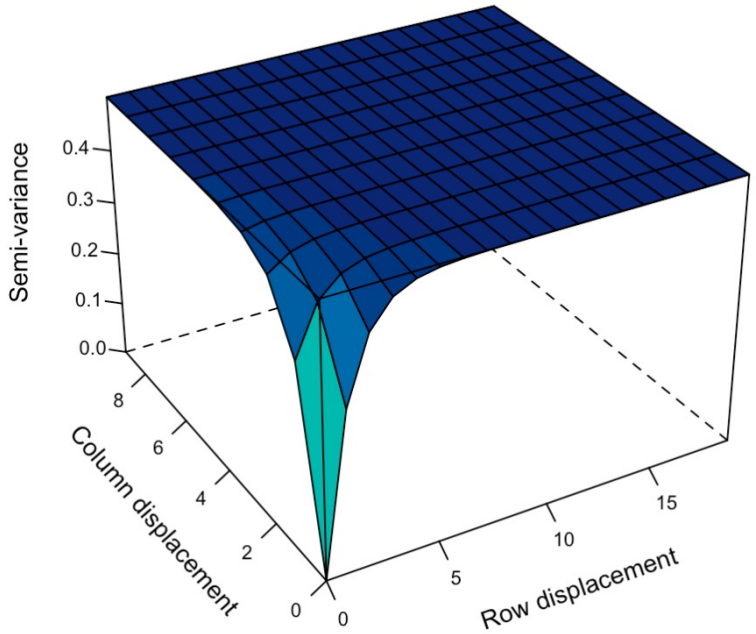
- **Autoregressive covariance model** + random error:

$$\mathbf{R} = \sigma_s^2 \, \boldsymbol{\Sigma}_{\mathbf{c}}(\rho_c) \otimes \boldsymbol{\Sigma}_{\mathbf{r}}(\rho_r) + \sigma_r^2 \mathbf{I}_n$$

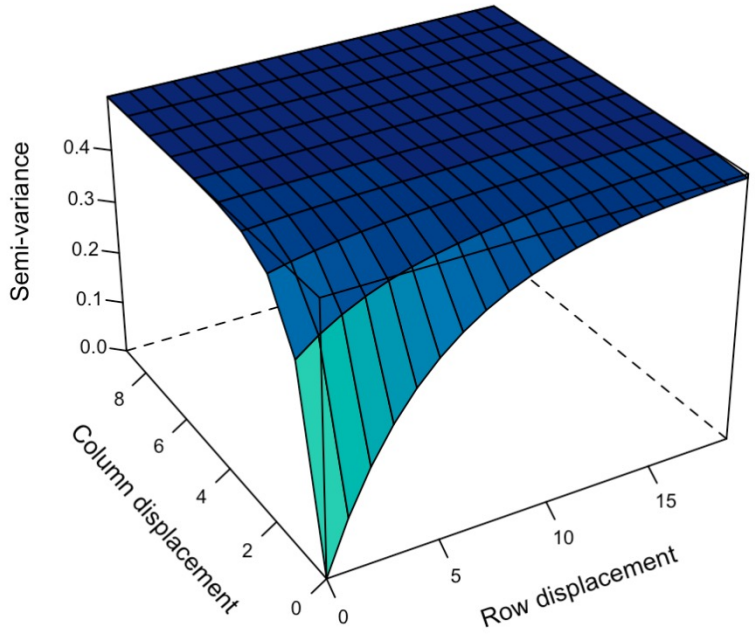
trend noise

Theoretical variogram:

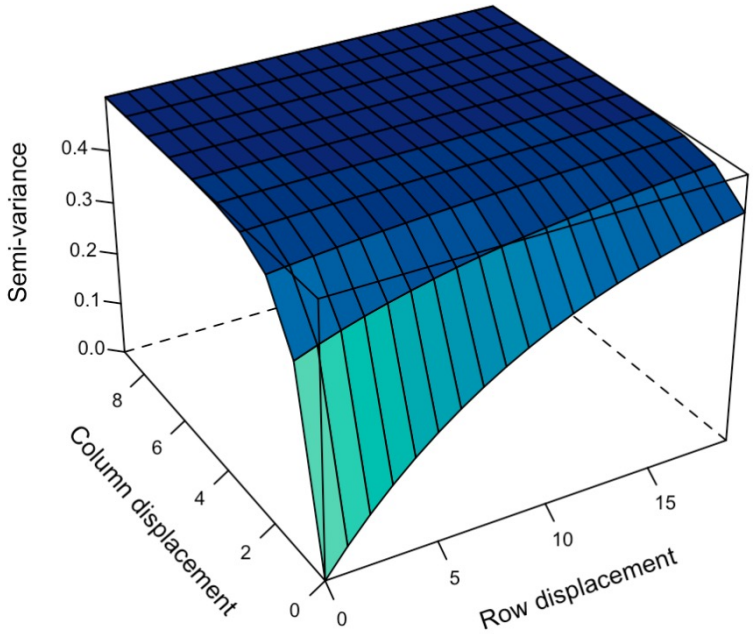
$\rho_c = 0.3,\ \rho_r = 0.4,\ \sigma_s^2 = 0.5$



$\rho_c = 0.3,\ \rho_r = 0.8,\ \sigma_s^2 = 0.5$



$\rho_c = 0.3,\ \rho_r = 0.9,\ \sigma_s^2 = 0.5$



Autoregressive covariance matrix

- **Autoregressive covariance model** + random error:

$$\mathbf{R} = \sigma_s^2 \, \mathbf{\Sigma_c}(\rho_c) \otimes \mathbf{\Sigma_r}(\rho_r) + \sigma_r^2 \mathbf{I}_n$$

trend noise

- Theoretical variogram (semi-variances):

$$v = \sigma_s^2 (1 - \rho_c^{|c_{i-j}|} \rho_r^{|r_{i-j}|})$$

  - $|c_{i-j}|$ is the absolute column displacement between plots i and j ($|c_{i-j}| = 0,1,\ldots,n_c - 1$), e.g. $|c_{i-j}| = 0$ for plots in the same column
  - $|r_{i-j}|$ is the absolute row displacement between plots i and j ($|r_{i-j}| = 0,1,\ldots,n_r - 1$), e.g. $|r_{i-j}| = 1$ for plots in adjacent rows

Autoregressive covariance matrix

- Autoregressive covariance model + **random error**:

$$\mathbf{R} = \sigma_s^2 \, \boldsymbol{\Sigma}_{\mathbf{c}}(\rho_c) \otimes \boldsymbol{\Sigma}_{\mathbf{r}}(\rho_r) + \sigma_r^2 \mathbf{I}_n$$

trend noise

- Random error term captures any remaining error variation not captured by the autoregressive covariance model

$$\mathbf{R} = \sigma_s^2 \begin{bmatrix} 1 & \rho_c & \rho_c^2 & \cdots & \rho_c^{n_c-1} \\ \rho_c & 1 & \rho_c & \rho_c^2 & \vdots \\ \rho_c^2 & \rho_c & 1 & \rho_c & \rho_c^2 \\ \vdots & \rho_c^2 & \rho_c & 1 & \rho_c \\ \rho_c^{n_c-1} & \cdots & \rho_c^2 & \rho_c & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho_r & \rho_r^2 & \cdots & \rho_r^{n_r-1} \\ \rho_r & 1 & \rho_r & \rho_r^2 & \vdots \\ \rho_r^2 & \rho_r & 1 & \rho_r & \rho_r^2 \\ \vdots & \rho_r^2 & \rho_r & 1 & \rho_r \\ \rho_r^{n_r-1} & \cdots & \rho_r^2 & \rho_r & 1 \end{bmatrix} + \boldsymbol{\sigma_r^2} \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \vdots & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}$$

Autoregressive covariance matrix

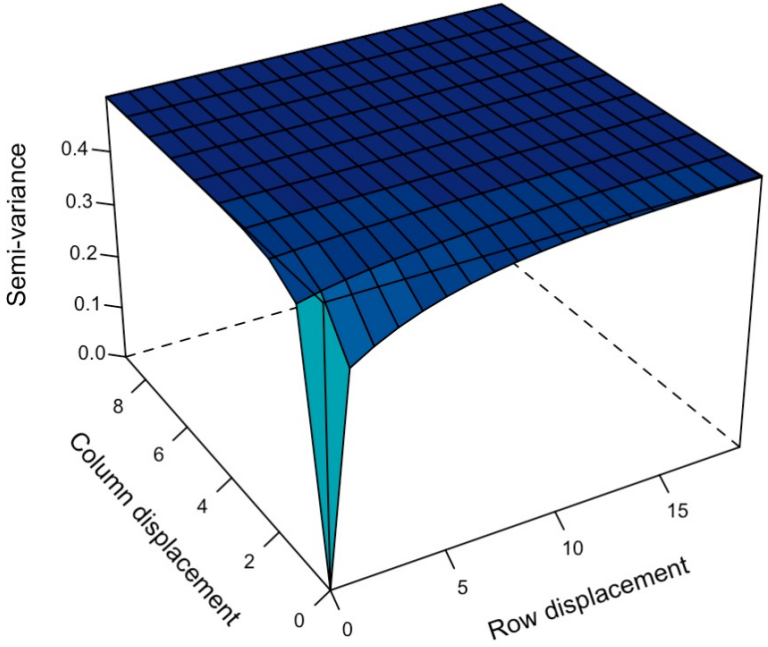
- Autoregressive covariance model + **random error**:

$$\mathbf{R} = \sigma_s^2 \, \boldsymbol{\Sigma}_{\mathbf{c}}(\rho_c) \otimes \boldsymbol{\Sigma}_{\mathbf{r}}(\rho_r) + \sigma_r^2 \mathbf{I}_n$$

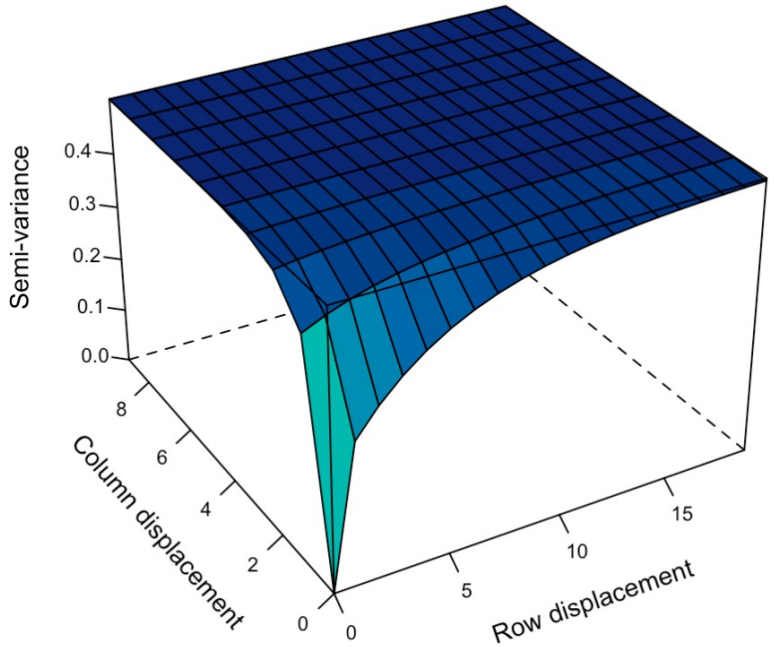
trend noise

Theoretical variogram:

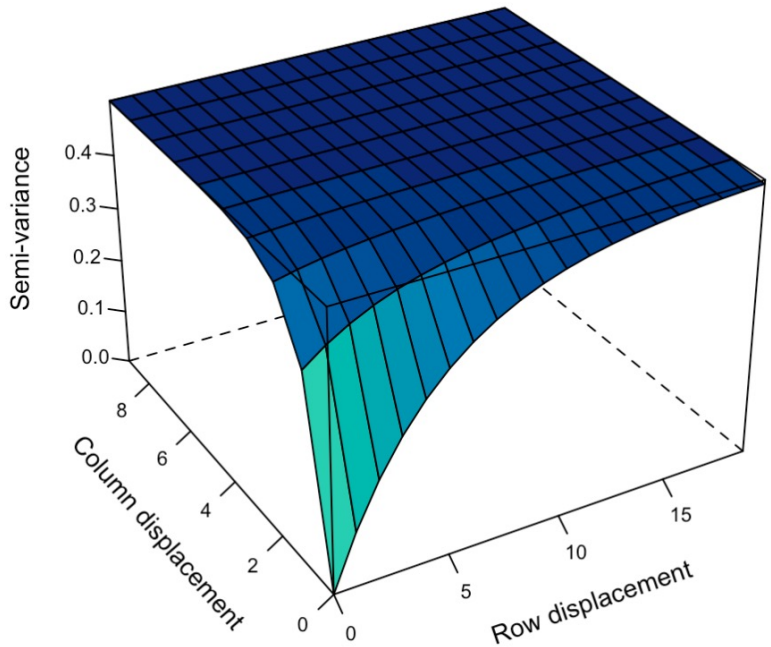
$\sigma_s^2/(\sigma_s^2 + \sigma_r^2) = 0.3$



$\sigma_s^2/(\sigma_s^2 + \sigma_r^2) = 0.6$



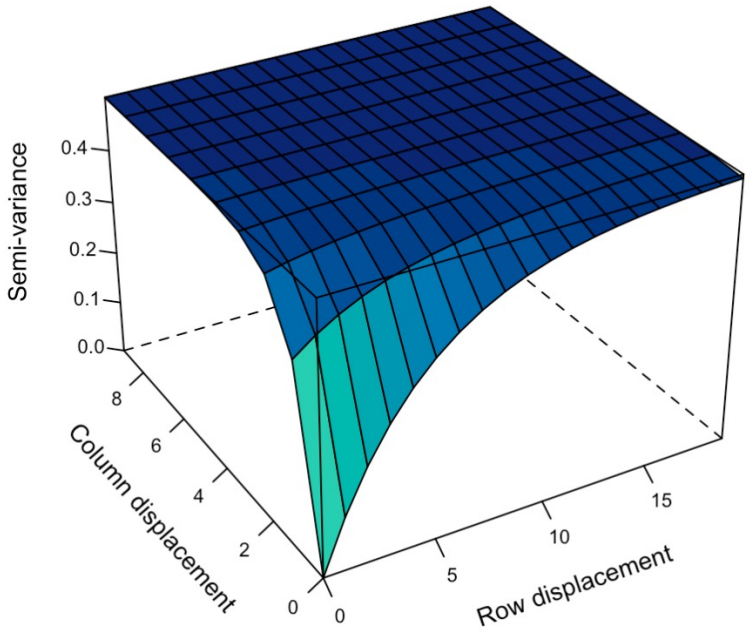
$\sigma_s^2/(\sigma_s^2 + \sigma_r^2) = 1$



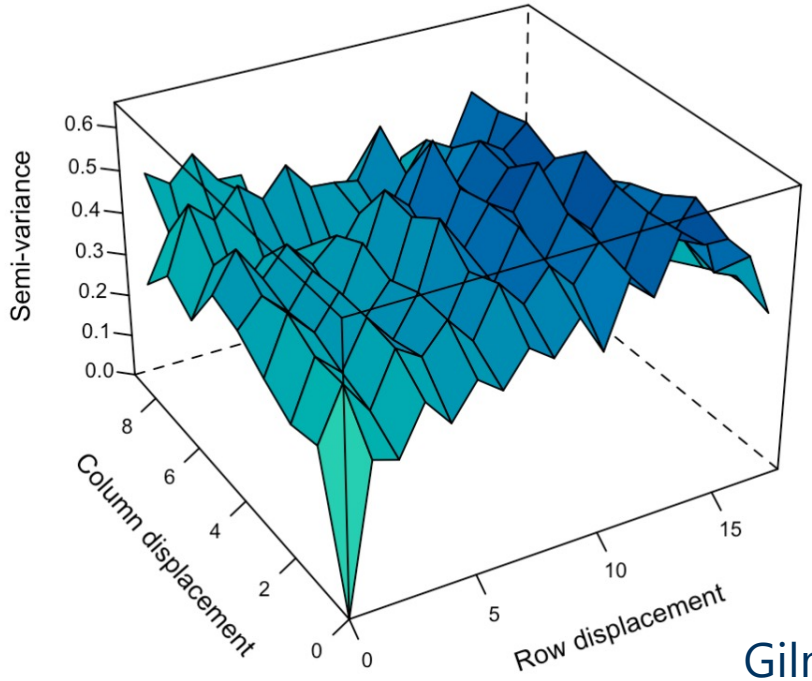
Accounting for extraneous variation

- Typically diagnosed by observing a sample variogram, which captures the average semi-variance between plots
- Accounted for by fitting additional fixed and random effects

Theoretical



Sample



Gilmour et al. (1997)

Lecture overview

1. **Design of plant breeding field trials**
   - Fundamental concepts of experimental design
   - Classical and model-based designs

2. **Linear mixed models for plants ←**
   - Complex residual variance structures ✓
   - Spatial variation ✓