Why data collection still matters: a case study from agriculture

R. A. Bailey
University of St Andrews

University of St Andrews
EPSRC Symposium: Discrete Mathematics and Big Data

(I) Variety trials for national lists

Every year, national agricultural organizations conduct trials on several varieties of several crops, each at several sites. This is done in many different countries, subject to the local regulations, in order to choose varieties of crops (such as cereals) to be included in the National List for recommendation to farmers.

(I) Variety trials for national lists

Every year, national agricultural organizations conduct trials on several varieties of several crops, each at several sites. This is done in many different countries, subject to the local regulations, in order to choose varieties of crops (such as cereals) to be included in the National List for recommendation to farmers.

A typical experiment may include up to 100 new varieties and 2–10 established "control" varieties for comparison. Each is grown in r plots, where $r = 2, 3$ or 4. Plots may be 2 m by 20 m, and are adjacent to other plots along the long edges.

(I) Variety trials for national lists

Every year, national agricultural organizations conduct trials on several varieties of several crops, each at several sites. This is done in many different countries, subject to the local regulations, in order to choose varieties of crops (such as cereals) to be included in the National List for recommendation to farmers.

A typical experiment may include up to 100 new varieties and 2–10 established "control" varieties for comparison. Each is grown in r plots, where $r = 2, 3$ or 4. Plots may be 2 m by 20 m, and are adjacent to other plots along the long edges.

A row of such plots forms a "complete" block, big enough to contain one plot for each variety. Data analysis allows for inherent differences (say, in fertility) between the blocks.

30 varieties in 4 complete blocks of 30 plots

G	V	3	N	C	A	K	R	T	1	P	I	Y	E	X	D	H	O	W	4	B	L	U	M	2	Z	Q	S	J	F
T	E	J	4	M	S	R	3	D	I	B	W	G	1	P	K	Y	U	N	F	Q	X	2	C	H	V	O	L	Z	A
T	G	F	2	O	X	Z	M	D	K	R	W	J	Y	C	U	P	H	A	3	S	L	E	1	N	4	B	Q	V	I
3	E	K	W	Q	T	N	Z	B	H	A	G	M	Y	S	C	U	I	1	O	V	J	2	P	D	X	4	F	R	L

Some notation

$$f(\omega) = \text{variety on plot } \omega.$$

Some notation

$$f(\omega) = \text{variety on plot } \omega.$$

$$Y_\omega = \text{response on plot } \omega.$$

Some notation

$$f(\omega) = \text{variety on plot } \omega.$$

$$Y_\omega = \text{response on plot } \omega.$$

$$\tau_i = \text{effect of variety } i.$$

$$\hat{\tau}_i = \text{estimate of } \tau_i \text{ from the data.}$$

Some notation

$$f(\omega) = \text{variety on plot } \omega.$$

$$Y_\omega = \text{response on plot } \omega.$$

$$\tau_i = \text{effect of variety } i.$$

$$\hat{\tau}_i = \text{estimate of } \tau_i \text{ from the data.}$$

Assume that

$$Y_\omega = \tau_{f(\omega)} + \text{stuff depending on plots.}$$

Some notation

$$f(\omega) = \text{variety on plot } \omega.$$

$$Y_\omega = \text{response on plot } \omega.$$

$$\tau_i = \text{effect of variety } i.$$

$$\hat{\tau}_i = \text{estimate of } \tau_i \text{ from the data.}$$

Assume that

$$Y_\omega = \tau_{f(\omega)} + \text{stuff depending on plots.}$$

We want to minimize

$$\sum_i \sum_{j \neq i} \mathrm{Var}(\hat{\tau}_i - \hat{\tau}_j),$$

excluding differences between pairs of control treatments.

Simple model allowing for heterogeneity within the field

$$Y_\omega = \tau_{f(\omega)} + \beta_{\text{block}(\omega)} + \epsilon_\omega$$

where

$\beta_{\text{block}(\omega)}$ is a constant depending on the block containing plot ω,

$$E(\epsilon_\omega = 0), \qquad \text{Var}(\epsilon_\omega) = \sigma^2,$$
$$\text{and} \quad \text{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \quad \text{if } \omega \neq \omega'.$$

Simple model allowing for heterogeneity within the field

$$Y_\omega = \tau_{f(\omega)} + \beta_{\text{block}(\omega)} + \epsilon_\omega$$

where

$\beta_{\text{block}(\omega)}$ is a constant depending on the block containing plot ω,

$$E(\epsilon_\omega = 0), \qquad \text{Var}(\epsilon_\omega) = \sigma^2,$$
$$\text{and} \quad \text{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \quad \text{if } \omega \neq \omega'.$$

This assumption leads to simple data analysis,
but may be unrealistic for large blocks.

Dividing each block into subblocks

40 years ago my colleagues Desmond Patterson, Emlyn Williams, Mike Talbot and Tony Hunter at the Agricultural Research Council Unit of Statistics (ARCUS) in Edinburgh proposed dividing each complete block into smaller subblocks, on the assumption that

$$Y_\omega = \tau_{f(\omega)} + \beta_{\text{block}(\omega)} + \gamma_{\text{subblock}(\omega)} + \epsilon_\omega,$$

where the γ might be either unknown constants or identically distributed mutually independent random variables.

Dividing each block into subblocks

40 years ago my colleagues Desmond Patterson, Emlyn Williams, Mike Talbot and Tony Hunter at the Agricultural Research Council Unit of Statistics (ARCUS) in Edinburgh proposed dividing each complete block into smaller subblocks, on the assumption that

$$Y_\omega = \tau_{f(\omega)} + \beta_{\text{block}(\omega)} + \gamma_{\text{subblock}(\omega)} + \epsilon_\omega,$$

where the γ might be either unknown constants or identically distributed mutually independent random variables.

They had to ensure that the subblocks could be put together to make up complete blocks, to comply with EU regulations.

They had to distribute custom-made software to enable people at the various sites to analyse their own data.

30 varieties in 4 blocks of 6 sub-blocks of 5 plots

G	V	3	N	C	A	K	R	T	1	P	I	Y	E	X	D	H	O	W	4	B	L	U	M	2	Z	Q	S	J	F
T	E	J	4	M	S	R	3	D	I	B	W	G	1	P	K	Y	U	N	F	Q	X	2	C	H	V	O	L	Z	A
T	G	F	2	O	X	Z	M	D	K	R	W	J	Y	C	U	P	H	A	3	S	L	E	1	N	4	B	Q	V	I
3	E	K	W	Q	T	N	Z	B	H	A	G	M	Y	S	C	U	I	1	O	V	J	2	P	D	X	4	F	R	L

Catalogue of designs

The ARCUS team created a catalogue of good designs, for between 20 and 100 varieties with 2 to 4 replications, which they distributed with the software.

They made the catalogue by assuming that good designs would have

- all treatments occurring the same number of times
- no pair of treatments occurring together in the same subblock more than once.

They found them by using a computer search over one particular method of construction.

Spatial heterogeneity

How should we allow for spatial heterogeneity within a site?

- Fixed effects of blocks and subblocks.
- Random effects of blocks and subblocks.
- Fixed or random effects of rows and columns.
- A fixed spatial trend,
  perhaps a low-dimensional polynomial in x and y.
- Spatial correlation between plots, where the correlation depends on both the x-distance and the y-distance.

Spatial heterogeneity

How should we allow for spatial heterogeneity within a site?

- Fixed effects of blocks and subblocks.
- Random effects of blocks and subblocks.
- Fixed or random effects of rows and columns.
- A fixed spatial trend,
  perhaps a low-dimensional polynomial in x and y.
- Spatial correlation between plots, where the correlation depends on both the x-distance and the y-distance.

This depends on the geography of the site,
its agricultural history over previous centuries,
and what response is being measured.

Row-column designs

The plots at a site usually form a rectangle.
Sometimes it is appropriate to use both rows and columns as (possibly incomplete) blocks.

Row-column designs

The plots at a site usually form a rectangle.
Sometimes it is appropriate to use both rows and columns as (possibly incomplete) blocks.

A statistician at DSIR in New Zealand advised that trials of cottton varieties should use row-column designs (and appropriate data analysis) rather than large complete blocks.

Row-column designs

The plots at a site usually form a rectangle.
Sometimes it is appropriate to use both rows and columns as (possibly incomplete) blocks.

A statistician at DSIR in New Zealand advised that trials of cottton varieties should use row-column designs (and appropriate data analysis) rather than large complete blocks.

This increased the precision so much that the trial replication was routinely reduced from 3 to 2.

(II) Early generation variety trials

Before new varieties get into national variety trials, they are developed by plant breeders. A typical early generation variety trial has over 100 new varieties and a few control varieties.

There is very little seed of the new varieties,
so about 20% of the plots are used for control varieties,
with the new varieties each allocated to a single plot.
The objective is to select a certain proportion (say 10%) of the new varieties for further development and breeding.

Thus comparisons with controls are not really important;
the control varieties are there to provide some estimate of the underlying spatial pattern.

Running example

There are 224 new varieties, with very little seed of each.

There are 280 plots available, in a 14×20 rectangle.

How do you design the experiment?

Simplest model

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \quad \text{if } \omega \neq \omega'.$$

Simplest model

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \quad \text{if } \omega \neq \omega'.$$

The design which minimizes

$$\sum_i \sum_{j \neq i} \mathrm{Var}(\hat{\tau}_i - \hat{\tau}_j),$$

over pairs of new varieties, has 2 plots for some new varieties, 1 plot for all other new varieties, and no controls.

Simplest model: example

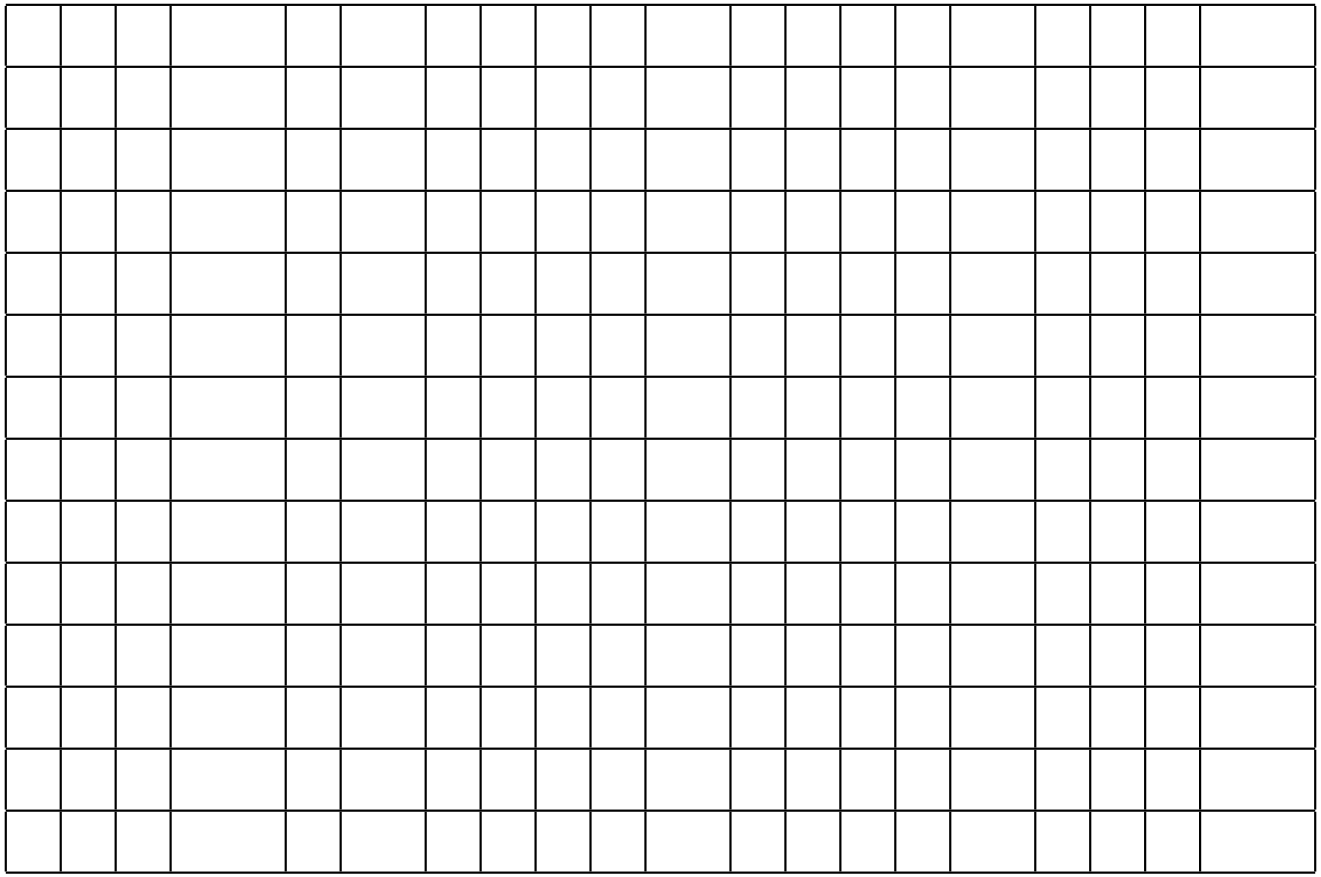
56 varieties have replication 2;

168 varieties have replication 1.

Simplest model: example

56 varieties have replication 2;

168 varieties have replication 1.



Simplest model: example

56 varieties have replication 2;

168 varieties have replication 1.

															1				
			2																
					1														
										2									

Simplest model: example

56 varieties have replication 2;

168 varieties have replication 1.

															1				
			2																
																			57
					1														
			58																
										2									

A breeder says . . .

Unfair!

The single plot with my variety was in an infertile part of the field.

Fixed spatial trend

$$Y_\omega = \tau_{f(\omega)} + g(\omega) + \epsilon_\omega$$

where

$g(\omega)$ is a two-dimensional low-degree polynomial in ω,

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \text{ if } \omega \neq \omega'.$$

Fixed spatial trend

$$Y_\omega = \tau_{f(\omega)} + g(\omega) + \epsilon_\omega$$

where

$g(\omega)$ is a two-dimensional low-degree polynomial in ω,

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \text{ if } \omega \neq \omega'.$$

Caliński, Mejza, ...:

Fixed spatial trend

$$Y_\omega = \tau_{f(\omega)} + g(\omega) + \epsilon_\omega$$

where

$g(\omega)$ is a two-dimensional low-degree polynomial in ω,

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \text{ if } \omega \neq \omega'.$$

Caliński, Mejza, ...:
use one plot for each new variety

Fixed spatial trend

$$Y_\omega = \tau_{f(\omega)} + g(\omega) + \epsilon_\omega$$

where

$g(\omega)$ is a two-dimensional low-degree polynomial in ω,

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \text{ if } \omega \neq \omega'.$$

Caliński, Mejza, . . . :
use one plot for each new variety
and several plots for a well-established but uninteresting
"control";

Fixed spatial trend

$$Y_\omega = \tau_{f(\omega)} + g(\omega) + \epsilon_\omega$$

where

$g(\omega)$ is a two-dimensional low-degree polynomial in ω,

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \text{ if } \omega \neq \omega'.$$

Caliński, Mejza, . . . :
use one plot for each new variety
and several plots for a well-established but uninteresting "control";
place the "control" plots in a grid;

Fixed spatial trend

$$Y_\omega = \tau_{f(\omega)} + g(\omega) + \epsilon_\omega$$

where

$g(\omega)$ is a two-dimensional low-degree polynomial in ω,

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \text{ if } \omega \neq \omega'.$$

Caliński, Mejza, . . . :
use one plot for each new variety
and several plots for a well-established but uninteresting "control";
place the "control" plots in a grid;
use the "control" responses to estimate the polynomial trend;

Fixed spatial trend

$$Y_\omega = \tau_{f(\omega)} + g(\omega) + \epsilon_\omega$$

where

$g(\omega)$ is a two-dimensional low-degree polynomial in ω,

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \text{ if } \omega \neq \omega'.$$

Caliński, Mejza, . . . :
use one plot for each new variety
and several plots for a well-established but uninteresting "control";
place the "control" plots in a grid;
use the "control" responses to estimate the polynomial trend;
estimate each variety effect by subtracting the trend value from its response.

Spatial trend: example

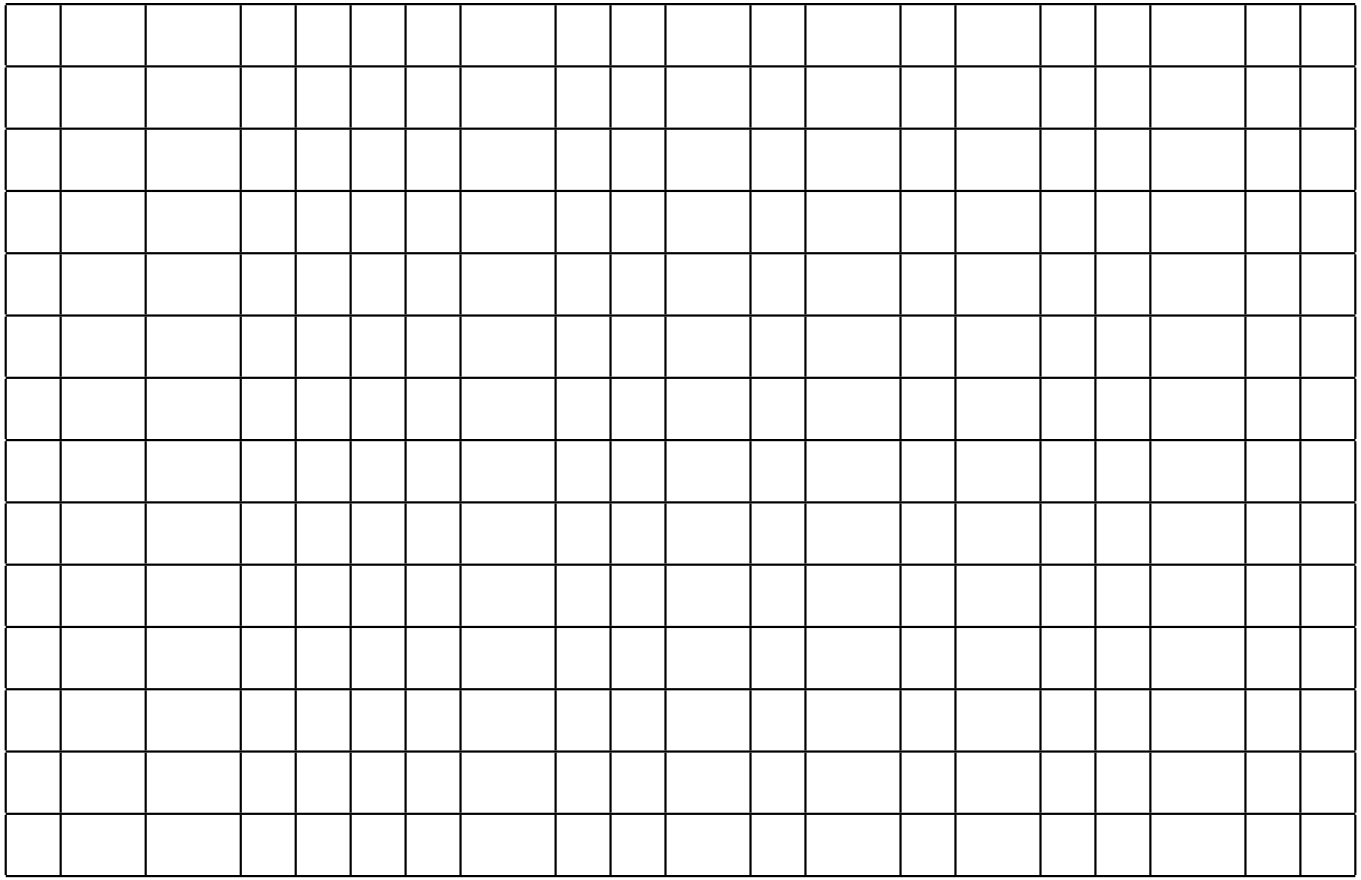
56 plots for "control"

224 new varieties have replication 1.

Spatial trend: example

56 plots for "control"

224 new varieties have replication 1.



Spatial trend: example

56 plots for "control"

224 new varieties have replication 1.

		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		

Spatial trend: example

56 plots for "control"

224 new varieties have replication 1.

		X					X					X					X		
		X					X			3		X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
	2	X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X		1			X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		

Spatial trend: example

56 plots for "control"

224 new varieties have replication 1.

		X					X					X					X		
		X					X			3		X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
	2	X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X		1			X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		
		X					X					X					X		

Controls are on every fifth plot, working along rows.

Spatial trend: example, another layout

56 plots for "control"

224 new varieties have replication 1.

Spatial trend: example, another layout

56 plots for "control"

224 new varieties have replication 1.

	X							X			X							X	
				X	X									X	X				
X									X	X									X
			X			X							X			X			
		X					X					X					X		
	X							X			X							X	
				X	X									X	X				
X									X	X									X
			X			X							X			X			
		X					X					X					X		
	X							X			X							X	
				X	X									X	X				
X									X	X									X
			X			X							X			X			

Controls are on every 5th plot, working boustrophedon along columns.

Spatial trend: example, a third layout

56 plots for "control"

224 new varieties have replication 1.

Spatial trend: example, a third layout

56 plots for "control"

224 new varieties have replication 1.

	X		X			X		X			X		X			X		X	
	X		X			X		X			X		X			X		X	
	X		X			X		X			X		X			X		X	
	X		X			X		X			X		X			X		X	
	X		X			X		X			X		X			X		X	
	X		X			X		X			X		X			X		X	
	X		X			X		X			X		X			X		X	

Controls are on a rectangular grid.

Spatial trend: example, what should we optimize?

56 plots for "control"

224 new varieties have replication 1.

Spatial trend: example, what should we optimize?

56 plots for "control"

224 new varieties have replication 1.

X	X				X	X							X	X				X	X
X	X				X	X							X	X				X	X
X	X																	X	X
X	X				X									X				X	X
									X	X									
									X	X									
X	X				X									X				X	X
X	X																	X	X
X	X				X	X							X	X				X	X
X	X				X	X							X	X				X	X

Controls are positioned to make the average variance of prediction small if the trend is a polynomial of degree three.

Spatial trend: example, what should we optimize/assume?

56 plots for "control"

224 new varieties have replication 1.

Spatial trend: example, what should we optimize/assume?

56 plots for "control"

224 new varieties have replication 1.

X	X	X						X	X	X	X						X	X	X
X	X	X															X	X	X
X																			X
X																			X
X									X	X									X
X									X	X									X
X									X	X									X
X									X	X									X
X																			X
X																			X
X	X	X															X	X	X
X	X	X						X	X	X	X						X	X	X

Controls are positioned to make the maximum variance of prediction small if the trend is a polynomial of degree two.

Assumptions about spatial trend

Slightly different assumptions about the exact form of the spatial trend
and slightly different criteria to optimize
can lead to very different designs.

Thanks to Bradley Jones,
who found these optimal designs by computer search.

Spatial trend: controls as covariates

Yates (1936), Atiqullah and Cox (1962) consider controls spread throughout the field. In their analysis, a weighted mean of the response on the nearest controls is used as a covariate, rather than being simply subtracted.

Spatial correlation

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega) = 0, \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

and $\mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'})$ depends on the spatial relationship between ω and ω'.

Spatial correlation

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega) = 0, \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

and $\mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'})$ depends on the spatial relationship between ω and ω'.

Kempton, Talbot, Besag, Martin, Eccleston, ...:

Spatial correlation

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega) = 0, \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

and $\mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'})$ depends on the spatial relationship between ω and ω'.

Kempton, Talbot, Besag, Martin, Eccleston, ...:
use one plot for each new variety and several plots for "control";

Spatial correlation

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega) = 0, \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

and $\mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'})$ depends on the spatial relationship between ω and ω'.

Kempton, Talbot, Besag, Martin, Eccleston, ...:
use one plot for each new variety and several plots for "control";
place the "control" plots in some kind of grid;

Spatial correlation

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega) = 0, \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

and $\mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'})$ depends on the spatial relationship between ω and ω'.

Kempton, Talbot, Besag, Martin, Eccleston, ...:
use one plot for each new variety and several plots for "control";
place the "control" plots in some kind of grid;
analyse the data with GLS or REML.

Spatial correlation

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega) = 0, \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

and $\mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'})$ depends on the spatial relationship between ω and ω'.

Kempton, Talbot, Besag, Martin, Eccleston, ...:
use one plot for each new variety and several plots for "control";
place the "control" plots in some kind of grid;
analyse the data with GLS or REML.

Cullis, Smith, Lim, Gilmour, Butler, Coombes, ...:

Spatial correlation

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega) = 0, \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

and $\mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'})$ depends on the spatial relationship between ω and ω'.

Kempton, Talbot, Besag, Martin, Eccleston, ...:
use one plot for each new variety and several plots for "control";
place the "control" plots in some kind of grid;
analyse the data with GLS or REML.

Cullis, Smith, Lim, Gilmour, Butler, Coombes, ...:
use 2 plots for some varieties and 1 plot for all other varieties,

Spatial correlation

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega) = 0, \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

and $\mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'})$ depends on the spatial relationship between ω and ω'.

Kempton, Talbot, Besag, Martin, Eccleston, ...:
use one plot for each new variety and several plots for "control";
place the "control" plots in some kind of grid;
analyse the data with GLS or REML.

Cullis, Smith, Lim, Gilmour, Butler, Coombes, ...:
use 2 plots for some varieties and 1 plot for all other varieties,
optimize the design by computer search,

Spatial correlation

$$Y_\omega = \tau_{f(\omega)} + \epsilon_\omega$$

where

$$E(\epsilon_\omega) = 0, \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

and $\mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'})$ depends on the spatial relationship between ω and ω'.

Kempton, Talbot, Besag, Martin, Eccleston, ...:
use one plot for each new variety and several plots for "control";
place the "control" plots in some kind of grid;
analyse the data with GLS or REML.

Cullis, Smith, Lim, Gilmour, Butler, Coombes, ...:
use 2 plots for some varieties and 1 plot for all other varieties,
optimize the design by computer search,
analyse the data with GLS or REML.

Blocks

The field is partitioned into homogeneous blocks.
(One block has all the stony plots; one block has all the plots near the rabbit warren, ...)

Blocks

The field is partitioned into homogeneous blocks.
(One block has all the stony plots; one block has all the plots near the rabbit warren, ...)

$$Y_\omega = \tau_{f(\omega)} + \beta_{h(\omega)} + \epsilon_\omega$$

where

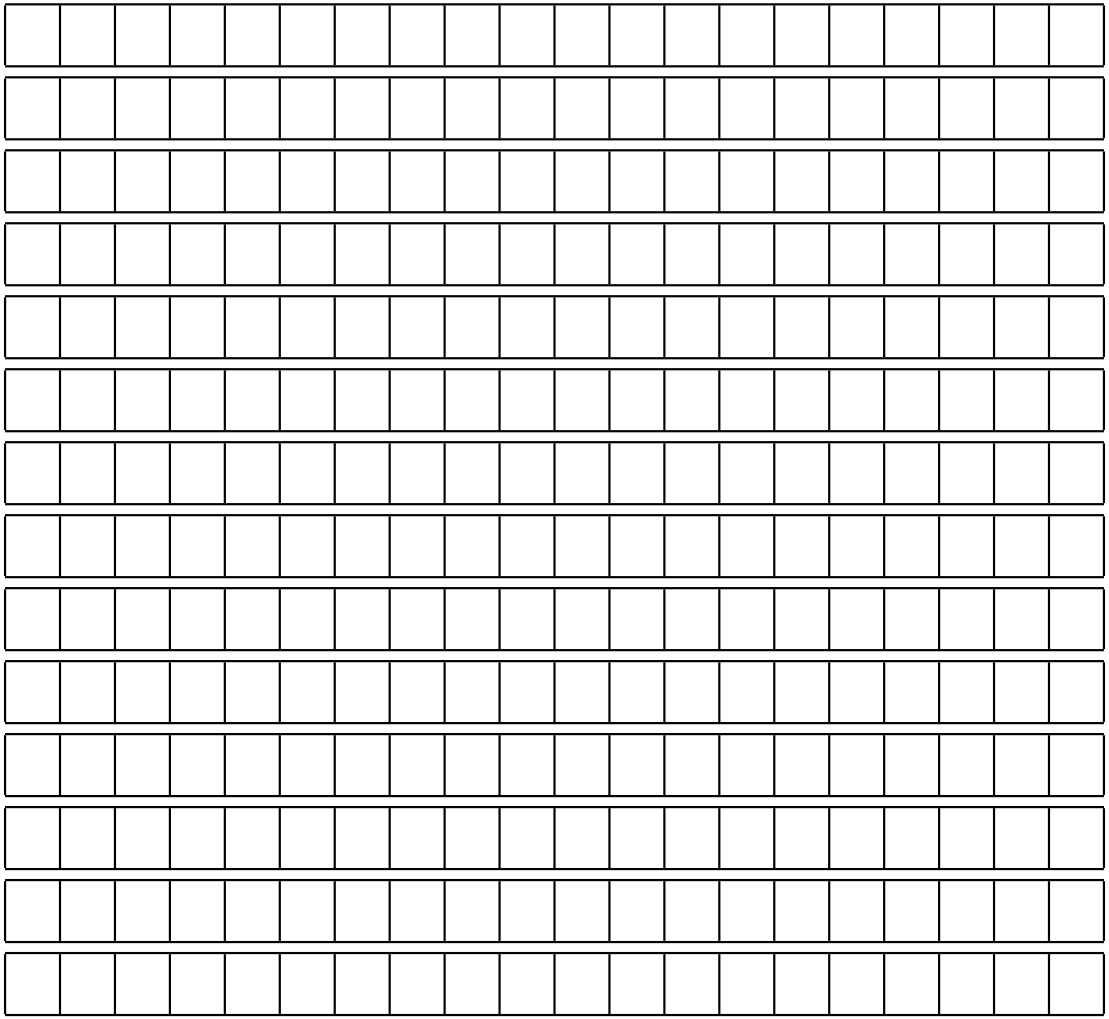
$$h(\omega) = \text{block containing } \omega,$$

$$E(\epsilon_\omega = 0), \qquad \mathrm{Var}(\epsilon_\omega) = \sigma^2,$$

$$\text{and} \quad \mathrm{Cov}(\epsilon_\omega, \epsilon_{\omega'}) = 0 \text{ if } \omega \neq \omega'.$$

Blocks: example

Rows are blocks, so there are 14 blocks, each with 20 plots.



Blocks: example, continued

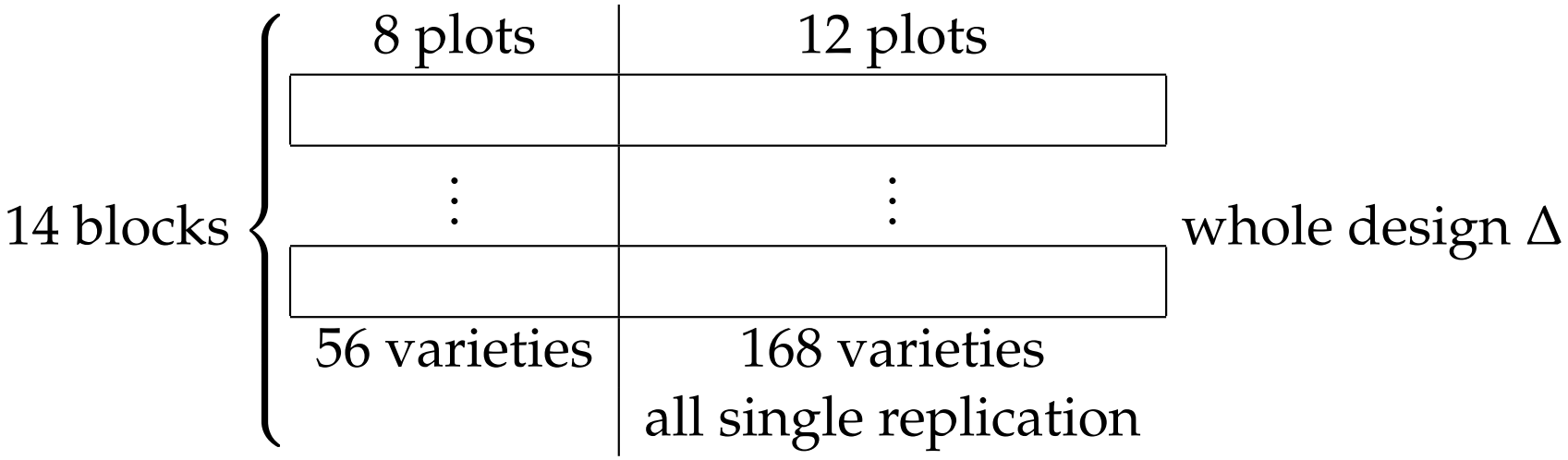
224 varieties in 14 blocks of size 20.

Blocks: example, continued

224 varieties in 14 blocks of size 20.
($280 - 224 = 56$ and $224 - 56 = 168$,
so at least 168 varieties must have single replication.)

Blocks: example, continued

224 varieties in 14 blocks of size 20.
($280 - 224 = 56$ and $224 - 56 = 168$,
so at least 168 varieties must have single replication.)



Subdesign Γ has 56 varieties
in 14 blocks of size 8.

Which is best? (224 varieties in 14 blocks of size 20)

Subdesign	Single replicate
8 plots 56 varieties with replication 2	168 varieties using 12 plots per block

Which is best? (224 varieties in 14 blocks of size 20)

Subdesign	Single replicate
8 plots 56 varieties with replication 2	168 varieties using 12 plots per block
7 plots 28 varieties with replication 2 14 varieties with replication 3	182 varieties using 13 plots per block
6 plots 28 varieties with replication 3	196 varieties using 14 plots per block

Which is best? (224 varieties in 14 blocks of size 20)

Subdesign	Single replicate
8 plots 56 varieties with replication 2	168 varieties using 12 plots per block
7 plots 28 varieties with replication 2 14 varieties with replication 3	182 varieties using 13 plots per block
6 plots 28 varieties with replication 3	196 varieties using 14 plots per block
5 plots 14 varieties with replication 5	210 varieties using 15 plots per block
4 plots 4 controls in every block	224 varieties using 16 plots per block

Which is best? (224 varieties in 14 blocks of size 20)

Subdesign	Single replicate
8 plots 56 varieties with replication 2	168 varieties using 12 plots per block
7 plots 28 varieties with replication 2 14 varieties with replication 3	182 varieties using 13 plots per block
6 plots 28 varieties with replication 3	196 varieties using 14 plots per block
5 plots 14 varieties with replication 5	210 varieties using 15 plots per block
4 plots 4 controls in every block	224 varieties using 16 plots per block

Recent work by RAB shows that, as the number of varieties or the number of blocks increases, there are phase changes in the best subdesign: starting with as many new varieties as possible with replication 2, these numbers go down and up respectively, until the subdesign consists entirely of controls.

Conclusion

The design of experiments for large numbers of varieties with very small average replication is still challenging.

Conclusion

The design of experiments for large numbers of varieties
with very small average replication is still challenging.

Large amounts of computer storage,
easy access to the stored data,
availability of software to undertake complicated analyses, ...
do not remove the need to design these experiments well.

Conclusion

The design of experiments for large numbers of varieties
with very small average replication is still challenging.

Large amounts of computer storage,
easy access to the stored data,
availability of software to undertake complicated analyses, ...
do not remove the need to design these experiments well.

Do not rush to collect big data without planning.