

A composite sampling strategy for the design and analysis of cereal resistance trials

Bev Gogel, The University of Adelaide
Alison Smith, University of Wollongong
Brian Cullis, University of Wollongong and CSIRO

Statistics for the Australian Grains Industry (SAGI) II

Alan McKay, Molecular Diagnostics Group, SARDI

Australasian Applied Statistics Conference
July 2011



Will cover

- Background biology
- Resistance trials we've been dealing with
 - and associated issues
- Composite sampling strategy
 - to deal with issues
- CCN resistance trial example
 - that we've applied strategy to
- How composite sampling performs in practice
- Some points / concluding remarks

Background

Cereal Cyst Nematodes (CCN)

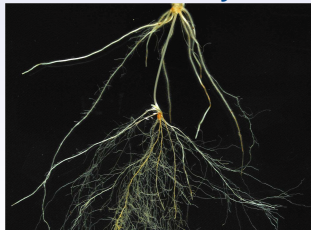
- microscopic worms
- lie dormant in soil over summer
- with onset of cold weather (drop to $< 14^{\circ}$) + moisture (rain events) ...
 - hatch and invade developing roots of seedlings
 - stimulate lateral roots which form knots
 - stop roots developing properly
- losses caused by CCN in SE Australia in wheat *alone* recently estimated at \$98 million each year

Background

Cereal Cyst Nematodes (CCN)

- microscopic worms
- lie dormant in soil over summer
- with onset of cold weather (drop to $< 14^{\circ}$) + moisture (rain events) ...
 - hatch and invade developing roots of seedlings
 - stimulate lateral roots which form knots
 - stop roots developing properly
- losses caused by CCN in SE Australia in wheat *alone* recently estimated at \$98 million each year

unaffected root system

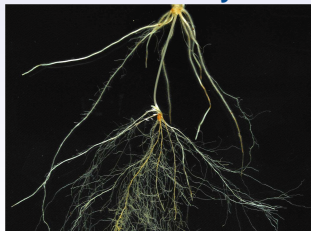


Background

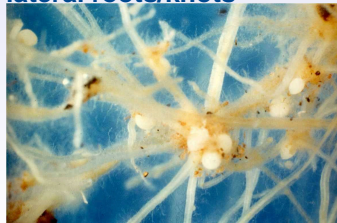
Cereal Cyst Nematodes (CCN)

- microscopic worms
- lie dormant in soil over summer
- with onset of cold weather (drop to $< 14^{\circ}$) + moisture (rain events) ...
 - hatch and invade developing roots of seedlings
 - stimulate lateral roots which form knots
 - stop roots developing properly
- losses caused by CCN in SE Australia in wheat *alone* recently estimated at \$98 million each year

unaffected root system



root system affected by CCN-
lateral roots/knots



Molecular Diagnostics Group, SARDI, Alan McKay

- contracted to provide information to NVT and plant breeders, on resistance and tolerance of varieties to CCN and other root lesion nematodes
 - resistance genes: affect ability of nematode to *multiply*
 - tolerance genes: affect ability of plant to yield when nematodes *attacking root system*
- do this through controlled, 3-phase
glasshouse-DNA storage-DNA extraction tests
 - run since 1970's *with improvements over time*
- now running field trials to
 - assess resistance/tolerance lines in field
 - using current farming systems/practices

CCN resistance trials

Resistance trials (focus here)

- designed as standard field trials
 - laid out in row \times range array of plots
 - reps aligned in ranges, say
- nematode counts taken
 - before *or as soon as possible after* seeding
 - immediately following harvest
- we then conduct linear mixed model analysis of \log_e -transformed final counts, that includes
 - blocking terms for randomisation
 - extraneous terms (linear row, range ...) where necessary
 - spatially correlated errors
 - we consider relationship between final and initial counts across varieties in random coefficient regression framework

... seems straightforward ... but

CCN resistance trials

... process to measure nematode levels in a single plot involves

- taking 20 soil cores and combining into single soil sample
 - intensive time consuming process
- processing sample in lab
 - BIG \$'s
- for every trial in 2010
 - sampled and processed every plot!



Molecular Diagnostics have therefore come to us

- not only to have data analysed
- but also seeking ways to increase efficiency in current experimental methods (both lab and field), for example,
 - reducing replicates per variety (currently 5-6)
 - reducing number of multi-phase tests run
 - in statistically valid manner
- because by reducing \$ cost per variety \implies they can ultimately test more varieties

Fortunately able to offer cost savings through *composite sampling strategy*

- work of Alison, Robin, Dave and Brian in Smith *et al.* (2011)
- offers compromise between extremes of
 - compositing all replicates for every variety
 - ▷ typical in grain quality testing
 - ▷ some traits expensive to measure
 - ▷ motivation for Smith paper
 - scoring each individual plot for each variety
 - ▷ has been practice of Molecular Diagnostics to date for resistance and tolerance trials
 - ▷ where we've been able to apply compositing strategies in paper

Sampling strategies

Smith paper presents 2 sampling strategies

subset selection

- pre-chose a subset of varieties ($\frac{1}{3} - \frac{1}{2}$ say)
 - lines of interest costly trait
- sample individual plots for this subset
- sample single rep of rest of lines
- analyse this data

1	Young	Young	Young	Young
2	Blue	Blue	Blue	Blue
3	Blue	Blue	Blue	Blue
4	Young	Blue	Blue	Blue
5	Young	Blue	Blue	Blue
6	Young	Blue	Blue	Blue
7	Blue	Blue	Blue	Blue
8	Blue	Blue	Blue	Blue
9	Blue	Blue	Blue	Blue
10	Blue	Blue	Blue	Blue
11	Blue	Blue	Blue	Blue
12	Blue	Blue	Blue	Blue
13	Blue	Blue	Blue	Blue

Sampling strategies

Smith paper presents 2 sampling strategies

subset selection

- pre-chose a subset of varieties ($\frac{1}{3} - \frac{1}{2}$ say)
 - lines of interest costly trait
- sample individual plots for this subset
- sample single rep of rest of lines
- analyse this data

Row	Day 1	Day 2	Day 3	Day 4
1	Young	Young	Young	Young
2	Blue	Blue	Blue	Blue
3	Blue	Blue	Blue	Blue
4	Blue	Blue	Blue	Blue
5	Blue	Blue	Blue	Blue
6	Blue	Blue	Blue	Blue
7	Blue	Blue	Blue	Blue
8	Blue	Blue	Blue	Blue
9	Blue	Blue	Blue	Blue
10	Blue	Blue	Blue	Blue
11	Blue	Blue	Blue	Blue
12	Blue	Blue	Blue	Blue
13	Blue	Blue	Blue	Blue

composite sampling

- pre-chose a subset of varieties
- sample individual plots for this subset
- combine soil samples across plots for remaining varieties and process these *composite* plots
- analyse this data

Sampling strategies

Smith paper presents 2 sampling strategies

subset selection

- pre-chose a subset of varieties ($\frac{1}{3} - \frac{1}{2}$ say)
 - lines of interest costly trait
- sample individual plots for this subset
- sample single rep of rest of lines
- analyse this data

Row	Day 1	Day 2	Day 3	Day 4
1	Variety 1	Variety 2	Variety 3	Variety 4
2	Variety 1	Variety 2	Variety 3	Variety 4
3	Variety 1	Variety 2	Variety 3	Variety 4
4	Variety 1	Variety 2	Variety 3	Variety 4
5	Variety 1	Variety 2	Variety 3	Variety 4
6	Variety 1	Variety 2	Variety 3	Variety 4
7	Variety 1	Variety 2	Variety 3	Variety 4
8	Variety 1	Variety 2	Variety 3	Variety 4
9	Variety 1	Variety 2	Variety 3	Variety 4
10	Variety 1	Variety 2	Variety 3	Variety 4
11	Variety 1	Variety 2	Variety 3	Variety 4
12	Variety 1	Variety 2	Variety 3	Variety 4
13	Variety 1	Variety 2	Variety 3	Variety 4

composite sampling

- pre-chose a subset of varieties
- sample individual plots for this subset
- combine soil samples across plots for remaining varieties and process these *composite* plots
- analyse this data

*proportion of complete set of plots
NOT used in recording data*

*ALL plots utilised in
measuring ALL traits*

SA CCN resistance trial: Molecular Diagnostics

Have applied compositing for first time

- full design
 - 4 reps, 26 entries, 13 row \times 8 bay array
 - 104 plots in total
 - reps in pairs of bays: 1&2, 3&4, 5&6, 7&8

1	Young	Conal	Olson	Waller	Pauline	Olson	Sharon	Olson
2	Ryan	Wade	Charles	Arnold	Texas	Northman	Conal	Waller
3	Olson	Olson	Olson	Olson	Olson	Olson	Olson	Olson
4	Arnold	Olson	Olson	Olson	Olson	Olson	Olson	Northman
5	Northman	Olson	Northman	Sharon	Olson	Olson	Arnold	Olson
6	Texas	Olson	Olson	Olson	Olson	Olson	Olson	Olson
7	Sharon	Northman	Olson	Olson	Waller	Olson	Olson	Olson
8	Waller	Olson	Conal	Olson	Olson	Olson	Olson	Olson
9	Olson	Olson	Olson	Olson	Olson	Olson	Olson	Olson
10	Olson	Olson	Olson	Olson	Olson	Olson	Olson	Olson
11	Olson	Olson	Olson	Olson	Olson	Olson	Olson	Olson
12	Olson	Olson	Olson	Olson	Olson	Olson	Olson	Olson
13	Olson	Olson	Olson	Olson	Olson	Olson	Olson	Olson
	1	2	3	4	5	6	7	8

SA CCN resistance trial: Molecular Diagnostics

Have applied compositing for first time

- full design
 - 4 reps, 26 entries, 13 row \times 8 bay array
 - 104 plots in total
 - reps in pairs of bays: 1&2, 3&4, 5&6, 7&8
- overlaid design with composite sampling strategy in which
 - full trial split into 2 subtrials: reps 1&2, 3&4
 - variety list split into 2 sublists: size 14 (A), 12 (B), such that
 - ▷ plots for A will be individually sampled in subtrial 1 (magenta plots)
 - ▷ plots for B composited (white plots)
 - ▷ plots for B will be individually sampled in subtrial 2 (magenta plots)
 - ▷ plots for A composited (white plots)
- effect of compositing to
 - reduce samples to process/variety from 4 \downarrow 3

1	Young	Conrad	Ellen	Widdell	Charles	Ernest	Shannon	Max
2	Rayn	Wend	Charles	Arnolds	Texas	Northman	Conrad	Widdell
3	Olson	Rayn	Ellen	Young	Charles	Arnolds	Rayn	Wend
4	Arnolds	John	Arnolds	Ernest	Paul	Rayn	Ernest	Northman
5	Northman	Paul	Northman	Shannon	Ernest	Charles	Arnolds	Paul
6	Texas	Paul	Paul	Paul	Sherry	Shannon	Paul	Ernest
7	Shannon	Northman	Yip	Charles	Widdell	Paul	Charles	John
8	Widdell	Charles	Conrad	Ernest	John	Yip	Texas	Charles
9	Ernest	Sherry	Ernest	Charles	Conrad	Ernest	Yip	Charles
10	Ernest	Arnolds	Ernest	Charles	Charles	Northman	Sherry	Paul
11	Paul	Shannon	John	Sherry	Charles	Ernest	Northman	Young
12	Yip	Charles	Northman	Young	Ernest	Charles	Charles	Arnolds
13	Charles	Charles	Paul	Ernest	Paul	Young	Shannon	Ernest
	1	2	3	4	5	6	7	8

1	Young	Conrad	Ellen	Widdell	Charles	Ernest	Shannon	Max
2	Rayn	Wend	Charles	Arnolds	Texas	Northman	Conrad	Widdell
3	Olson	Rayn	Ellen	Young	Charles	Arnolds	Rayn	Wend
4	Arnolds	John	Arnolds	Ernest	Paul	Rayn	Ernest	Northman
5	Northman	Paul	Northman	Shannon	Ernest	Charles	Arnolds	Paul
6	Texas	Paul	Paul	Paul	Sherry	Shannon	Paul	Ernest
7	Shannon	Northman	Yip	Charles	Widdell	Paul	Charles	John
8	Widdell	Charles	Conrad	Ernest	John	Yip	Texas	Charles
9	Ernest	Sherry	Ernest	Charles	Conrad	Ernest	Yip	Charles
10	Ernest	Arnolds	Ernest	Charles	Charles	Northman	Sherry	Paul
11	Paul	Shannon	John	Sherry	Charles	Ernest	Northman	Young
12	Yip	Charles	Northman	Young	Ernest	Charles	Charles	Arnolds
13	Charles	Charles	Paul	Ernest	Paul	Young	Shannon	Ernest
	1	2	3	4	5	6	7	8

SA CCN resistance trial: Molecular Diagnostics

some points

- CCN trial designed as 2 embedded partially replicated designs
 - 2 reps vars in A, 1 rep of B *restricted to rows 1-10* subtrial 1
 - 2 reps vars in B, 1 rep of A *restricted to rows 4-13* subtrial 2
- jury out as to whether
 - embedded p -rep design more optimal than
 - random allocation of varieties throughout all 13 rows within reps within subtrials
- need more research into this

1	Young	Levon	Yield	Yield	Shelby	Shelby	Shelby	Shelby	Shelby	Shelby	Shelby	Shelby
2	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
3	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
4	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
5	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
6	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
7	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
8	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
9	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
10	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
11	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
12	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield
13	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield	Yield

How does the analysis work?

For analysis of complete field trial data, we'd typically

- conduct linear mixed model analysis of trait

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e}$$

- $\boldsymbol{\tau}$ = fixed effects with design matrix \mathbf{X} ,
- \mathbf{u}_g = random variety effects with design matrix \mathbf{Z}_g ,
- \mathbf{u}_p = peripheral effects with design matrix \mathbf{Z}_p ,
- \mathbf{e} = spatially correlated errors

How does the analysis work?

For compositing have

- data for transformation of original experimental units down to the number of samples
- 104 plots \Downarrow 78 samples
 - 52 for the single plots of 2 reps for each variety
 - 26 for the composite of 2 reps for each variety
- essentially analyse

$$\begin{aligned}z &= Dy \\ &= DX\tau + DZ_g u_g + DZ_p u_p + De\end{aligned}$$

- D = transformation or *compositing* matrix (no. samples \times no. plots = 78×104)
- effectively maps to
 - ▷ original plot data for lines where individual plots scored
 - ▷ average over plots for lines whose plots have been composited

How does composite sampling perform in practice?

Smith paper presents results of simulation study

- consider 2 design types \times 8 methods for obtaining predicted variety effects
 - context of canola MET
 - methods centred about
 - ▷ the *full* dataset
 - ▷ a *subset* of the complete dataset
 - ▷ a *mixture* of composite and individual plots
- calculate correlation between true and predicted variety effects for each method \times design type

How does composite sampling perform in practice?

Results of simulation

- bias
 - no significant bias for any parameters for any method
 - ▷ including compositing method
 - ▷ first time this result reported in literature
- precision of variance parameter estimates
 - standard errors always lower for compositing methods *cf* subsetting (reduced samples)
- precision of spatial parameter estimates
 - converse true
 - due to smearing when plots combined
- correlation between true and predicted (focus)
 - compositing methods performed substantially better than subsetting methods

Some results: SA CCN resistance data

- initial nematode counts have just come in
 - while expected composite data
 - they ended up sampling *all* plots just after sowing
- been able to compare analysis of full dataset... with ASReml-R code

```
ccnfull.asr <- asreml(lccninit ~ 1, random = ~ Repl + Range,  
rcov = ~ ar1(Column):ar1(Row), data=ccnfull.df1)
```

- to analysis of partially composited dataset ... with code

```
pcompccn.asr <- asreml(z ~ 1,  
random=~ grp('Repl') + grp('Range') + str( grp('Plot'), ~ ar1v(8):ar1(13)),  
family=asreml.gaussian(dispersion=.0001),data=ccnpc.df,maxit=30,  
R.param=gamms.df,G.param=gamms.df,  
control=asreml.control(group=list(Repl=bpos,Range=cpos,Plot=ppos)))
```

Some results: SA CCN resistance data

- variance and spatial parameter estimates

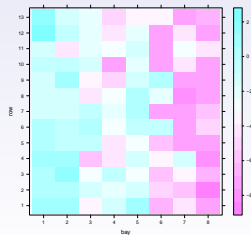
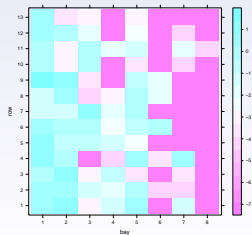
term	full	partially composited
replicate	3.72	4.95
range	2.47	1.44
error	3.91	4.02
spatial		
range	0.08	-0.01
row	0.17	0.16

- pleasingly!* all estimates very similar



Some results: SA CCN resistance data

- level plots of predicted initial counts for each plot in full field layout based on analyses of *full* and *composite* datasets
- compositing has done pretty well in mapping distribution of initial counts
 - while primary application for us
 - ▷ accurately estimating variety effects
 - example demonstrates a second application of compositing
 - ▷ *mapping the distribution of a trait*
 - heavy metal example, Alison



Some points and concluding remarks

- composite sampling relatively straightforward for 2 rep trials
- as number of reps increases so do possibilities for compositing
- raises important questions in terms of optimum design and sampling schemes
- further research needed in these areas

Smith, A. B, Thompson, R., Butler, D.G. and Cullis, B. R. *The design and analysis of variety trials using mixtures of composite and individual plot samples*. Applied Statistics (2011) **60**, Part 3, pp. 437-455