Spatial Analysis of field Experiments

Arthur Gilmour Arthur.Gilmour@cargovale.com.au December 2020

Introduction

These notes pertain to ASReml and Echidna.

Field trials are typically laid out in a rectangular grid of spatially correlated plots. **Spatial analysis** has been used for 30 years to better predict genotype performance. We will consider combining information from multiple trials after discussing the analysis of a single experiment.

Background

A **uniformly trial** is where a field is divided into grid of separate plots. Each plot is sown with the same treatment/variety but we observe that at harvest, they do not all give the same yield. The differences in yield may be completely irregular but often we see a pattern; a clump of plots may be higher, or lower than the others. Or there may appear to be rows or columns of higher/lower yield. These differences have many potential sources. The soil type may change over the field, the moisture content may differ due to ponding or soil depth, nutrient levels may differ because of past events (a track across the field, a big tree was removed, or there was a fire), an insect infestation, a windstorm, our equipment may have been faulty (a drill line blocked in the combine) or our driving crooked. Some of these sources of variation have a pattern to them, some may be considered intrinsic and some have been caused by what we have done.

In this context, we want to see whether some treatments/varieties/genotypes/lines (we use the terms equivalently from a design point of view) yield better than others, so we plant different varieties in the different plots. How can we tell which differences are due to variety and which are due to all the other things that cause plots to give different yields.

The first strategy for this problem was to replicate that plots, to have several plots of each treatment. If they are distributed at random, then we can measure the variance among the plots of the same treatment compared with the variance between the plots of different treatments and if the latter is greater than the former, then conclude there is variance attributable to treatments. The theory validating this approach depends on treatments being assigned to plots at random.

We have noted there is a pattern to some sources of variation. Our analysis may be more efficient of it can accommodate this variation. Consequently, plant breeders have developed a series of different models for analysis of field trials.

Unreplicated design where each treatment is applied to a single plot. Then there is no way to distinguish intrinsic variation from treatment variation. Consequently, any experiment needs to have some if not all treatments replicated. Then, a simple analysis of variance will compare the differences (variance) between treatments with the differences (variance) between plots with the same treatment.

Completely Random design is when the replicates are randomly arranged across the field. However, we have noted that some of the field variation appears as patches, or in blocks, and it may be better if we could separate out the variation due to blocks. So, **randomised complete block** designs were devised

but these have two issues. If there are many treatments, the blocks will be large and typically we do not know *a priori* where the patches in the field will appear, so the blocks may not coincide at all with these patches. Therefore, researchers developed **incomplete block designs** and **balanced incomplete block designs.** Here the blocks are smaller, but it was realised that while better, more was needed.

This lead to the development of **neighour** designs. Really this is changing the focus from the planned layout of the experiment (the design) to the analysis using the principle that neighbouring plots are likely to be more similar. Various approaches were tried, but the endpoint has been the wide adoption of what is commonly called (in the literature of field experimentation) **spatial analysis.**

A single trial

Tessa wrote:

"A few years ago, I attended the course "Genetic analysis using ASReml4.0" at Wageningen University. In this course, you gave a few examples on how to conduct spatial analysis. Recently, I started working for Rijk Zwaan where it is quite common to perform spatial corrections, which was not the case in my previous work. Hence, I try to gain more knowledge on this topic. I ran several analyses with ASReml and I have a few questions which I hope you can answer.

Assume the following model:

Y ~ mu !r genotype
residual ar1v(row):ar1(col)

In the .asr file, I will see the following results:

Row AR_R
Row AR_V
Col AR_R

I have difficulties to understand what is exactly produced in the .asr output. Is it two spatial correlation coefficients, and a variance component?

Moreover, I mostly see that the spatial corrections is included in the residual part but sometimes people use the following model (e.g. when there are missing observations):

Y ~ mu !r genotype ar1(row):ar1(col)
residual units

Could you help me understand the differences in interpretation between both models and what, in your opinion, is the preferred method?

"

The context for this query is the spatial analysis of a field variety trial. Consider the SHF (Slate Hall Farm) example distributed with the programs. This trial was grown in 1976 according to the standard protocols of the day as part of a national plant breeding program. Here we have yields from six reps of 25 genotypes. The field layout was a balanced incomplete block design. The reps were contiguous in a 3 × 2 arrangement and within reps, the treatments were in a 5 × 5 grid so the total experiment has 15 rows and 10 columns.

The data file begins

```
Replicate,colblk,rowblk,variety,yield,row,column
1,1,1,1,1003,1,1
1,1,2,2,1356,2,1
1,1,3,4,1412,3,1
1,1,4,3,1239,4,1
1,1,5,5,1508,5,1
2,11,6,19,1967,6,1
2,11,7,23,1572,7,1
2,11,8,2,1969,8,1
2,11,9,6,1747,9,1
2,11,10,15,1598,10,1
3,21,11,18,1630,11,1
3,21,12,25,1633,12,1
3,21,13,9,1255,13,1
3,21,14,11,1277,14,1
3,21,15,2,1572,15,1
1,2,1,6,1531,1,2
1,2,2,7,1540,2,2
1,2,3,9,1250,3,2
1,2,4,8,1658,4,2
1,2,5,10,1185,5,2
…
```

And has been sorted rows (1:15) within columns (1:10). The first data field codes for the replicate (1:6), the second and third code for the incomplete blocks (1:30, 5 rows and 5 columns in each replicate), the fourth codes the variety (1:25), the fifth is the yield (the units are not given but may be lb/acre or kh/hectare. Typically, the varieties would be coded using actual names. Note that the row/column coding refer to actual field positions in contrast to the coding of replicate, column block, row block and variety where the coding is arbitrary.

With this coding, we can fit several spatial models which reflect different assumptions about the residuals.

Randomised Complete Block

For many analyses, ASReml and Echidna accept the same input. Echidna was written from scratch starting in 2017 to have basically the same functionality as ASReml and produce equivalent output. They need two files. The first is the data file (in this case the ASCII text file called shf.asd partly displayed above). The second is the command file.

The command file describes the data file to ASReml/Echidna so that the program interprets the data appropriately and then specifies the model to be fitted. It also is an ASCII text file with filename extension .as (.es). For a randomised Complete block analysis of this data, the command file might be

```
!WORK 2  !REN  !ARG  !OUT
TITLE: shf  !DOPART $1
 # Replicate,colblk,rowblk,variety,yield,row,column ...
 # 1,1,1,1,1003,1,1 ...
Replicate  *     # 1
colblk    *      # 1
rowblk     *    # 1
variety  *     # 1
yield          # 1003
row       *     # 1
column    *     # 1

shf.ASD !SKIP 1
yield ~ mu variety !r  Rep
residual units
```

Since the data file identifies the data fields with simple names on the first line, running ASReml/Echidna on the data file will produce a basic command file but it typically requires revision so we explain what is here.

On the first line, !WORK 2 !REN !ARG !OUT

!WORK 2	requests 2 Gbyte workspace (memory)
!REN !ARG	allows multiple models specified in the job to produce distinct output files based on the arguments of !ARG (if present, see later)
!OUT	puts the output files in separate folders

On the second line, TITLE: shf !DOPART $1

TITLE: shf is text used as a title for the run

!DOPART $1	picks up the 1st argument (after !ARG on the first line, if present) and processes the subsequent lines indicated by that argument using !PART statements

```
# Replicate,colblk,rowblk,variety,yield,row,column ...
# 1,1,1,1,1003,1,1 ...
```

are comment lines (indicated by #) displaying the top of the data file

```
Replicate  *     # 1
colblk    *      # 1
rowblk     *    # 1
variety  *     # 1
yield          # 1003
row       *     # 1
column    *     # 1
```

lists the 7 variables in the data in order using the names obtained from the data file (but the names used here do not need to literally match those used in the data file). All except *yield* are factor variables coded 1:*n* where n is the number of classes in the factor; *yield* is a simple variate. The distinction is indicated by the asterisk (*) beside the factor variable names (which is not present for *yield*).

```
 shf.ASD !SKIP 1
```

specifies the data file, and that the first line (the variable names) is to be skipped. It is distinguished from the preceding variable names by having a DOT (.) in the filename; DOT and most other non alphanumeric characters are not permitted in variable names.

These lines specify the data, how it is to be interpreted (factor or variate). The next 2 lines specify the model to be fitted.

```
yield ~ mu variety !r  Rep
residual units
```

specifies the fitting of a randomised complete block model using the variable names just defined and a few special codes/names

yield is the variable to be analysed

~ can be read as 'is modelled as' and separates the dependent variable from the 'independent' or prediction variables.

mu is a reserved word standing for the 'constant' or 'intercept' in the model

variety is the factor defined above by this name

!r says that following terms will be fitted as random (mu and variety are fitted as fixed effects in this case)

Rep is the Replicate factor defined above; a truncated form of the variable name is permitted provided no ambiguity ensues. Since no variance structure is specified for Rep, it is fitted as independent effects with common variance which is to be estimated.

residual at the start of a new line is a keyward indicating that the model for the residual follows.

units is a reserved word standing for a factor with a level for each observation. Since no variance structure is specified for Rep, it is fitted as independent effects with common variance which is to be estimated.

Since 'residual units' is the default residual structure, this line could have been omitted.

This command file was called shf.es and running Echidna on it puts several output files in the folder shf. These are all ascii text files which can be viewed an any basic text editor (popular ones are ConText, notepad, notepad++, emacs and vi). If run in ASReml, the output files have different file extensions and different layout but contain equivalent information.

.esr/.asr file

The .esr/.asr file contains run details (user, date and time, filenames), a data summary and the primary model fitting output.

```
Echidna 1.37   24 Nov 2020 Windows      2.7 Gbyte  at Wed Dec  2 15:50:13 2020
 Licensed to Arthur(Arthur@cargovale.com.au)
 TITLE: shf
 Folder: E:\MMX-II\Ex\Tests

 Data File: shf.ASD

 Summary of 150 data records

 Variable   Levels Miss Zero     Min     Max    Distribution or Mn SD Sk Kt
 Replicate      6    0    0       1       6   25 25 25 25 25 25
 colblk        30    0    0       1      30
 rowblk        30    0    0       1      30
 Variety       25    0    0       1      25
 yield          1    0    0  917.00  2119.00  1470.44   232.31   0.15  -0.16
```

```
row          15   0   0      1      15
column       10   0   0      1      10   15 15 15 15 15 15 15 15 15 15

Note: Using !DOPART 2
Note: Model is fitting 32 equations, DENSE portion has 26 equations.
* This job may use 4 processor threads. *
  1 LogL= -744.30  0.3633E+05      125 DF
  2 LogL= -743.80  0.3572E+05      125 DF
  3 LogL= -743.41  0.3502E+05      125 DF
  4 LogL= -743.34  0.3472E+05      125 DF
  5 LogL= -743.34  0.3467E+05      125 DF
  6 LogL= -743.34  0.3466E+05      125 DF

Akaike Information Criterion   1490.68 (assuming 2 parameters).
Bayesian Information Criterion 1496.34

      Analysis of yield

                Wald F statistics
Source of Variation        NumDF    DenDF    F-inc          P-inc
mu                             1          1216.29
Variety                       24             3.06

Model_Term              Order    Gamma       Sigma    Z_ratio  %C
Replicate                   6  0.267696     9279.60       1.38   0 P
Residual_units            150   1.00000      34664.6       7.75
 Replicate                      6 effects fitted.
Finished: Wed Dec  2 15:50:14 2020 LogL Converged     E:\MMX-II\Ex\Tests/shf2/shf
```

It is essential to check that the program has read the data properly. This is done by reviewing the data summary.

The LogL sequence shows the model has converged.

The Wald F table shows an F statistic of 3.06 for the Variety effects.

The variance components table reports a replicate variance component of 9280 and a residual variance of 34665.

The **.esx/.res** file contains iteration specific values for the variance parameters, a plot of residuals and often other information about the model fitted. A residual plot is also provided as a graphics file ready for inclusion in a report document.

The **.ess/.sln** file reports all the fitted effects in the model.

The **.esy/.yht** file contains fitted values and residuals for each data point.

Balanced Incomplete block analysis

This experiment was designed as a balanced incomplete block and that model can be fitted by adding extra lines to our job file. Change

```
yield ~ mu variety !r  Rep
residual units
```

To

```
!PART 1    # Randomised complete block analysis
yield ~ mu variety !r  Rep
residual units

!PART 2    # balanced incomplete block analysis
yield ~ mu variety !r  Rep rowblk  colblk
residual units

!PART 3    # basic spatial analysis
yield ~ mu variety !r  Rep
residual ar1(row).ar1(col)

!PART 4    # BIB spatial analysis
yield ~ mu variety !r  Rep rowblk  colblk
residual ar1(row).ar1(col)
```

and insert 2 as the argument to !ARG on the first line. The .esr file from this model contains

```
   7 LogL= -707.79   8062.         125 DF

 Akaike Information Criterion   1423.57 (assuming 4 parameters).
 Bayesian Information Criterion 1434.88

        Analysis of yield

                  Wald F statistics
Source of Variation        NumDF     DenDF     F-inc          P-inc
 mu                        1          1216.99
 variety                   24            8.84

 Model_Term                Order     Gamma       Sigma     Z_ratio  %C
 Replicate                   6  0.528713      4262.41       0.62   0 P
 rowblk                     30   1.83725      14811.6       3.04   0 P
 colblk                     30   1.93444      15595.1       3.06   0 P
 Residual_units            150   1.00000      8061.85       6.01
  Replicate                      6 effects fitted.
  rowblk                        30 effects fitted.
  colblk                        30 effects fitted.
```

First note the LogL has increased 35.55 (from -743.34 to 0707.79) with just 2 extra parameters and so this is a significantly better fitting model. It attempts to model the spatial variation by a rather complicated sum pf rowblk and colblk effects. The replicate variance component drops from 9280 to 4262 is really no longer significant; it dropped, its variation would move in the blk effects. Also notice that the F statistic for the Variety effects has increased from 3.06 to 8.84 indicating much greater confidence in the estimated Variety effects.

This was the best that could be dome in 1976.

Basic spatial analysis

The main papers developing spatial analysis now implemented in ASReml/Echidna were those of Gleeson and Cullis (1987) and Cullis and Gleeson (1991).

Following on from Gilmour et al (1997) who expounded on the use of a variogram to explore the spatial variation, spatial analysis in the broad sense covers all possible models that are based on the physical layout of the plots indexed by row and column. The most popular base model assumes and auto-regressive correlation structure across rows and across columns. While autoregression has mainly been developed in a time context which is directional (what happens today depends to a large extent on what happened yesterday), the resulting correlation structure does not require dependence on one direction. Gleeson and Cullis (1987) considered the correlation in one field dimension. Cullis and Gleeson (1991) used a direct product concept to extend the autocorrelation into two dimensions.

In the Slate Hall example, the plots are arranged in a 10 × 15 regular grid (10 rows by 15 columns, all plots the same size). The variance of the residuals can be written as a direct product $\sigma^2 \, \mathbf{\Sigma}_r(\rho_r) \times \mathbf{\Sigma}_c(\rho_c)$ where σ^2 is the variance scaling the whole structure, $\mathbf{\Sigma}_r(\rho_r)$ is a 10 × 10 correlation matrix where the correlations decrease in powers of the correlation parameter as you move away from the diagonal (the first row of the matrix has elements [1 ρ_r^1 ρ_r^2 ρ_r^3 ρ_r^4 ρ_r^5 ρ_r^6 ρ_r^7 ρ_r^8 ρ_r^9]), × is the direct product operator and $\mathbf{\Sigma}_c(\rho_c)$ is like $\mathbf{\Sigma}_r(\rho_r)$ but pertaining to columns and of order 15.

An advantage of this particular correlation structure is that its inverse is a tri-diagonal matrix which means the estimation process is efficient, even in very large field trials.

This structure for the residuals is specified to the software by writing
residual ar1(row).ar1(col)

Change the argument to !ARG to 3 on the first line and the resulting analysis report is:

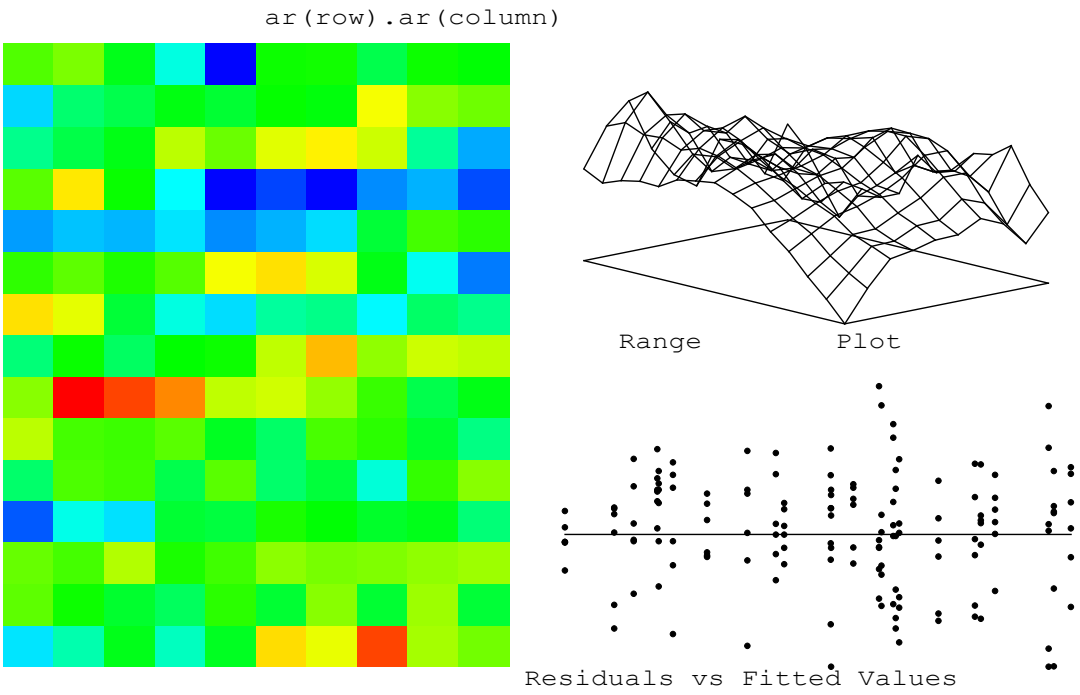
```
 ...
  7 LogL= -700.32  0.3870E+05      125 DF

Akaike Information Criterion   1408.65 (assuming 4 parameters).
Bayesian Information Criterion 1419.96

      Analysis of yield

                Wald F statistics
Source of Variation        NumDF     DenDF     F-inc          P-inc
mu                         1           853.84
variety                    24           13.02

Model_Term                Order     Gamma       Sigma     Z_ratio  %C
Replicate                  6  0.138503E-04  0.536001       5.03 -87 P
ar1(row).ar1(col)              150 effects
ar1(row)                  15  0.683679      0.683679      10.92   0 P
ar1(col)                  10  0.457673      0.457673       5.58   0 P
Residual_units             150   1.00000      38699.5        5.03
 Replicate                      6 effects fitted.
```

This model has 4 variance parameters, the same number as model 2, but has a higher LogL (-700.32 rather than -707.79) and so fits the data better. The F statistic for Variety is also higher (13.02 rather than 8.84) and so the variety effects have higher precision.

There is negligible variation associated with replicate effects (dropping Replicate from the model would not reduce the LogL). The "gamma" value associated with ar1(row) is actually the correlation parameter (ρ_r) and is quite high indicating strong spatial variation, and similarly for ar1(col).

Several comparisons of model fits across a series of trials were conducted in the 90's and generally showed that more than half the trials exhibited spatial variation and in most cases, the ar1 x ar1 model outperformed the BIB model as it did here.

This analysis utilises the actual spatial layout of the plots and an expanded residual graphics file is produced.



The 'heatmap' shows where the high plots (dark blue) and low plots (red) are. The top right shows a variogram based on the residuals. The difference between neighbouring residuals is substantially less than for plots further apart. The bottom right shows the usual plot of residuals against fitted values.

Balanced Incomplete block analysis with correlated errors

Change the argument to !ARG to 4 on the first line. Running this model reports a LogL of -699.96 which represents a LogL gain of 0.36 with 2 extra parameters. So, adding the BIB block structure does not improved the model fit.

Discussion on Slate Hall model comparison

Comparing these four models:

	LogL	Residual variance	Number of variance parameters	F ratio for Variety effects
Randomised Complete Block	-743.34	34665	2	3.06
balanced incomplete block	-707.79	8062	4	8.83
basic spatial analysis	-700.32	38715	4 (actually 3 since Replicate is 0.0)	13.02
BIB spatial analysis	-699.96	3604	6 (actually 4 since Replicate and rowblk are nearly 0.0)	13.50

There is substantial spatial variation in this data as indicated by the fact that the last 3 models have a substantially higher LogL than the first model. But notice that the basic spatial analysis has a higher LogL than the BIB analysis and combining the 2 produces effectively no further gain.

Several reviews of field experiments have shown that field experiments tend to have spatial variation and that the 'autoregressive' spatial model is usually sufficient to model the spatial variation.

Review

A basic model for a field trial might be:

```
Y ~ mu !r genotype
residual units
```

which assumes the residuals are independent (S^2 I) and fits genotype as random but experience shows field plots are usually not independent and this observation led to the now common spatial model

```
Y ~ mu !r genotype
residual ar1v(row):ar1(col)
```

where the residual variance structure is assumed to be σ^2 C_r x C_c where σ^2 is the residual variance, C_r (C_c) is an autoregressive correlation structure across rows (columns) of the plots, and these are in direct product. The autoregressive correlation structure has a single parameter which is the correlation between immediate neighbours; the correlation (being <1) reduces as the power of the distance. A separate correlation is specified for rows and columns because plots are typically not square and one source of correlation is plot management which tends to be applied across columns or across rows.

The reported variance components are

Row AR_R	the row autocorrelation parameter (C_r)
Row AR_V	the residual variance (S^2)
Col AR_R	the column autocorrelation parameter (C_c)

Note that if C_r and C_c are zero, this reduces to the independent residual variance model.

Now there are many ways we could specify a more complicated residual variance structure (Gilmour, A.R., Cullis, B.R. and Verbyla, A. (1997). Accounting for Natural and Extraneous Variation

in the Analysis of Field Experiments. Journal of Agricultural, Biological and Environmental Statistics 2, 269-293.)

One is to assume the residual variance structure is S^2_i I + S^2_c C_r x C_c in which the residuals are split into two components (a correlated part and an independent part). The independent part is then often called the nugget variance. It is exactly the same process as when genetic variance is estimated in an animal model; the total variance is partitioned into a genetic (correlated) part and an (uncorrelated) residual. In forestry, we can have two correlated parts (genetic and spatial).

This model is what is fitted by

```
Y ~ mu !r genotype ar1(row):ar1(col)
residual units
```

or

```
Y ~ mu !r genotype units
residual ar1v(row):ar1(col)
```

Missing plots

You raise the matter of missing plots. It is necessary for the total grid of plots to be present to fit the correlated spatial residual structure. This requires the missing plots be estimated and so the respective models are written

```
Y ~ mu mv !r genotype
residual ar1v(row):ar1(col)
```
and

```
Y ~ mu mv !r genotype units
residual ar1v(row):ar1(col)
```

Extraneous variation

Gilmour et al (1997) distinguished between natural and extraneous sources of spatial variation and showed how experimental operations had contributed to the latter. The data was from a wheat trial conducted by Gil Hollamby in South Australia involving 3 replicates of 107 varieties designed as a randomised complete block.

The RCB analysis reports

```
 7 LogL= -1235.22  0.1336E+05      223 DF

Akaike Information Criterion   2474.44 (assuming 2 parameters).
Bayesian Information Criterion 2481.26

      Analysis of Yield
```

Wald F statistics

Source of Variation	NumDF	DenDF	F-inc	P-inc
mu	1	82.54		
Variety	106	1.44		

Model_Term	Order	Gamma	Sigma	Z_ratio	%C
Replicate	3	0.953283	12736.1	1.01	0 P
Residual_units	330	1.00000	13360.3	10.51	

The basic spatial AR1 × AR1 analysis discussed above reports

9 LogL= -1101.58 0.1556E+05 223 DF

Akaike Information Criterion 2211.17 (assuming 4 parameters).
Bayesian Information Criterion 2224.80

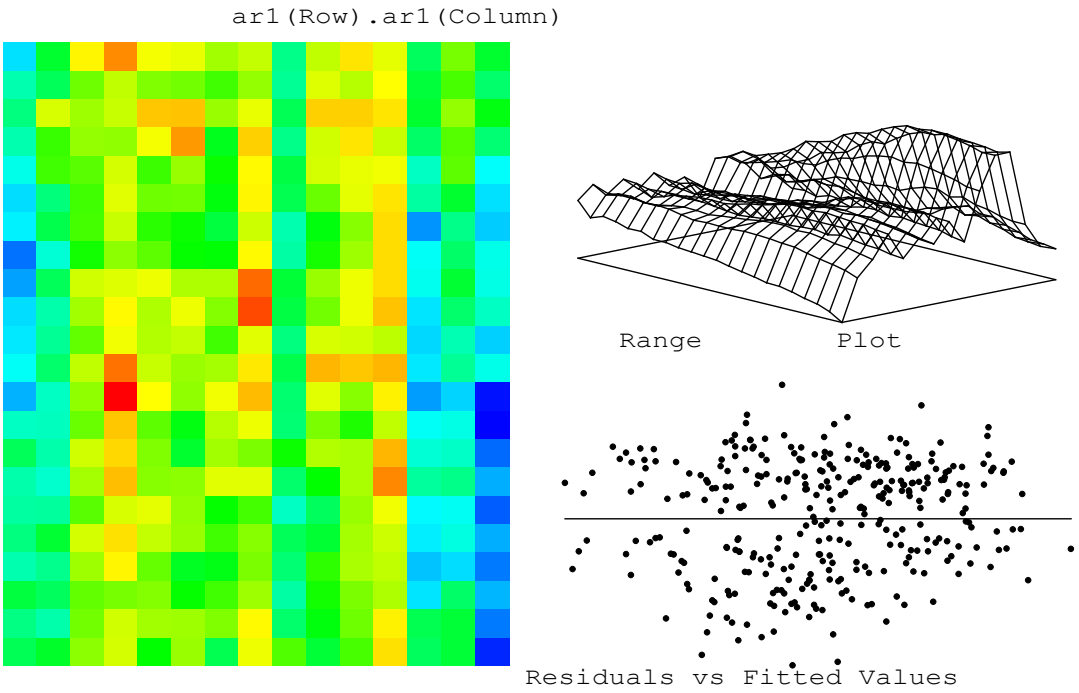
Analysis of Yield

Wald F statistics

Source of Variation	NumDF	DenDF	F-inc	P-inc
mu	1	95.66		
Variety	106	6.90		

Model_Term	Order	Gamma	Sigma	Z_ratio	%C
Replicate	3	0.501648	7806.23	0.79	0 P
ar1(Row).ar1(Column)	330 effects				
ar1(Row)	22	0.881500	0.881500	40.44	0 P
ar1(Column)	15	0.387177	0.387177	5.63	0 P
Residual_units	330	1.00000	15561.2	5.25	

The LogL increased 133.64, a highly significant amount, and the F value for Variety increased from 1.44 to 6.90. Also note the very high row auto-correlation of 0.88. To try and understand the spatial variation, we can inspect the heatmap of the residuals and the variogram.



It is not difficult to see column effects in the heat map, even blocks of three columns. This pattern is unlikely to be natural to the plots and is likely caused by the experimenter.

The variogram plots the variance of the difference between residuals at various distances apart. The origin is the 0,0 point where the variance is zero since there is no difference of a point with itself (lag 0). Range is Row. Moving left from the origin, the variance slowly increases as the distance reflecting a typical stationary autoregressive process. However, the fact that it continues to increase rather than plateau suggests a trend across rows. Moving right we see large differences between close neighbours and that plots 4 columns apart are more similar than plots 3 columns.

Too investigate further, we add random row and column effects.

9 LogL= -1078.60 3018. 223 DF

Akaike Information Criterion 2169.21 (assuming 6 parameters).
Bayesian Information Criterion 2189.65

Analysis of Yield

Wald F statistics

Source of Variation	NumDF	DenDF	F-inc	P-inc
mu	1	82.34		
Variety	106	5.58		

Model_Term	Order	Gamma	Sigma	Z_ratio	%C
Replicate	3	3.40134	10264.3	0.81	0 P
Column	15	3.84523	11603.9	2.37	0 P
Row	22	0.175692	530.193	2.00	0 P

ar1(Col).ar1(Row)		330 effects			
ar1(Col)	15	0.194670	0.194670	2.11	0 P
ar1(Row)	22	0.477400	0.477400	6.87	0 P
Residual_units	330	1.00000	3017.73	7.58	

Now Replicate is a classification of columns and together they explain considerable variation. The effects are: -131 -69 62 125 72 132 99 183 0 117 3 40 -214 -147 -264

The explanation proposed by Gilmour et al (1997) is that it reflects inaccurate plot length trimming creating a pattern 123412341234123 that is evident n the variogram.

Row effects are smaller but still significant.
-4.6 1.1 -9.9 -22.3 6.8 -22.0 8.6 2.3 -23.9 -4.4 23.7 -15.7 16.7 20.2 -22.9 -22.1 1.9 1.4 16.0 42.5 -5.7 26.0

As well as a trend across the rows, there is an effect of the combine used. It actually sowed 3 plots at once and was used a serpentine manner generating a pattern LLMRRMLLMRRMLLMRRMLLMR (Left,Middle,Right) where the L plots tend to yield high and the M plots low.

Note that the LogL from this model is significantly (22.98) higher than model 2 but the Variety F ratio has dropped from 6.9 to 5.6. So Model 3 does a better job at modelling the variance is probably not much better at estimating fixed Variety effects.

Gilmour et al (1997) recommend fitting the identified patterns as fixed effects but I believe the main advantage of this analysis is to highlight the need for good experimental procedures which avoid these issues.

Automated analysis

When analysing many trials in an automated process, it is not feasible manually explore alternative models. Some use model 3 as standard but if it fails, drop back to a simpler. To drop row/column from model 3 (giving model 2) is to be preferred over dropping the AR1 × AR1 residual correlation structure because model 2 will accommodate a wider range of structures.

Some are concerned about over-fitting. This will mainly be a problem of fixed terms (such as lin(row), lin(col)) are added to accommodate spatial variation when they are not needed. However, using random terms such as row and column factors were the variance components are estimated (by REML) rarely results in over-fitting because if they are not needed, the variance component will be small and the resulting adjustement is negligible. The same applies to the residual autocorrelation. If the correlation parameter is not significant, the resulting adjustment of variety effects will be negligible.

The exception is in unreplicated trials. When an 'animal' model is fitted to individual tree data (i.e. a pedigree file is used to create a relationship matrix connecting all trees) and a spatial (AR1 × AR1) residual model is fitted, it is necessary to also fit an independent units term to properly estimate the genetic component.

Fixed versus Random

The preceeding analyses have fitted Variety as a fixed effect. This is appropriate in a replicated experiment where there is (roughly) equal information on each genotype and the purpose is to describe

the outcome of the experiment. However, plant breeding is usually about predicting future performance (selecting genotypes for advancement or for industry release). Then, it is more apprpriate to treat varieties as random effects, especially when information (replication) is not equal. Random effects have a narrower range than fixed effects because they a shrunken to reflect the size of the corresponding variance component.

In the early stages of a breeding program, seed is limited and trials are unreplicated (except for a grid of check plots of standard virieties) or partially replicated (maybe a third are replicated). In this situation, genotypes need to be fitted as random effects to get the Best Linear Unbiassed Predictions of the genetic potential of the lines. Indeed a pedigree may be used to define genetic links among them. An AR1 × AR1 residual model is then most advantageous.

Genotypes are also fitted as random effects when we analyse a series of related trials together. Then we are able to model the correlation (or covariance) of genetic performance among trials.

Two stage analysis

So far we have discussed analysis of individual trials but the reality is that plant breeding typically involves planting many similar trials over a region and over years as the aim is to predict performance across a region in future years.

With a few trials, the data can be combined and a multi-environment spatial analysis performed. A possible model statement might be:

```
Yield ~ mu Env !r diag(Env).Row diag(Env).Column xfa1(Env).Genotype
residual sat(Env).ar1v(Row).ar1(Column)
```

However, this is not feasible in general when there may be hundreds of trials to combine. In this case, it is common to perform a weighted analysis of adjusted data. First, each trial is analysed independently and the variety means are predicted along with weights to be used in the combined analysis. The required predict statement is typically like:

```
Yield ~ mu Genotype !r Row + Column
residual ar1v(Row).ar1(Column)
predict Genotype !TWOSTAGEWEIGHTS
```

The predicted values and their weights are typically saved to a data base. Then for a combined analysis, the values from the appropriate trial extracted and analysed together.

However, in the two stage approach, the means taken forward must be based on fixed effects and as we have noted, this model is not appropriate for unreplicated and partially replicated trials. In that case, we first fit genotypes as random and then use the estimated spatial variance parameters in a second model with genotype fixed. For example

```
!OUT  !NO  !REN !ARG 1 2
Slate Hall 1976 Cereal trial !DOPART $1
rep     6
latrow  30
latcol  30
variety 25
yield
```

```
 fldrow  15
 fldcol  10
shf.asd !SKIP 1

!PART 1     #Fitting AR1.AR1     Gamma Scale
yield ~ mu mv !r var fldr fldc
res ar1(fldr).ar1v(fldc)

!PART 2 #Fitting AR1.AR1
!HOLD 1:100
!CONTINUE shfhld1\shfhld.rsv
!MAXIT 1
yield ~ mu var mv !r fldr  fldc
res ar1(fldr).ar1v(fldc)
predict var !TWOSTAGEWTS
```

Combined analysis

As an example, we consider the analysis of 87 Lupin trials. The basic ASReml command file is

```
!WORK 2 !DE !LOG !NO !continue !R !arg  1 2 3 4  !OUT

Title: ALBUS_2stage.            !DOPART $1
#trial,year,region,variety,yield,rep,weight,ems
#KFA02BURU,2002,NSW,KIEV-MUTANT,0.873,3,2136.562,0.0010000
 trial  !A
 year  !I
 region  !A
 variety  !A
 yield
 rep  *
 weight   !*0.025
 ems

ALBUS.csv  !SKIP 1   !MAXIT  150  !SLN

!PART 1 2 3 4   # 2911.48  3053.73  3153.92  3230.94
 yield !wt=weight !DISP 0.025 ~ mu trial !r xfa$1(trial).var
```

Model (PART 1) in Echidna reports

```
Data File: ALBUS.csv

 Summary of 2019 data records

 Variable   Levels Miss Zero      Min      Max    Distribution or Mn SD Sk Kt
 trial          87    0    0        1       87
 year            5    0    0        1        5   337 438 518 430 296
 region          3    0    0        1        3   1940 24 55
 variety       203    0    0        1      203
 yield           1    0    0     0.09     5.61     1.73     1.08   0.56  -0.50
 rep             6    0    0        1        6   201 294 1485 18 6 15
 weight          1    0    0     0.19   335.37    14.75    28.27   4.84  29.70
```

```
 ems             1    0    0  0.00037  0.39100  0.02491  0.04671   5.95  42.38

 Note: Using !DOPART 1
 Note: Model is fitting 17952 equations, DENSE portion has 88 equations.

    1 LogL= 2056.94      1932 DF
    2 LogL= 2563.48      1932 DF
    3 LogL= 2635.87      1932 DF
    4 LogL= 2684.64      1932 DF
…
   49 LogL= 2911.06      1932 DF
   50 LogL= 2911.11      1932 DF

 Akaike Information Criterion   -5474.23 (assuming 174 parameters).
 Bayesian Information Criterion -4505.69

          Analysis of yield
          using weights in weight

                         Wald F statistics
Source of Variation           NumDF     DenDF     F-inc              P-inc
 mu                               1            22303.96
 trial                           86              810.25

 Model_Term                     Order     Gamma         Sigma     Z_ratio  %C
 xfa1(trial).var                    17864 effects
 xfa1(trial)_V              0   1  88  0.741962E-02  0.741962E-02    3.23   0 P
 xfa1(trial)_V              0   2  88  0.537544E-02  0.537544E-02    1.35   0 P
 xfa1(trial)_V              0   3  88  0.277613E-02  0.277613E-02    2.82   0 P
 xfa1(trial)_V              0   4  88  0.635824E-02  0.635824E-02    2.98   0 P
…
xfa1(trial)_L              1  86  88  0.965742E-01  0.965742E-01    0.63   1 P
 xfa1(trial)_L              1  87  88  0.190312E-01  0.190312E-01    1.71   0 P
 Residual_units                2019   1.00000       1.00000        0.00
```

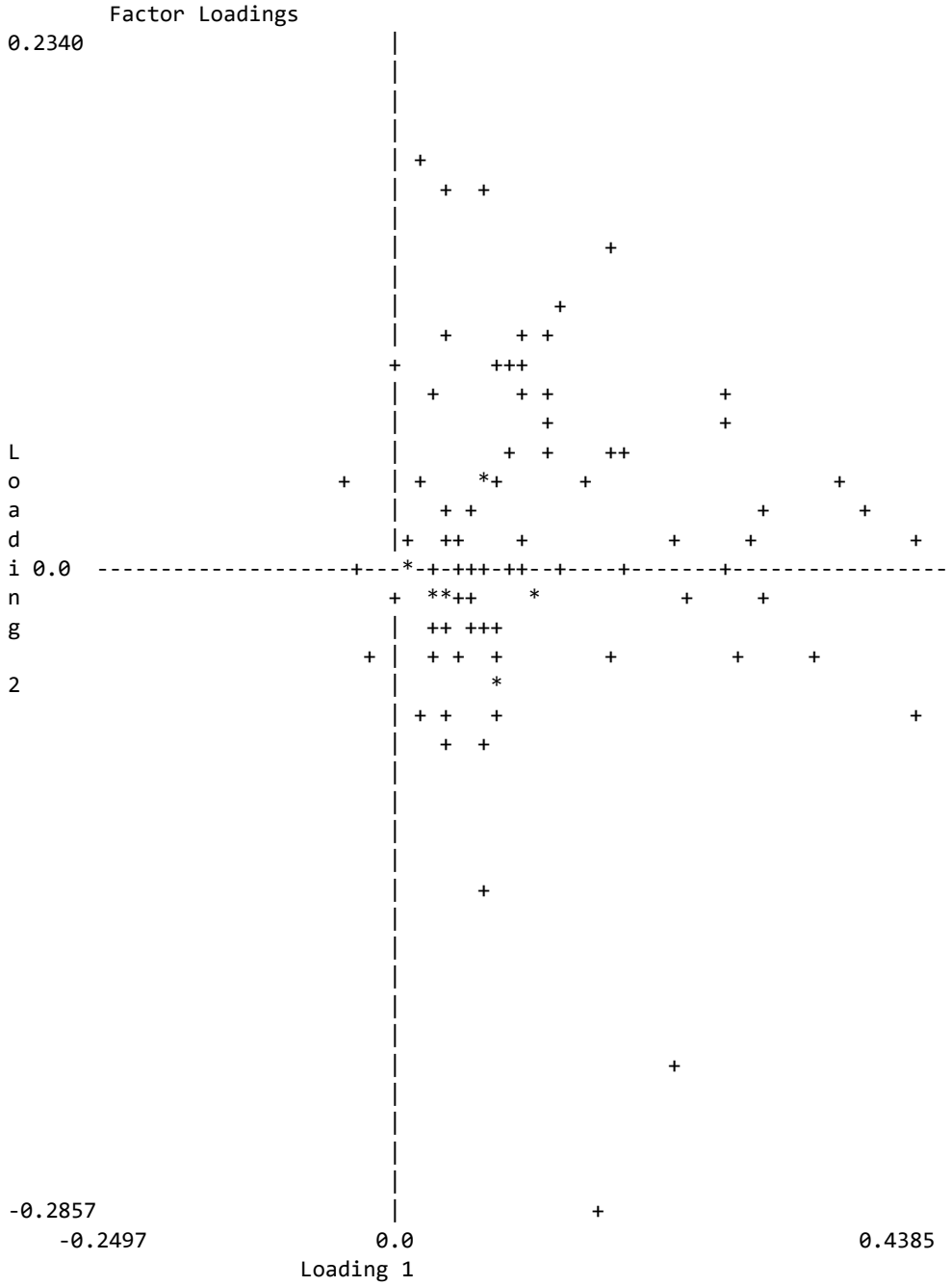
The data summary shows there are 2019 means in the data covering 87 trials conducted over 3 regions (but mainly just 1) and 5 years and representing 203 varieties. Therefore the trial × variety table has only 11.4% cells occupied. A single factor factor analytic model has been fitted. This models the across trial variance matrix as $\boldsymbol{\Gamma\Gamma'} + \boldsymbol{\Psi}$ where $\boldsymbol{\Gamma}$ is a vector of loadings so that $\boldsymbol{\Gamma\Gamma'}$ provides the covariances and $\boldsymbol{\Psi}$ is a diagonal matrix of specific variances (the genotype × environment component of the variety variance). The BLUPs for the FACTOR reported in the .esy file represent the common variety effect over all trials and the loadings scale the common effect for each trial.

Of course there may be structure that is not picked up in this single factor model and so we also fit the model with 2, 3 and 4 factors. Since the two-way table is relatively sparse, these models do not always converge to the same point. The input file reports some LogL values (2911.48 3053.73 3153.92 3230.94) obtained some years ago in ASReml. The values obtained with Echidna 1.41 on 29 Jan 2020 were 2911.11, 3030.35, 3142.76 and 3222.58. The LogL increases are 129, 112 and 80 representing 86, 85 and 84 unconstrained parameters each. So the 4 th factor is probably not significant.

If you want an overall ranking of genotype performance, you would use the Variety BLUPs for the factor from the first model.

If you wanted to identify trials which stood apart from the consensus, you could calculate diag(**Ψ)/(** diag(**Ψ)+ Γ**^{2}).

If you wanted to classify the trials, you would plot $\mathbf{\Gamma_1}$ against $\mathbf{\Gamma}_2$ and the other loading vectors.



Conclusion

This is just a basic introduction to the topics discussed. The data sets discussed (shf.dat, jabes.dat and albus.csv) are in the Echidna Examples folder distributed with the program.

References

Cullis, B. R., & Gleeson, A. C. (1991). Spatial analysis of field experiments – an extension to two dimensions. *Biometrics*, **47**, 1449–1460. https ://doi.org/10.2307/2532398

Gilmour, A.R., Cullis, B.R. and Verbyla, A. (1997). Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. *Journal of Agricultural, Biological and Environmental Statistics* **2,** 269-293.

Gleeson, A. C., & Cullis, B. R. (1987). Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. *Biometrics*, **43**, 277–288. https ://doi.org/10.2307/2531812