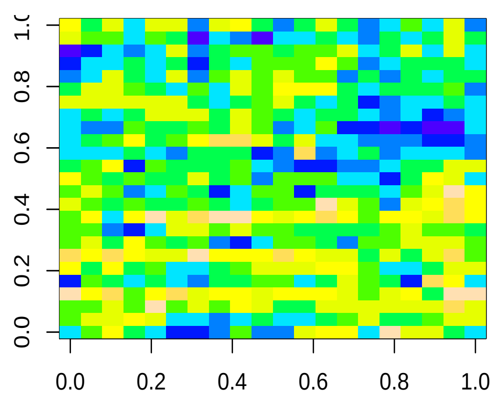
Spatial analysis of a designed experiment

- RA Fisher's 3 principles of experimental design
  - randomization ⇒ unbiased estimate of treatment effect
  - replication ⇒ unbiased estimate of error variance
  - blocking = "local control of variation" ⇒ eliminate unwanted sources of variation
- Any experiment: experimental units (eu's) are not identical
  - Lots of sources of variation:
- Field experiment: variation is often spatially structured
  - elevation / moisture / drought; soil type
  - proximity to field edge

Uniformity trials

- Uniformity trial
  - Common in early - mid 20'th century in US and UK
  - "Experiment without a treatment"
  - before using an experimental field, plant a crop, harvest in small plots
  - look at how variation is structured across the field
- Usually find that high yielding plots are surrounded by other high-yielding plots
- and similarly for low-yielding plots.
- E.g. Mercer and Hall wheat yield data (next slide)
  - Classic data set (1910 study, 1911 paper)
  - Experimental field planted in wheat. No treatments - all the same
  - Harvested in small plots: 3.3m E-W, 2.51m N-S
  - ca 2-fold variation in yield
  - Key point: variation is spatially correlated

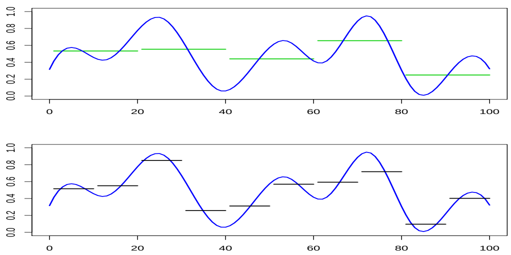


Blocking

- Traditional approach is to control unwanted variability by blocking
- Blocks are groups of similar experimental units
  - human study: group by sex and age-group (e.g. male, age 20-29)
  - field study: group based on knowledge of field, almost always adjacent plots
    e.g. low part of field = one block, high part = second block
  - very useful. Typical efficiency = 110% - 120%
    i.e. Var CRD = 1.1 Var RCBD - 1.2 Var RCBD
  - Alternatively, 10 replicates in an RCBD have same precision as 11-12 replicates in a CRD

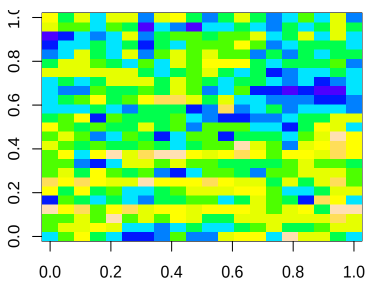
Blocking: practical issues

- 3 issues/problems:
- 1) analysis model is constant block effect (same for all plots w/i that block). But, variation may be smoother



Blocking: practical issues

- 2) often hard to know where to place blocks (e.g. M-H data)



Blocking: practical issues

- 3) want blocks to be small, but may need to be large to include all trts
  - small blocks more likely to be homogeneous
  - this (in my mind) is why one rep per trt and block is so common
- Plant breeders often have very large numbers of treatments
- trts = varieties of plant, often 128 or 256.
  - often use very ingenious incomplete block designs
  - Complete block: all treatments in each block
  - Breeders: subset treatments in each block, e.g. 16 trts per incomplete block
  - often "resolvable": collection of small blocks makes a large block that has all trts
  - 8 incomplete blocks, each with 16 diff trts, includes all 128 trts

Consequences of spatial correlation

- Nearby plots are clearly similar to each other.
- I sometimes see the argument that the usual ANOVA on a field experiment is wrong because of the spatial correlation
- Remember: ANOVA makes three assumptions:
  - errors are normally distributed
  - errors have equal variance
  - errors are independent
- Which is most important??

Consequences of spatial correlation

- A: independence
- So, you do an experiment on plots that are spatially correlated
  Is ANOVA wrong, because it violates the independence assumption?

Consequences of spatial correlation

- I say no, at least for a designed experiment:
  - the independence comes from the random assignment of treatments to plots.
  - So there is nothing wrong about ignoring spatial correlation
  - But accounting for spatial correlation may be a better analysis
    - increased precision of estimates
    - increased power of tests
- analogous to sampling spatially organized things
  - a simple random sample, assuming independence, is just fine
  - OLS estimates (usual estimates) and tests are NOT wrong
- Note: very different from ignoring subsampling or repeated meas.
- Why is correlation among rep. meas. a problem, but spatial correlation is not?
  - Treatment randomly assigned in spatial study
  - Time not randomly assigned in rep. meas.

GLS and eGLS

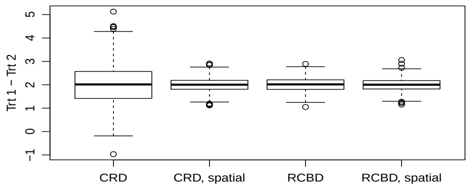
- GLS estimates that account for spatial correlation will be better
- When spatial correlation model and parameters known
- Obvious problems:
  - don't know the form of spatial correlation (what model?)
  - and certainly don't know correlation parameters
  - (e.g., nugget, range, sill)
- need to estimate these from data
- If want to be accurate, call the procedure: eGLS
- Algorithm:
  1. Assume independence (to get started)
  2. estimate fixed effect parameters
  3. use residuals to estimate VC parameters
  4. repeat 2, 3 until convergence

eGLS in practice

- Various practical concerns
- 1) Frequentist inference conditions on estimated VC parameters
  - Ignores uncertainty in the VC parameters
  - Go Bayes if want to incorporate VC uncertainty
- 2) small sample distribution eGLS estimates not known
  - Approximate as T with an estimated df
  - Satterthwaite approximation (Giesbrecht and Burns 1985 extension)
- 3) Estimates of VC parameters are biased when obs have spatial correlation
  - Kenward-Rogers method implements a bias correction to VC estimates
  - then applies Satterthwaite
- Lots of ad-hoc adjustments
- But works relatively well (in simulation studies)
  - unless study has very few replicates

Illustration

- Consider a field, spatially correlated plots (details don't matter)
- Design a study to compare 5 treatments, 50 plots
- Consider two experimental designs: CRD or blocks (RCBD)
- and two analyses: usual or spatial analysis
  - Generate data from "a study" using a design (CRD, RCBD)
  - Estimate parameters using a model (usual, spatial), repeat 1000 times
  - Focus on estimated difference between two treatments: $\hat{\mu}_1 - \hat{\mu}_2$



Illustration

Design	Analysis	Average	sd=se	$\sqrt{\text{Ave est Var}}$	ratio
CRD	–	2.010	0.870	0.875	1.006
"	spatial	2.003	0.282	0.244	0.868
RCBD	–	2.006	0.299	0.302	1.011
"	spatial	2.003	0.266	0.254	0.953

- Bias of estimates: Compare ave. estimate to truth (=2.00)
  - Ignoring spatial correlation still unbiased
- Precision of estimates: look at sd of estimates
  - Blocking substantially increases precision (much more than 10%)
  - Spatial analysis further increases precision
    - A lot for CRD, a bit for RCBD
- Is the precision well estimated (equiv. of se = sd$/\sqrt{n}$)
  - Non-spatial analysis: fine (ratios close to 1)
  - Spatial analyses: tends to underestimate se

Papadakis's method

- Old method - original paper in 1937
- Concept:
  - Use neighbors to "predict" what an obs. would be like if no treatment
  - Use this value as a covariate in model to remove "local" spatial variation
- Details:
  - Fit preliminary model $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
  - Estimate residuals = $\hat{\varepsilon}_{ij} = Y_{ij} - (\hat{\mu} + \hat{\tau}_i)$, i.e. obs. - trt mean
  - calculate $\bar{r}_{ij}$ = average residual for each obs. by ave. resids of neighbors
  - do not include residual for self
  - include $\bar{r}_{ij}$ as a covariate in the model

$$Y_{ij} = \mu + \tau_i + \beta\,\bar{r}_{ij} + \gamma_{ij}$$

Papadakis method

- Divides "error" into two components:
  - a) contribution of neighbors: $\beta\,\bar{r}_{ij}$
  - b) "intrinsic error": γ_{ij}
- Consequences:
  - obs. surrounded by "high" neighbors (relative to their treatment means) are expected to be high
  - surrounded by "low" expected to be low
  - if $\hat{\beta} \approx 0$, little to no spatial correlation
    - neither blocking nor Papadakis adjustment very useful
- Not easily implemented in modern software
  - Can exploit relationship between Papadakis and spatial models for areal data
  - Modern version: nearest neighbor adjustment

Spatial linear model

- Notation: i: Treatment, j: Replicate

$$\begin{aligned} Y_{ij} &= \mu + \tau_i + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim mvN(\mathbf{0}, \mathbf{\Sigma}) \end{aligned}$$

- Just like the usual linear model (ANOVA or regression or combination), but errors are correlated
- VC matrix, $\mathbf{\Sigma}$, usually specified as a geostatistical model
  - VC matrix is for plot errors, not observations
  - Parameterized in terms of correlation or covariance
- Common to assume equal variances and to use one of the usual variogram models
  - e.g. Exponential, Spherical, Matern

Spatial Linear Model

- Notice this is very much like models for repeated measures data
  - Treatments assigned to subjects,
  - each subject measured more than once
  - observations on same subject probably correlated
  - 402 discussed various models for those correlations
  - correlation depends on time lag between observations
- Spatial is similar; now correlation depends on distance, not time lag
- Can add additional trend to the fixed effects model

$$\begin{aligned} Y_{ij} &= \mu + \tau_i + \beta_E \, Easting_{ij} + \beta_N \, Northing_{ij} + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim mvN(\mathbf{0}, \mathbf{\Sigma}) \end{aligned}$$

- Or block effects (α_j)

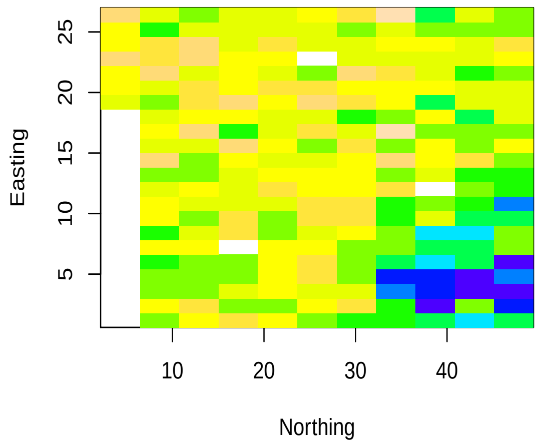
$$\begin{aligned} Y_{ij} &= \mu + \tau_i + \alpha_j + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim mvN(\mathbf{0}, \mathbf{\Sigma}) \end{aligned}$$

Cullis Gleeson model

- Developed for agricultural data: crop planted in rows, plots on a grid
  - does not have to be a square grid; can have different spacing along rows from across rows
- Cullis and Gleeson (1991, Biometrics) model accounts for correlation in two dimensions
  - AR(1) model with one correlation along the rows (x coordinate)
  - AR(1) model with different correlation across the rows (y coordinate)
  - Cor $\varepsilon(\boldsymbol{s}_1)$, $\varepsilon(\boldsymbol{s}_2) = \rho_x^{\Delta x} \rho_y^{\Delta y}$
- Implemented in ASREML
- Can fit in SAS as anisotropic exponential covariance
- R doesn't (as far as I can tell) include anisotropic correlation models
  - Fudge it when ρ_x / ρ_y known by rescaling coordinates
  - Same trick as used with geometric anisotropy

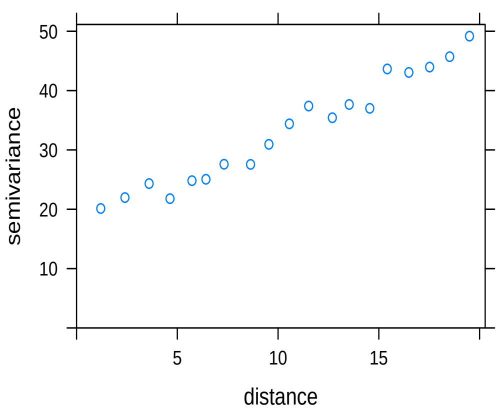
Example: Alliance wheat trial

- Variety trial in wheat. 56 varieties, 4 reps of each, blocked
- Exploratory analysis:
  - Fit a model **without blocks**
  - (we are interested in the spatial effect, after all)
  - plot residuals for each location
- see next slide
- spatial pattern in residuals is obvious

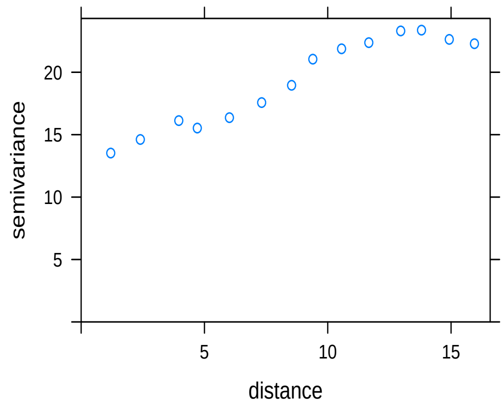


Calculating GLS estimates

- Remember that GLS estimates require a $\mathbf{\Sigma}$ matrix
- Two ways to estimate $\mathbf{\Sigma}$:
  - Variogram on OLS residuals
  - Estimate $\mathbf{\Sigma}$ and β together (REML)
- 1) Estimate a variogram from residuals **is an approximation**
  - Why is this an approximation?
  - A: residuals are negatively correlated
  - often ignored when error df/n close to 1
  - not so here! because only 4 reps / entry
  - error df/n = 0.75



- Variogram looks linear, even when extend max lag distance to 20
- Suggests a spatial trend
- Add Northing, Easting, and their product to model
- Product because of blob of extreme residuals in one corner of field
- Variogram of those residuals looks much nicer
- Could use this variogram to estimate $\mathbf{\Sigma}$, then use GLS to estimate $\hat{\beta}$
- But, notice the problem:
  - $\hat{\mathbf{\Sigma}}$ based on OLS residuals
  - GLS estimates of $\hat{\beta}$ are not the same as the OLS estimates
  - so residuals are not the same
  - so $\hat{\mathbf{\Sigma}}$ will change
- And, have the df/n issue
- Alternative is to simultaneously estimate $\mathbf{\Sigma}$ and β, using REML to account for fixed effects (df/n issue)



REML estimation of variances and related quantities

- Maximum likelihood is great, but estimates are often biased
- Example: $Y_i \sim N(\mu, \sigma^2)$
  - ML estimate of σ^2 is $\frac{1}{n}\Sigma(Y_i - \hat{\mu})^2$
  - Unbiased estimate is $\frac{1}{n-1}\Sigma(Y_i - \hat{\mu})^2$
  - subtracting 1 "accounts" for using the data to estimate $\hat{\mu}$ before estimating $\hat{\sigma}^2$
  - can be a serious issue when n small, or many fixed effect parameters
- If estimate k fixed effect parameters, unbiased est. is $\frac{1}{n-k}\Sigma(Y_i - \hat{\mu})^2$
- Basic idea for a solution known in 1937 (Bartlett), but widely popularized in early 1970's (Patterson and Thompson)
- Since Patterson and Thompson, known as REML = REstricted ML or REsidual ML

REML

- Concept:
  - Calculate residuals for each obs. using fixed effects model
  - When the fixed effects have k d.f., the residuals have $n - k$ df.
  - If you give me $n - k$ residuals, I know the values of the remaining k residuals
    - Simple example: $Y \sim N(\mu, \sigma^2)$. Residuals must satisfy $\Sigma_i(Y_i - \hat{\mu}) = 0$, so if you give me the first $n - 1$ residuals, I know the value of the last residual: $Y_n - \mu = -\Sigma_{i=1}^{n-1}(Y_i - \mu)$.
  - So change the data: replace the n obs. by $n - k$ residuals.
    - no loss of information because the remaining k values are known
  - Then do ML on the $n - k$ residuals
  - For the simple example: $\hat{\sigma}^2 = \frac{1}{n-1}\Sigma(Y_i - \hat{\mu})^2$, which is the unbiased estimate!
  - In general, $\hat{\sigma}^2 = \frac{1}{n-k}\Sigma(Y_i - \hat{\mu})^2$, which is (again) the unbiased estimate!
  - REML accounts for the "loss of df due to estimating fixed effects"

Alliance: results from spatial linear model

- error variance is smaller than in the non-spatial analysis
- more precise estimates of treatment differences
- parameters of fitted semi-variogram different from empirical sv
- Reinforces earlier point about residuals
- R (and SAS) use REML to estimate variogram parameters.
  - REML accounts for the "loss of df from fitting fixed effects"
  - empirical sv does not
- so use empirical sv only to get an idea of starting values
- One other huge difference between REML and empirical sv estimate
  - REML: uses all the data (all distances)
  - Empirical variogram fitting: only shorter distances

More points about REML

- REML is an easy way to fit many spatial models
  - estimates of variances/covariances not always unbiased
  - but usually less biased than ML estimates
- How to choose the best model?
- Can calculate an AIC statistic from the REML lnL
  - AIC = -2 lnL + 2k
  - Interpret just like usual AIC:
  - Smaller is better (good fit to data with a simple model)
- BUT, REML/AIC only evaluates correlation models!
- must use same fixed effects model for all models
  - That's because AIC only comparable when models fit to the same data
  - Changing the fixed effects model changes the residuals, so changes the data that REML uses
  - Very common mistake!

More points about REML

- AIC only compares the specified set of models
  - Consider the following results for 3 correlation models:

Model	AIC
A	104.2
B	100.1
C	108.4

  - Tempting to say "B" is the correct correlation model
  - NO. You only know that B is the best among the set you evaluated
  - There could easily be a model D with AIC = 75.7 that fits much better than any you considered.
- Use diagnostics to check for anisotropy, outliers, and equal variances
  - Most models assume isotropy, no outliers, equal variances
- Or ignore, because an approximate spatial model is usually good enough

Accounting for spatial correlation

- Three common approaches
  - 1) geostatistical model: either point or areal data
  - 2) Simultaneous Autoregressive (SAR) Model for areal data
  - 3) Conditional Autoregressive (CAR) Model for areal data
- We've just talked about the geostatistical model
- More choices for areal data
- Choice reflects training / background of the user as much as reality
  - Statisticians: tend towards geostatistical approaches
    - use CAR models in Bayesian analysis
  - Spatial econometricians: almost exclusively SAR/CAR models

Final thoughts on spatial ANOVA/regression

- 1) Everything I've said about ANOVA models applies in straight-forward fashion to regression models
  - Both are specific choices of $\boldsymbol{X}$ in $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$
- 2) One thing to be aware of:
  estimates of $\hat{\beta}$ can change whan you change the correlation model
- In the usual (independence) ANOVA/regression model:
  - estimate the trt. means or the $\hat{\beta}$'s
  - use these to estimate the variance σ^2
  - but, the estimates do not depend on the variance
- In spatial (or more generally, most correlated data models), GLS estimates of $\hat{\boldsymbol{\beta}}$ depends on $\boldsymbol{\Sigma}$.
- e.g. in a plant breeder variety trial, adj. for spatial correlation may change ranking of varieties (because trt means are different).
- If you believe you have a good model for the spatial correlation, GLS ranking is better

Final thoughts on spatial ANOVA/regression

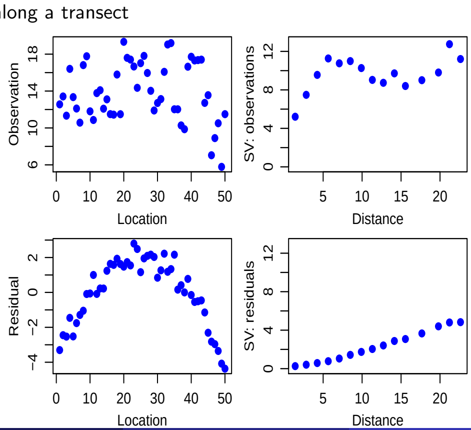
- 3) calculating d.f. for treatment means or comparisons of trt. means
  - In simple problems (independent data), d.f. $= n - k$
- not so when obs. are correlated. If + correlation, each obs. is less than 1 new piece of information
- A very difficult problem.
- One approach: refuse to compute df (At least one R package)
- Current best, but not great, for models with correlated observations
  Kenward-Rogers adjustment.
  Spilke et al. 2010, Plant Breeding 129: 590-598
  - ddfm=kr in SAS
  - pbkrtest package in R. Does not work with nlme models (e.g., with spatial correlation)
  - not too big an issue if many error df
- KR also adjusts variances to reduce bias

Final thoughts on spatial ANOVA/regression

- 4) When are spatial models likely to work well?
  - No published guidance (that I know about)
  - My thoughts:
    - At least 10 treatments, at least 5 reps per trt
    - and small scale = patchy spatial variation
    - May also need to remove blocks from fixed effect part of model "fights" with the spatial correlation
- 5) Remember there is a crucial difference between observations and residuals
  - Spatial models are for the residuals
  - Observations may have a very different pattern (next page)
  - Especially when treatments have large effects

Observations or residuals?

50 plots along a transect



Final thoughts on spatial ANOVA/regression

- 6, 7) What if the X variable is spatially correlated?
  - Very common in observational data
  - Has a couple of consequences
- 6) Spatial correlation in observations arises because X is correlated
  - Observations, Y, are spatially correlated
  - X is spatially correlated
  - residuals after regressing Y on X have no spatial correlation
- No need to adjust for spatial correlation; X has "taken care of it"
- Often not completely so
  - Still some "left over" spatial correlation in the residuals

Final thoughts on spatial ANOVA/regression

- One view of spatial correlation:
  - an omitted spatially correlated X variable accounts for that "left over" correlation
  - Spatial correlation is a surrogate for all omitted variables
  - spatial correlation model is equivalent to a model with independent errors and a "spatial X"
- Moran eigenvector maps (Griffith and Peres-Neto 2006, Ecology 87:2603-2613)
  - takes this idea one step further
  - construct a small set of new variables that account for the spatial correlation
  - i.e., move the spatial correlation from the VC matrix of errors into the fixed effect part of the model
  - essentially a principal components analysis

Final thoughts on spatial ANOVA/regression

- 7) Sometimes get bad news when you include a spatial correlation
  - Independent errors: regression coefficient for "your favorite" X is large and precise
  - Spatial correl.: now small, large se. Your favorite effect has vanished!
- Just like what happens when 2 X variables are highly correlated (multicollinearity)
- Spatial: "your favorite" X is highly correlated with the "spatial X"
- Difficult issues with interpretation
- Various approaches, appropriate practice is not settled
- Last two points primarily concern regression
  - Could be an issue for ANOVA, but only if treatments are very poorly distributed across the study area