SELECTING ENVIRONMENT COVARIATES TO EXPLAIN GXE: a comparison of cyclic forward regression and subset regression approaches.

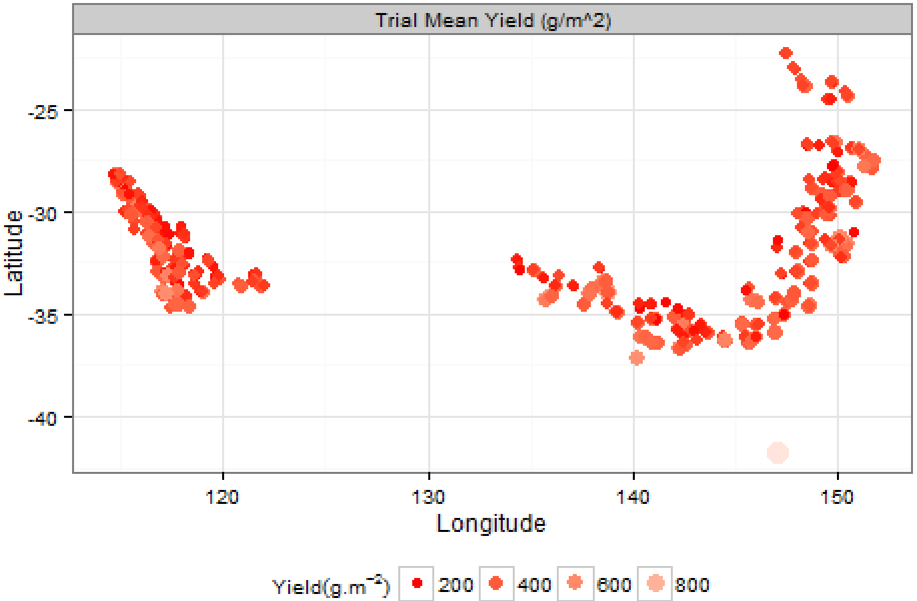
Ky L. Mathews, Arthur Gilmour, Bangyou Zheng, Scott Chapman

CSIRO

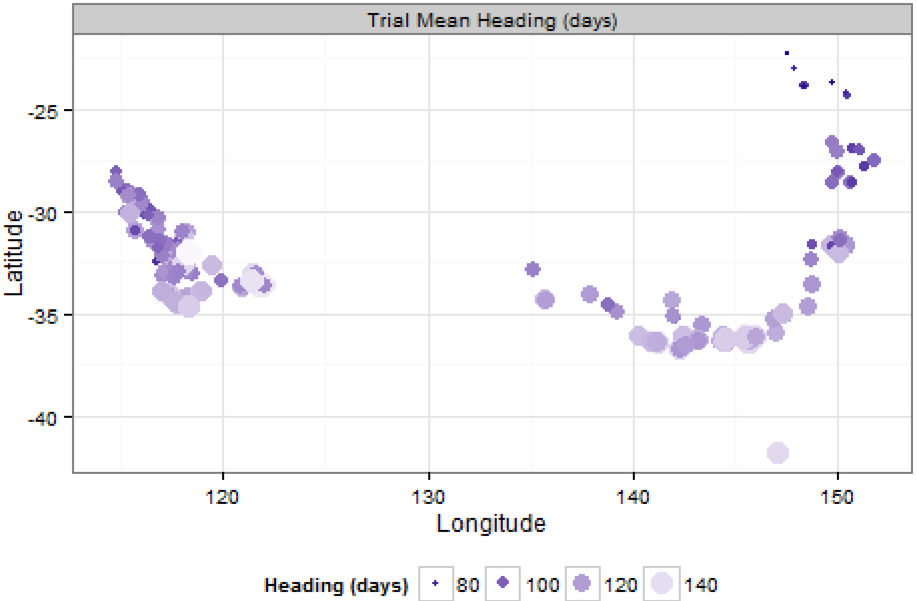
GRDC Grains Research & Development Corporation
Your GRDC working with you

Australian wheat belt characteristics

Yield (g.m^{-2})



Phenology (flowering)



Weather variables

rainfall (mm)

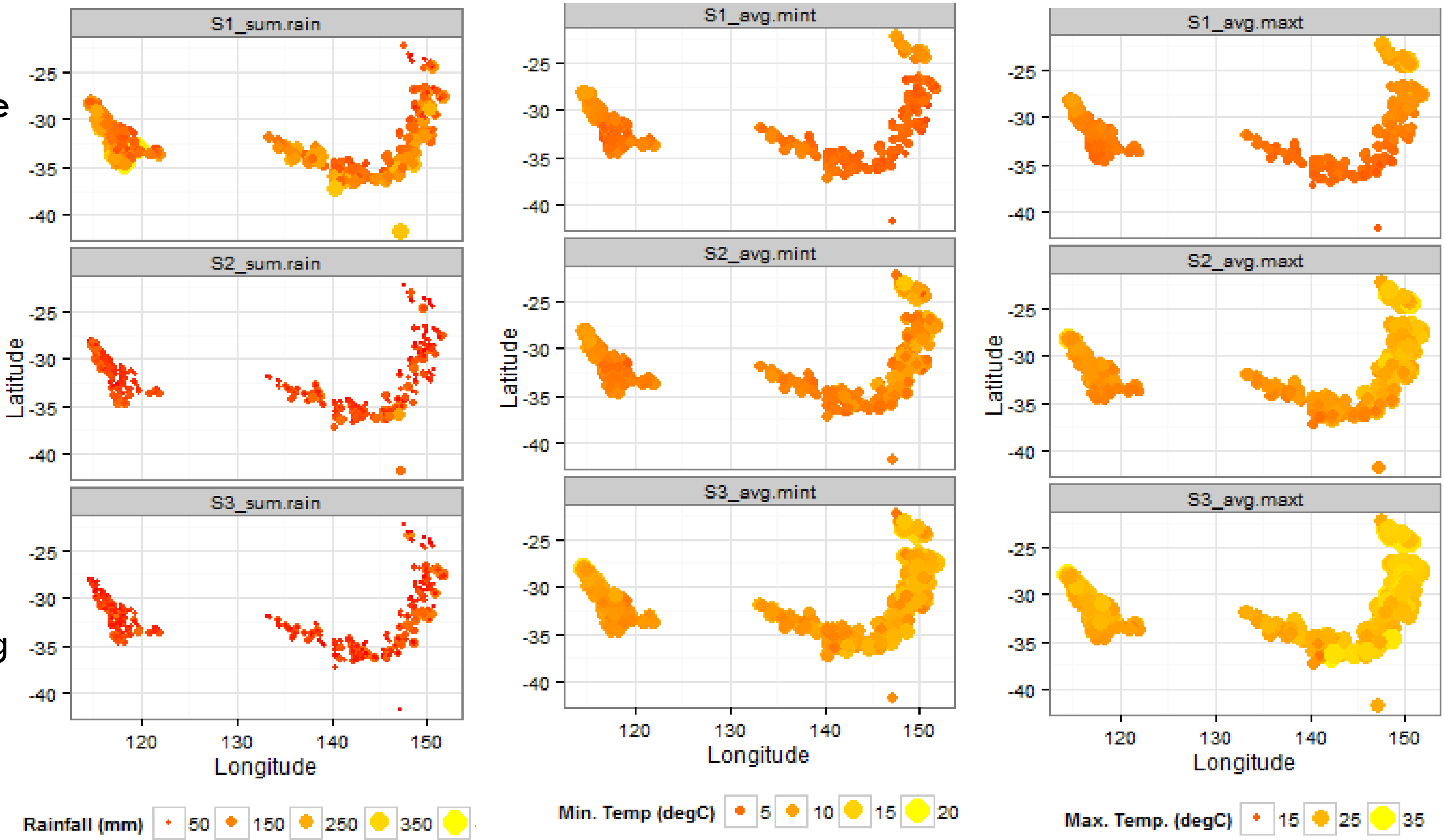
min. temperature

max. temperature

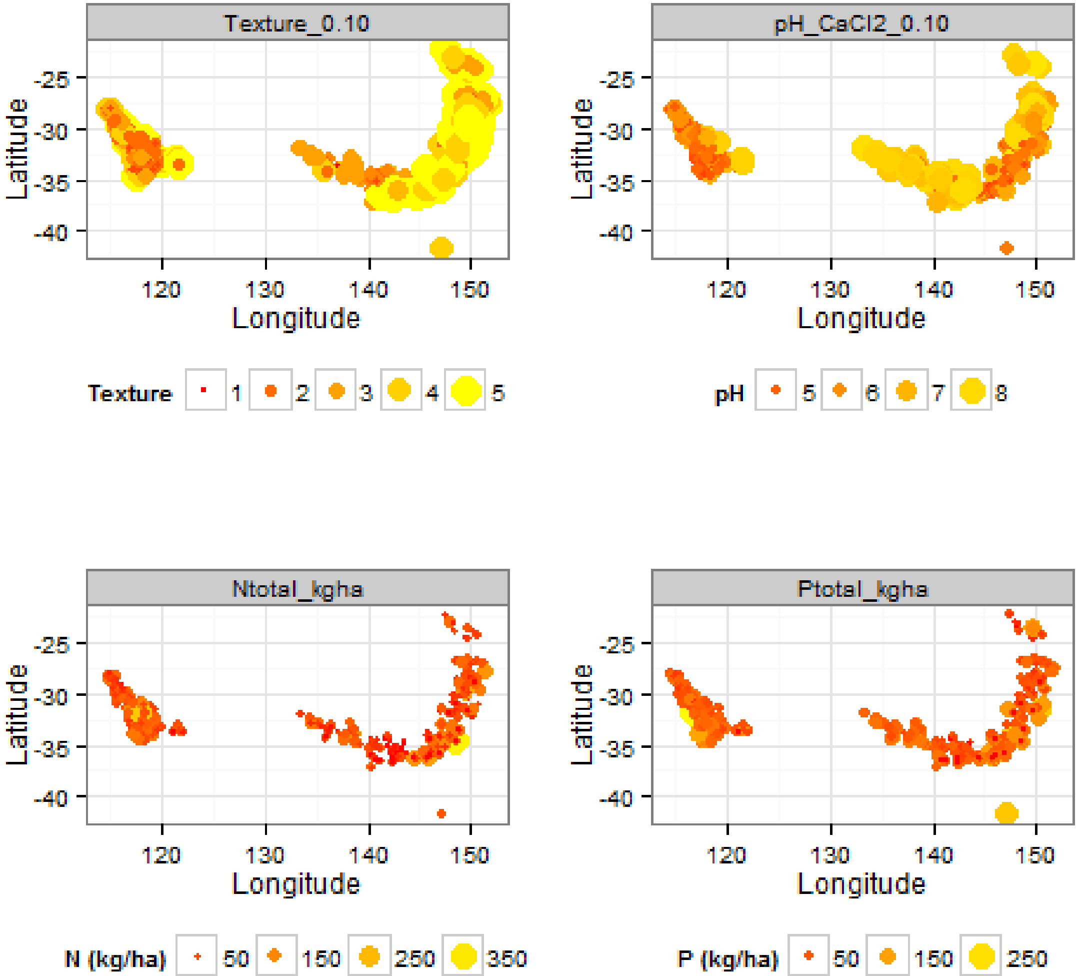
vegetative

flowering

grainfilling

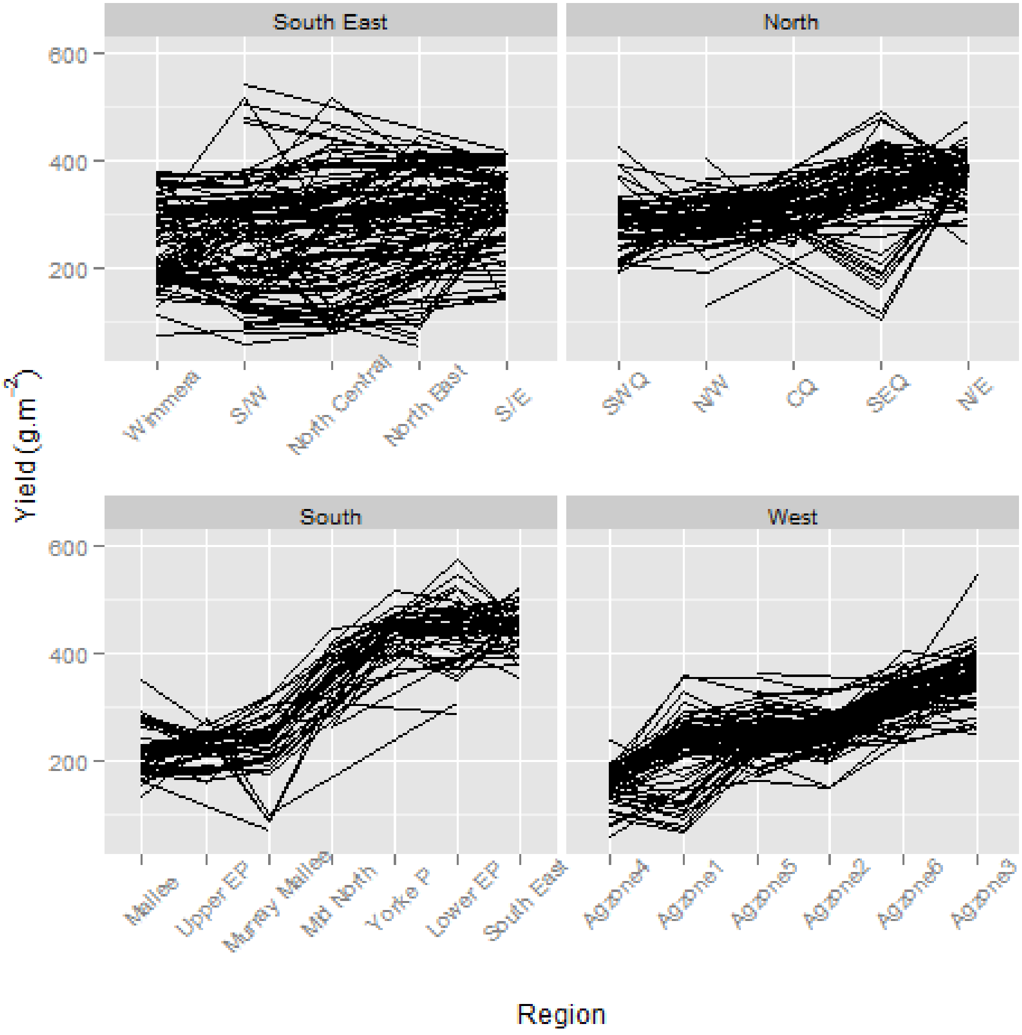


Soil variables



- lots of variability
- nutrient poor
- toxin rich
- hard to measure

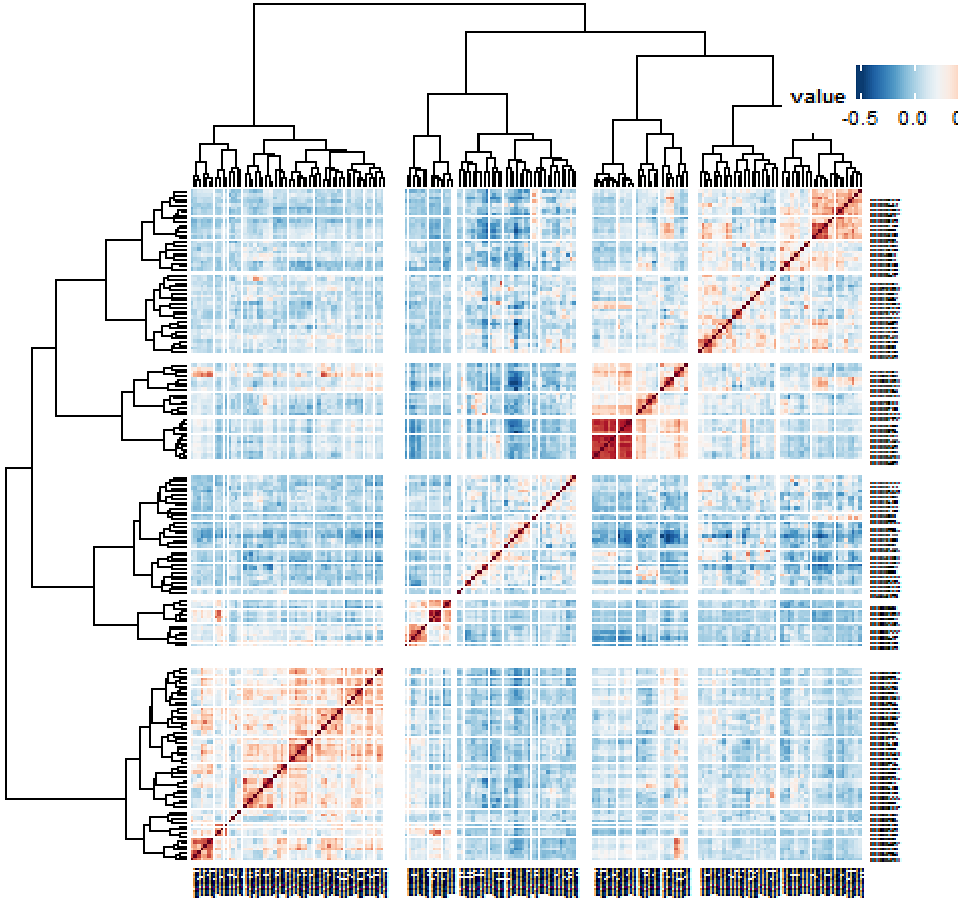
Genotype by Environment Interaction



National Variety Trials
933 trials
187 locations; 24 regions
162 varieties
9 years – 2005-13

Region by variety mean yields showing GxE typical of the Australian wheat belt.

Trial variance-covariance heterogeneity



Clustered phenotypic correlation matrix of the South East dataset

180 trials, 71 varieties

We cannot ignore either the variance or the covariance heterogeneity.

Mixed model framework

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \boldsymbol{\eta}$$

- □ Factor analytic

$$\mathbf{u} = (\boldsymbol{\Lambda} \otimes \mathbf{I}_m)\mathbf{f} + \boldsymbol{\delta} \text{ and}$$

$$\mathrm{var}(\mathbf{u}) = (\Lambda\Lambda' + \psi) \otimes I_m$$

Smith et al (2001) ANZJS & Biometrics
Smith et al (2015) TAG

- □ ASREML-R v 3.0-1 asreml object 4.0jy on the R platform
- □ Starting line: FA of order $k = 1$. Usually the larger k the more variance-covariance is explained. However, here we stop at $k = 1$ so that we have some variability to use the environmental covariates with.
- □ Goalpost: FA1 + V:TMY as an indicator of how much can be explained by environmental covariates – modern day Finlay-Wilkinson model?

Environmental Covariates

- Weather:
  - Rainfall
  - Temperature – min/max/mean
  - Radiation
- Soil
  - Physical – Texture, plant available water capacity
  - Chemical - N, P, K, S, Zn, pH, Total Exchangeable Cations
  - Biological
- Primary and secondary sets of covariates - PCA
- Gene based phenology model to provide us with growth stages, e.g. pre-sowing, vegetative, flowering, grainfilling
- Indices – the sky is the limit!

Cyclic forward regression

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{Z}^{*}\mathbf{u}^{*} + \boldsymbol{\eta}$$

where $\mathbf{u}^{*}$ is the $tmc \times 1$ vector of random variety effects for each environment and each covariate with associated design matrix $\boldsymbol{Z}^{*}$.

There are 4 steps in each cycle:

1. Note the REML LogLikelihood for the current (base) model
2. Obtain the REML LogLikelihood for each candidate covariate (Z^{*}) by adding the random term $Z^{*} \times$ Variety to the current base model
3. Identify several candidate covariates which make a large (and significant) increase in the LogLikelihood but which are not variants of each other
4. Add this set of candidates and then drop any that do not contribute significantly to the enlarged model

□ **STOP** when the REML Loglikelihood is not significant.

□ Output Variety by covariate BLUPS for interpretation.

Cyclic forward regression properties

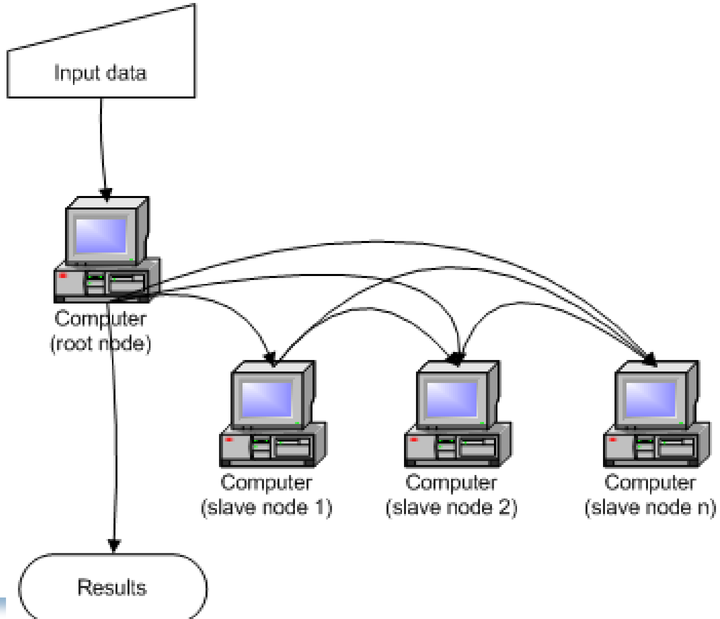
- Heuristic
- Statistician labour intensive
- West: $n_e = 50$, $n_v = 50$ with 15 covariates took 2 days
- SouthEast: $n_e = 141$, $n_v = 71$ with 12 covariates took 5 days
- Forward regression in that mostly terms are not dropped from the model but this is not hard and fast.
- Nested models

Subset regression

- Fits every possible combination of the c environmental covariates, i.e. $\sum_{i=1}^{c} {}^{c}C_i$
  - for 12 covariates = 4095 models!
  - For 15 covariates = $32,767$ models!!!
- The FA1 + covariate model is fitted for each covariate individually. The resulting gammas from these models are used as initial starting values for the $c > 1$ models.
- Each model is then fitted simultaneously across a computer cluster, the results returned and collated together into one file.
- Models with a covariate term that hits the boundary are removed
- The remaining models are inspected for the lowest AIC/BIC

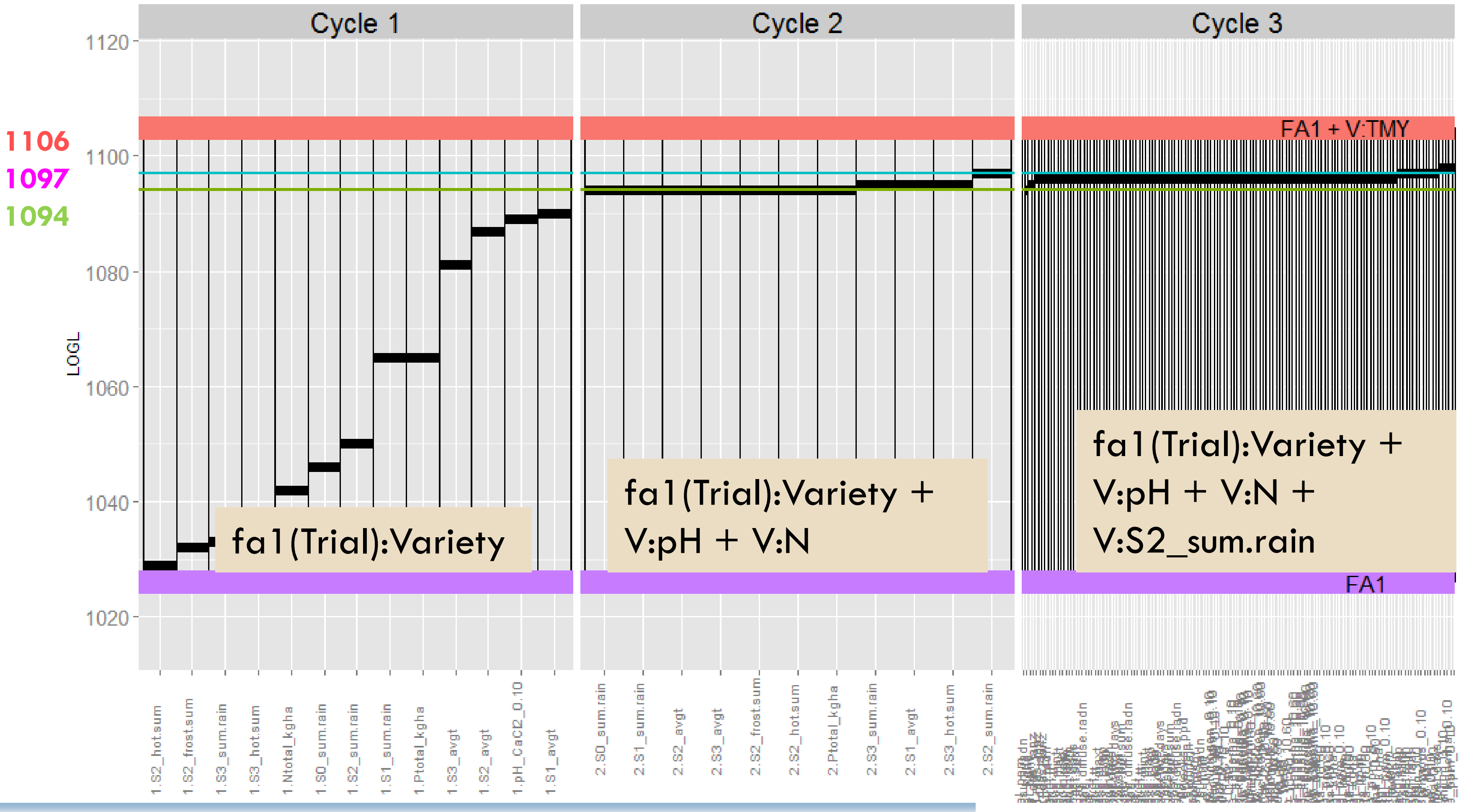
Subset regression selection

- Distribution models into high performance cluster
  - Condor for computing non-intensive models (WA)
    - <10 min for 32767 models with 8000+ cores
  - HPC for computing intensive models (NSW & VIC)
    - 1 hour for the $n_e = 141$, $n_v = 71$ 4095 models with 200 cores
  - Outputs: 1 .csv file 7MB



Cyclic Regression Example - West

- $n_e = 50$, $n_v = 50$, 45% balance, 47% of the GxE expl. by FA1



fa1(Trial):Variety

fa1(Trial):Variety + V:pH + V:N

fa1(Trial):Variety + V:pH + V:N + V:S2_sum.rain

Subset Regression Example - West

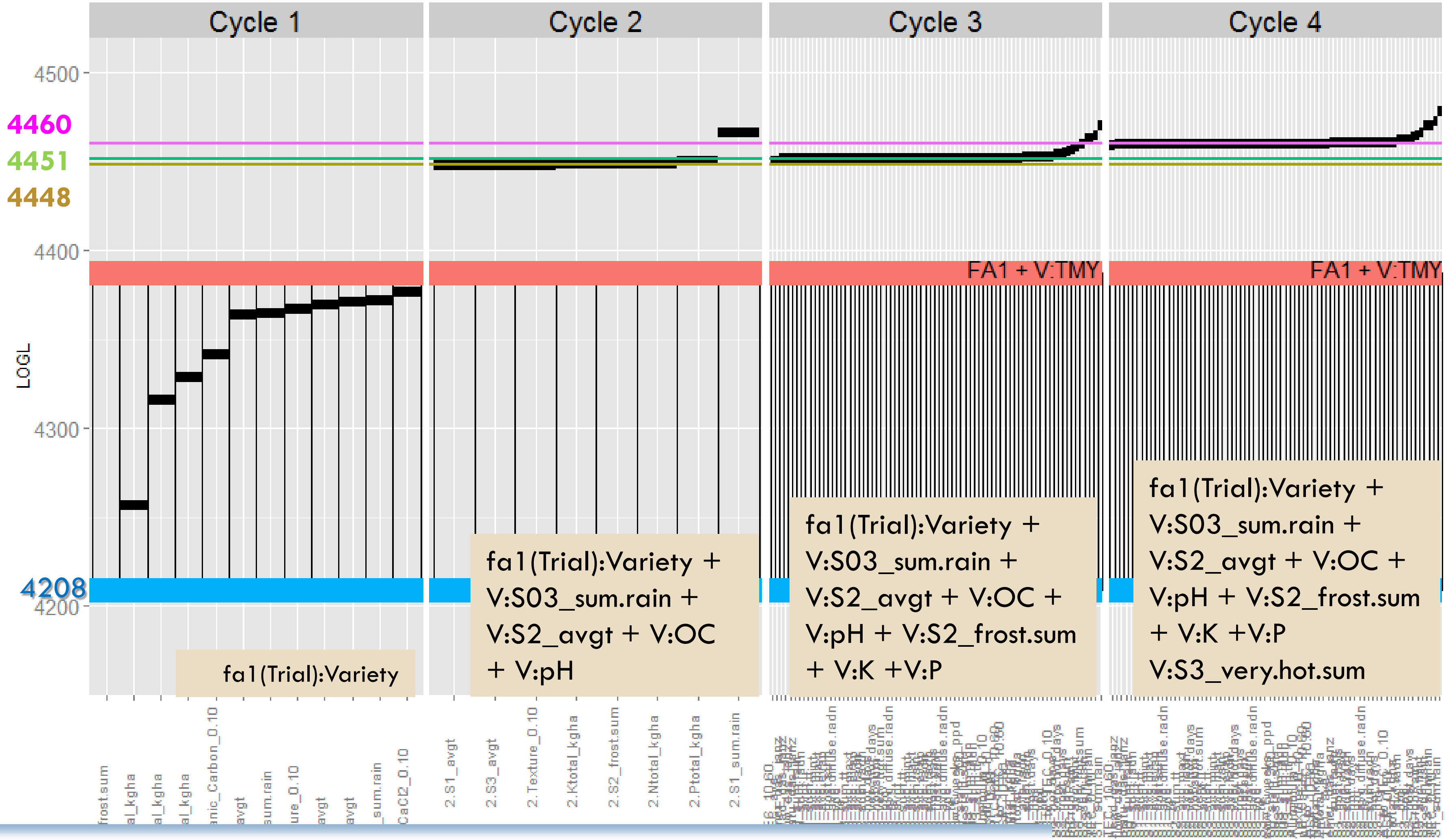
- □ 15 covariates ⟹ 32, 767 different models
- □ Discard any models with boundary terms (1563 models to inspect)
- □ Four models with the same minimum AIC

Model	LogL	AIC	BIC
S2_sum.rain + N + pH	**1097**	**-1985**	**-1468**
S2_sum.rain + S2_frost.sum + N + pH	1097	-1985	-1462
S0_sum.rain + S2_sum.rain + N + pH	1097	-1985	-1462
S0_sum.rain + S2_sum.rain + S1_avgt + N	1097	-1985	-1462

- □ Check correlations between environmental covariates

Cyclic Regression Example – South East

- □ $n_e = 141$, $n_v = 71$, 49% balance; FA1 explains 33%



fa1(Trial):Variety

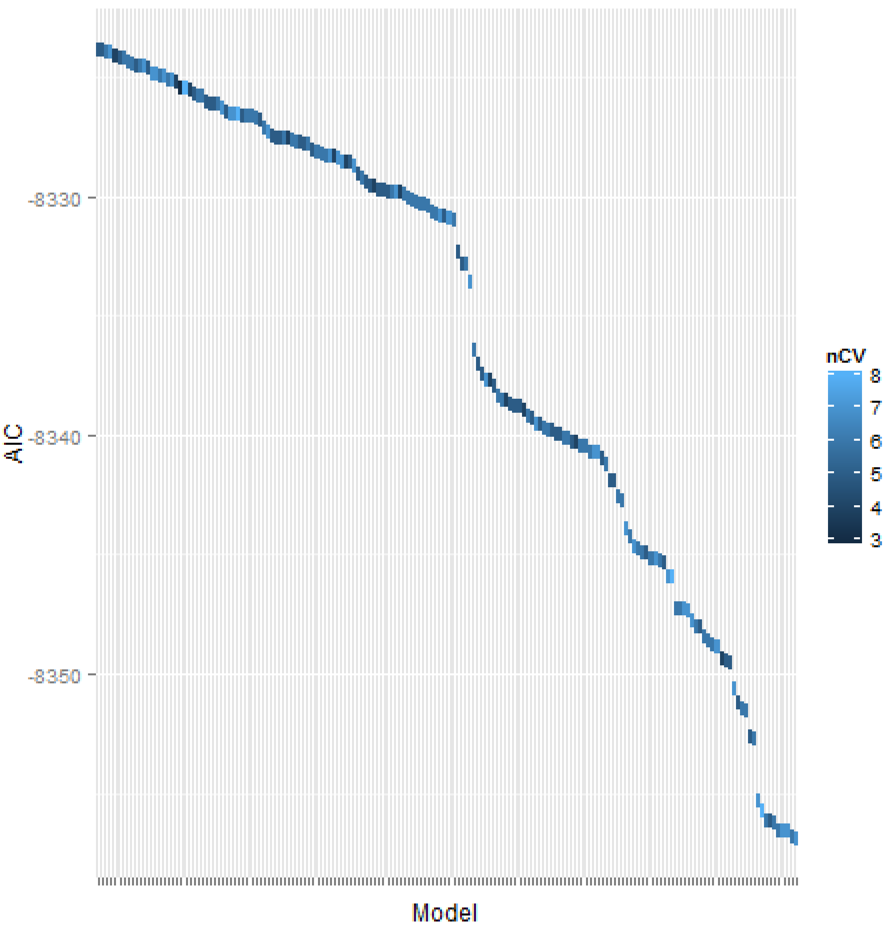
fa1(Trial):Variety + V:S03_sum.rain + V:S2_avgt + V:OC + V:pH

fa1(Trial):Variety + V:S03_sum.rain + V:S2_avgt + V:OC + V:pH + V:S2_frost.sum + V:K +V:P

fa1(Trial):Variety + V:S03_sum.rain + V:S2_avgt + V:OC + V:pH + V:S2_frost.sum + V:K +V:P V:S3_very.hot.sum

Subset Regression Example – South East

- 12 covariates → 4095 models
- Discard any models with boundary terms (1443 models to inspect)
- Best model using AIC is: (LOGL = 4468) FA1 + V:S03_sum.rain + V:S2_avgt + V:OC + V:pH + **V:P + V:K** + V:S1_sum.rain
- Best model using BIC is (LOGL = 4466) FA1 + V:S03_sum.rain + V:S2_avgt + V:OC + V:pH + V:S1_sum.rain
- Plot is 170 models with AIC < cyclic
- Still decisions that need to be made with brains!



Summary of methods

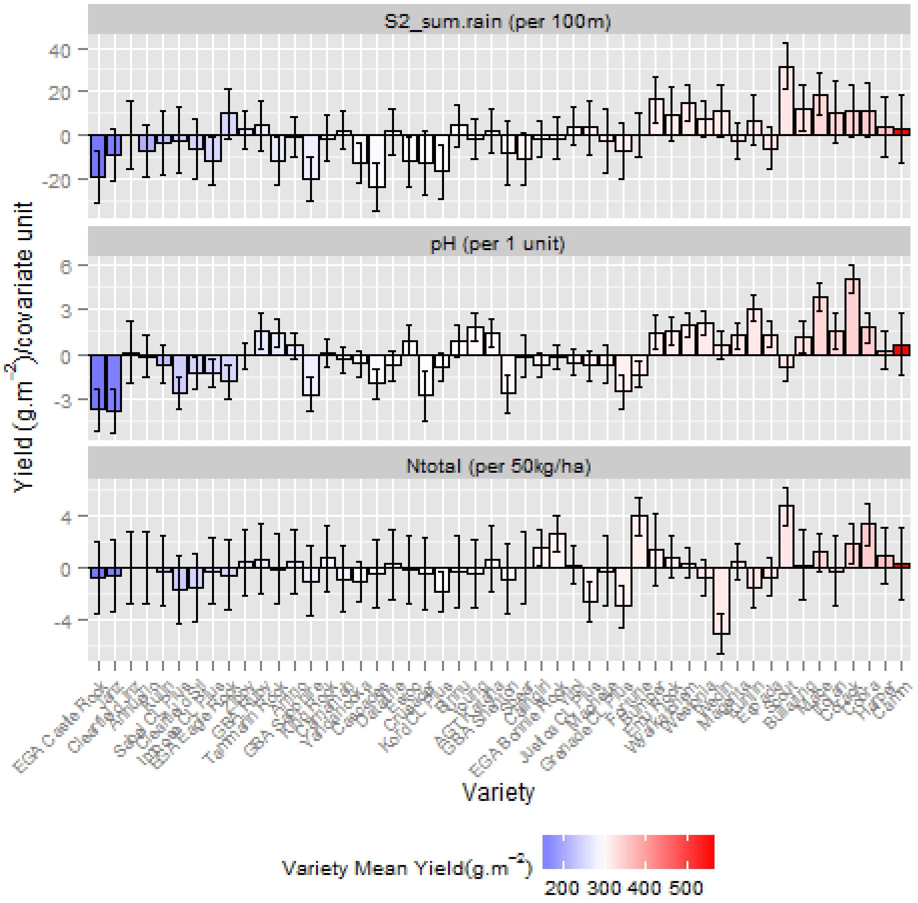
Cyclic Regression

- slower but perhaps a better understanding of the dataset?
- more dependent on the choices made by the statistician/scientist
- no. covariates not limited

Subset Regression

- faster
- tendency to select models with fewer terms which may or may not be helpful
- can always re-run terms in the model
- can use the cyclic regression mindset to make selections
- run lots of models you don’t need – but does it matter???
- no. covariates limited by the system

Model Outputs (West Results)



- □ Variety by covariate predictions
- □ Identify varieties that are sensitive to a particular covariate.
- □ E.g. for every 100mm increase in flowering rain Scout yields 31g.m^{-2} more than the average; for every increase in pH unit it decreases by 0.9g.m^{-2} and for every 50kg/ha increase in avail. N it increases by 4.7g.m^{-2}.

Future work (i.e. questions to answer)

- What’s the best model?
  - Model selection criterion – AIC/BIC/other?
  - Percentage variance explained
- Model relationship between environmental covariates to accommodate the between covariate correlations
- Model traits such as rainfall better, p-splines?
- Predictive power still an issue, e.g. we want to predict how varieties respond if there is a dry mid-season.
- Incorporate flowering information as this is a key driver and used by breeders/growers to manage environmental risk
- Important to work with both physiologists and breeders to make sure it’s not purely data driven

Acknowledgements

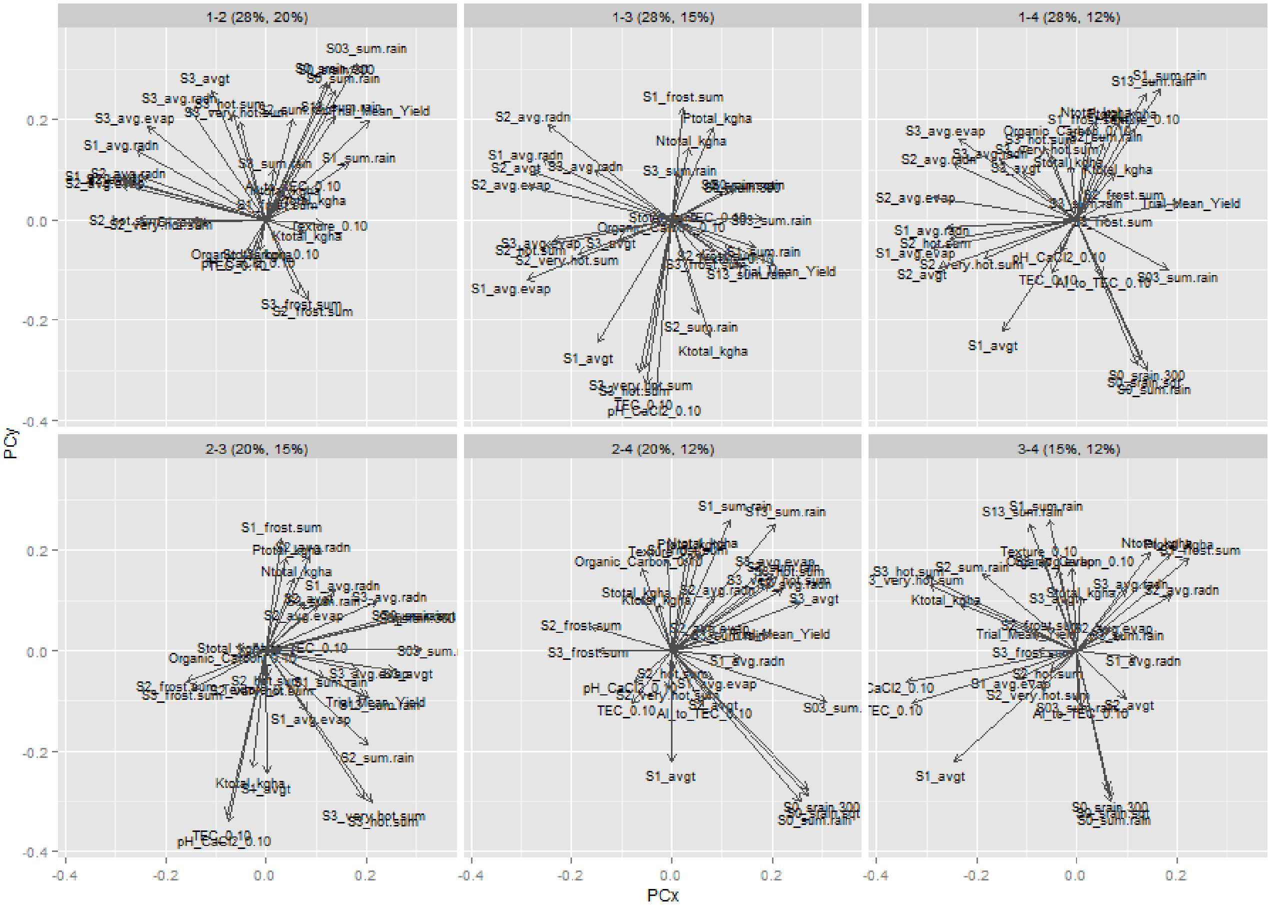
- Neale Sutton (ACAS)
- Co-authors
- GRDC
  - CSA00027: Adding value to GRDC's National Variety Trial network (Zvi Hochman)

CSIRO

GRDC Grains Research & Development Corporation

Your GRDC working with you

Biplots from PCA of env covariates


1-2 (28%, 20%)
1-3 (28%, 15%)
1-4 (28%, 12%)
2-3 (20%, 15%)
2-4 (20%, 12%)
3-4 (15%, 12%)
PCy
PCx
S03_sum.rain
S3_avgt
S3_avg.radn
S3_avg.evap
S1_avg.radn
S3_very.hot.sum
Trial_Mean_Yield
S1_sum.rain
S3_sum.rain
Texture_0.10
Ktotal_kgha
S3_frost.sum
S2_frost.sum
S2_very.hot.sum
S1_frost.sum
Ptotal_kgha
Ntotal_kgha
S2_avg.radn
S2_avgt
S2_avg.evap
S1_avg.evap
S1_avgt
S2_sum.rain
S13_sum.rain
TEC_0.10
pH_CaCl2_0.10
S0_grain.300
S0_sum.rain
Organic_Carbon_0.10
Stotal_kgha
S2_hot.sum
S3_hot.sum
S1_hot.sum
Al_to_TEC_0.10
Stotal_kgha
S1_frost.sum