CORTEVA™
agriscience

Maize Yield Predictions

Results from the 2022 G2F prediction competition

Alencar Xavier[1,2], Wesley Barber[1], Cristiano Zimmer[1], Fabiani Rocha[1], Ignacio Trucillo[1], Abelardo de la Vega[1], Jim Holland[3], Jacob Washburn[3], Jose Varella[3], Natalia de Leon[4,] Dayane Lima[4], David Ertl[5], Joseph Gage[6], Qiuyue Chen[6], Cinta Romay[7].

1 Corteva Agrisciences; 2 Purdue University; 3 USDA-ARS; 4 UW Madison; 5 National Corn Growers Associations; 6 NCSU; 7 Cornell;

2023 Maize Genetics Conference

Outline

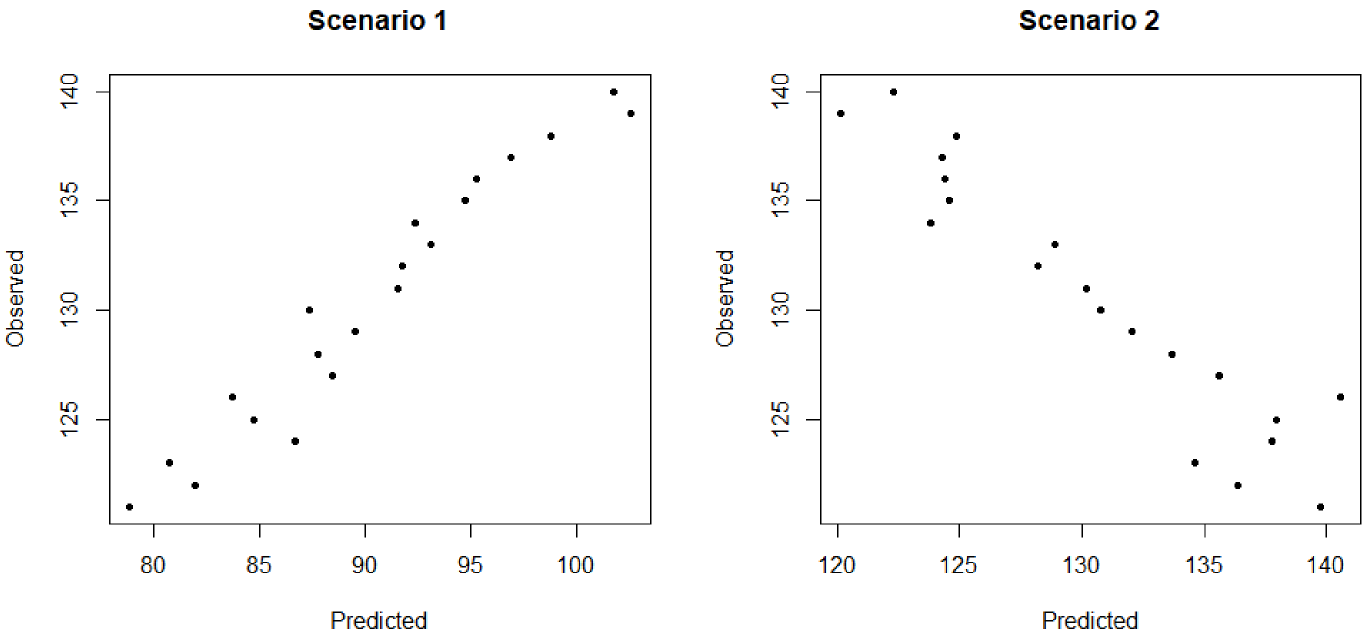
1. **Introduction**
    - Our team
    - Evaluation criterion
2. **Data**
    - Information
    - Population (TPG, TPE)
3. **Modeling and QC**
    - What to model
    - Genomic information
    - Statistical model
4. **Statistical models**
    - Model A
    - Model B
    - Submissions
5. **Conclusion**

1. **Introduction**
    - Our team
    - Evaluation criterion
2. **Data**
    - Information
    - Population (TPG, TPE)
3. **Modeling and QC**
    - What to model
    - Genomic information
    - Statistical model
4. **Statistical models**
    - Model A
    - Model B
    - Submissions
5. **Conclusion**

CLAC team
(Corteva Latin America Corn)

- Alencar Xavier – Breeding Analyst LAAF, adj. prof. Purdue 🇧🇷
- Wesley Barber – Safrinha corn breeder, EZ lead (Brazil, south) 🇺🇸
- Cristiano Zimmer – Tropical corn breeder (Brazil, central) 🇧🇷
- Fabiani Rocha – Subtropical corn breeder (Brazil, south) 🇧🇷
- Ignacio Trucillo – Temperate corn breeder (Argentina) 🇦🇷

Evaluation criterion

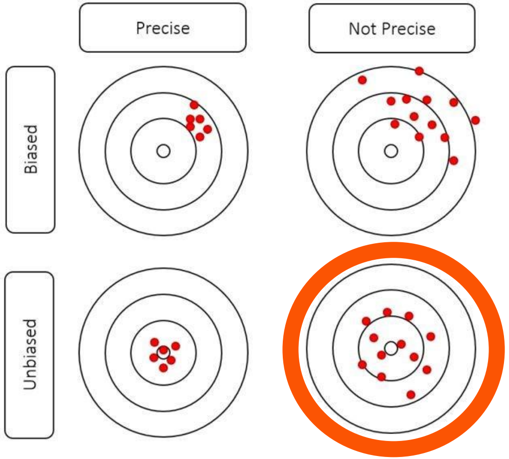

Scenario 1

Scenario 2

Some considerations

- Ranking matter less than getting environmental means correctly

- Shrinkage is also our enemy
  - Genomic BLUPs must be rescaled into phenotypic variance
  - Environmental means predicted via machine learning too

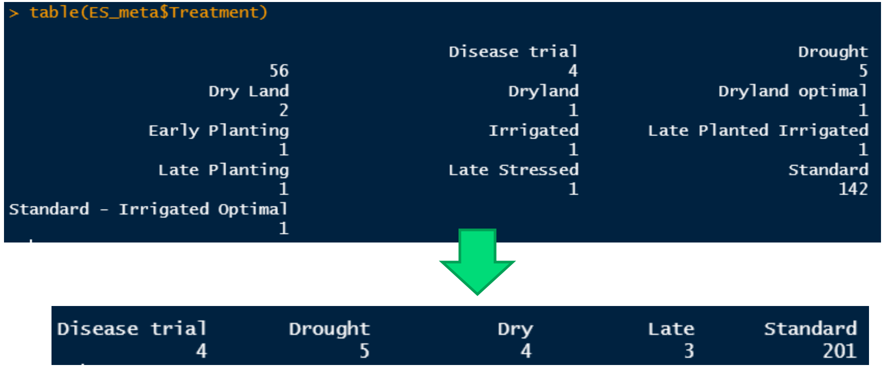


Picture source

1. **Introduction**
   - Our team
   - Evaluation criterion
2. **Data**
   - Information
   - Population (TPG, TPE)
3. **Modeling and QC**
   - What to model
   - Genomic information
   - Statistical model
4. **Statistical models**
   - Model A
   - Model B
   - Submissions
5. **Conclusion**

Information

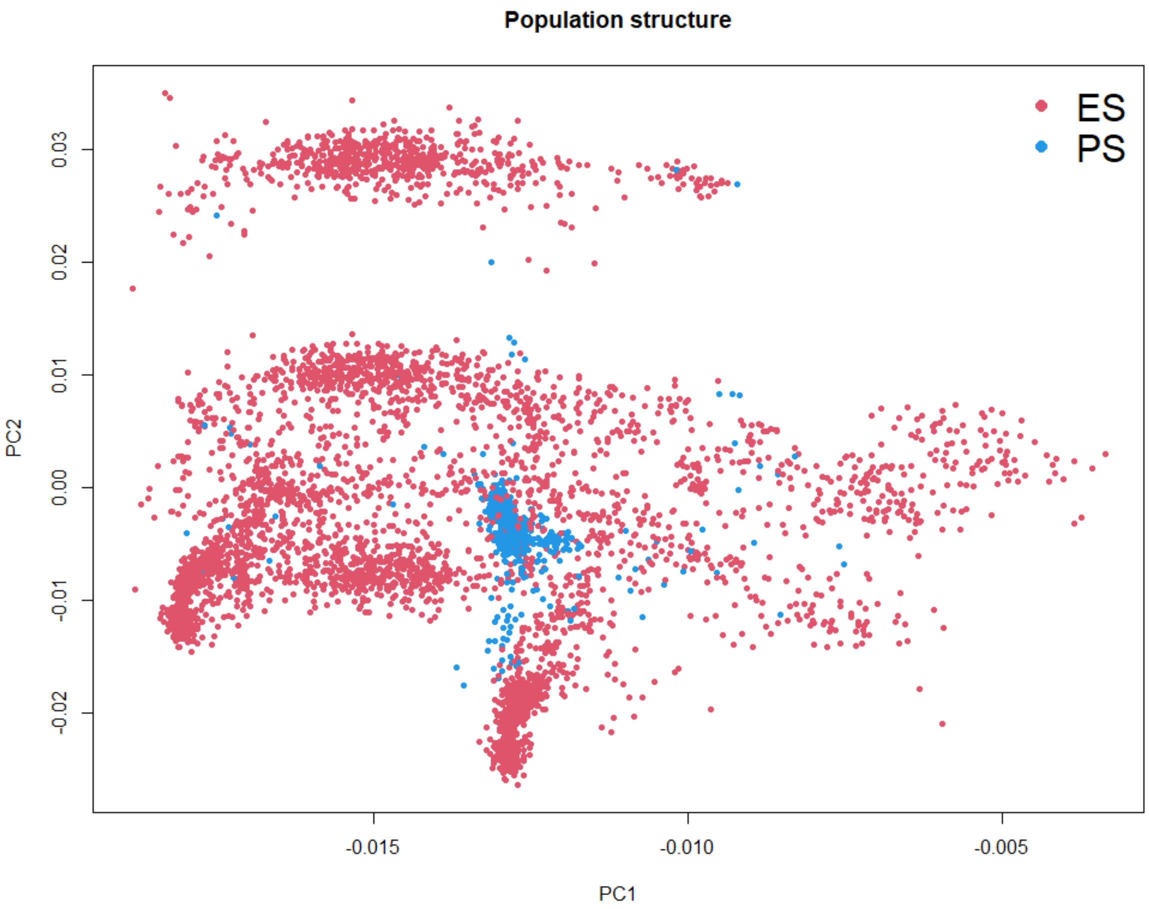
- Resources:
  - Genomic information (400K markers, 150K after QC)
  - Environmental covariates (EnvRtype), weather from NASA power, soil data
  - Meta data: Irrigation, Treatment



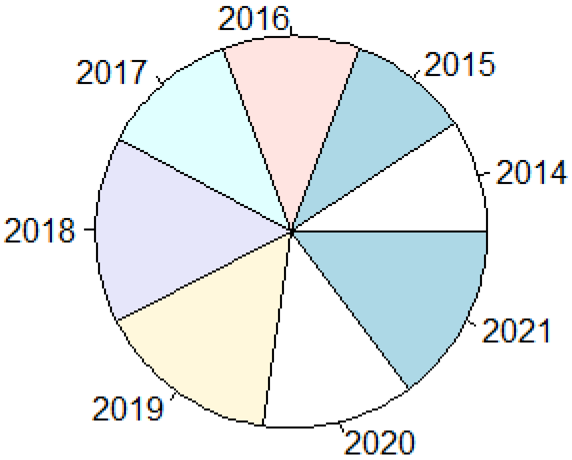
```
> table(ES_meta$Treatment)

                                            Disease trial                      Drought
                          56                            4                            5
                    Dry Land                      Dryland              Dryland optimal
                           2                            1                            1
              Early Planting                    Irrigated       Late Planted Irrigated
                           1                            1                            1
               Late Planting                Late Stressed                     Standard
                           1                            1                          142
Standard - Irrigated Optimal
                           1
```

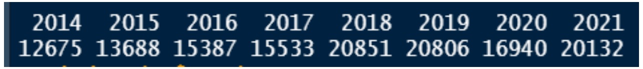
```
Disease trial         Drought             Dry            Late        Standard
            4               5               4               3             201
```

TPG

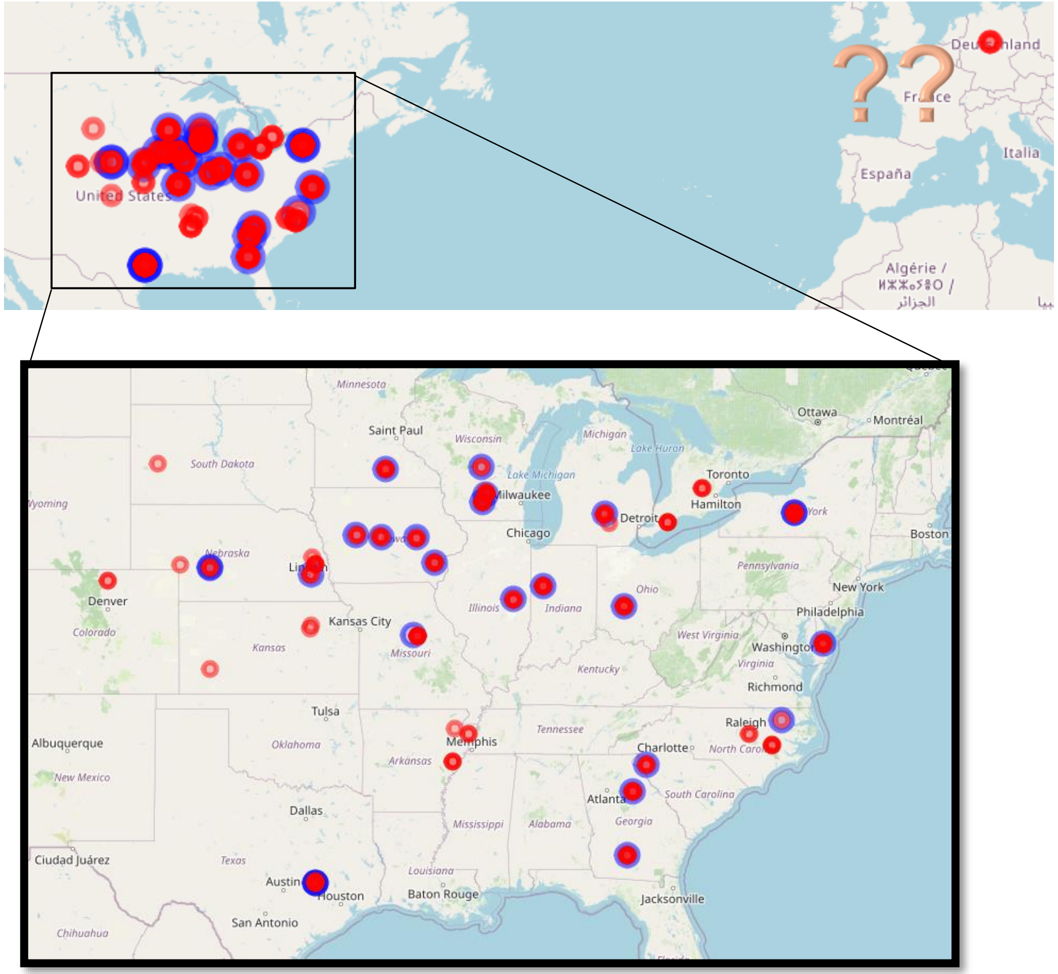

Population structure
PC2
0.03
0.02
0.01
0.00
-0.01
-0.02
PC1
-0.015
-0.010
-0.005
● ES
● PS

TPE





```
 2014  2015  2016  2017  2018  2019  2020  2021
12675 13688 15387 15533 20851 20806 16940 20132
```



1. **Introduction**
    - Our team
    - Evaluation criterion
2. **Data**
    - Information
    - Population (TPG, TPE)
3. **Modeling and QC**
    - What to model
    - Genomic information
    - Statistical model
4. **Statistical models**
    - Model A
    - Model B
    - Submissions
5. **Conclusion**

What to model?

$$y|E_i = \mu_i + g|E_i$$

$$\text{Phenotype @ i}^{\text{th}}\text{ Loc } = \text{i}^{\text{th}}\text{ Loc Mean} + \text{Genetic effect @ i}^{\text{th}}\text{ Loc}$$

- Prediction set:
  - 11555 observations
  - 26 locations
  - 548 genotypes (43 were observed in the ES)
  - Ranging from 336 to 530 GE/AOI

Genomic information

- Just too much SNP data for ~5K hybrids (**not parents, no GCA**)

- QC’ed based on LD and MAF
  - MAF of 0.05 in both training and prediction sets
  - Reduced data to ~150K

- **Need reparameterization!!**

Genomic information

(Kernel trick + Kernel-to-X trick)

ArcCos relationship matrix

$$A = f(Z)$$

where:
Z (5K x 300K)
A (5K x 5K)

EVD

$$A = UD^2U'$$
$$A = (UD)(UD)'$$
$$A = MM'$$

where $M = UD$ (5K x 5K)

- So that

$$y = \mu + Ma + e$$
$$\text{var}(y) = MM'\sigma_g^2 + I\sigma_e^2$$

= Ridge regression

$= A\sigma_g^2 + I\sigma_e^2 = \text{GBLUP}$

Formatting the data

- **State, Stations**: Obtained from Env names: OHH1_2020 = Ohio, H1

- **Pooling levels**
  - **Previous crop** (peanut, soy, etc. = “Legume”)
  - **Treatment** (early, irrigated, Standard = “Standard”)
  - **Irrigation**: Irrigated in meta data or treatment says it is irrigated
  - Level with too few locations were set as unknown

Loc with irrigation:

- 3 of 217 locs in ES
- 9 of 26 locs in PS

- **Yield outliers**: More than 3 Std Dev within location

Prediction model

Our final predictions were the average of two models

$$\hat{\mathbf{y}} = \frac{\hat{\mathbf{y}}_{\mathrm{A}} + \hat{\mathbf{y}}_{\mathrm{B}}}{2}$$

Model A – Simple univariate GBLUP, no spatial

Model B – Highly processed/engineered GBLUP model

1. **Introduction**
   - Our team
   - Evaluation criterion
2. **Data**
   - Information
   - Population (TPG, TPE)
3. **Modeling and QC**
   - What to model
   - Genomic information
   - Statistical model
4. **Statistical models**
   - Model A
   - Model B
   - Submissions
5. **Conclusion**

Model A

- Simple GBLUP

$$y = X\beta + Zu + e$$
$$u \sim N(0, MM'\sigma_u^2)$$
$$e \sim N(0, I\sigma_e^2)$$

Terms

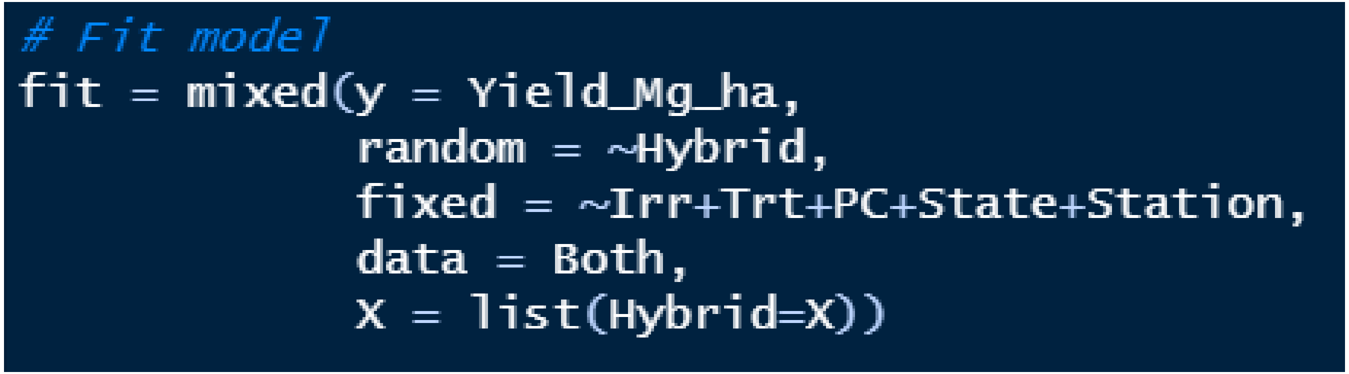
- Fixed terms ($X\beta$): State, station, treatment, irrigation, previous crop
- Random (Zu): Hybrid
- No year component

Model A

- Fixed terms were found significant via LRT

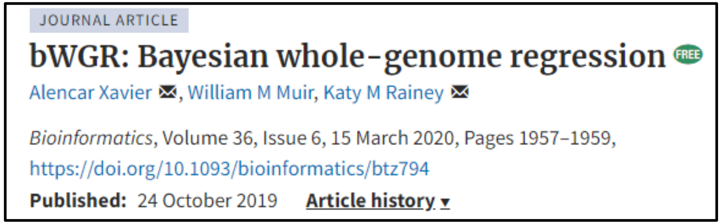
  - H_0: Yield = Env + Hyb
  - H_1: Yield = Env + Hyb + PC (p-value H0-H1: 0.0082**)
  - H_2: Yield = Env + Hyb + PC + Irr (p-value H1-H2 : 0.0002*)
  - H_3: Yield = Env + Hyb + PC + Irr + Trt (p-value H2-H3: 0.09454.)

Model A



```
# Fit model
fit = mixed(y = Yield_Mg_ha,
            random = ~Hybrid,
            fixed = ~Irr+Trt+PC+State+Station,
            data = Both,
            X = list(Hybrid=X))
```

R package bWGR



JOURNAL ARTICLE

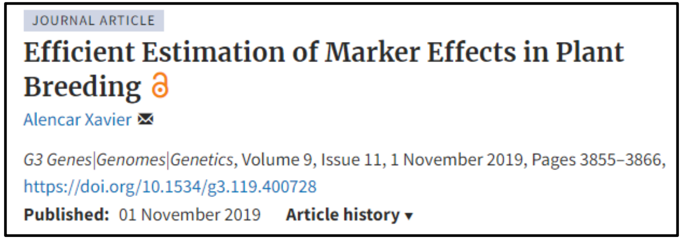
bWGR: Bayesian whole-genome regression FREE

Alencar Xavier ✉, William M Muir, Katy M Rainey ✉

Bioinformatics, Volume 36, Issue 6, 15 March 2020, Pages 1957–1959, https://doi.org/10.1093/bioinformatics/btz794

Published: 24 October 2019 **Article history ▾**

Method / Solver



JOURNAL ARTICLE

Efficient Estimation of Marker Effects in Plant Breeding

Alencar Xavier ✉

G3 Genes|Genomes|Genetics, Volume 9, Issue 11, 1 November 2019, Pages 3855–3866, https://doi.org/10.1534/g3.119.400728

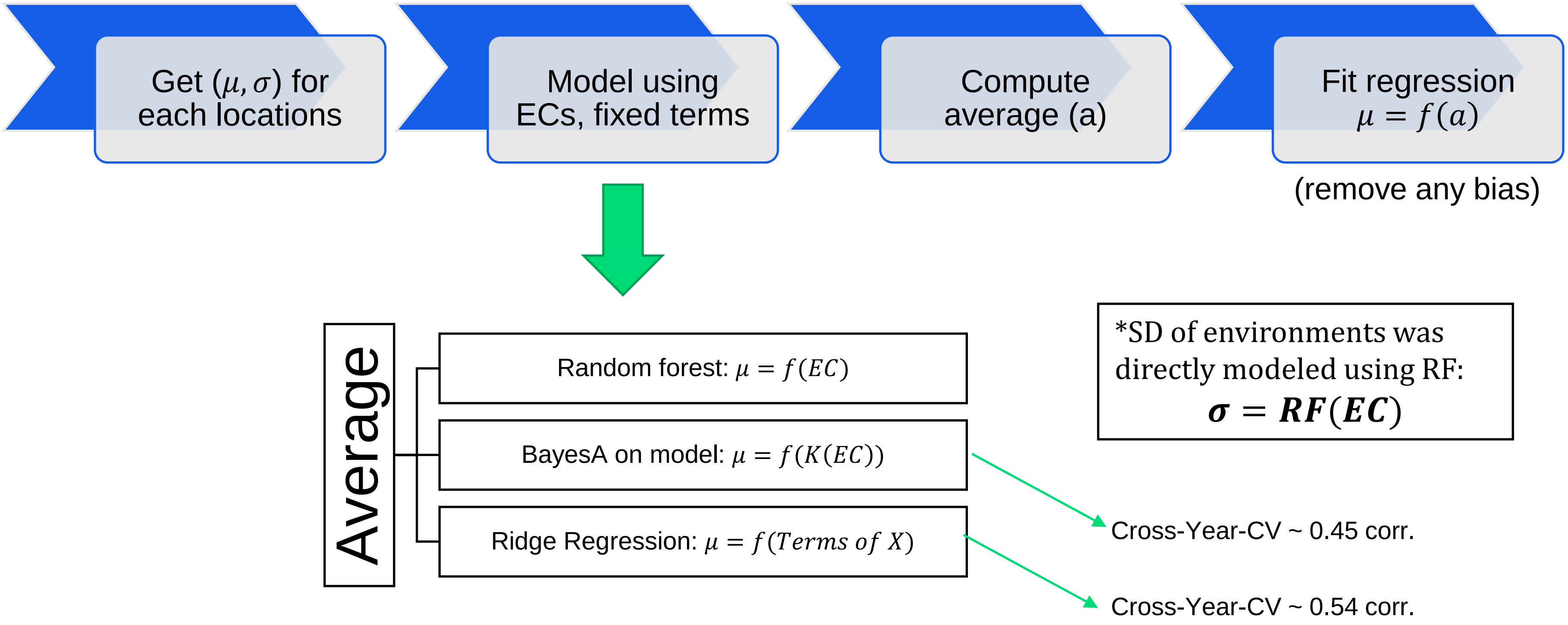
Published: 01 November 2019 **Article history ▾**

Model B

$$y = \mu + g\sigma + e$$

- Three separate models:
  - Location mean (μ) - f(EC, meta data)
  - Location variance (σ) - f(EC)
  - Normalized hybrid prediction (g) - f(SNPs, spatial)

Model B – Predict mean and SD



Get (μ, σ) for each locations

Model using ECs, fixed terms

Compute average (a)

Fit regression $\mu = f(a)$

(remove any bias)

Average

Random forest: $\mu = f(EC)$

BayesA on model: $\mu = f(K(EC))$

Ridge Regression: $\mu = f(Terms\ of\ X)$

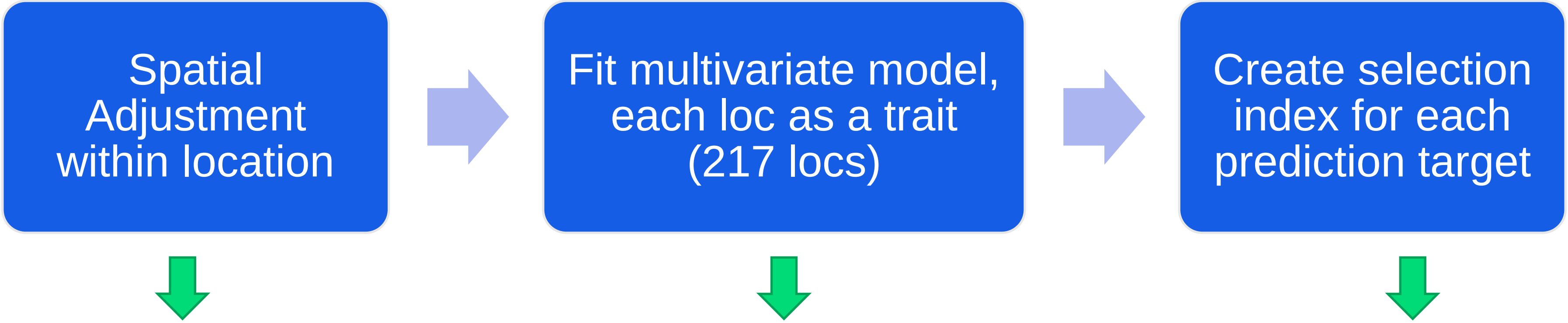
Cross-Year-CV ~ 0.45 corr.

Cross-Year-CV ~ 0.54 corr.

*SD of environments was directly modeled using RF:

$$\boldsymbol{\sigma = RF(EC)}$$

Model B – Hybrid prediction

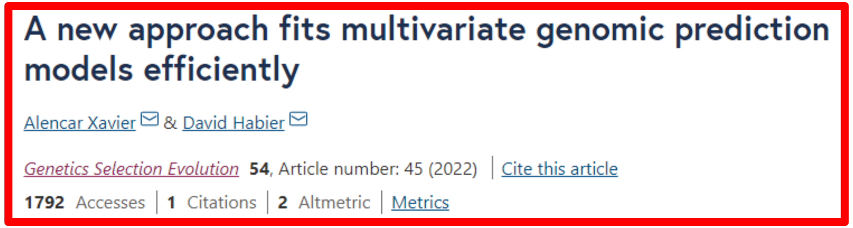


Spatial Adjustment within location

Fit multivariate model, each loc as a trait (217 locs)

Create selection index for each prediction target

$f(Row, Column)$



A new approach fits multivariate genomic prediction models efficiently

Alencar Xavier & David Habier

Genetics Selection Evolution **54**, Article number: 45 (2022) | Cite this article

1792 Accesses | **1** Citations | **2** Altmetric | Metrics

$f(Acc, State, Station)$

(Available in the R package bWGR)

Model B – Hybrid prediction

$$y = \mu + \mathbf{X}\beta + e$$

- Where $y = \{y_1, y_2, \dots, y_K\}$, $\mu = \{\mu_1, \mu_2, \dots, \mu_K\}$, $\beta = \{\beta_1, \beta_2, \dots, \beta_K\}$,

  $e = \{e_1, e_2, \dots, e_K\}$, $Z = \text{BlockDiag}\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_K\}$
- Variances:

$$\Sigma_\beta = \begin{bmatrix} \sigma^2_{\beta(1)} & \cdots & \sigma_{\beta(1,K)} \\ \vdots & \ddots & \vdots \\ \sigma_{\beta(K,1)} & \cdots & \sigma^2_{\beta(K)} \end{bmatrix} \text{ and } \Sigma_e = \begin{bmatrix} \sigma^2_{e(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{e(K)} \end{bmatrix}$$

More information on the approach:

Slides - https://github.com/alenxav/Lectures/blob/master/UIUC_2022/AX_UIUC_2022.09.16.pdf
Paper - https://gsejournal.biomedcentral.com/articles/10.1186/s12711-022-00730-w

Model B – Hybrid prediction

- Selection index:
  - We calculated the deterministic accuracy ($\mathbf{A}$) between every pair of training environment and testing environment

$$\mathrm{a_{PS|ES}} = \mathrm{cor}(\mathrm{g_{PS}}, \hat{\mathrm{g}}_{\mathrm{ES}}) = \frac{\mathrm{cov}(\mathrm{g_{PS}}, \hat{\mathrm{g}}_{\mathrm{ES}})}{\mathrm{sd}(\mathrm{g_{PS}})\mathrm{sd}(\hat{\mathrm{g}}_{\mathrm{ES}})} = \sqrt{\frac{\mathrm{cov}(\mathrm{g_{PS}}, \hat{\mathrm{g}}_{\mathrm{ES}})}{\mathrm{v}(\hat{\mathrm{g}}_{\mathrm{ES}})}} = \sqrt{\frac{\mathrm{G_{PS,ES}V_{ES}^{-1}G_{ES,PS}}}{\mathrm{G_{PS,PS}}}}$$

  - Index

$$\mathrm{g_{PS|ES_i}} = 0.1\ \mathrm{a_{PS|ES_i}} + 2\ \mathrm{a_{PS|ES_i}}(\text{if same state}) + 2\ \mathrm{a_{PS|ES_i}}(\text{if same station})$$

The index is where/how we are capturing GxE!!

Why would multivariate be any better?

Simple (bivariate) model:

INFORMATION GAIN

$$\mathbf{y} = \mathbf{g} + \mathbf{e}$$

$$Var\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma^2_{a_2} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma^2_{e_2} \end{bmatrix}$$

Why would multivariate be any better?

$$y = Zg + e, \qquad y \sim N(0, V)$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

- Covariance structure

$$V = G \otimes \Sigma_a + I \otimes \Sigma_e = G \otimes \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma_{a_2}^2 \end{bmatrix} + I \otimes \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma_{e_2}^2 \end{bmatrix}$$

- Mixed model equation

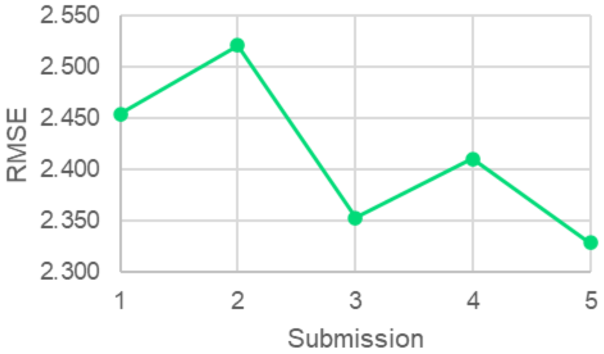
$$\begin{bmatrix} Z_1'\Sigma_e^{11}Z_1 + G^{-1}\Sigma_a^{11} & Z_1'\Sigma_e^{12}Z_2 + G^{-1}\Sigma_a^{12} \\ Z_2'\Sigma_e^{12}Z_1 + G^{-1}\Sigma_a^{12} & Z_2'\Sigma_e^{11}Z_2 + G^{-1}\Sigma_a^{22} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} Z_1'(\Sigma_e^{11}y_1 + \Sigma_e^{12}y_2) \\ Z_2'(\Sigma_e^{22}y_2 + \Sigma_e^{12}y_1) \end{bmatrix}$$

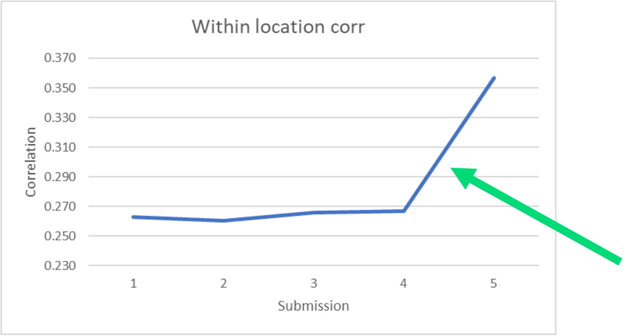
- **Univariate vs bivariate**

$$\begin{aligned} g_1 &= (Z_1'\Sigma_e^{11}Z_1 + G^{-1}\Sigma_a^{11})^{-1}(Z_1'\Sigma_e^{11}y_1) \\ g_1|g_2 &= (Z_1'\Sigma_e^{11}Z_1 + G^{-1}\Sigma_a^{11})^{-1}(Z_1'(\Sigma_e^{11}y_1 + \overbrace{\Sigma_e^{12}y_2) - (Z_1'\Sigma_e^{12}Z_2 + G^{-1}\Sigma_a^{12}}^{\text{INFORMATION GAIN}})g_2) \end{aligned}$$

Submissions

Submission	RMSE	Description
S1	2.454	GBLUP cooked with minor QC and some ECs
S2	2.521	QC'ed data for GEBVs and location means; No Ecs
S3	2.353	Average S1 and S4
S4	2.410	QC'ed GEBVs; Non-QC'ed location means; No ECs
S5	2.239	Average S1 and uncooked univariate GBLUP





Impact of averaging with simple GBLUP

Submissions

Ranking with other metrics
(post-competition analysis)

Realized results

Team Name	Within RMSE
CLAC	2.329
igorkf	2.345
phenomaize	2.374
UCD_MegaLMM	2.387
CGM	2.391
breedingteam	2.398
Purdue	2.402
SmAL	2.425
ML_APT	2.472
MPB_Group	2.544

Ranking with alternative metrics

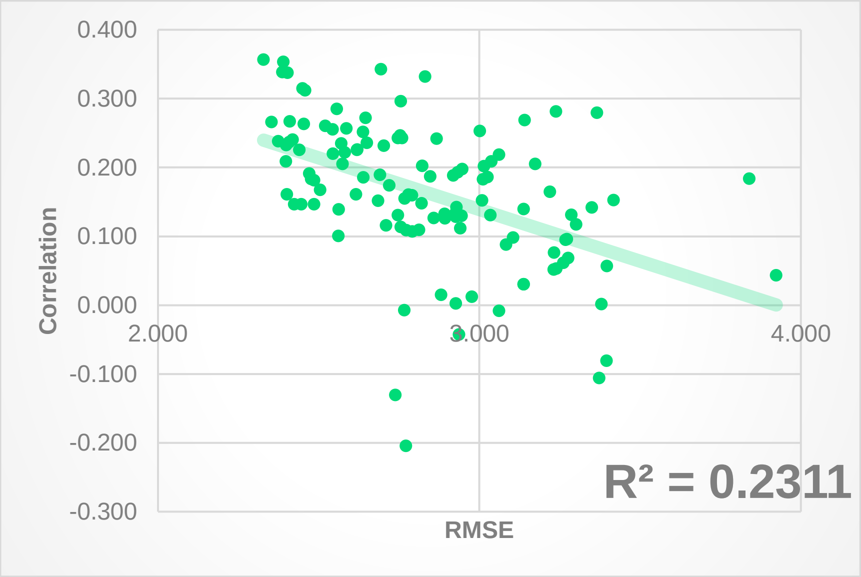
Team Name	Cor Within Loc
CLAC	0.357
CGM	0.353
MPB_Group	0.342
UCD_MegaLMM	0.338
SmAL	0.285
DeepCropVision	0.281
CropEnthusiast	0.279
AllModelsAreWrong	0.272
DataJanitors	0.256
supermanwasd	0.243

Team Name	Cor Across Loc
breedingteam	0.650
DataJanitors	0.644
CLAC	0.631
Purdue	0.631
UCD_MegaLMM	0.628
phenomaize	0.617
igorkf	0.600
CGM	0.587
SmAL	0.586
AllModelsAreWrong	0.575

(Doing well because of CLAC's 5th submission)

Source: Jacob Washburn, Jose Ignacio Varela, Alencar Xavier

RMSE vs Corr



(Sorted by RMSE)

Team Name	RMSE	WL Corr	Subm
CLAC	2.329	0.357	**5**
igorkf	2.345	-	4
CLAC	2.353	0.266	3
igorkf	2.355	-	2
phenomaize	2.374	0.238	6
UCD_MegaLMM	2.387	0.338	3
CGM	2.391	0.353	1
breedingteam	2.398	0.208	1
breedingteam	2.399	0.232	4
Purdue	2.402	0.161	1
UCD_MegaLMM	2.404	0.337	5
breedingteam	2.408	0.237	6
CLAC	2.410	0.267	4
igorkf	2.414	-	5
phenomaize	2.419	-	5
breedingteam	2.420	0.240	5
SmAL	2.425	0.146	4
igorkf	2.441	0.225	3
SmAL	2.446	0.146	1
phenomaize	2.448	-	4
…			

1. **Introduction**
    - Our team
    - Evaluation criterion
2. **Data**
    - Information
    - Population (TPG, TPE)
3. **Modeling and QC**
    - What to model
    - Genomic information
    - Statistical model
4. **Statistical models**
    - Model A
    - Model B
    - Submissions
5. **Conclusion**

Thank you for your attention!

Final remarks:

1) The evaluation metric values locations means
2) ES-PS shared same locations environments
3) Our best model was an average of two GBLUPs

Questions??

Alencar Xavier

Alencar.Xavier@Corteva.com

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF