Statistical description of the Corteva Latin America Corn (CLAC) model for the G2F GxE prediction competition

Alencar Xavier <alenxav@gmail.com>; Barber, Wesley <wesley.barber@corteva.com>; Zimmer, Cristiano <cristiano.zimmer@corteva.com>; Trucillo Silva, Ignacio <ignacio.trucillo@corteva.com>; Rocha, Fabiani <fabiani.rocha@corteva.com>

The predictions were the average of two models, A and B:

$$\hat{y} = \frac{\hat{y}_A + \hat{y}_B}{2}$$

Model A consisted of a univariate GBLUP

$$\begin{gathered} y = Xb + Zu + e \\ u \sim N(0, K\sigma_u^2) \\ e \sim N(0, I\sigma_e^2) \end{gathered}$$

where the design matrix of fixed effects (X) was built with location meta-data: station, irrigation, and previous crop but no year effect or year-location combination; the relationship matrix (K) was calculated as an arccosine kernel. Predictions from model A were generated as

$$\hat{y}_A | \hat{u}_{22} = X\hat{b} + Z\hat{u}_{22}$$

Where $\hat{u}_{22}$ correspond to the predicted genomic values estimated for the 2022 data.

Model B consisted of location specific model using environmental variables (W) and the location meta-data (X) to predict the environmental means (μ), while relying on an index from a multivariate GBLUP to predict the genetic component. Thus, for the k^{th} location, the model was

$$y_k = \mu_k + u_k + e_k$$

The environment means model fitted the mean yield of environments (μ) as a function of environmental variables (W) and the location meta-data (X). The statistical model can be generally described as

$$\mu = f(X, W) + e$$

The function of environmental variables and meta data was to first get a biased composite predictor as the average of three sub-models

$$\mu_0 = \frac{RF(W) + RR(W) + LM(X)}{3}$$

where RF(W) is a random forest of the environmental factors, RR(W) is a ridge regression of environmental factors, and LM(X) is a least-squares of the location meta-data. Subsequently, the unbiased estimator is fit using a linear regression to remove the shrinkage from the composite prediction. Thus,

$$\mu = b_0 + \hat{\mu}_0 b_1 + e$$

where b_0 is the intercept and b_1 is the slope. To obtain the 2022 prediction, the sub-models of the estimated composited predictor were used to fit the 2022 data and subsequently fit into the linear model meant to mitigate the bias. Thus:

$$\hat{\mu}_{22} = \hat{b}_0 + \hat{\mu}_{0(22)} \hat{b}_1$$

Model B's estimated genomic value were inferred from a selection index. The predictions start from fitting an unstructured GxE model

$$\begin{gathered} y^* = g + e \\ g = \Sigma_g \otimes K \\ e = \Sigma_e \otimes I \end{gathered}$$

where y^* are normalized and spatially adjusted phenotypic values; g and e are the genetic effect and residuals of each corresponding phenotype, respectively. The genetic covariance matrix Σ_g contains the variances of each environment in the diagonal and covariances between pair of environments in the off diagonal. The residual covariance Σ_e is a diagonal matrix containing the residual variance for each environment. The output of the model consists of predictions of every individual in every environment.

The genetic merit for the k^{th} 2022 location (u_k) was estimated as a linear combination of the genomic values of the observed locations as:

$$u_k = \sigma_k \sum_{i=1}^{I} g_i w_{i,k}$$

where the scalar $w_{i,k}$ corresponds to the weight of the i^{th} location to predict the k^{th} location of 2022, and σ_k is the predicted standard deviation for the k^{th} location. The genotypic standard deviation of 2022 locations were predicted from the phenotypic standard deviation of location as a function of environmental covariates using random forest. Thus:

$$\sigma = RF(W) + e$$

The weights ($w_{i,k}$) were based on the deterministic accuracy of i^{th} predicting k^{th} location and the geographical location, such that locations in the same state and station would have higher weights than location further away, and locations with individuals more related to those in the k^{th} location will also have higher weights.

The final prediction of model B is:

$$y_k = \hat{\mu}_k + \hat{u}_k$$

Implementation

All computation were done in R. Linear mixed models were fit using the R package bWGR 2.1 (Xavier et al. 2020). The univariate model is described by Xavier (2019) and the multivariate model by Xavier and Habier (2022). The random forest used the R package ranger (Wright and Ziegler 2015).

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.

Xavier, A., & Habier, D. (2022). A new approach fits multivariate genomic prediction models efficiently. *Genetics Selection Evolution*, *54*(1), 1-15.

Xavier, A., Muir, W. M., & Rainey, K. M. (2020). bWGR: Bayesian whole-genome regression. Bioinformatics 36(6). btz764.

Xavier, A. (2019). Efficient estimation of marker effects in plant breeding. *G3: Genes, Genomes, Genetics*, *9*(11), 3855-3866.