PEDRO CÉSAR DE OLIVEIRA RIBEIRO

USING HISTORICAL PHENOTYPIC DATA AND ENVIRONMENTAL INFORMATION FOR GENOMIC PREDICTION IN BIOMASS SORGHUM

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

Orientador: Pedro Crescêncio Souza Carneiro
Coorientadores: Maria Marta Pastina
Rafael Augusto da Costa Parrella
Kaio Olímpio das Graças Dias

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
R484u
2022

Ribeiro, Pedro César de Oliveira, 1990-
Using historical phenotypic data and environmental information for genomic prediction in biomass sorghum / Pedro César de Oliveira Ribeiro. – Viçosa, MG, 2022.
1 tese eletrônica (78 f.): il. (algumas color.).

Texto em inglês.
Inclui anexos.
Inclui apêndice.
Orientador: Pedro Crescêncio Souza Carneiro.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Biologia, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.630
Modo de acesso: World Wide Web.

1. *Sorghum bicolor* - Melhoramento genético. 2. *Sorghum bicolor* - Seleção. 3. Genômica. I. Carneiro, Pedro Crescêncio Souza, 1966-. II. Universidade Federal de Viçosa. Departamento de Biologia. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 633.172

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

PEDRO CÉSAR DE OLIVEIRA RIBEIRO

USING HISTORICAL PHENOTYPIC DATA AND ENVIRONMENTAL INFORMATION FOR GENOMIC PREDICTION IN BIOMASS SORGHUM

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 18 de fevereiro de 2022.

Assentimento:

Pedro César de Oliveira Ribeiro
Autor

Pedro Crescêncio Souza Carneiro
Orientador

Aos meus pais e irmã.

AGRADECIMENTOS

A Deus.

Minha família, em especial aos meus pais Edmar e Ana, minha irmã Ediana, minha afilhada Luísa, meu sobrinho Heitor e minha Noiva Ruane.

Ao meu orientador, amigo e conselheiro Pedro Crescêncio Souza Carneiro.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Genética e Melhoramento.

À EMBRAPA milho e sorgo, em especial aos Coorientadores Rafael Parrella e Maria Marta pelos conselhos e apoio.

Ao Coorientador e grande Amigo Professor Kaio, pelas inúmeras conversas e conselhos.

A todos envolvidos de forma direta ou indireta com este trabalho, em especial a toda equipe de pesquisa de melhoramento de sorgo, o qual serei eternamente grato a equipe "GALPÃO do SORGO".

Ao professor Diego Jarquin, pela grande receptividade durante minha estadia em Nebraska, na cidade de Lincoln na Universidade of Nebraska -UNL. Thanks Doctor Jarquin, for all discussion about genomic selection, mixed models, and environmental information.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

ABSTRACT

RIBEIRO, Pedro Cesar de Oliveira, D.Sc., Universidade Federal de Viçosa, February, 2022. **Using historical phenotypic data and environmental information for genomic prediction in biomass sorghum.** Adviser: Pedro Crescêncio Souza Carneiro. Co-advisers: Maria Marta Pastina, Rafael Augusto da Costa Parrella and Kaio Olímpio das Graças Dias.

Energy sorghum is currently considered a promising alternative crop for generating bioenergy, due to its high biomass yield, has high concentrations of fermentable sugars in its stalks and desirable agro-industrial features, such as short growth cycle, high calorific value, and total crop mechanization. Therefore, the breeding programs are interested in developing new hybrids of sweet and biomass sorghum for a wide range of environmental conditions, however, the cost of the field-testing is expensive and time-consuming. The genomic selection (GS) is a powerful tool that allows breeders to predict the performance of new hybrids in yet to observe untested environments. GS models coupled with the use of environmental covariables (ECs) have the potential to enhance selection accuracy in breeding programs. Models of GS with ECs had potential to increase the accuracy of selection of breeding program. Thus, the goals of this study were to use historical data from VCU's energy sorghum tests from the Embrapa Maize and Sorghum breeding program, associated with molecular marker data and one set of environmental covariables for the prediction of untested energy sorghum hybrids and identification. of mega environments for the cultivation of energy sorghum hybrids.

Keywords: *Sorghum bicolor*. Breeding plants. Genomic selection.

RESUMO

RIBEIRO, Pedro Cesar de Oliveira, D.Sc., Universidade Federal de Viçosa, fevereiro de 2022. **Uso de dados fenotípicos históricos e covariáveis climáticas para a predição genômica de híbridos de sorgo energia em múltiplos ambientes**. Orientador: Pedro Crescêncio Souza Carneiro. Coorientadores: Maria Marta Pastina, Rafael Augusto da Costa Parrella e Kaio Olímpio das Graças Dias.

Atualmente, o sorgo é considerado uma alternativa promissora para a geração de bioenergia, devido ao seu alto rendimento de biomassa, caldo com açúcares que são altamente fermentáveis e às características agroindustriais desejáveis, como ciclo curto, alto poder calorífico e colheita mecanizada. Sendo assim, os programas de melhoramento estão interessados em desenvolver novos híbridos de sorgo sacarino e biomassa para uma ampla gama de condições ambientais; no entanto, o custo do teste de campo é caro e demorado. A Seleção Genômica (GS) é uma ferramenta poderosa que permite aos melhoristas prever o desempenho de novos híbridos em ambientes ainda não testados. Modelos de GS aliados ao uso de covariáveis ambientais (ECs) têm o potencial de aumentar a precisão da seleção em programas de melhoramento. Sendo assim os conjuntos de dados históricos de melhoramento, compostos por vários anos e locais, informação molecular e covariáveis ambientais, pode implementar a predição de híbridos de sorgo energia. Permitindo predições mais acuradas de híbridos ainda não testados em ambientes já avaliados, ou ainda a predição de genótipos não avaliados em ambientes também não avaliados pelo programa de melhoramento. Assim, o objetivo deste trabalho é a utilização de dados históricos dos ensaios VCU's de sorgo energia do programa de melhoramento da Embrapa Milho e Sorgo, associados aos dados de marcadores um conjunto de covariáveis climáticas para a predição de híbridos não realizados de sorgo energia e identificação de mega ambientes para a cultura do sorgo energia.

Palavras-chave: *Sorghum bicolor*. Melhoramento genético. Seleção genômica.

SUMÁRIO

GENERAL INTRODUCTION

Biomass sorghum (Sorghum bicolor (L.) Moench) is a specie cultivated around the world, which presents certain intrinsic advantages for bioenergy purposes (Appiah-Nkansah et al., 2019). The sorghum [Sorghum bicolor (L.) Moench] has multiple uses in the generation of bioenergy, such as first and second-generation ethanol, cogeneration of energy through the combustion of dry biomass, and production of biogas, and it has the potential for the supply chain of the bioenergy sector (Rezende and Richardson, 2017; Gomes et al., 2019). In Brazil, the bioenergy sorghum crop has a short production cycle with no need for adjustments in the bioethanol industrial process.

The Brazilian Agricultural Research Corporation (Embrapa Maize and Sorghum) has developed sorghum hybrids according to the three lines scheme, A B, and R-lines (Ribeiro et al., 2020). These hybrids developed to Embrapa have been tested in a multi-locations and years. This historical phenotypic data is essential to validate the performance and provide recommendations for developing hybrids (Smith et al., 2005; Malosetti et al., 2013).

To provide recommendations for releasing new hybrids (Smith et al., 2001, 2015) biomass sorghum breeding programs are interested in developing superior hybrids for various environmental conditions. But phenotyping hybrids in multiple environments might be expensive. Therefore, genomic selection became a powerful tool for predicting the performance of new combinations in untested environments using environmental covariates (ECs) (Jarquin et al., 2014; Crossa et al., 2017, Crossa et al., 2021). Therefore, genomic prediction holds the potential to accelerate genetic gains in breeding programs for diverse crops (Lopez-Cruz et al., 2015; Bernal-Vasquez et al., 2017; Dias et al., 2018; Islam et al., 2020). Nonetheless, in biomass sorghum, a few studies have been conducted (Yu et al., 2016; de Oliveira et al., 2018; Fernandes et al., 2018; Rice and Lipka, 2019).

To release a stable and high-performance cultivar of sweet sorghum for tropical environments, it is crucial to understand genotype-by-environment interactions (GEI) (Souza, et al., 2021). The crucial step required to release a cultivar with stable and high-performance is the development field trials in multi-environment trials (MET) to validate the performance and recommendation of cultivars (Smith et al., 2001, 2005, 2015). One example of MET is the value for cultivation and use (VCU). The plant breeding program develops cultivars with high performance under different environmental conditions. However, in Brazil, the breeder's biomass sorghum faces great difficulty due to its extensive territorial area and the diversity of climatic conditions.

According to Jarquin et al. (2014), most of the effects of genes could be modeled by regressing phenotypes on genetic markers and on environmental covariates like; temperature, soil moisture, and solar radiation. Plus, the G × E can be modeled using interactions between genetic markers and environmental covariates. Besides, the modern genotyping methods, molecular markers information, and technologies have become high dimensional, also, as climatic and agronomic information systems, and environmental information (Bernardo 2010; Crossa 2012).

Meuwissen et al. (2001) proposed whole genome regression (WGR) methods to solve the limitations of QTL-based models. Using these models, it is possible to capture major-effect genes and the contribution of genomic regions with small effects. Moreover, the gain selection can perform predictive power based on pedigree methods, extending to multi-trait multi-environment settings using these modern estimation procedures, and plant breeding data (Crossa et al. 2010, Heslot et al. 2012).

The G-BLUP can interpret as a reaction norm model (Su et al., 2006) Where genetic and environmental gradients are described using linear regression on genetic markers on ECs.

Nonetheless, the prediction accuracy was verified for two prediction problems. The first problem was called CV1, in which models are used to predict the performance of lines that have never been evaluated in field trials, and a second problem CV2, in which all lines have at least one field evaluation available, and the prediction problem was that of predicting performance across environments, in an incomplete trial (Jarquin et al., 2014; Crossa et al., 2021).

Jarquin et al. (2014) concluded that models that accounted for the interaction of effects of markers, environments, and environments covariates could increase the predictive correlation by almost 35% on average from CV1, and 17% from CV2. That means gains in prediction accuracy by combining those tools.

Thus, the goals of this study were to use historical data from VCU's energy sorghum tests from the Embrapa Maize and Sorghum breeding program, associated with 9molecular marker data and one set of environmental covariables for the prediction of untested energy sorghum hybrids and identification. of mega environments for the cultivation of energy sorghum hybrids.

REFERENCES

Appiah-Nkansah, N.B., Li, J., Rooney, W., & Wang, D. 2019. A review of sweet sorghum as a viable renewable bioenergy crop and its techno-economic analysis. Renewable Energy, 143, 1121-1132.

Bernal-Vasquez, A. M., Gordillo, A., Schmidt, M., & Piepho, H. P. 2017. Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC genetics, 18(1), 1-17.

Bernardo R.N. 2010. Breeding for quantitative traits in plants, 2nd edn. Stemma Press, Woodbury.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., & Varshney, R. K. 2017. Genomic selection in plant breeding: methods, models, and perspectives. Trends in plant science, 22(11), 961-975.

Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O.A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A., & Bentley, A.R. 2021. The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. Frontiers in plant science, 12, 651480.

Crossa J. 2012. From genotype × environment interaction to gene × environment interaction. Curr Genomics 13:225–244. doi:10.2174/138920212800543066.

Dias, K. O. D. G., Gezan, S. A., Guimarães, C. T., Nazarian, A., da Costa e Silva, L., Parentoni, S. N., ... & Pastina, M. M. 2018. Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity, 121(1), 24-37.

Fernandes, S.B., Dias, K.O., Ferreira, D.F., & Brown, P.J. 2018. Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. Theoretical and applied genetics, 131(3), 747-755.

Gomes, L., F. De Almeida, R. Augusto, M. Lúcia, F. Simeone, et al. 2019. Biomass and Bioenergy Composition and growth of sorghum biomass genotypes for ethanol production. 122(June 2018): 343–348. doi: 10.1016/j.biombioe.2019.01.030.

Heslot N, Yang H-P, Sorrells ME, Jannink J-L. 2012. Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146. doi:10.2135/cropsci2011.06.0297.

Islam, M. S., Fang, D. D., Jenkins, J. N., Guo, J., McCarty, J. C., & Jones, D. C. 2020. Evaluation of genomic selection methods for predicting fiber quality traits in Upland cotton. Molecular Genetics and Genomics, 295(1), 67-79.

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., ... & de los Campos, G. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and applied genetics, 127(3), 595-607.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J. L., ... & de los Campos, G. 2015. Increased prediction accuracy in wheat breeding trials using a marker×environment interaction genomic selection model. G3: Genes, Genomes, Genetics, 5(4), 569-582.

Malosetti, M., Ribaut, J.M., & van Eeuwijk, F.A. 2013. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. Frontiers in physiology, 4, 44.

Meuwissen. T.H.E., Hayes. B.J., Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Oliveira Ribeiro, P.C., de Souza Marçal, T., Oliveira, I.C.M., Schaffert, R.E., Carneiro, P.C.S., de Oliveira, A.B., & da Costa Parrella, R.A. 2020. Insight into genetic potential of male sterile sweet sorghum A-lines for agroindustrial traits using tester R-lines. Industrial crops and products, 153, 112577.

Oliveira, A. A., Pastina, M. M., de Souza, V. F., da Costa Parrella, R. A., Noda, R. W., Simeone, M. L. F., ... & Margarido, G. R. A. 2018. Genomic prediction applied to high-biomass sorghum for bioenergy production. Molecular Breeding, 38(4), 1-16.

Rezende, M.L., & Richardson, J.W. 2017. Risk analysis of using sweet sorghum for ethanol production in southeastern Brazil. Biomass and bioenergy, 97, 100-107.

Rice, B., & Lipka, A.E. 2019. Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. The Plant Genome, 12(1), 180052.

Smith, A.B., Cullis, B.R., & Thompson, R. 2005. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. The Journal of Agricultural Science, 143(6), 449-462.

Smith, A., B. Cullis, and A. Gilmour. 2001. The analysis of crop variety evaluation data in Australia. Aust. New Zeal. J. Stat. 43(2): 129–145.

Smith, A.B., A. Ganesalingam, H. Kuchel, and B.R. Cullis. 2015. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. Theor. Appl. Genet. 128(1): 55–72. doi: 10.1007/s00122-014-2412-x.

Souza, V.F.D., Ribeiro, P.C.D.O., Vieira Junior, I.C., Oliveira, I.C.M., Damasceno, C.M.B., Schaffert, R.E., & Pastina, M.M. 2021. Exploring genotype× environment interaction in sweet sorghum under tropical environments. Agronomy Journal, 113(4), 3005-3018.

Su G, Madsen P, Lund MS, Sorensen D, Korsgaard IR Jensen J. 2006. Bayesian analysis of the linear reaction norm model with unknown covariates. J Anim Sci 84 (7)(July):1651–1657. doi:10. 2527/jas.2005-517

Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S. E., & Yu, J. 2016. Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nature Plants, 2(10), 1-7.

CAPITULO 1

PREDICTING THE VALUE FOR CULTIVATION AND USE OF SWEET SORGHUM CULTIVARS TRIALS WITH ZONE EFFECTS VIA CLIMATE DATA

ABSTRACT

The second-largest producer of bioethanol in the world is Brazil that needs a growing demand for the development of alternative fuels from renewable sources. Thus, sweet sorghum has a potential for bioethanol production. The crucial step required to release a cultivar with stable and high-performance is the development field trials in Multi-environment trials (MET) to validate the performance and recommendation of cultivars. One example of MET is the value for cultivation and use (VCU). The plant breeding program develops cultivars with high performance under different environmental conditions. However, in Brazil, the breeder's sweet sorghum faces great difficulty due to its extensive territorial area and the diversity of climatic conditions. The identification of target populations of environments or zones can help the breeding programs to recommend the cultivars. The environmental covariables can enable the definition zones and use this information in sweet sorghum breeding programs. Another discussion in analysis of VCU to sweet sorghum about the effects of cultivar, and locations, which should be regarded as fixed or random, and the impact in the rank of the cultivars. Thus, the goals of this study were i) to investigate the impact of the cluster zone effects via climate data in the prediction of VCU trials; ii) to compare the effect of modeling the different linear mixed models that use either BLUE or BLUP under unweighted and weighted two-stage analyses. We used two datasets for sweet sorghum (line data and hybrid) to meet our goals and evaluated tons of Brix per hectare. To define the breeding zones for sorghum locations under tropical conditions, using environmental covariables. We compared 17 linear mixed models with different effects fixed or random. We used a cross-validation scheme to compare and select the best model and assess the accuracy of estimates of differences between cultivars in various locations. We used a mean squared error of prediction (MSEP) proposed by Piepho (1998). Results show models all fixed in all most case outperformance, other models. The inclusion of zone effects in the model reduced the MSEP for most cases. Cultivar rankings change in both data sets. Our results suggest that models with (co)variance structure outperformed the model with fixed effects, currently adopted model in the sweet sorghum program. Besides, target populations of environments, stratified through environmental covariates increased the model precision based on cross-validation scenarios. Based on two sweet sorghum datasets, our result showed that delimitation of the target population of environments via environmental covariable could increase cultivar effects prediction under tropical conditions. We performed thorough provided insights into the analysis of VCU in sweet sorghum to assess the performance of

random or fixed-cultivar-effects models, which also shows the advantage of BLUP in sweet sorghum breeding program routine.

Keywords: *Sorghum bicolor.* Bioethanol. Target population environment. BLUE and BLUP.

INTRODUCTION

Among the countless challenges of developing innovation in bioethanol production processes are the release of raw materials (i.e. cultivars) to supply the sugar-energy sector demand and needs (Moncada et al., 2018). Brazil is the second-largest producer of bioethanol in the world (Renewable Fuels Association, 2019) mainly using sugarcane crops. Through the bioenergy sorghum improvement program of Embrapa Maize and Sorghum and several partnerships with research institutes and universities have developed works with the culture of sweet sorghum (da Silva et al., 2017; Ribeiro et al., 2020; Leite et al., 2020) that stands out for presenting desirable traits in a bioethanol production process (Appiah-Nkansah et al., 2019). A study suggests that sweet sorghum can reduce the production cost in the sector chain, nevertheless improvements (Rezende and Richardson, 2017).

In the mind of the numerous steps involved in a breeding program, field trials in multi-locations to validate the performance and recommendation of the developed experimental cultivars are essential for breeding programs to recommend these cultivars (Smith et al., 2005; Malosetti et al., 2013). Multi-environment trials (MET) aim to evaluate cultivar performance across locations and regions, with the final goals of recommending a new cultivar (Smith et al., 2001, 2005, 2015a). One example of MET is the value or cultivation and use (VCU) trials that are annually performed by the company's breeding for legal registration and commercial recommendation of newly developed cultivars. VCU trials increase confidence in select and recommend cultivars for specific and broad adaptation. In Brazil, the Embrapa energy sorghum-breeding program for legal registration and commercial recommendation of newly developed cultivars carries out annually VCU trials for sweet sorghum.

The linear mixed models have become popular in plant breeding, and in the analysis of MET, and shows among the advantages the ability to assume genotype and location to be as fixed or random in yield performance trials (Smith et al., 2001, 2005, 2015a; Piepho et al.,

2008). The issue of whether the main effects, i.e., cultivar and locations, should be regarded as fixed or random is extremely important in plant breeding. The choice in determining whether the cultivar should have fixed or random has been a contrast of thoughts (Smith et al., 2005). The estimators eBLUP's (*Best Linear Unbiased Prediction*), in general, have less variance than those of the eBLUE's (*Best Linear Unbiased Estimation*) is resulting in more reliable estimates that can determine differences in the classification of genotypic and minimizes selection errors, mainly in unbalanced data (Henderson, 1975; Duarte and Vencovsky, 2001; Piepho et al., 2008; Gezan et al., 2017). On the other hand, when genotypes are assumed as fixed, with unbalanced data, the rank of cultivars can be harmed, and the use of random effects predictions is more realistic (Smith et al., 2005).

The main goal of a plant breeding program is to develop cultivars with high performance under different environmental conditions (Yan et al., 2000; Gauch et al., 2008; Malosetti et al., 2013). Recently, the selection of sweet sorghum cultivars bases on the individual performance of BLUEs (de Souza et al., 2013; de Figueiredo et al., 2015; Eculica et al., 2019). The breeders of sweet sorghum face great difficulty in achieving this goal in Brazil due to its extensive territorial area and the diversity of climatic conditions found in tropical conditions. The environmental conditions can be divided into: predictable (ex: soil, geographic position) and unpredictable (i.e. fluctuations in weather) (Allard and Bradshaw, 1964). In this way, the identification of mega environments, that is, a subdivision of the territory into target populations of environments (TPE) or zones, is a key step in breeding programs and can help in recommending cultivars (Yan et al., 2000; Gauch et al., 2008; Malosetti et al., 2013; Van Eeuwijk et al., 2016; González-Barrios et al., 2019). Zones can be a group of environments that produce a similar ranking of the cultivars, which has been proposed as an efficient alternative to minimize the presence of Genotype by Environment (GEI) (Piepho and Möhring, 2005; González Barrios et al., 2019; Dias et al., 2020). And, once that the environmental covariates

have a biological meaning able to understand the environment, can be used to drive GEI in predictability information to the TPE (Voltas et al., 1999; Ortiz et al., 2007). In this context, the use of environmental covariables can enable the definition of these zones and help the improvement programs of sweet sorghum.

In the sweet sorghum program, the cultivar's recommendation delineated by the average performance across VCU trials conducted in Brazil. However, in most cases, the recommendation does not consider the zone information and the advantage of a mixed model approach for recommendation purposes. Therefore, the goals of our study were: i) to investigate the impact of the cluster zone effects via climate data in the prediction of VCU trials; ii) to compare the effect of modeling the different linear mixed models that use either BLUE or BLUP under unweighted and weighted two-stage analyses.

MATERIAL AND METHODS

Plant material and experimental design

We used two datasets for sweet sorghum to meet our goals. The first dataset (hereafter called line data) comprises phenotypic data of 41 sweet sorghum genotypes, 32 elite inbred lines, and nine hybrids evaluated in 38 late-stage breeding trials conducted across 19 locations over 2009 to 2014 (Figure 1). The second dataset, (hereafter called hybrid data), comprises 89 sweet sorghum genotypes, being 83 hybrids, and six elite inbred lines, evaluated in 33 late-stage breeding trials across 13 locations over 2014 to 2018 (Figure 1). Hereafter all plant material (i.e line and hybrid) called genetic material. As expected in a breeding routine, genotypes are unbalanced over the years due to eliminating the worst ones and inclusion of new ones each year (Figure 2). Also, in the two datasets, locations are not constant over the years (Tables S1-S3).

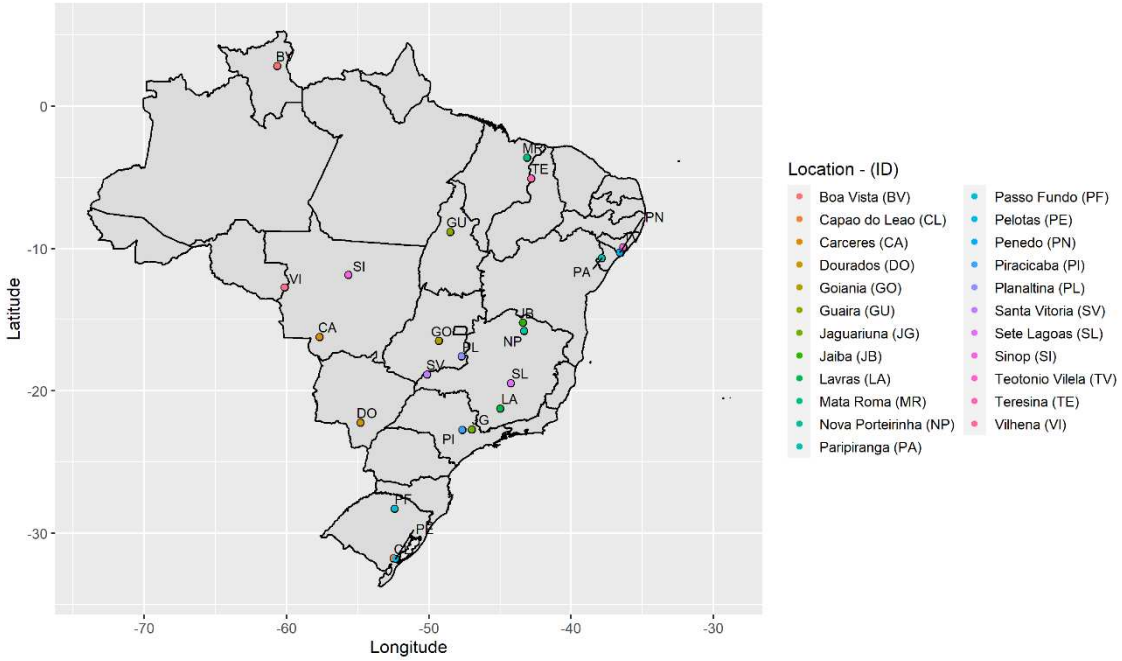

Figure 1. The 23 tropical locations used from 2009 to 2018 for the valor of cultivation and use (VCU) trial for line and hybrid datasets. Locations points: green: line and hybrid data, red: only line data, black: only hybrid data.

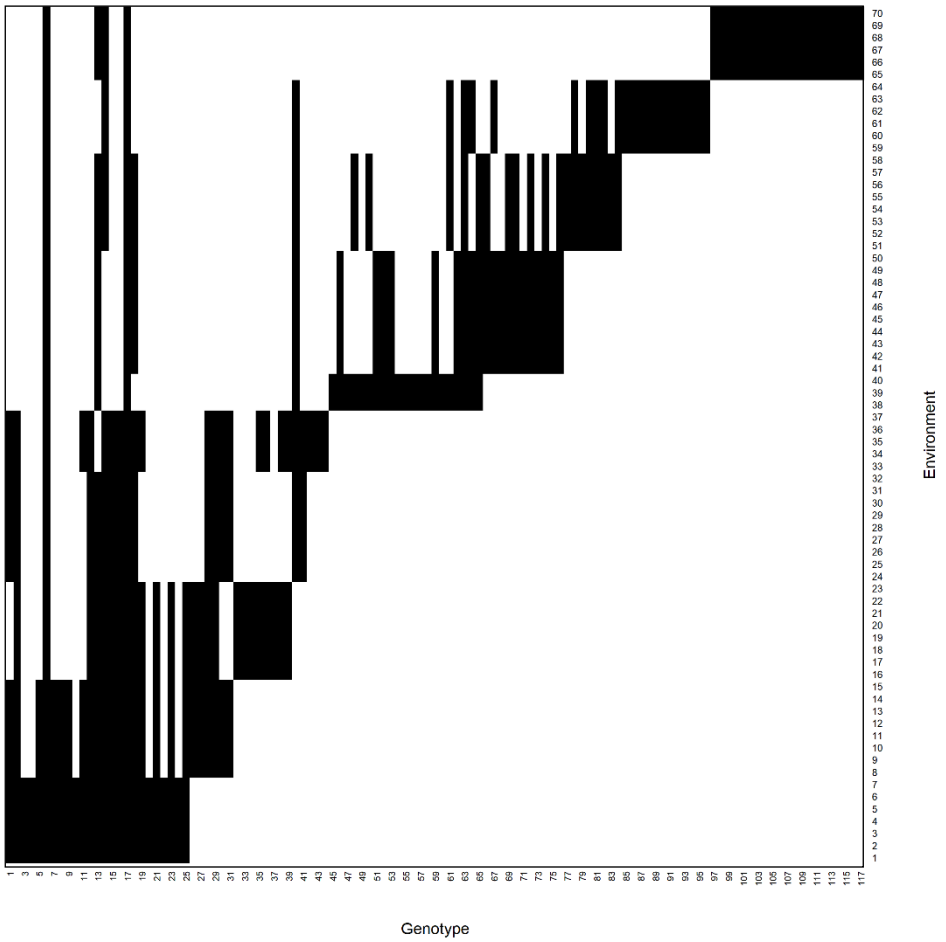

Figure 2. Unbalance data over the years. Environments codify and describe tables S1 and S2.

In the field, each trial was arranged as an alpha lattice design with three replications. In all years, 25 sweet sorghum genotypes were assessed per trial considering five genotypes per block in the lattice design. For the line data in 2014, 16 genotypes were tested considering a block size of four. The experimental plot consisted of two five-meter rows spaced by 0.7 m, with a population density of approximately 120,000 plants ha^{-1}.

We evaluated tons of Brix per hectare (TBH) by the index: $TBH = \frac{FBY \times TSS}{100}$, where FBY is fresh biomass yield, assessed as the weight of whole plants in each plot, and converted to tons per hectare. The TSS is the total soluble solids measured in the stalk juice and determined in ºBrix by automatic digital refractometer.

Defining breeding zones

To define the breeding zones for sweet sorghum locations under tropical conditions, using climate data. The first step the determined the beginning of the harvest season according to the Ministry of Agriculture (zonal climate recommended - MINISTÉRIO DA AGRICULTURA 2019), for each tested location in the VCUs trials from 2009 to 2018. Thereby, starting the beginning of the crop in each location, 150 days were counted, referring to the biomass sorghum cycle, and used as a reference to collect the environmental covariables data.

Were collected climate data that tends to affect sweet sorghum performance, such as maximum, mean, and minimum temperature (°C), solar radiation (in MJ m^{-2} day^{-1}), precipitation (mm day^{-1}), relative humidity (%), wind speed (m s^{-1}), altitude (m), latitude, and longitude. The environmental covariates we obtained from the database of the National Aeronautics and Space Administration Prediction of Worldwide Energy Resource (NASA POWER) project (https://power.larc.nasa.gov/data-access-viewer/). To cluster the locations into target populations of environments (TPE), we used the k-means algorithm considering the

average value of each environmental covariable for each site. Also, we tested different numbers of TPE and selected the optimum number based on "wss" (Total Within Sum of Square) of the k-means (Hartigan and Wong, 1979).

Statistical analyses

Single-environment trial analyses

We fitted a linear mixed models using the statistical package ASReml-R v.4 (Butler et al., 2018) into the R software v. 3.6 (2020), the first stage model was as follow:

$$y = R + C:B.R \quad [1]$$

where R is the replicate effect, B is the block within a replicate effect, and C is the cultivar effect. The fixed effects are specified before the colon and the random effects after the colon. The dot between two factors indicates either nested of an interaction effect, according to the main effect's inclusion. The intercept, and the residual error term are implicit.

We estimated the generalized heritability (H^2) (Cullis et al., 2006) for each location using the following equations: $H^2 = 1 - \left[vdBLUP/\left(2 \times \sigma_g^2\right)\right]$, where, $\sigma_g^2$is the genetic variance and vdBLUP is the average pairwise prediction error variance of genotype effects.

Multi- environment trial analyses

In the second stage, we fitted models across locations for each year, using the adjusted means of cultivars from the first stage. Below we described a compound symmetry model considering the zone effect:

$$\overline{Y} = Z:L + C + Z:C + L:C + Z:L$$

where $\overline{Y}$ is the adjusted mean for a cultivar from stage I, C is the cultivar effect, Z is the zone effect, and L is a location effect. The fixed effects are specified before the colon and the random effects after the colon. The dot between two factors indicates a nested effect. The intercept and the residual error term are implicit.

Also, we compared 17 linear mixed models categorized into four groups (Table 1). The first group is the only one that considers genetic material as a fixed effect (MF) and is a current model used in the Sweet Sorghum program from Embrapa. The second group comprises models with no zone effects (MNZ), while the third group considers the zone effects (MZ) defined used environmental covariate. The last group is one stage (MOS).

Table 1. The 17 statistical models used in the single-year series cross-validation.

Model Name†	Fixed terms‡	Random terms‡	VCOV*
MF 1	G + L#	-	$\mathbf{R} = \sigma^2\mathbf{I}_{nxn}$
MNZ 2	L	G:L	$\mathbf{G}_{GL} = \sigma^2_{gl}\mathbf{I}_{nxn};\ \mathbf{R} = \mathbf{\Sigma}_{nxn}$
MNZ 3	L	G:L	$\mathbf{G}_{GL} = \sigma^2_{gl(n)}\mathbf{I}_{nxn};\mathbf{R} = \mathbf{\Sigma}_{nxn}$
MNZ 4	L	G:L	$\mathbf{G}_{GL} = (\sigma^2_g\mathbf{J}_{nxn} + \sigma^2_{gl}\mathbf{I}_{nxn});\mathbf{R} = \mathbf{\Sigma}_{nxn}$
MNZ 5	L	G:L	$\mathbf{G}_{GL} = [\Lambda\Lambda^T + \psi]\otimes\mathbf{I}_{nxn};\mathbf{R} = \mathbf{\Sigma}_{nxn}$
MNZ 6	L	G	$\mathbf{G}_G = \boldsymbol{\sigma^2_{gl}}\mathbf{I}_{nxn};\ \mathbf{R} = \boldsymbol{\sigma^2}\mathbf{I}_{nxn}$
MNZ 7	-	L + G:L	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\ \mathbf{G}_{GL} = \sigma^2_{gl}\mathbf{I}_{nxn};\ \mathbf{R} = \mathbf{\Sigma}_{nxn}$
MNZ 8	-	L + G:L	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\mathbf{G}_{GL} = \sigma^2_{gl(n)}\mathbf{I}_{nxn};\ \mathbf{R} = \mathbf{\Sigma}_{nxn}$
MNZ 9	-	L + G:L	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\mathbf{G}_{GL} = (\sigma^2_g\mathbf{J}_{nxn} + \sigma^2_{gl}\mathbf{I}_{nxn});\mathbf{R} = \mathbf{\Sigma}_{nxn}$
MNZ 10	-	L + G:L	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\ \mathbf{G}_{GL} = [\Lambda\Lambda^T + \psi]\otimes\mathbf{I}_{nxn};\mathbf{R} = \mathbf{\Sigma}_{nxn}$
MNZ 11	-	G+ L	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\ \mathbf{G}_G = \sigma^2_g\mathbf{I}_{nxn};\ \mathbf{R} = \sigma^2\mathbf{I}_{nxn}$
MZ 12	Z	L + C:L + C:Z	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\mathbf{G}_{GL} = \sigma^2_{gl}\mathbf{I}_{nxn};\mathbf{G}_{GZ} = \sigma^2_{gz}\mathbf{I}_{nxn};\ \mathbf{R} = \mathbf{\Sigma}_{nxn}$
MZ 13	Z	L + C:L + C:Z	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\ \mathbf{G}_{GL} = \sigma^2_{gl}\mathbf{I}_{nxn};\mathbf{G}_{GZ} = \sigma^2_{gz(l)}\mathbf{I}_{nxn};\mathbf{R} = \mathbf{\Sigma}_{nxn}$
MZ 14	Z	L + C:L + CZ	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\mathbf{G}_{GL} = \sigma^2_{gl}\mathbf{I}_{nxn};\mathbf{G}_{GZ} = (\sigma^2_g\mathbf{J}_{nxn} + \sigma^2_{gz(l)}\mathbf{I}_{nxn});\mathbf{R} = \mathbf{\Sigma}_{nxn}$
MZ 15	Z	L + C:L + C:Z	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\mathbf{G}_{GL} = \sigma^2_{gl}\mathbf{I}_{nxn};\ \mathbf{G}_{GZ} = [\Lambda\Lambda^T + \psi]\otimes\mathbf{I}_{nxn};\mathbf{R} = \mathbf{\Sigma}_{nxn}$
MZ 16	Z	C + L	$\mathbf{G}_L = \sigma^2_l\mathbf{I}_{nxn};\ \mathbf{G}_c = \sigma^2_g\mathbf{I}_{nxn};\ \mathbf{R} = \sigma^2\mathbf{I}_{nxn}$
MOS 17	L + R:L	L:R:B + C + C:L	$\mathbf{G}_B = \sigma^2_B\mathbf{I}_{nxn};\mathbf{G}_c = \sigma^2_C\mathbf{I}_{nxn};\mathbf{G}_{GL} = \sigma^2_{CL}\mathbf{I}_{nxn};\ \mathbf{R} = \mathbf{\Sigma}_{nxn}$

† MF, model fixed; MNZ, model no-zone, MZ, model zone, MOS, model one-stage,‡ G, genotype; L, Location; Z, zone; R, repetition; B, block; # The current-practice model; * $\mathbf{I}_{nxn}$: is identity matrices with their corresponding orders *n* locations, $\mathbf{\Sigma}$ is a (*nxn*) diagonal matrix, in which the diagonal elements are given by the inverse of the variance and covariance matrix of the adjusted means of cultivars in each location, $\mathbf{J}_{nxn}$ is a matrix with 1's their corresponding orders *nxn* locations

Obtained the variance components based on a compound symmetry model as outlined above, the significance of these components was access by LRTs (Likelihood Ratio Tests) considering α = 0.05, and by the chi-square statistic (χ^2) test with *v* degrees of freedom, where *v* is the difference between the number of parameters of the compared models (Gilmour et al., 2015) and were performed for single and multi-environment trial analyses for TBH. Models with the same group (i.e., same fixed effects) were compared based on their AIC - Akaike Information Criterion (Akaike, 1974). It is also important to highlight that were only consider single-year data in all analyses. The main reason is due to the low overlap of genotypes over the years.

Cross-Validation

We used a cross-validation (CV) scheme to compare and select the best model (Table 1) in both datasets. For this, a leave-one-out strategy was used, being one location out at a time for the validation set and the remaining locations as a training set. For example, in 2015, for the hybrid data were nine locations, eight folds were used as a training set and one as a validation set. Each location was used once as a validation set.

To assess the accuracy of estimates of differences between cultivars in various locations, we used a mean squared error of prediction (MSEP) proposed by Piepho (1998).

$$MSEP = \frac{\sum_{k=1}^{J}\sum_{l=1}^{I}\sum_{k\neq l}^{I}[Y_{kl} - Y_{k'l} - (Z_{kl} - Z_{k'l})]^2}{JI(I-1)}$$

where: $Y_{kl} - Y_{k'l}$ is a difference between observed values, $(Z_{kl} - Z_{k'l})$ is a difference between predictive values, and JI are the number of locations and genotypes, respectively. The CV study was conducted by R software v.4.0 (R Core team, 2020).

According to the formula above, the MSEP is calculated based on the measurement difference between cultivars in different environments and assesses the accuracy of estimates

this difference. It is worth mentioning that the differences between cultivars in various environments are the main goal in trials annually performed in breeding programs (Piepho, 1998).

RESULTS

The phenotypic means for TBH across trials ranged from 3.83 t ha^{-1} in CL 2012 to 12.32 t ha^{-1} in GU 2010, with an overall mean of 7.83 t ha^{-1} in the line data (Figure 2). Hybrid data values of TBH ranged from 4.09 t ha^{-1} in PE 2015 to 18.57 t ha^{-1} in NP 2014, with an overall mean of 9.85 t ha^{-1} (Figure 2). Based on the Likelihood Ratio Test (LRT, considering $\alpha = 0.05$), the cultivar component of variance differs significantly from zero for some environments (Tables S4 and S5). The heterogeneity of residual variances was observed in the two datasets, as verified by genetic and residual s from the single environment, each location in each year (Table S4 and S5). Generalized heritability values ranging from 0.27 to 0.97 in both datasets, with mean was 0.70 and 0.75 for line and hybrid data, respectively, which represents a high accuracy of experimental trials.

In analyses multi-environment the component represents environments (σ_L^2), and cultivar (σ_C^2) shows significantly in both data sets in all years, the same result we found to the interaction Cultivar-By-Environment Effects (σ_{CL}^2). The same did not see to the component interaction Cultivar-By-Zone (σ_{CZ}^2), where σ_{CZ}^2 showed significance only in three years in both datasets (Table 2). Highlights that we have grouped environments in the zone (predictability) according to historical climate data during the harvest season sweet sorghum in tropical conditions via k-means and selected the optimum number based on "wss" (Total Within Sum of Square) equal 3 (Figure 3).

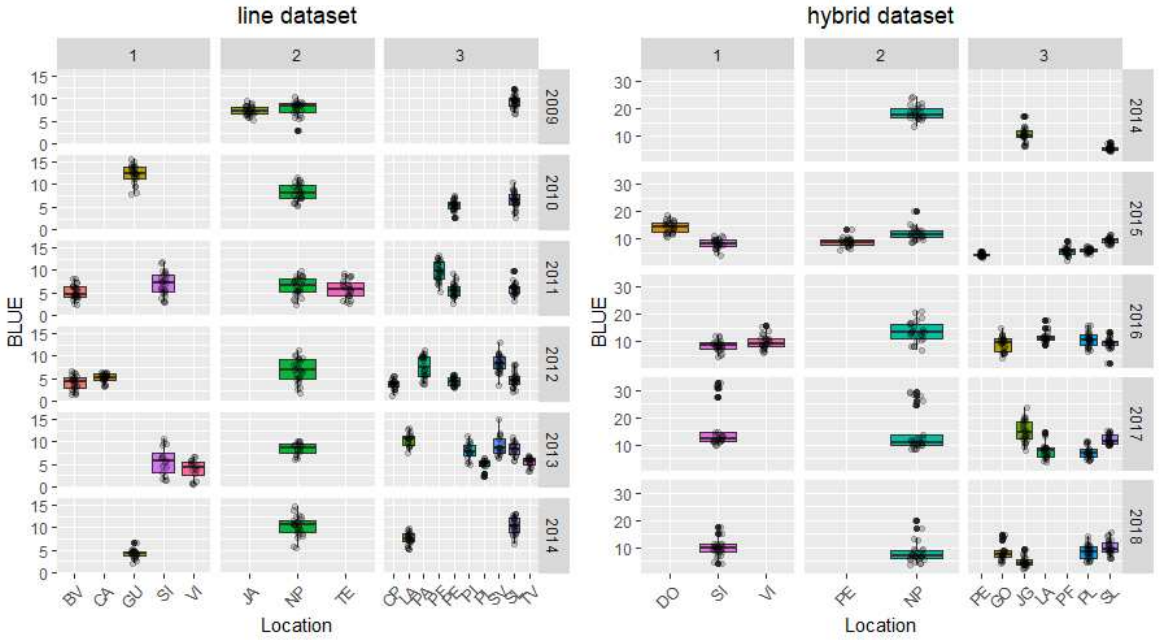

Figure 2. Boxplot of adjusted means of each environment used for the Embrapa's sweet sorghum valor cultivation and use (VCU) trials, BV: Boa Vista, CA: Cárceres, CP: Capão do Leão, DO: Dourados, GO: Goiânia, GU: Guaíra, JG: Jaguariúna, JA: Jaíba, LA: Lavras, MR: Mata Roma, NP: Nova Porteirinha, PA: Paripiranga, PF: Passo Fundo, PE: Pelotas, PN: Penedo, PI: Piracicaba, PL: Planaltina, SV: Santa Vitória, SL: Sete Lagoas, SI: Sinop, TV: Teotônio Vilela, TE: Teresina, VI: Vilela.

Table 2. Likelihood Ratio Test (LRT) estimate based to for the line and hybrid data, L represents environments (σ_L^2), C represents cultivar (σ_C^2), CZ Cultivar-By-Zone (σ_{CZ}^2), and CL Cultivar-By-Environment Effects (σ_{CL}^2). Estimates of the variance of the components to Environments (σ_L^2), Cultivar (σ_C^2), Cultivar-By-Zone (σ_{CZ}^2), and Cultivar-By-Environment Effects (σ_{CL}^2).

Data Set	Year	σ_L^2	σ_C^2	σ_{CZ}^2	σ_{CL}^2
Line	2009	0.29*	187.88*	0.00ns	92.99*
	2010	4.61*	178.59*	0.00ns	15.42*
	2011	61.56*	405.56*	0.05ns	83.16*
	2012	139.47*	534.61*	0.26ns	40.20*
	2013	88.34*	1550.90*	5.41*	88.54*
	2014	119.28*	167.52*	0.94ns	2.35*
Hybrid	2014	46.62*	88.94*	0.00ns	48.33*
	2015	319.85*	236.64*	8.33*	30.77*
	2016	52.86*	683.19*	0.16ns	171.19*
	2017	70.45*	22266.00**	27.56*	634.74*
	2018	30.36*	1579.50*	0.00ns	591.81**

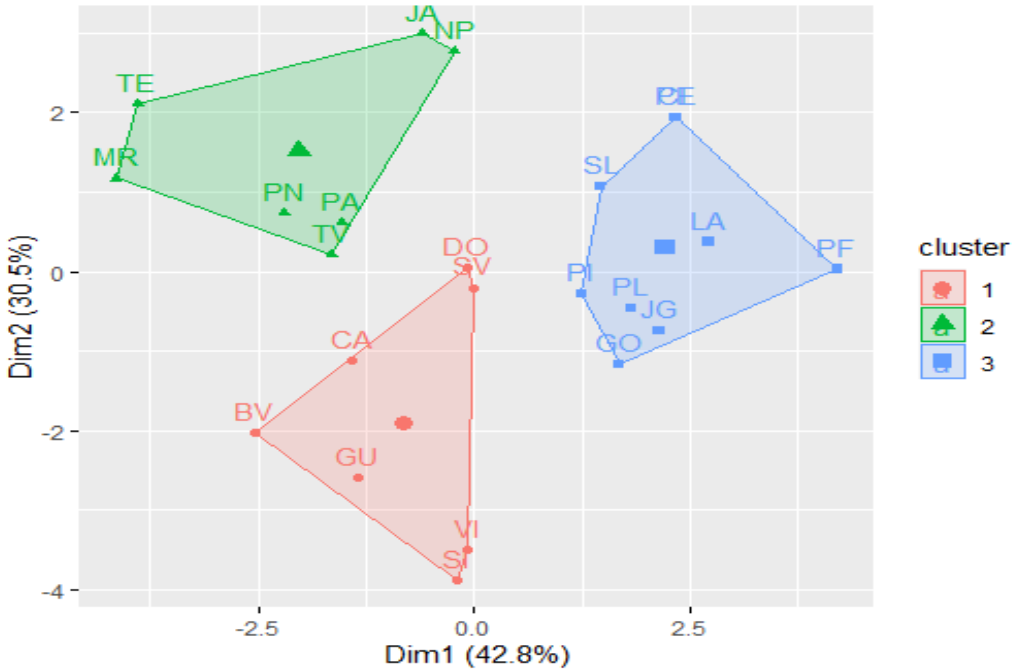

Figure 3. Plot-cluster of Zone of the 22 locations used for the Embrapa's sweet sorghum valor cultivation and use (VCU) trials defined by k-means. BV: Boa Vista, CA: Cárceres, DO: Dourados, GO: Goiânia, GU: Guaíra, JG: Jaguariúna, JA: Jaíba, LA: Lavras, MR: Mata Roma, NP: Nova Porteirinha, PA: Paripiranga, PF: Passo Fundo, PE: Pelotas, PN: Penedo, PI: Piracicaba, PL: Planaltina, SV: Santa Vitória, SL: Sete Lagoas, SI: Sinop, TV: Teotônio Vilela, TE: Teresina, VI: Vilhela.

In cross-validation, the currently used fixed model for VCU recommendation had the highest MSEP in most cases (Table 3). Smaller MSEP represents models with greater accuracy in estimating differences between cultivars in various environments since it predicted yield differences most accurately in the validation set. Models with environment fixed in all most case outperformance the models with environment random (Table 3). The MSEP each year of the line data and hybrid data and the MSEP average for both data show this (Table 3). The models MNZ 2 and 3 are in the four top models in MSEP average, that is, they had the lowest values of MSEP. And the models MNZ 8, 7, and 10 performed worst among all investigated models for both data, that is, the largest MSEP average. For the models with locations fixed, the MNZ 3 is the best model in both datasets, line, and hybrid data. When, considering the models with locations random, MNZ 9 is the best model in both datasets.

The inclusion of zone effects in the model reduced the MSEP for most cases (Table 3). The models with zones were the best model in five years out of a total of eight evaluated years

(Table 3). It is worth mentioning that the years 2009, 2010, and 2014 do not have zone effect due has a representation of only two zones (2009 and 2014) or one location per zone (2010), it is not possible to perform the analyses. Also, in general, the compound symmetry one-stage model (MOS17) had better performance than the equivalent two-stage model (MNZ4). However, as expected, the differences between the MSEP of the two models are small for most case.

Table 3. The Mean squared error of prediction (MSEP) of 17 statistical models used in the single-year series cross-validation.

		Line data							Hybrid data					
		2009	2010	2011	2012	2013	2014	Mean	2014	2015	2016	2017	2018	Mean
Fixed	MSF 1	4.97	5.11	6.00	4.21	5.84	3.78	4.99	11.79	5.31	11.20	40.03	14.05	16.48
No Zone	MNZ 2	4.45	**4.71**	5.85	4.09	5.65	3.56	4.72	11.45	5.15	10.76	39.70	13.74	16.16
	MNZ 3	4.36	4.73	**5.83**	4.13	5.69	3.63	4.73	10.24	5.11	10.78	39.53	**13.72**	15.88
	MNZ 4	4.63	4.82	5.88	3.98	5.69	3.47	4.75	12.17	5.09	10.89	39.93	14.11	16.44
	MNZ 5	4.57	4.87	5.90	4.08	5.69	3.52	4.77	10.95	5.13	10.88	39.89	13.84	16.14
	MNZ 6	4.40	4.70	5.94	4.15	5.86	3.61	4.78	12.60	5.18	11.17	40.93	14.26	16.83
	MNZ 7	**3.82**	6.11	7.01	6.13	7.09	5.11	5.88	11.05	5.79	12.69	52.22	17.66	19.88
	MNZ 8	**3.82**	6.11	7.01	6.13	7.09	5.11	5.88	11.05	5.79	12.69	52.22	17.66	19.88
	MNZ 9	4.46	**4.71**	5.87	3.97	5.66	3.50	4.70	12.08	5.10	10.90	40.80	14.40	16.66
	MNZ 10	4.42	5.18	6.30	5.03	6.13	4.38	5.24	12.32	5.24	12.87	51.28	14.63	19.27
	MNZ 11	4.40	4.70	5.95	4.15	5.86	3.61	4.78	12.60	5.18	11.17	40.93	14.23	16.82
Zone	MZ 12	-	-	6.02	4.11	5.73	3.48	4.83	-	5.07	11.31	39.21	15.80	17.84
	MZ 13	-	-	6.02	4.30	**5.55**	3.57	4.86	-	**5.00**	**10.60**	**39.10**	15.11	17.45
	MZ 14	-	**-**	5.87	3.96	5.80	**3.45**	4.77	-	5.09	10.84	40.95	14.51	17.84
	MZ 15	-	-	5.93	4.32	5.57	3.54	4.84	-	5.11	11.31	43.91	14.62	18.73
	MZ 16	-	-	5.95	4.15	5.90	3.62	4.87	-	5.20	11.10	42.00	14.30	18.15
One Stage	MOS 17	4.61	4.80	5.89	**3.94**	5.57	3.45	4.71	13.20	5.11	10.83	41.40	14.05	16.92

MSF, model fixed; MNZ, model no-zone, MZ, model zone, MOS, model one-stage; Bold values present the model with lowest MSEP in each year

Table 4 are represented the ranking of cultivars, between the current-practice model, MSF 1, and the best model in the last year of evaluation, according to the MSEP, to the line and hybrids data, MZ 15 and MNZ 3, respectively. According to the line data, the best cultivar in the MSF 1 model was genotype CMSXS5008, and in the MZ 15, genotype BRS506. Which shows the difference in the ranking of genotypes between both models. Were selected five best cultivars, because of the intensity of selection adopted to the sweet sorghum program. According to the MSF1 model, the best genotypes were CMSXS5008, BRS506, BRS511, CMSXS5007, and CV568, respectively, and according to the MZ15 model, the best genotypes were BR506, BRS511, CMSXS5008, CMSXS5009, and CMSXS630, respectively. On the other hand, five genotypes were the worst cultivars (i.e., minor predict value), for both models, but the ranking was not the same.

In the hybrid data, there is a different ranking between the MSF1 and MNZ 3 models, however, the selection of the top five cultivars (CMSXS5501A x CMSXS5021, CMSXS5507A x CMSXS5021, CMSXS5501A x CMSXS5022, CMSXS5507A x CMSXS5022 and CMSXS5508A x CMSXS5022) was the same for both models (Table 4). According to the MSF1 model, the best genotypes were CMSXS5501A x CMSXS5021, CMSXS5507A x CMSXS5021, CMSXS5501A x CMSXS5022, CMSXS5508A x CMSXS5022 and CMSXS5507A x CMSXS5022, respectively, and according to the MNZ 3 model, the best genotypes were CMSXS5501A x CMSXS5021, CMSXS5507A x CMSXS5021, CMSXS5508A x CMSXS5022, CMSXS5501A x CMSXS5022, and CMSXS5508A x CMSXS5022, respectively. And for the worst cultivars, considering the same selection intensity, the rank was the same between both models, being the CMSXS5503A x CMSXS5014, CMSXS5505A x CMSXS5012, CMSXS5507A x CMSXS643, CMSXS5507A

x CMSXS5015, and CMSXS5507A x CMSXS5014, respectively, the genotypes with the minor predict values.

Table 4. Rank of the predict values of the cultivars to trait tons of Brix per hectare (TBH) according to line and hybrid data in comparing models: the model single fixed (MSF 1) is the current-practice model in sweet sorghum program and the model zone (MZ 15) and model no-zone (MNZ 3) model with smaller Mean Squared Error of Prediction (MSEP) for the last year of each data set.

Line Data					Hybrid Data				
	MSF 1		MZ 15			MSF 1		MNZ 3	
Rank	Genotype		Genotype		Rank	Genotype		Genotype	
1	CMSXS5008	9.7	BRS506	9.5	1	CMSXS5501AxCMSXS5021	13.2	CMSXS5507AxCMSXS5021	12.12
2	BRS506	9.6	BRS511	9.5	2	CMSXS5507AxCMSXS5021	12.5	CMSXS5501AxCMSXS5021	11.98
3	BRS511	9.4	CMSXS5008	9.4	3	CMSXS5501AxCMSXS5022	11.5	CMSXS5508AxCMSXS5022	10.82
4	CMSXS5007	9.2	CMSXS5009	9.4	4	CMSXS5508AxCMSXS5022	11.4	CMSXS5501AxCMSXS5022	10.62
5	CV568	9.2	CMSXS630	9.1	5	CMSXS5507AxCMSXS5022	11.1	CMSXS5507AxCMSXS5022	10.33
6	CMSXS643	9.1	CV568	9.1	6	CMSXS5507AxCMSXS5017	10.2	CMSXS5507AxCMSXS5017	9.91
7	CV198	9	CMSXS5006	9	7	CMSXS5503A xCMSXS5016	8.82	CMSXS5502AxCMSXS5019	8.77
8	CMSXS630	9	CV198	8.9	8	CMSXS5502AxCMSXS5019	8.79	CMSXS5503AxCMSXS5016	8.75
9	CMSXS646	9	CMSXS643	8.9	9	CMSXS646	8.22	CMSXS46	8.23
10	CMSXS5009	9	CMSXS647	8.8	10	115	8.07	115	8.07
11	CMSXS647	8.9	CMSXS646	8.8	11	CMSXS5503AxCMSXS5019	7.83	CMSXS5503A xCMSXS5019	7.9
12	BRS509	8.7	CMSXS648	8.5	12	BRS511	7.43	BRS511	7.57
13	CMSXS648	8.4	BRS509	8.5	13	100	7.4	BRS508	7.51
14	Sugargraze	8.2	Sugargraze	8.3	14	CMSXS643	7.39	100	7.51
15	CMSXS639	8.1	CMSXS639	8.2	15	BRS508	7.37	CMSXS643	7.45
16	CMSXS629	7.7	CMSXS629	7.9	16	CMSXS5507AxCMSXS646	7.35	CMSXS5507AxCMSXS646	7.44
17	V82392	7.5	V82391	7.8	17	CMSXS5504AxCMSXS5015	7.17	CMSXS5504AxCMSXS5015	7.24

18	V82391	7.4	CMSXS5003	7.7	18	CMSXS5506AxCMSXS5015	7.05	CMSXS5506AxCMSXS5015	7.12
19	CMSXS5003	7.1	V82392	7.5	19	CMSXS5502AxCMSXS5015	7.04	CMSXS5502AxCMSXS5015	7.1
20	V82393	6.8	CMSXS5010	7.4	20	CMSXS5503AxCMSXS5015	6.93	CMSXS5503AxCMSXS5015	7.02
21	CMSXS5010	6.7	V82393	7.2	21	CMSXS5503AxCMSXS5014	6.85	CMSXS5503AxCMSXS5014	6.93
22	CMSXS644	6.7	CMSXS644	6.9	22	CMSXS5505AxCMSXS5012	6.72	CMSXS5505AxCMSXS5012	6.88
23	CMSXS5004	6.4	CMSXS5004	6.8	23	CMSXS5507A x CMSXS643	6.61	CMSXS5507Ax CMSXS643	6.74
24	CMSXS5006	5.5	CMSXS5006	6.2	24	CMSXS5507AxCMSXS5015	6.35	CMSXS5507AxCMSXS5015	6.57
25					25	CMSXS5507AxCMSXS5014	5.5	CMSXS5507AxCMSXS5014	5.77

DISCUSSION

In this study, we investigated the impact of the cluster zone effects via climate data in the prediction of VCU trials and compared the effect of modeling the different linear mixed models that use either BLUE or BLUP under unweighted and weighted two-stage analyses. Our results suggest that models with (co)variance structure outperformed the model with fixed effects, currently adopted model in the sweet sorghum program. In addition, target populations of environments, stratified through environmental covariates increased the model precision based on cross-validation scenarios.

Component σ^2_{CZ} showed significance only in three years in both datasets (Table 2). The result is expected because the sweet sorghum breeding program of Embrapa maize and sorghum made the recommendation based on general mean and recommends the cultivars for all country (de Figueiredo et al., 2015; Eculica et al., 2019), without adopting zone information. Were defined the zones from information climate data (unpredictability) and geographic position (predictability) (Allard and Bradshaw, 1964; Mirzawan. et al., 1994), and were grouped environments in the zone (predictability) according to historical climate data during the harvest season sweet sorghum in tropical conditions (Figure 3). Highlights, in group G2 formed by k-means, the municipalities: Jaíba, Nova Porteirinha, Teresina, Mata Roma, Penedo, Paripiranga, Teotônio Vilela, are predominantly from the region northeast and northeast of stage of Minas Gerais, that are characterized by being in a zone of Brazilian caatinga vegetation and transition region cerrado/caatinga. In group G3, the grouped municipalities southeast (Lavras, Piracicaba, Jaguariuna and Sete Lagoas) and central-west region (Goiania and Planaltina) are similar characteristics. On the other hand, the results showed the of k-means in grouped similar locations and the importance of modeling the zone effect in the sweet sorghum breeding program.

It is worth mentioning that this is the first study creating zones with only climate data in sweet sorghum. Second Piepho and Möhring (2005), and Van Eeuwijk et al. (2016), the stratification in zones can help to minimize the problem with interaction, mainly under tropical conditions (Dias et al., 2020). Oliveira et al. (2020) reported in studies with biomass sorghum, a high correlation between the sites evaluated in VCU trials by Embrapa, this correlation between environments can be helped to define mega-environments, which confirms the need for zoning for Embrapa's bioenergy sorghum breeding program. When trials or environments are grouped based on phenotypic or environmental information from the trials, are formed a mega-environments, or, sometimes, adaptation zones (Gauch and Zobel, 1997; Yan et al., 2000; Malosetti et al., 2013; Van Eeuwijk et al., 2016). This produce a TPE defined to a set of environments in which we want our inference and predictions to be precise and valid (Van Eeuwijk et al., 2016). Example of TPE use in breeding programs were reported in winter wheat (*Triticum aestivum* L.), spring barley (*Hordeum vulgare* L.) (Buntaran et al., 2019, 2020a), cereal growing areas in Australia (Smith et al., 2015).

The currently used fixed model for VCU recommendation in the Embrapa's breeding program, had the highest MSEP (Table 3). Models with (co)variance structure outperformed compared to MSF 1. In all the scenarios tested via cross-validation, the inclusion of the zone effects, defining from climate data, improved the MSEP. The years 2009, 2010, and 2014 can't calculate the MSEP because has a representation of only two zones or show one location for each zone. The process of cross-validation accomplished doesn't allow. Models with environment fixed in all most case outperformance the models with environment random (Table 2). However, in the breeding programs, it is not wished to predict environmental effects, therefore environmental effects are fixed in VCU trail of sweet sorghum. It is known that the critical area, in this context, is the definition of regions (Smith et al., 2001). For this purpose, the inclusion of zone effects in sweet sorghum can be helpful in the recommendation of

cultivars and adopt strategy the design of mega-environments. This inclusion can improve the selection response and optimize the use of resources (González Barrios et al., 2019). Thus, the sweet sorghum areas of Brazil can subdivide into target breeding zones similar propose to Dias et al. (2020) the maize growing areas of Brazil can be subdivided into target breeding zones widely explored by breeders. For these reasons, we concluded that for the routine analysis of VCU, models that recover information from correlated zones should be used instead of the currently used model. Therefore, this study shows that edaphoclimatic covariates can use successfully to improve the predictions. Further studies using molecular information and data from multiply years deserve attention in this topic. The use of zones can direct breeders to discard locations in future evaluations, optimizing the resources to evaluate other untested (or poorly tested) geographic regions in Brazil.

Stage-wise analysis using weights provides a measure of uncertainty for the means, being a good approximation or similar results to a single-stage model depending on the weights method (Smith et al., 2001; Piepho et al., 2012). However, one-stage, MOS 16, and two-stage models (MNZ4) showed similar results for MSEP (Table 2). Gogel et al. (2018) compared one-stage versus two-stage, and Damesa et al. (2017) analyzed of series experiments and reported similar results to those found in our study. The inclusion of weights improved the models and reported a better MSEP. Similar conclusions were already reported in simulated data (Piepho et al., 2012), cross-validation with zone effects (Buntaran et al. 2020), and maize hybrids in tropical conditions (Dias et al., 2020). The dataset used in this study of unbalanced historical data and reported heterogeneity of variances, reinforcing the importance use weights into account the uncertainty regarding adjusted means, similar conclusion to Smith et al. (2015), Damesa et al. (2017), and Gogel et al. (2018) of used weights in analysis MET.

The BLUP has been used for selecting plants in breeding programs with success (Barbosa et al., 2005; Piepho et al., 2008; Slater et al., 2014; Chavarría-Perez et al., 2020). In

the present study, models with genotype random (location fixed or random) and models with zone effects outperformed the current model, genotype fixed, adopted to the sweet sorghum breeding program in both datasets. Duarte and Vencovsky (2001) proposed to reflect the process of estimation (BLUE) and prediction (BLUP) of treatment average or random-effects cultivars in trials with block designs in plant breeding and verified the importance of this process to determine the rank of genotypes. We observed the different rankings in both datasets when we compared the best model in 2014 and the current model in the program. In the line data, for example, using the model MZ 15 the genotype BRS506 was the best cultivar, already using the current model (MSF1) the best cultivar was another, CMSXS5008. The cultivar BRS506 is one line from the sweet sorghum program, selected during years and great potential to produce hybrids with high performance, and the cultivar CMSXS5008 was developed in crosses involved in Brandes and Wray, the most of these CMSXS lines development to Embrapa used these parents (da Silva et al., 2017), however, the cultivar CMSX5008 was discarded for the sweet sorghum program for not meeting the desirable traits. In the model MSF1, the cultivar BRS506 was the second better cultivar, highlights that CMSXS630 shows in the MZ 15 and the genotype CV568 in the MSF1; the CV568 is a hybrid of the private company and discard because it doesn't have great potential and, the CMSXS630 is a line development to Embrapa and shows a great traits for ethanol production, so that models with BLUP show more reliable estimates that can determine differences in the classification of genotypic and minimizes selection errors.

In the hybrid data, the top five cultivars were same in compared both models in the year 2018. However, the positions changed, and change the BLUP show more reliable estimates and determine differences in rank.

CONCLUSION

In this study, based on two sweet sorghum datasets, our result showed that delimitation of the target population of environments via environmental covariable could increase cultivar effects prediction under tropical conditions. Also, as expected, the use of weights stage-wise analysis showed similar performance of the one-stage model. Moreover, we performed thorough provided insights into the analysis of VCU in sweet sorghum to assess the performance of random or fixed-cultivar-effects models, also shows the advantage of BLUP in sweet sorghum breeding program routine.

REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. EEE Trans. Autom. Control 19(6): 716–723.

Allard, R.W., and A.D. Bradshaw. 1964. Implications of Genotype-Environmental Interactions in Applied Plant Breeding 1. Crop Sci. 4(5): 503–508.

Barbosa, M.H.P., M.D.V. de Resende, J.A. Bressiani, L.C.I. da Silveira, and L.A. Peternelli. 2005. Selection of sugarcane families and parents by Reml/Blup Crop Breeding and Applied Brazilian Society of Plant Breeding. Printed in Brazil Selection of sugarcane families and parents by Reml/Blup.

Buntaran, H., H. Piepho, J. Hagman, and J. Forkman. 2019. A Cross-Validation of Statistical Models for Zoned-Based Prediction in Cultivar Testing. Crop Sci. 59(4): 1544–1553. doi: 10.2135/cropsci2018.10.0642.

Buntaran, H., H. Piepho, P. Schmidt, J. Rydén, M. Halling, et al. 2020a. Cross-validation of stage-wise mixed-model analysis of Swedish variety trials with winter wheat and spring barley. Crop Sci.: csc2.20177. doi: 10.1002/csc2.20177.

Buntaran, H., H. Piepho, P. Schmidt, J. Rydén, M. Halling, et al. 2020b. Cross-validation of

stagewise mixed-model analysis of Swedish variety trials with winter wheat and spring barley. Crop Sci.: csc2.20177. doi: 10.1002/csc2.20177.

Butler, D.G., B.R. Cullis, A.R. Gilmour, B.J. Gogel, and R. Thompson. 2018. ASReml-R Reference Manual Version 4 ASReml estimates variance components under a general linear mixed model by residual maximum likelihood (REML).

Chavarría-Perez, L.M., W. Giordani, K.O.G. Dias, Z.P. Costa, C.A.M. Ribeiro, et al. 2020. Improving yield and fruit quality traits in sweet passion fruit: Evidence for genotype by environment interaction and selection of promising genotypes (P.E. Teodoro, editor). PLoS One 15(5): e0232818. doi: 10.1371/journal.pone.0232818.

Cullis, B.R., A.B. Smith, and N.E. Coombes. 2006. On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11(4): 381–393. doi: 10.1198/108571106X154443.

Damesa, T.M., J. Möhring, M. Worku, and H.-P. Piepho. 2017. One Step at a Time: Stage-Wise Analysis of a Series of Experiments. Agron. J. 109(3): 845–857. doi: 10.2134/agronj2016.07.0395.

Dias, K.O.G., H.P. Piepho, L.J.M. Guimarães, P.E.O. Guimarães, S.N. Parentoni, et al. 2020. Novel strategies for genomic prediction of untested single-cross maize hybrids using unbalanced historical data. Theor. Appl. Genet. 133(2): 443–455. doi: 10.1007/s00122-019-03475-1.

Duarte, J.B., and R. Vencovsky. 2001. Estimação e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. Sci. Agric. 58(1): 109–117. doi: 10.1590/S0103-90162001000100017.

Eculica, G.C., P.C.D.O. Ribeiro, A.B. De Oliveira, N.N.L.D. Parrella, P.S.D.S. Leite, et al.

2019. Adaptability and stability of saccharine sorghum cultivars. African J. Agric. Res. 14(31): 1432–1442. doi: 10.5897/AJAR2019.14043.

Van Eeuwijk, F.A., D. V Bustos-Korts, and M. Malosetti. 2016. What should students in plant breeding know about the statistical aspects of genotype$\times$ environment interactions? Crop Sci. 56(5): 2119–2140.

de Figueiredo, U.J., J.A.R. Nunes, R.A.C. Parrella, E.D. Souza, A.R. da Silva, et al. 2015. Adaptability and stability of genotypes of sweet sorghum by GGEBiplot and Toler methods. Genet. Mol. Res. 14(3). doi: 10.4238/2015.September.22.15.

Gauch, H.G., and R.W. Zobel. 1997. Identifying Mega-Environments and Targeting Genotypes. Crop Sci. 37(2): 311–326. doi: 10.2135/cropsci1997.0011183X003700020002x.

Gogel, B., A. Smith, and B. Cullis. 2018. Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. Euphytica 214(2). doi: 10.1007/s10681-018-2116-4.

González Barrios, P., L. Díaz-García, and L. Gutiérrez. 2019. Mega-Environmental Design: using genotype by environment interaction to optimize resources for cultivar testing. Crop Sci. doi: 10.2135/cropsci2018.11.0692.

Hartigan, J.A., and M.A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm.

Malosetti, M., J.M. Ribaut, and F.A. van Eeuwijk. 2013. The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. Front. Physiol. 4 MAR(March): 1–17. doi: 10.3389/fphys.2013.00044.

Mirzawan., P.D.N., M. Cooper, I.H. DeLacy, and D.M. Hogarth. 1994. Retrospective analysis of the relationships among the test environments of the Southern Queensland sugarcane

breeding programme. Theor. Appl. Genet. 88: 707–716.

Oliveira, I.C.M., J.H.S. Guilhen, P.C. de O. Ribeiro, S.A. Gezan, R.E. Schaffert, et al. 2020. Genotype-by-environment interaction and yield stability analysis of biomass sorghum hybrids using factor analytic models and environmental covariates. F. Crop. Res. 257(August). doi: 10.1016/j.fcr.2020.107929.

Piepho, H.P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theor. Appl. Genet. 97(1–2): 195–201. doi: 10.1007/s001220050885.

Piepho, H.P., J. Möhring, A.E. Melchinger, and A. Büchse. 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161(1–2): 209–228. doi: 10.1007/s10681-007-9449-8.

Piepho, H.P., J. Möhring, T. Schulz-Streeck, and J.O. Ogutu. 2012. A stage-wise approach for the analysis of multi-environment trials. Biometrical J. 54(6): 844–860. doi: 10.1002/bimj.201100219.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

da Silva, M.J., M.M. Pastina, V.F. de Souza, R.E. Schaffert, P.C.S. Carneiro, et al. 2017. Phenotypic and molecular characterization of sweet sorghum accessions for bioenergy production (J.-M. Lacape, editor). PLoS One 12(8): e0183504. doi: 10.1371/journal.pone.0183504.

Slater, A.T., G.M. Wilson, N.O.I. Cogan, J.W. Forster, and B.J. Hayes. 2014. Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. Theor. Appl. Genet. 127(4): 809–820. doi: 10.1007/s00122-013-2258-7.

Smith, A., B. Cullis, and A. Gilmour. 2001. The analysis of crop variety evaluation data in Australia. Aust. New Zeal. J. Stat. 43(2): 129–145.

Smith, A.B., A. Ganesalingam, H. Kuchel, and B.R. Cullis. 2015a. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. Theor. Appl. Genet. 128(1): 55–72. doi: 10.1007/s00122-014-2412-x.

Smith, A.B., A. Ganesalingam, H. Kuchel, and B.R. Cullis. 2015b. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. Theor. Appl. Genet. 128(1). doi: 10.1007/s00122-014-2412-x.

Yan, W., L.A. Hunt, Q. Sheng, and Z. Szlavnics. 2000. Cultivar evaluation and mega-environment investigation based on the GGE biplot. Crop Sci. 40(3): 597–605. doi: 10.2135/cropsci2000.403597x.

SUPPLEMENTARY TABLE INFORMATION

Table S1. Description of the locations tested for line data Value for Cultivation and Use trials and their respective years and locations.

Environments	ID*	Latitude	Longitude	Altitude	2009	2010	2011	2012	2013	2014
Boa Vista	BV	-2.82 S	-50.67 W	85			8		21	
Cáceres	CA	-16.20 S	-57.68 W	459				16		
Capão do Leão	CL	-31.76 S	-52.48 W	21				15		
Dourados	DO	-22.22 S	-54.81 W	437						36
Guaira	GU	-8.83 S	-48.50 W	259		4				33
Jaíba	JÁ	-15.20 S	-43.40 W	457	1					
Lavras	LA	-21.25 S	-45.00 W	919					23	35
Nova Porteirinha	NP	-15.78 S	-43.30 W	518	2	5	9	17	24	34
Paripiranga	PA	-10.69 S	-37.85 W	430					22	
Passo Fundo	PF	-28.26 S	-52.41 W	687			10			
Pelotas	PE	-31.77 S	-52.34 W	7		6	11	18		
Piracicaba	PI	-22.73 S	-47.65 W	528					25	
Planaltina	PL	-17.58 S	-47.71 W	971					26	
Santa Vitória	SV	-18.85 S	-50.12 W	498				19	27	
Sete Lagoas	SL	-19.47 S	-44.25 W	751	3	7	12	20	28	32
Sinop	SI	-11.85 S	-55.65 W	384			13		29	
Teotônio Vilela	TV	-9.91 S	-36.35 W	156					30	
Teresina	TE	-5.09 S	-42.80 W	87			14			
Vilhena	VI	-12.73	-60.14 W	612					31	

*ID: coded to environments.

Table S2. Description of the locations tested for hybrid data Value for Cultivation and Use trials and their respective years and locations.

Environments	ID*	Latitude	Longitude	Altitude	2014	2015	2016	2017	2018
Goiânia	GO	-16.49 S	-49.31 W	749			47		64
Jaguariúna	JG	-22.71 S	-46.99 W	732	39			55	62
Lavras	LA	-21.25 S	-45.00 W	919		40	48	56	
Nova Porteirinha	NP	-15.78 S	-43.30 W	518	38	41	49	57	63
Passo Fundo	PF	-28.26 S	-52.41 W	687		42			
Pelotas	PE	-31.77 S	-52.34 W	7		43			
Planaltina	PL	-17.58 S	-47.71 W	971		44	51	59	66
Sete Lagoas	SL	-19.47 S	-44.25 W	751	37	45	52	58	61
Sinop	SN	-11.85 S	-55.65 W	384		46	53	60	65
Vilhena	VI	-12.73	-60.14 W	612			54		

*ID: coded to environments.

Table S3. Number of common genotypes between year in the diagonal and under the diagonal, and number of similar genotypes environments per year.

Line data							Hybrid data					
Year	2009	2010	2011	2012	2013	2014	Year	2014	2015	2016	2017	2018
2009	25	25	19	13	10	11	2014	25	13	9	6	3
2010		25	19	13	10	11	2015		25	13	6	3
2011			25	16	13	15	2016			25	9	4
2012				25	11	17	2017				25	3
2013					25	15	2018					25
2014						16	-	-	-	-	-	-
NL*	4	4	5	9	8	5		3	10	9	6	6

* Number of environments

Table S4. Estimates of the phenotypic mean (Mean); block (σ_b^2), genetic (σ_g^2), and residual (σ_e^2) variance components; and generalized heritability (H^2). In parentheses test LRT for genotype (LRT). Estimates for the 37 trials conducted by the Embrapa's sweet sorghum breeding program to evaluate tons of Brix per hectare (TBH).

Year	ID	Mean	σ_b^2	σ_g^2	σ_e^2	H^2
2009	JB	7.36	0.43	0.68 (8.48*)	1.03	0.63
	NP	7.87	0.07	2.27 (38.81*)	0.83	0.89
	SL	9.36	0.00	1.53 (15.69*)	1.60	0.74
2010	GU	12.32	0.00	3.19 (20.17*)	2.65	0.78
	NP	8.35	0.23	2.59 (20.41*)	1.69	0.80
	PE	5.46	0.39	0.87 (11.19*)	1.04	0.68
	SL	6.53	0.52	2.50 (13.53*)	2.65	0.72
2011	BV	5.23	0.89	2.01 (21.04*)	1.26	0.80
	NP	6.44	0.07	3.83 (48.85*)	1.06	0.91
	PF	9.79	0.78	2.75 (5.33*)	6.42	0.54
	PE	5.64	1.90	0.93 (2.24*)	2.36	0.42
	SL	5.95	0.52	0.93 (5.64*)	2.02	0.55
	SI	6.90	0.20	5.70 (33.69*)	2.58	0.86
	TE	6.08	0.00	3.10 (12.5*)	2.79	0.62
2012	BV	4.09	0.28	1.84 (18.88*)	1.24	0.77
	CL	3.82	0.03	1.03 (40.15*)	0.37	0.89
	CA	5.18	0.26	0.51 (7.87*)	0.88	0.61
	PA	7.63	0.49	5.12(36.22*)	1.79	0.86
	PE	4.47	1.19	0.65 (7.71*)	0.51	0.65
	NP	7.08	1.36	5.45 (25.01*)	2.04	0.86
	SV	8.27	0.00	2.23 (6.50*)	4.82	0.58
	SL	4.84	0.95	2.24 (20.47*)	1.61	0.78
2013	LA	10.38	0.00	1.28 (3.27*)	3.23	0.54
	NP	8.44	0.11	1.19 (6.77*)	1.68	0.67
	PI	7.91	0.22	2.47 (8.78*)	2.70	0.72
	PL	4.95	0.03	1.40 (63.78*)	0.09	0.98
	SV	9.17	1.04	3.60 (6.65*)	3.95	0.71
	SL	8.34	0.14	1.67 (8.55*)	1.86	0.72
	SI	5.62	0.39	8.61 (36.29*)	1.40	0.94
	TV	5.49	0.21	0.82 (6.25*)	1.03	0.68
	VI	3.95	0.05	4.39 (67.38*)	0.25	0.98
2014	SL	10.40	0.81	2.69 (13.93*)	2.76	0.72
	GU	4.23	0.01	0.59 (9.89*)	0.89	0.67
	NP	10.35	0.00	3.77 (17,20*)	3.90	0.75
	LA	7.54	0.00	1.39 (21.50*)	1.14	0.79

*, nsSignificantand no-significantaccordingtothex^2 test ($\alpha = 0.05$ respectivament); BV: Boa Vista, CA: Cárceres, CP: Capão do Leão, DO: Dourados, GU: Guaíra, JA: Jaíba, LA: Lavras, NP: Nova Porteirinha, PA:Paripiranga, PF: Passo Fundo, PE: Pelotas,PI: Piracicaba, PL: Planaltina, SV: Santa Vitória, SL: Sete Lagoas, SI: Sinop, TV:

Teotônio Vilela, VI: Vilhena.

Table S5. Estimates of the phenotypic mean (Mean); block (σ_b^2), genetic (σ_g^2), and residual (σ_e^2) variance components; and generalized heritability (H^2). In parentheses test LRT for genotype (LRT). Estimates for the 33 trials conducted by the Embrapa's sweet sorghum breeding program to evaluate tons of Brix per hectare (TBH).

Year	ID	Mean	σ_b^2	σ_g^2	σ_e^2	H^2
2014	SL	5.61	0.40	0.41 (7.35*)	0.67	0.61
	NP	18.57	5.67	3.30 (3.22*)	8.34	0.51
	JG	10.50	0.61	6.99 (23.84*)	4.03	0.83
2015	PN	8.81	0.61	1.82 (18.97*)	1.17	0.80
	DO	14.35	3.38	2.99 (7.24*)	5.18	0.60
	JG	7.97	1.09	1.92 (1.61ns)	9.95	0.35
	LA	11.42	0.00	0.51 (0.45ns)	5.76	0.21
	NP	11.94	0.00	4.82 (20.49*)	3.95	0.79
	PF	5.31	0.00	0.99 (2.74*)	3.87	0.43
	PE	4.09	0.01	0.27 (14.64*)	0.30	0.73
	PL	5.84	0.08	0.47 (21.1*)	0.33	0.79
	SL	9.47	0.00	0.59 (4.15*)	1.76	0.50
	SI	8.23	0.49	1.93 (7.08*)	3.58	0.60
2016	GO	9.30	1.11	6.21 (24.84*)	3.94	0.81
	LA	11.92	2.81	2.88 (5.65*)	6.02	0.56
	NP	13.62	1.19	12.41(30.82*)	5.72	0.86
	PL	6.97	0.00	0.23 (1.02ns)	1.61	0.30
	PE	10.80	0.24	6.50 (45.01*)	1.85	0.91
	SL	9.45	0.97	4.08 (23.67*)	1.87	0.85
	SI	8.46	0.15	3.31 (40.31*)	1.19	0.89
	VI	10.01	0.26	6.15 (37.62*)	2.41	0.88
2017	JG	14.96	0.24	15.32 (92.30*)	1.32	0.97
	NP	14.51	2.46	55.99 (80.53*)	4.96	0.97
	SL	11.99	0.20	2.29 (16.85*)	3.63	0.65
	LA	7.81	0.32	7.22 (35.28*)	3.01	0.86
	PL	7.06	0.20	4.92 (80.14*)	0.50	0.96
	SI	16.41	0.00	63.23 (60.44*)	11.67	0.94
2018	SL	9.98	0.70	6.42 (30.86*)	2.80	0.86
	JG	4.65	0.55	1.82 (10.51*)	1.48	0.67
	NP	8.05	0.46	14.78 (73.13*)	1.82	0.96
	GO	8.42	0.33	8.95 (67.13*)	1.36	0.95
	SI	9.82	0.00	7.78 (34.51*)	3.61	0.87
	PL	8.47	0.00	6.65 (81.21*)	0.83	0.96

*, nsSignificantand no-significantaccordingtothex^2 test ($\alpha = 0.05$ respectivament);DO: Dourados, GO: Goiânia, JG: Jaguariúna, LA: Lavras, NP: Nova Porteirinha, PF: Passo Fundo, PE: Pelotas, PN: Penedo, PL: Planaltina, SL: Sete Lagoas, SI: Sinop,VI: Vilhena.

CAPITULO 2

HYBRID PREDICTION OF BIOMASS SORGHUM USING CLIMATIC DATA VIA COMBINING ABILITY MODELS

ABSTRACT

The biomass of hybrid sorghum is used in the generation of bioenergy. The breeding programs are interested in developing superior hybrids for a wide range of environmental conditions; however, the cost of field-testing is expensive and time-consuming. Genomic Selection (GS) is a powerful tool that allows breeders to predict the performance of new hybrids in yet to observe untested environments. GS models coupled with the use of environmental covariables (ECs) have the potential to enhance selection accuracy in breeding programs. The goals of this study were to: i) evaluate GS models to predict the biomass of untested sorghum hybrids in tested/untested environments, using a historical dataset of the Embrapa's breeding program; ii) compare the effect of modeling different environmental kinship matrices with ECs; iii) identify mega-environments for sorghum hybrids based only on ECs. Evaluations were conducted using seven different models including main effects of environments; general combining ability (GCA) and specific combining ability (SCA) terms, interactions between genetic effects (GCA and SCA) with environments and ECs for different environmental kinship matrices based on ECs. The incorporation of GCA and SCA and interaction with environmental factors showed improvements in predictive ability for cross-validation CV1 (18%), CV2 (17%), and CV0 (8%) in comparison to the baseline model. Also, it was possible to identify mega-environments for energy sorghum cultivation in Brazil based only on climatic data in evaluation sites of Embrapa.

Keywords: Historical data. Predictive ability. Environmental factors.

INTRODUCTION

The biomass of sorghum [Sorghum bicolor (L.) Moench] has multiple uses in the generation of bioenergy, such as first and second-generation ethanol, cogeneration of energy through the combustion of dry biomass, production of biogas, and it has the potential of supply chain of bioenergy sector (Rezende and Richardson, 2017; Gomes et al., 2019). Sorghum features include fast growth, high yield potential for bioenergy purposes (Appiah-Nkansah et al., 2019). Sorghum is an autogamous specie and the discovery of the cytoplasmic male sterility further allowed the production of hybrids (Smith and Frederiksen, 2000; Pfeiffer et al., 2010). Hybrid improvement in sorghum faces challenges such as the development of elite parental lines (A-lines (sterile line) and R-lines (male line)), that result in hybrids with highly fresh biomass yield, due the highly correlated to dry biomass yield, which is the most important trait for cogeneration bioenergy. Besides that, significant genotype-by-environment (GxE) interactions resulting in inconsistent response patterns under tropical conditions (Oliveira et al., 2020).

The sorghum breeding program of the Brazilian Agricultural Research Corporation (Embrapa Maize and Sorghum) has developed sorghum hybrids according to the three lines scheme, where the A and B-lines are isogenic, and they only differ in their cytoplasm source. The A-line is sterile and the B-line has a normal cytoplasm, and the R-lines are fertility restorers (Ribeiro et al., 2020). In any breeding program, field trials in multi-locations are essential to validate the performance and provide recommendations for developing hybrids (Smith et al., 2005; Malosetti et al., 2013). The final goal of evaluating hybrid performance across locations in multi-environment trials (MET), is to provide recommendations for releasing new hybrid genotypes (Smith et al., 2001, 2005, 2015).

Breeding programs are interested in developing superior hybrids for a wide range of environmental conditions; but, phenotyping hybrids in multiple environments might be quite expensive (Smith et al., 2001, 2005, 2015). Genomic selection became a powerful tool for predicting the performance of new hybrids in untested environments using and environmental information (Crossa et al., 2017, Jarquin et al., 2021, Crossa et al., 2021). Therefore, genomic prediction holds the potential to accelerate genetic gains in breeding programs for diverse crops (Lopez-Cruz et al., 2015; Bernal-Vasquez et al., 2017; Dias et al., 2018a; Islam et al., 2019; Jarquin et al., 2020). Nonetheless, in biomass sorghum a few studies have been conducted (Yu et al., 2016; de Oliveira et al., 2018; Fernandes et al., 2018; Rice and Lipka, 2019).

Recently, studies have shown strategies involving the inclusion of ECs in genomic prediction models (Jarquín et al., 2014a; Pérez-Rodríguez et al., 2015; Basnet et al., 2019; Millet et al., 2019; Costa-Neto et al., 2021a). However, the incorporation of high dimensional genetic and environmental data is not an easy task (Pérez-Rodríguez et al., 2015), and identifying hidden patterns and specific factors under ECs in field conditions remain challenging, besides identifying critical environmental determinants (Li et al., 2018a).

Models that account simultaneously for genomic relationships between genotypes, via general combining ability (CGA) and specific combining ability (SCA) terms, environments, and environmental covariables (ECs) hold the potential for improving the predictive ability of tested and untested genotypes in observed and unobserved environments (Costa-Neto et al., 2021a; Fonseca et al., 2021; Jarquin et al., 2021). However, these have not been implemented yet in sorghum breeding programs for biomass prediction.

Therefore, the goals of our study were: i) evaluate GS models to predict the biomass of untested sorghum hybrids in tested/untested environments, using a historical dataset of the Embrapa's breeding program; ii) compare the effect of modeling different environmental

kinship matrices with ECs; iii) identify mega-environments for sorghum hybrids based only on ECs

MATERIAL AND METHODS

Sorghum experimental design

For this study, we used historical phenotypic data of Value-Cultive-Use (VCU) from the Embrapa sorghum breeding program: 221 hybrids derived from crossing 46 A-lines and 25 R-lines. For all lines involved in the crosses in this study molecular marker information was available. The lines were genotyped for markers using the genotyping-by-sequencing (GBS) technique (Elshire et al., 2011). Conventional quality control was applied on the molecular markers to ensure the quality of the data. SNPs with more than 25% of missing values, or with a minor allele frequency lower than 5%, or with more than 5% of heterozygous genotypes were discarded. The number of SNP markers that remained in the final analysis was 4,298.

The sorghum hybrids were evaluated in 64 field trials (combination Year-Location) (Table 1) at 14 locations in Brazil, representing nine different states distributed within the Southeastern (Minas Gerais, Rio de Janeiro, and São Paulo), Midwest (Federal District, Goias, Mato Grosso do Sul, and Mato Grosso) and Southern (Rio Grande do Sul) regions of Brazil (Figure 1). The experiments were established between 2011 and 2019. The historical data set is composed of preliminary and late-stage breeding trials. Each trial was planted in a lattice design with three replicates, where the plots consisted of two rows that were five meters long, having 0.7 meters between rows. The plant population was 110,000 plants per hectare. Fresh biomass yield (FBY, ton ha^{-1}) was evaluated, and it was obtained by weighing two rows of each plot and converted to ton ha^{-1}.

Table 1. Description of the site tested for historical data set of hybrids from the Embrapa sorghum breeding and their respective years.

Site\Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
Caceres				18						1
C.Goytacazes					28	36	47	53		4
Dourados				19	29	37				3
Dracena				20						1
Goiania			9	21	30	38	48	54	58	7
Jaguariuna									59	1
Lavras			10	22	31	39	49			5
N.Porteirinha	1*	5	11	15#,23	32	40		55	60	9
Pelotas	2			14	33	41				4
Planaltina						42	50		61	3
S.Vitoria		7	14							2
S.Lagoas	3	4#,6	8#,13	16#,17#,26	27#,35	45	46#,52	56	64	15
Sinop			12	25	34	43	51	57	62	7
Vilhena						44			63	2
N of fields	3	4	7	12	9	10	7	5	7	64

C.Goytacazes: Campus do Goytacazes, N.Porteirinha: Nova Porteirinha, S. Vitoria: Santa Vitoria, S.Lagoas: Sete Lagoas, *each combination Location/Year receive one code number, #preliminary breeding trials.

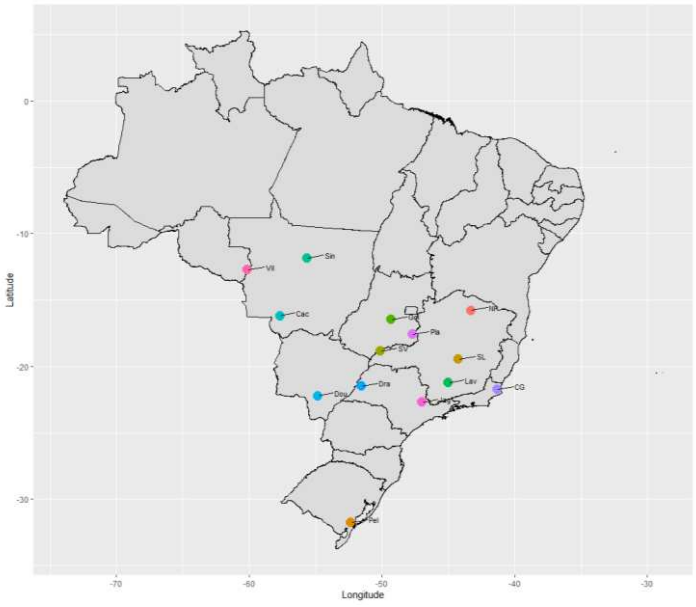

Figure 1. The locations of the sorghum field trials of the Embrapa's breeding program. Cac: Caceres, CG: Campo Grande, Dra: Dracena, Dou: Dourados, Goi: Goiania, Jag: Jagauriuna, Lav: Lavras, NP: Nova Porteirinha, Pel: Pelotas, Pla: Planaltina, SL: Sete lagoas, SV: Santa Vitoria, Sin: Sinop, Vil: Vilhela.

Environmental Covariates

Environmental information comprises data collected between the planting and the harvest seasons for each test location. The planting date was considered as day one in all

environments and the smaller cycle of season in the field trial (Table S1) offered the opportunity to compare crop performance under similar weather conditions. Geographical coordinates (latitude and longitude) and altitude information were collected for each trial. In addition, the environmental covariates were retrieved and processed using the R package *EnvRtype* (Costa-Neto et al., 2021b) (Table 2). This data processing relates radiation, temperature and atmospheric demands that influence the photosensitive in sorghum (Rooney and Aydin, 1999).

Table 2: Environmental covariables (EC's).

	Environmental Covariable	Unit
1	Elevation	M
2	Longitude	○
3	Latitude	○
4	Mean air temperature at 2 m above the surface of the earth	°C day^{-1}
5	Maximum air temperature at 2 m above the surface of the earth	°C day^{-1}
6	Minimum air temperature at 2 m above the surface of the earth	°C day^{-1}
7	Rainfall precipitation	mm day^{-1}
8	Wind speed at 10 m above the surface of the earth	m/s
9	relative air humidity	%
10	The dew point	°C day^{-1}
11	Downward Thermal Infrared Radiative Flux	MJ m^{-2} day^{-1}
12	Insolation Incident on a Horizontal Surface	MJ m^{-2} day^{-1}
13	Top-of-atmosphere Insolation	MJ m^{-2} day^{-1}
14	Daylight's hours	N, hours
15	Actual duration of sunshine	n, hours
16	Extraterrestrial radiation	MJ m^{-2} day^{-1}
17	Solar Radiation	m^{-2} day^{-1}
18	Vapor pressure deficit	kPa d^{-1}
19	Slope of saturation vapor pressure curve	SP, in kPa °C day^{-1}
20	Potential Evapotranspiration	mm day^{-1}
21	Growing Degree Day	MJ m^{-2} day^{-1}
22	Effect of temperature on radiation use efficiency	-
23	Daily Temperature Range	°C day^{-1}

Statistical analyses

First-stage: Single-environmental trial analyses

Within environments, the following linear mixed model used for accounting for the experimental design for FBY:

$$y_{hjk} = \mu + r_q + b_{jq} + g_h + e_{hjq}$$

where y is a $hjq \times 1$ vector of phenotypes for genotype h block j for replicate q . The μ is the overall mean; r_q is a fixed effects of the q^{th} replicates; b_{jq} is a random effects of the jq^{th}blocks within replicates, with $b \sim N(0, I\sigma_b^2)$; g is a fixed effects of the h^{th} hybrid, with $g \sim N(0, I\sigma_a^2)$; and e_{hjq} is a random error assumed identically (*idd*) and also normally distribution such that, with $e_{hjq} \sim N(0, I\sigma_e^2)$, where σ_e^2 denotes the error variance.

For each environment, we estimate the broad-sense heritability (H^2) on entry means bases as described with the following equation:

$$H^2 = \frac{\sigma_h^2}{\sigma_h^2 + \frac{\sigma_e^2}{r}}$$

where: σ_h^2 , σ_e^2 and r are the genetic variance of hybrid, the residual variance, and the number of replicates in each environment, respectively.

Second-stage: Genomic Prediction Models

To implement Genome Prediction (GP) analysis the genetic relationship matrix between the 221 hybrids is modeled by considering the marker information from the 46 A-lines and 25 R-lines (GRMs, VanRaden 2008) via general and specific combining ability (GCA and SCA) terms. The GCA of A-line and R-line were built using their corresponding marker profiles (Bernardo, 1994; Technow et al., 2014; Kadam et al., 2016), and the SCA, which is the interaction effects of crossing a A-line and an R-line (Acosta-Pech et al., 2017).

Model M1: General Combining Ability Model

This model uses the genomic information from the A-lines and R-lines *via* the GCA of the parents involved in the cross via two genetic scores, and it also includes an environmental effect, and an error term.

Here, for the h^{th} hybrid we denoted the corresponding scores (MARCADOR) for the parental A-line and R-line, g_{a_j} and g_{r_i} respectively, these being the linear combinations

between a $(i = 1,2, \dots, m)$ and r $(i = 1,2, \dots, m)$ markers and the corresponding maker effects. With $g_{a_j} = \sum_{m=1}^{p} x_{a\Im} b_{ma}$ and $g_{r_i} = \sum_{m=1}^{p} x_{rim} b_{mr}$; $X_a = \{X_{aim}\}$, $X_r = \{X_{rim}\}$ are the corresponding inbred marker matrices for the A-line and R-line (Bernardo, 1994; Technow et al., 2014; Kadam et al., 2016); b_{ma}, b_{mr}, are the corresponding effects of the m^{th} marker for A-line and R-line such that $b_{ma}N(0, \sigma_{ba}^2)$ and $b_{mr}N(0, \sigma_{br}^2)$, σ_{ba}^2 and σ_{br}^2 acting as the variance components.

$$\overline{y}_{ht} = \mu + E_t + g_{a_j} + g_{r_i} + e_{ht}$$

where, $\overline{y}_{ht}$ is the mean for each hybrid at each location (i.e., the BLUE computed in the first stage), μ is the common mean, E_t is the main effect of the t^{th}environments such that $E_t N(0, \sigma_E^2)$, $g_a = \{g_{aj}\}N(0, G_a\sigma_A^2)$; $g_r = \{g_{rj}\}N(0, G_r\sigma_R^2)$; with $G_a = Z_a G_a Z_a'$, $G_a = \frac{X_a X_a'}{p}$, $G_r = Z_r G_r Z_r'$, $G_r = \frac{X_r X_r'}{p}$, Z_a, Z_r are the corresponding incidence matrices that connect phenotypes with A-lines and R-lines ($g_{a_j}$and g_{r_j}), σ_A^2, σ_R^2, are the corresponding variance components of the parental effects, incidence matrices that connect phenotypes with A-lines and R-lines, and $e_{ht}N(0, \sigma_e^2)$ with σ_e^2 being the variance component associated with the residuals.

Model M2: General Plus Specific Combining Ability Model

This model is an extension of M1 that includes the Specific Combining Ability (SCA) interaction effect for a specific A-line and an R-line for each hybrid (Acosta-Pech et al., 2017). The SCA ($g_{h_{ij}}$) was modeled with the cell-by-cell product between the covariance structures derived from the A-line and R-line such that $g_{h_{ij}} = \left\{g_{a_j x r_i}\right\} N(0, G_h\sigma_h^2)$, where $G_h = G_a{}^{\circ}G_r$ and σ_H^2 is the corresponding variance component, and ° donates the cell by cell product between the A-line matrix and the R-line matrix. Combining the assumptions and terms, the linear predictor becomes:

$$\overline{y}_{ht} = \mu + E_t + g_{a_j} + g_{r_i} + g_{h_{ij}} + e_{ht}$$

Model M3: General Plus Specific Combining Ability including Interactions with Environments Model

This model allows specific genomic effects of the same genotype in different environments. The interactions between g_{a_j}, g_{r_i}, $g_{h_{ij}}$ terms and the environment is included via a reaction norm model (Jarquín et al., 2014a) as follows:

$$\overline{y}_{ht} = \mu + E_t + g_{a_j} + g_{r_i} + g_{h_{ij}} + gE_{a_jt} + gE_{r_it} + gE_{h_{ijt}} + e_{ht}$$

where $gE_a = \left\{gE_{ta_j}\right\} \sim N\left(0, (\boldsymbol{Z_a G_a Z'_a}) \circ (\boldsymbol{Z_e Z'_e})\sigma^2_{gE_a}\right)$, $gE_r = \left\{gE_{tr_j}\right\} \sim N\left(0, (\boldsymbol{Z_r G_r Z'_r}) \circ (\boldsymbol{Z_e Z'_e})\sigma^2_{gE_r}\right)$, $gE_h = \left\{gE_{th_{ij}}\right\} \sim N\left(0, (\boldsymbol{Z_a G_a Z'_a}) \circ (\boldsymbol{Z_r G_r Z'_r}) \circ (\boldsymbol{Z_e Z'_e})\sigma^2_{gE_h}\right)$, with $\sigma^2_{gE_a}$, $\sigma^2_{gE_r}$ and $\sigma^2_{gE_h}$ as the corresponding variance components of the interaction between the parental A-lines, R-lines and A×R-lines with the environments. The other components of the model are defined earlier.

Model M4: General Plus Specific Combining Ability in Interaction with Environmental Covariables Model

This model considers the interaction between g_{a_j}, g_{r_j}, $g_{h_{ij}}$and Environments Covariables (ECs) (Jarquín et al., 2014a; Basnet et al., 2019). The resulting linear predictor is as follows:

$$\overline{y}_{ht} = \mu + E_t + g_{a_j} + g_{r_j} + g_{h_{ij}} + gW_{ta_j} + gW_{tr_i} + gW_{th_{ij}} + e_{ht}$$

Where $gW_{a_j} \sim N\left(0, (\boldsymbol{Z_a G_a Z'_a}) \circ (\boldsymbol{Z_e \Omega Z'_e})\sigma^2_{gW_{taj}}\right)$, $gW_{r_i} \sim N\left(0, (\boldsymbol{Z_r G_r Z'_r}) \circ (\boldsymbol{Z_e \Omega Z'_e})\sigma^2_{gW_{trj}}\right)$, $gW_{h_{ij}} \sim N\left(0, (\boldsymbol{Z_a G_a Z'_a}) \circ (\boldsymbol{Z_r G_r Z'_r}) \circ (\boldsymbol{Z_e \Omega Z'_e})\sigma^2_{gW_{th_{ij}}}\right)$, with $\sigma^2_{gW_{taj}}$, $\sigma^2_{\mathrm{g}W_{trj}}$ and $\sigma^2_{gW_{thij}}$ as the corresponding variance components of the interaction terms, and $\Omega = \frac{WW'}{q}$ is the environmental relationship matrix whose entries describe the environmental similarities between pairs of environments, where *W* is a matrix of *q* environmental covariables.

Model M5: General Plus Specific Combining Ability with Interactions with Environments and Environmental Covariables Model

This model is the combination of the M3 and M4 models as described as follows:

$$\overline{y}_{ht} = \mu + E_t + g_{a_j} + g_{r_i} + g_{h_{ij}} + gE_{ta_j} + gE_{tr_i} + gE_{th_{ij}} + gW_{ta_j} + gW_{tr_i} + gW_{th_{ij}} + e_{ht}$$

where all the terms are as previously defined above.

Model M6: The linear predictor of this model is the same than M5; however, the environmental relationship matrix is computed differently. For each EC a linear regression on the environmental mean was implemented to find the most crucial window of time for explaining trait performance, adopted the idea by Li et al. (2018b). i.e., we checked different starting days (s) and different window sizes or days (l): $(1/l)\sum_{s}^{s+l}(Ecs)$. The correlation coefficients between the ECs and the FBY means environments were computed for each window sizes. Then the environmental relationship matrix was computed using the environmental information of each ECs corresponding to the window sizes that returned the highest coefficient of determination.

Model M7: In this model the average mean of the window size from the previous model was used for constructing the environmental relationship matrix).

Cross-Validation Scheme

A cross-validation (CV) study was conducted to assess the predictive ability of the different models. Four different CV schemes that have been detailed elsewhere (Jarquin et al., 2017) were used: CV1, CV2, CV0, and CV00. The objective of these CV schemes is to represent realistic scenarios that sorghum breeders face predicting hybrid performance.

The CV1 mimics the scenario of predicting a set of hybrids that have not been observed in any environment. The CV2 attempts to predict field trials (i.e., some hybrids have been evaluated in some environments but not tested in others). In both CV schemes (CV1 and CV2), a five-folds partition was implemented to generate the training and testing sets. The dataset was randomly divided into five subsets, with 80% of the hybrids assigned to the trained set and 20%

to the testing for CV1 while for CV2 phenotypes were assigned to training and testing sets. We repeated the procedure ten times, and the results were averaged and reported.

The CV0 scheme predicts the performance of hybrids in unobserved environments. The CV00 predicted untested hybrids in unobserved environments. In both cross-validation schemes, the number of folds correspond to the number of environments (i.e., 64). The procedure does not involve random partitioning and thus, it is implemented only once.

Assessing Predictive Ability

The predictive ability was assessed with the Person's correlations (*r*) between predicted and observed values within environments. The average predictive ability across environments was computed according to Tiezzi et al. (2017):

$$r_{\varphi} = \frac{\sum_{i=1}^{I} \frac{r_i}{V(r|i)}}{\sum_{i=1}^{I} \frac{1}{V(r|i)}}$$

where r_i is the Pearson's correlation between the predicted and observed values at the i^{th} environment, $V(r_i) = \frac{1-r_i^2}{n_i-2}$ is the sampling variance and n_i is the number of observations in the i^{th} environment.

The described models above (M1-M7) were fitted using the Bayesian Generalized Linear Regression (BGLR) R package (Perez-Rodriguez and de los Campos, 2014).

RESULTS

The phenotypic means for FBY across trials ranged from 27.95 t ha^{-1} (28 - Campus do Goytacazes year 2015) to 105.15 t ha^{-1}, (32 - Nova Porteirinha 2015) with an overall mean of 70.65 t ha^{-1} (Figure 2). Generalized heritability values ranged from 0.11 (56 – Sete Lagoas year

2018) to 0.97 (61 – Planaltina year 2019), with a mean of 0.82, which represents a high accuracy of experimental trials (Figure 3).

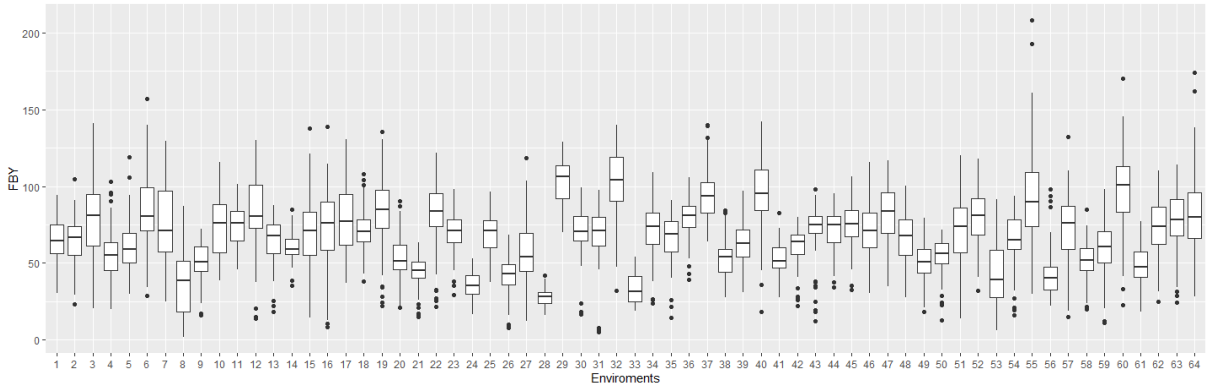

Figure 2. Boxplot of average mean of Fresh Biomass Yield (FBY, t ha^{-1}) of 64 Environments (Location x Year). Number of environments (1-64) defined in Table 1.

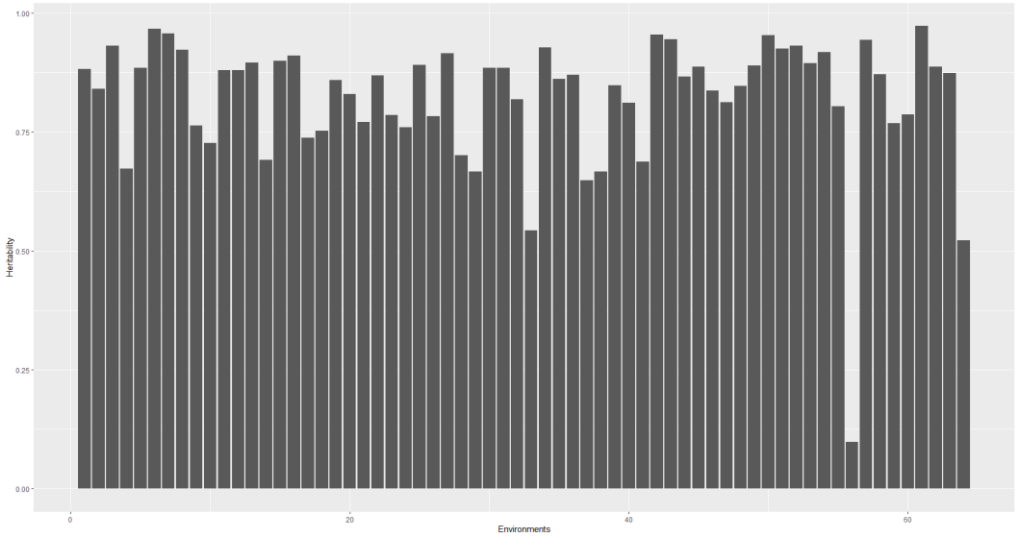

Figure 3. Heritability of 64 Environments (Location x Year). Number of environments (1-64) defined in Table 1.

Analysis of the Variance Components

Table 3 presents the variance components for the different models. The environmental term $(\sigma|E^2)$ captured the largest percentage of the phenotypic variability (46.87-50.76%) for all models. Also, theσ_r^2 corresponding to the main effect of the markers of the R-lines explains a higher percentage of variability compared to σ_a^2 (A-lines). When the SCA was included in models M2-M7, it explained a small percentage of the phenotypic variability (1.05 - 1.58%) (Table 3).

In M5, when all interactions were included the percentage of variability explained by the interactions $\sigma^2_{gE_a}$ and $\sigma^2_{gW_a}$ was similar (2.85%). The same occurred with the interactions $\sigma^2_{gE_r}$ and $\sigma^2_{gW_r}$ (3.58), $\sigma^2_{gE_h}$ and $\sigma^2_{gW_h}$ (2.45) (Table 3). However, when the reduced Ω matrix was considered, with M6 and M7, the amount of variability explained by the interactions involving weather data was reduced by approximately half.

Table 3) Percentage of the variability explained by models M1-M7 (and variance components).

Models	Main Effects				Interaction Effects						σ^2_e
	σ^2_E	σ^2_a	σ^2_r	σ^2_H	$\sigma^2_{Eg_a}$	$\sigma^2_{Eg_r}$	$\sigma^2_{Eg_h}$	$\sigma^2_{Wg_a}$	$\sigma^2_{Wg_r}$	$\sigma^2_{Wg_h}$	
	49.55	4.16	29.26								17.04
M1	(0.65)	(0.05)	(0.38)								(0.17)
	50.20	2.62	28.82	1.58							16.79
M2	(0.65)	(0.03)	(0.37)	(0.02)							(0.17)
	49.82	2.77	25.40	1.18	3.90	4.08	3.35				9.51
M3	(0.63)	(0.03)	(0.32)	(0.01)	(0.05)	(0.04)	(0.03)				(0.14)
	50.76	2.61	24.95	1.20				2.10	3.15	1.69	13.53
M4	(0.65)	(0.03)	(0.32)	(0.02)				(0.02)	(0.03)	(0.02)	(0.10)
	46.87	2.58	22.92	1.05	2.85	3.58	2.45	2.85	3.58	2.45	8.82
M5	(0.63)	(0.03)	(0.31)	(0.01)	(0.04)	(0.04)	(0.02)	(0.03)	(0.04)	(0.02)	(0.11)
	47.93	2.69	26.28	1.26	2.70	2.97	2.51	1.82	1.11	1.74	8.99
M6	(0.61)	(0.03)	(0.33)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.01)	(0.02)	(0.11)
	48.53	2.74	24.51	1.26	3.32	3.93	2.71	1.49	0.59	1.32	9.53
M7	(0.60)	(0.03)	(0.30)	(0.02)	(0.04)	(0.04)	(0.03)	(0.01)	(0.00)	(0.01)	(0.11)

M1: Model 1; M2: Model 2; M3: Model 3; M4: Model 4; M5: Model 5; M6: Model 6; and M7: Model 7. σ^2_E**:** represents the effects of environments. σ^2_a and σ^2_r: main effects of inbred markers accounting for A-line and R-line (GCA); σ^2_H: is the interaction between inbred markers for A-line and R-line (SCA); $\sigma^2_{Eg_a}$ and $\sigma^2_{Eg_r}$: are interactions between inbreed markers and environments; $\sigma^2_{Eg_h}$: is the interaction between SCA effects and environments; $\sigma^2_{Wg_a}$ and $\sigma^2_{Wg_r}$: are interactions between inbreed markers and Environmental Covariates (ECs); $\sigma^2_{Wg_h}$: is the interactions between SCA effects and ECs;σ^2_e: residual effects

Comparison of Models Under Different Scenarios

CV1

Table 4 presents the average correlation across environments for all cross-validation schemes. The objective of CV1 is to predict new or untested hybrids in already observed environments (CV1). This is possible because these genotypes are genetically related to other hybrids observed in the same or in other environments. The average correlation across trials for the reference model M1 was 0.46. The inclusion of the SCA component (M2) gave similar results to the baseline model. However, when the interactions were included (M3 - M7) the predicted increased and showed results 0.55, 0.51, 0.55, 0.56, and 0.56, respectively. Among

the models including interactions M4 was the worse. Nevertheless, this model was superior to the baseline model (Table 4).

Table 4) The average correlations across 64 environments for the seven models (M1-M7) for the four cross-validation schemes (CV1, CV2, CV0, CV00).

Model	CV1	CV2	CV0	CV00
M1	0.46	0.51	0.52	0.39
M2	0.47	0.52	0.53	0.39
M3	0.55	0.59	0.53	0.36
M4	0.51	0.55	0.48	0.34
M5	0.55	0.60	0.52	0.32
M6	**0.56**	**0.61**	0.55	0.38
M7	**0.56**	**0.61**	**0.56**	**0.40**

M1: Model 1; M2: Model 2; M3: Model 3; M4: Model 4; M5: Model 5; M6: Model 6; and M7: Model 7, CV1; CV2; CV0; CV00: cross-validation scheme described above. The highest average correlation is in bold**.**

CV2

When attempting to predict incomplete field trials (i.e., some hybrids have been evaluated in some environments but not in others), M1 presented a weighted average correlation of 0.51. The inclusion of the specific combining ability of the parents (M2) enhanced predictive ability to 0.52. With the inclusion of interactions (M3 – M5) with the environments and/or ECs, the correlation between predicted and observed values was substantially improved to 0.59, 0.55, and 0.60, respectively. Also, when the modified kinship matrix of the environment was introduced with M6 and M7, the correlation between predicted and observed values was slightly improved to 0.61 in both models (Table 4). The combination of interactions (M5) gave similar results to the interaction with environments only (M3). As in the previous CV scheme, M4 returned worse results than those from M3. The M6 and M7 had similar results to M3 and M5, but M7 had a smaller mean square error (Table 5). However, compared to the baseline model (M1) the models M6 and M7 outperformed in both cases.

CV0

CV0 considers the predictions of hybrids in novel unobserved environments but already tested in other environments. The weighted average predictive ability of the M1 was 0.52. The inclusion of the SCA term (M2) and the interactions between the GCA and SCA with environments (M3) and M5 have similar results to the baseline model, the M4 that included interaction between GCA and SCA with ECs slightly decreased the predictive ability in comparison with M1. However, with the models that consider the modified environmental kinship (M6 and M7) the predictive ability was improved to 0.55 and 0.56 (Table 4). Perhaps, the data reduction of the environmental data in the Ω matrix reduced the noise slightly increasing the signal for explaining trait variability.

Table 5) Mean Square Error (MSE) across 64 environments for the seven models.

	MSE			
Model	CV1	CV2	CV0	CV00
M1	126.26	120.48	447.79	455.84
M2	126.73	118.85	451.13	452.14
M3	117.35	108.91	**434.46**	468.73
M4	120.75	114.17	46575	485.27
M5	117.64	109.30	448.32	464.16
M6	117.75	109.49	460.80	451.30
M7	**116.43**	**108.11**	452.42	**447.23**

M1: Model 1; M2: Model 2; M3: Model 3; M4: Model 4; M5: Model 5; M6: Model 6; and M7: Model 7, CV1; CV2; CV0; CV00: cross-validation scheme described above. The smallest Mean Square Error is in bold**.**

CV00

The predictions of untested hybrids in unobserved environments are the most interesting prediction scenario of all breeding programs. Here, the baseline model (M1) and the model that includes the SCA term returned a correlation of 0.39. The inclusion of the interaction terms (M3 – M5) returned worst results than the baseline model (M1). Thus, modeling the interaction between ECs and genotypes did not improve predictability. However, M7 (0.40) had similar

results to M1 and M2. However, M7 showed smaller mean square error compared with all other models (Table 5). In addition, M7 showed high correlations in the other three cross-validation schemes (CV1, CV2, and CV0) and the smallest mean square errors in CV1 and CV2.

DISCUSSION

In this study, we showed that genomic prediction models could be improved by considering the interactions of GCA and SCA components with environmental factors, and by identifying a common window of days in the growing season (i.e., reduce the dimension of the Ω matrix). Our results suggest a large genetic variation for GCA of R-lines in all models. Similar results were reported in diallelic models for biomass sorghum discussed the importance of this effects (Oliveira et al., 2019) (Table 4). For this study we used information on 225 hybrids that were developed from 46 A-lines and 25 R-lines. Principal components analysis of the genomic relationship matrix shows that the first two principal components explained 53.36% of the genetic variance for the parental lines with a larger variance of the R-line group.

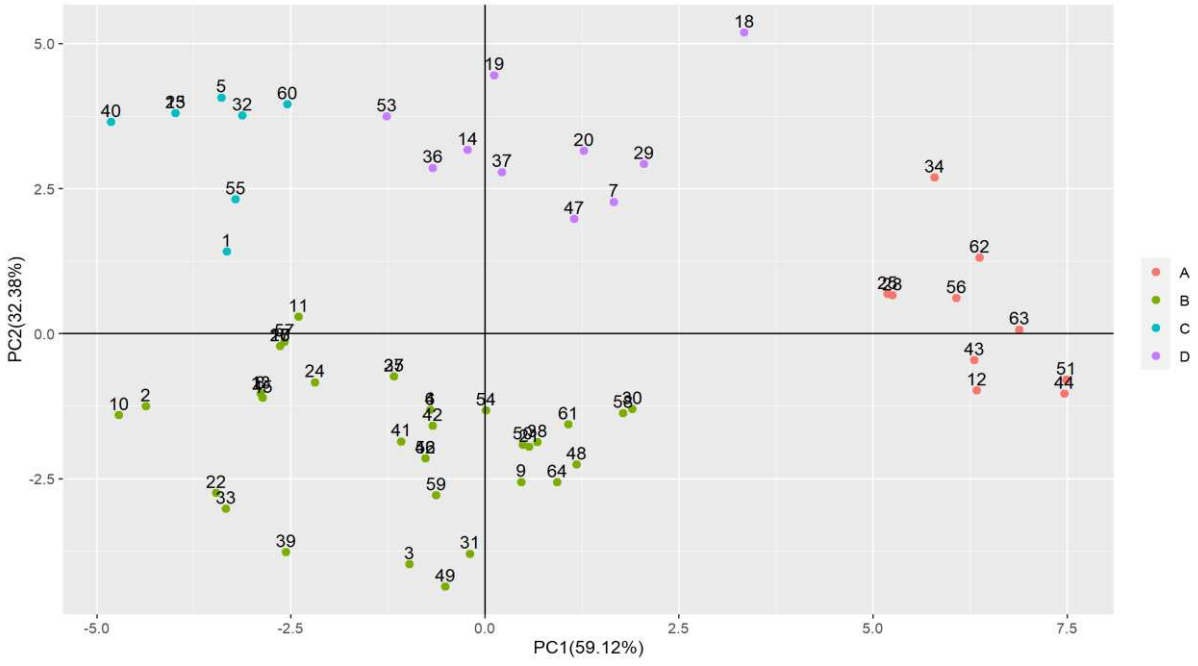

Figure 3. PC1 vs PC2 plot. The first two principal components of the Ω matrix (only considering environmental covariables), accounted for nearly 91.50% of the variation. Number of environments (1-64) defined in Table 1. A: mega-environment 1; B: mega-environment 2; C: mega-environment 3; D: mega-environment 4.

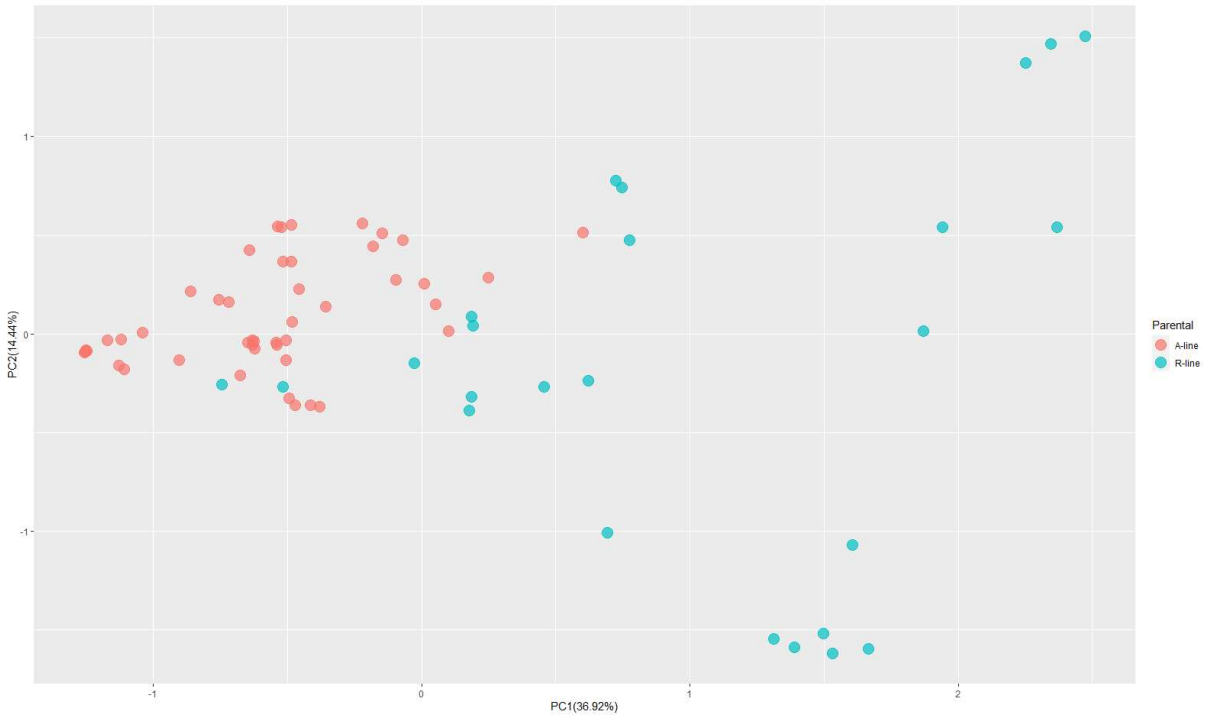

Figure 4) The first two principal components of the kinship matrix of parental lines, accounted for nearly 53.36% of the genetic variation.

With M3, we observed improvements in the predictive ability when the interaction terms were added to the model. The models with environmental interactions outperformed the baseline model for CV1, CV2, and CV0. These results varied somewhat depending on the cross-validation scheme. However, only in the CV00, the M3 had the worst performance compared to the baseline model. These results highlighted that the inclusion of the interaction component was beneficial for improving predictive ability for GS models. Similar results were observed in maize (*Zea mays L.*) (Dias et al., 2018; Jarquin et al., 2021), pearl millet (*Pennisetum glaucum*) (Jarquin et al., 2020), and wheat (*Triticum aestivum*) (Lopez-Cruz et al., 2015).

The results of M4 were in line with some results in the literature (Jarquín et al., 2014) where the models that included environmental covariables had null improvement in predictive ability through modeling the ECs compared to the baseline model. There are three possible reasons why M4 did not outperform M1 in all cases. First, the number of ECs included in the model did not sufficiently explain similarities among pairs of environments concerning sorghum biomass growth and development. Second, as we observed the environmental kinship, 36 of 64 environments were in the same mega-environment. Third, the amount of genetic

diversity (R-lines) may have played an important role, as was the case in this study. Although, this study had a greater number of environments compared with those from Jarquín et al., (2014) no significant improvements were observed when the number of environments was increased.

The part of the data used in this analysis was collected from the trials of VCU. The goal of this trial was to test hybrids under different environmental conditions. Nevertheless, this type of dataset produces unbalanced genotypes; this is a perfect dataset for testing models in different cross-validation schemes mimicking realistic scenarios of interest in breeding programs. Therefore, models that account simultaneously for genomic relationships between genotypes, environments, and ECs are expected to improve the predictive ability of tested and untested genotypes in observed and unobserved environments (Jarquin et al., 2021). Some studies have discussed strategies that involve the inclusion of ECs in genomic prediction models (Pérez-Rodríguez et al. 2015; Millet et al. 2019; Basnet et al. 2019, Costa-Neto et al 2021). The reaction norm model described by Jarquín et al. (2014a) shows that the incorporation of these effects increase prediction accuracy. We used this concept in models M4-M7. However, incorporating high dimensional genetic and environmental data is not an easy task (Pérez-Rodríguez et al., 2015). Model M5 outperformed the M1-M4 in CV1 and CV2, showing some advantages of using ECs and molecular markers in predicting tested genotypes in observed environments, and in predicting newly developed genotypes in observed environments.

Models M6 and M7 increased predictive ability in CV1, CV2, CV0. Likewise, model M7 increased predictive ability in CV00. Moreover, M7 showed smaller MSE in CV1, CV2, and CV00. This model works with specific windows of time for each EC during the growing season. For this dataset, we found that reducing the dimensionality of the environmental data is one good way to consider working with ECs. Considering a specific window of time of the ECs we were trying to work with the time interval that most affect the prediction of sorghum biomass, similarly as how it was proposed by Li et al. (2018a) for grain sorghum and one

environmental covariable (photothermal time). However, here we were working with 23 ECs, which made it difficult to account for the window because each ECs have different windows of time (Table S2). These results suggest that we should study the information of specific windows of time that most affect the prediction of sorghum biomass, and hence can increase the predictability. Therefore, new studies with specific time intervals for each ECs should be considered.

Another use of this model allows us the construction of mega-environments. We selected M7, because this model shows a smaller percentage of unexplained variance, i.e., residual variance, high prediction accuracy in all scenarios, and small mean square error in CV1, CV2, and CV00 (Table 3, 4, and 5). In this present study, we defined four mega-environments. The identification of the mega environments under tropical conditions can help the breeding programs to work with the genotype by environmental interaction (Yan et al., 2000; Gauch et al., 2008; Malosetti et al., 2013; Van Eeuwijk et al., 2016; González Barrios et al., 2019).

CONCLUSIONS

This study showed the advantages of including the GCA and SCA models and their interactions with environments and environmental covariables when predicting untested biomass sorghum hybrids in tested/untested environments for fresh biomass yield. Here, we modeled different kinship matrix of the ECs and showed some improvement of predictive ability. The identification of mega-environments for energy sorghum cultivation in Brazil based on climatic data of the evaluation sites of Embrapa maize and sorghum experiments was also accomplished.

REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. EEE Trans. Autom.

Control 19(6): 716–723.

Allard, R.W., and A.D. Bradshaw. 1964. Implications of Genotype-Environmental Interactions in Applied Plant Breeding 1. Crop Sci. 4(5): 503–508.

Barbosa, M.H.P., M.D.V. de Resende, J.A. Bressiani, L.C.I. da Silveira, and L.A. Peternelli. 2005. Selection of sugarcane families and parents by Reml/Blup Crop Breeding and Applied Brazilian Society of Plant Breeding. Printed in Brazil Selection of sugarcane families and parents by Reml/Blup.

Buntaran, H., H. Piepho, J. Hagman, and J. Forkman. 2019. A Cross-Validation of Statistical Models for Zoned-Based Prediction in Cultivar Testing. Crop Sci. 59(4): 1544–1553. doi: 10.2135/cropsci2018.10.0642.

Buntaran, H., H. Piepho, P. Schmidt, J. Rydén, M. Halling, et al. 2020a. Cross-validation of stage-wise mixed-model analysis of Swedish variety trials with winter wheat and spring barley. Crop Sci.: csc2.20177. doi: 10.1002/csc2.20177.

Buntaran, H., H. Piepho, P. Schmidt, J. Rydén, M. Halling, et al. 2020b. Cross-validation of stagewise mixed-model analysis of Swedish variety trials with winter wheat and spring barley. Crop Sci.: csc2.20177. doi: 10.1002/csc2.20177.

Butler, D.G., B.R. Cullis, A.R. Gilmour, B.J. Gogel, and R. Thompson. 2018. ASReml-R Reference Manual Version 4 ASReml estimates variance components under a general linear mixed model by residual maximum likelihood (REML).

Chavarría-Perez, L.M., W. Giordani, K.O.G. Dias, Z.P. Costa, C.A.M. Ribeiro, et al. 2020. Improving yield and fruit quality traits in sweet passion fruit: Evidence for genotype by environment interaction and selection of promising genotypes (P.E. Teodoro, editor). PLoS One 15(5): e0232818. doi: 10.1371/journal.pone.0232818.

Cullis, B.R., A.B. Smith, and N.E. Coombes. 2006. On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11(4): 381–393. doi: 10.1198/108571106X154443.

Damesa, T.M., J. Möhring, M. Worku, and H.-P. Piepho. 2017. One Step at a Time: Stage-Wise Analysis of a Series of Experiments. Agron. J. 109(3): 845–857. doi: 10.2134/agronj2016.07.0395.

Dias, K.O.G., H.P. Piepho, L.J.M. Guimarães, P.E.O. Guimarães, S.N. Parentoni, et al. 2020. Novel strategies for genomic prediction of untested single-cross maize hybrids using unbalanced historical data. Theor. Appl. Genet. 133(2): 443–455. doi: 10.1007/s00122-019-03475-1.

Duarte, J.B., and R. Vencovsky. 2001. Estimação e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. Sci. Agric. 58(1): 109–117. doi: 10.1590/S0103-90162001000100017.

Eculica, G.C., P.C.D.O. Ribeiro, A.B. De Oliveira, N.N.L.D. Parrella, P.S.D.S. Leite, et al. 2019. Adaptability and stability of saccharine sorghum cultivars. African J. Agric. Res. 14(31): 1432–1442. doi: 10.5897/AJAR2019.14043.

Van Eeuwijk, F.A., D. V Bustos-Korts, and M. Malosetti. 2016. What should students in plant breeding know about the statistical aspects of genotype$\times$ environment interactions? Crop Sci. 56(5): 2119–2140.

de Figueiredo, U.J., J.A.R. Nunes, R.A.C. Parrella, E.D. Souza, A.R. da Silva, et al. 2015. Adaptability and stability of genotypes of sweet sorghum by GGEBiplot and Toler methods. Genet. Mol. Res. 14(3). doi: 10.4238/2015.September.22.15.

Gauch, H.G., and R.W. Zobel. 1997. Identifying Mega-Environments and Targeting

Genotypes. Crop Sci. 37(2): 311–326. doi: 10.2135/cropsci1997.0011183X003700020002x.

Gogel, B., A. Smith, and B. Cullis. 2018. Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. Euphytica 214(2). doi: 10.1007/s10681-018-2116-4.

González Barrios, P., L. Díaz-García, and L. Gutiérrez. 2019. Mega-Environmental Design: using genotype by environment interaction to optimize resources for cultivar testing. Crop Sci. doi: 10.2135/cropsci2018.11.0692.

Hartigan, J.A., and M.A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm.

Malosetti, M., J.M. Ribaut, and F.A. van Eeuwijk. 2013. The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. Front. Physiol. 4 MAR(March): 1–17. doi: 10.3389/fphys.2013.00044.

Mirzawan., P.D.N., M. Cooper, I.H. DeLacy, and D.M. Hogarth. 1994. Retrospective analysis of the relationships among the test environments of the Southern Queensland sugarcane breeding programme. Theor. Appl. Genet. 88: 707–716.

Oliveira, I.C.M., J.H.S. Guilhen, P.C. de O. Ribeiro, S.A. Gezan, R.E. Schaffert, et al. 2020. Genotype-by-environment interaction and yield stability analysis of biomass sorghum hybrids using factor analytic models and environmental covariates. F. Crop. Res. 257(August). doi: 10.1016/j.fcr.2020.107929.

Piepho, H.P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theor. Appl. Genet. 97(1–2): 195–201. doi: 10.1007/s001220050885.

Piepho, H.P., J. Möhring, A.E. Melchinger, and A. Büchse. 2008. BLUP for phenotypic

selection in plant breeding and variety testing. Euphytica 161(1–2): 209–228. doi: 10.1007/s10681-007-9449-8.

Piepho, H.P., J. Möhring, T. Schulz-Streeck, and J.O. Ogutu. 2012. A stage-wise approach for the analysis of multi-environment trials. Biometrical J. 54(6): 844–860. doi: 10.1002/bimj.201100219.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

da Silva, M.J., M.M. Pastina, V.F. de Souza, R.E. Schaffert, P.C.S. Carneiro, et al. 2017. Phenotypic and molecular characterization of sweet sorghum accessions for bioenergy production (J.-M. Lacape, editor). PLoS One 12(8): e0183504. doi: 10.1371/journal.pone.0183504.

Slater, A.T., G.M. Wilson, N.O.I. Cogan, J.W. Forster, and B.J. Hayes. 2014. Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. Theor. Appl. Genet. 127(4): 809–820. doi: 10.1007/s00122-013-2258-7.

Smith, A., B. Cullis, and A. Gilmour. 2001. The analysis of crop variety evaluation data in Australia. Aust. New Zeal. J. Stat. 43(2): 129–145.

Smith, A.B., A. Ganesalingam, H. Kuchel, and B.R. Cullis. 2015a. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. Theor. Appl. Genet. 128(1): 55–72. doi: 10.1007/s00122-014-2412-x.

Smith, A.B., A. Ganesalingam, H. Kuchel, and B.R. Cullis. 2015b. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. Theor. Appl. Genet. 128(1). doi: 10.1007/s00122-014-2412-x.

Yan, W., L.A. Hunt, Q. Sheng, and Z. Szlavnics. 2000. Cultivar evaluation and mega-

environment investigation based on the GGE biplot. Crop Sci. 40(3): 597–605. doi: 10.2135/cropsci2000.403597x.

SUPPLEMENTARY MATERIAL

Table S1: Cycle days of the field trail

Trail	Site	Stage*	YEAR	Planting	Harvest	Cycle (days)
1	Nova Porteirinha	F	2011	11/7/2011	5/3/2012	178
2	Pelotas	F	2011	11/23/2011	5/21/2012	180
3	Sete Lagoas	F	2011	12/4/2011	5/19/2012	167
4	Sete Lagoas	P	2012	12/4/2012	5/22/2013	169
5	Nova Porteirinha	F	2012	12/7/2012	5/16/2013	160
6	Sete Lagoas	F	2012	12/4/2012	5/22/2013	169
7	Santa Vitoria	F	2012	11/20/2012	5/21/2013	182
8	Sete Lagoas	P	2013	12/5/2013	5/21/2014	167
9	Goiania	F	2013	11/21/2013	5/22/2014	182
10	Lavras	F	2013	11/29/2013	5/20/2014	172
11	Nova Porteirinha	F	2013	11/20/2013	5/22/2014	183
12	Sinop	F	2013	11/19/2013	5/26/2014	188
13	Sete Lagoas	F	2013	12/5/2013	5/21/2014	167
14	Santa Vitoria	F	2013	11/20/2013	5/22/2014	183
15	Nova Porteirinha	P	2014	11/22/2014	5/20/2015	179
16	Sete Lagoas	P	2014	12/4/2014	5/11/2015	158
17	Sete Lagoas	P	2014	12/4/2014	5/11/2015	158
18	Caceres	F	2014	12/5/2014	6/19/2015	196
19	Dourados	F	2014	11/6/2014	5/13/2015	188
20	Dracena	F	2014	12/18/2014	5/20/2015	153
21	Goiania	F	2014	11/22/2014	5/5/2015	164
22	Lavras	F	2014	11/15/2014	4/29/2015	165
23	Nova Porteirinha	F	2014	11/22/2014	5/20/2015	179
24	Pelotas	F	2014	12/6/2014	3/25/2015	109
25	Sinop	F	2014	11/6/2014	4/10/2015	155
26	Sete Lagoas	F	2014	12/4/2014	5/11/2015	158
27	Sete Lagoas	P	2015	12/16/2015	5/30/2016	166
28	Campos do Goytacazes	F	2016	3/1/2016	7/14/2016	135
29	Dourados	F	2015	12/9/2015	4/18/2016	131
30	Goiania	F	2015	11/26/2015	6/10/2016	197
31	Lavras	F	2015	11/10/2015	5/2/2016	174
32	Nova Porteirinha	F	2015	10/27/2015	5/21/2016	207
33	Pelotas	F	2015	11/6/2015	5/5/2016	181
34	Sinop	F	2015	11/26/2015	5/16/2016	172
35	Sete Lagoas	F	2015	12/16/2015	5/30/2016	166
36	Campos do Goytacazes	F	2016	11/30/2016	5/25/2017	176
37	Dourados	F	2016	11/14/2016	4/20/2017	157
38	Goiania	F	2016	12/1/2016	4/26/2017	146
39	Lavras	F	2016	11/25/2016	5/30/2017	186
40	Nova Porteirinha	F	2016	11/22/2016	5/10/2017	169
41	Pelotas	F	2016	12/9/2016	4/12/2017	124
42	Planatina	F	2016	11/9/2016	5/16/2017	188
43	Sinop	F	2016	11/23/2016	5/2/2017	160

44	Vilhela	F	2016	12/9/2016	4/26/2017	138
45	Sete Lagoas	F	2016	11/22/2016	5/11/2017	170
46	Sete Lagoas	P	2017	10/26/2017	5/4/2018	190
47	Campos do Goytacazes	F	2017	11/28/2017	5/11/2018	164
48	Goiania	F	2017	12/21/2017	6/13/2018	174
49	Lavras	F	2017	12/5/2017	5/17/2018	163
50	Planatina	F	2017	11/21/2017	5/23/2018	183
51	Sinop	F	2017	11/18/2017	5/1/2018	164
52	Sete Lagoas	F	2017	10/26/2017	5/4/2018	190
53	Campos do Goytacazes	F	2018	11/1/2018	6/18/2019	229
54	Goiania	F	2018	11/29/2018	5/8/2019	160
55	Nova Porteirinha	F	2018	10/22/2018	5/27/2019	217
56	Sete Lagoas	F	2018	12/19/2018	5/16/2019	148
57	Sinop	F	2018	11/16/2018	4/24/2019	159
58	Goiania	F	2019	12/10/2019	5/1/2020	143
59	Jaguariuna	F	2019	11/26/2019	6/8/2020	195
60	Nova Porteirinha	F	2019	11/7/2019	5/10/2020	185
61	Planatina	F	2019	11/13/2019	4/7/2020	146
62	Sinop	F	2019	11/20/2019	5/6/2020	168
63	Vilhela	F	2019	11/18/2019	4/28/2020	162
64	Sete Lagoas	F	2019	12/9/2019	5/11/2020	154

*F: Final, P: preliminary

Table S2: The Window of each ECs, with the day Start and End, and Correlation coefficients (R^2)

	Environmental Covariable	Day Start	Day end	R^2
4	Mean air temperature at 2 m above the surface of the earth	60	62	0.17
5	Maximum air temperature at 2 m above the surface of the earth	3	114	0.19
6	Minimum air temperature at 2 m above the surface of the earth	62	63	0.09
7	Rainfall precipitation	4	121	0.11
8	Wind speed at 10 m above the surface of the earth	61	64	0.13
9	Relative air humidity	5	120	0.13
10	The dew point	7	118	0.04
11	Downward Thermal Infrared Radiative Flux	12	86	0.21
12	Insolation Incident on a Horizontal Surface	1	124	0.11
13	Top-of-atmosphere Insolation	19	106	0.2
14	Daylight's hours	13	94	0.15
15	Actual duration of sunshine	6	118	0.11
16	Extraterrestrial radiation	5	120	0.14
17	Solar Radiation	62	63	0.18
18	Vapor pressure deficit	3	116	0.09
19	Slope of saturation vapor pressure curve	30	95	0.09
20	Potential Evapotranspiration	30	95	0.09
21	Growing Degree Day	18	106	0.13
22	Effect of temperature on radiation use efficiency	62	63	0.18

23	Daily Temperature Range	2	116	0.12

Table S3: The number of times each location appearance in each Mega-environment.

Location	MG1	MG2	MG3	MG4	N
Caceres	-	-	-	1	1
Campo Grande	1	-	-	3	4
Dourados	-	-	-	3	3
Dracena	-	-	-	1	1
Goiania	-	7	-	-	7
Jaguariuna	-	1	-	-	1
Lavras	-		-	-	5
Nova Porteirinha	-	1	8	-	9
Pelotas	-	4	-	-	4
Planaltina	-	3	-	-	3
Santa Vitoria	-	-	-	2	2
Sete Lagoas	-	15	-	-	15
Sinop	7	-	-	-	7
Vilhela	2	-	-	-	2
Total	10	36	8	10	64

M1: mega-enviromental 1, M2: mega-enviromental 2; M3: mega-enviromental 3; M4: mega-enviromental 4; N: total environment (combination Year x Location).

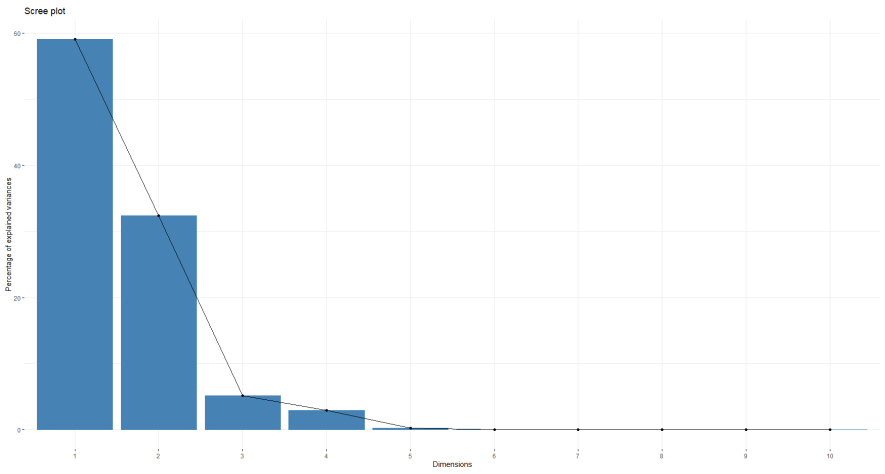

Figure S1) The number of clusters was defined from the plot of the percentage of explained variances.

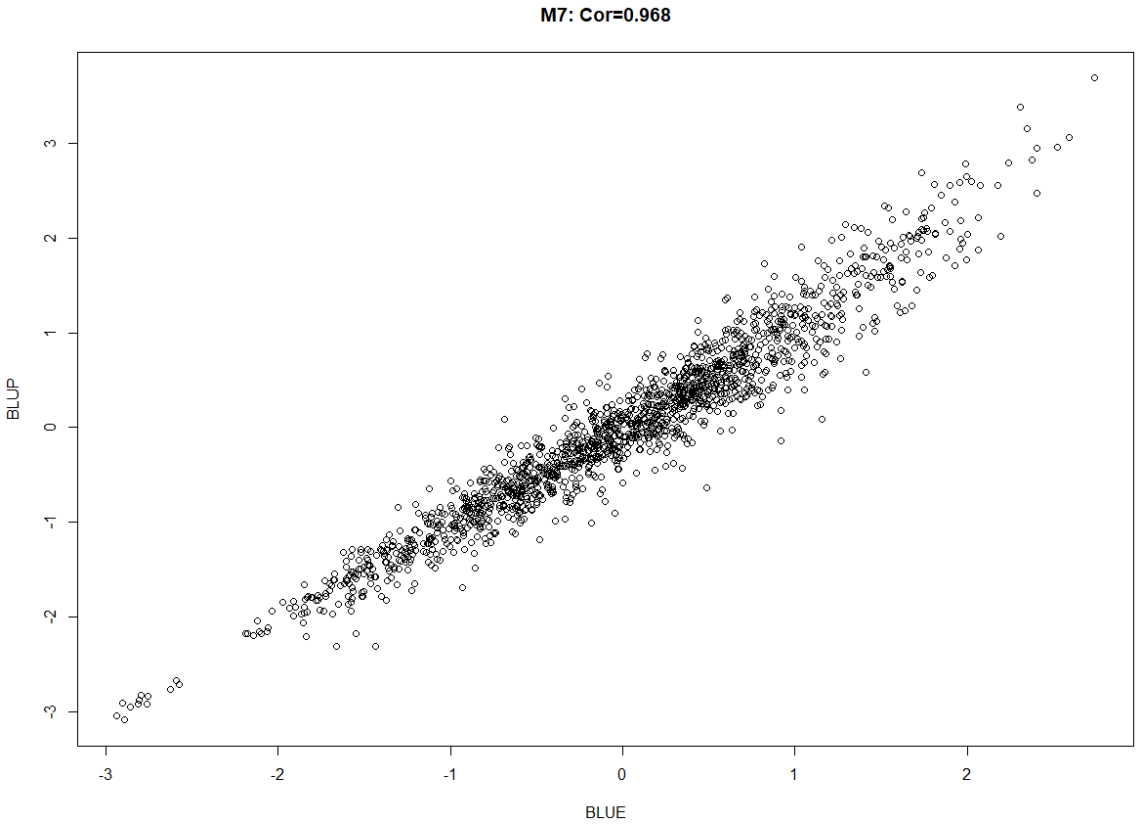

Figure S2) Correlation coefficient between the BLUE of the first stage and the BLUP of the second stage, for the M7