



**FRANCIELLY DE CÁSSIA PEREIRA**

**CHALLENGES IN PHENOTYPIC DATA ANALYSIS AND GENOMIC  
SELECTION APPLICATION FOR MAIZE HYBRID PREDICTION IN  
TROPICAL REGIONS**

**LAVRAS – MG**

**2021**

**FRANCIELLY DE CÁSSIA PEREIRA**

**CHALLENGES IN PHENOTYPIC DATA ANALYSIS AND GENOMIC SELECTION  
APPLICATION FOR MAIZE HYBRID PREDICTION IN TROPICAL REGIONS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para obtenção do título de Doutor.

**Prof. Dr. Renzo Garcia Von Pinho**

**Orientador**

**LAVRAS – MG**

**2021**

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca  
Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Pereira, Francielly de Cássia.

Challenges in phenotypic data analysis and genomic selection  
application for maize hybrid prediction in tropical regions /

Francielly de Cássia Pereira. - 21.

84 p. : il.

Orientador(a): Renzo Garcia Von Pinho.

Tese (doutorado) - Universidade Federal de Lavras, 21.

Bibliografia.

1. Phenotypic data analysis. 2. Genomic selection. 3. Zea mays.  
I. Von Pinho, Renzo Garcia. II. Título.

**FRANCIELLY DE CÁSSIA PEREIRA**

**CHALLENGES IN PHENOTYPIC DATA ANALYSIS AND GENOMIC SELECTION  
APPLICATION FOR MAIZE HYBRID PREDICTION IN TROPICAL REGIONS**

**DESAFIOS NA ANÁLISE FENOTÍPICA DE DADOS E APLICAÇÃO DA SELEÇÃO  
GENÔMICA PARA PREDIÇÃO DE HÍBRIDOS DE MILHO EM REGIÕES  
TROPICAIS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração Genética e Melhoramento de Plantas para obtenção do título de Doutor.

APROVADA em 29/01/2021

Dr. Renzo Garcia Von Pinho	UFLA
Dr. Marcio Fernando Ribeiro de Resende Junior	UF-Florida
Dr. Adriano Teodoro Bruzi	UFLA
Dra. Magno Antonio Patto Ramalho	UFLA
Dr. José Maria Villela Pádua	UFLA



Prof. Dr. Renzo Garcia Von Pinho

Orientador

**LAVRAS – MG  
2020**

*A Deus, por ser minha fortaleza nos momentos difíceis,*

OFEREÇO

*Aos meus pais, José Heitor e Clara, a minha irmã Priscilla e ao meu noivo Ricardo, pelos conselhos,  
incentivo e paciência durante essa caminhada,*

DEDICO

## AGRADECIMENTOS

À Deus por ser a base de tudo em minha vida, por não permitir que eu desistisse nos momentos difíceis e por manter minha fé inabalável quando eu precisei.

Aos meus pais, José Heitor e Clara, que dedicaram suas vidas para que eu concluísse meus estudos, por todas as privações, por todo suporte, incentivo e carinho para que eu chegasse até aqui, serei eternamente grata a vocês.

À minha irmã, Priscilla que desde o início dos nossos estudos sempre foi um exemplo para mim, pelas palavras de incentivo e pelos conselhos durante essa jornada.

Ao meu noivo Ricardo, por todo amor dedicado, pela compreensão nos momentos que estive ausente, pela paciência nos momentos difíceis, por me incentivar e sempre enxergar o lado bom da vida, pelo companheirismo nesses anos, por acreditar em mim e me mostrar que eu poderia alcançar mais do que eu imaginava, enfim por ser essa pessoa maravilhosa, meu porto seguro, sem você eu não teria conseguido.

À Universidade Federal de Lavras e ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, pela oportunidade, suporte e estrutura para que eu obtivesse o título de doutora.

Ao professor Renzo pela orientação, paciência e conselhos durante esse tempo. A todos os professores do departamento de biologia que dividiram seus conhecimentos e contribuíram para minha formação, em especial ao professor Magno, que é um exemplo de profissional para mim, por toda ajuda, conselhos, disposição e entusiasmo em nossas conversas. Ao professor Marcio Balestre (in memoriam) pela coorientação. A todos os funcionários do departamento de biologia que de alguma forma me auxiliaram durante esse período.

Aos membros da banca por aceitarem o convite e pelas valiosas contribuições.

Aos amigos, que tornaram essa caminhada mais tranquila e prazerosa.

A Universidade da Florida onde eu realizei parte do meu doutorado, em especial ao Professor Márcio Resende, pela oportunidade de estar em seu laboratório, pela orientação durante o período que estive lá, por ser um exemplo de pesquisador. Obrigada por acreditar em mim e me conceder a experiência para expandir meus conhecimentos em genética e melhoramento. Agradeço a todo time do seu laboratório pelo suporte e ótimo convívio durante esse tempo.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de financiamento 001 e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq / MCTIC pela concessão da bolsa de estudos.

A todos que contribuíram direta ou indiretamente para a conclusão desse trabalho e para o meu crescimento profissional nesses quatro anos, meus sinceros agradecimentos.

## RESUMO GERAL

O processo de obtenção de híbridos simples (HS) de milho é muito dinâmico. Nas fases iniciais, inúmeras combinações genotípicas são geradas. À medida que as etapas avançam, muitos HS vão sendo descartados ao passo que o número de repetições e locais de avaliação aumentam. Assim, é comum que os dados oriundos de programas de melhoramento de milho sejam muito desbalanceados, devido à baixa coincidência dos HS que são avaliados ao longo dos diferentes locais e anos. Nos últimos anos, a seleção genômica tem sido proposta como uma ferramenta para acelerar os ganhos genéticos e reduzir os custos na seleção e recomendação de HS de milho. Contudo, resta a dúvida se os dados obtidos desses experimentos altamente desbalanceados, podem contribuir para aumentar a acurácia dos modelos preditivos. Assim, o objetivo desse estudo foi, em uma primeira etapa, analisar criticamente os dados de produtividade de grãos de 2770 híbridos de milho, oriundos da avaliação de experimentos conduzidos em diferentes anos e épocas de semeadura e verificar os impactos das condições de desbalanceamento na estimativa de parâmetros genéticos e fenotípicos. Adicionalmente em um segundo momento, utilizando o mesmo conjunto de dados, foi comparada a capacidade preditiva do modelo GBLUP (Genomic Best Linear Unbiased Prediction) considerando efeitos aditivos e de dominância. Para isso, as 447 linhagens parentais foram genotipadas utilizando 23.153 marcadores Darts. Os resultados evidenciam o grande desafio de analisar as informações de todas as safras sob condições de alto desbalanceamento e significativo efeito da interação genótipo por ambientes. Esses fatores comprometem as estimativas dos componentes de variância, da herdabilidade e dos valores genéticos dos indivíduos e conseqüentemente podem afetar as acurácias preditivas dos modelos de seleção genômica. As análises envolvendo informações genômicas demonstraram que é possível obter ganhos genéticos com a predição de HS não avaliados e que a inclusão dos efeitos de dominância no modelo GBLUP melhora sua capacidade preditiva.

Palavras – Chaves: Melhoramento de Plantas; Ganho genético; Desbalanceamento; Seleção Genômica; Parâmetros genéticos e fenotípicos.



## ABSTRACT

The process to obtain maize single-cross hybrids (SH) is very dynamic. In the initial stages, numerous genotypic combinations are produced. As the process progresses, many SH are discarded, and the number of repetitions and sites that these SH are evaluated increases. Thus, it is common that data from maize breeding programs are very unbalanced due to the low coincidence of SH that are evaluated over different sites and years. In the recent years, genomic selection has been proposed as a tool to accelerate the genetic gains and to reduce the cost with the selection and the recommendation of maize SH. However, a pending question is if the data obtained from these highly unbalanced experiments, can contribute to increase the accuracy of predictive models. Therefore, the purpose of the present study, in a first moment, was to critically analyze the grain yield data of 2770 maize hybrids, from experiments conducted in different years and sowing seasons; and to verify the impacts of unbalanced designs to the estimates of genetic and phenotypic parameters. Additionally, in a second step, using the same data set, the predictive capacity of the GBLUP (Genomic Best Linear Unbiased Prediction) model was compared considering additive and dominance effects. For this, the 447 parental lines were genotyped using 23,153 Darts markers. The results show the complexities of analyzing information from all crop seasons under conditions of high experimental unbalance and significant effect of genotype by environment interaction. These factors compromise the estimates of variance components, heritability, genetic values of individuals and, consequently, may affect the predictive accuracy of genomic selection models. Nonetheless, the analysis involving genomic information showed that it is possible to obtain genetic gains with the prediction of SH not evaluated and that the inclusion of dominance effects in the GBLUP model can improve its predictive ability.

**Keywords:** Plant breeding; Genetic gain; Unbalanced; Genomic Selection; Genetic and Phenotypic Parameters.

## LISTA DE FIGURAS

Segunda Parte.....	15
ARTIGO 1 - MEGA-ENVIRONMENT ANALYSIS OF MAIZE BREEDING DATA FROM BRAZIL .....	15
Figure 1: Map of Brazil, showing the locations where experiments for evaluation of single cross hybrids of maize were conducted in the central and southern regions. Points in black, red, orange, blue, and pink correspond to experiments set up in the first crop season; points in green, purple, brown and yellow correspond to experiments set up in the second crop season. Crop seasons are labeled by their year of sowing (2011 - 2014), followed by their crop season (first crop, s1, or second crop, s2) and region (West Center, C, or South, S). .....	19
Figure 2: The diagonal of the heat plot corresponds to the number of hybrids evaluated in each crop season. Upper diagonal heat plot indicates the number of genotypes in common that were evaluated in the pairs of environments while lower diagonal indicates the estimates of the correlation between the mean values of the SH in common in the crop seasons two by two. Crop seasons are labeled by their year of sowing (2011 - 2014), followed by their crop season (first crop, s1, or second crop, s2) and region (West Center, C, or South, S). .....	26
Figure 3: A) Boxplot representing the variation between the upper and lower EBLUP limit of the predictions of the single cross hybrids evaluated in each crop season and possible outliers. B) Heritability estimates for grain yield over nine crop seasons using three different methods (standard, Cullis, and Holland-Piepho).....	28
ARTIGO 2 – EMPREGO DE EFEITOS ADITIVOS E DE DOMINÂNCIA PARA PREDIÇÃO DE HÍBRIDOS DE MILHO EM MEGA AMBIENTES DE CULTIVO .....	45
Figura 1: Análise de componentes principais dos 2770 híbridos simples de milho avaliados utilizando as matrizes de parentesco aditivo (A) e de dominância (B). Heatmap das matrizes de parentesco aditivo (C) e de dominância (D). Cada célula representa a relação entre os pares de híbridos. ....	<b>Error! Bookmark not defined.</b>
Figura 2: Boxplot das acurácias de predição dentro de cada safra para os modelos aditivo (A) e aditivo-dominante (AD). .....	59

Figura 3: Ganho realizado com a seleção (%), obtidos dentro de cada safra e entre safras, para os modelos **GBLUP\_A** e **GBLUP\_AD** (em negrito) e ganho realizado real com a seleção (%), obtidos entre as safras, a partir das estimativas **BLUP** dos HS avaliados em cada safra (em itálico), considerando a mesma época de semeadura e diferentes épocas de semeadura. .... 64

## LISTA DE TABELAS

Segunda Parte.....	15
ARTIGO 1 - MEGA-ENVIRONMENT ANALYSIS OF MAIZE BREEDING DATA FROM BRAZIL .....	15
Table 1: Description of the crop season, year, abbreviation, experimental design, number of single cross hybrids (SH), locations, experiments (EXP) and replications (REP) in each crop season.....	20
Table 2: Overall mean, lower limit (LL), upper limit (UL) of mean grain yield of single cross maize hybrids in the experiments and environments of each crop season in t.ha <sup>-1</sup> and correlation between the mean value and the EBLUP of each hybrid in the nine crop seasons evaluated (r). .....	27
ARTIGO 2 – EMPREGO DE EFEITOS ADITIVOS E DE DOMINÂNCIA PARA PREDIÇÃO DE HÍBRIDOS DE MILHO EM MEGA AMBIENTES DE CULTIVO .....	45
Figura 1: Análise de componentes principais dos 2770 híbridos simples de milho avaliados utilizando as matrizes de parentesco aditivo (A) e de dominância (B). Heatmap das matrizes de parentesco aditivo (C) e de dominância (D). Cada célula representa a relação entre os pares de híbridos. ....	<b>Error! Bookmark not defined.</b>
Figura 2: Boxplot das acurácias de predição dentro de cada safra para os modelos aditivo (A) e aditivo-dominante (AD). .....	59
Figura 3: Ganho realizado com a seleção (%), obtidos dentro de cada safra e entre safras, para os modelos GBLUP_A e GBLUP_AD (em negrito) e ganho realizado real com a seleção (%), obtidos entre as safras, a partir das estimativas BLUP dos HS avaliados em cada safra (em itálico), considerando a mesma época de semeadura e diferentes épocas de semeadura. ....	64
Tabela 1: Estimativas dos parâmetros genéticos obtidos via regressão linear bayesiana generalizada, para as nove safras avaliadas, incluindo efeitos aditivos (GBLUP_A) e aditivo-dominante (GBLUP_AD). .....	58
Tabela 2: Correlação entre os valores preditos e observados em diferentes anos, obtidos considerando os modelos GBLUP_A e GBLUP_AD (em negrito), para as primeiras e	

segundas safras da região centro, considerando as estratégias de validação E1, E2 e E3 e correlação fenotípica entre as médias dos HS coincidentes nas safras i e j (em itálico). 60

Tabela 3: Correlação entre os valores preditos e observados em diferentes anos, obtidos considerando os modelos **GBLUP\_A** e **GBLUP\_AD** (em negrito), para as primeiras e segundas safras e entre as primeiras e segundas safras da região sul, considerando as estratégias de validação E1, E2 e E3 e correlação fenotípica entre as médias dos HS coincidentes nas safras i e j (em itálico). ..... 61

Tabela 4: Correlação entre os valores preditos e observados em diferentes anos, obtidos considerando os modelos **GBLUP\_A** e **GBLUP\_AD** (em negrito), entre as primeiras e segundas safras da região centro, considerando as estratégias de validação E1, E2 e E3 e correlação fenotípica entre as médias dos HS coincidentes nas safras i e j (em itálico). 62

## SUMARIO

PRIMEIRA PARTE.....	9
INTRODUÇÃO GERAL.....	9
REFERÊNCIAS.....	12
SEGUNDA PARTE - ARTIGOS .....	15
ARTIGO 1 - MEGA-ENVIRONMENT ANALYSIS OF MAIZE BREEDING DATA FROM BRAZIL.....	15
1 INTRODUCTION .....	17
2 MATERIALS AND METHODS.....	18
2.1 Genetic material, experimental design, and environments .....	18
2.2 Statistical analyses.....	20
3 RESULTS .....	25
4 DISCUSSION.....	31
5 REFERENCES .....	38
ARTIGO 2 – EMPREGO DE EFEITOS ADITIVOS E DE DOMINÂNCIA PARA PREDIÇÃO DE HÍBRIDOS DE MILHO EM MEGA AMBIENTES DE CULTIVO.....	45
1 INTRODUÇÃO .....	47
2 MATERIAL E MÉTODOS .....	49
2.1 Dados fenotípicos.....	49
2.2 Dados genotípicos .....	52
2.3 Predição dos valores genéticos genômicos .....	52
2.4 Validação dentro e entre safras .....	55
2.5 Medidas para verificar a eficiência dos modelos GBLUP .....	56
3 RESULTADOS.....	57

3.1	Parâmetros genéticos e dados genômicos .....	57
3.2	Predição genômica .....	59
4	DISCUSSÃO .....	65
5	REFERÊNCIAS.....	71
	MATERIAL SUPLEMENTAR.....	76

## **PRIMEIRA PARTE**

### **INTRODUÇÃO GERAL**

O milho (*Zea mays*) é considerado a cultura com maior importância econômica no mundo em termos de nutrição, fonte de fibras, combustível e alimentação animal. Atualmente apresenta produção mundial de 1,14 bilhões de toneladas, sendo os Estados Unidos e China os maiores produtores desse grão (USDA, 2020). O Brasil é considerado o segundo maior exportador e o terceiro maior produtor mundial de milho. Nos anos agrícolas de 2019/2020, foram cultivados 17,5 milhões de hectares com milho (*Zea mays*), sendo a segunda safra responsável por mais de 70% da área plantada no país (CONAB, 2020).

Ao longo dos anos a produção de milho sofreu um grande incremento devido a várias razões como o uso de fertilizantes, o aumento da densidade de plantas, a aplicação de herbicidas e o desenvolvimento de novos equipamentos agrícolas para a condução das lavouras. Somado a esses fatores, o melhoramento genético também teve papel fundamental para alcançar os números atuais, pois possibilitou a seleção de genótipos adaptados a condições ambientais específicas (CROW, 1998).

Atualmente, as empresas produtoras de sementes híbridas de milho geram milhares de combinações genotípicas a cada ano. Para que a seleção e recomendação desses híbridos para os diferentes ambientes seja feita de forma segura, é necessário que elas sejam testadas em centenas de repetições (TROYER, 1996; GAFFNEY et al., 2015). Devido a quantidade de combinações híbridas que são produzidas, dificilmente um mesmo híbrido será avaliado em todos os ambientes (locais, anos e safras), o que deve prejudicar a decisão no momento da recomendação. Muitas vezes a eliminação pode ter sido precipitada, pois o determinado híbrido em teste não teve a chance de mostrar o seu real potencial. Adicionalmente, há um excessivo desbalanceamento dos experimentos que prejudica as análises e conseqüentemente a recomendação desses híbridos.

Sabe-se que as condições ambientais que envolvem a primeira e a segunda safra no Brasil, são bem distintas em relação a temperatura e a distribuição de chuvas, caracterizando os climas



tropicais e subtropicais. O desafio para recomendar híbridos de milho para as condições brasileiras é ainda maior que em regiões temperadas, pois além das diferentes condições climáticas, é necessário levar em consideração os diferentes biomas (pampas, mata atlântica e cerrado) que o milho é cultivado em toda extensão territorial.

Diversas ferramentas têm sido utilizadas para auxiliar o melhorista na identificação e seleção dos melhores híbridos, dentre elas destaca-se a seleção genômica ampla que foi proposta por Meuwissen et al. (2001) a partir da viabilidade de mapeamento genético dos cromossomos de muitas espécies. A seleção genômica é uma forma de seleção assistida por marcadores que ao invés de empregar apenas marcadores que possuem uma correlação significativa predefinida com uma característica, utiliza-se informações de marcadores cobrindo todo o genoma para calcular o mérito genético de cada genótipo, ou seja, os valores genéticos genômicos estimados (GEBV) (MEUWISSEN et al., 2001).

Enquanto os métodos tradicionais utilizam o fenótipo para inferir sobre o genótipo, a seleção genômica aplica o efeito genético pré-estimado em uma amostra da população para inferir sobre o fenótipo que será expresso nos candidatos a seleção (RESENDE et al., 2012). O processo central dessa ferramenta é baseado na estimativa dos GEBVs para indivíduos com apenas dados genotípicos, utilizando um modelo que foi treinado em uma população com dados fenotípicos e genotípicos (MEUWISSEN et al., 2001).

O método de seleção genômica ampla permaneceu discreto por alguns anos devido ao fato dos marcadores moleculares disponíveis no passado serem caros e restritos. No entanto, com o desenvolvimento e baixo custo dos marcadores do tipo SNP (single nucleotide polymorphism), a técnica se tornou atrativa para ser empregada em programas de melhoramento genético públicos e privados (MEUWISSEN, 2007; HEFFNER; SORRELLS; JANNINK, 2009; DESTA; ORTIZ, 2014; LADO et al, 2016; DIAS et al., 2018; AMES; BERNADO, 2020).

Relatos na literatura demonstram o sucesso e as vantagens de se empregar a GWS no melhoramento de milho, para predizer os valores genotípicos de híbridos não avaliados, possibilitando a redução de custos e o incremento no ganho com a seleção (BERNARDO; YU, 2007; MASSMAN; JUNG; BERNARDO, 2013; DOS SANTOS et al., 2016; CANTELMO et al. 2017; LYRA et al. 2017; DIAS et al. 2018a; FRITSCHÉ-NETO et al. 2018; DE OLIVEIRA et al., 2020; KRAUSE et al., 2020). Contudo, é importante destacar que a seleção genômica não elimina a fenotipagem, uma vez que para treinar os modelos de predição, dados fenotípicos acurados são necessários (HEFFNER; SORRELLS; JANNINK, 2009).

Para que a seleção genômica possa ser incorporada na rotina dos programas de melhoramento, além da obtenção de dados fenotípicos acurados, alguns outros fatores têm sido destacados na literatura como fundamentais para melhorar a predição de genótipos nos diferentes ambientes de cultivo, dentre eles destacam-se a interação genótipos por ambientes e os efeitos não aditivos (DESTA; ORTIZ, 2014; CUERVAS et al., 2016; LADO et al., 2016; DOS SANTOS et al., 2016; FERRÃO et al., 2018, DIAS et al., 2018a; DIAS et al., 2018b; DE OLIVEIRA et al., 2020; KRAUSE et al., 2020). Assim, o objetivo desse trabalho foi realizar uma análise crítica de experimentos de produtividade de híbridos de milho e verificar a influência dos efeitos de dominância no modelo de seleção genômica, sob um alto nível de desbalanceamento e em diferentes anos e épocas de semeadura.

## REFERÊNCIAS

- AMES Nicholas C.; BERNARDO, Rex. 2020. Genomewide predictions as a substitute for a portion of phenotyping in maize. **Crop Science**, Saint Paul, p. 1-9, 2020.
- BERNARDO Rex; YU J. Prospects for Genomewide Selection for Quantitative Traits in Maize. **Crop Science**, v.47, p.1082-1090, 2007.
- CANTELMO N. F.; VON PINHO R. G.; BALESTRE M. Genome-wide prediction for maize single-cross hybrids using the GBLUP model and validation in different crop seasons. **Molecular Breeding**, v.37, p.1-13, 2017.
- CONAB. Companhia Nacional de Abastecimento Acompanhamento da safra brasileira de grãos. **Terceiro levantamento, Brasília**, Disponível em <http://www.conab.gov.br>. Acesso em: 27 de dez.e 2020.
- CROW, James F. 90 years ago: the beginning of hybrid maize. **Genetics**, v.148, p.923-928, 1998.
- CUEVAS J. et al. Genomic prediction of genotype X environment interaction kernel regression models, **Plant genome**, v.9, p. 1-20, 2016.
- DESTA, Zeratsion Abera; ORTIZ, Rodomiro. Genomic selection: genome-wide prediction in plant improvement. **Trends in plant science**, v.19, p.592-601, 2014.
- DE OLIVEIRA, A.A. et al. Genomic prediction applied to multiple traits and environments in second season maize hybrids. **Heredity**, v.125, p.60–72 , 2020.
- DIAS KODG et al. Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. **Heredity** v.121, p.24–37, 2018.
- DIAS, KODG et al. Estimating genotype × environment interaction for and genetic correlations among drought tolerance traits in maize via factor analytic multiplicative mixed models. **Crop Science**, v. 58(1), p.72, 2018.

DOS SANTOS JPR et al. Inclusion of dominance effects in the multivariate GBLUP model. **PLoS ONE**, v.11, p.1–21, 2016

FERRÃO, L.F.V. et al. Integration of dominance and marker X environment Interactions into maize genomic prediction models. **bioRxiv**, 2018.

FRITSCHÉ-NETO R; AKDEMIR D; JANNINK JL Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. **Theoretical Applied Genetics**, v.131, p.1153–1162, 2018.

GAFFNEY, J. et al. Industry-Scale Evaluation of Maize Hybrids Selected for Increased Yield in Drought-Stress Conditions of the US Corn Belt. **Crop Science**, v.55(4), p. 1608-1618, 2015.

HEFFNER E. L.; SORRELLS M. E.; JANNINK J. L. Genomic selection for crop improvement. **Crop Science**, v.49(1), p.1-12, 2009.

KRAUSE M. D. et al. Boosting predictive ability of tropical maize hybrids via genotype by environment interaction under multivariate GBLUP models. **Crop Science**, v.60, p.3049-3065, 2020.

LADO, Bettina et al. Modeling genotype X environment interaction for genomic selection with unbalanced data from a wheat breeding program. **Crop Science**, v.56, p.2165-2179, 2016.

LYRA DH et al. Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. **Molecular Breeding**, v.37, p.1–14, 2017.

MASSMAN J. M.; JUNG H. G.; BERNARDO Rex Genomewide Selection versus Marker-assisted Recurrent Selection to Improve Grain Yield and Stover-quality Traits for Cellulosic Ethanol in Maize. **Crop Science**, v.53, p.58-66, 2013.

MEUWISSEN T. H. E.; HAYES B. J.; GODDARD. M. E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**, v.157, p.1819-1829, 2001.

MEUWISSEN, T. Genomic selection: marker assisted selection on a genome wide scale. **Journal of animal Breeding and genetics**, v.124, p.321-322, 2007.

RESENDE M. D. V. et al. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada e Estatística Espacial**. Viçosa, MG: Ed. UFV, 2012.

TROYER, A. F. Breeding widely adapted, popular maize hybrids. **Euphytica**, v.92, p.163–174, 1996

USDA World Agricultural Production. **Circular Series**. 2020. Disponível em: <https://apps.fas.usda.gov/psdonline/circulars/production.pdf>. Acesso em 27 de dez. 2020.

**SEGUNDA PARTE - ARTIGOS****ARTIGO 1 - MEGA-ENVIRONMENT ANALYSIS OF MAIZE BREEDING DATA  
FROM BRAZIL**

Artigo redigido conforme as normas NBR 6022 (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2003).

## **MEGA-ENVIRONMENT ANALYSIS OF MAIZE BREEDING DATA FROM BRAZIL**

### **ABSTRACT**

The development and recommendation of single cross maize hybrids (SH) to be used in extensive land areas (mega-environments), and in different crop seasons requires many experiments under numerous environmental conditions. The question we asked is if the data from these multi-environment experiments are sufficient to identify the best hybrid combinations. The aim of this study was to critically analyze the phenotype data of experiments of yield, established by a large seed producing company, under a high level of unbalance data. Data from evaluation of 2770 SH were used from experiments conducted over four years, involving the first and second crop seasons, in 50 locations of different years and regions of Brazil. Different types of analysis were carried out and genetic and non-genetic components were estimated, with emphasis on the different interactions of the SH with the environments. Results showed that the coincidence of common hybrids in these experiments is normally small. The estimates of the correlations between the hybrids coinciding in each pair of environments is of low magnitude. The hybrid  $\times$  crop season interaction was always expressive; however, the interactions of hybrids and other environmental variables were also important. Under these conditions, alternatives were discussed for making with the information obtained from the experiments, can be more efficient in the process to obtain new hybrids by breeding programas.

**KEYWORDS:** Genotype x Environment Interaction; Unbalance Data; Hybrid Recommendation Process; Variance Components; Plant Breeding.

## 1 INTRODUCTION

Two maize crop seasons are common in Brazil per year. The first crop occurs from September to December, while the second season is from January to April. The environmental conditions between these two crop seasons are quite distinct in relation to temperature and rain distribution. In addition, farmers' use of technology in maize growing is quite diversified. This makes selection of hybrids for recommendation under these different conditions a much greater challenge than that in temperate regions, for example.

In order for a breeder to be successful in the identification of hybrids adapted to the mega-environment of maize growing, the hybrids obtained annually must be broadly evaluated. Clearly, these evaluations will only be successful if the experiments are conducted in the greatest number of environments possible. Experience in respect shows that secure recommendation was only possible through middle-term results coming from hundreds of replications (TROYER, 1996; GAFFNEY et al., 2015). However, companies obtain thousands of hybrids annually, which makes testing in multiple replications difficult. In this way, the same hybrid will rarely be evaluated in all the environments, resulting in excessive unbalance data and consequently hindering the decision making at the time of recommendation.

In many situations, these experiments are used to evaluate the possibility of employing genomic selection in prediction of potentially superior hybrid combinations. In this sense, the more accurate the model is, the greater the association of future performance of the hybrid will be through the response of the genotyped line. Previous experiences show that the effect of the hybrid  $\times$  environment interaction greatly complicates the prediction process. Because of this interaction, the responses of the hybrids do not coincide in the diverse environments evaluated. An alternative is to include this effect in the predictive models to obtain more accurate information (LADO et al.,



2016; FERRÃO et al., 2020, DIAS et al., 2018, MONTESINOS-LÓPEZ et al., 2019; KRAUSE et al., 2020). The question is whether the hybrid  $\times$  environment interaction component obtained from highly unbalanced experiments can contribute to the predictive models.

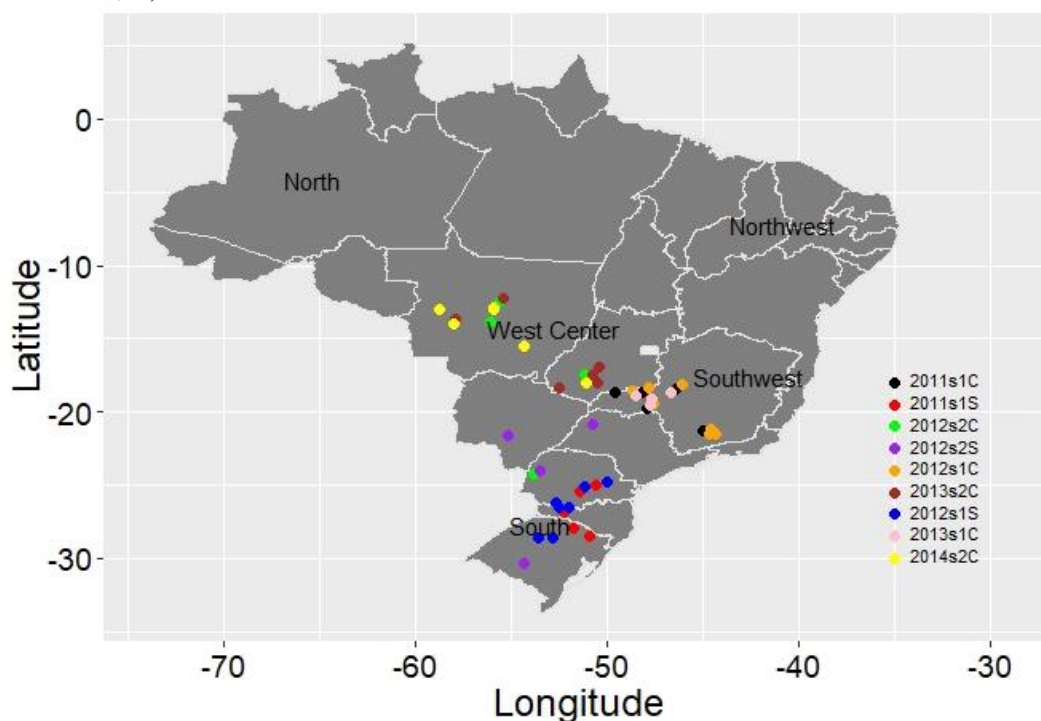
Thus, the purpose of the present study was to analyze the phenotypic data from yield experiments of different crop years and seasons, estimate genetic and phenotypic parameters under a high level of unbalance and comment the impact of these conditions on breeder decisions to selection maize hybrids.

## **2 MATERIALS AND METHODS**

### **2.1 Genetic material, experimental design, and environments**

Grain yield data ( $\text{t}\cdot\text{ha}^{-1}$ ), kindly provided by a Brazilian seed company, were used in this study. These data were obtained over 4 years with two crop seasons per year including numerous locations in the central and southern regions of Brazil (Figure 1). During this period, 2770 SH of maize were evaluated. These hybrids originated from crosses of 447 lines coming from different tropical, subtropical, and temperate regions around the world. Due to the breeding program in question being a line introgression program, the number of SH common to the two crop seasons, as well as the number of experiments and of treatments evaluated per experiment in each location, was quite variable (Table 1 and Figure 2).

Figure 1: Map of Brazil, showing the locations where experiments for evaluation of single cross hybrids of maize were conducted in the central and southern regions. Points in black, red, orange, blue, and pink correspond to experiments set up in the first crop season; points in green, purple, brown and yellow correspond to experiments set up in the second crop season. Crop seasons are labeled by their year of sowing (2011 - 2014), followed by their crop season (first crop, s1, or second crop, s2) and region (West Center, C, or South, S).



Source: Author (2021).

Randomized block (RBD) and incomplete block (IBD) experimental designs were used for evaluation of the hybrids, with two or three replications. The plots consisted of four 5 meters rows with a 0.7 meters between-row spacing. Different experiments were set up within each crop season in the same location. The experiments within each location were connected through check varieties in common since the hybrids evaluated in each experiment were not necessarily the same. Additional information regarding the number of hybrids, the locations, experiments, replications, and experimental design adopted in each crop season is provided in Table 1. Each crop season was

identified by an abbreviation that corresponds to the year of sowing (2011, 2012, 2013, or 2014), followed by the crop season (first crop, s1, or second crop, s2) and region (West Center, C, or South, S) (Figure 1 and Table 1).

Table 1: Description of the crop season, year, abbreviation, experimental design, number of single cross hybrids (SH), locations, experiments (EXP) and replications (REP) in each crop season.

Crop season	Year	Abbreviation	Design	SH	Location	EXP	REP
West Center 1st crop	2011/2012	2011s1C	RBD	1012	5	28	3
South 1st crop	2011/2012	2011s1S	RBD	783	5	22	3
West Center 2nd crop	2012	2012s2C	RBD	1041	6	24	3
South 2nd crop	2012	2012s2S	RBD	1040	4	22	3
West Center 1st crop	2012/2013	2012s1C	IBD	441	7	13	3
South 1st crop	2012/2013	2012s1S	IBD	725	7	19	2
West Center 2nd crop	2013	2013s2C	IBD	411	6	11	3
West Center 1st crop	2013/2014	2013s1C	IBD	810	4	20	2
West Center 2nd crop	2014	2014s2C	IBD	853	6	21	2

Source: Author (2021).

## 2.2 Statistical analyses

For better characterization of the dataset, the overall mean per crop season and the variation among mean values of the SH in the different experiments and, subsequently, in the locations were estimated. Considering only the data from the 2011s1C crop season for the purpose of making inferences regarding what occurs among locations in the same crop season, genetic variance ( $\sigma_g^2$ ) among the hybrids evaluated, variance of the error ( $\sigma_e^2$ ), and heritability ( $h^2$ ) were estimated in each location using the following statistical model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_{ge}\mathbf{u}_{ge} + \mathbf{Z}_b\mathbf{u}_b + \mathbf{e} \quad (1)$$

where  $\mathbf{y}$  is the vector of phenotypic observations;  $\boldsymbol{\tau}$  is the vector of fixed effect of the experiment;  $\mathbf{u}_g$  is the vector of random genotypic effects of hybrids, with  $\mathbf{u}_g \sim N(0, \sigma_g^2 \mathbf{I}_g)$ ;  $\mathbf{u}_{ge}$  is the vector of random effects of the hybrid by experiment interaction, with  $\mathbf{u}_{ge} \sim N(0, \sigma_{ge}^2 \mathbf{I}_{ge})$ ;  $\mathbf{u}_b$  is the vector of random effect of replication within experiments, with  $\mathbf{u}_b \sim N(0, \sigma_b^2 \mathbf{I}_b)$ ;  $\mathbf{e}$  is the vector of random errors, with  $\mathbf{e} \sim N(0, \sigma_e^2 \mathbf{I}_n)$ ;  $\mathbf{X}$ ,  $\mathbf{Z}_g$ ,  $\mathbf{Z}_{ge}$ , and  $\mathbf{Z}_b$  are the incidence matrices associated with the vectors  $\boldsymbol{\tau}$ ,  $\mathbf{u}_g$ ,  $\mathbf{u}_{ge}$ , and  $\mathbf{u}_b$ ;  $\sigma_g^2$ ,  $\sigma_{ge}^2$ ,  $\sigma_b^2$ , and  $\sigma_e^2$  are the variance components associated with the vectors  $\mathbf{u}_g$ ,  $\mathbf{u}_{ge}$ ,  $\mathbf{u}_b$ , and  $\mathbf{e}$ ; and  $\mathbf{I}_g$ ,  $\mathbf{I}_{ge}$ ,  $\mathbf{I}_b$ , and  $\mathbf{I}_n$  are the identity matrices associated with the vectors  $\mathbf{u}_g$ ,  $\mathbf{u}_{ge}$ ,  $\mathbf{u}_b$ , and  $\mathbf{e}$ .

In each crop season, the grain yield data were analyzed through a mixed models approach considering the model according to the experimental design adopted:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_{gl}\mathbf{u}_{gl} + \mathbf{Z}_b\mathbf{u}_b + \mathbf{e} \quad (2)$$

where  $\mathbf{y}$  is the vector of phenotypic observations;  $\boldsymbol{\tau}$  is the vector of fixed effects (experimental RBD: location and experiment within location; experimental IBD: location, experiment within location, and replication within experiment and location);  $\mathbf{u}_g$  is the vector of random genotypic effects of hybrids, with  $\mathbf{u}_g \sim N(0, \sigma_g^2 \mathbf{I}_g)$ ;  $\mathbf{u}_{gl}$  is the vector of random effects of the hybrid by location interaction, with  $\mathbf{u}_{gl} \sim N(0, \sigma_{gl}^2 \mathbf{I}_{gl})$ ;  $\mathbf{u}_b$  is the vector of random effects (experimental RBD: block within experiment and location; experimental IBD: block within replication, experiment, and location, with  $\mathbf{u}_b \sim N(0, \sigma_b^2 \mathbf{I}_b)$ ;  $\mathbf{e}$  is the vector of random errors, with  $\mathbf{e} \sim N(0, \mathbf{R})$ ;  $\mathbf{X}$ ,  $\mathbf{Z}_g$ ,  $\mathbf{Z}_{gl}$ , and  $\mathbf{Z}_b$  are the incidence matrices associated with the vectors  $\boldsymbol{\tau}$ ,  $\mathbf{u}_g$ ,  $\mathbf{u}_{gl}$ , and  $\mathbf{u}_b$ ;  $\sigma_g^2$ ,  $\sigma_{gl}^2$ ,  $\sigma_b^2$ , and  $\sigma_e^2$  are the variance components associated with the vectors  $\mathbf{u}_g$ ,  $\mathbf{u}_{gl}$ ,  $\mathbf{u}_b$ , and  $\mathbf{e}$ ; and  $\mathbf{I}_g$ ,  $\mathbf{I}_{gl}$ ,  $\mathbf{I}_b$ , and  $\mathbf{I}_n$  are the identity matrices associated with the vectors  $\mathbf{u}_g$ ,  $\mathbf{u}_{gl}$ ,  $\mathbf{u}_b$ , and  $\mathbf{e}$ . The residual (co)variance

matrix, with the aim of modeling the effect of location within each crop season, adopted a diagonal block variance structure, using the identity matrix ( $\mathbf{R} = \bigoplus_{l=1}^t \sigma_{e_l}^2 \mathbf{I}_{n_l}$ ).

Previously, alternative methodologies of unstructured variance covariance matrix were tested to try to model the genetic correlation between all environment pairs. These matrices allow a better understanding of the genetic structure and evaluate the stability of genotypes in mega-environments. To this end, the genetic and residual effects were considered as  $\mathbf{u}_g \sim MVN(0, \sigma_g^2 \mathbf{I} \otimes \Sigma_l)$  e  $\mathbf{e} \sim MVN(0, \sigma_e^2 \mathbf{I} \otimes \mathbf{R}_l)$ .  $\Sigma_l$  is the VCOV matrix for the additive genetic effects in the  $l$  environments and  $\mathbf{R}_l$  represents the VCOV matrix for the residual effects in the  $l$  environments. In this case, the main environment effects were implicitly modeled and an unstructured form for the genetic  $\Sigma_l$  and residual  $\mathbf{R}_l$  VCOV matrix was assumed. Because of the large number of sites evaluated in each season ( $>5$ ) the convergence of these unstructured matrices was difficult. In this way, the diagonal block variance structure was adopted, as described above, to model the genetic and residual effects in this study.

The variance components associated with the random effects were obtained using the residual maximum likelihood method (REML) (PATTERSON; THOMPSON, 1971) and their significance levels were verified by the likelihood ratio test. To make inferences regarding the occurrence of interaction, the correlation ( $r_{qs}$ ) among the mean values of the SH coinciding in the  $q$  and  $s$  crop seasons was estimated. The estimator used was similar to that presented by Steel, Torrie and Dickey (1997):

$$r_{qs} = \frac{cov(\overline{HS}_{iq} \times \overline{HS}_{is})}{\sqrt{\sigma_{HS_{iq}}^2 \times \sigma_{HS_{is}}^2}} \quad (3)$$

where  $\overline{HS}_{iq}$  is the mean of the single cross hybrid  $i$  in crop season  $q$ ;  $\overline{HS}_{is}$  is the mean of the single cross hybrid  $i$  in crop season  $s$ ;  $\sigma_{HS_{iq}}^2$  is the variance of the single cross hybrid  $i$  in crop season  $q$ ; and  $\sigma_{HS_{is}}^2$  is the variance of the single cross hybrid  $i$  in crop season  $s$ .

In addition, the effect of the hybrid  $\times$  crop season interaction was also verified through the coincidence of the genotypes selected based on the mean of two environments (considering different combinations of year, region, and sowing time) in relation to selection based on the mean of each environment individually. For that purpose, the maize yield data from each location within the crop seasons were fitted regarding the effect of blocking and replication, according to the design adopted in each situation, to obtain the EBLUE (Empirical Best Linear Unbiased Estimation) estimates. These estimates were used to carry out the individual analyses of each crop season and also the combined analyses of the environments two by two. Using the EBLUP (Empirical Best Linear Unbiased Prediction) predictions, the ten best hybrids in the mean of the environments and also in each one of the environments were selected.

Furthermore, in each crop season, the correlation between the mean value and the EBLUP of the SH was estimated, and the estimates of heritability were obtained using the following estimators:

- a) Standard method using the expression presented by Falconer and Mackay (1996) that presupposes balanced data and independent genetic effects:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{g \times l}^2}{l} + \frac{\sigma_e^2}{l \times r}} \quad (4)$$

- b) Holland et al. (2003) cited by Piepho and Möhring (2007) recommended for cases of unbalance and fixed effects of genotype:

$$h_H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \bar{v}_{EBLUE}/2} \quad (5)$$

where  $\bar{v}_{EBLUE}$  means the mean variance of the difference of the fitted mean values of two treatments (EBLUE).

c) Cullis et al. (2006), also estimated in cases of unbalance and random effect of genotype:

$$h_C^2 = 1 - \frac{\bar{v}_{EBLUP}}{2\sigma_g^2} \quad (6)$$

where  $\bar{v}_{EBLUP}$  means the mean variance of the difference of two EBLUPs.

Finally, combined analysis was carried out involving the single cross hybrids common to two crop seasons using different alternatives. The crop seasons chosen were 2011s1C/2012s2C and 2013s1C/2014s2C. These crop seasons were evaluated in different years and also differed in experimental accuracy and in the number of hybrids that coincided among them. In all the alternatives adopted, selection was made of the 20 best SH common to the two crop seasons chosen (2011s1C/2012s2C and 2013s1C/2014s2C). Combined analysis based on the EBLUE estimates that were obtained for the hybrids in each environment was performed a) involving only the SH common to the two environments and common residual variance; b) considering all the SH present in the two environments, i.e., also those that were eliminated in the year, crop season, or location and common residual variance; c) involving only the SH common to the two environments and considering a residual variance different for each environment using the diagonal matrix, and d) considering all the SH present in the two environments and residual variance different for each environment.

### 3 RESULTS

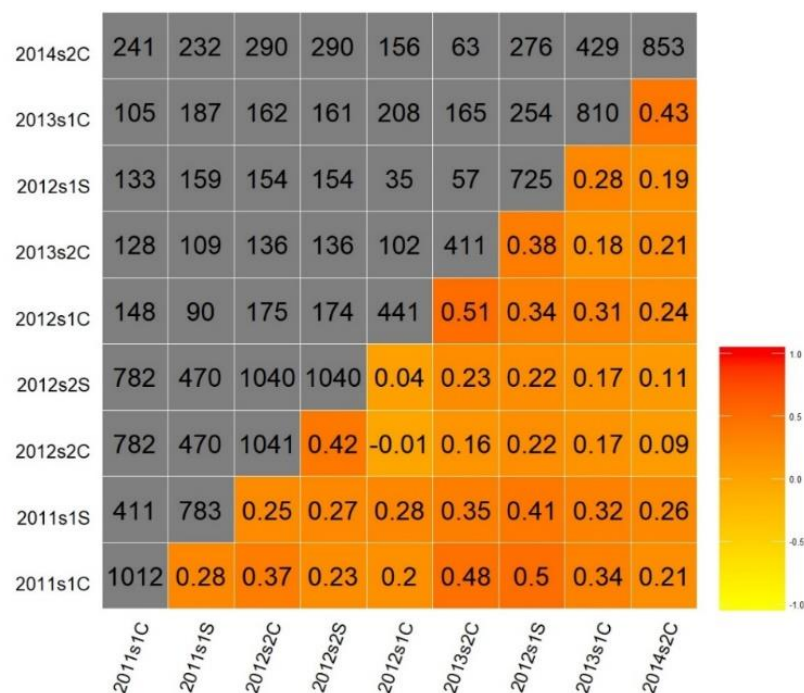
The dataset evaluated is typical of breeding programs with the objective of line introgression to obtain new SH. As many lines do not adapt well to Brazilian climate conditions, an unbalance was observed in the number of SH, locations, experiments, and replications evaluated over the crop seasons (Table 1).

The number of SH that were repeated among the crop seasons varied widely. Comparing the 2011s1S crop season and the 2012s1S crop season, of the 783 SH evaluated in the first year, only 159 proceeded, i.e., high selection intensity was applied and only 20% of the SH evaluated in 2011s1S were allocated to the experiments in the following crop season. An even more complex scenario was observed between the 2012s2C and 2013s2C crop seasons, where only 13% of the hybrids evaluated in 2012s2C advanced to the 2013s2C crop season. The same observation is valid for other years when comparing crop seasons and/or regions, and it becomes clear that there is great difficulty in evaluating data from different crop seasons in a combined manner (Figure 2).

The overall mean of the crop seasons ranged from 3.9 to 9.7 ( $\text{t}\cdot\text{ha}^{-1}$ ), and the first crop seasons ( $8.3 \text{ t}\cdot\text{ha}^{-1}$ ) were 1.56 times higher yielding than the second crop seasons ( $5.3 \text{ t}\cdot\text{ha}^{-1}$ ), regardless of the region evaluated. The variation among the crop seasons was high; for example, for the West Center region, the yield in the first crop season in 2011/2012 was 58% greater than in the second crop season. However, in 2012/2013, this superiority was much lower, only 18%, but returned to a higher level in 2013/2014 at 51%, once more showing the discrepancy among the crop seasons evaluated (Table 2).



Figure 2: The diagonal of the heat plot corresponds to the number of hybrids evaluated in each crop season. Upper diagonal heat plot indicates the number of genotypes in common that were evaluated in the pairs of environments while lower diagonal indicates the estimates of the correlation between the mean values of the SH in common in the crop seasons two by two. Crop seasons are labeled by their year of sowing (2011 - 2014), followed by their crop season (first crop, s1, or second crop, s2) and region (West Center, C, or South, S).



Source: Author (2021).

Within each crop season, a greater variation in mean yield of the SH tested was observed among the experiments than among the locations. The amplitude of variation of the experiments in relation to the mean was up to 110%, as is the case of the 2011s1S crop season ( $[(11.2 - 2.7) / 7.6] = 1.10$ ). The variation of the lower and upper limits among the locations in relation to the overall mean was under 55%, except in the 2011s1S and 2014s2C crop seasons. It is important to

highlight that the greater variation among experiments is a result of the effect of locations and also of the different SH evaluated among the experiments (Table 2).

Table 2: Overall mean, lower limit (LL), upper limit (UL) of mean grain yield of single cross maize hybrids in the experiments and environments of each crop season in t.ha<sup>-1</sup> and correlation between the mean value and the EBLUP of each hybrid in the nine crop seasons evaluated (r).

<b>Crop Season</b>	<b>LL and UL of the mean yield of the experiments</b>	<b>LL and UL of the mean yield of the locations</b>	<b>Overall mean of the crop seasons</b>	<b>r</b>
2011s1C	6.0 - 9.9	6.5 - 9.3	8.5	0.80
2011s1S	2.7 - 11.1	4.8 - 9.8	7.6	0.77
2012s2C	3.4 - 8.2	4.2 - 7.6	6.0	0.69
2012s2S	2.7 - 5.6	3.3 - 5.2	3.9	0.62
2012s1C	4.8 - 7.6	5.7 - 7.6	7.2	0.94
2012s1S	6.3 - 10.3	7.9 - 11.4	9.7	0.90
2013s2C	3.6 - 5.9	4.6 - 7.2	5.6	0.96
2013s1C	5.9 - 9.2	6.6 - 10.1	8.5	0.99
2014s2C	1.9 - 7.3	2.9 - 7.9	5.7	0.97

Source: Author (2021).

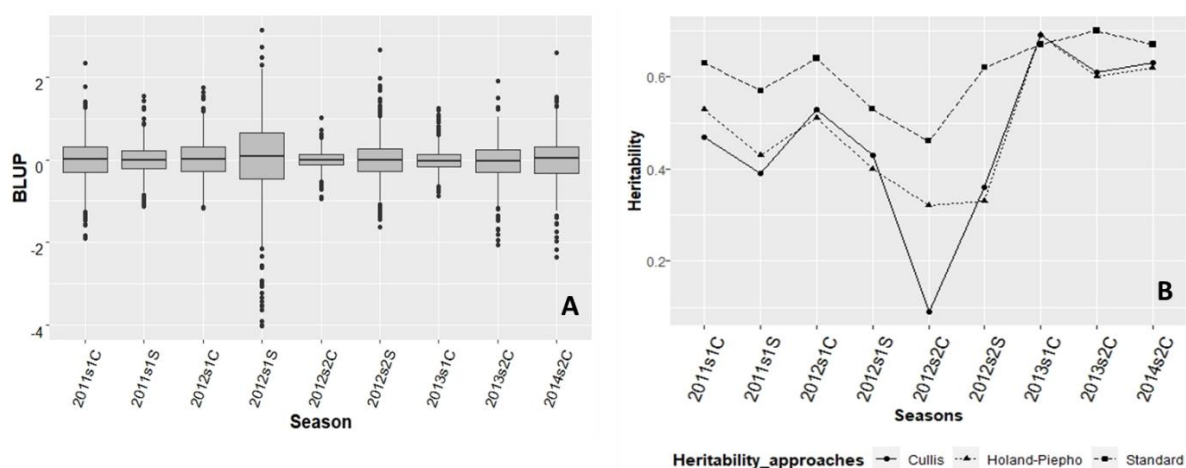
The grain yield data were fitted through the mixed models/REML approach. The estimates of correlations between the mean of the SH and their EBLUPs was greater than the 0.9 involving the crop seasons of more recent years. It follows that under these conditions, an optimal association between the mean values and the EBLUPs of the SH was obtained (Table 2).

In all the crop seasons, hybrids were observed with discrepant performance in relation to the set evaluated. The 2012s1S crop season exhibited the widest amplitude of variation, which was associated with the highest mean yield value (9.7 t.ha<sup>-1</sup>). In that crop season, the EBLUPs of the hybrids ranged from -4.0 to 3.1. In the 2012s2C crop season, lower amplitude of variation and a

mean value of  $6.0 \text{ t}\cdot\text{ha}^{-1}$  was found, with EBLUPs ranging from  $-0.7$  to  $1.0$  (Table 2 and Figure 3A).

To more easily make inferences regarding what happens among the locations within the same crop season, each location in the 2011s1C crop season was analyzed in detail. In that year, a total of 28 different experiments were set up; however, the number of experiments evaluated in each location was different. The mean yield variation of the experiments in each location was at most 44% ( $6.1$  to  $8.8 \text{ t}\cdot\text{ha}^{-1}$ ), and the overall mean of the five locations evaluated was from  $6.5$  to  $9.2 \text{ t}\cdot\text{ha}^{-1}$ . Variation was observed in the magnitude of the estimates of genetic variance and of standard heritability among the five locations. The variance of the hybrid  $\times$  experiment interaction was greater than the genetic variance in all the locations evaluated in this season, except for one of locations where both variances were similar.

Figure 3: A) Boxplot representing the variation between the upper and lower EBLUP limit of the predictions of the single cross hybrids evaluated in each crop season and possible outliers. B) Heritability estimates for grain yield over nine crop seasons using three different methods (standard, Cullis, and Holland-Piepho).



Fonte: Author (2021).

The estimates of the genetic variance components ( $\sigma_g^2$ ) involving all the experiments and locations obtained in each crop season were significant by the likelihood ratio test, indicating the presence of genetic variability and the possibility of selection among the hybrids. The magnitude of the variance of the hybrid by environment interaction in relation to genetic variation was expressive, reflecting the hybrid performance that did not coincide across the environments. The  $\sigma_{g \times l}^2 / \sigma_e^2$  ratio ranged from 0.12 (2013s1C) to 4.47 (2012s2C) among the crop seasons, accentuating what was commented.

The expressive existence of the hybrid  $\times$  environment interaction within each crop season was also found through the estimates of the correlations ( $r_{qs}$ ) of the mean performance of the SH coinciding in the crop seasons two by two. The estimates of  $r_{qs}$  across the combinations of crop seasons ranged from -0.01 to 0.51. In the pairs of environments 2012s2C/2012s1C, 2012s2S/2012s1C, 2012s2C/2014s2C, and 2012s2S/2014s2C, in which the lowest estimates of  $r$  were observed, high intensity of selection applied was also found in the SH evaluated from one crop season to another (Figure 2).

The presence of the interaction was also highlighted by the coincidence among the ten best SH selected in the mean of the EBLUPs of the two crop seasons in relation to their relative performance in each crop season. When the same region and year of evaluation were considered, i.e., the response of the first and second crop season, the coincidence varied between the sowing times. As expected, the greatest coincidence in most cases was in the first crop season. However, even under these conditions, in the 2011/2012 and 2013/2014 crop years, the coincidence was less than 50%. These results, once more showing, the difficulty of moving toward recommendation of new SH involving different regions, sowing times, and crop years.

As the experiments were unbalanced, the heritability ( $h^2$ ) in the mean of the hybrids within each crop season was estimated considering three strategies: i) standard method according to Falconer and Mackay (1996), ii) method according to Cullis et al. (2006), and iii) method according to Holland et al. (2003), cited by Piepho and Möhring (2007). Strategy i, prescribing the use of balanced data, ranged from 0.46 to 0.69 among the crop seasons, and as expected, was always superior to the estimates of  $h^2$  obtained by the other strategies, except in the 2013s1C crop season (Figure 3B).

The estimates of heritabilities proposed by Cullis et al. (2006) and by Holland et al. (2003); Piepho and Möhring (2007) were always of similar magnitude, except in the 2012s2C crop season, and ranged from 0.09 to 0.69 among the crop seasons evaluated. However, the coincidence in the estimates of  $h^2$  of the three strategies was greater in the experiments conducted in recent years (Figure 3A).

In carrying out combined analysis involving the SH common to two or more environments, there are some alternatives. One is involving only the SH in common and the other would be considering all the SH, i.e., also those that were eliminated in some environments. It is also possible to carry out the analyses considering residual variance in common or heterogeneous residual variance. In this study, combined analysis was performed considering the combinations of the 2011s1C/2012s2C and 2013s1C/2014s2C crop seasons. These pairs were chosen as they consisted of data from the first and second crop seasons of different years that also differed in experimental accuracy and in the number of coinciding hybrids.

For the combination of 2011s1C/2012s2C that involves the first and second crop seasons in the 2011/2012 crop year, the coincidence of the 20 best SH, which is what most interests breeders, changed significantly. Of the 20 best SH ranked in the analysis involving all the hybrids

common to the two crop seasons, only five remained when the analysis was performed considering only the SH common to the two crop seasons. This information is valid for both cases, when the homogeneous or heterogeneous residual variance is considered.

Different results were observed for the combination of the first and second crop seasons in the 2013/2014 crop year (2013s1C/2014s2C combination) in relation to previous crop seasons. Coincidence in identification of the 20 best SH, involving all the SH or only the SH common to two crop seasons, was total upon using the same residual variance. However, when the residual variance used was heterogeneous, though the coincidence was high, it was not total (15 SH in 20 SH).

#### **4 DISCUSSION**

The challenge common to all companies is evaluating a large number of SH annually for the purpose of recommending those that have the best performance for farmers. The information coming from these evaluations is often unbalanced in relation to the number of SH, of replications, of experiments, and of locations, which may compromise the choice of the best hybrid. In addition, the breeder needs to deal with the hybrid  $\times$  environment interaction in seeking greater reliability in future recommendations since this interaction is a complicating factor in the performance of the SH evaluated.

The results obtained in this analysis showed considerable substitution of hybrids among the different environmental conditions (Figure 2). This low coincidence among the hybrids evaluated is explainable because if a determined SH evaluated did not have good results under certain conditions, why reevaluate it under other conditions? Other plausible explanations would be the difficulty of continuing to evaluate a SH that has some agronomic trait other than grain yield

that would make its future recommendation unviable, as well as the lack of adaptation of the lines to tropical conditions, which impedes the production of hybrid seeds in large quantity for evaluation of the SH in different environments. Thus, it is expected above all between crop seasons that the low coincidence among the SH evaluated is a reality that likely will not change.

The low coincidence among the hybrids evaluated in the different environments makes it difficult to estimate genetic and phenotypic parameters, especially genetic variance and, above all, the SH  $\times$  environment interaction. This difficulty has been reported in the literature by diverse authors in recent years (SMITH et al., 2001; MÖHRING; PIEPHO, 2009; SMITH et al., 2015; NUVUNGA et al., 2015; DA SILVA et al., 2019).

To deal with unbalanced data, some proposals have been implemented more recently for analysis of experiments with plants using, for example, analysis in two steps, in which weighting is considered in the second step in accordance with the number of replications, with the experimental design, and with residual variance (SMITH et al., 2001; MÖHRING; PIEPHO, 2009, WELHAM et al., 2010, PIEPHO et al., 2012). Other alternatives are multiplicative models (SMITH et al., 2015, NUVUNGA et al., 2015), sequential analysis, which considers all the hybrids evaluated in the previous generations (PIEPHO; MÖHRING, 2006) and models that consider the use of heterogeneous residual variance (EDWARDS; JANNINK, 2006; SO; EDWARDS, 2011; ORELLANA et al., 2014; HU et al., 2014; ANDRADE et al., 2016; DA SILVA et al, 2019).

Due to the wide variation in analytical possibilities for unbalanced experiments in multi-environments, in this study, individual analyses were initially performed in each location within each crop season. Due to the great volume of information, we chose to present only the results in reference to the 2011s1C crop season. In this analysis, the importance of the SH  $\times$  experiment

interaction was clear, even in a single location. This was possible because some SH were present in more experiments.

After that, the yield data from each crop season were fitted through mixed models regarding the block and location effect, seeking to obtain the best estimates of the genetic value of each hybrid. In conditions as observed in this study, wherein a huge number of hybrids were evaluated in many environments across the years, the use of approaches like unstructured VCOV matrices and factorial analytic models have been adopted, once these structures allow to model different genetic variance to each site and different covariances between pairs of environments evaluated (SMITH et al., 2001; BURGUEÑO et al., 2012; KRAUSE et al., 2020; DE OLIVEIRA et al., 2020).

In this study, unstructured VCOV structures were tested to better understand the correlation between environments by including the genotype by environment interaction in the model. However, models that include unstructured VCOV matrices shows computational difficulties to converge due a huge number of parameters to be estimated under high unbalance level. The factor analytic structure is an alternative approach to deal with these limitations (SMITH et al., 2001; KELLY et al., 2007; DIAS et al., 2018b). Due to the difficulty of convergence of models and computational limitations, in this study we adopted VCOV simpler structures to model the effect of location within each crop season. Despite that, always is possible, it is important to assume models with more complex structures, as previously mentioned.

In situations in which the dataset exhibits considerable unbalance, a way of checking the fit of the model is to correlate the mean values and the EBLUPs of the SH. In general, the estimates of correlation in most of the crop seasons evaluated were high, especially in more recent years, showing that in many situations, when the experiments are well conducted, the mean can be



considered a good indication of the performance of the SH, even under unbalanced conditions (Figure 2).

A significant effect of the SH  $\times$  environment (locations and experiments within locations) interaction was found in all the crop seasons evaluated. In these cases, the SH  $\times$  environment interaction component was greater than the genetic variance component in most of the crop seasons. This is very frequent in most of the situations in which various hybrids are evaluated in the same crop season (TONK et al., 2011; NZUVE et al., 2013; NDHLELA et al., 2014; MENGESHA et al., 2019). In the conditions evaluated, the significant effect of the interaction is expected due to the expressive environmental variation of numerous factors, such as climate, soil fertility, and management practices that occurs in the different locations in which maize experiments are conducted (NOIA JUNIOR et al., 2019; EMBRAPA, 2020).

It should be emphasized that environmental variations under tropical and subtropical conditions are more expressive than those normally observed under temperate conditions. This environmental variation is even more challenging since it is largely unpredictable (EEUWIJK et al., 2016). Given this situation, the great challenge of breeders is identifying hybrids that are more adapted and stable under these growing conditions. For that purpose, numerous methods have been proposed in the literature over the past fifty years (EBERHART; RUSSEL, 1966; WRICKE; WEBER, 1986; GAUCH; ZOBEL, 1988; PIEPHO, 1997; YAN et al., 2000, SMITH et al., 2015, NUVUNGA et al., 2015); most recently, the use of mixed models has been proposed above all, according to a survey performed by Eeuwijk et al. (2016).

In addition, in Brazil, variation in environmental factors in the second crop season is much more expressive than in the first, especially due to drought stress or heat stress (DA SILVA et al., 2019, DA SILVA et al., 2018), and so a difference in mean yield between the crop seasons is

expected. In spite of that, this yield difference has diminished through the choice of more adapted hybrids and the use of greater technology in crop fields. The great challenge for seed production companies currently has been identifying hybrids adapted to both growing conditions. The results obtained using this dataset show that finding a hybrid with wide adaptation to different climatic regions is a challenging factor for breeders because of the enormous contribution of the SH  $\times$  crop season interaction (Figure 2).

In the present study, the effect of the interaction on SH performance can be observed through estimation of the correlation between the mean of the hybrids in common across the crop seasons (Figure 2). Another option for the study of the interaction with greater importance for breeders was the coincidence of the hybrids selected considering two or more environments. The low magnitudes of the estimates of correlation and the low coincidences observed show that in most of the cases evaluated, the response to the interaction was of a complex nature and in some cases it was probably not even linear, making identification of the best hybrid difficult. Results similar to these are discussed by Eeuwijk et al. (2016) through graph illustrations involving the yield of the genotype and environmental quality.

The heritability ( $h^2$ ) estimate is a key parameter in plant breeding because it is associated with predictive measurement of success in selection. It has been estimated by the ratio between the part of genetic variance exploited by the genotypes evaluated and the phenotypic variance of the selection unit applied (FALCONER AND MACKAY, 1996; BERNADO, 2010). However, with the increased use of mixed models to attenuate the effects of unbalanced data, new options of  $h^2$  estimates have been proposed (CULLIS et al., 2006; PIEPHO AND MOHRING, 2007; SCHMIDT et al., 2019).

In this study,  $h^2$  was estimated by three procedures. Especially in the first crop seasons, the estimates did not greatly coincide. However, in more recent crop seasons, there was greater coincidence. It should be emphasized that, as was expected in all cases, the absolute value of  $h^2$  in the standard method was superior to the other two (CULLIS et al., (2006); SCHMIDT et al., 2019) (Figure 3A). This occurs because in the standard method, phenotypic variance is estimated considering that there is no variation in the number of replications and locations, for example. This discrepancy in the estimates of  $h^2$  has also frequently been observed in other conditions (PIEPHO AND MOHRING, 2007; SCHMIDT et al., 2019).

Regardless of the method used, the  $h^2$  estimates, in most cases, were considered of medium magnitude, which is a favorable condition for selection of SH, based on the overall mean of each crop season. It should be emphasized that when selection is made among SH, all the genetic variance is used, i.e., additive, dominant, and epistatic variance (HALLAUER et al., 1998; SOUZA JUNIOR, 2007).

The proposal of using all the hybrids in analyses and not only those in common in both crop seasons, as proposed by Piepho; Möhring (2006), did not prove to be effective in relation to the use of only the SH in common. The fact of considering homogeneous residual variance or not leads to a difference above all when the  $h^2$  are of lower magnitude and, therefore, it is difficult to decide in the latter case from the results obtained in this study that the use of heterogeneous residual variance is more appropriate, since there is no way to prove which ranking is more trustworthy. In the literature, however, there are numerous reports that the use of heterogeneous variance is more advisable (EDWARDS; JANNINK, 2006; SO; EDWARDS, 2011; ORELLANA et al., 2014; HU et al., 2014; ANDRADE et al., 2016; DA SILVA et al, 2019).

From the above, it is clear that the possibility of selecting general hybrids for different growing seasons is very difficult. The possibilities of identifying hybrids that stand out under both conditions can be increased when using experiments with a smaller number of hybrids, with check varieties that are common to the experiments, with more replications and evaluations in the greatest number of locations possible, as has frequently been reported in the literature (TROYER, 1996; COOPER et al., 2014; GAFFNEY et al., 2015).

In the current period of “plant breeding 4.0”, the need for evaluating hybrids considering various replications is not disregarded. In addition, the proposal considers the use of other information, such as climate, soil, geographic coordinates, phenological data, molecular markers, and the possibilities that exist in current analytical terms to identify the SH with best performance (WALLACE et al., 2016; ERSOZ et al., 2019; RAMSTEIN et al., 2019). Obtaining accurate experiments is especially important in the molecular marker validation phase. Without accurate experiments, it is impossible to find trustworthy associations between the phenotype and the molecular marker.

More recently, the use of genomic selection models, including the effect of the genotype  $\times$  environment interaction, have frequently been reported as a tool to accelerate the selection process and improve the accuracy between the predicted value and observed value in breeding programs (CUEVAS et al., 2016; LADO et al., 2016; FERRÃO et al., 2018, DIAS et al., 2018a, MONTESINOS-LÓPEZ et al., 2019, MONTEVERDE et al. 2019; AMES AND BERNADO, 2020). Nevertheless, it is clear that the interaction information will only effectively contribute to improve the predictive capability of the models if the interaction component used includes not only the genetic variation but also the future possibilities of environmental variation.

The analysis presented here, were carried out to better understand what happens with this data set. In addition, provide subsidies for the genomic prediction study, within will be carried out in a subsequent step, based on the genotyping of the parental lines of the hybrids evaluated in this study. Studies published recently in the literature, involving the prediction of hybrids under different environmental conditions, suggest that the inclusion of the component genotype  $\times$  environment interaction in genomic prediction models, may improve hybrids predictions if the environmental component is reliable (KRAUSE et al., 2020; DE OLIVEIRA et al., 2020). The question remains, given that the experiments are very unbalanced, if the component of the interaction to be used in the model will be able to improve its predictive capability, since, as found in this study, the component of the hybrid  $\times$  environment interaction is very expressive.

Therefore, it should be highlighted that in breeding programs of any species, the most important step is the final evaluation of the lines/hybrids. Recommendation of a cultivar with low accuracy of evaluation is a huge risk, not only economically, but also for the image of the company. The risk will only be reduced if, as already emphasized, the experiments are not only conducted in various environments, but are also as accurate as possible.

## 5 REFERENCES

AMES Nicholas C.; BERNARDO, Rex. 2020. Genomewide predictions as a substitute for a portion of phenotyping in maize. **Crop Science**, Saint Paul, p. 1-9, 2020.

ANDRADE, Vinícius T. et al. Statistical modeling implications for coffee progenies selection. **Euphytica**, v.207(1), p. 177–189, 2016.

BERNARDO, R. **Breeding for quantitative traits in Plants**. Ed. 2. Woodbury: Stemma Press, 2010. 400p.

BURGUEÑO, Juan et al. Genomic prediction of breeding values when modeling genotype X environment interaction using pedigree and dense molecular markers. **Crop Science**. v. 52(2), p. 707–719, 2012.

COOPER, Mark et al. Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. **Crop and Pasture Science**, v. 65(4), p. 311-336, 2014.

CULLIS, B. R.; SMITH, A. B.; COOMBES, N. E. On the design of early generation variety trials with correlated data. **Journal of Agricultural, Biological, and Environmental Statistics**, v.11(4), p. 381–393, 2006.

DA SILVA Andrea M. C.; DALLACORT R.; BARBIERI J. D.; TIEPPO R. C. Impacts of Future Climate Predictions on Second Season Maize in an Agrosystem on a Biome Transition Region in Mato Grosso State. **Revista Brasileira de Meteorologia**, v.34(2), p.335-347, 2019.

DA SILVA Andrea, M.C et al. Variability and limitations of maize production in Brazil: potential yield, water-limited yield and yield gaps. **Agricultural Systems**, v.165, p.264–273, 2018.

DA SILVA, C. P. et al. Heterogeneity of Variances in the Bayesian AMMI Model for Multienvironment Trial Studies. **Crop Science**, v.59, p. 1-18, 2019.

DE OLIVEIRA, A.A et al. Genomic prediction applied to multiple traits and environments in second season maize hybrids. 2020. **Heredity**, 2020.

DIAS, Kaio et al. Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. **Heredity**, 2018.

DIAS, Kaio et al. Estimating genotype X environment interaction for and genetic correlations among drought tolerance traits in maize via factor analytic multiplicative mixed models. **Crop Science**, v.58(1), p. 72-83, 2018.

EBERHART, S. A.; RUSSELL, W. A. Stability parameters for comparing varieties. **Crop Science**, v.6, p.36-40, 1966.

EEUWIJK, F.; BUSTOS-KORTS, D; MALOSETTI, M. 2016. What Should Students in Plant Breeding Know About the Statistical Aspects of Genotype X Environment Interactions?. **Crop Science**, v56, p.2119-2140, 2016.

EDWARDS, J. W.; JANNINK, J.L. Bayesian Modeling of Heterogeneous Error and Genotype X Environment Interaction Variances. **Crop Science**, v.46(2), 820-833, 2006.

EMBRAPA. Empresa Brasileira de Agropecuária. Available at: <https://www.spo.cnptia.embrapa.br>. [Accessed July 06, 2020].

ERSOZ, E.S.; MARTIN, N.F.; STAPLETON, A.E. 2019. On to the Next Chapter for Crop Breeding: Convergence with Data Science. **Crop Science**, 2019.

FALCONER, D.S.; MACKAY, T.F.C. 1996. **Introduction to quantitative genetics**. 4<sup>a</sup> Ed. London: Longman, 1996.

FERRÃO, L.F.V. et al. Integration of Dominance and Marker x Environment Interactions into Maize Genomic Prediction Models. **Crop Science**, 2020.

GAFFNEY, J. et al. Industry-Scale Evaluation of Maize Hybrids Selected for Increased Yield in Drought-Stress Conditions of the US Corn Belt. **Crop Science**, v.55(4), p.1608-1618, 2015.

GAUCH, H.G.; ZOBEL, R.W. Predictive and postdictive success of statistical analyses of yield trials. **Theoretical and Applied Genetics**, V.76, p.1–10, 1988.

HALLAUER, A.R; MIRANDA FILHO, J.B. **Quantitative Genetics in maize breeding**. 2 edition, Ames: Iowa State University Press, 1998.

HOLLAND, J.; BNYQUIST, W. E.; CERVANTES-MARTINEZ, C. T. Estimating and interpreting heritability for plant breeding: an update. **Plant Breed. Rev.** 22: 9–112, 2003

HU, X.; YAN, S.; LI, S. The influence of error variance variation on analysis of genotype stability in multi-environment trials. **Field Crops Research**, v.156, p. 84–90, 2014.

KELLY, A.M. et al. The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. **Crop Science**, v. 47, p.1063–1070, 2007.

KRAUSE, M. D. et al. Boosting predictive ability of tropical maize hybrids via genotype by environment interaction under multivariate GBLUP models. **Crop Science**, v.60, p.3049-3065, 2020.

LADO, B. et al. Modeling genotype X environment interaction for genomic selection with unbalanced data from a wheat breeding program. **Crop Science**, v.56, p.2165–2179, 2016.

MENGESHA, W.; MENKIR, A.; MESEKA, S. 2019. Factor analysis to investigate genotype and genotype X environment interaction effects on pro-vitamin A content and yield in maize synthetics. **Euphytica**, v.215, p.1-15, 2019.

MOHRING, J.; PIEPHO, H.P. Comparison of weighting in two-stage analysis of plant breeding trials. **Crop Science**, v.49, p.1977–1988, 2009.

MONTEVERDE, E. et al. Integrating Molecular Markers and Environmental Covariates To Interpret Genotype by Environment Interaction in Rice (*Oryza sativa* L.) Grown in Subtropical Areas. **G3: GENES, GENOMES, GENETICS**, v.9, p.1519-1531, 2019.

MONTESINOS-LÓPEZ, O. A. et al. Multi-Trait, Multi-Environment Genomic Prediction of Durum Wheat With Genomic Best Linear Unbiased Predictor and Deep Learning Methods. **Frontiers Plant Science**, 2019.



NDHLELA, T. et al. Genotype X Environment Interaction of Maize Grain Yield Using AMMI Biplots. **Crop Science**, v.54(5), 2014.

NÓIA JUNIOR, R.D.S.; SENTELHAS, P.C. Soybean–maize off–season double crop system in Brazil as affected by El Niño Southern Oscillation phases. **Agric. Syst.** 173, 254–267, 2019.

NUVUNGA, J. et al. Factor analysis using mixed models of multi-environment trials with different levels of unbalancing. **Genetics and Molecular Research**, v.14, p.14262-14278, 2015.

NZUVE, F. et al. Analysis of Genotype x Environment Interaction for Grain Yield in Maize Hybrids. **Journal of Agricultural Science**, v.5, p.75-85, 2013.

ORELLANA, M.; EDWARDS, J.; CARRIQUIRY, A. Heterogeneous Variances in Multi-Environment Yield Trials for Corn Hybrids. **Crop Science**, v.54, p.1048-1056, 2014

PATTERSON, H.D.; THOMPSON, R. Recovery of interblock information when block sizes are unequal. **Biometrika**, v.58, p.545-554, 1971.

PIEPHO, H.P. Analyzing genotype-environment data by mixed models with multiplicative terms. **Biometrics**, v.53, p.761–766, 1997.

PIEPHO, H.; MOHRING, J, Selection in Cultivar – Is It Ignorable? **Crop Science**, Madison, v.46, n.6, p. 192-201, 2006.

PIEPHO, H.P.; MÖHRING, J. Computing heritability and selection response from unbalanced plant breeding trials. **Genetics**, v.177, p.1881–1888, 2007.

PIEPHO, H.P. et al. A stagewise approach for the analysis of multi-environment trials. **Biometrical Journal**, v.54, p.1-17, 2012.

RAMSTEIN, G.P.; JENSEN, S.E.; BUCKLER, E.S. Breaking the curse of dimensionality to identify causal variants in Breeding 4. **Theoretical and Applied Genetics**, v.132, p.559–567, 2019.

SMITH, A. B. et al. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. **Theoretical and Applied Genetics**, v.128, p.55-72, 2015.

SMITH, A.; CULLIS B. R.; THOMPSON R. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. **Biometrics**, v.57(4), p.1138–1147, 2001.

SMITH, A. B.; CULLIS, B. R.; GILMOUR, A. The analysis of crop variety evaluation data in Australia. Australian and New Zealand, **Journal of Statistics**, v.43, p.129-145, 2001.

SCHMIDT, P. et al. Estimating Broad-Sense Heritability with Unbalanced Data from Agricultural Cultivar Trials. **Crop Science**, v.59, p.525-536, 2019.

SO, Y.-S.; EDWARDS, J. 1. Predictive Ability Assessment of Linear Mixed Models in Multienvironment Trials in Corn. **Crop Science**, v.51, 2011.

SOUZA, JR. C.L. **Melhoramento de espécies alógamas**. In: NASS LL. (ed.) Recursos genéticos vegetais. Brasília: Embrapa Recursos Genéticos e Biotecnologia. 8, 2007.

STEEL, R.G.D.; TORRIE, J.H.; DICKEY, D.A. **Principles and procedures of statistics: A Biometrical Approach**. New York: McGraw-Hill Book Company, 1997.

TONK, F. A.; ILKER, E.; TOSUN, M. Evaluation of genotype x environment interactions in maize hybrids using GGE biplot analysis. Crop Breeding **Applied Biotechnology**, Viçosa, v. 11, p. 01-09, 2011.

TROYER, A. F. Breeding widely adapted, popular maize hybrids. **Euphytica**, v.92(1-2), p.163–174, 1996.

WALLACE, J. G.; RODGERS-MELNICK, E.; BUCKLER, E. S. 2016. **On the Road to Breeding 4.0: Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics. Annual Review of Genetics**, 52(1). doi:10.1146/annurev-genet-120116-024846, 2016.

WELHAM, S. et al. B. A comparison of analysis methods for late-stage variety evaluation trials. **Australian & New Zealand Journal**, v.52, p.125–149, 2010.

WRICKE, G.; WEBER, W.E. **Quantitative genetics and selection in plant breeding. Walter de Gruyter**, Berlin & New York, pp. 406, 1986.

YAN, W. et al. Cultivar evaluation and mega-environment investigation based on the GGE biplot. **Crop Science**, v.40, p.597–605, 2000.

**ARTIGO 2 – EMPREGO DE EFEITOS ADITIVOS E DE DOMINÂNCIA PARA  
PREDIÇÃO DE HÍBRIDOS DE MILHO EM REGIÕES TROPICAIS**

Artigo redigido conforme as normas NBR 6022 (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2003).

## RESUMO

Na última década, a seleção genômica vem sendo amplamente adotada em programas de melhoramento de milho como uma ferramenta vantajosa para aumentar os ganhos genéticos por meio da predição de híbridos. Diante disso, esse estudo foi realizado com o objetivo de verificar a influência dos efeitos de dominância no modelo preditivo, sob condições de alto desbalanceamento em diferentes anos e épocas de semeadura. Foram utilizados dados de 2770 híbridos de milho, avaliados ao longo de nove safras (combinação de época de semeadura e ano) quanto a produtividade de grãos. A informação genotípica desses híbridos foi obtida por meio genotipagem das 447 linhagens parentais utilizando 23.153 marcadores Darts. A análise de seleção genômica foi realizada utilizando o modelo GBLUP (Genomic Best Linear Unbiased Prediction) com a inclusão das matrizes de parentesco aditivo (GBLUP\_A) e de dominância (GBLUP\_AD). A herdabilidade e as correlações entre os valores preditos e observados variaram consideravelmente ao longo dos anos, demonstrando os desafios de lidar com dados desbalanceados. Ambos os modelos foram eficientes em prever híbridos em ambientes não avaliados e possibilitaram incrementos no ganho com a seleção, quando comparado com a seleção baseada apenas nas informações fenotípicas.

**PALAVRAS-CHAVE:** Seleção genômica; Efeitos não aditivos; Ganho com a seleção; GBLUP; Desbalanceamento.

## 1 INTRODUÇÃO

Na última década, a seleção genômica passou a ser amplamente utilizada em programas de melhoramento de plantas. Dentre os inúmeros benefícios dessa ferramenta, diversos trabalhos destacam a possibilidade de prever o desempenho de genótipos não avaliados, o aumento do ganho com a seleção e a redução dos custos com a fenotipagem, do número de locais a serem avaliados e do tempo para seleção de indivíduos superiores (MEUWISSEN, 2007; BERNARDO, 2009; DESTA; ORTIZ, 2014; LADO et al, 2016; DIAS et al., 2018; FERRÃO et al, 2020; FRITSCHENETO et al. 2018; AMES; BERNADO, 2020; KRAUSE ET AL., 2020).

A aplicação da seleção genômica no melhoramento de milho tem sido cada vez mais frequente (DOS SANTOS et al., 2016; LYRA et al. 2017; DIAS et al., 2018; FRITSCHENETO et al. 2018; DE OLIVEIRA et al., 2020). Uma das vantagens dos programas de melhoramento de milho, é a possibilidade de explorar o vigor híbrido ou a heterose, que pode ser entendida como a superioridade do híbrido em relação a média dos pais (SHULL, 1952). O uso da heterose na obtenção de híbridos só é possível pois, grande parte da variância genética observada é devido a variância de dominância. (HALLAUER; MIRANDA FILHO, 1998; BERNARDO, 2020). Em teoria, o efeito de dominância é considerado como uma interação entre os alelos de um mesmo gene, podendo ser definido como a diferença entre valor genotípico do heterozigoto em relação a média dos homozigotos (FALCONER; MACKAY, 1996).

A inclusão dos efeitos de dominância nos modelos tem sido considerado fator crucial para melhorar a acurácia preditiva e a qualidade das componentes de variância, uma vez que as variâncias aditiva e de dominância podem ter seus efeitos confundidos em modelos cuja apenas os efeitos aditivos são explorados (HILL et al., 2008; TECHNOW et al. 2014; DOS SANTOS et al., 2016; DIAS et al., 2018; LIU et al., 2019). Apesar da estimativa dos efeitos aditivos e de

dominância auxiliar no entendimento da herança genética da característica alvo e a definir as estratégias de melhoramento, muitas análises genéticas ainda se concentram apenas nas estimativas dos efeitos genéticos aditivos ou totais, isso porque a estimativa dos efeitos não aditivos (dominância e epistáticos) muitas vezes é difícil, exigem modelos específicos e um grande número de observações devido à falta de ortogonalidade (DIAS et al., 2018).

Adicionalmente, outros fatores como a estrutura da população, o desbalanceamento excessivo entre os ensaios de avaliação da performance dos híbridos e a presença da interação genótipos por ambientes, requerem atenção e de acordo com a literatura sempre que possível devem ser levados em consideração nos modelos de seleção genômica pois podem afetar diretamente as estimativas da acurácia preditiva (SCHOPP et al., 2017; ZHANG et al., 2017a; CANTELMO et al., 2017; DE OLIVEIRA et al., 2018; DIAS et al., 2018; ZHANG et al., 2019; KRAUSE ET AL., 2020). Frente a esses desafios, o objetivo desse estudo foi comparar a habilidade preditiva dos modelos de seleção genômica com a inclusão apenas de efeitos aditivos e dos efeitos aditivos e de dominância, utilizando uma alta densidade de marcadores, em diferentes regiões brasileiras, épocas de semeadura e anos, sob condições de alto desbalanceamento.

## 2 MATERIAL E MÉTODOS

### 2.1 Dados fenotípicos

Foram utilizados dados de produtividade de grãos ( $t\cdot ha^{-1}$ ), de 2770 híbridos de simples (HS) de milho, oriundos de um programa de introgressão de linhagens de uma empresa produtora de sementes. Esses HS foram resultantes do cruzamento de 447 linhagens provenientes de diferentes regiões tropicais, subtropicais e temperadas ao redor do mundo. O germoplasma em estudo é bastante diversificado e foi estratificado em dez grupos heteróticos (Iodent, Lancaster, Suwan, Non-Stiff Stalk Synthetic (NSSS), Stiff Stalk Synthetic (SSS), Tropical Dent, Tropical Flint, Amarillo Dentado, Femeas DK-Argentina and Machos NK-Argentina). Esses grupos heteróticos, foram subdivididos de acordo com o país de origem ou de acordo com o genitor que originou cada linhagem (CANTELMO et al., 2017).

Os HS foram avaliados ao longo de quatro anos (2011-2014), duas épocas de semeadura (1ª safra e 2ª safra) e em duas regiões distintas do Brasil (Sul e Centro oeste), totalizando nove safras (combinação de ano, época de semeadura e região). As safras foram identificadas com uma sigla que corresponde ao ano (2011, 2012, 2013 ou 2014), seguido pela época de semeadura (primeira safra: s1 ou segunda safra: s2) e região (centro-oeste: C ou sul: S). Os delineamentos experimentais utilizados para a avaliação dos híbridos foram o de blocos casualizados (DBC) e de blocos incompletos (DBI) com duas ou três repetições. As parcelas foram constituídas de quatro linhas de 5m e espaçamento de 0,7 m entre linhas. O número de HS, locais e experimentos avaliados por safra foi bastante variável, caracterizando um conjunto de dados altamente desbalanceado (Material suplementar 1).



Em cada safra, foi realizada a análise conjunta dos dados utilizando um modelo linear misto para obter as estimativas BLUE (Best Linear Unbiased Estimators) para cada HS. O ajuste dos fenótipos y foi dado pelo seguinte modelo:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_{ga}\mathbf{u}_{ga} + \mathbf{Z}_b\mathbf{u}_b + \mathbf{e} \quad (2)$$

em que:  $\mathbf{y}$  é o vetor de observações fenotípicas;  $\boldsymbol{\tau}$  é o vetor de efeitos fixos de (design experimental DBC: local e experimento dentro de local; design experimental DBI: local e repetição dentro de local);  $\mathbf{u}_g$  é o vetor de efeitos genotípicos fixo de híbridos;  $\mathbf{u}_{ga}$  é o vetor de efeitos fixos da interação híbridos por locais, com  $\mathbf{u}_{ga} \sim N(0, \sigma_{ga}^2 \mathbf{I}_{ga})$ ;  $\mathbf{u}_b$  é o vetor de efeitos aleatórios de (design experimental DBC: repetição dentro de experimento e local; design experimental DBI: bloco dentro de repetição e ambiente, com  $\mathbf{u}_b \sim N(0, \sigma_b^2 \mathbf{I}_b)$ ;  $\mathbf{e}$  é o vetor de erros aleatórios, com  $\mathbf{e} \sim N(0, \mathbf{R})$ ;  $\mathbf{X}$ ,  $\mathbf{Z}_g$ ,  $\mathbf{Z}_{ga}$  e  $\mathbf{Z}_b$  são as matrizes de incidência associadas aos vetores  $\boldsymbol{\tau}$ ,  $\mathbf{u}_g$ ,  $\mathbf{u}_{ga}$  e  $\mathbf{u}_b$ ;  $\sigma_{ga}^2$ ,  $\sigma_b^2$  e  $\sigma_e^2$  são os componentes de variância associadas aos vetores  $\mathbf{u}_{ga}$ ,  $\mathbf{u}_b$  e  $\mathbf{e}$ ;  $\mathbf{I}_{ga}$ ,  $\mathbf{I}_b$  e  $\mathbf{I}_n$  são as matrizes identidade associadas aos vetores  $\mathbf{u}_{ga}$ ,  $\mathbf{u}_b$  e  $\mathbf{e}$ . A matriz de (co)variância residual, visando modelar o efeito de local dentro de cada safra, assumiu uma estrutura de variância bloco diagonal empregando a matriz identidade ( $\mathbf{R} = \bigoplus_{l=1}^t \sigma_{e_l}^2 \mathbf{I}_{n_l}$ ).

A significância dos efeitos fixos foi testada por meio do teste de Wald (W), enquanto a significância dos efeitos aleatórios foi verificada pelo teste da máxima verossimilhança residual:

$$W = \frac{\hat{\theta}}{DP(\hat{\theta})} ; \quad D = 2 * [\log L_R - \log L_F] \sim \chi_{f_R - f_L}^2$$

Em que  $\theta$  representa os efeitos estimados para os efeitos fixos; DP é o desvio padrão da  $\theta$ ; Foi considerado  $H_0: \theta = 0$ ;  $H_1: \theta \neq 0$ ;  $\log L_R$  é a estimativa da máxima verossimilhança do modelo sem o efeito a ser testado;  $\log L_F$  é a estimativa da máxima verossimilhança residual do modelo

completo;  $\chi_{f_R-f_L}^2$  é uma distribuição de  $\chi^2$  com  $f_R - f_L$  graus de liberdade, que representa a diferença nos parâmetros entre o modelo completo ( $f_R$ ) e modelo reduzido ( $f_L$ ).

As estimativas BLUE foram utilizadas nos modelos de predição para obter os valores genéticos genômicos preditos associados a cada HS. Para cada safra, também foi ajustado um modelo de linear misto para obter as predições BLUPs (Best Linear Unbiased Predictor). Esses valores foram utilizados posteriormente para verificar a acurácia entre os valores genéticos genômicos preditos e os valores observados. Para tanto, o seguinte modelo foi adotado:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_{ga}\mathbf{u}_{ga} + \mathbf{Z}_b\mathbf{u}_b + \mathbf{e} \quad (3)$$

em que:  $\mathbf{y}$  é o vetor de observações fenotípicas;  $\boldsymbol{\tau}$  é o vetor de efeitos fixos de (design experimental DBC: local e experimento dentro de local; design experimental DBI: local e repetição dentro de local);  $\mathbf{u}_g$  é o vetor de efeitos genotípicos aleatórios de híbridos, com  $\mathbf{u}_g \sim N(0, \sigma_g^2 \mathbf{I}_g)$ ;  $\mathbf{u}_{ga}$  é o vetor de efeitos aleatórios da interação híbridos por locais, com  $\mathbf{u}_{ga} \sim N(0, \sigma_{ga}^2 \mathbf{I}_{ga})$ ;  $\mathbf{u}_b$  é o vetor de efeitos aleatórios de (design experimental DBC: repetição dentro de experimento e local; design experimental DBI: bloco dentro de repetição e ambiente, com  $\mathbf{u}_b \sim N(0, \sigma_b^2 \mathbf{I}_b)$ ;  $\mathbf{e}$  é o vetor de erros aleatórios, com  $\mathbf{e} \sim N(0, \mathbf{R})$ ;  $\mathbf{X}$ ,  $\mathbf{Z}_g$ ,  $\mathbf{Z}_{ga}$  e  $\mathbf{Z}_b$  são as matrizes de incidência associadas aos vetores  $\boldsymbol{\tau}$ ,  $\mathbf{u}_g$ ,  $\mathbf{u}_{ga}$  e  $\mathbf{u}_b$ ;  $\sigma_g^2$ ,  $\sigma_{ga}^2$ ,  $\sigma_b^2$  e  $\sigma_e^2$  são os componentes de variância associadas aos vetores  $\mathbf{u}_g$ ,  $\mathbf{u}_{ga}$ ,  $\mathbf{u}_b$  e  $\mathbf{e}$ ;  $\mathbf{I}_g$ ,  $\mathbf{I}_{ga}$ ,  $\mathbf{I}_b$  e  $\mathbf{I}_n$  são as matrizes identidade associadas aos vetores  $\mathbf{u}_g$ ,  $\mathbf{u}_{ga}$ ,  $\mathbf{u}_b$  e  $\mathbf{e}$ . A matriz de (co)variância residual, visando modelar o efeito de local dentro de cada safra, assumiu uma estrutura de variância bloco diagonal empregando a matriz identidade ( $\mathbf{R} = \bigoplus_{l=1}^t \sigma_{e_l}^2 \mathbf{I}_{n_l}$ ).

O método utilizado para estimar as variâncias foi o da máxima verossimilhança residual (REML) (PATTERSON; THOMPSON, 1971). A estimativa da herdabilidade no sentido amplo

em cada safra, foi obtida pela expressão apresentada por Falconer and Mackay (1996) utilizando os componentes de variância extraídos do modelo 2:

$$H_F^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{g \times a}^2}{l} + \frac{\sigma_e^2}{l \times r}}$$

Todas as análises fenotípicas foram realizadas utilizando o pacote ASReml-R (BUTLER et al., 2017).

## 2.2 Dados genotípicos

A genotipagem das 447 linhagens foi realizada seguindo o protocolo da empresa Diversity Arrays Technology utilizando 23.153 marcadores Darts. Todas as marcas com frequência mínima do alelo (MAF) menor que 5% foram descartadas. A imputação de dados de marcadores perdidos foi feita com base na média (AMADEU et al., 2016). Os 21.277 marcadores Darts resultantes foram utilizados para construir a matriz de parentesco genômico dos híbridos com base na combinação genotípica das duas linhagens parentais.

A matriz de marcadores moleculares das linhagens foi obtida utilizando código 1 para presença do alelo  $t$  na marca  $m$  na linhagem  $i$  e 0 para ausência do alelo  $t$ . Já a matriz de marcadores dos híbridos foi dada pela soma das matrizes de marcadores das linhagens parentais previamente genotipadas, sendo uma marca codificada em 0 na ausência dos alelos  $t$  em ambos os parentais (genótipo homocigoto recessivo –  $ab$ ), 1 na presença do alelo  $t$  em um parental e na ausência do alelo  $t$  em outro parental (genótipo heterocigoto -  $Ab, aB$ ) e 2 na presença do alelo  $t$  em ambos os parentais (genótipo homocigoto dominante -  $AB$ ).

## 2.3 Predição dos valores genéticos genômicos

A análise para obtenção dos valores genéticos genômicos preditos foi realizada em duas etapas devido à complexidade e a demanda computacional para ajustar os dados dos multi-ambientes em apenas uma etapa. Como previamente descrito, as informações fenotípicas foram corrigidas para efeito de local, experimento e design experimental. Posteriormente, os seguintes modelos foram testados:

a) Modelo GBLUP aditivo (GBLUP\_A):

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e},$$

em que  $\mathbf{y}$  é o vetor de observações fenotípicas ajustadas BLUE  $n \times 1$ , onde  $n$  é o número de observações fenotípicas ajustadas BLUE dos híbridos simples em cada safra;  $\mathbf{X}$  é a matriz de incidência dos efeitos fixos;  $\hat{\boldsymbol{\beta}}$  é o vetor de efeitos fixos;  $\mathbf{Z}$  é a matriz de incidência dos híbridos simples  $n \times n$ ;  $\boldsymbol{\alpha}$  é o vetor de efeitos aditivos  $m \times 1$ , em que  $m$  é o número de efeitos;  $\mathbf{e}$  é o vetor dos resíduos  $n \times 1$ .  $\mathbf{y}$ ,  $\boldsymbol{\alpha}$  e  $\mathbf{e}$  foram assumidos como:  $\mathbf{y} \sim \text{MVN}(0, V)$ ,  $\boldsymbol{\alpha} \sim \text{MVN}(0, G\boldsymbol{\alpha})$  e  $\mathbf{e} \sim \text{MVN}(0, \mathbf{R})$ , em que  $G\boldsymbol{\alpha} = A\hat{\sigma}_a^2$ ,  $\mathbf{R} = I\hat{\sigma}_e^2$ .  $I$  é uma matriz identidade  $n \times n$ ,  $A$  é a matriz de efeitos aditivos baseado na matriz de parentesco genômico ( $n \times n$ ), construída pelo método de VanRaden (VANRADEN, 2008).

b) Modelo GBLUP aditivo e dominante (GBLUP\_AD):

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\delta} + \mathbf{e},$$

$\mathbf{y}$ ,  $\mathbf{X}\hat{\boldsymbol{\beta}}$ ,  $\mathbf{Z}$  e  $\mathbf{e}$  são equivalentes aos termos descritos no modelo GBLUP aditivo. Semelhante a  $\boldsymbol{\alpha}$ ,  $\boldsymbol{\delta}$  corresponde ao vetor de efeito de dominância  $m \times 1$ ;  $\boldsymbol{\delta}$  é assumido  $\boldsymbol{\delta} \sim \text{MVN}(0, G\boldsymbol{\delta})$ , em que  $G\boldsymbol{\delta} = D\hat{\sigma}_d^2$ .  $D$  é a matriz de efeitos de dominância baseado na matriz de parentesco genômico ( $n \times n$ ), construída pelo método proposto por Vitezica et al. (2013).

As matrizes de parentesco genômico aditivo e de dominância foram obtidas pelas seguintes equações:

$$A = \frac{ZZ'}{2\sum_i^m p_i q_i} \quad D = \frac{WW'}{4\sum_i^m p_i^2 q_i^2},$$

Em que p e q são as frequências alélicas no loco i e Z é a matriz de incidência dos efeitos aditivos das marcas e W é a matriz de incidência dos efeitos de dominância das marcas. A matriz Z foi codificada como 0 para indivíduos homozigotos  $A_1A_1$ , 1 para o heterozigoto  $A_1A_2$  e 2 para o homozigoto  $A_2A_2$ .

Os componentes de variância aditiva ( $\sigma^2A$ ), de dominância ( $\sigma^2D$ ) e residual ( $\sigma^2e$ ) e seus respectivos erros padrões, foram estimadas por meio da regressão linear bayesiana generalizada. A herdabilidade no sentido amplo ( $H^2$ ) e no sentido restrito ( $h^2$ ) e a proporção da variância genética explicada pelos efeitos de dominância ( $d^2$ ) foram estimados pelas seguintes expressões:

$$H^2 = h^2 + d^2$$

$$d^2 = \frac{\sigma^2D}{\sigma^2A + \sigma^2D}$$

$$h^2 = \frac{\sigma^2A}{\sigma^2A + \sigma^2D + \sigma^2e}$$

Em que  $\sigma^2A$  é a variâncias aditiva;  $\sigma^2D$  é a variância de dominância;  $\sigma^2e$  é a variância residual ( $\sigma^2e$ ).

Os modelos GBLUP\_A e GBLUP\_AD, foram implementados no pacote BGLR do software R (PEREZ; DE LOS CAMPOS, 2014). Para isso foram utilizadas um total de 40000 interações, assumindo um período de “burn-in” de 10000 e um intervalo de amostragem (thin) de cinco interações. A convergência dos modelos foi verificada pelas autocorrelações e por meio de uma

análise visual gráfica das estimativas das variâncias aditiva, de dominância e residual obtidas em cada uma das interações (Material suplementar 2).

#### **2.4 Validação dentro e entre safras**

Os modelos descritos, foram comparados baseando-se em suas habilidades preditivas, calculada pela correlação entre os valores preditos e os valores fenotípicos observados. Dentro de cada safra, os modelos foram ajustados para a população de treinamento e os valores preditos foram correlacionados com os valores observados da população de validação. Para isso, o conjunto de híbridos avaliados em cada safra foi dividido em cinco grupos aleatórios. Quatro dos cinco grupos, foram utilizados como população de treinamento e o grupo remanescente como população de validação. Por meio desse esquema, os valores genéticos genômicos preditos foram obtidos para todos os indivíduos avaliados.

Entre as diferentes safras, a eficiência dos modelos em prever novos genótipos na safra  $j$  com base na avaliação fenotípica da safra  $i$ , foi verificada pela correlação direta entre os valores genéticos genômicos e os valores fenotípicos observados (EBLUP). Essa estratégia foi adotada como um esquema de validação real, uma vez que os híbridos já haviam sido avaliados ao longo das safras nos diferentes anos agrícolas. Para isso, três situações diferentes foram adotadas:

- i) Correlação entre os valores preditos e observados considerando todos os genótipos avaliados na safra  $j$  (E1);
- ii) Correlação entre os valores preditos e observados considerando apenas os genótipos comuns entre as safras  $i$  e  $j$  (E2).
- iii) Correlação entre os valores preditos e observados considerando apenas os genótipos não coincidentes entre as duas safras  $i$  e  $j$  (E3);

## 2.5 Medidas para verificar a eficiência dos modelos preditivos

Visando verificar a eficiência dos modelos GBLUP\_A e GBLUP\_AD em prever genótipos superiores entre e dentro das safras, foi estimado o ganho realizado com a seleção. Para isso, realizou-se o ranqueamento dos 10% melhores genótipos preditos na safra i considerando os valores genéticos genômicos preditos e posteriormente a seleção dos mesmos 10% na safra j com base em suas estimativas BLUP. O ganho realizado em porcentagem foi dado pela seguinte expressão:

$$GR(\%) = \frac{\overline{BLUP_{IJ}}}{\mu_{fenotípica}} * 100$$

Em que  $\overline{BLUP_{IJ}}$  é a média dos BLUP's dos genótipos selecionados na safra j com base no ranqueamento dos mesmos genótipos na safra i a partir dos valores genético genômicos preditos.  $\mu_{fenotípica}$  média fenotípica da safra j.

A eficiência dos modelos preditivos ainda foi verificada ao comparar os resultados obtidos pela estratégia E2 (utilizando dados genômicos) com: a) correlação fenotípica entre as médias dos HS coincidentes nas safras i e j; b) estimativa do ganho realizado considerando o desempenho fenotípico dos 10% melhores genótipos safra i e seleção dos mesmos 10% na safra j com base em suas estimativas BLUP.

### 3 RESULTADOS

#### 3.1 Parâmetros genéticos e dados genômicos

A análise fenotípica do presente conjunto de dados foi realizada amplamente no capítulo anterior, visando entender melhor o que acontece entre e dentro das nove safras avaliadas a nível de desbalanceamento e da interação genótipos por ambientes. Alguns pontos relevantes foram comentados novamente nessa secção, que buscou dar maior ênfase na análise e interpretação dos dados genotípicos.

Foram observadas diferenças significativas para os efeitos de híbridos, locais, interação híbridos por locais pelo teste de Wald ( $P < 0,05$ ) e pelo teste da máxima verossimilhança residual ( $P < 0,05$ ), para produtividade de grãos em todas as safras. A média fenotípica ao longo das nove safras avaliadas, esteve no intervalo e 3.9 a 9.7 ( $t \cdot ha^{-1}$ ). A estimativa da herdabilidade no sentido amplo foi de média magnitude entre as safras, variando de 0,43 a 0,70, refletindo uma boa confiabilidade das avaliações fenotípicas. A variância aditiva estimada em cada safra, variou de 0,11 a 1,99, enquanto a variância de dominância variou de 0,16 a 1,16. A proporção da variância genética total explicada pelos efeitos de dominância foi superior a contribuição dos efeitos aditivos para a maioria das safras (Tabela 1).

As herdabilidades do sentido amplo e restrito variaram consideravelmente ao longo dos anos e das diferentes épocas de semeadura. As estimativas de  $h^2$  e  $H^2$  da primeira safra foram sempre superiores as da segunda safra para ambas regiões avaliadas, com exceção da safra 2014s2C cuja  $h^2$  superou a estimativa da safra 2013s1C. Considerando o modelo GBLUP\_AD, verificou-se um decréscimo considerável na estimativa da variância aditiva e conseqüentemente da  $h^2$ , essa



redução foi de até 40% quando comparado ao modelo considerando apenas efeitos aditivos (GBLUP\_A) (Tabela 1).

Tabela 1: Estimativas dos parâmetros genéticos obtidos via regressão linear bayesiana generalizada, para as nove safras avaliadas, incluindo efeitos aditivos (GBLUP\_A) e aditivo-dominante (GBLUP\_AD).

Safra	GBLUP_A			GBLUP_AD						
	$\sigma^2 A^*$	$\sigma^2 e$	$h^2$	$\sigma^2 A$	$\sigma^2 D$	$\sigma^2 e$	$h^2$	$H^2$	$d^2$	$H_F^2$
2011s1C	0,96 (0,31)	4,70 (0,21)	0,17	0,65 (0,27)	0,41 (0,11)	4,65 (0,22)	0,11	0,18	0,07	0,63
2011s1S	0,56 (0,18)	3,63 (0,19)	0,13	0,34 (0,15)	0,70 (0,21)	3,23 (0,21)	0,08	0,24	0,16	0,57
2012s2C	0,18 (0,06)	1,61 (0,07)	0,10	0,12 (0,05)	0,16 (0,04)	1,55 (0,07)	0,07	0,15	0,08	0,46
2012s2S	0,17 (0,05)	1,42 (0,06)	0,11	0,11 (0,04)	0,21 (0,06)	1,32 (0,07)	0,06	0,20	0,12	0,62
2012s1C	2,57 (0,79)	4,80 (0,35)	0,34	1,99 (0,69)	1,16 (0,40)	4,29 (0,37)	0,27	0,42	0,15	0,64
2013s2C	0,57 (0,21)	2,73 (0,21)	0,17	0,38 (0,17)	0,45 (0,17)	2,57 (0,21)	0,11	0,24	0,13	0,70
2012s1S	0,50 (0,17)	3,92 (0,22)	0,11	0,31 (0,13)	0,67 (0,21)	3,56 (0,22)	0,07	0,22	0,15	0,53
2013s1C	0,46 (0,14)	2,80 (0,15)	0,14	0,31 (0,12)	0,43 (0,14)	2,59 (0,16)	0,09	0,22	0,13	0,67
2014s2C	0,66 (0,18)	3,08 (0,16)	0,18	0,42 (0,16)	0,48 (0,15)	2,86 (0,16)	0,11	0,24	0,13	0,67

\*Variâncias aditiva ( $\sigma^2 A$ ), de dominância ( $\sigma^2 D$ ) e residual ( $\sigma^2 e$ ) e seus respectivos erros padrão (em parêntese). Herdabilidade fenotípica ( $H_F^2$ );  $h^2$  herdabilidade no sentido restrito;  $H^2$  herdabilidade no sentido amplo;  $d^2$  porção da variância genética total explicada pelos efeitos de dominância.

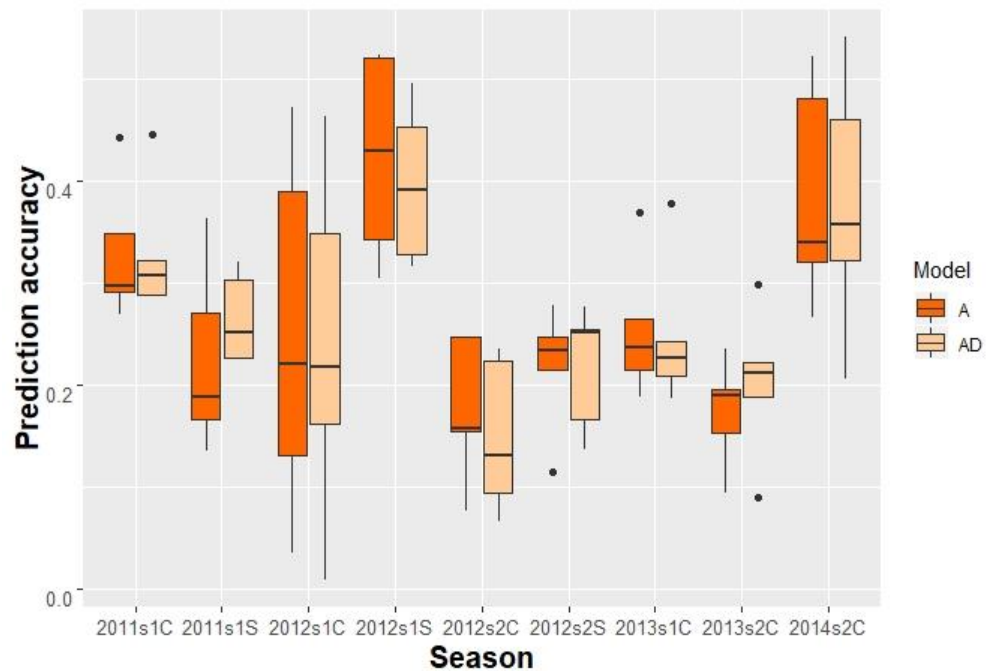
Fonte : Do autor (2020)

A matriz de marcadores dos híbridos foi obtida por meio da combinação dos alelos das linhagens parentais e a imputação dos marcadores perdidos foi realizada por meio da média. Os dados de marcadores foram filtrados considerando uma MAF mínima de 0,05, resultando na eliminação de 1876 marcadores DARTs do painel. As marcas resultantes foram utilizadas para obtenção das matrizes de parentesco genômico aditivo e de dominância.

### 3.2 Predição genômica

A herdabilidade genômica no sentido amplo foi em média de baixa magnitude entre as safras, sendo o maior valor observado de 0,42 na safra 2012s1C (Tabela 1). As estimativas da acurácia preditiva para os modelos GBLUP\_A e GBLUP\_AD obtidos dentro de cada safra utilizando o esquema de validação cruzada, variaram consideravelmente entre os anos/época de semeadura. Foram observados valores nos intervalos de 0,17 a 0,42 (GBLUP\_A) e de 0,15 a 0,39 (GBLUP\_AD) (Figura 2).

Figura 1: Boxplot das acurácias de predição dentro de cada safra para os modelos aditivo (A) e aditivo-dominante (AD).



Fonte: Do autor (2021).

A predição de híbridos entre as safras foi realizada considerando três estratégias de validação (E1, E2 e E3). A predição dos híbridos foi realizada a princípio dentro de mesma região e época de semeadura e posteriormente dentro da mesma região, porém considerando diferentes épocas de semeadura. Essa estratégia foi adotada pois a correlação fenotípica entre híbridos

coincidentes ao longo das diferentes safras (épocas de semeadura, região e ano) foi de baixa magnitude, principalmente entre regiões e épocas de semeadura diferentes, variando de -0.01 a 0.51.

Para as predições realizadas dentro de uma mesma região e época de semeadura, o modelo GBLUP\_AD foi superior ao modelo GBLUP\_A em 60% dos casos analisados e semelhante a ele em 16% das situações (Tabelas 2 e 3). Dentro da região centro considerando a mesma época de semeadura, as correlações entre os valores preditos e observados tiveram uma amplitude de variação de 0,09 a 0,43 para o modelo GBLUP\_A e de 0,13 a 0,45 para o modelo GBLUP\_AD. A estratégia de validação E2, em que apenas os genótipos comuns as safras i e j foram levados em consideração para estimar a habilidade preditiva do modelo, foi superior as estratégias E1 e E3 (Tabela 2).

Tabela 2: Correlação entre os valores preditos e observados em diferentes anos, obtidos considerando os modelos GBLUP\_A e GBLUP\_AD (em negrito), para as primeiras e segundas safras da região centro, considerando as estratégias de validação E1, E2 e E3 e correlação fenotípica entre as médias dos HS coincidentes nas safras i e j (em itálico).

	1ª Safra		Estratégias de validação	2ª Safra	
	2012s1C	2013s1C		2013s2C	2014s2C
2011s1C	0,3532 <b>0,3569</b>	0,1969 <b>0,2205</b>	E1	2012s2C 0,2375 <b>0,2557</b>	0,2269 <b>0,2539</b>
2012s1C	- -	0,2324 <b>0,2262</b>		2013s2C - <b>0,1596</b>	0,2028 <b>0,1596</b>
2011s1C	0,4311 <b>0,4578</b>	0,2263 <b>0,2318</b>	E2	2012s2C 0,2525 <b>0,2960</b>	0,2744 <b>0,2975</b>
2012s1C	<i>0,2002</i> - -	<i>0,3417</i> 0,0964 <b>0,1177</b>		2013s2C - <b>0,2885</b>	<i>0,0901</i> 0,1585 <b>0,2885</b>
2011s1C	0,3128 <b>0,3045</b>	0,1892 <b>0,2039</b>	E3	2012s2C 0,2037 <b>0,2153</b>	0,1939 <b>0,2118</b>
2012s1C	- -	0,2318 <b>0,2147</b>		2013s2C - <b>0,1359</b>	0,1807 <b>0,1359</b>

Fonte: Do autor (2021).

As correlações obtidas para região sul dentro de uma mesma época de semeadura, foram em média de menor magnitude do que as obtidas na região centro, variando de -0,008 a 0,31 para o modelo GBLUP\_A e de 0,02 a 0,29 para o modelo GBLUP\_AD. Para essa região, o modelo GBLUP\_AD não se mostrou superior ao modelo GBLUP\_A. Semelhante ao observado na região centro, a estratégia de validação E2 foi a que propiciou maior correlação entre os valores preditos e observados, para ambos modelos (Tabela 3).

Tabela 3: Correlação entre os valores preditos e observados em diferentes anos, obtidos considerando os modelos GBLUP\_A e GBLUP\_AD (em negrito), para as primeiras e segundas safras e entre as primeiras e segundas safras da região sul, considerando as estratégias de validação E1, E2 e E3 e correlação fenotípica entre as médias dos HS coincidentes nas safras i e j (em itálico).

	2012s2S	2012s1S	Estratégia de validação
2011s1S	-0,0211 <b>0,0641</b>	0,1129 <b>0,1028</b>	E1
2012s2S	- -	0,2888 <b>0,2894</b>	
2011s1S	-0,0088 <b>0,0996</b> <i>0,2745</i>	0,2465 <b>0,2830</b> <i>0,4079</i>	E2
2012s2S	- -	0,3139 <b>0,2714</b> <i>0,2157</i>	
2011s1S	-0,0312 <b>0,0235</b>	0,0673 <b>0,0382</b>	E3
2012s2S	- -	0,2904 <b>0,2955</b>	

Fonte: Do autor (2021).

A habilidade preditiva dos modelos GBLUP\_A e GBLUP\_AD também foi verificada entre as safras de uma mesma região, considerando diferentes épocas de semeadura (s1 e s2) (Tabelas 2 e 4). A estratégia de validação E2 foi a que resultou em habilidades preditivas de maior magnitude para região centro (Tabela 4). Para região sul, foram observadas estimativas próximas a zero ou

negativas, para todas as estratégias de validação adotadas considerando diferentes épocas de semeadura (Tabela 3).

Tabela 4: Correlação entre os valores preditos e observados em diferentes anos, obtidos considerando os modelos GBLUP\_A e GBLUP\_AD (em negrito), entre as primeiras e segundas safras da região centro, considerando as estratégias de validação E1, E2 e E3 e correlação fenotípica entre as médias dos HS coincidentes nas safras i e j (em itálico).

1ª safra predizendo 2ª safra			2ª safra predizendo 1ª safra				
	2012s2C	2013s2C	2014s2C		2012s1C	2013s1C	
2011s1C	0,1989	0,5010	0,2826	E1	2012s2C	0,1117	0,2250
	<b>0,2152</b>	<b>0,4653</b>	<b>0,3265</b>			<b>0,0989</b>	<b>0,2185</b>
2012s1C	-	0,2893	0,265	E1	2013s2C	-	0,0388
	-	<b>0,2546</b>	<b>0,2714</b>			-	<b>-0,0057</b>
2013s1C	-	-	0,1714	E1	-	-	-
	-	-	<b>0,1551</b>			-	-
2011s1C	0,2196	0,4839	0,3193	E2	2012s2C	0,0364	0,2549
	<b>0,2495</b>	<b>0,5042</b>	<b>0,3410</b>			<b>-0,0142</b>	<b>0,241</b>
2012s1C	<i>0,3623</i>	<i>0,4754</i>	<i>0,2085</i>	E2	2013s2C	<i>-0,0270</i>	<i>0,1673</i>
	-	0,3504	0,2656			-	0,2352
2013s1C	-	<b>0,3482</b>	<b>0,2705</b>	E2	2013s2C	-	<b>0,1585</b>
	-	<i>0,5097</i>	<i>0,2398</i>			-	<i>0,1802</i>
2011s1C	-	-	0,2792	E3	-	-	-
	-	-	<b>0,2586</b>			-	-
2012s1C	0,1528	0,4572	0,2671	E3	2012s2C	0,146	0,2102
	<b>0,1285</b>	<b>0,3951</b>	<b>0,3159</b>			<b>0,1574</b>	<b>0,2007</b>
2013s1C	-	0,2723	0,2495	E3	2013s2C	-	0,1807
	-	<b>0,2238</b>	<b>0,253</b>			-	<b>0,1359</b>
2011s1C	-	-	0,0856	E3	-	-	-
	-	-	<b>0,0643</b>			-	-

Fonte: Do autor (2021).

As estimativas de correlação entre os valores preditos e observados entre e dentro de safras foram de baixa magnitude em todas as estratégias adotadas, porém condizentes com as estimativas da herdabilidade em cada safra. As maiores correlações foram observadas nas safras em que a herdabilidade de fenotípica foi de maior magnitude 0,64 (2012s1C) e 0,67 (2013s2C) (Tabelas 1, 2 e 4). De maneira geral, a estratégia de validação E1, onde a correlação entre valores preditos e observados foi realizada considerando todos os genótipos avaliados na safra j, retornou estimativas

semelhantes a estratégia E2. Por outro lado, como esperado a estratégia E3 foi a que resultou, em média, nas menores estimativas de correlação entre valores preditos e observados, evidenciando a dificuldade de prever híbridos em diferentes anos, principalmente sem considerar o efeito ambiental no modelo preditivo. (Tabelas 2, 3 e 4).

Para verificar a relevância dos modelos GBLUP\_A e GBLUP\_AD, a capacidade preditiva dos modelos, obtidas pela estratégia E2, foi comparada com a correlação fenotípica entre a média dos HS comuns as safras i e j. Para ambas as regiões e épocas de semeadura, foi observado que os modelos GBLUP, superaram o valor fenotípico real observado em 10 das 18 combinações de safras analisadas. Surpreendentemente, em algumas situações, o modelo preditivo foi 50 % (2011s1C/2012s1C) e até 70% (2011s1C/2014s2C) superior ao valor fenotípico observado (Tabelas 2, 3 e 4).

A eficiência dos modelos preditivos GBLUP\_A e GBLUP\_AD também foi verificada pela estimativa do ganho realizado com a seleção (GR). O modelo GBLUP\_A se mostrou superior ao modelo GBLUP\_AD para as análises realizadas dentro de cada safra. Ao contrário disso, quando as análises envolveram diferentes safras e épocas de semeadura, o modelo GBLUP\_AD foi considerado mais eficiente para prever genótipos superiores. As estimativas do GR variaram de 0 a 10,72 %, indicando que é possível obter ganhos com a seleção, baseando-se apenas na predição de HS não avaliados. As maiores estimativas do GR com a seleção foram obtidas entre safras considerando diferentes épocas de semeadura. Em algumas situações o GR foi negativo, ou seja, as predições obtidas pelos modelos GBLUP\_A e GBLUP\_AD foram ineficientes para selecionar HS mais produtivos (Figura 3).

Figura 2: Ganho realizado com a seleção (%), obtidos dentro de cada safra e entre safras, para os modelos GBLUP\_A e GBLUP\_AD (em negrito) e ganho realizado real com a seleção (%), obtidos entre as safras, a partir das estimativas BLUP dos HS avaliados em cada safra (em itálico), considerando a mesma época de semeadura e diferentes épocas de semeadura.

	2011s1C	2012s2C	2012s1C	2013s2C	2013s1C	2014s2C		2011s1S	2012s2S	2012s1S	
<b>2011s1C</b>	2.22	1.15	5.62	5.99	3.53	6.34		<b>2011s1S</b>	1.41	0.67	-0.45
	<b>1.96</b>	<b>1.24</b>	<b>7.34</b>	<b>6.79</b>	<b>7.10</b>	<b>10.72</b>			<b>1.56</b>	<b>0.95</b>	<b>-0.15</b>
	<i>4.86</i>	<i>4.05</i>	<i>4.20</i>	<i>8.58</i>	<i>7.55</i>	<i>5.63</i>			<i>8.33</i>		
<b>2012s2C</b>	0.39			-3.31		2.03		<b>2012s2S</b>	3.31	4.40	
	<b>0.34</b>			<b>1.38</b>		<b>2.92</b>			<b>3.47</b>	<b>3.21</b>	
<b>2012s1C</b>			5.25	7.17	4.25	4.51		<b>2012s1S</b>		-0.14	
			<b>3.97</b>	<b>5.47</b>	<b>3.52</b>	<b>3.54</b>			<b>-0.23</b>		
			<i>9.49</i>	<i>8.77</i>	<i>9.06</i>	<i>11.78</i>					
<b>2013s2C</b>				4.96		0.11					
				<b>5.54</b>		<b>1.70</b>					
<b>2013s1C</b>											
					-0.49	-0.10					
					<b>0.87</b>	<b>0.16</b>					
<b>2014s2C</b>						5.55					
						<b>5.21</b>					

	Dentro da safra
	Entre safras/mesma época de plantio
	Entre safras/diferentes épocas de plantio

Fonte: Do autor (2021)

As estimativas de ganho realizado obtidas a partir das predições dos modelos GBLUP, foram comparadas com o ganho real observado entre as safras. Na maioria das combinações de safras que foram avaliadas, o ganho real foi superior ao GR obtido por meio dos valores preditos. Apesar disso, foi possível verificar combinações (2011s1C/2012s1C, 2011s1C/2013s2C, 2011s1C/2013s1C, 2011s1C/2014s2C, 2012s2C/2013s2C, 2012s2S/2012s1S) em que os modelos preditivos GBLUP foram mais eficientes, superando o ganho real estimado apenas com base na informação fenotípica (Figura 3).

## 4 DISCUSSÃO

Na última década, a seleção genômica passou a ser adotada frequentemente, em programas de melhoramento de milho, como uma ferramenta vantajosa que permite a redução do tempo e dos recursos dispendidos na seleção e recomendação de híbridos superiores (BERNARDO, 2009; DOS SANTOS et al., 2016; CANTELMO et al. 2017; LYRA et al. 2017; DIAS et al. 2018; FRITSCHENETO et al. 2018; DE OLIVEIRA et al., 2020; KRAUSE et al., 2020). Neste estudo, a seleção genômica foi aplicada visando comparar a habilidade preditiva e a eficiência dos modelos GBLUP, em prever híbridos simples não avaliados, considerando os efeitos aditivos e não aditivos, em diferentes cenários ao longo de quatro anos.

O conjunto de dados em estudo caracteriza perfeitamente o cenário das empresas produtoras de sementes híbridas de milho, uma vez que uma infinidade de combinações híbridas foi testada em diferentes anos e épocas de semeadura, com baixa repetibilidade e coincidência desses genótipos ao longo dos ambientes avaliados. Apesar de refletir as situações práticas que ocorrem no dia a dia, esse cenário de alto desbalanceamento somado a significativa interação genótipos x ambientes, pode comprometer a estimativa dos componentes de variância, da herdabilidade, dos valores genéticos dos indivíduos, como demonstrado no capítulo anterior.

A inclusão de efeitos de dominância em modelos preditivos é de grande importância para programas de melhoramento cujo foco principal são populações de polinização cruzada e/ou produção de híbridos (ZENG et al. 2013; NISHIO; SATOH 2014; DIAS et al., 2018; BERNARDO, 2020). Neste trabalho, verificou-se que a inclusão dos efeitos de dominância no modelo resultou na redução da herdabilidade no sentido restrito. Resultados semelhantes foram observados por Muñoz et al. (2014), Kumar et al. (2015) e Dias et al. (2018).



A redução na estimativa da herdabilidade no sentido amplo era esperada pois em espécies alógamas como milho, sabe-se que a variância de dominância é responsável por grande parte da variação genética observada (HALLAUER; MIRANDA FILHO, 1998; BERNARDO, 2020). A ausência dos efeitos de dominância no modelo resulta em estimativas superestimadas como as observadas (Tabela 1). De acordo com Bernardo (2020), obter uma estimativa mais acurada da  $h^2$  é importante pois os efeitos de dominância e epistáticos que são capturados pela  $H^2$  são considerados incoerentes, uma vez que eles não são transmitidos dos parentais para a descendência.

A superioridade da  $\sigma^2D$  em relação a  $\sigma^2A$  observada pode ser um indicativo da presença de sobredominância ou de que a frequência alélica dos locos responsáveis pela manifestação fenotípica do caráter em estudo, seja extrema. A menos que essas condições sejam observadas, a  $\sigma^2A$  será superior a  $\sigma^2D$ , uma vez que a  $\sigma^2D$  corresponde a menor variação possível entre os valores genotípicos (BERNARDO, 2020).

De acordo com a teoria da genética quantitativa, tanto a  $\sigma^2A$  quanto a  $\sigma^2D$  são função dos efeitos de dominância (FISHER, 1918):  $\sigma_A^2 = \sum_k 2p_k(1 - p_k)[a_k + d_k(q_k - p_k)]^2$ ;  $\sigma_D^2 = \sum_k (2p_k(1 - p_k)d_k)^2$ . Dependendo das frequências alélicas, parte dos efeitos de dominância podem ser estimados como variância aditiva, mesmo na presença de efeitos não aditivos (HILL et al., 2008). Portanto, é esperado que ambas as variâncias exibam forte confundimento, principalmente quando uma delas não estiver presente no modelo preditivo (DOS SANTOS et al., 2016).

A inclusão de efeitos não aditivos em modelos de predição, tem sido extensivamente estudada nos últimos anos. De acordo com a literatura, considerar efeitos de dominância para predição de genótipos não avaliados, tem se mostrado uma prática vantajosa para programas de melhoramento cuja cultura em estudo permite explorar os benefícios da heterose. As vantagens de

se considerar os efeitos não aditivos nos modelos preditivos são mais evidentes quando a herdabilidade do caráter em estudo é baixa (DA et al., 2014; MUÑOZ et al., 2014; DOS SANTOS et al., 2016; VIEIRA et al., 2017; DIAS et al., 2018; LIU; CHEN, 2018; LIU et al., 2019).

No presente trabalho, a capacidade preditiva dos modelos GBLUP foi verificada em dois cenários distintos, dentro de cada safra envolvendo os locais avaliados em uma mesma época de semeadura, e entre safras, onde os fatores ano e/ou diferentes épocas de semeadura, foram adicionados. Dentro de cada safra, a capacidade preditiva dos modelos foi de baixa magnitude. Esse fato era esperado pois a produtividade de grãos é um caráter controlado por muitos genes e altamente influenciado pelo ambiente, estando sempre associado a baixas estimativas da herdabilidade, como aqui demonstrado. Apesar disso, os resultados observados no presente estudo coincidem com os apresentados recentemente por Dias et al. (2018) e de Oliveira et al. (2020).

A seleção de híbridos adaptados a mega ambientes é uma prática que já vem sendo bastante adotada ao longo dos anos para as condições brasileiras de cultivo. Por isso, as análises entre safras foram realizadas a princípio para cada região (Centro/Sul) e época de semeadura (S1 e S2) e posteriormente envolvendo diferentes épocas de semeadura. Essa estratificação se deu, pois, as condições ambientais entre as duas regiões e época de semeadura são muito variáveis em relação a solo, temperatura, umidade, fotoperíodo e precipitação.

Similar ao observado em cada safra, a capacidade preditiva dos modelos GBLUP\_A e GBLUP\_AD também foi de baixa magnitude entre as safras. A estrutura da população, pode ser considerada um dos fatores responsáveis pelas baixas acurácias preditivas observadas, principalmente em cenários em que as análises envolveram diferentes safras e uma grande diversidade de híbridos. Outros fatores que também podem afetar acurácia preditiva dos modelos são o tamanho da população, a densidade de marcadores, a herdabilidade, os modelos estatísticos

e a similaridade entre as populações de treinamento e validação (SCHOPP et al., 2017; ZHANG et al., 2017a; CANTELMO et al., 2017; DE OLIVEIRA et al., 2018; DIAS et al., 2018; ZHANG et al., 2019).

De maneira geral, o modelo GBLUP\_AD foi superior ou retornou estimativas semelhantes ao modelo GBLUP\_A em aproximadamente 50% condições avaliadas. Dias et al. (2018) também verificou que efeitos não aditivos tem uma importante contribuição para predição genômica. Segundo esses autores, a inclusão de efeitos de dominância no modelo melhora consideravelmente a seleção de híbridos não testados para tolerância a seca. De acordo com relatos da literatura, considerar efeitos de dominância no modelo estatístico pode incrementar a correlação entre os valores preditos e ausentes em até 20% e conseqüentemente acelerar o processo de seleção nos programas de melhoramento (DA et al., 2014; DOS SANTOS et al., 2018; LIU et al., 2019).

Dentre os cenários que foram analisados, a estratégia de validação que considerou apenas os híbridos coincidentes entre as safras (E2), resultou em maiores acurácias preditivas, quando comparado com E1 e E3. Apesar da magnitude das acurácias preditivas terem sido baixas, mesmo com a inclusão do efeito de dominância no modelo estatístico, os resultados obtidos utilizando essa estratégia foram encorajadores, uma vez que em muitas situações a correlação entre os valores preditos e observados foi superior a correlação fenotípica entre as médias dos HS coincidentes nas safras avaliadas. Em outras palavras, cabe dizer que o emprego da seleção genômica para predição de híbridos não avaliados, superou a seleção com base na avaliação fenotípica em muitas situações.

A eficiência da seleção genômica também foi comprovada ao comparar o ganho realizado com a seleção utilizando a informação genômica e o ganho real com base apenas na avaliação fenotípica. Surpreendentemente, dentro das safras o modelo GBLUP\_A possibilitou maiores ganhos, apesar disso, nas análises entre as safras que envolveram efeitos de época de semeadura e

ano, o modelo GBLUP\_AD foi superior. Em muitos cenários, a seleção genômica retornou ganhos maiores que a seleção fenotípica, contudo, com base nos resultados observados ainda resta a dúvida se o emprego da seleção genômica pode de fato substituir a avaliação fenotípica de híbridos de milho.

É claro que os recursos salvos com a predição de híbridos em um único ambiente, podem ser utilizados para gerar novas populações, obter dados genotípicos de linhagens para prever o desempenho de novos híbridos, avaliar mais genótipos em estágios avançados do processo de melhoramento e conseqüentemente aumentar os ganhos com a seleção ou simplesmente reduzir os recursos despendidos nos programas de melhoramento (KRAUSE et al., 2020). De acordo com a Krchov and Bernardo (2015), a redução do trabalho dispendido com a fenotipagem pode levar a uma melhor qualidade dos dados coletados a campo e potencializar a eficiência da seleção.

Com base nas análises dos dados fenotípicos desse trabalho, verificou-se um significativo efeito da interação genótipo por ambientes, que foi confirmado pelas baixas estimativas de correlação entre os anos safras. É importante comentar que o efeito da interação é resultado da manifestação dos alelos ora como favorável em um ambiente e ora como desfavorável em outro ambiente. Portanto, a magnitude das correlações entre os ambientes é também outro fator a ser considerado nos modelos preditivos, o qual pode influenciar diretamente a habilidade dos modelos de emprestar informações de ambientes correlacionados (KRAUSE et al., 2020).

No presente estudo, o efeito da interação genótipos por ambientes não foi incluso nos modelos estatísticos. Contudo, sabe-se que esse efeito pode afetar consideravelmente a eficiência da seleção genômica. Estudos recentes apontam as vantagens de se incluir as matrizes de correlações ambientais nos modelos preditivos, dentre elas está o incremento da habilidade de

prever genótipos em ambientes não avaliados (LÓPEZ-CRUZ et al., 2015; CUERVAS et al., 2016; LADO et al., 2016; LYRA et al. 2017; DE OLIVEIRA et al., 2020; KRAUSE et al., 2020).

De acordo com Dias et al. (2018), modelar todos os efeitos de locais e anos quando os híbridos são avaliados em estágios avançados dos programas de melhoramento, pode ser um problema. Uma opção seria realizar as análises em duas etapas, primeiro ajustando o efeito de cada ambiente, como aqui realizado, e posteriormente combinar as médias estimadas em uma segunda análise maior que inclua as correlações ambientais. Os resultados dessas análises têm apontado que mesmo em condições de alto nível de desbalanceamento dos dados, modelos que capturam a correlação entre ambientes e a informação genômica podem ser uma ferramenta valiosa para prever o valor genético de indivíduos não testados (DE OLIVEIRA et al., 2020; KRAUSE et al., 2020)

As respostas encontradas neste estudo com a inclusão dos efeitos de dominância foram encorajadoras e demonstram como a seleção genômica pode contribuir para reduzir custo e acelerar o processo de melhoramento. Apesar disso, em algumas situações não foram observadas diferenças significativas entre as estimativas com os dados fenotípicos reais e a predição genômica (modelos GBLUP\_A e GBLUP\_AD), fato que pode ser justificado pelas baixas estimativas da herdabilidade e das correlações entre os ambientes. Sendo assim, é de grande importância que estudos futuros com o mesmo conjunto de dados, inclua os efeitos ambientais nos modelos preditivos, por meio de artifícios que requerem uma menor demanda computacional como a realização das análises em duas etapas, uso de estruturas de fator analítico, “deep learning” em modelos de predição genômica multivariado, entre outros (MONTESINOS-LÓPEZ et al., 2018; LIU et al., 2019; DIAS et al., 2018; DE OLIVEIRA et al., 2020; KRAUSE et al., 2020). Com a adoção dessas estratégias, espera-

se uma melhora na capacidade preditivas dos modelos, principalmente em situações que as análises envolvam a predição de híbridos entre diferentes épocas de semeadura e anos.

## 5 REFERÊNCIAS

AMES Nicholas C.; BERNARDO, Rex. 2020. Genomewide predictions as a substitute for a portion of phenotyping in maize. **Crop Science**, Saint Paul, p. 1-9, 2020.

AMADEU R. R et al. AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: a Blueberry Example, **Plant Genome**, v.9, p.1–10, 2016.

BERNARDO, R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. **Heredity**, v.125, p.375–385, 2020.

BERNARDO R **Breeding for quantitative traits in plants**, 3rd edn. Stemma Press, Woodbury, Minnesota, 2020.

BERNARDO, R. Genomewide selection for rapid introgression of exotic germplasm in maize, **Crop Science**, v.49, p.419–425, 2009.

BUTLER, D. G. et al. **ASReml-R Reference Manual** Version 4. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, 2017.

CANTELMO N. F.; VON PINHO R. G.; BALESTRE M. Genomic analysis of maize lines introduced in the early stages of a breeding programme. **Plant Breeding**, v.136(6), p.845–860, 2017.

CANTELMO N. F.; VON PINHO R. G.; BALESTRE M. Genome-wide prediction for maize single-cross hybrids using the GBLUP model and validation in different crop seasons. **Molecular Breeding**, v.37, p.1-13, 2017.

CUEVAS J. et al. Genomic prediction of genotype X environment interaction kernel regression models, **Plant genome**, v.9, p. 1-20, 2016.

DA Y et al. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. **PLoS One**, v.9, p.1-11, 2014.

DESTA, Zeratsion Abera; ORTIZ, Rodomiro. Genomic selection: genome-wide prediction in plant improvement. **Trends in plant science**, v.19, p. 592-601, 2014.

DE OLIVEIRA, A.A et al. Genomic prediction applied to high-biomass sorghum for bioenergy production. **Molecular Breeding**, v.38, p.38-49, 2018.

DE OLIVEIRA, A.A. et al. Genomic prediction applied to multiple traits and environments in second season maize hybrids. **Heredity**, v.125, p.60–72 , 2020.

DIAS KODG et al. Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. **Heredity** v.121, p.24–37, 2018.

DOS SANTOS JPR et al. Inclusion of dominance effects in the multivariate GBLUP model. **PLoS ONE**, v.11, p.1–21, 2016

FALCONER, D.S.; MACKAY, T.F.C. **Introduction to quantitative genetics**. 4<sup>a</sup> Ed. London: Longman. 463 p., 1996.

FERRÃO, L.F.V. et al. Integration of Dominance and Marker x Environment Interactions into Maize Genomic Prediction Models. **Crop Science**, 2020.

FISHER RA **The correlation between relatives on the supposition of Mendelian inheritance**. *Trans R Soc Edinb* 52:399–433, 1918.

FRITSCHÉ-NETO R; AKDEMIR D; JANNINK JL Accuracy of genomic selection to predict maize single-crosses obtained through different mating designs. **Theoretical Applied Genetics**, v.131, p.1153–1162, 2018.

HALLAUER, A.R; MIRANDA FILHO, J.B. **Quantitative Genetics in maize breeding**. Ames: Iowa State University Press, 2 edition, 1998.

HILL W; GODDARD M; VISSCHER P Data and theory point to mainly additive genetic variance for complex traits. **PLoS Genetics**, v.4, p.1–10, 2008.

KRAUSE, M. D. et al. 2020. Boosting predictive ability of tropical maize hybrids via genotype by environment interaction under multivariate GBLUP models. **Crop Science**, v.60, p.3049-3065, 2020.

KRCHOV, L.-M.; BERNARDO Relative efficiency of genomewide selection for testcross performance of doubled haploid lines in a maize breeding program. **Crop Science**, v.55(5), p.2091–2099, 2015.

KUMAR S et al. Genome-enabled estimates of additive and non-additive genetic variances and prediction of apple phenotypes across environments. **G3 Genes Genomes Genetics**, v.5, p.2711–2718, 2015.

LADO, B. et al. (2016). Modeling genotype X environment interaction for genomic selection with unbalanced data from a wheat breeding program. **Crop Science**, v.56, p.1–15, 2-16.

LIU, H.; CHEN, GB. A new genomic prediction method with additive-dominance effects in the least-squares framework. **Heredity**, v.121, p.196–204, 2018.

LYRA, D.H et al. Multi-trait genomic prediction for nitrogen response indices in tropical maize hybrids. **Molecular Breeding**, v.37, p.1–14, 2017.

LÓPEZ-CRUZ, M. A. et al. Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. **G3 Genes Genomes Genetics**, v.5, p.569-582, 2015.

MEUWISSEN, T. Genomic selection: marker assisted selection on a genome wide scale. **Journal of animal Breeding and genetics**, v.124, p.321-322, 2007.

MONTESINOS-LÓPEZ OA et al. Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. **G3 Gene Genome Genetics**, v.8, p.3829–3840, 2018.



MUÑOZ, P.R. et al. Unraveling additive from nonadditive effects using genomic relationship matrices. **Genetics** 198: 1759–1768. doi: 10.1534/genetics.114.171322, 2014.

NISHIO, M.; SATOH M. Including Dominance Effects in the Genomic BLUP Method for Genomic Evaluation. **PLoS One**, v.9, 2014.

PATTERSON, H.D.; THOMPSON, R. Recovery of interblock information when block sizes are unequal. **Biometrika**, v.58, p.545-554, 1971.

PEREZ P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, 198(2):483-U463, 2014.

SCHOPP, P. et al. Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium. **Genetics**, v.205, p.441–454, 2017.

SHULL GH **Beginnings of the Heterosis concept**. In: Gowen JW (ed) *Heterosis*, Chap 2. Iowa State College Press, Ames, Iowa, pp 14–48. 1952.

TECHNOW F et al. Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. **Genetics**, v.197, p.1343–1355, 2014.

VANRADEN, P. M. Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science**, Champaign, v. 91, n. 11, p. 4414-4423, Nov. 2008.

VIEIRA I et al. Assessing non-additive effects in gblup model. **Genetics and Molecular Research**, v.16(2), p.1–21, 2017.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the Additive and Dominant Variance and Covariance of Individuals within the Genomic Selection Scope. **Genetics**, Bethesda, v. 195, n. 4, p. 1223-1230, Dec. 2013.

ZHANG, A. et al. Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations, **Frontiers in Plant Science**, v.8, p.1-12, 2017.

ZHANG, H. et al. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations, **Frontiers in Plant Science**, v.10, 2019.

ZENG, J. et al. 2013 Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action, **Genetics Selection Evolution**, v.45, p.1-17, 2013.

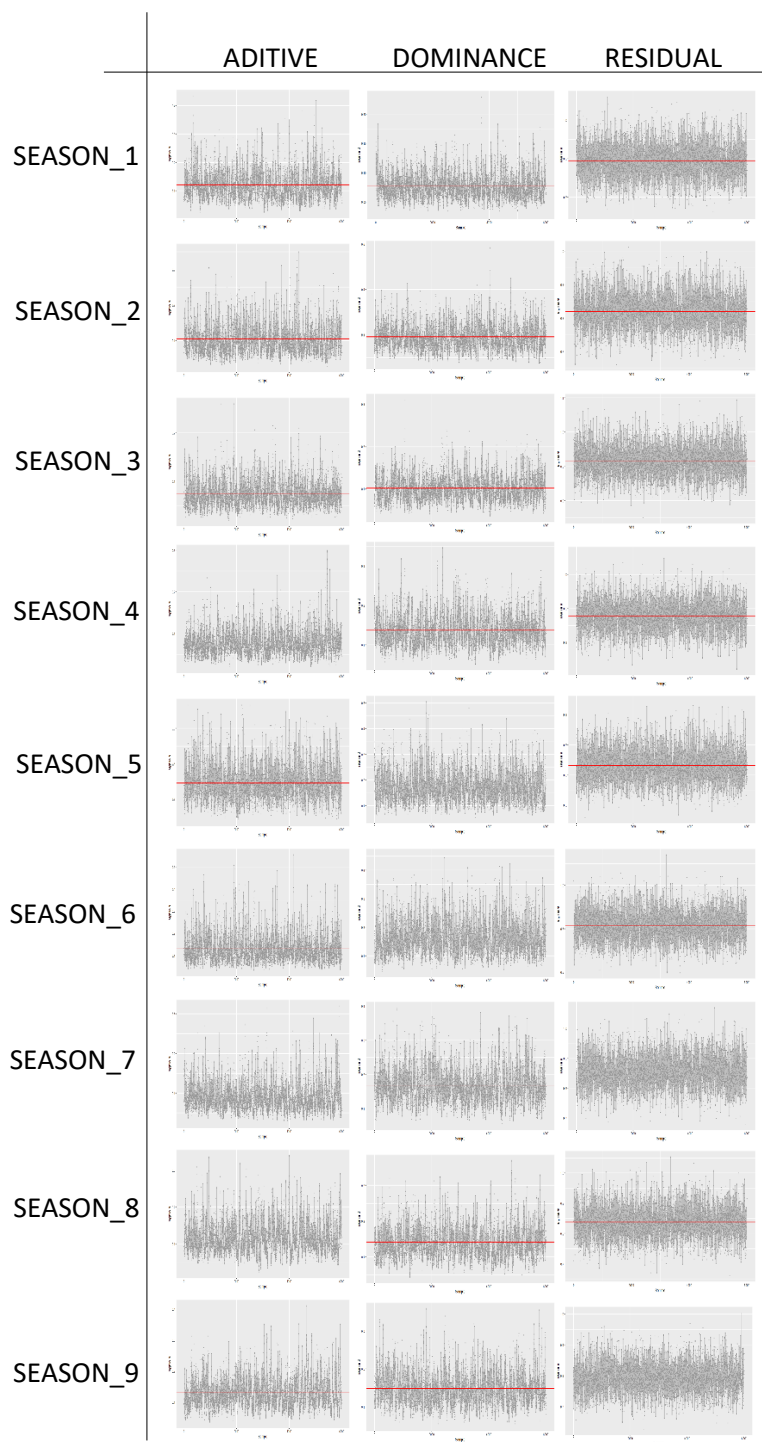
## MATERIAL SUPLEMENTAR

Material Suplementar 1: Descrição do número de híbridos simples (HS), locais, repetições (REP), sigla e design experimental adotados em cada safra (EXP).

Safra	Ano	Sigla	HS	Locais	EXP	Design	REP
Centro 1 <sup>a</sup> safra	2011/2012	2011s1C	1012	5	28	DBC	3
Sul 1 <sup>a</sup> safra	2011/2012	2011s1S	783	5	22	DBC	3
Centro 2 <sup>a</sup> safra	2012	2012s2C	1041	6	24	DBC	3
Sul 2 <sup>a</sup> safra	2012	2012s2S	1040	4	22	DBC	3
Centro 1 <sup>a</sup> safra	2012/2013	2012s1C	441	7	13	DBI	3
Sul 1 <sup>a</sup> safra	2012/2013	2012s1S	725	7	19	DBI	2
Centro 2 <sup>a</sup> safra	2013	2013s2C	411	6	11	DBI	3
Centro 1 <sup>a</sup> safra	2013/2014	2013s1C	810	4	20	DBI	2
Centro 2 <sup>a</sup> safra	2014	2014s2C	853	6	21	DBI	2

Fonte: Do autor (2021).

Material Suplementar 2: Estudo de convergência do modelo de predição aditivo-dominante com base na estima das variâncias aditiva, de dominância e residual, obtidas nas 4000 interações considerando um intervalo de amostragem (thin) igual a 5.



Fonte: Do autor (2021).