SAULO FABRÍCIO DA SILVA CHAVES

STATISTICAL GENETICS TOOLS FOR EMPOWERED DATA-DRIVEN DECISIONS

Thesis submitted to the Genetics and Breeding Graduate Program of the Universidade Federal de Viçosa in a partial fulfilment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Luiz Antônio dos Santos Dias

Co-advisers: Kaio Olimpio G. Dias
Rodrigo Silva Alves

VIÇOSA - MINAS GERAIS

2024

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

T
C512s
2024

Chaves, Saulo Fabrício da Silva, 1997-
Statistical genetics tools for empowered data-driven decions / Saulo Fabrício da Silva Chaves. – Viçosa, MG, 2024.
1 tese eletrônica (208 f.): il. (algumas color.).

Orientador: Luiz Antônio dos Santos Dias.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Agronomia, 2024.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2024.122
Modo de acesso: World Wide Web.

1. Plantas - Melhoramento genético. 2. Genética quantitativa. I. Dias, Luiz Antônio dos Santos, 1957-. II. Universidade Federal de Viçosa. Departamento de Agronomia. Programa Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 631.52

SAULO FABRÍCIO DA SILVA CHAVES

STATISTICAL GENETICS TOOLS FOR EMPOWERED DATA-DRIVEN DECISIONS

Thesis submitted to the Genetics and Breeding Graduate Program of the Universidade Federal de Viçosa in a partial fulfilment of the requirements for the degree of *Doctor Scientiae*

APPROVED: March 18, 2024.

Assent:

gov.br Documento assinado digitalmente
SAULO FABRICIO DA SILVA CHAVES
Data: 26/03/2024 16:47:19-0300
Verifique em https://validar.iti.gov.br

Saulo Fabrício da Silva Chaves
Author

gov.br Documento assinado digitalmente
LUIZ ANTONIO DOS SANTOS DIAS
Data: 26/03/2024 20:06:36-0300
Verifique em https://validar.iti.gov.br

Luiz Antônio dos Santos Dias
Adviser

To Rubens and Ana Chaves

ACKNOWLEDGEMENTS

To Rubens and Ana Chaves, my dear parents, whose effort was vital for me to reach my goals. This is all for them.

To Ariane Carneiro, my beloved lifemate, who was always there to ease pains and celebrate victories. My little piece of home.

To the friendships built throughout these years in Viçosa. A special thanks to Filipe and Mauricio, two of the best guys I met in both personal and professional ways.

To the three scientists who taught me about plant breeding, data analysis and statistical genetics: Rafael Alves, Rodrigo Alves and Kaio Dias. Their support and willingness to share their vast knowledge was a beacon for me and led me to the place I am now. There are no words to express how grateful I am for their belief in me, even in moments of self-doubt.

To my adviser, Professor Luiz Dias, a mentor since my master's graduate, who provided all the means for me to reach my objectives from the moment I arrived at UFV.

To all people who directly and indirectly contributed to my formation: my family and friends from Belém, professors from UFRA and the Graduate Program in Genetics and Breeding of UFV, friends from the Biometry laboratory, Statistical Genetics and Computational Biology laboratory, and Agroenergy laboratory, co-authors of the papers (chapters) of this thesis, and the evaluation committee.

To the institutions whom I had the opportunity to work with, and whose data I used in the chapters of this thesis: Embrapa Amazônia Oriental, Embrapa Milho e Sorgo, CENIBRA, CMPC, and CEPLAC.

All papers within this thesis were done with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Financing code 001.

˝If you want to overcome the whole world, overcome yourself˝

Fyodor Dostoevsky, *Demons*

ABSTRACT

CHAVES, Saulo Fabrício da Silva, D.Sc., Universidade Federal de Viçosa, March, 2024. **Statistical genetics tools for empowered data-driven decisions**. Adviser: Luiz Antônio dos Santos Dias.

The pressure to accelerate results in plant breeding programs is intensifying. Conversely, there is a concerning decline in the genetic diversity of staple crops, making it increasingly difficult to achieve genetic gains. Consequently, efficient resource allocation within breeding programs requires the strategic implementation of statistical genetics tools. This shift necessitates data-driven decision-making, placing professionals proficient in this toolkit at a significant advantage for addressing both traditional and emerging challenges. This thesis serves as a practical demonstration of utilizing statistical genetics in various plant breeding endeavours. Divided into six chapters, each with distinct objectives, the work showcases a range of applications. In Chapter 1, we determined the optimal number of harvests for selection in cacao breeding, considering both recommendation and recombination. Chapter 2 explores the application of covariance structure modelling in two common scenarios of perennial plant breeding: multi-harvest and multi-site data analysis. Chapter 3 demonstrates the use of factor analytic mixed models in maize breeding, including the incorporation of selection tools for streamlined decision-making. Notably, this chapter highlights the advantage of seasonal selection for achieving greater genetic gains compared to a combined approach. In Chapter 4, we evaluated the efficacy of the reciprocal recurrent selection (RRS) scheme within a eucalyptus breeding program. This chapter acknowledges the extended timeframe associated with RRS but also demonstrates its success in enhancing the hybrid population. Additionally, the chapter emphasizes the importance of considering dominance effects during the selection process. Chapter 5 offers a comprehensive tutorial on conducting linear mixed model analyses in perennial plant breeding. The chapter covers various analyses, including individual trials, multi-environment trials, spatial analysis, and competition analysis. Finally, Chapter 6 introduces the `R` package `ProbBreed`, which utilizes Bayesian principles and probabilistic concepts to support selection in multi-environment trials. `ProbBreed` estimates the risk associated with selecting candidates, empowering more informed decision-making. This chapter also introduces a novel multi-location-year model and compares the outcomes of `ProbBreed` and `ASReml-R` using simulated data. By showcasing the applications of statistical genetics tools and facilitating knowledge sharing through open-source code and reproducible examples, this thesis

emphasizes the versatility and importance of this field in tackling diverse challenges within the dynamic field of plant breeding.

Keywords: Data Analysis. Linear Mixed Models. Bayesian models. Genotype-by-Environment interaction. Spatial Analysis. Reciprocal Recurrent Selection

RESUMO

CHAVES, Saulo Fabrício da Silva, D.Sc., Universidade Federal de Viçosa, março de 2024. **Ferramentas de genética estatística para decisões qualificadas baseadas em dados**. Orientador: Luiz Antônio dos Santos Dias.

A pressão para acelerar os resultados dos programas de melhoramento de plantas está a intensificar-se. Em contraste, há um declínio preocupante na diversidade genética das principais culturas, tornando cada vez mais difícil obter ganhos genéticos consistentes. Consequentemente, a alocação eficiente de recursos nos programas de melhoramento requer a implementação estratégica de ferramentas de genética-estatística. Esta mudança exige uma tomada de decisão baseada em dados, colocando os profissionais proficientes neste conjunto de ferramentas numa vantagem significativa para enfrentar desafios tradicionais e emergentes. Esta tese serve como uma demonstração prática da utilização da genética-estatística em várias vertentes dod melhoramento de plantas. Dividido em seis capítulos, cada um com objetivos distintos, o trabalho apresenta uma gama de aplicações. No Capítulo 1 determinamos o número ideal de colheitas para seleção no melhoramento de cacau, considerando recomendação e recombinação. O Capítulo 2 explora a aplicação da modelagem de estrutura de covariância em dois cenários comuns de melhoramento de plantas perenes: análise de dados de múltiplas colheitas e de múltiplos locais. O Capítulo 3 demonstra o uso de modelos fator analítico no melhoramento de milho, incluindo a incorporação de ferramentas de seleção para agilizar a tomada de decisões. Notavelmente, este capítulo destaca a vantagem da seleção sazonal para alcançar maiores ganhos genéticos em comparação com uma abordagem combinada. No Capítulo 4 avaliamos a eficácia do esquema de seleção recorrente recíproca (SRR) dentro de um programa de melhoramento genético de eucalipto. Este capítulo reconhece o prazo alargado associado ao SRR, mas também demonstra o seu sucesso no aumento da população híbrida. Além disso, o capítulo enfatiza a importância de considerar os efeitos de dominância durante o processo de seleção. O Capítulo 5 oferece um tutorial abrangente sobre a condução de análises baseadas em modelos lineares mistos no melhoramento de plantas perenes. O capítulo cobre várias análises, incluindo ensaios individuais, ensaios multiambientais, análise espacial e análise de concorrência. Finalmente, o Capítulo 6 apresenta o pacote do R `ProbBreed`, que utiliza princípios bayesianos e conceitos probabilísticos para apoiar a seleção em ensaios multiambientes. `ProbBreed` estima o risco associado à seleção de candidatos, capacitando uma tomada de decisão mais informada. Este capítulo também apresenta um novo modelo de

múltiplos anos e locais e compara os resultados dos pacotes `ProbBreed` e `ASReml-R` usando dados simulados. Ao mostrar as aplicações de ferramentas de genética-estatística e facilitar o compartilhamento de conhecimento através de códigos e exemplos reproduzíveis, esta tese enfatiza a versatilidade e a importância deste campo no enfrentamento de diversos desafios no dinâmico campo do melhoramento de plantas.

Palavras-chave: Análise de dados. Modelos Lineares Mistos. Modelos Bayesianos. Interação Genótipos-Ambientes. Análise Espacial. Seleção Recorrente Recíproca

LIST OF FIGURES

LIST OF TABLES

SUMMARY

INTRODUCTION

Wallace et al. (2018) proposed a framework dividing plant breeding into four stages. Breeding 1.0 emerged alongside the origins of agriculture and persisted until the work of Fisher (1919), who bridged the gap between Mendelian and biometrical schools of thought, laying the foundation for quantitative genetics. This ushered in the era of Breeding 2.0. Breeding 3.0 dawned with the advent of early efforts in constructing linkage maps using molecular data. This approach gradually evolved into genome-wide association studies (GWAS) and genome-wide prediction (GWP), as high-throughput genome sequencing platforms became available (ELSHIRE et al., 2011). Breeding 4.0, as outlined by Wallace et al. (2018), signifies a shift towards more comprehensive data acquisition practices, incorporating tools such as genome editing, high-throughput phenotyping, and machine learning. The authors assert that plant breeding is currently undergoing a transition from Breeding 3.0 to Breeding 4.0, and they further explore the opportunities and challenges associated with accelerating and refining this transition. It is noteworthy that the knowledge is cumulative, i.e., there is not a formal end to a stage, so another can begin. The democratization of these tools is out of the scope of this text, but it is worth mentioning that there is a gap between the Breeding stages of developed and developing countries.

A factor changed the course of plant breeding history beginning from Breeding 2.0: statistical genetics. Until then, breeding was based solely on empirical knowledge, and results were achieved less efficiently. Thus, the pivotal role of statistical genetics in shaping the trajectory of plant breeding cannot be overstated. This versatile field bridges the gap between two crucial areas, employing statistical methods to elucidate genetic phenomena (BALDING et al., 2019). Statistical genetics seamlessly integrates with the concepts of quantitative genetics, which were initially demonstrated mathematically and observed phenotypically. Later, the advent of molecular data provided the crucial genetic validation for these concepts. These advancements allowed plant breeding to play an important role in the green revolution and keep the world fed until these days, tackling the pessimist prediction of Thomas Malthus.

The "quantitative genetic golden era" of Breeding 2.0 witnessed the establishment of numerous groundbreaking concepts that continue to hold significant importance in modern breeding practices. These include recurrent selection schemes pioneered by Hull (1945) and Comstock et al. (1949), development of specific trial designs to estimate additive, dominant, and epistatic variance components (COMSTOCK; ROBINSON, 1952, 1948), the establishment of covariance relationships between relatives (COCKERHAM, 1973; KEMPTHORNE et al., 1954), development of selection indices and selection methods under genotype-by-environment interaction (GEI) (FINLAY; WILKINSON, 1963; HAZEL, 1943; HAZEL; LUSH, 1942; SHUKLA, 1972). Furthermore, the foundational work of Patterson and Thompson (1971) and Henderson (1975) laid the groundwork for linear mixed model analyses, an invaluable tool widely used in modern plant breeding. Breeding 3.0 also saw significant advancements in quantitative genetics, including the development of QTL mapping methods (LANDER et al., 1987; MACKAY, 2001; ZENG, 1994), the use of genome-wide information for selection and prediction, incorporating both frequentist and Bayesian approaches (BERNARDO, 1994; HABIER et al., 2007; MEUWISSEN et al., 2001); and advances in constructing kinship kernels for application in statistical genetic models (VANRADEN, 2008; VITEZICA et al., 2013; YANG et al., 2010).

The evolution of plant breeding necessitates the integration of sophisticated statistical genetics tools, as there is no “modern plant breeding program” without it. As we delve deep into the “big data” era, characterized by the democratization of high-throughput platforms (genomics, phenomics, enviromics, metabolomics, etc.), data-driven decision-making is becoming increasingly paramount. Consequently, a comprehensive understanding of statistical genetics is an essential component of the modern plant breeder’s toolkit and must be an integral part of breeders’ formation (BERNARDO, 2020). This expertise empowers professionals to leverage statistical genetics as a powerful tool to tackle both longstanding and emerging challenges within the field. By embracing this evolving landscape, breeders are well-positioned to navigate the future stages of plant breeding.

To demonstrate the versatility of statistical genetics in handling problems of varying complexities, this thesis comprises six chapters arranged in ascending order of complexity. The first four chapters draw upon data from various ongoing breeding programmes and address real-life challenges or propose improvements to inform decision-making. The final two chapters take a more instructional approach, providing detailed demonstrations, including the relevant R

code, of how to conduct data analysis using cutting-edge methods within both frequentist and Bayesian frameworks. Notably, the final chapter references an `R` package developed by our own research group.

In Chapter **1**, we employed a repeatability model to establish the optimal number of harvest years required for selection in both recombination and recommendation stages of cacao breeding (*Theobroma cacao* L.) (CHAVES et al., 2022b). Before the publication of this paper, few studies had addressed this topic, focusing solely on yield traits such as the number of healthy fruits (CARVALHO et al., 2002; DIAS; KAGEYAMA, 1998; TAHI et al., 2019). Our research was the first to utilise a mixed model framework to determine the minimum evaluation period for witches' broom resistance and dry seed weight. Notably, a subsequent study from our group pioneered the use of random regression models to analyse multi-harvest data in cacao (ALVES et al., 2024). This approach is particularly advantageous when dealing with longitudinal data, facilitating a comprehensive assessment of the genotype-by-harvest interaction.

Chapter **2** explores how covariance structure modelling can enhance the precision of data-driven decisions in perennial plant breeding (CHAVES et al., 2022a). We focus on two common scenarios: orchard species requiring repeated measures (multi-harvest) for reliable selection, exemplified by the cupuassu tree [*Theobroma grandiflorum* Willd. (Ex. Spreng) Schumm.]; and timber species typically evaluated in diverse locations (multi-environmental trials, MET), exemplified by eucalyptus (*Eucalyptus* spp. L'Hér.). It is important to note that the methodologies implemented in this study are not restricted to perennial plant species and can be effectively applied to short-cycle species as well. Notably, this chapter offers a unique contribution by integrating spatial analysis with longitudinal analysis modelling. We employed the three-way autoregressive structure outlined by Smith et al. (2007) to model the residual covariance. Finally, the chapter concludes by discussing the practical implications of covariance structure modelling and its potential to improve decision-making in perennial plant breeding. Building upon our previous work, it is noteworthy that a second paper from our group further explored covariance structure modelling in the context of longitudinal data (CHAVES et al., 2023a). This study pioneered the application of Factor Analytic Selection Tools (FAST, see next paragraph for more details) developed by Smith and Cullis (2018), originally intended for multi-environmental trial data, for selection in a multi-harvest scenario. Furthermore, we introduced a novel metric – the average semivariance ratio (ASR) – to quantify the explanatory power of Factor Analytic (FA) mixed models. This metric draws inspiration from the ideas

presented by Piepho (2019).

Chapter **3**, the only chapter utilizing data from an annual species (maize, *Zea mays* L.), placed a major emphasis on Factor Analytic (FA) models. This chapter pursued two primary objectives: firstly, to demonstrate the potential of FA models becoming a standardized procedure within maize breeding, primarily through the application of FAST; and secondly, to investigate whether selection based on individual seasons could prove more advantageous than a combined approach (CHAVES et al., 2023b). FAST encompasses three metrics – overall performance (OP), root mean squared deviation (RMSD, a measure of global stability), and responsiveness (RE, a measure of specific stability) – that offer summaries of FA model complexity, thereby facilitating informed decision-making. This study marked the first instance of FAST implementation within a maize ongoing breeding program. Additionally, our findings indicated that seasonal selection yielded greater gains compared to simultaneous selection for both seasons. This aligns perfectly with biological rationale, as genotypes cultivated in the second season require distinct characteristics, such as precocity and abiotic stress tolerance, compared to their first-season counterparts. This distinction stems from the Brazilian maize second season's shorter timeframe and increased susceptibility to drought (or frost, depending on location). A subsequent study by our group combined FAST, geographical information systems (GIS), partial least squares regression, and enviromics to propose a novel methodology – GIS-FA (ARAÚJO et al., 2024). This method facilitates selection in both tested and untested environments while enabling in-depth exploration of genotype-by-environment interaction (GEI) through thematic maps. GIS-FA holds promise as a tool for optimizing genotype allocation and cost reduction, as predicted outcomes can potentially replace the need for physical trials.

Chapter **4** presents a comprehensive analysis of a 26-year-old reciprocal recurrent selection (RRS) breeding program for eucalyptus. This chapter's primary objective was to evaluate the efficacy of RRS in enhancing the hybrid population (*Eucalyptus grandis* W.Hill × *Eucalyptus urophylla* S.T.Blake) and quantify the genetic gain achieved after two RRS cycles. Additionally, the research assessed the impact of RRS on specific population parameters, including genetic variance and average dominance degree. To the best of our knowledge, this represents the first study to utilize real data to assess the RRS strategy in eucalyptus breeding. Our findings revealed a 28.5% and 12.3% gain in high-altitude and low-altitude regions, respectively. Furthermore, a detailed investigation highlighted that neglecting dominance effects during

selection can significantly hinder selection efficiency. The chapter concludes by discussing potential avenues to enhance RRS efficacy in eucalyptus, including reducing the cyclical timeframe or incorporating additional data sources such as genomics and competition analysis. A recent study by our group explored how competition can potentially disrupt selection processes, leading to a reduced correlation between observed field trial performance and the *per se* performance of selected trees in other trials or commercial plantations (FERREIRA et al., 2023). This study proposed a novel classification system for candidates based on their competition capacity. Additionally, it introduced a novel method for defining optimal clonal mixtures, aiming to increase genetic diversity in the field and, consequently, enhance economic security. This method also facilitates the optimization of field composition by selecting clones that demonstrate harmonious coexistence.

Chapter **5** embodies our group's resolute commitment to knowledge sharing, recognizing that potential is limited by the refusal to disseminate it. This chapter serves as a practical guide for conducting linear mixed model analyses in perennial plant breeding using the `ASReml-R` library (THE VSNI TEAM, 2023). Its unique value lies in the inclusion of reproducible `R` code snippets, readily accessible to readers via a public dataset available on GitHub (`https://github.com/saulo-chaves/MHT_MET_MM`). This dataset was previously used in Chapter 2 (CHAVES et al., 2022a). Chapter 5 guides users through a comprehensive range of analyses, including individual trial analyses, multi-environment trial analyses (incorporating both covariance structure modelling and FA models), spatial analyses, and competition analysis. The FA section delves into the particulars of factor rotation, model selection, the utilization of latent regression plots, and FAST. Notably, the chapter also functions as a repository of pertinent references on the discussed topics, facilitating not only the replication of the analyses but also their comprehension and interpretation.

Chapter **6** delves into the application of our newly developed `R` package, `ProbBreed` (CHAVES et al., 2024). This package aims to democratize the method proposed by Dias et al. (2022), which utilizes a Bayesian framework and might prove challenging to implement for individuals without advanced programming skills. While the execution itself is not trivial, the underlying concept is clear: leveraging Hamiltonian Monte Carlo discretized samples from the posterior distributions of fitted Bayesian models to simulate various trial outcomes. Within each sample, questions like "What is the risk of selecting a particular candidate?", "How likely is a candidate to outperform a commercial check?", and "What are the chances

of a genotype performing similarly across environments?" are posed and answered through probabilistic principles. Notably, due to its probabilistic basis, the outputs generated by `ProbBreed` inherently incorporate an essential factor for breeders: risk. In simpler terms, the package empowers informed decision-making by minimizing the risks associated with inaccurate recommendations. Additionally, beyond providing a comprehensive tutorial for performing analyses using `ProbBreed` with real data (including `R` code) from a publicly available dataset (KRAUSE et al., 2023), the chapter also introduces a novel multi-environment-year model and presents a comparative analysis of outcomes generated by `ProbBreed` and `ASReml-R` using simulated data.

The key takeaway is that data-driven decision-making in plant breeding finds its foundation in the concepts of statistical genetics. This dynamic field continuously evolves to develop novel solutions for both existing and new challenges within the field.

References

ALVES, A. K. S.; CHAVES, S. F. S.; ARAÚJO, M. S.; MALIKOUSKI, R. G.; ALMEIDA, C. M. V. C.; DIAS, L. A. S. Improving multi-harvest data analysis in cacao breeding using random regression. **Euphytica**, v. 220, n. 1, p. 7, 2024.

ARAÚJO, M. S.; CHAVES, S. F. S.; DIAS, L. A. S.; FERREIRA, F. M.; PEREIRA, G. R.; BEZERRA, A. R. G.; ALVES, R. S.; HEINEMANN, A. B.; BRESEGHELLO, F.; CARNEIRO, P. C. S.; KRAUSE, M. D.; COSTA-NETO, G.; DIAS, K. O. G. GIS-FA: an approach to integrating thematic maps, factor-analytic, and envirotyping for cultivar targeting. **Theoretical and Applied Genetics**, v. 137, n. 4, p. 80, 2024.

BALDING, D.; MOLTKE, I.; MARIONI, J. (Eds.). **Handbook of Statistical Genomics: Two Volume Set**. 1. ed.: Wiley, 2019.

BERNARDO, R. Prediction of maize single-cross performance using RFLPs and information from related hybrids. **Crop Science**, v. 34, n. 1, cropsci1994.0011183x003400010003x, 1994.

BERNARDO, R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. **Heredity**, v. 125, n. 6, p. 375–385, 2020.

CARVALHO, C.; CRUZ, C.; ALMEIDA, C.; MACHADO, P. Yield repeatability and evaluation period in hybrid cocoa assessment. **Crop Breeding and Applied Biotechnology**, v. 2, n. 1, p. 149–156, 2002.

CHAVES, S. F. S.; ALVES, R. S.; DIAS, L. A. S.; ALVES, R. M.; DIAS, K. O. G.; EVANGELISTA, J. S. P. C. Analysis of repeated measures data through mixed models: An application in *Theobroma grandiflorum* breeding. **Crop Science**, v. 63, n. 4, p. 2131–2144, 2023.

CHAVES, S. F. S.; EVANGELISTA, J. S. P. C.; ALVES, R. S.; FERREIRA, F. M.; DIAS, L. A. S.; ALVES, R. M.; DIAS, K. O. G.; BHERING, L. L. Application of linear mixed models for multiple harvest/site trial analyses in perennial plant breeding. **Tree Genetics & Genomes**, v. 18, n. 6, p. 44, 2022.

CHAVES, S. F. S.; EVANGELISTA, J. S. P. C.; TRINDADE, R. S.; DIAS, L. A. S.; GUIMARÃES, P. E.; GUIMARÃES, L. J. M.; ALVES, R. S.; BHERING, L. L.; DIAS, K. O. G. Employing factor analytic tools for selecting high-performance and stable tropical maize hybrids. **Crop Science**, v. 63, n. 3, p. 1114–1125, 2023.

CHAVES, S. F. S.; KRAUSE, M. D.; DIAS, L. A. S.; GARCIA, A. A. F.; DIAS, K. O. G. ProbBreed: a novel tool for calculating the risk of cultivar recommendation in multienvironment trials. **G3 Genes|Genomes|Genetics**, v. 14, n. 3, jkae013, 2024.

CHAVES, S. F. S.; DIAS, L. A. S.; ALVES, R. S.; ALVES, R. M.; JOSÉ, A. R. M.; ALMEIDA, C. M. V. C. d. Number of harvest years and selection for productivity, witches' broom resistance, stability, and adaptability in cacao. **Agronomy Journal**, v. 114, n. 6, p. 3234–3245, 2022.

COCKERHAM, C. C. Analyses of gene frequencies. **Genetics**, v. 74, n. 4, p. 679–700, 1973.

COMSTOCK, R. E.; ROBINSON, H. F. Estimation of average dominance of genes. In: GOWEN, J. W. (Ed.). **Heterosis: A record of researches directed toward explaining and utilizing the vigor of hybrids**. Ames: Iowa State College Press, 1952. P. 494–516.

COMSTOCK, R. E.; ROBINSON, H. F. The components of genetic variance in populations of biparental progenies and their use in estimating the average degree of dominance. **Biometrics**, v. 4, n. 4, p. 254–266, 1948.

COMSTOCK, R. E.; ROBINSON, H. F.; HARVEY, P. H. A breeding procedure designed to make maximum use of both general and specific combining ability. **Agronomy Journal**, v. 41, n. 8, p. 360–367, 1949.

DIAS, K. O. G.; SANTOS, J. P. R.; KRAUSE, M. D.; PIEPHO, H.-P.; GUIMARÃES, L. J. M.; PASTINA, M. M.; GARCIA, A. A. F. Leveraging probability concepts for cultivar recommendation in multi-environment trials. **Theoretical and Applied Genetics**, v. 135, n. 4, p. 1385–1399, 2022.

DIAS, L. A. S.; KAGEYAMA, P. Y. Repeatability and minimum harvest period of cacao (*Theobroma cacao* L.) in Southern Bahia. **Euphytica**, v. 102, p. 29–35, 1998.

ELSHIRE, R. J.; GLAUBITZ, J. C.; SUN, Q.; POLAND, J. A.; KAWAMOTO, K.; BUCKLER, E. S.; MITCHELL, S. E. A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. **PLOS ONE**, v. 6, n. 5, e19379, 2011.

FERREIRA, F. M.; CHAVES, S. F. S.; BHERING, L. L.; ALVES, R. S.; TAKAHASHI, E. K.; SOUSA, J. E.; RESENDE, M. D. V.; LEITE, F. P.; GEZAN, S. A.; VIANA, J. M. S.; FERNANDES, S. B.; DIAS, K. O. G. A novel strategy to predict clonal composites by jointly modeling spatial variation and genetic competition. **Forest Ecology and Management**, v. 548, p. 121393, 2023.

FINLAY, K. W.; WILKINSON, G. N. The analysis of adaptation in a plant-breeding programme. **Australian Journal of Agricultural Research**, v. 14, n. 6, p. 742, 1963.

FISHER, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. **Earth and Environmental Science Transactions of The Royal Society of Edinburgh**, v. 52, n. 2, p. 399–433, 1919.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The impact of genetic relationship information on genome-assisted breeding values. **Genetics**, v. 177, n. 4, p. 2389–2397, 2007.

HAZEL, L. N. The genetic basis for constructing selection indexes. **Genetics**, v. 28, n. 6, p. 476–490, 1943.

HAZEL, L. N.; LUSH, J. L. The efficiency of three methods of selection. **Journal of Heredity**, v. 33, n. 11, p. 393–399, 1942.

HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a selection model. **Biometrics**, v. 31, n. 2, p. 423, 1975.

HULL, F. H. Recurrent selection for specific combining ability in corn. **Agronomy Journal**, v. 37, n. 2, p. 134–145, 1945.

KEMPTHORNE, O.; FISHER, R. A.; MATTHEWS, B. H. C. The correlation between relatives in a random mating population. **Proceedings of the Royal Society of London. Series B - Biological Sciences**, v. 143, n. 910, p. 103–113, 1954.

KRAUSE, M. D.; DIAS, K. O. G.; SINGH, A. K.; BEAVIS, W. D. Using soybean historical field trial data to study genotype by environment variation and identify mega-environments with the integration of genetic and non-genetic factors. **bioRxiv : the preprint server for biology**, 2023.

LANDER, E. S.; GREEN, P.; ABRAHAMSON, J.; BARLOW, A.; DALY, M. J.; LINCOLN, S. E.; NEWBURG, L. MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. **Genomics**, v. 1, n. 2, p. 174–181, 1987.

MACKAY, T. F. C. The genetic architecture of quantitative traits. **Annual Review of Genetics**, v. 35, n. 1, p. 303–339, 2001.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, 2001.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545–554, 1971.

PIEPHO, H.-P. A coefficient of determination (R^2) for generalized linear mixed models. **Biometrical Journal**, v. 61, n. 4, p. 860–872, 2019.

SHUKLA, G. K. Some statistical aspects of partitioning genotype-environmental components of variability. **Heredity**, v. 29, n. 2, p. 237–245, 1972.

SMITH, A. B.; CULLIS, B. R. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. **Euphytica**, v. 214, n. 8, p. 143, 2018.

SMITH, A. B.; STRINGER, J. K.; WEI, X.; CULLIS, B. R. Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. **Euphytica**, v. 157, n. 1, p. 253–266, 2007.

TAHI, M.; TREBISSOU, C.; RIBEYRE, F.; GUIRAUD, B. S.; POKOU, D. N.; CILAS, C. Variation in yield over time in a cacao factorial mating design: changes in heritability and longitudinal data analyses over 13 consecutive years. **Euphytica**, v. 215, n. 6, p. 106, 2019.

THE VSNI TEAM. **asreml: Fits Linear Mixed Models using REML**. 2023.

VANRADEN, P. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics**, v. 195, n. 4, p. 1223–1230, 2013.

WALLACE, J. G.; RODGERS-MELNICK, E.; BUCKLER, E. S. On the road to breeding 4.0: Unraveling the good, the bad, and the boring of crop quantitative genomics. **Annual Review of Genetics**, v. 52, n. 1, p. 421–444, 2018.

YANG, J.; BENYAMIN, B.; MCEVOY, B. P.; GORDON, S.; HENDERS, A. K.; NYHOLT, D. R.; MADDEN, P. A.; HEATH, A. C.; MARTIN, N. G.; MONTGOMERY, G. W.; GODDARD, M. E.; VISSCHER, P. M. Common SNPs explain a large proportion of the heritability for human height. **Nature Genetics**, v. 42, n. 7, p. 565–569, 2010.

ZENG, Z. B. Precision mapping of quantitative trait loci. **Genetics**, v. 136, n. 4, p. 1457–1468, 1994.

CHAPTER 1

NUMBER OF HARVEST YEARS AND SELECTION FOR PRODUCTIVITY, WITCHES' BROOM RESISTANCE, STABILITY, AND ADAPTABILITY IN CACAO

Published article: CHAVES, S. F. S.; DIAS, L. A. S.; ALVES, R. S.; ALVES, R. M.; JOSÉ, A. R. M.; ALMEIDA, C. M. V. C. Number of harvest years and selection for productivity, witches' broom resistance, stability, and adaptability in cacao. **Agronomy Journal**, v. 114, n. 6, p. 3234–3245, 2022.

A suitable statistical model for the analysis of repeated measures is a prerequisite for an accurate genetic selection of cacao (*Theobroma cacao* L.) genotypes. Thus, the objectives of this study were to (a) evaluate homoscedastic and heteroscedastic repeatability models, (b) estimate the optimal number of harvest years for genetic selection, and (c) identify and select biparental crosses with high productivity, witches' broom (*Moniliophthora perniciosa* Stahel & Philips-Mora) resistance, stability, and adaptability. Twenty biparental crosses were evaluated in a complete randomized blocks design, with seven replications and 12 trees per plot for 10 harvest years. We evaluated the number of healthy fruits (NHF), number of symptomatic fruits for witches' broom (NWBF), dry bean weight (DBW), and fruit index (FI). Variance components were estimated by restricted maximum likelihood, and genotypic values were predicted by best linear unbiased prediction. The best-fitted repeatability model for each trait was indicated by Akaike's information criterion. The additive index was used for simultaneous selection for productivity, witches' broom resistance, stability, and adaptability. The heteroscedastic repeatability model showed the best fit for all traits. For NHF, DBW, and FI, two and seven harvest years were sufficient for selection aiming at recombination and recommendation, respectively. For NWBF, evaluated only from the seventh year onwards, one harvest year was enough for recombination and three for recommendation. Five biparental crosses with high productivity, witches' broom resistance, stability, and adaptability were selected for recombination and/or for competition trials for recommendation. Such results can

help define efficient strategies and optimize cacao breeding programs.

1 Introduction

Cacao (*Theobroma cacao* L.) is a perennial fruit tree of great socioeconomic value, especially for developing countries in tropical regions of the world. Latin America is the centre of origin and cacao has been cultivated there for over 5,000 yr (DIAS, 2001; MOTAMAYOR et al., 2008; THOMAS et al., 2012; ZARRILLO et al., 2018). Today it is an important crop in South and Central America, Southeast Asia, Melanesia, and West Africa (BEKELE; PHILLIPS-MORA, 2019; WICKRAMASURIYA; DUNWELL, 2018).

Despite its relative success, the Brazilian cacao industry faces serious difficulties that range from socio-environmental issues to phytosanitary problems, such as the occurrence of diseases: witches' broom (*Moniliophthora perniciosa* Stahel & Philips-Mora), black pod (*Phytophthora* spp.), and ceratocystis wilt (*Ceratocystis cacaofunesta* Engelbrecht & TC Harr) (BAILEY et al., 2018; BENJAMIN et al., 2016; RODRIGUES et al., 2020). An economically viable and ecologically correct strategy to overcome these difficulties is genetic breeding. This strategy focuses on the development of genotypes with high dry bean yield, adequate vegetative development, and resistance to the main pests and diseases (BEKELE; PHILLIPS-MORA, 2019; DIAS, 2001). Cacao is an allogamous species, with variable rates of sporophytic genetic inter- and self-incompatibility. To capitalize on these peculiarities, intrapopulation recurrent selection has been successfully used, which operates with recurrent cycles of crossings and selection for the development of superior genotypes (DUVAL et al., 2017; MUSTIGA et al., 2018).

The duration of each breeding cycle (12 yr on average) associated with the longevity of plants in the field, generates and requires the use of repeated measures on the same individual for an efficient genetic selection. The repeated measures data has particularities which must be considered in the genetic evaluation, such as correlation and heterogeneity of variances between measures (ANDRADE et al., 2016; VERBYLA et al., 2021). The repeatability model allows for the estimation of the optimal number of harvest years for genetic selection, optimizing time, and consequently, increasing selection gains (FERREIRA et al., 2021). It also allows for the study of the temporal stability and adaptability of genetic values (ALVES et al., 2018) by the use of the genotypes × harvest years interaction (GHI) effects, which is especially important for

a crop that can produce for over 80 yr. Temporal stability and adaptability are also important in the current context of climate change (ATLIN et al., 2017; CHAVES et al., 2021; DIAS; KAGEYAMA, 1998; SCHROTH et al., 2016).

Determining the optimal number of harvest years for genetic selection and evaluating GHI have been considered in previous studies (CARVALHO et al., 2003, 2002; DIAS et al., 2001; DIAS; KAGEYAMA, 1997; DIAS et al., 2003). Selection decisions from these studies were based on estimates obtained from the least squares method (ANOVA). However, the ANOVA-based methods are inefficient with imbalanced data and/or variance heterogeneity over the harvest years, all relatively common scenarios when it comes to repeated measures in perennial species like cacao. Thus, the restricted maximum likelihood/best linear unbiased prediction (REML/BLUP) (HENDERSON, 1975; PATTERSON; THOMPSON, 1971) is the best procedure (in the frequentist context) for genetic evaluation and selection, as it has greater robustness and flexibility to deal with these problems (PIEPHO et al., 2008; RESENDE, 2002; RESENDE; ALVES, 2020).

The objectives of this study were to (a) evaluate homoscedastic and heteroscedastic repeatability models for determining the better-fit model, (b) to estimate the optimal number of harvest years for genetic selection, and (c) to identify and select biparental crosses with high productivity, witches' broom resistance, stability, and adaptability.

2 Materials and methods

2.1 Experimental conditions

The trial was established at the Ouro Preto Experimental Station (ESEOP), in the municipality of Ouro Preto D'Oeste (10° 42′ 30″ S, 62° 13′ 30″ W), state of Rondônia, Brazilian Amazon. The region has an Am-type climate, according to Köppen's classification, with an annual average temperature of 25.6 $^{\circ}C$ and relative air humidity of 89%. The soil is medium deep, with good drainage and layer differentiation, with medium fertility.

The cacao trees, spaced 3 × 3 m apart, were planted in an agroforestry system typical of the region, with temporary shading provided by banana tree (*Musa* spp.), installed in the same spacing as the cacao tree, and cassava (*Manihot esculenta* Crantz), with four propagules per cacao tree. The definitive shading was provided by *Erythrina glauca*, in 24× 24 m. Around the experimental area, two rows of cacao trees were set up as an external border. The

experimental management followed good cultivation practices for cacao trees established in regional agroforestry systems (SILVA NETO et al., 2013).

2.2 Biparental crosses and trial design

Twenty biparental crosses were used in this study (Table 1). Each family was represented by 12 plants in a plot planted in a randomized complete block design arrangement. The plots consisted of three rows of four cacaos trees each and were replicated seven times. All plants in one plot are individuals originating from seeds, derived from the same cross. Except for E1H02, which is a biparental cross of an Upper Amazon clone (SCA 6) with a Trinitarian clone (ICS 1), all the others are crosses between Upper Amazon and Lower Amazon clones, aiming to gather complementary characteristics in the progenies (CARVALHO et al., 2003).

Table 1: Maternal and paternal clones and their respective geographic origin of the evaluated cacao biparental crosses

Biparental cross	Maternal		Paternal	
	Clone	Geographic origin	Clone	Geographic origin
E1H01	IMC 67	Iquitos, Peru	SIC 813	Bahia, Brazil
E1H02	SCA 6	Equador	ICS 1	River State, Trinindad
E1H03	IMC 67	Iquitos, Peru	CA 4	Amazonas, Brazil
E1H04	IMC 67	Iquitos, Peru	BE 8	Pará, Brazil
E1H05	IMC 67	Iquitos, Peru	BE 9	Pará, Brazil
E1H06	IMC 67	Iquitos, Peru	SIAL 169	Bahia, Brazil
E1H07	POUND 7	Iquitos, Peru	BE 10	Pará, Brazil
E1H08	POUND 7	Iquitos, Peru	MA 12	Amazonas, Brazil
E1H09	POUND 7	Iquitos, Peru	SIC 864	Bahia, Brazil
E1H10	POUND 7	Iquitos, Peru	MA 15	Amazonas, Brazil
E1H11	POUND 12	Iquitos, Peru	SIC 329	Bahia, Brazil
E1H12	POUND 12	Iquitos, Peru	MA 14	Amazonas, Brazil
E1H13	POUND 12	Iquitos, Peru	SIAL 505	Bahia, Brazil
E1H14	POUND 12	Iquitos, Peru	SIC 831	Bahia, Brazil
E1H15	PA 150	Parinari, Peru	SIC 328	Bahia, Brazil
E1H16	PA 150	Parinari, Peru	SIAL 325	Bahia, Brazil
E1H17	PA 150	Parinari, Peru	MA 11	Amazonas, Brazil
E1H18	PA 150	Parinari, Peru	SIC 864	Bahia, Brazil
E1H19	SCA 6	Equador	BE 9	Pará, Brazil
E1H20	SCA 6	Equador	BE10	Pará, Brazil

2.3 Traits evaluated

The following traits were evaluated on a plot basis: number of healthy fruits (NHF), number of symptomatic fruits for witches' broom (NWBF), and wet seed weight (WSW, in kg). Dry bean weight (DBW, in kg) was obtained by multiplying WSW by a correction factor (0.4). With the variables NHF and DBW, the fruit index (FI) was estimated, which indicates how many fruits are needed to produce 1 kg of dried beans, as follows (DIAS, 2001):

$$FI = \frac{1}{DBW/NHF} \tag{1}$$

Data collection started in the second year of the experiment, at the end of the cacao's juvenile period and beginning of its production and went up to the 12th year, totalling 10 consecutive years. In this study, the sum of all monthly evaluations carried out in a year was considered a harvest year. To evaluate the resistance to witches' broom, the infection of the disease occurred naturally, so that the genotypes could manifest the mechanisms of horizontal resistance. However, the trial was witches' broom-free until the fifth harvest year, the seventh year after planting. Therefore, data regarding this trait began to be collected only from this year onwards.

2.4 Statistical analyses

The REML method (PATTERSON; THOMPSON, 1971) was used to estimate the variance components, and the BLUP method (HENDERSON, 1975) was used for the prediction of genotypic values. The following equation determined the complete repeatability model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{p} + \mathbf{Z}_3\mathbf{gh} + \boldsymbol{\varepsilon} \tag{2}$$

where $\mathbf{y}$ is the vector of phenotypic data; $\mathbf{b}$ is the vector of the effects of the harvest years–repetition combinations (assumed to be fixed), added to the overall mean; $\mathbf{g}$ is the vector of genotypic effects, assumed to be random [$\mathbf{g} \sim N(\mathbf{0}, \sigma_g^2\mathbf{I})$, where σ_g^2 is the genotypic variance between biparental crosses]; $\mathbf{p}$ is the vector of the permanent environmental effects, assumed to be random [$\mathbf{p} \sim N(\mathbf{0}, \sigma_p^2\mathbf{I})$, where σ_p^2 is the variance of the permanent environmental effects]; $\mathbf{gh}$ is the vector of random GHI effects [$\mathbf{gh} \sim N(\mathbf{0}, \sigma_{gh}^2\mathbf{I})$, where σ_{gh}^2 is the variance of the GHI]; and $\boldsymbol{\varepsilon}$ is the vector of random residuals [$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$, where σ_ε^2 is the residual variance for

the homoscedastic model and $\sigma^2_\varepsilon \sim N(\mathbf{0}, \mathbf{R} = \oplus^H_{h=1}\sigma^2_\varepsilon \mathbf{I}_H)$, where $\mathbf{R}$ is a diagonal matrix of residual variance for the heteroscedastic model, and $\mathbf{I}_H$ is an identity matrix of order H, which represents the total number of harvest years (10)]. The capital letters $\mathbf{X}$, $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ are the incidence matrices for **b**, **g**, **p**, and **gh**, respectively.

The models with different residual variance (homogeneous and heterogeneous) structures were tested by the Akaike information criterion (AIC) (AKAIKE, 1974), given by the following equation:

$$AIC = -2logL + 2t \tag{3}$$

where $logL$ is the logarithm of the maximum of the restricted maximum likelihood function and t is the number of estimated parameters.

The significance of the random effects was tested using confidence intervals (CI_x) (GILMOUR et al., 2015), given by the following equation:

$$CI_x = \sigma^2_x \pm (1.96 \times SE_x) \tag{4}$$

where σ^2_x is the variance estimate of the random effect x, EP_x is the standard error of this estimate, and 1.96 is the tabulated value of the Student's t-test, with $P < 0.05$, for infinite degrees of freedom. if $CI_x \leq 0$, the effect was considered to be non-significant. Despite not being a formal hypothesis test, it yields equivalent results (RESENDE et al., 2014).

The following variance component and genetic parameters (phenotypic variance, $\sigma^2_{f_h}$; selective accuracy, $r_{\hat{g}g}$; generalized heritability, H^2_g; repeatability coefficient, ρ; coefficient of determination of permanent environmental effects, c^2_p; coefficient of determination of GHI effects, c^2_{gh}; and type B genotypic correlation across all harvest years, r^2_{gh}) were estimated by the following equations, respectively (CULLIS et al., 2006; GEZAN et al., 2017; MRODE, 2014; RESENDE et al., 2014):

$$\sigma^2_{f_h} = \sigma^2_g + \sigma^2_{gh} + \sigma^2_p + \sigma^2_{e_h} \tag{5}$$

$$H^2_g = 1 - \frac{\overline{V[\Delta(g_i - g_{i'})]_h}}{2 \times \sigma^2_g} \tag{6}$$

$$r_{\hat{g}g} = \sqrt{1 - \frac{PEV}{\sigma^2_g}} \tag{7}$$

$$\rho = \frac{\sigma^2_g + \sigma^2_p}{\sigma^2_f} \tag{8}$$

$$c_p^2 = \frac{\sigma_p^2}{\sigma_f^2} \tag{9}$$

$$c_{gh}^2 = \frac{\sigma_{gh}^2}{\sigma_f^2} \tag{10}$$

$$r_{gh}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{gh}^2} \tag{11}$$

where $\sigma_{e_h}^2$ is the residual variance of the h-th harvest year (unique for the homoscedastic model and variable for the heteroscedastic model), $\overline{V[\Delta(g_i - g_{i'})]}_h$ is the mean-variance of the difference between two BLUPs in the h-th harvest year, and PEV is the prediction error variance.

The pairwise genotypic correlations (Pearson correlation) between harvest years ($r_{hh'}$) for the same trait were estimated by the following equation:

$$r_{hh'} = \frac{Cov(GV_h, GV_{h'})}{[V(GV_h) \times V(GV_{h'})]^{1/2}} \tag{12}$$

where GV_h is the genotypic value in the h-th harvest year

The genotypic correlations ($r_{c_1c_2}$) between pairs of traits were estimated by the following equation:

$$r_{c_1c_2} = \frac{Cov(c_1, c_2)}{[V(c_1) \times V(c_2)]^{1/2}} \tag{13}$$

where c_1 and c_2 represent two traits.

The efficiency (E) and accuracy (r_m) with the use of H harvest years regarding the use of only one were estimated by the following equations, respectively (RESENDE et al., 2014):

$$E = \left\{ \frac{H}{[1 + (H-1)\rho]} \right\}^{1/2} \tag{14}$$

$$r_m = \left[\frac{H\rho}{H\rho + 1 - \rho} \right]^{1/2}$$

To identify biparental crosses with high stability, the Harmonic Mean of Genotypic Values ($HMGV_i$) was estimated for each (i-th) biparental cross (ALVES et al., 2018):

$$HMGV_i = \frac{H}{\sum_{h=1}^{H} {}^{1}/_{GV_{ih}}} \tag{15}$$

where GV_{ih} is the genotypic value (BLUP) of the i-th biparental cross in the h-th harvest year.

To identify biparental crosses with high stability, the Relative Performance of the Genotypic Values ($RPGV_i$) was estimated for each (i-th) biparental cross (ALVES et al., 2018):

$$RPGV_i = \frac{1}{H}\frac{\sum_{h=1}^{H} GV_{ih}}{\mu_h} \tag{16}$$

where μ_h is the mean in the h-th harvest year.

To identify biparental crosses with high productivity and witches' broom resistance, stability, and adaptability, the Harmonic Mean of the Relative Performance of the Genotypic Values ($HMRPGV_i$) was calculated for each (i-th) biparental cross (ALVES et al., 2018):

$$HMRPGV_i = \frac{H}{\sum_{h=1}^{H} \frac{1}{GV_{ih}/\mu_h}}$$

To select simultaneously for the four traits, the Additive Index (AI) was used (RESENDE, 2016):

$$AI_i = \sum_{1}^{4} \mathrm{CV}_{gc} \frac{HMRPGV_{ic}}{\sigma_{gc}} \tag{17}$$

where CV_g is the genotypic coefficient of variation [$CV_{gc} = {}^{100\sigma_{gc}}/_{\mu}$], used as weight and σ_{gc} is the genotypic standard deviation of the trait c. Positive weights ($+CV_g$) were assigned to NHF and DBW, and negative weights ($-CV_g$) to NWBF and FI.

Once the selection was made, genetic gains were estimated by:

$$SG\% = \left(\frac{GV_s}{\mu}\right) \tag{18}$$

where GV_s is the sum of the selected genotypes' genotypic values and μ is the general mean.

Statistical analyses were performed using the `SELEGEN REML/BLUP` (RESENDE, 2016) software and the `ASReml-R` package, version 4.1 (BUTLER et al., 2018) at the R environment (R CORE TEAM, 2023).

3 Results

The repeatability model with heterogeneous residual variances was more suitable than the repeatability model with homogeneous residual variance for all traits, as it had the lowest AIC

values (Table 2).

Table 2: Repeatability models with homogeneous (1) and heterogeneous (2) residual variances, number of estimated parameters (t), and Akaike information criterion (AIC) for the traits number of healthy fruits (NHF), number of fruits with witches' broom symptoms (NWBF), dry bean weight (DBW), and fruit index (FI) evaluated in 20 cacao biparental crosses

Model	t	AIC				Accuracy			
		NHF	DBW	NWBF	FI	NHF	DBW	NWBF	FI
1	4	11978.07	3,222.94	7,485.49	5,089.07	0.88	0.84	0.81	0.76
2	9	11701.18	2,875.58	7,466.58	4,953.64	0.91	0.89	0.90	0.87

All random effects of the repeatability model were significant, except for the GHI effect for FI. This behaviour was reflected in type B genotypic correlations across harvests, 0.91, 0.64, 0.77 and 0.7 for FI, NHF, NWBF and DBW, respectively (Table 3).

When considering the heteroscedastic repeatability model, the residual variances are individualized for each harvest (Table 3), allowing us to visualize the fluctuation of the residual effects in the phenotypic manifestation (Figure 1). As a result, the phenotypic variances and all parameters derived from these variances, such as the generalized heritability and coefficients of determination of permanent environment and GHI effects, as well as the repeatability coefficient, are also individualized (Figure 1). The behaviour of residual variances was similar for NHF and DBW, with increasing values until the sixth harvest year, where a sharp drop was registered, followed by slight fluctuations in the following harvests (Table 3). The residual variances for NWBF suggest a cyclical behaviour of the witches' broom disease. A less intense cyclic behaviour was also observed for FI.

Generalized heritability estimates for NHF, DBW, NWBF, and FI were 0.76, 0.65, 0.82, and 0.91, respectively. This parameter had greater variation for NHF (0.59 to 0.89) and DBW (0.38 to 0.84) and remained more uniform for NWBF (0.79 to 0.85) and FI (0.89 to 0.93; Figure 1a). A clear pattern was observed for the heritability: intense variations in the five first harvest years and stabilization in the other five. The repeatability coefficient had sharper fluctuations for all four traits, with mean values of 0.40, 0.39, 0.60, and 0.47, for NHF (0.25 to 0.55), DBW (0.19 to 0.57), NWBF (0.52 to 0.69), and FI (0.29 to 0.65), respectively (Figure 1b). The permanent environmental effects presented greater influences on NHF, DBW, and NWBF (Figure 1c). The GHI effects were more intense in NHF and DBW (Figure 1d).

Although the GHI effects were significant, the pairwise genotypic correlations between harvest years were high for all traits, except for the first and last harvest years of NHF and

Table 3: Variance component estimates (E) and genetic parameters (P) estimated from the model with heterogeneous residual variances for the traits number of healthy fruits (NHF), number of fruits with witches' broom symptoms (NWBF), dry bean weight (DBW), and fruit index (FI) evaluated in 20 cacao biparental crosses.

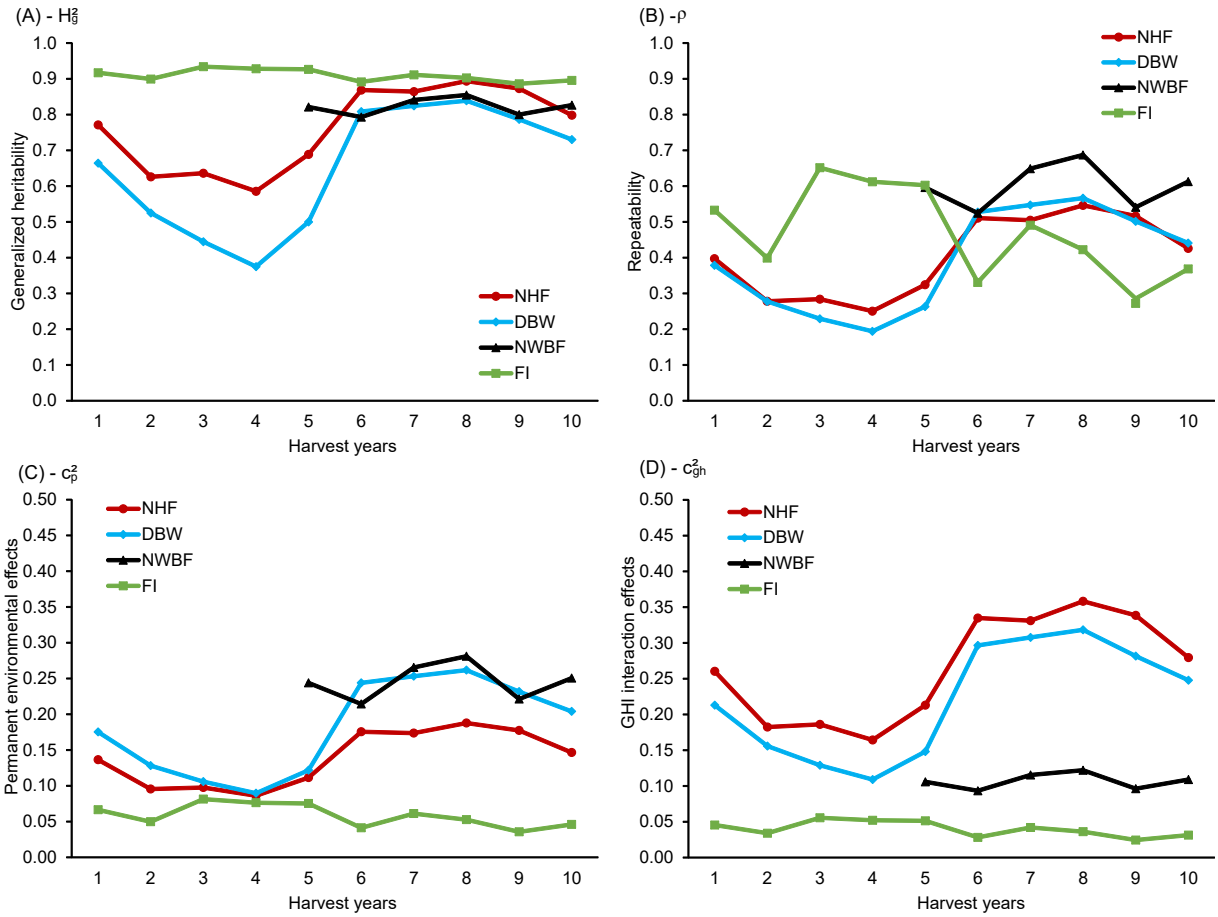
E/P[a]	Harvest year	NHF	DBW	NWBF	FI
σ^2_g	1 to 10	1153.68 ± 877.18	1.02 ± 0.82	2701.22 ± 2010.12	8.96 ± 3.04
σ^2_p	1 to 10	604.46 ± 179.79	0.88 ± 0.26	1870.28 ± 582.85	1.28 ± 0.58
σ^2_{gh}	1 to 10	1152.91 ± 305.66	1.08 ± 0.32	812.24 ± 325.73	0.87 ± 1.71
σ^2_ε	1	1516.89	2.05	–	8.09
	2	3410.97	3.89	–	14.56
	3	3284.08	5.34	–	4.6
	4	4103.83	6.86	–	5.62
	5	2504.78	4.26	2286.33	5.87
	6	532.76	0.64	3336.43	19.84
	7	570.81	0.51	1662.63	9.74
	8	308.59	0.39	1267.62	13.13
	9	494.48	0.83	3066.24	24.73
	10	1213.98	1.35	2078.77	16.65
σ^2_f	1	4427.94	5.03	–	19.2
	2	6322.02	6.87	–	25.66
	3	6195.14	8.31	–	15.7
	4	7014.89	9.83	–	16.72
	5	5415.83	7.24	7670.06	16.98
	6	3443.81	3.61	8720.16	30.95
	7	3481.86	3.48	7046.37	20.84
	8	3219.64	3.37	6651.35	24.24
	9	3405.53	3.8	8449.98	35.83
	10	4125.03	4.32	7462.50	27.76
r_{gh}	1 to 10	0.5	0.49	0.77	0.91
μ	1 to 10	138.84	5.16	214.26	27.38

[a] σ^2_g is the genotypic variance between parental crosses, σ^2_p is the permanent environmental effects variance, σ^2_{gh} is the genotypes × harvest years interaction (GHI) variance; $\sigma^2_{\varepsilon_h}$ is the residual variance of the h-th harvest year; σ^2_{fh} is the phenotypic variance of the h-th harvest year, r_{gh} is the type B genotypic correlation across harvest years, and μ is the phenotypic mean

DBW (Figure 2). This is an indication that GHI was simple (non-crossover). As suggested by previous parameters, NWBF and FI have higher temporal uniformity than the other two traits.

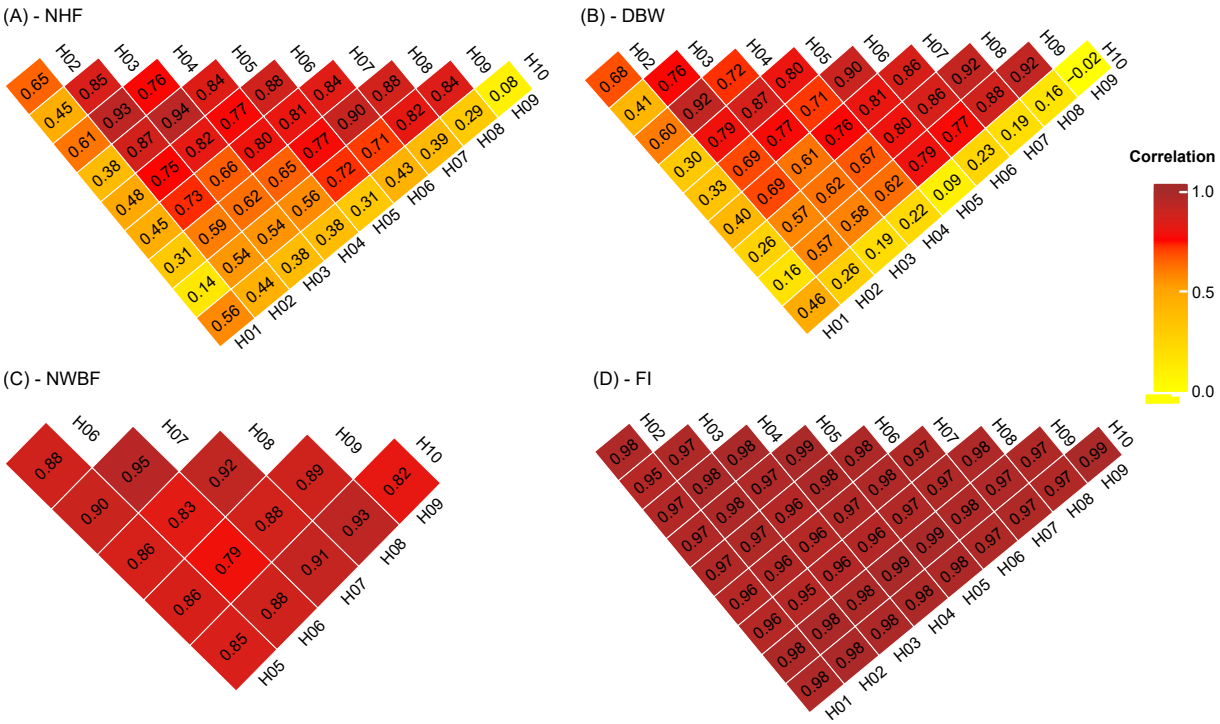
The different patterns observed between the repeatability coefficients for the traits under study suggest a differential behaviour in terms of accuracy and selection efficiencies over the harvest years (Figure 3). The interpretation of these two parameters allows for estimating the

Figure 1: Estimates of genetic parameters in each harvest year: (a) generalized heritability (H_g^2); (b) repeatability coefficient (ρ); (c) coefficient of determination of permanent environmental effects (c_p^2); and (d) coefficient of determination of genotypes × harvest years interaction (GHI) effects (c_{gh}^2), for the traits number of healthy fruits (NHF), dry bean weights (DBW), number of fruits with witches' broom symptoms (NWBF), and fruit index (FI) evaluated in 20 cacao biparental crosses



optimal number of harvest years for selecting aiming at recombination (minimum accuracy of 0.7) and recommendation (minimum accuracy of 0.9), as Resende and Duarte (2007) suggested. The number of healthy fruits, DBW, and FI reached an accuracy of 0.7 in the second harvest year, meaning that in only a couple of harvest years, breeders can select for recombination and advancement of the breeding cycle (Figure 3a). For NWBF, only a single measurement (in this case, in the fifth harvest year) can be considered for performing the selection for recombination (Figure 3a). This fact was also reflected in the efficiency, which grows at lower rates with the increase in the number of harvest years (Figure 3b). When the objective is the recommendation of biparental crosses, NHF, DBW, and NWBF coincidently pointed to the seventh harvest year (Figure 3a). Fruit index, as a more stable trait, reached the minimum recommended value in

Figure 2: Genotypic correlations between harvest years (HY) for the traits (a) number of healthy fruits (NHF), (b) dry bean weights (DBW), (c) number of fruits with witches' broom symptoms (NWBF), (d) and fruit index (FI) evaluated in 20 cacao biparental crosses



the fifth harvest year (Figure 3a). In other words, considering the seventh harvest year as the optimum for the recommendation, the breeder can reduce the breeding cycle by 25%.

Similar patterns between NHF and DBW can be explained by the high genotypic correlation between these two traits (Table 4). In fact, among the evaluated traits, NHF and DBW are the only traits with a high and positive genotypic correlation, which allows indirect selection, that is, selection in one will provide indirect gain in the other. It should be noted that NHF allows for a quick and less costly evaluation than DBW since the determination of this trait requires the breaking of the fruits, followed by fermentation and drying of the seeds.

There was 90% agreement regarding the order of the biparental crosses with the highest genotypic values and the greatest HMRPGV, and for the five best this coincidence was 100% for the four traits. For this reason, we chose to present only the performance of biparental crosses regarding HMRPGV (Table 5).

According to the Additive Index, the biparental crosses E1H02, E1H04, E1H07, E1H15, and E1H20 (Table 5) stood out. Their selection will provide gains of 21.1, 19.0, 2.3%, and 0.2% for NHF, DBW, NWBF and FI, respectively.

Figure 3: Accuracy (a) and efficiency (b) with the use of several harvest years regarding the use of only one for the traits number of healthy fruits (NHF), dry bean weight (DBW), number of fruits with witches' broom symptoms (NWBF), and fruit index (FI) evaluated in 20 cacao biparental crosses

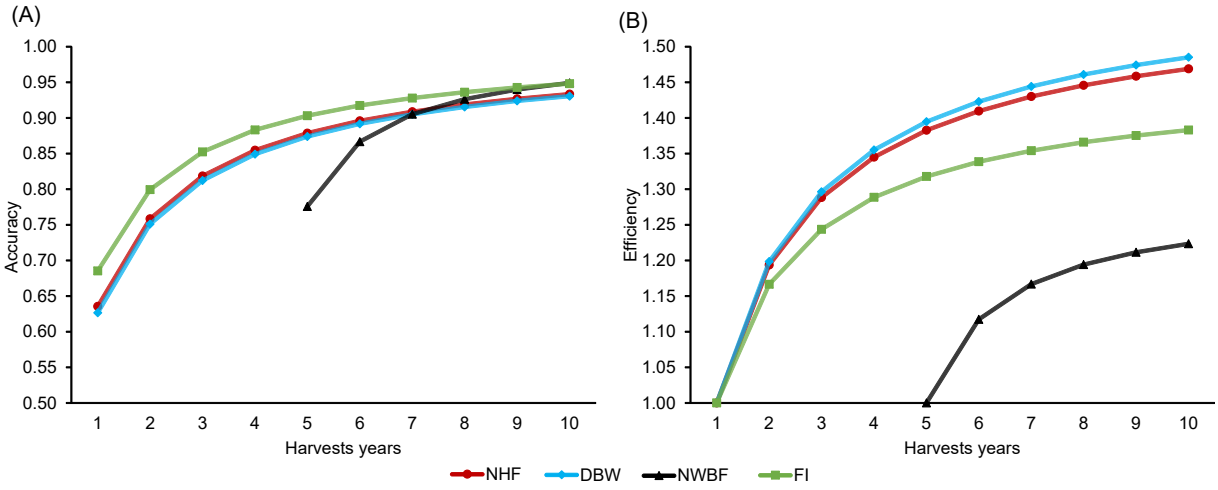


Table 4: Genotypic correlations (above the diagonal) and associated p-values (below the diagonal) between the traits number of healthy fruits (NHF), number of fruits with witches' broom symptoms (NWBF), dry bean weight (DBW), and fruit index (FI) evaluated in 20 cacao biparental crosses.

Traits	NHF	DBW	NWBF[a]	FI
NHF	–	.91*	.11	.50*
DBW	.00	–	.00	.11
NWBF[a]	.65	.99	–	.28
FI	.03	.64	.23	–

[a] Assessed from the fifth harvest onwards
* Significant at the 5% level by the Student's t-test

4 Discussion

Perennial fruit trees, such as cacao, are subject to constant environmental changes, especially of climatic origin, during the production cycle (LAHIVE et al., 2019). These conditions can promote heterogeneity of residual variances between harvest years (TAHI et al., 2019). This explains the better fit of the heteroscedastic model used in this study and highlights the importance of modelling the residual variances. By modelling, one can elevate the selective accuracy, as observed herein for all traits.

Furthermore, greater safety and sustainability for the crop are provided by selecting genotypes that are not only productive and resistant, but with high temporal stability and adaptability of genotypic values, that is, that have the resilience to adapt to different climatic

Table 5: Harmonic mean of the relative performance of genotypic values (HMRPGV) and additive index (AI) for the traits number of healthy fruits (NHF), number of fruits with witches' broom symptoms (NWBF), dry bean weight (DBW), and fruit index (FI) evaluated in 20 cacao biparental crosses (BP)

BP	NHF	DBW	NWBF	FI	AI
E1H01	0.914	0.987	1.020	0.942	0.159
E1H02[a]	1.500	1.496	0.803	0.983	0.261
E1H03	0.563	0.785	0.685	0.831	0.123
E1H04[a]	0.985	1.136	0.729	0.890	0.191
E1H05	0.766	0.920	0.853	0.904	0.147
E1H06	0.650	0.761	0.643	0.936	0.115
E1H07[a]	1.056	1.093	1.324	0.977	0.177
E1H08	0.303	0.454	1.052	0.926	0.051
E1H09	0.958	1.079	1.618	0.922	0.175
E1H10	0.532	0.743	1.087	0.857	0.111
E1H11	1.279	1.089	1.323	1.139	0.172
E1H12	0.598	0.489	1.065	1.188	0.051
E1H13	0.874	0.801	1.137	1.105	0.116
E1H14	1.141	1.028	1.230	1.078	0.162
E1H15[a]	1.256	1.290	0.781	0.957	0.220
E1H16	0.946	0.957	0.826	1.014	0.151
E1H17	1.144	1.081	0.997	1.035	0.175
E1H18	0.940	0.965	0.645	0.988	0.155
E1H19	1.235	1.060	0.824	1.119	0.170
E1H20[a]	1.556	1.274	1.100	1.203	0.209

[a] Top five biparental crosses according to the additive index.

conditions (DIAS; KAGEYAMA, 1998; FARRELL et al., 2018; SCHROTH et al., 2016). It is important to bear in mind that the producer can cultivate the same plant for decades in the same local. This peculiarity explains the importance of studying temporal stability and adaptability (DIAS; KAGEYAMA, 1998).

In general, crossing between cacao genotypes from different geographic origins results in high genetic variability in the hybrid population (LÓPEZ et al., 2021; RODRIGUES et al., 2020). The significance of the GHI effects, for three of the four traits, suggests differential performance of the biparental crosses over the harvest years. This differentiated performance was also observed by Benjamin et al. (2016) for resistance to witches' broom and Tahi et al. (2019) for the production of healthy fruits. The concomitant analysis of the results of this study with those of the previously cited studies suggests that the phenotypic expression, both for production and for resistance to witches' broom, depends not only on the presence of favourable alleles but also on the macro- and micro-environmental experimental conditions, which varies

year by year. On the other hand, the type B genotypic correlation across harvest years observed herein was moderate for NHF and high for the other traits, suggesting that the GHI effects do not result in ranking modifications (GEZAN et al., 2017; RESENDE; ALVES, 2020).

The generalized heritability coefficients of the evaluated traits were mostly > 0.50, offering good selection perspectives. Heritability is a parameter that depends not only on genetic variability but also on the influence of uncontrolled environmental factors in the trial (FALCONER; MACKAY, 1996; SCHMIDT et al., 2019). This parameter's variation follows the residual variation over the harvest years. This denotes that selective accuracies, that is, correlations between predicted and true genotypic values (MRODE, 2014), can also vary.

The same interpretation can be extended to the repeatability coefficient. There was a high range of values for the four traits over the years, that is, coefficients considered low ($\rho \leq 0.3$) and high ($\rho \geq 0.6$) were estimated. According to Resende and Alves (2020) and considering the average values, NHF, DBW, and FI showed median repeatability, and NWBF had high values of this parameter. Medium to high values for this coefficient denote good genotypes' ability to perform similarly over time (FERREIRA et al., 2021; MALIKOUSKI et al., 2021). This was attested by analysing the pairwise genotypic correlations between harvest years, high for all traits. These high genotypic correlations are based on the use of a repeatability model, efficient when this assumption is met (MRODE, 2014).

The constancy of the biparental crosses' performance throughout time reflects not only in the relation between harvest years but also in the selective accuracy. As several harvest years are evaluated and considered in the analysis, the precision and accuracy of the selection can be improved (FERREIRA et al., 2020). However, recall the breeder's equation [$G = ir\sigma_g^2/t$, where G is the selection gain, i is the selection intensity, r is the accuracy, and t is the time], that is, the higher the time spent to select, the lesser the gains per time unit. In that context, the results presented in this work offer relevant information for cacao breeding. Considering an average cycle duration of 12 yr (2 yr for the juvenile period included), cacao breeders can save 3 yr if the goal is to select the recommendation for all traits. When the objective is the recombination of genotypes, the economy of time—and resources—goes even further, as only two harvest years are enough for NHF, DBW, and FI, agreeing with the results of Tahi et al. (2019). The trait NWBF demands more time only if the breeder opts for the natural infection of the plants.

In Brazil, early works recommended that the first measurement be carried out only when the plants reached yield maturity when there was less fluctuation in the genotypes' performance

(CARVALHO et al., 2003, 2002; DIAS; KAGEYAMA, 1998). Indeed, the present study can attest to this behaviour by analysing jointly the lack of complexity of GHI, high correlations for all traits, and the heritability stabilization from the sixth harvest year onwards. Thus, as these studies suggest, it is unnecessary to evaluate several harvest years at this stage of maturity to perform the genetic selection. Those authors also recommend two consecutive harvest years for selection, but only after the pre-climax period (i.e., before the sixth harvest year) has passed. The present study shows that this period must be taken into account to abbreviate even further the evaluation period. Furthermore, the results achieved herein were obtained via REML/BLUP, which grants this study greater reliability, as it is more suitable for situations normally found in multi-harvest trials with cacao. Therefore, such results may help to optimize intrapopulation recurrent selection in cacao breeding programs.

Although simple, the GHI was significant for NHF, NWBF, and DBW. This significance justifies the use of a methodology that capitalizes on the GHI effects since the objective was to elevate the selective accuracy and select the most stable and adaptable biparental crosses. The HMRPGV, which ranks the biparental crosses based on stability, adaptability, and productivity and/or resistance simultaneously, proved to be adequate to identify the genotypes that met these requirements, showing that those with greater genotypic values are also those with greater stability and adaptability, a fact also observed by Evangelista et al. (2020) and Peixoto et al. (2021). The selection of these biparental crosses will incorporate into the breeding program resilient genotypes, capable of adapting to micro- and macro-environmental variations, which, ultimately, will promote greater security for producers.

In cacao breeding, there is a need to select not only productive genotypes but ones that are resistant to the main pests and diseases, that have vegetative characteristics favourable to sustainable management (i.e., appropriate size) and that produce beans with qualitative aspects that meet certain niches (BEKELE; PHILLIPS-MORA, 2019). Consequently, several traits are evaluated, and selection indexes are usually needed (JAIMEZ et al., 2020; MUSTIGA et al., 2018). In this study, there is a distinction between the best biparental crosses for each trait individually, especially concerning yield and resistance traits. These traits have non-significant genotypic correlation, meaning that the selection for one trait may not prioritize the best biparental crosses for the other. One way to select simultaneously for all traits is to use selection indexes. The use of the Additive Index together with the HMRPGV allowed the identification of the five best biparental crosses: E1H02, E1H04, E1H07, E1H15, and E1H20.

In addition to satisfactory performance for the four evaluated traits, these biparental crosses have high temporal stability and adaptability, and therefore, are suitable to be incorporated into the breeding program for future hybridizations and/or used in competition trials for recommendation.

References

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

ALVES, R. S.; PEIXOTO, L. A.; TEODORO, P. E.; SILVA, L. A.; RODRIGUES, E. V.; RESENDE, M. D. V.; LAVIOLA, B. G.; BHERING, L. L. Selection of *Jatropha curcas* families based on temporal stability and adaptability of genetic values. **Industrial Crops and Products**, v. 119, p. 290–293, 2018.

ANDRADE, V. T.; GONÇALVES, F. M. A.; NUNES, J. A. R.; BOTELHO, C. E. Statistical modeling implications for coffee progenies selection. **Euphytica**, v. 207, n. 1, p. 177–189, 2016.

ATLIN, G. N.; CAIRNS, J. E.; DAS, B. Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. **Global Food Security**, v. 12, p. 31–37, 2017.

BAILEY, B. A.; EVANS, H. C.; PHILLIPS-MORA, W.; ALI, S. S.; MEINHARDT, L. W. *Moniliophthora roreri*, causal agent of cacao frosty pod rot: Frosty pod rot of cacao. **Molecular Plant Pathology**, v. 19, n. 7, p. 1580–1594, 2018.

BEKELE, F. L.; PHILLIPS-MORA, W. Cacao (*Theobroma cacao* L.) Breeding. In: AL-KHAYRI, J. M.; JAIN, S. M.; JOHNSON, D. V. (Eds.). **Advances in Plant Breeding Strategies: Industrial and Food Crops**. Cham: Springer International Publishing, 2019. P. 409–487.

BENJAMIN, C. S.; LUZ, E. D. M. N.; SANTOS, W. O.; PIRES, J. L. Cacao families and parents selected as resistant to natural infection of *Moniliophthora perniciosa*. **Crop Breeding and Applied Biotechnology**, v. 16, p. 141–146, 2016.

BUTLER, D. G.; CULLIS, B. R.; GILMOUR, A. R.; GOGEL, B. J.; THOMPSON, R. **ASReml-R reference manual Version 4**. Hemel Hempstead, UK: VSN International, 2018.

CARVALHO, C.; ALMEIDA, C.; CRUZ, C.; MACHADO, P. Hybrid cocoa tree adaptability and yield temporal stability in Rondônia State, Brazil. **Crop Breeding and Applied Biotechnology**, v. 3, n. 3, p. 237–244, 2003.

CARVALHO, C.; CRUZ, C.; ALMEIDA, C.; MACHADO, P. Yield repeatability and evaluation period in hybrid cocoa assessment. **Crop Breeding and Applied Biotechnology**, v. 2, n. 1, p. 149–156, 2002.

CHAVES, S. F. S.; ALVES, R. M.; ALVES, R. S.; SEBBENN, A. M.; RESENDE, M. D. V.; DIAS, L. A. S. *Theobroma grandiflorum* breeding optimization based on repeatability, stability and adaptability information. **Euphytica**, v. 217, n. 12, p. 211, 2021.

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, n. 4, p. 381–393, 2006.

DIAS, L. A. S. **Melhoramento genético do cacaueiro**. Goiás: FUNAPE-UFG, 2001.

DIAS, L. A. S.; CRUZ, C. D.; CARNEIRO, P. C. S. Analysis of experiments with repeated measures. **Ingenic Newsletter**, v. 6, p. 29–31, 2001.

DIAS, L. A. S.; KAGEYAMA, P. Y. Temporal stability of multivariate genetic divergence in cacao (*Theobroma cacao* L.) in Southern Bahia conditions. **Euphytica**, v. 93, n. 2, p. 181–187, 1997.

DIAS, L. A. S.; KAGEYAMA, P. Y. Repeatability and minimum harvest period of cacao (*Theobroma cacao* L.) in Southern Bahia. **Euphytica**, v. 102, p. 29–35, 1998.

DIAS, L. A. S.; SOUZA, C. A. S.; AUGUSTO, S. G.; SIQUEIRA, P. R.; MÜLLER, M. W. Período mínimo de colheita para avaliação de cultivares de cacau em Linhares-ES. **Revista Árvore**, v. 27, n. 4, p. 495–501, 2003.

DUVAL, A.; GEZAN, S. A.; MUSTIGA, G.; STACK, C.; MARELLI, J.-P.; CHAPARRO, J.; LIVINGSTONE, D.; ROYAERT, S.; MOTAMAYOR, J. C. Genetic Parameters and the Impact of Off-Types for *Theobroma cacao* L. in a Breeding Program in Brazil. **Frontiers in Plant Science**, v. 8, p. 2059, 2017.

EVANGELISTA, J. S. P. C.; ALVES, R. S.; PEIXOTO, M. A.; RESENDE, M. D. V.; TEODORO, P. E.; SILVA, F. L.; BHERING, L. L. Soybean productivity, stability, and adaptability through mixed model methodology. **Ciência Rural**, v. 51, 2020.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics**. 4. ed. Harlow, England: Pearson Prentice Hall, 1996.

FARRELL, A. D.; RHINEY, K.; EITZINGER, A.; UMAHARAN, P. Climate adaptation in a minor crop species: is the cocoa breeding network prepared for climate change? **Agroecology and Sustainable Food Systems**, v. 42, n. 7, p. 812–833, 2018.

FERREIRA, F. M.; ROCHA, J. R. A. S. C.; ALVES, R. S.; ELIZEU, A. M.; BENITES, F. R. G.; RESENDE, M. D. V.; SOUZA SOBRINHO, F.; BHERING, L. L. Estimates of repeatability coefficients and optimum number of measures for genetic selection of *Cynodon spp*. **Euphytica**, v. 216, n. 5, p. 70, 2020.

FERREIRA, F. M.; ROCHA, J. R. A. S. C.; BHERING, L. L.; FERNANDES, F. D.; LÉDO, F. J. S.; RANGEL, J. H. A.; KOPP, M.; CÂMARA, T. M. M.; SILVA, V. Q. R.; MACHADO, J. C. Optimal harvest number and genotypic evaluation of total dry biomass, stability, and adaptability of elephant grass clones for bioenergy purposes. **Biomass and Bioenergy**, v. 149, p. 106104, 2021.

GEZAN, S. A.; CARVALHO, M. P.; SHERRILL, J. Statistical methods to explore genotype-by-environment interaction for loblolly pine clonal trials. **Tree Genetics & Genomes**, v. 13, n. 1, p. 1, 2017.

GILMOUR, A. R.; GOGEL, B. J.; CULLIS, B. R.; WELHAM, S. J.; THOMPSON, R. **ASReml User Guide Release 4.1 Structural Specification**. Hemel Hempstead, UK: VSN International, 2015.

HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a selection model. **Biometrics**, v. 31, n. 2, p. 423, 1975.

JAIMEZ, R. E.; VERA, D. I.; MORA, A.; LOOR, R. G.; BAILEY, B. A. A disease and production index (DPI) for selection of cacao (*Theobroma cacao*) clones highly productive and tolerant to pod rot diseases. **Plant Pathology**, v. 69, n. 4, p. 698–712, 2020.

LAHIVE, F.; HADLEY, P.; DAYMOND, A. J. The physiological responses of cacao to the environment and the implications for climate change resilience. A review. **Agronomy for Sustainable Development**, v. 39, n. 1, p. 5, 2019.

LÓPEZ, M. E.; RAMÍREZ, O. A.; DUBÓN, A.; RIBEIRO, T. H. C.; DÍAZ, F. J.; CHALFUN-JUNIOR, A. Sexual compatibility in cacao clones drives arrangements in the field leading to high yield. **Scientia Horticulturae**, v. 287, p. 110276, 2021.

MALIKOUSKI, R. G.; PEIXOTO, M. A.; MORAIS, A. L.; ELIZEU, A. M.; ROCHA, J. R. A. S. C.; ZUCOLOTO, M.; BHERING, L. L. Repeatability coefficient estimates and optimum number of harvests in graft/rootstock combinations for 'tahiti' acid lime. **Acta Scientiarum. Agronomy**, v. 43, e51740–e51740, 2021.

MOTAMAYOR, J. C.; LACHENAUD, P.; MOTA, J. W. S.; LOOR, R.; KUHN, D. N.; BROWN, J. S.; SCHNELL, R. J. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). **PLOS ONE**, v. 3, n. 10, e3311, 2008.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. 3rd ed. Boston, MA: CABI, 2014.

MUSTIGA, G. M.; GEZAN, S. A.; PHILLIPS-MORA, W.; ARCINIEGAS-LEAL, A.; MATA-QUIRÓS, A.; MOTAMAYOR, J. C. Phenotypic description of *Theobroma cacao* L. for yield and vigor traits from 34 hybrid families in Costa Rica based on the genetic basis of the parental population. **Frontiers in Plant Science**, v. 9, p. 808, 2018.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545–554, 1971.

PEIXOTO, M. A.; EVANGELISTA, J. S. P. C.; ALVES, R. S.; FARIAS, F. J. C.; CARVALHO, L. P.; TEODORO, L. P. R.; TEODORO, P. E.; BHERING, L. L. Models for optimizing selection based on adaptability and stability of cotton genotypes. **Ciência Rural**, v. 51, n. 5, e20200530, 2021.

PIEPHO, H.-P.; MÖHRING, J.; MELCHINGER, A. E.; BÜCHSE, A. BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, v. 161, n. 1-2, p. 209–228, 2008.

R CORE TEAM. **R: A Language and environment for statistical computing**. Viena, Áustria: R Foundation for Statistical Computing, 2023.

RESENDE, M. D. V. **Genética biométrica e Estatística no melhoramento de plantas perenes**. 1. ed. Brasilia: Embrapa Informação Tecnológica, 2002.

RESENDE, M. D. V. Software Selegen-REML/BLUP: a useful tool for plant breeding. **Crop Breeding and Applied Biotechnology**, v. 16, n. 4, p. 330–339, 2016.

RESENDE, M. D. V.; ALVES, R. S. Linear, generalized, hierarchical, bayesian and random regression mixed models in genetic/genomics in plant breeding. **Functional Plant Breeding Journal**, v. 2, n. 2, p. 1–31, 2020.

RESENDE, M. D. V.; DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, v. 37, n. 3, p. 182–194, 2007.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão, aleatória, seleção genômica, QTL, GWAS, estatística espacial e temporal, competição, sobrevivência.** Viçosa, MG, Brazil: UFV, 2014.

RODRIGUES, G. S.; PIRES, J. L.; LUZ, E. D. M. N. Association of genes from different sources of resistance to major cacao diseases. **Revista Ceres**, v. 67, n. 5, p. 383–394, 2020.

SCHMIDT, P.; HARTUNG, J.; BENNEWITZ, J.; PIEPHO, H.-P. Heritability in plant breeding on a genotype-difference basis. **Genetics**, v. 212, n. 4, p. 991–1008, 2019.

SCHROTH, G.; LÄDERACH, P.; MARTINEZ-VALLE, A. I.; BUNN, C.; JASSOGNE, L. Vulnerability to climate change of cocoa in West Africa: Patterns, opportunities and limits to adaptation. **Science of The Total Environment**, v. 556, p. 231–241, 2016.

SILVA NETO, P. J.; MATOS, P. G. G.; MARTINS, A. C. S.; SILVA, A. P. **Manual técnico do cacaueiro para a Amazônia Brasileira**. 1. ed. Belém, Pará: CEPLAC/SUEPA, 2013.

TAHI, M.; TREBISSOU, C.; RIBEYRE, F.; GUIRAUD, B. S.; POKOU, D. N.; CILAS, C. Variation in yield over time in a cacao factorial mating design: changes in heritability and longitudinal data analyses over 13 consecutive years. **Euphytica**, v. 215, n. 6, p. 106, 2019.

THOMAS, E.; ZONNEVELD, M.; LOO, J.; HODGKIN, T.; GALLUZZI, G.; ETTEN, J. Present spatial diversity patterns of *Theobroma cacao* L. in the neotropics reflect genetic differentiation in pleistocene refugia followed by human-influenced dispersal. **PLOS ONE**, v. 7, n. 10, e47676, 2012.

VERBYLA, A. P.; FAVERI, J.; DEERY, D. M.; REBETZKE, G. J. Modelling temporal genetic and spatio-temporal residual effects for high-throughput phenotyping data. **Australian & New Zealand Journal of Statistics**, v. 63, n. 2, p. 284–308, 2021.

WICKRAMASURIYA, A. M.; DUNWELL, J. M. Cacao biotechnology: current status and future prospects. **Plant Biotechnology Journal**, v. 16, n. 1, p. 4–17, 2018.

ZARRILLO, S.; GAIKWAD, N.; LANAUD, C.; POWIS, T.; VIOT, C.; LESUR, I.; FOUET, O.; ARGOUT, X.; GUICHOUX, E.; SALIN, F.; SOLORZANO, R. L.; BOUCHEZ, O.; VIGNES, H.; SEVERTS, P.; HURTADO, J.; YEPEZ, A.; GRIVETTI, L.; BLAKE, M.; VALDEZ, F. The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. **Nature Ecology & Evolution**, v. 2, n. 12, p. 1879–1888, 2018.

CHAPTER 2

APPLICATION OF LINEAR MIXED MODELS FOR MULTIPLE HARVEST/SITE TRIAL ANALYSES IN PERENNIAL PLANT BREEDING

Published article: CHAVES, S. F. S.; EVANGELISTA, J. S. P. C.; ALVES, R. S.; FERREIRA, F. M.; DIAS, L. A. S.; ALVES, R. M.; DIAS, K. O. G.; BHERING, L. L. Application of linear mixed models for multiple harvest/site trial analyses in perennial plant breeding. **Tree Genetics & Genomes**, v. 18, n. 6, p. 44, 2022.

The optimization of perennial plant breeding necessarily involves the evaluation of multi-harvest and/or multi-site trials. In these situations, modelling covariance structures can elevate accuracy. This study aimed to evaluate different covariance structures for multi-harvest and multi-site trial analyses, using two datasets (D1 and D2). In D1, 25 hybrids of *Theobroma grandiflorum* were evaluated in a complete randomized block design, during twelve consecutive harvest years. In D2, 215 clones of *Eucalyptus* spp. were evaluated in a complete randomized block design, in four sites. For both datasets, the covariance structures of the random effects were modelled, and their adequacy was tested by the Akaike and Bayesian information criteria. From the selected model, the variance components and genetic parameters were estimated. We also compared the expected genetic gains and the rankings of genotypes based on the genotypic values provided by the basic and the selected models. For D1, the third-order factor analytic model was the most suitable for genetic effects, while for D2, the unstructured model showed the best fit for such effects. The models provided a better insight into the variances dynamics over the harvest years/sites. The genetic gains were 3.52 percentage points higher in D1 and did not change in D2. Despite similar results, the standard model, modelled with covariance structures that assume homogeneity of covariances, was not the most statistically appropriate model for D2 according to the information criteria. Therefore, the modelling of covariance structures can and should be used in the genetic evaluation of perennial plants.

1 Introduction

Perennial species have been studied as an alternative for more sustainable production systems (CREWS et al., 2016; CREWS; CATTANI, 2018). These crops do not have to be planted every year, which may conserve the physical and chemical properties of the soil (e.g., water uptake efficiency, nutrient and soil retention, carbon balance, and microbiome functions) and reduce the necessity of some agricultural interventions such as manual or chemical weeding (BAKER, 2017; CHAVES et al., 2020; WEISSHUHN et al., 2017). However, the challenges for the development of productive perennial cultivars are even greater than those found for short-cycle species, given their exposure to fluctuating weather (ISABEL et al., 2020; RAI; SHEKHAWAT, 2014). Perennial varieties can be impacted by frosts, drought, winds, heat stress, and pressure from disease and insect pests, which can mislead the selection of superior genotypes. Furthermore, the use of appropriate statistical models that can capture the impacts of this environmental seasonality is essential for a more accurate genetic evaluation of perennials. It is important to bear in mind that well-designed and properly conducted trials are prerequisites for the successful application of statistical models to rank genotypes in a breeding program.

Perennial plant breeders usually deal with one of two situations: the first one is related to species that produce periodically (e.g., orchard species); the second one is related to species that are grown for several years and harvested at the end of the cycle (e.g., timber species). Hence, the datasets available for the breeder's decision-making will consist of measurements of a genotype over harvests (multi-harvest) or sites (multi-site), or even both.

Multi-harvest trial (MHT) and multi-site trial (MST) data contain genetic and non-genetic sources of variation that can complicate the identification of the best genotypes (SINGH et al., 2013). Additionally, it is plausible to assume that phenotypes on the same genotype are correlated over harvests and/or sites (FERRÃO et al., 2017; KRAUSE et al., 2020; PIEPHO; ECKL, 2014). Fortunately, it is possible to find patterns for these sources of variation by modelling the covariance structures between harvests and/or sites. Classical methods, such as the ones based on least squares, do not deal with heterogeneity, as they rely on the assumption that the variance is homogeneous and independent (ARMSTRONG, 2017; HAVERKAMP; BEAUDUCEL, 2017), which is often untrue in these studies (CROWDER; HAND, 2020; EEUWIJK et al., 2019; HU et al., 2013). On the other hand, mixed models allow us to deal with heterogeneity and dependency between harvests and/or sites by fitting covariance structures for the random effects of the statistical model (MALOSETTI et al., 2013).

In MHT data analysis, a better fitting and more insightful model is achieved when a multilevel linear model (MLM) with an unstructured covariance matrix (UN) is used, compared to MLM with compound-symmetry covariance matrix (CS) or a repeated measures analysis of variance (rANOVA), especially when a substantial number of missing values is present (HAVERKAMP; BEAUDUCEL, 2017). For MST, the UN covariance matrix provides the most complete information, despite the overparameterization when a large set of environments is being analysed or there is much imbalance in the testing of all genotypes at each site (GEZAN et al., 2017). The factor analytic (FAk) models are useful alternatives in these situations (PIEPHO, 1997) since they allow the interpretation of high-dimensional data and the study of stability via the factor loadings and scores (PIEPHO, 1998; SMITH et al., 2001). However, depending on the population breeding history and the true patterns of covariance, the use of more parameterized models may not be advantageous. Therefore, the best way to find patterns for some of these sources of heterogeneity is by testing many covariance structures for the random effects of the statistical model (ISIK et al., 2017a, chapter 8).

The present study was carried out through the analyses of two datasets (D1 and D2). D1 refers to a cupuassu tree [*Theobroma grandiflorum* Willd. (Ex. Spreng.) Schumm.] population, assessed in twelve consecutive harvest years, and D2 refers to a eucalyptus (*Eucalyptus* spp.) population composed of interspecific hybrids, evaluated in four sites. Our objectives were (i) to study the implications of modelling the covariance structures of the random effects in the estimation of variance components and prediction of genotypic values and (ii) to compare the genetic gains from two models with different covariance structures for the random effects: the standard model, modelled with covariance structures that assume homogeneity of covariances, and the best-fitted model, modelled with different covariance structures, using two datasets from perennial plant breeding programs (MHT and MST).

2 Material and Methods

2.1 Datasets

Two datasets were analysed herein. Coded as D1 and D2, they refer to the cupuassu tree and eucalyptus populations, respectively. The former refers to MHT, and the latter refers to MST.

Cupuassu tree dataset (D1)

In D1, 25 cupuassu tree hybrids were evaluated in a single-site trial, arranged in a complete randomized block design with five blocks and three plants per plot. The plants within a plot are full-sib, i.e., individuals that came from seeds of the same crossing. The hybrids were planted in February 2005, at 6×4 m spacing. Data collection began 3 years after planting at the plot level. Since the cupuassu tree produces from December to mid-June in the trial's region, we divided the data collection into four evaluations for better sampling. The sum of all evaluations represents the data from a harvest year. We repeated this dynamic in all the 12 harvest years. In the mathematical notation in the following topics, v is the number of cupuassu tree genotypes ($i = 1, 2, \ldots, 25$), q is the number of blocks ($r = 1, 2, \ldots, 5$), and h is the number of harvest years ($c = 1, 2, \ldots, 12$). The evaluated trait was average fruit yield per plant, obtained by multiplying the number of fruits per tree by the average weight of the fruits. The geographic coordinates and average climatic conditions of the harvest years are presented in Table 1.

Table 1: Geographic location (latitude, longitude and altitude) and annual rainfall for the harvests and sites evaluated in the experiments corresponding to the datasets of *Theobroma grandiflorum* (D1) and *Eucalyptus* (D2), respectively.

Dataset	Harvests (H)/Site (S)	Geographic coordinates		Altitude (m)	Rainfall (mm)
		Latitude	Longitude		
D1	H01 to H12	02º32'53.1" S	48º15'52.8" W	45	2716
D2	S1	30°11'09" S	52°00'10" W	141	1422
	S2	30°29'45" S	52°19'35" W	378	1564
	S3	30°27'19" S	52°39'53" W	250	1368
	S4	30°14'46" S	53°49'7" W	301	1133

Eucalyptus tree dataset (D2)

The D2 population consists of 215 eucalyptus (*Eucalyptus* spp.) interspecific hybrid clones. These data were used and made publicly available by (ALVES et al., 2020). The experiment was carried out in four sites, using a complete randomized block design with one plant per plot (single-tree plot), and 30 blocks per site. In the mathematical notation below, t represents the number of eucalyptus genotypes ($l = 1, 2, \ldots, 215$), b is the number of blocks ($w = 1, 2, \ldots, 30$), and m is the number of sites ($j = 1, 2, 3, 4$). The trees were planted at a spacing of 3.5×2.6 m. Thirty-six months after planting, the diameter at breast height (DBH, in

cm) was measured using a diameter tape. The geographic coordinates and climatic conditions of the crops are presented in Table 1.

2.2 Statistical analyses

The analyses were performed using a linear mixed model, and the variance components were estimated by residual maximum likelihood (REML) (PATTERSON; THOMPSON, 1971). The genotypic values were predicted by best linear unbiased prediction (BLUP) (HENDERSON, 1975).

Cupuassu tree dataset (D1)

The following model considering multiple harvests was fitted by:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}_1\mathbf{a} + \mathbf{X}_2\mathbf{d} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{p} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the $n \times 1$ vector of harvests yield, where n is the number of observations; μ is the intercept, linked to $\mathbf{y}$ by a $n \times 1$ vector of ones ($\mathbf{1}$); $\mathbf{a}$ is the $h \times 1$ vector of the fixed effects of harvest years, with a $n \times h$ design matrix ($\mathbf{X}_1$); $\mathbf{d}$ is the $qh \times 1$ vector of the fixed effects of blocks within harvest years, with an $n \times qh$ incidence matrix ($\mathbf{X}_2$); $\mathbf{g}$ is the $vh \times 1$ vector of genotypic random effects within harvest years [$\mathbf{g} \sim \text{MVN}(0, \mathbf{G} \otimes \mathbf{I}_v)$], with $\mathbf{Z}_1$ as a $n \times vh$ matrix linking $\mathbf{g}$ to $\mathbf{y}$; $\mathbf{p}$ is the $vq \times 1$ vector of permanent environmental random effects [$\mathbf{p} \sim \text{MVN}(\mathbf{0}, \sigma_p \mathbf{I}_{vq})$], with a $n \times vq$ design matrix ($\mathbf{Z}_2$); and $\mathbf{e}$ is the $n \times 1$ vector of residual random effects [$\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{R} \otimes \mathbf{I}_n)$]. $\mathbf{G}$ and $\mathbf{R}$ are covariance matrices of the genotypic and residual effects, respectively. $\otimes$ is a Kronecker product, and $\mathbf{I}$ is an identity matrix whose order is represented by its subscript. Note that for the covariance of $\mathbf{g}$, $\mathbf{I}_v$ would be substituted by a $v \times v$ numerator relationship matrix if we had the pedigree information available (PIEPHO et al., 2008). The fitted model (Eq. 1) was the basis for modelling $\mathbf{G}$ and $\mathbf{R}$, creating other models. We opted to fix the permanent environmental effect covariance structure as $\sigma_p \mathbf{I}_{vq}$ based on the recommendations of (SMITH et al., 2007).

Eucalyptus dataset (D2)

The base-line model fitted for the multi-site data was:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}_1\mathbf{s} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{f} + \mathbf{e} \quad (2)$$

where $\mathbf{y}$ is the $n \times 1$ vector of tree DBH in each site; μ is the intercept multiplied by the $n \times 1$ vector of ones ($\mathbf{1}$); $\mathbf{s}$ is the $m \times 1$ vector of the fixed effects of sites, with a $n \times m$ design matrix ($\mathbf{X}_1$); $\mathbf{g}$ is the $tm \times 1$ vector of genotypic random effects within sites [$\mathbf{g} \sim \text{MVN}(\mathbf{0}, \mathbf{G} \otimes \mathbf{I}_t)$], accompanied by its $n \times tm$ incidence matrix $\mathbf{Z}_1$; $\mathbf{f}$ is the $bm \times 1$ vector of random block effects within site [$\mathbf{f} \sim \text{MVN}(\mathbf{0}, \mathbf{B} \otimes \mathbf{I}_t)$], with a $n \times bm$ design matrix ($\mathbf{Z}_2$); and $\mathbf{e}$ is the $n \times 1$ vector of residual random effects [$\mathbf{e} \sim \text{MVN}(0, \mathbf{R} \otimes \mathbf{I}_n)$]. $\mathbf{I}_t$ could be changed by a $t \times t$ numerator relationship matrix if the breeder has pedigree records (PIEPHO et al., 2008). Like in Eq. 1, $\mathbf{G}$ and $\mathbf{R}$ are covariance matrices of the genotypic and residual effects, respectively. Note from Eq. 2 that we considered the block effects as random, so it also has its covariance matrix ($\mathbf{B}$). The decision to consider it as a random effect is based on a previous test using the corrected Bayesian Information Criterion (BIC, SCHWARZ, 1978), proposed by Verbyla (2019), which supports the comparison of models with different fixed effects. In this test, the model that had the block effect as random had a lower value (data not shown). Equation 2 was the basis for modelling the covariance structures of the random effects for D2.

2.3 Modelling random effects

The modelling was performed sequentially, i.e., the covariance matrix of a random effect was modelled after identifying the best structure for a predecessor effect, as done by Faveri et al. (2015). The starting point was the residual covariance matrix. In the base-line models for D1 and D2, $\mathbf{R}$ has the identity covariance structure (IDV) form, which assumes homogeneity of variances and absence of covariances.

In D2, we tested the diagonal covariance structure (DIAG) for $\mathbf{R}$. This structure assumes heterogeneous residual variances between sites and the absence of covariances between site pairs. We tested the same structure for the block effects (Table 2).

In D1, the grid information was available, so the starting point was to test if there was a relevant spatial tendency in the field. We accounted for the spatial dependency using a separable first-order autoregressive (AR1) process in both row and column dimensions (GILMOUR et

Table 2: Covariance structures used for modelling the random effects' covariance matrices (marked with a ● in the last columns) of the models fitted for D1 (multi-harvest data) and D2 (multi-site data)

Name	Variance structure†	Acronym	D1		D2		
			G	R	G	B	R
Identity variance	$\sigma^2_{e/b}\mathbf{I}_{n/b}$	IDV		●		●	●
Compound symmetry	$\left(\sigma^2_g\mathbf{J} + \sigma^2_{gei}\mathbf{I}_{h/m}\right) \otimes \mathbf{I}_{v/t}$	CS	●		●		
Diagonal	$\mathbf{D} \otimes \mathbf{I}_{n/b}$	DIAG		●		●	●
Heterogeneous compound symmetry	$\left\{ \sqrt{\mathbf{D}}[\mathbf{I}_{h/m} + \rho(\mathbf{J} - \mathbf{I}_{h/m})] \sqrt{\mathbf{D}} \otimes \mathbf{I}_{v/t} \right\}$	CORH	●		●		
Homogeneous autoregressive	$\sigma^2_{g/e}\mathbf{I}_h + \mathbf{\Sigma} \otimes \mathbf{I}_{v/n}$	AR	●		●		
Heterogeneous autoregressive	$\sqrt{\mathbf{D}}\mathbf{\Sigma}\sqrt{\mathbf{D}} \otimes \mathbf{I}_{v/n}$	AR1H	●		●		
Unstructured	$\mathbf{\Gamma} \otimes \mathbf{I}_{v/t}$	UN	●		●		
Factor analytic	$(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}) \otimes \mathbf{I}_{v/t}$	FA	●		●		

† σ^2_x is the modelled random effect variance [x = genotype (g), genotype-by-environment interaction (gei), block (b), or residual (e)]; $\mathbf{I}_x$ is an identity matrix [x = number of genotypes in D1 (v) or D2 (t), number of harvest years in D1 (h) or sites in D2 (m), number of blocks in D2 (b), or number of observations in both D1 an D2 (n)]; $\mathbf{J}$ is an $h \times h$ or $m \times m$ all-ones matrix for D1 and D2, respectively; $\mathbf{D}$ is an $h \times h$ or $m \times m$ diagonal matrix whose elements are σ^2_x, for D1 and D2, respectively; ρ is a correlation coefficient; $\mathbf{\Sigma}$ is an $h \times h$ or $m \times m$ matrix of autocorrelations for D1 and D2, respectively; $\mathbf{\Gamma}$ is an $h \times h$ or $m \times m$ full covariance matrix for D1 and D2, respectively; $\mathbf{\Lambda}$ is an $h \times k$ or $m \times k$ matrix of factorial loadings, where k is the number of factors, for D1 and D2, respectively; $\mathbf{\Psi}$ is an $h \times h$ or $m \times m$ diagonal matrix of specific variances, for D1 and D2, respectively; and $\otimes$ is the Kronecker product

al., 1997). Following the procedure described by Smith et al. (2007), a particular harvest year would have the residual covariance matrix as $\mathbf{R}_c = \sigma^2_{e_c}\mathbf{\Sigma}_{Cc} \otimes \mathbf{\Sigma}_{Rc}$, in which $\sigma^2_{e_c}$ is the residual variance of the c-th harvest year, and $\mathbf{\Sigma}_{C_c}$ and $\mathbf{\Sigma}_{R_c}$ are the spatial correlation matrices for the column and row dimensions in c-th harvest year, respectively. Expanding this model to a multi-harvest scenario, $\mathbf{R}$ is composed of a three-way process (FAVERI et al., 2015; SMITH et al., 2007) with the residual covariance matrix between harvest years ($\mathbf{R}_h$) and the two-way autoregressive process represented by the spatial correlation matrices: $\mathbf{R} = \mathbf{R}_h \otimes \mathbf{\Sigma}_C \otimes \mathbf{\Sigma}_R$. In this context, $\mathbf{R}_h$ is the matrix to be tested with different structures. In our

study, we modelled $\mathbf{R}_h$ with the DIAG (previously defined) and the first-order homogeneous and heterogeneous autoregressive structures (AR1 and AR1H, respectively) (Table 2). The idea behind the temporal autoregressive structures for $\mathbf{R}_h$ is the same as the spatial autoregressive: the more distant (in time) the harvests, the lower the correlation between them (WOLFINGER, 1996).

The genotypic effects were modelled using compound symmetry, both homogenous (CS) and heterogeneous (CORH); unstructured (UN); and factor analytic (FAk) models. The CS was the structure present in the baseline model for D1 and D2 and assumes the homogeneity of variances and covariances, in addition to attributing a variance for the genotypic main effect and the genotype-by-environment (harvest year or site) interaction effect. CORH assumes the heterogeneity of variances and covariances, associating them with a correlation coefficient. UN structure assumes heterogeneous variances for each harvest/site and covariance between harvest/site pairs, being the most complete and parameterized model. The FAk structure is a parsimonious alternative to the UN (KELLY et al., 2007; MEYER, 2009) and estimates the heterogeneous covariances using the factors, a set of latent variables that capture the common variance between harvest years/sites (PIEPHO, 1997). Here, we tested first-, second-, and third-order FA models, being the FA1, FA2, and FA3, respectively. Therefore, twelve mixed models were fitted for D1, and eight mixed models were fitted for D2, considering different covariance structures for the models' random effects (Table 2).

The models fitted for each dataset (D1 and D2) had the same fixed effects. Therefore, their goodness-of-fit was tested by the Akaike (AIC, AKAIKE, 1974) and Bayesian information criteria. The models with lower AIC and BIC provide better fitness for the data. In the FAk models, we also calculated the percentage of accumulated variance explained by the k factors ($\overline{\%\mathrm{V}}$, CULLIS et al., 2014):

$$\overline{\%\mathrm{V}} = \frac{\mathrm{Trace}\left(\mathbf{\Lambda}^{*}\mathbf{\Lambda}^{*\prime}\right)}{\mathrm{Trace}\left(\mathbf{\Lambda}^{*}\mathbf{\Lambda}^{*\prime}+\mathbf{\Psi}\right)}100 \tag{3}$$

where $\mathbf{\Lambda}^{*}$ is the $h \times k$ or $m \times k$ matrix of rotated loadings, for D1 and D2, respectively. One must rotate the matrix of loadings if $k > 2$. More details about the rotation process in FAk models can be found at (CULLIS et al., 2010). Trace is the summation of diagonal elements of a matrix. The best model was selected considering AIC, BIC, and $\overline{\%\mathrm{V}}$ values (if the model had a FAk structure). After the selection of the most adequate models for D1 and D2, the significance of the random effects was tested by the likelihood ratio test (LRT) considering the chi-square

statistic with one degree of freedom and 0.95 confidence level.

2.4 Genetic and non-genetic parameters

For both datasets (D1 and D2), we estimated the individual phenotypic variances (sum of the variance components estimated by the model), generalized heritability (H_g^2, CULLIS et al., 2006) and accuracy (r, MRODE, 2014, chapter 3) from the chosen model. H_g^2 and r were estimated using the following equations:

$$H_g^2 = 1 - \frac{\overline{\Delta}_{\text{BLUP}}}{2\sigma_g^2} \tag{4}$$

$$\text{r} = \sqrt{1 - \left(\frac{PEV}{\sigma_g^2}\right)} \tag{5}$$

where $\overline{\Delta}_{\text{BLUP}}$ is the average pairwise prediction error variance of the genotypes, σ_g^2 is the genotypic variance, and PEV is the prediction error variance per se, obtained from the diagonal of the coefficients matrix' generalized inverse from the mixed models equation.

We also estimated the genetic correlation between pairs of harvests/sites (ϱ). in the FAk model, ϱ is estimated using the factor loadings ($\lambda_{\kappa_{c/j}}$, CULLIS et al., 2010):

$$\varrho = \frac{\sum_{\kappa=1}^{k} \lambda_{\kappa_{c/j}} \lambda_{\kappa_{c'/j'}}}{\sqrt{\sigma^2_{g_{c/j}} \sigma^2_{g_{c'/j'}}}} \tag{6}$$

where $\lambda^*_{\kappa_{c/j}}$ is the rotated factor loading of the κ-th factor in the c-th, or j-th harvest/site. In the conventional models, we used the covariances themselves for estimating ϱ:

$$\varrho = \frac{\sigma_{g_{cc'/jj'}}}{\sqrt{\sigma^2_{g_{c/j}} \sigma^2_{g_{c'/j'}}}} \tag{7}$$

where $\sigma_{g_{cc'/jj'}}$ is the genotypic covariance between pairs of harvests/sites.

The expected selection gains (SG%) were calculated considering the best 7 (28%) and 30 (14%) genotypes for D1 and D2, respectively:

$$\text{SG\%} = \left(\frac{\overline{g_s} - \overline{g_t}}{\overline{x}}\right) * 100 \tag{8}$$

where $\overline{g_s}$ is the mean BLUP of the selected genotypes, and $\overline{g_t}$ is the mean BLUP of the evaluated

population and $\overline{x}$ is the phenotypic mean.

The correlations between ranks (Spearman correlation, $r_{MM'}$) were calculated to compare the ranking of the best-fitting model and the basic model (without modelling the covariance structures).

All the analyses were performed in the `R` environment (R CORE TEAM, 2023, version 4.2.1), using the `ASReml-R` package (BUTLER et al., 2018, version 4.1). The figures were made using the `ggplot2` (WICKHAM, 2016) and `ComplexHeatmap` (GU et al., 2016) packages. The data sets and the scripts are available at `https://github.com/saulo-chaves/MHT_MET_MM`.

3 Results

3.1 Model selection

The information criteria differed in the definition of the best-fitting model for both datasets (Tables 3 and 4). In D1, AIC indicated the twelfth model (MHT12), while BIC indicated the seventh model (MHT7) (Table 3). MHT7 has the highest accuracy (0.92), compared to the base model MHT1 (0.82) and the AIC-chosen model MHT12 (0.89). On the other hand, MHT12 provides a higher expected genetic gain (16.89%) than MHT1 and MHT7 (13.37 and 13.46%, respectively). Since the FA3 structure in MHT12's **G** enables a better insight into the genetic covariances over harvest years (99% of accumulated explained variance) than the AR1 structure in MHT7's **G**, we chose MHT12 for estimating the variance components and predicting genotypic values.

In D2, the fourth (MST4) and fifth (MST5) models showed the best fit according to BIC and AIC, respectively (Table 4). In this case, both models had the same value for accuracy. MST5 provided a somewhat higher expected genetic gain (11.37%) than MST4 (11.36%). Furthermore, the UN structure in MET5's **G** is the most complete in terms of information on genetic covariances between sites, although it makes the model less parsimonious and may overfit as indicated by the BIC. Since D2 has few sites, using the UN structure will not be a problem, so we chose MST5 to estimate the genetic parameters and predict the BLUPs.

Table 3: Values of Akaike (AIC) and Bayesian (BIC) information criteria, mean accuracy ($\overline{r}$), and expected genetic gains (SG%) of the twelve models [multi-harvest trial (MHT)] fitted for the cupuassu tree dataset (D1), and percentage of accumulated variance ($\overline{\%V}$) explained by factor analytic models

Model	Covariance matrix[†]		AIC	BIC	$\overline{r}$	SG%	$\overline{\%V}$
	G	R					
MHT1	CS	IDV	7532.04	7553.13	0.82	13.37	-
MHT2	CS	IDV ⊗ AR1 ⊗ AR1	7514.29	7545.93	0.82	13.60	-
MHT3	CS	DIAG ⊗ AR1 ⊗ AR1	7255.07	7460.70	0.91	13.42	-
MHT4	CS	AR1 ⊗ AR1 ⊗ AR1	7502.35	7539.26	0.83	13.61	-
MHT5	CS	AR1H ⊗ AR1 ⊗ AR1	7196.43	7291.33	0.88	11.50	-
MHT6	CSH	AR1H ⊗ AR1 ⊗ AR1	7196.96	7349.86	0.78	15.32	-
MHT7	AR1	AR1H ⊗ AR1 ⊗ AR1	7182.60	**7277.50**	0.92	13.46	-
MHT8	AR1H	AR1H ⊗ AR1 ⊗ AR1	7187.72	7340.62	0.83	14.79	-
MHT9[‡]	US	AR1H ⊗ AR1 ⊗ AR1	-	-	-	-	-
MHT10	FA1	AR1H ⊗ AR1 ⊗ AR1	7187.01	7376.82	0.83	15.21	67.36
MHT11	FA2	AR1H ⊗ AR1 ⊗ AR1	7176.79	7445.68	0.87	16.53	95.85
MHT12	FA3	AR1H ⊗ AR1 ⊗ AR1	**7156.17**	7419.79	0.89	16.89	99.02

Data in bold are the best-fitted models according to AIC and BIC

[†] See Table 2

[‡] Did not converge

Table 4: Values of Akaike (AIC) and Bayesian (BIC) information criteria, mean accuracy ($\overline{r}$), and expected genetic gains (SG%) of the eight models [multi-site trials (MST)] fitted for the eucalyptus dataset (D2), and percentage of accumulated variance ($\overline{\%V}$) explained by the factor analytic models

Model	Covariance matrix[†]			AIC	BIC	$\overline{r}$	SG%	$\overline{\%V}$
	G	B	R					
MST1	CS	IDV	IDV	57063.29	57095.34	0.98	11.37	-
MST2	CS	IDV	DIAG	56015.16	56071.24	0.98	11.17	-
MST3	CS	DIAG	DIAG	56009.31	56089.43	0.95	11.18	-
MST4	CORH	DIAG	DIAG	55939.21	**56043.37**	0.95	11.29	-
MST5	UN	DIAG	DIAG	**55905.33**	56049.55	0.95	11.36	-
MST6	FA1	DIAG	DIAG	55921.11	56049.29	0.95	11.37	55.15%
MST7	FA2	DIAG	DIAG	55907.33	56059.56	0.95	11.36	67.38%
MST8	FA3	DIAG	DIAG	55911.33	56079.58	0.98	11.36	67.97%

Data in bold are the best-fitted models according to AIC and BIC

[†] See Table 2

3.2 Variance components and genetic parameters

According to the LRT, at a 5% significance level, all random effects were significant in D1 and D2 (data not shown). In both datasets, the oscillations of the variance components

estimated from the selected models over the harvest years/sites were evident (MHT12 and MST5, Tables 5 and 6, respectively). This is because MHT12 and MST5 accounted for the covariance heterogeneity in the data sets. It is important to highlight that since we have chosen an FA3 structure to model the genotypic effects in D1, the loadings had to be rotated before estimating the genotypic variances, according to the process described by (CULLIS et al., 2010). This is necessary because when $k > 1$, the constraint imposed in the matrix of loadings to keep its identifiability hinders the biological meaning of the loadings.

Table 5: Variance component estimates/genetic parameters (E/P) estimated by the twelfth model (MHT12) for the cupuassu tree dataset (D1)

E/P	H01	H02	H03	H04	H05	H06	H07	H08	H09	H10	H11	H12
σ^2_g	2.7	5.9	8.17	25.7	17.54	18.59	26.94	15.46	9.41	14.86	14.26	7.92
σ^2_p	1.97											
σ^2_e	3.91	9.31	27.62	72.3	51.38	41.18	67.89	58.22	100.36	94.76	76.01	42.19
ρ_h	0.22											
ρ_c	0.04											
ρ_r	0.15											
σ^2_f	8.58	17.18	37.76	100	70.89	61.74	96.8	75.65	111.74	11.59	92.24	52.08
μ	3.57	6.97	10.62	27.4	22.77	17.57	24.08	21.13	17.28	28.92	18.49	17.06
H^2_g	0.77	0.84	0.86	0.83	0.86	0.86	0.82	0.83	0.83	0.81	0.86	0.74
r	0.87	0.91	0.91	0.9	0.91	0.92	0.89	0.89	0.85	0.86	0.9	0.83

H represents a harvest year, σ^2_g is the genotypic variance, σ^2_p is the permanent environmental effects variance, σ^2_e is the residual variance, ρ_h is the temporal autocorrelation coefficient, ρ_c and ρ_r are the spatial autocorrelation coefficients at column and row directions, σ^2_f is the phenotypic variance, μ is the phenotypic mean, H^2_g is the generalized heritability and r is the accuracy.,

Table 6: Variance component estimates/genetic parameters (E/P) estimated by the fifth model (MST5) for the eucalyptus dataset (D2)

E/P	S1	S2	S3	S4
σ^2_g	1.37	2.2	0.91	2.55
σ^2_r	0.13	0.1	0.22	0.04
σ^2_e	4.9	3.44	2.88	6.51
σ^2_f	6.4	5.74	4.01	9.1
μ	14.11	13.21	12.64	14.55
H^2_g	0.79	0.9	0.81	0.82
r	0.94	0.97	0.95	0.95

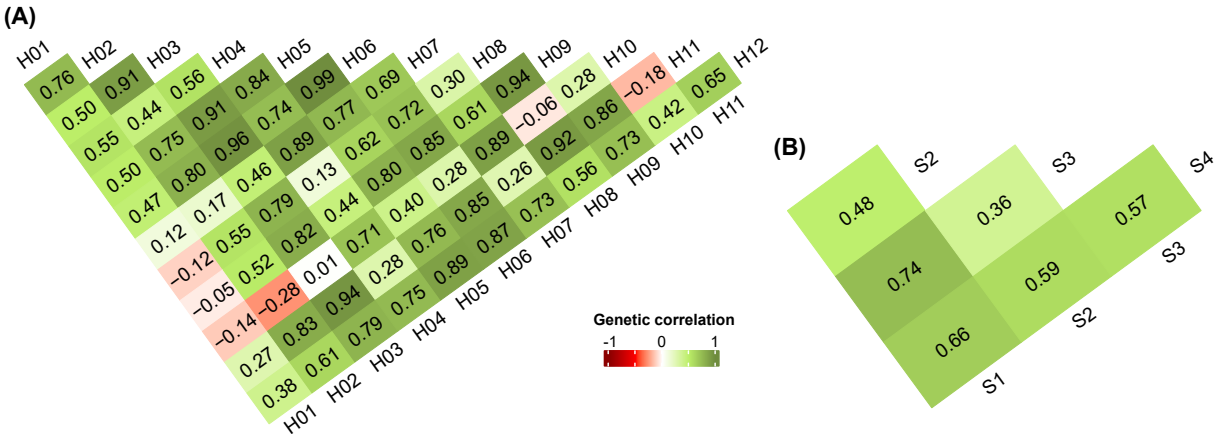
S represents a site, σ^2_g is the genotypic variance, σ^2_r is the variance of block effects, σ^2_e is the residual variance, σ^2_f is the phenotypic variance, μ is the phenotypic mean, H^2_g is the generalized heritability and r is the accuracy.,

In D1, the generalized heritability ranged from 0.74 to 0.86 (mean of 0.83), and the accuracy ranged from 0.83 to 0.91 (mean of 0.89). The relationship between harvest years is evidenced by the temporal autocorrelation coefficient ($\rho_h = 0.22$). In addition, MHT12 provides a general

spatial adjustment, represented by the spatial autocorrelation coefficient at both row at column dimensions ($\rho_r = 0.15$ and $\rho_c = 0.04$, respectively). In D2, the minimum and maximum generalized heritability were 0.79 and 0.90, respectively (mean of 0.82). The accuracy had a smaller variation, with the difference between minimum (0.94) and maximum (0.97) values of 0.3 and an average of 0.95.

In D1, there was a high variability in the magnitude of the genetic correlations between harvest years (ranging from 0.28 to 0.99). In other words, there is a differential allelic expression throughout the harvests (Fig. 1A). The genetic correlations between sites in D2 were mostly moderate (ranging from 0.36 to 0.74), proving that the site effect is relevant for the genotypic behaviour (Fig. 1B).

Figure 1: Genetic correlations between pairs of harvest years (H01 to H12) in D1 (**A**) and sites (S1 to S4) in D2 (**B**)

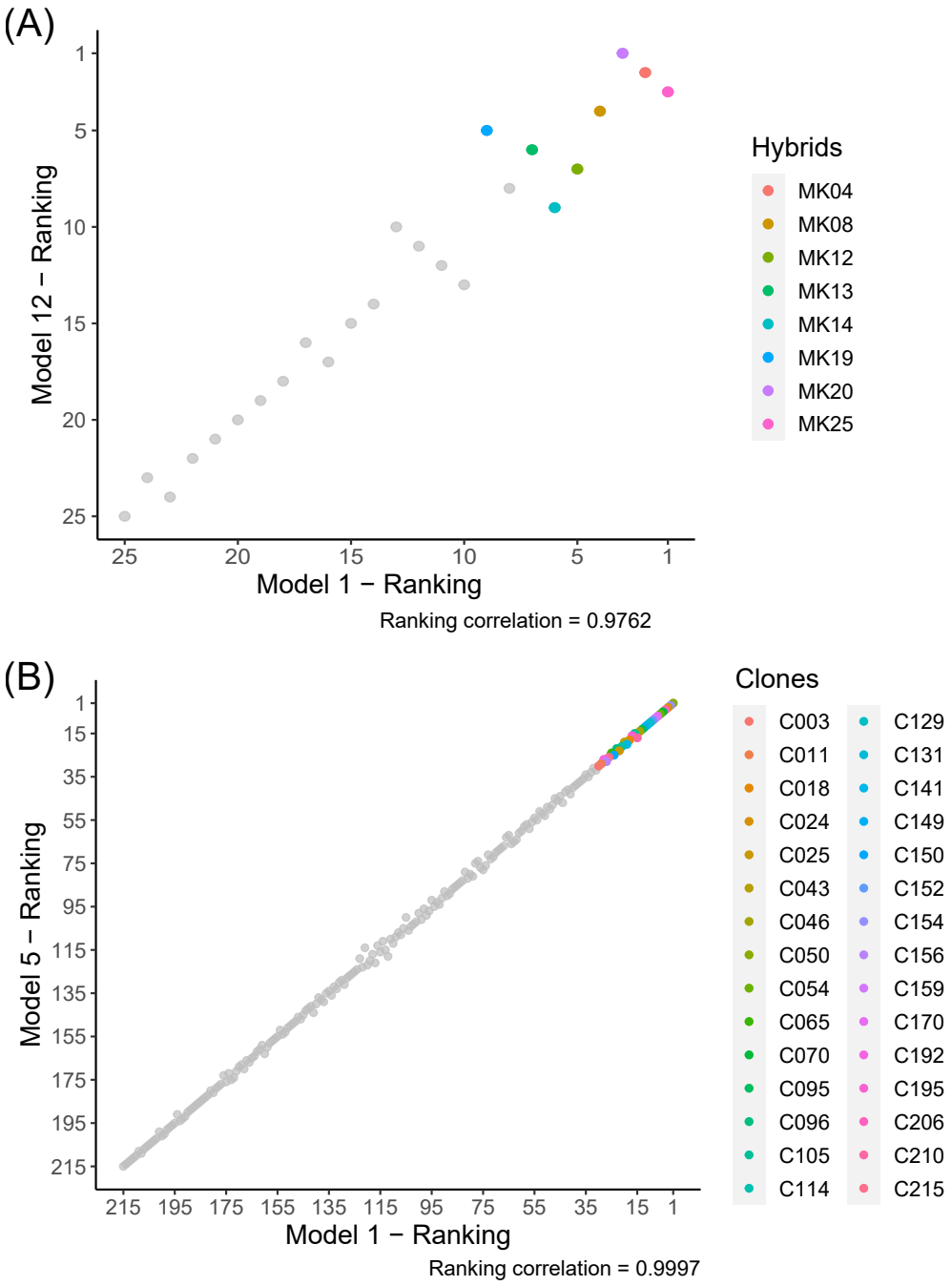


3.3 Genotypic values and genetic gains

In D1, MK04, MK08, MK12, MK13, MK19, MK20, and MK25 were the hybrids with the highest genotypic value, according to MHT12. If one used the ranking provided by MHT1, there would be only one change: MK14 substituting MK19 (Fig. 2A). Note that this difference represents a 3.52-point increase in the expected selection gain percentage, i.e., from 13.37% in MHT1 up to 16.89% in MHT12.

In D2, there were no differences between the top 30 clones targeted in the selection (Fig. 2B), meaning that the real gain is similar between the two models, although the expected gain is somewhat smaller in MST5. Joint analysis of the rank correlations and Fig. 2 showed the subtle differences between the genotype ranks from the genotypic values obtained by the basic

Figure 2: Comparison of the genotypes' ranks based on the predicted genotypic values obtained by the basic (MHT1 and MST1) and best fit (MHT12 and MST5) models and correlation between them in D1 (A) and D2 (B). The selected genotypes are highlighted



(MHT1 and MST1) and selected (MHT10 and MST5) models in both datasets.

4 Discussion

4.1 Model selection

Trials evaluated over time and/or space subject treatments to different environmental conditions (FERRÃO et al., 2017; PASTINA et al., 2012). These trials are fundamental for a more accurate selection, enabling the incorporation and recommendation of genotypes resilient to environmental adversities and responsive to environmental improvement (ELIAS et al., 2016). This fact has greater importance in perennial plants, which remain in the field for long periods, exposed to the most diverse environmental variations related to factors outside the control of the breeder, such as climate change (ATLIN et al., 2017; ELLI et al., 2020; LAHIVE et al., 2019). However, such experimental strategies lose efficiency if they are not accompanied by adequate statistical models, capable of highlighting the peculiarities of each harvest/site (COBB et al., 2019; STRINGER et al., 2017).

In this study, the AIC and BIC criteria indicated different models as the best fit. The criteria for the selection of models have different principles and therefore do not always lead to the selection of the same model (DRTON; PLUMMER, 2017). AIC ranks models based on an efficiency criterion (CAVANAUGH; NEATH, 2019), and BIC is known to be a consistency criterion (NEATH; CAVANAUGH, 2012). From a practical point of view, AIC tends to select more parameterized models, while BIC prioritizes those that are less parameterized (ISIK et al., 2017b). This pattern is proved by our results. Here, we prioritized the most informative models, i.e., the ones that could provide a complete analysis of the population dynamics over the harvest years/sites. Under these criteria, we selected the models appointed by the AIC in D1 and D2. In the context of FAk models, AIC and BIC may not be the best option. For example, note that BIC did not appoint any FAk model as the best fitting in either data set. The same can happen with AIC, as observed by previous studies (ISIK et al., 2017b; ZHANG et al., 2020). Then, a third criterion can be used as an auxiliary: the percentage of variance explained by each factor (CULLIS et al., 2014). MHT12, the chosen model for D1, explained 99% of the total variance of the data set.

The two datasets illustrate two situations usually encountered by perennial plant breeders. In drawing a parallel between the chosen models, it is evident that in simpler situations, such as D2, where only four sites were evaluated, the UN structure is indeed the most complete and the one that best explains the data (MONTEVERDE et al., 2018; MRODE, 2014). On the

other hand, when the data is more complex, such as D1, in which twelve consecutive harvest years were evaluated, employing UN may not result in the best fit, as proven in the present study. In these cases, factor analytic (FAk) structures can be advantageous (DIAS et al., 2018; SHALIZI; ISIK, 2019; SMITH et al., 2015). The principle of FAk models is the reduction of data dimensionality taking into account the covariance between harvest year/sites for the formation of factors, a set of latent variables (PIEPHO, 1997; PIEPHO, 1998; SMITH et al., 2001). The use of factor loadings and scores can simplify the results for breeders to make selections with high performance and stability (SMITH; CULLIS, 2018; SMITH et al., 2021).

4.2 Variance components and genetic parameters

Variations between harvest years or sites are due to genetic (differential expression of genes) and environmental (management, edaphoclimatic conditions, among others) factors (HALLDORSDOTTIR; BINDER, 2017; MELO et al., 2020; SNOWDON et al., 2021). Such variations are heterogeneous, as reflected by the estimates of variance components and genetic parameters. When modelling the covariance structures, such effects are better explained, and the residual variance is reduced. This allows the increasing of heritability and accuracy, parameters indicative of success with selection (MRODE, 2014; SCHMIDT et al., 2019). As in the present study, several authors verified the superiority of the UN model compared to the basic model (compound symmetry model) in MET analyses (SHALIZI; ISIK, 2019). This is also valid for the FAk models, either in multi-site or multi-harvest contexts (GEZAN et al., 2017; VERBYLA et al., 2021). In summary, assuming homogeneity of covariances (genetic, residual) between sites and harvest years is to simplify a relationship that often has nuances that only the modelling of covariance structures can detect.

In D1, we also accounted for the spatial tendencies in the field. The presence of spatial dependency was attested by the lowering of the AIC and BIC values, proving that the model is better fitted when we consider the spatial effect. These tendencies can be natural, caused by soil gradients, topography, or microclimate; or man-made, caused by unequal irrigation and fertilization, machine passaging, or any deed related to the trial conduction (GILMOUR et al., 1997). Here, we join the spatial autocorrelation with the temporal autocorrelation (WOLFINGER, 1996) in a three-way process as used by Smith et al. (2007) and Faveri et al. (2015). In the last instance, the goal is to have better control of the effects that can influence the genotypes' performance, diminishing the residual effects and elevating the importance of

the genotypic effect on the phenotypic expression. Our study's results attest to this fact. Thus, we reiterate the recommendations of Gilmour et al. (1997) and Velazco et al. (2017) to test for the presence of spatial tendencies in the field, whenever possible.

The relative complexity of the datasets used in this study can be illustrated by the genetic correlations between harvests (D1) and sites (D2). In D1 the genetic correlations showed that harvest years can be greatly associated or do not have any association whatsoever (Fig. 1A). At first glance, this pattern can be counterintuitive since data is collected in the same individuals over the harvest years. Nevertheless, there are two main causes for this variation: (i) although the correlation is genetic, there is a differential gene expression when the individual faces different climate conditions, which characterizes the genotype-by-harvest interaction (REDPATH et al., 2021), and (ii) in perennial fruit trees such as cupuassu tree (D1), the maturity stage of plants is also a source of variation (CHAVES et al., 2021). For example, the genes that were expressed in the first three harvest years are not necessarily the same as those expressed in the fourth harvest year onwards (note the difference between these periods in Table 4).

Conversely, the main cause for the moderate genetic correlation between sites in D2 is the genotype-by-site interaction itself. In other words, the different conditions of climate and soil between sites cause a differential gene expression in the selection candidates (MALOSETTI et al., 2013). Nevertheless, the differences in the environmental characteristics are not remarkable enough to provoke a major decrease in the correlations, probably due to the geographical proximity. For example, Silva et al. (2019) also evaluated four sites, but with a large distance between them, and observed correlations far lower than the ones we have found in this study. Note that in both data sets, we are accounting for both additive and non-additive effects. Previous studies showed that the non-additive effects are more influenced by the environment than the additive effects (BERLIN et al., 2019; HUNT et al., 2020). In other words, the additive genetic covariance between harvest years/sites would be greater than the total genetic variance estimated in our study if we had the pedigree information.

4.3 Genotypic values and genetic gains

The main objective of the modelling of covariance structures is to increase the reliability of genetic selection. In D1, the selection gains obtained by the basic model were lower (13.37%) when compared to those obtained by MHT12 (16.89%). In D2, there was a complete coincidence in genotype ranking, which will cause the real genetic gain to be the same, whether

the breeder selects the base model or the best-fitting model. The contrast between D1 and D2 is related to the complexity of each data set. D2 is composed of only four sites, with a moderate genotype-by-site interaction (Fig. 1B), and with many clones in evaluation (215). Note from Fig. 2B that there are changes in the rankings between MST1 and MST5, but none of these changes occurred in the top 30 genotypes. In this situation, breeders could use both MST1 and MST5 to base the selection. Nevertheless, as we stated previously, our goal was to choose a model that could provide tools for studying the covariance dynamics between sites. Conversely, D1 is a more complex dataset, containing twelve harvest years with high genotype-by-harvest interaction effects (Fig. 1A), but with few hybrids in evaluation. Even though the ranking correlation was also high for D1, there were changes all along the ranking when comparing MHT1 to MHT12 (Fig. 1A). In this case, neglecting the modelling of covariance structures could lead to erroneous selection (SOUZA et al., 2021).

In summary, this study showed that modelling the covariance structures and identifying the model with the best fit is fundamental for genetic evaluation in the breeding of perennial plants, whether orchard or timber species. For D1, the model with the best fit generated more reliable results, providing an increase in genetic gains of 3.52 points. For D2, modelling the covariance structures had more discrete advantages, since the accuracy and the genetic gains were alike. In a similar situation, we still advocate for the modelling since (i) one cannot conclude a priori if modelling the covariance structures of the random effects will have major or little benefits, and (ii) even though the results were similar, the base-line model, MST1, was not the most statistically appropriate model according to the information criteria.

References

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

ALVES, R. S.; RESENDE, M. D. V.; AZEVEDO, C. F.; SILVA, F. F.; ROCHA, J. R. A. S. C.; NUNES, A. C. P.; CARNEIRO, A. P. S.; SANTOS, G. A. Optimization of *Eucalyptus* breeding through random regression models allowing for reaction norms in response to environmental gradients. **Tree Genetics & Genomes**, v. 16, n. 2, p. 38, 2020.

ARMSTRONG, R. A. Recommendations for analysis of repeated-measures designs: testing and correcting for sphericity and use of manova and mixed model analysis. **Ophthalmic and Physiological Optics**, v. 37, n. 5, p. 585–593, 2017.

ATLIN, G. N.; CAIRNS, J. E.; DAS, B. Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. **Global Food Security**, v. 12, p. 31–37, 2017.

BAKER, B. Can modern agriculture be sustainable?: Perennial polyculture holds promise. **BioScience**, v. 67, n. 4, p. 325–331, 2017.

BERLIN, M.; JANSSON, G.; HÖGBERG, K.-A.; HELMERSSON, A. Analysis of non-additive genetic effects in Norway spruce. **Tree Genetics & Genomes**, v. 15, n. 3, p. 42, 2019.

BUTLER, D. G.; CULLIS, B. R.; GILMOUR, A. R.; GOGEL, B. J.; THOMPSON, R. **ASReml-R reference manual Version 4**. Hemel Hempstead, UK: VSN International, 2018.

CAVANAUGH, J. E.; NEATH, A. A. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. **WIREs Computational Statistics**, v. 11, n. 3, e1460, 2019.

CHAVES, S. F. S.; ALVES, R. M.; ALVES, R. S.; SEBBENN, A. M.; RESENDE, M. D. V.; DIAS, L. A. S. *Theobroma grandiflorum* breeding optimization based on repeatability, stability and adaptability information. **Euphytica**, v. 217, n. 12, p. 211, 2021.

CHAVES, S. F. S.; GAMA, M. A. P.; ALVES, R. M.; OLIVEIRA, R. P.; PEDROZA NETO, J. L.; LIMA, V. M. N. Evaluation of physicochemical attributes of a yellow latosol under agroforestry system as compared to secondary forest in the Eastern Amazon. **Agroforestry Systems**, v. 94, n. 5, p. 1903–1912, 2020.

COBB, J. N.; JUMA, R. U.; BISWAS, P. S.; ARBELAEZ, J. D.; RUTKOSKI, J.; ATLIN, G.; HAGEN, T.; QUINN, M.; NG, E. H. Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. **Theoretical and Applied Genetics**, v. 132, n. 3, p. 627–645, 2019.

CREWS, T. E.; BLESH, J.; CULMAN, S. W.; HAYES, R. C.; JENSEN, E. S.; MACK, M. C.; PEOPLES, M. B.; SCHIPANSKI, M. E. Going where no grains have gone before: From early to mid-succession. **Agriculture, Ecosystems & Environment**, v. 223, p. 223–238, 2016.

CREWS, T. E.; CATTANI, D. J. Strategies, advances, and challenges in breeding perennial grain crops. **Sustainability**, v. 10, n. 7, p. 2192, 2018.

CROWDER, M. J.; HAND, D. J. **Analysis of repeated measures**. New York: Routledge, 2020.

CULLIS, B. R.; JEFFERSON, P.; THOMPSON, R.; SMITH, A. B. Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a *Pinus radiata* breeding programme. **Theoretical and Applied Genetics**, v. 127, n. 10, p. 2193–2210, 2014.

CULLIS, B. R.; SMITH, A. B.; BEECK, C. P.; COWLING, W. A. Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. **Genome**, v. 53, n. 11, p. 1002–1016, 2010.

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, n. 4, p. 381–393, 2006.

DIAS, K. O. G.; GEZAN, S. A.; GUIMARÃES, C. T.; PARENTONI, S. N.; GUIMARÃES, P. E. O.; CARNEIRO, N. P.; PORTUGAL, A. F.; BASTOS, E. A.; CARDOSO, M. J.; ANONI, C. O.; MAGALHÃES, J. V.; SOUZA, J. C.; GUIMARÃES, L. J. M.; PASTINA, M. M. Estimating genotype ×environment interaction for and genetic correlations among drought tolerance traits in maize via factor analytic multiplicative mixed models. **Crop Science**, v. 58, n. 1, p. 72–83, 2018.

DRTON, M.; PLUMMER, M. A Bayesian information criterion for singular models. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 79, n. 2, p. 323–380, 2017.

EEUWIJK, F. A. van; BUSTOS-KORTS, D.; MILLET, E. J.; BOER, M. P.; KRUIJER, W.; THOMPSON, A.; MALOSETTI, M.; IWATA, H.; QUIROZ, R.; KUPPE, C.; MULLER, O.; BLAZAKIS, K. N.; YU, K.; TARDIEU, F.; CHAPMAN, S. C. Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. **Plant Science**, v. 282, p. 23–39, 2019.

ELIAS, A. A.; ROBBINS, K. R.; DOERGE, R.; TUINSTRA, M. R. Half a century of studying genotype × environment interactions in plant breeding experiments. **Crop Science**, v. 56, n. 5, p. 2090–2105, 2016.

ELLI, E. F.; SENTELHAS, P. C.; BENDER, F. D. Impacts and uncertainties of climate change projections on *Eucalyptus* plantations productivity across Brazil. **Forest Ecology and Management**, v. 474, p. 118365, 2020.

FAVERI, J.; VERBYLA, A. P.; PITCHFORD, W. S.; VENKATANAGAPPA, S.; CULLIS, B. R. Statistical methods for analysis of multi-harvest data from perennial pasture variety selection trials. **Crop and Pasture Science**, v. 66, n. 9, p. 947, 2015.

FERRÃO, L. F. V.; FERRÃO, R. G.; FERRÃO, M. A. G.; FRANCISCO, A.; GARCIA, A. A. F. A mixed model to multiple harvest-location trials applied to genomic prediction in *Coffea canephora*. **Tree Genetics & Genomes**, v. 13, n. 5, p. 95, 2017.

GEZAN, S. A.; CARVALHO, M. P.; SHERRILL, J. Statistical methods to explore genotype-by-environment interaction for loblolly pine clonal trials. **Tree Genetics & Genomes**, v. 13, n. 1, p. 1, 2017.

GILMOUR, A. R.; CULLIS, B. R.; VERBYLA, A. P. Accounting for natural and extraneous variation in the analysis of field experiments. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 2, n. 3, p. 269–293, 1997.

GU, Z.; EILS, R.; SCHLESNER, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. **Bioinformatics (Oxford, England)**, v. 32, n. 18, p. 2847–2849, 2016.

HALLDORSDOTTIR, T.; BINDER, E. B. Gene × Environment Interactions: from molecular mechanisms to behavior. **Annual Review of Psychology**, v. 68, n. 1, p. 215–241, 2017.

HAVERKAMP, N.; BEAUDUCEL, A. Violation of the sphericity assumption and its effect on type-I error rates in repeated measures ANOVA and Multi-Level Linear Models (MLM). **Frontiers in Psychology**, v. 8, p. 1841, 2017.

HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a selection model. **Biometrics**, v. 31, n. 2, p. 423, 1975.

HU, X.; YAN, S.; SHEN, K. Heterogeneity of error variance and its influence on genotype comparison in multi-location trials. **Field Crops Research**, v. 149, p. 322–328, 2013.

HUNT, C. H.; HAYES, B. J.; EEUWIJK, F. A. van; MACE, E. S.; JORDAN, D. R. Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships. **Theoretical and Applied Genetics**, v. 133, n. 3, p. 1009–1018, 2020.

ISABEL, N.; HOLLIDAY, J. A.; AITKEN, S. N. Forest genomics: Advancing climate adaptation, forest health, productivity, and conservation. **Evolutionary Applications**, v. 13, n. 1, p. 3–10, 2020.

ISIK, F.; HOLLAND, J.; MALTECCA, C. **Genetic Data Analysis for Plant and Animal Breeding**. Cham: Springer International Publishing, 2017.

ISIK, F.; HOLLAND, J.; MALTECCA, C. Multi Environmental Trials. In: ISIK, F.; HOLLAND, J.; MALTECCA, C. (Eds.). **Genetic Data Analysis for Plant and Animal Breeding**. Cham: Springer International Publishing, 2017. P. 227–262.

KELLY, A. M.; SMITH, A. B.; ECCLESTON, J. A.; CULLIS, B. R. The accuracy of varietal selection using Factor Analytic Models for Multi-Environment plant breeding trials. **Crop Science**, v. 47, n. 3, p. 1063–1070, 2007.

KRAUSE, M. D.; DIAS, K. O. G.; SANTOS, J.; OLIVEIRA, A. A.; GUIMARÃES, L. J. M.; PASTINA, M. M.; MARGARIDO, G. R. A.; GARCIA, A. A. F. Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models. **Crop Science**, v. 60, n. 6, p. 3049–3065, 2020.

LAHIVE, F.; HADLEY, P.; DAYMOND, A. J. The physiological responses of cacao to the environment and the implications for climate change resilience. A review. **Agronomy for Sustainable Development**, v. 39, n. 1, p. 5, 2019.

MALOSETTI, M.; RIBAUT, J.-M.; EEUWIJK, F. A. van. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. **Frontiers in Physiology**, v. 4, p. 44, 2013.

MELO, V. L.; MARÇAL, T. S.; ROCHA, J. R. A. S. C.; ANJOS, R. S. R.; CARNEIRO, P. C. S.; CARNEIRO, J. E. S. Modeling (co)variance structures for genetic and non-genetic effects in the selection of common bean progenies. **Euphytica**, v. 216, n. 5, p. 77, 2020.

MEYER, K. Factor-analytic models for genotype × environment type problems and structured covariance matrices. **Genetics Selection Evolution**, v. 41, n. 1, p. 21, 2009.

MONTEVERDE, E.; ROSAS, J. E.; BLANCO, P.; PÉREZ DE VIDA, F.; BONNECARRÈRE, V.; QUERO, G.; GUTIERREZ, L.; MCCOUCH, S. Multienvironment models increase prediction accuracy of complex traits in advanced breeding lines of rice. **Crop Science**, v. 58, n. 4, p. 1519–1530, 2018.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. 3rd ed. Boston, MA: CABI, 2014.

NEATH, A. A.; CAVANAUGH, J. E. The Bayesian information criterion: background, derivation, and applications. **WIREs Computational Statistics**, v. 4, n. 2, p. 199–203, 2012.

PASTINA, M. M.; MALOSETTI, M.; GAZAFFI, R.; MOLLINARI, M.; MARGARIDO, G. R. A.; OLIVEIRA, K. M.; PINTO, L. R.; SOUZA, A. P.; EEUWIJK, F. A. van; GARCIA, A. A. F. A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. **Theoretical and Applied Genetics**, v. 124, n. 5, p. 835–849, 2012.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545–554, 1971.

PIEPHO, H.-P. Analyzing genotype-environment data by mixed models with multiplicative terms. **Biometrics**, v. 53, n. 2, p. 761–766, 1997.

PIEPHO, H.-P. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures: **Theoretical and Applied Genetics**, v. 97, n. 1-2, p. 195–201, 1998.

PIEPHO, H.-P.; ECKL, T. Analysis of series of variety trials with perennial crops. **Grass and Forage Science**, v. 69, n. 3, p. 431–440, 2014.

PIEPHO, H.-P.; MÖHRING, J.; MELCHINGER, A. E.; BÜCHSE, A. BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, v. 161, n. 1-2, p. 209–228, 2008.

R CORE TEAM. **R: A Language and environment for statistical computing**. Viena, Áustria: R Foundation for Statistical Computing, 2023.

RAI, M. K.; SHEKHAWAT, N. S. Recent advances in genetic engineering for improvement of fruit crops. **Plant Cell, Tissue and Organ Culture (PCTOC)**, v. 116, n. 1, p. 1–15, 2014.

REDPATH, L. E.; GUMPERTZ, M.; BALLINGTON, J. R.; BASSIL, N.; ASHRAFI, H. Genotype, environment, year, and harvest effects on fruit quality traits of five blueberry (*Vaccinium corymbosum* L.) cultivars. **Agronomy**, v. 11, n. 9, p. 1788, 2021.

SCHMIDT, P.; HARTUNG, J.; BENNEWITZ, J.; PIEPHO, H.-P. Heritability in plant breeding on a genotype-difference basis. **Genetics**, v. 212, n. 4, p. 991–1008, 2019.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, v. 6, n. 2, p. 461–464, 1978.

SHALIZI, M. N.; ISIK, F. Genetic parameter estimates and GxE interaction in a large cloned population of *Pinus taeda* L. **Tree Genetics & Genomes**, v. 15, n. 3, p. 46, 2019.

SILVA, P. H. M.; MARCO, M.; ALVARES, C. A.; LEE, D.; MORAES, M. L. T.; PAULA, R. C. d. Selection of *Eucalyptus grandis* families across contrasting environmental conditions. **Crop Breeding and Applied Biotechnology**, v. 19, p. 47–54, 2019.

SINGH, M.; TADESSE, W.; SARKER, A.; MAALOUF, F.; IMTIAZ, M.; CAPETTINI, F.; NACHIT, M. Capturing the heterogeneity of the error variances of a group of genotypes in crop cultivar trials. **Crop Science**, v. 53, n. 3, p. 811–818, 2013.

SMITH, A. B.; CULLIS, B. R. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. **Euphytica**, v. 214, n. 8, p. 143, 2018.

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

SMITH, A. B.; GANESALINGAM, A.; KUCHEL, H.; CULLIS, B. R. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. **Theoretical and Applied Genetics**, v. 128, n. 1, p. 55–72, 2015.

SMITH, A. B.; NORMAN, A.; KUCHEL, H.; CULLIS, B. R. Plant variety selection using interaction classes derived from factor analytic linear mixed models: models with independent variety effects. **Frontiers in Plant Science**, v. 12, p. 1857, 2021.

SMITH, A. B.; STRINGER, J. K.; WEI, X.; CULLIS, B. R. Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. **Euphytica**, v. 157, n. 1, p. 253–266, 2007.

SNOWDON, R. J.; WITTKOP, B.; CHEN, T.-W.; STAHL, A. Crop adaptation to climate change as a consequence of long-term breeding. **Theoretical and Applied Genetics**, v. 134, n. 6, p. 1613–1623, 2021.

SOUZA, V. F.; RIBEIRO, P. C. O.; VIEIRA JÚNIOR, I. C.; OLIVEIRA, I. C. M.; DAMASCENO, C. M. B.; SCHAFFERT, R. E.; PARRELLA, R. A. C.; DIAS, K. O. G.; PASTINA, M. M. Exploring genotype × environment interaction in sweet sorghum under tropical environments. **Agronomy Journal**, v. 113, n. 4, p. 3005–3018, 2021.

STRINGER, J. K.; ATKIN, F. C.; GEZAN, S. A. Statistical approaches in plant breeding: maximising the use of the genetic information. In: GENETIC Improvement of Tropical Crops. Cham: Springer International Publishing, 2017. P. 3–17.

VELAZCO, J. G.; RODRÍGUEZ-ÁLVAREZ, M. X.; BOER, M. P.; JORDAN, D. R.; EILERS, P. H. C.; MALOSETTI, M.; EEUWIJK, F. A. van. Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. **Theoretical and Applied Genetics**, v. 130, n. 7, p. 1375–1392, 2017.

VERBYLA, A. P.; FAVERI, J.; DEERY, D. M.; REBETZKE, G. J. Modelling temporal genetic and spatio-temporal residual effects for high-throughput phenotyping data. **Australian & New Zealand Journal of Statistics**, v. 63, n. 2, p. 284–308, 2021.

VERBYLA, A. P. A note on model selection using information criteria for general linear models estimated using REML. **Australian & New Zealand Journal of Statistics**, v. 61, n. 1, p. 39–50, 2019.

WEISSHUHN, P.; RECKLING, M.; STACHOW, U.; WIGGERING, H. Supporting agricultural ecosystem services through the integration of perennial polycultures into crop rotations. **Sustainability**, v. 9, n. 12, p. 2267, 2017.

WICKHAM, H. **ggplot2: Elegant graphics for data analysis**. 2. ed. Cham: Springer, 2016.

WOLFINGER, R. D. Heterogeneous variance: covariance structures for repeated measures. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 1, n. 2, p. 205–230, 1996.

ZHANG, R.; HAN, D.; HU, X. Analyzing the performance of corn in China using a factor-analytic variance-covariance structure with multiple factors. **Crop Science**, v. 60, n. 1, p. 190–201, 2020.

CHAPTER 3

EMPLOYING THE FACTOR ANALYTIC TOOLS FOR SELECTING HIGH-PERFORMANCE AND STABLE TROPICAL MAIZE HYBRIDS

Published article: CHAVES, S. F. S.; EVANGELISTA, J. S. P. C.; TRINDADE, R. S.; DIAS, L. A. S.; GUIMARÃES, P. E.; GUIMARÃES, L. J. M.; ALVES, R. S.; BHERING, L. L.; DIAS, K. O. G. Employing factor analytic tools for selecting high-performance and stable tropical maize hybrids. **Crop Science**, v. 63, n. 3, p. 1114–1125, 2023.

Genotype-by-environment interaction (GEI) is a major concern in tropical maize breeding. Here, we used the Factor Analytic Mixed Models (FAMM) to study the GEI in a tropical maize data set. Our goal was to select high-performance and stable hybrids using the Factor Analytic Selection Tools (FAST), derived from the FAMM outputs. Since the data set comprised two different planting seasons, we also investigated the differences in the genetic gains between a season-wise selection and a general selection (i.e. for both seasons simultaneously). The trials were installed in a lattice design with 36 hybrids, two replicates and blocks with six plants. We evaluated 53 hybrids and seven checks in 48 environments, represented by the combination of locations, years and seasons. We fitted the FAMMs with a different number of factors. The best-fitted FAMM was selected considering a parsimony-explained variance balance. With the best FAMM for both season-wise and general analyses, we estimated the selection tools: overall performance (OP), root mean squared deviation ($RMSD$), which represents general stability; and responsiveness (RE), which represents specific stability. The selected hybrids had the highest OP amongst the ones with $RMSD < 0.5$ (the lower the $RMSD$, the better). Only two and three hybrids were selected considering both joint and season-wise analyses. One may reach higher genetic gains by employing FAMM and using FAST in each season rather than jointly analysing both seasons.

1 Introduction

Different phenotypic responses of maize (*Zea mays* L.) genotypes to the distinct environments is a challenging task for cultivar recommendations in tropical and temperate environments. Tropical environments are more prone to outbreaks of pests and diseases and have major environmental differences between locations, years and seasons within years (PATERNIANI, 1990; SANDHU; DHILLON, 2021; SILVA et al., 2022). The differential genotypic responses to these geographical and temporal variations compose the genotype-by-environment interaction (GEI), a hindering agent for cultivar recommendation (CROSSA, 2012; EEUWIJK et al., 2016; MALOSETTI et al., 2013). Then, exploring GEI is essential for diminishing the risks of decision-making, i.e. selecting and recommending cultivars.

Exploring GEI based on an appropriate stability analysis can leverage the recommendation of adapted crops to fit agricultural environment needs (DIAS et al., 2022; WALLACE et al., 2018). Early methods used ordinary least squares (OLS)-based approaches to study GEI (MALOSETTI et al., 2013). A method that has been thoroughly used in plant breeding is the regression of the genotypes' performance on the environmental index, in which the general or specific stability of a genotype is based on its putative reaction norm (FINLAY; WILKINSON, 1963). Nonetheless, the attempt to explain such a complex phenomenon based on a data-built covariate (environmental index) may hinder the efficiency and bias the results, since the regressor is subject to the same errors as the response variable (CROSSA, 1990; MALOSETTI et al., 2013; SMITH et al., 2005). Multiplicative methods such as AMMI (Additive Main effects and Multiplicative Interaction effects) (GAUCH, 1988) and GGE (Genotype + Genotype-by-environment) biplot (YAN et al., 2000) try to fill this gap by exploring the multiple facets of GEI using multiplicative terms, exploiting the biplots for graphical support. The biplots provide detailed information about the GEI patterns in the data. On the other hand, those plots can be complex and hard to interpret in large datasets, hampering the selection and recommendation of candidates (DIAS et al., 2022; SMITH; CULLIS, 2018).

The aforementioned methods have deficiencies linked to the OLS base: the data must be balanced, the variances are assumed to be homogeneous and the genotypic effect is considered to be fixed (CROSSA, 2012; SMITH et al., 2005). Conversely, a mixed model approach surpasses these drawbacks, providing more reliable estimates to interpret GEI (EEUWIJK et al., 2016). Piepho (1997) proposed the Factor Analytic Mixed Models (FAMM), which was later given a thorough account by Piepho (1998) and Smith et al. (2001). FAMM considers the

genetic covariance heterogeneity, does not require genetic or statistical balance, and benefits from the advantages of considering the genotypic effect as random (GOGEL et al., 2018; SMITH et al., 2005, 2001).

FAMM itself does not provide the succinct results required to expedite the decision-making, pivotal in a plant breeding program routine. In this sense, Cullis et al. (2014) leveraged the multiple regressions within the FAMM to build the latent regression plots, in which the y-axis represents the genotypic values and the x-axis is the k factor loadings. In the latent regression, the k factor scores are the regression coefficients, which dictate the slopes of the lines that represent the genotype's performance and stability. Despite recalling an intuitive manner for interpreting GEI, latent regressions have limited utility for selection, especially when a lot of candidates are being considered. A more straightforward method was proposed by Smith and Cullis (2018), the so-called Factor Analytic Selection Tools (FAST), which compiles the FAMM outputs into a set of estimates that can be used directly as a base for selection. These estimates are Overall Performance (OP), which measures the performance of the candidates for the evaluated trait; Root Mean Squared Deviation ($RMSD$), which is a measure of general stability; and Responsiveness (RE) to the k factor, which represents the specific stability, depending on the loadings sign in the k factor. A genotype that has high performance and general stability has a high OP, and $RMSD$ and RE next to the nullity. Biplots with these estimates can aid the selection. The seminal work used data from a *Pinus radiata* D. Don. breeding program (SMITH; CULLIS, 2018). Lately, other breeders used FAST for decision-making, e.g. (SJOBERG et al., 2021), who employed FAST for thoroughly studying the GEI and recommended *Triticum aestivum* L. varieties, and Tolhurst et al. (2019) whose work expanded the FAST to a genomic selection context, also in *T. aestivum*.

To the best of our knowledge, studies employing FAST for recommendation in tropical maize breeding remain scarce. Here, we apply the FAST in a late-stage dataset of tropical maize evaluated in a target population of environments in two crop seasons. Note that these seasons have distinct environmental conditions. The first season tends to have a larger photoperiod, water availability, and solar radiation, i.e. better conditions for maize cultivation. On the other hand, the second season usually has a shorter planting window and a higher risk of yield losses, given the frequent occurrence of drought (DIAS, F. S. et al., 2019; NÓIA JÚNIOR; SENTELHAS, 2019). Thus, our goals were to select high-performance hybrids with high stability using FAST and compare the genetic gains between a season-specific and a general

selection (for both seasons).

2 Material and Methods

2.1 Plant material and experimental conditions

The data refers to late-stage breeding trials of Embrapa Milho e Sorgo (Brazilian Agricultural Research Corporation, Maize and Sorghum unit). The trials were installed in a lattice design, with 36 hybrids, two replicates and blocks with six plants. Two lines of four-meter length, spaced by 0.8 m composed each experimental plot.

In total, we evaluated 53 hybrids (H1 to H53) and seven checks (C54 to C60) in two consecutive years, two different seasons within a year and 27 locations across the tropical breeding regions in Brazil (Figure 1). Here, we assigned an "environment" as the combination of years, seasons and locations, totalling 48 environments. Twelve hybrids and 15 environments are common between years (Figure 2). The trials were conducted following maize's recommended agronomical practices for each location. We evaluated the grain yield trait (GY, adjusted to 13% of grain moisture, converted to t ha^{-1}) at the plot level.

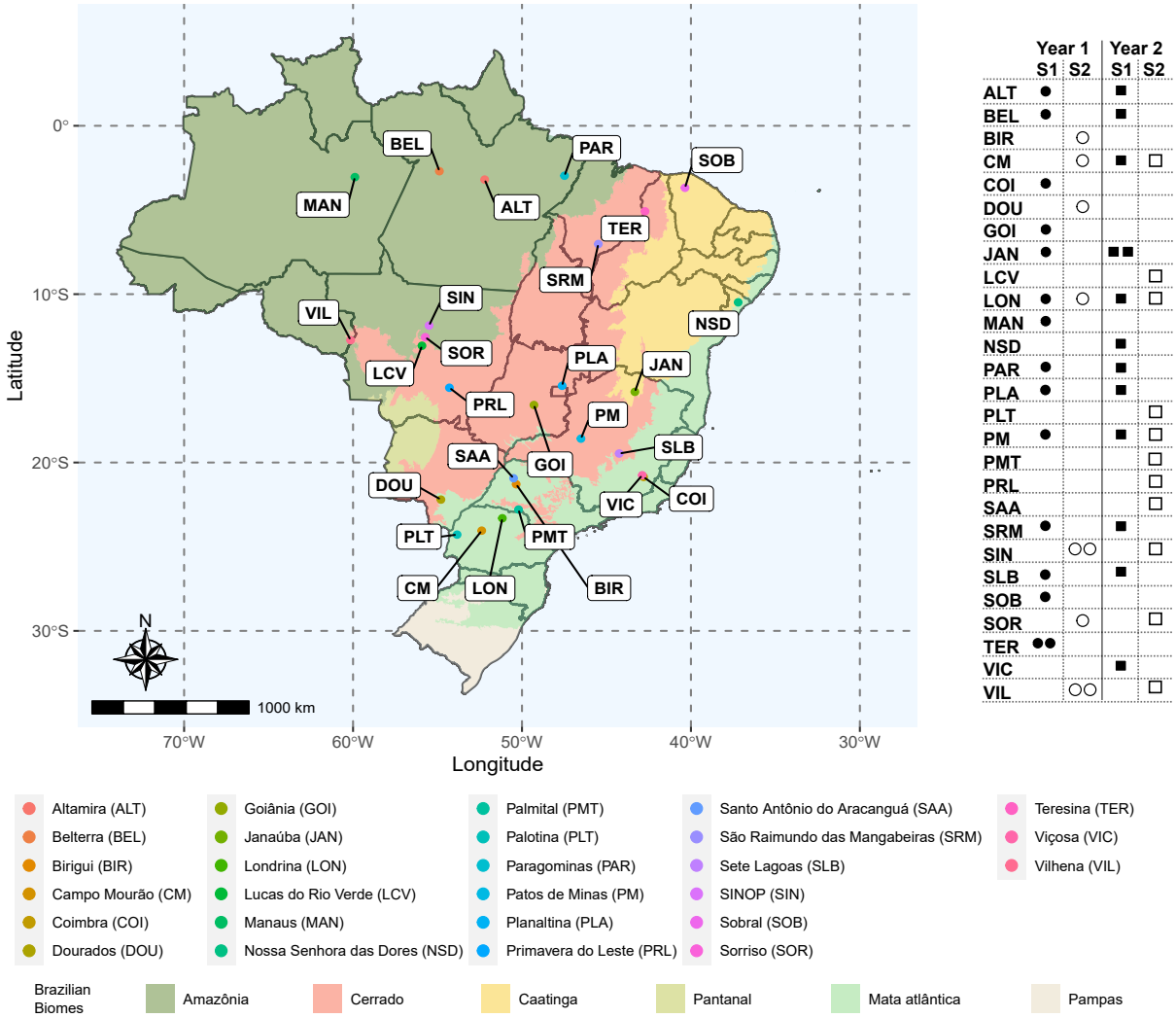
2.2 Statistical analyses

We performed all the statistical analyses employing the linear mixed models, using the Residual Maximum Likelihood (REML, PATTERSON; THOMPSON, 1971) method to estimate the variance components. Those were acquired through the iteration process using the mixed model equations (HENDERSON, 1975). We performed a single-step Multi-Environment Trials (MET) analysis. The model that evaluates the n data of the h hybrids allocated in q blocks within the r replicates in m environments is:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}_1\mathbf{a} + \mathbf{X}_2\mathbf{r} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{b} + \mathbf{e} \tag{1}$$

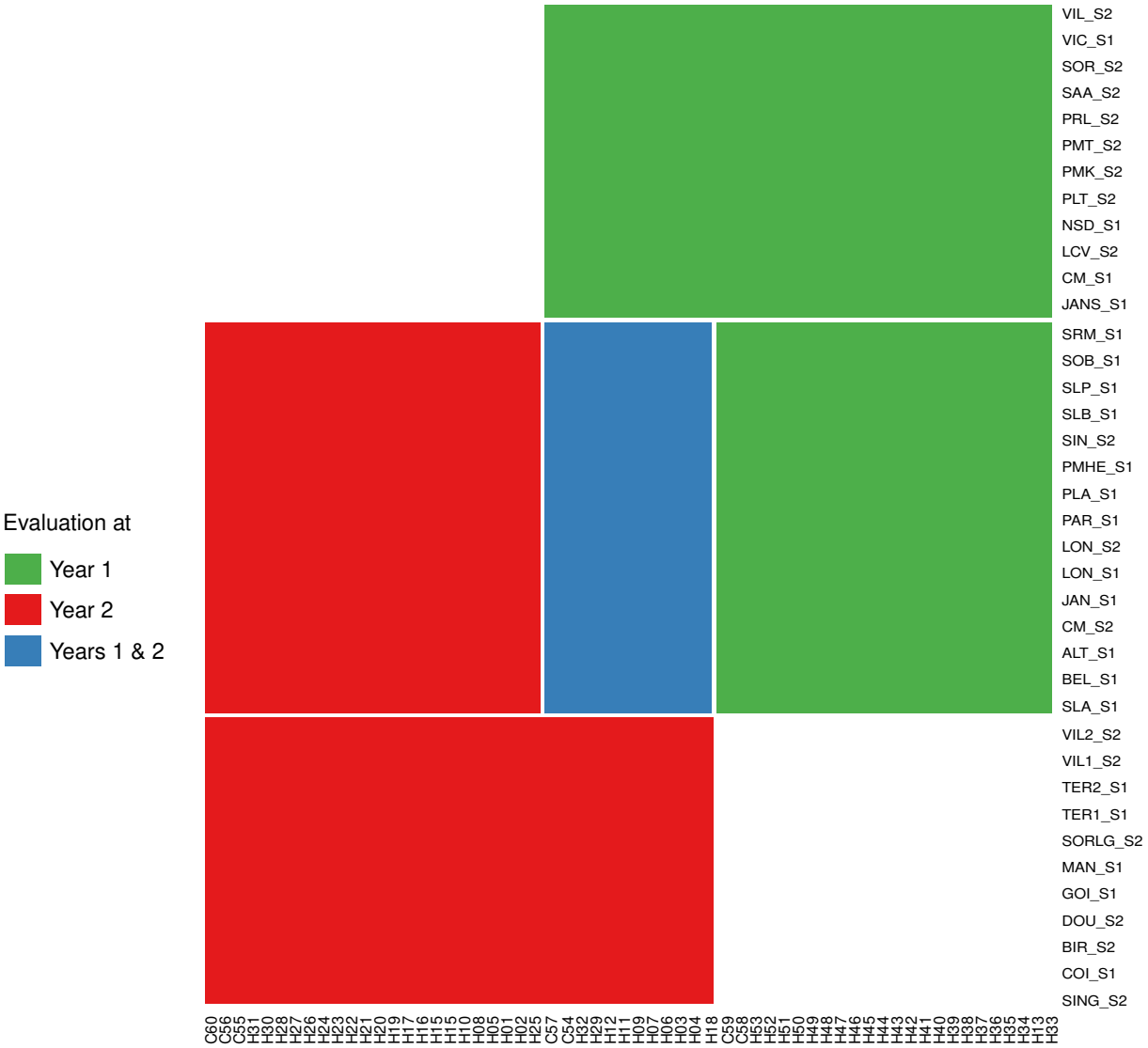
where $\mathbf{y}^{(n\times 1)}$ is the vector of phenotypes, μ is the intercept, $\mathbf{a}^{(m\times 1)}$ is the vector of fixed effects of environments, $\mathbf{r}^{(mr\times 1)}$ is the vector of fixed effects of replicates within environments; $\mathbf{g}^{(mh\times 1)}$ is the vector of random genotypic effects within environment [$\mathbf{g} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_g \otimes \mathbf{I}_h)$]; $\mathbf{b}^{(mrq\times 1)}$ is the vector of random block effects within replicates and environments [$\mathbf{b} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_b) \otimes \mathbf{I}_q$], and $\mathbf{e}^{(n\times 1)}$ is the vector of random error effects [$\mathbf{e} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_e) \otimes \mathbf{I}_n$]. $\mathbf{X}_1^{(n\times m)}$, $\mathbf{X}_2^{(n\times mr)}$,

Figure 1: Location (map), year and season (caption on the right) of each trial



$\mathbf{Z}_1^{(n\times mh)}$ and $\mathbf{Z}_2^{(n\times mhq)}$ are the design matrices for their respective effects and $\mathbf{1}^{(n\times 1)}$ is a vector of ones. $\mathbf{\Sigma}_g$, $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_e$ are the covariance matrices of the genotypic, block and residual effects, respectively. $\mathbf{\Sigma}_b$ and $\mathbf{\Sigma}_e$ were modelled as diagonal matrices, so $\mathbf{\Sigma}_b = \oplus_{j=1}^{m}\sigma_{b_j}^2\mathbf{I}_m$ and $\mathbf{\Sigma}_e = \oplus_{j=1}^{m}\sigma_{e_j}^2\mathbf{I}_m$, where σ_b^2 and σ_e^2 are variance components for the block and error effects, respectively, $\oplus$ is a direct sum and $\mathbf{I}_m$ is an identity matrix whose dimension is the number of environments. Consequently, we are considering that each environment has independent error and block variance components. For modelling $\mathbf{\Sigma_g}$, we used a Factor Analytic Mixed Model (FAMM, PIEPHO, 1997; SMITH et al., 2001). Thus, $\mathbf{\Sigma_g} = (\mathbf{\Lambda\Lambda'}) + \mathbf{\Psi}$, in which $\mathbf{\Lambda}^{(m\times\kappa)}$ is the matrix of loadings, with κ being the number of factors ($\mathbf{\Lambda} = \{\lambda_{k_j}\}$, where λ_{k_j} is a factor loading of the j-th environment in the k-th factor), $\mathbf{\Psi}^{(m\times m)}$ is the diagonal matrix of specific variances ($\mathbf{\Psi} = \{\psi_j\}$, where ψ_j is the specific variance of the j-th environment), i.e. not explained by any

Figure 2: Years (colours), locations and seasons (y-axis) in which each hybrid (x-axis) was evaluated.



factors.

In the FAMM context, we can write $\mathbf{g}$ as:

$$\mathbf{g} = (\mathbf{\Lambda} \otimes \mathbf{I}_h)\mathbf{f} + \boldsymbol{\delta} \tag{2}$$

where $\mathbf{f}^{(h\kappa \times 1)}$ is the vector of genotypic scores and $\boldsymbol{\delta}^{(hm \times 1)}$ is the vector of lack of fit effects. Note that $V(\boldsymbol{\delta}) = \mathbf{\Psi} \otimes \mathbf{I}$, so the lack of fit is associated with the specific variances (SMITH et al., 2001). This allows us to separate the effects within $\mathbf{g}$ into common $[(\mathbf{\Lambda} \otimes \mathbf{I}_h)\mathbf{f}]$ and specific effects ($\boldsymbol{\delta}$). By disregarding the specific effects, $\mathbf{g}$ has a broader interpretation and prioritizes predictability for common GEI effects rather than isolated environmental effects (CULLIS et

al., 2010; SMITH et al., 2015). Therefore, we based the selection only on the common effects.

When $\kappa > 1$, the non-uniqueness of $\mathbf{\Lambda}$ hinders its estimation, so it is mathematically convenient to impose constraints. Here, the upper diagonal values of $\mathbf{\Lambda}$ were considered zero. This is the standard constraint of `ASReml-R` (BUTLER et al., 2018), the package we used to perform the statistical analyses (version 4.1) within the `R` software environment, version 4.2.1. To retrieve the biological meaningfulness, it is fundamental to perform a rotation, i.e. make the columns of $\mathbf{\Lambda}$ orthogonal. Here, we performed the Singular Value Decomposition rotation [see Cullis et al. (2010) for more details], obtaining $\mathbf{\Lambda}^*$ and $\mathbf{f}^*$, i.e. the rotated loadings and scores. With $\mathbf{\Lambda}^*$ and $\mathbf{f}^*$ we estimated the genetic covariances and the predicted genotypic values, as demonstrated in the previous paragraph.

We defined the number of factors considering a parsimony-explained variance balance. The first criterion was the AIC (Akaike Information Criterion, AKAIKE, 1974), given by:

$$AIC = -2logL + 2p \tag{3}$$

where L is the likelihood estimate and p is the number of estimated parameters.

In the FAMM context, AIC may not reflect the ideal model (ISIK et al., 2017; SMITH et al., 2015). An auxiliary parameter such as the percentage of explained variance by the latent variables should aid in the model selection. Thus, we calculated the explained variance by all κ factors ($\overline{v}$) and by each factor in each environment (v_{k_j}). The following equations give these parameters, respectively (SMITH et al., 2015):

$$\overline{v} = \frac{\text{Trace}(\mathbf{\Lambda}\mathbf{\Lambda}')}{\text{Trace}(\mathbf{\Sigma}_g)} \tag{4}$$

$$v_{k_j} = \frac{\lambda_{k_j}^2}{\sum_{k=1}^{\kappa} \lambda_{k_j}^2 + \psi_j} \tag{5}$$

where Trace is the sum of the diagonal elements.

From the best-fitted model, we estimated the environment-wise accuracies (r, MRODE, 2014):

$$r = \sqrt{1 - \frac{PEV}{\sigma_g^2}} \tag{6}$$

where PEV is the prediction error variance, obtained from the diagonal of the coefficient matrix's inverse, and σ_g^2 is the genotypic variance

We estimated the generalized heritabilities for each environment (H^2), given by (CULLIS et al., 2006):

$$H^2 = 1 - \frac{\overline{V(\Delta)}}{2\sigma_g^2} \tag{7}$$

where $\overline{V}(\Delta)$ is the mean pairwise prediction error variance.

We also calculated the experimental coefficient of variance (CV_e), given by:

$$CV_e = \frac{\sqrt{\sigma_e^2}}{\mu} \tag{8}$$

Lastly, we calculated the type B genetic correlations, given by ($\rho_{g_{jj'}}$, CULLIS et al., 2014):

$$\rho_{g_{jj'}} = \frac{\sum_{k=1}^{\kappa} \lambda_{k_j}^2 \lambda_{k_{j'}}^2}{\sqrt{\sigma_{g_j}^2 \sigma_{g_{j'}}^2}} \tag{9}$$

we represented the correlations in a *heatmap* and associated them with a dendrogram, which clustered the environments using the "complete" method. For this, we used the `ComplexHeatmap` package, version 2.12.0 (GU et al., 2016).

Factor analytic selection tools

Here, we employed the selection tools proposed by Smith and Cullis (2018). Recall that $\mathbf{g} = (\mathbf{\Lambda}^* \otimes \mathbf{I}_h)\mathbf{f}^*$, so $g_{ij} = \lambda_{1_j}^* f_{1_i}^* + \lambda_{2_j}^* f_{2_i}^* + \lambda_{3_j}^* f_{3_i}^* + ... + \lambda_{k_j}^* f_{k_i}^*$ and therefore $g_{ij} = \lambda_{1_j}^* f_{1_i}^* + \epsilon_{ij}$, with $\epsilon_{ij} = \lambda_{2_j}^* f_{2_i}^* + \lambda_{3_j}^* f_{3_i}^* + ... + \lambda_{k_j}^* f_{k_i}^*$. This relation is useful to calculate the overall performance (OP_i) and the root mean squared deviation ($RMSD_i$), which were the parameters we used to select the high-performance and stable hybrids. Those are acquired by the following equations (SMITH; CULLIS, 2018):

$$OP_i = \frac{1}{m} \sum_{j=1}^{m} \lambda_{1_j}^* f_{1_i}^* \tag{10}$$

$$RMSD_i = \sqrt{\frac{1}{m} \sum_{j=1}^{m} \epsilon_{ji}^2} \tag{11}$$

bear in mind that the higher the OP and the lower the $RMSD$, the better.

For studying the selected hybrids' specific stability, we calculated their responsiveness to

all factors but the first, and the latent regressions. This is because the first factor indicates performance and the remainder, stability (CULLIS et al., 2014). The responsiveness (RE_{ik}) was calculated using the following equation (SMITH; CULLIS, 2018):

$$RE_{k_i} = (\overline{\lambda}^*_{k+} - \overline{\lambda}^*_{k-}) f^*_{k_i} \tag{12}$$

where $\overline{\lambda}^*_{k+}$ and $\overline{\lambda}^*_{k-}$ are the mean positive and negative loadings of the k-th factor, respectively. We performed the analyses in the whole dataset (both seasons, 48 environments) and for each season (28 environments in the first season and 20 environments in the second season). The plots that supported the selection were built in the `ggplot2` package (WICKHAM, 2016).

Genetic gains

After selecting we predicted the genetic gains, given by the following equation:

$$GG = \frac{\sum_{i=1}^{t} g_{ij}}{\mu} \tag{13}$$

where t is the number of selected hybrids. Following Embrapa's standard procedure, we selected the top eight (13%) hybrids regarding both performance and stability (DIAS et al., 2020). The criteria for selection were the eight hybrids with the highest OP and had a $RMSD$ lower than 0.5. We compared the gains in the global and season-wise analyses by exchanging the selected hybrids between analyses, i.e. estimating the GG of an analysis (e.g. global) using the selected hybrids of another (e.g. only the first season).

3 Results

We prioritized the model with the lowest AIC amongst the ones that explained at least 70% of the total variance (Figure 3). Thus, the model with four factors (FA4) was the choice for the global analysis (Figure 3A) and for the analysis regarding only the first season, whereas the FA3 met the criteria above for the analysis regarding only the second season (Figures 3B and 3C). Even though FA3 had an explained variance of 73% in the first season analysis, half of the first factor loadings' signs were negative. This division would compromise FAST meaningfulness. Smith and Cullis (2018) recommend that the majority of the first factor loadings are positive as a prerequisite for using FAST. We met this requisite using the FA4, a model with a somewhat

higher AIC, but a considerably higher explained variance (79%). There was a highly variable factor-environment relationship (Figure 3). Note that some environments had low explained variance. This fact is related to the percentage of variance not explained by the model.

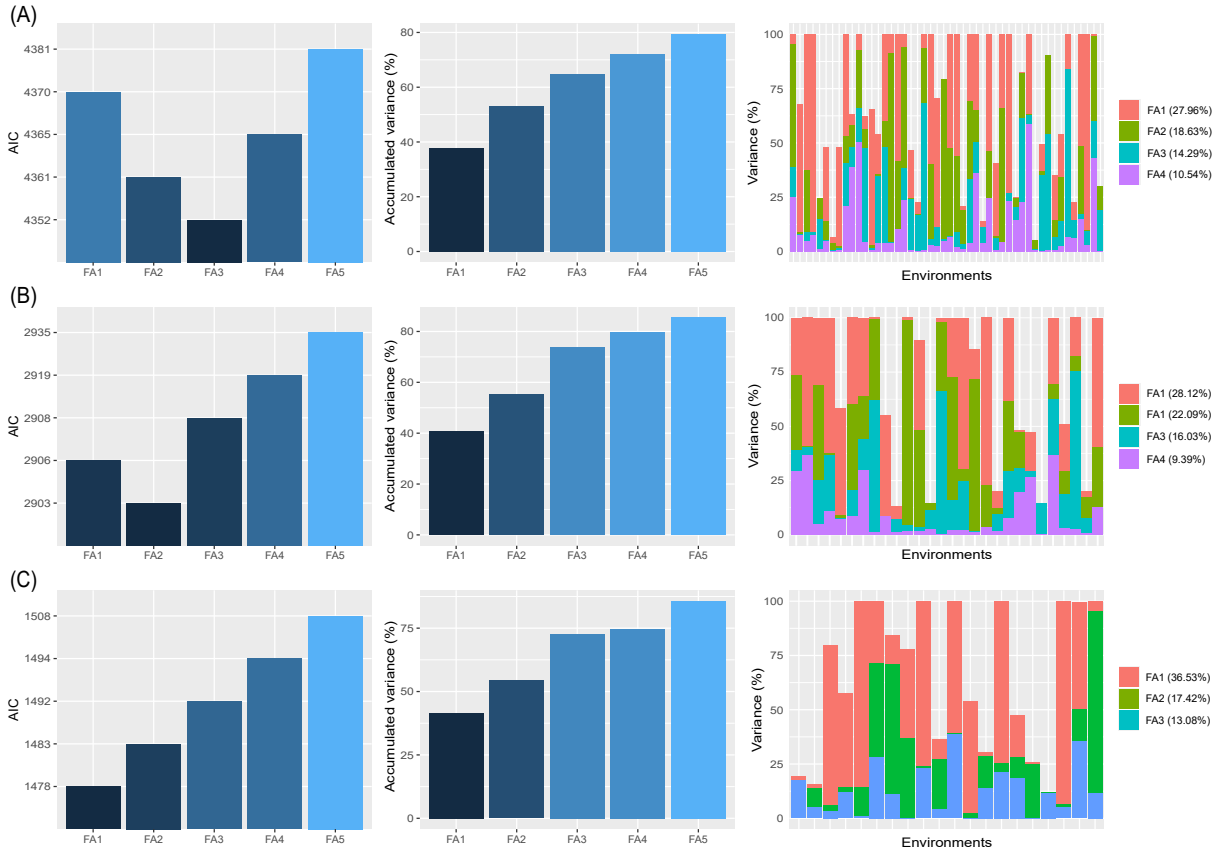


Figure 3: Akaike Information Criterion (AIC), percentage of explained total variance of each Factor Analytic Mixed Model and by factor in each environment in the best-fitted model (model that explains more than 70% of the total variance with the lowest AIC) regarding both seasons (A, 48 environments), only season one (B, 28 environments) and only season two (C, 20 environments)

Most environments showed high accuracies and heritabilities, and low experimental coefficients of variation (CV_e) (Figure 4). The genetic parameters were slightly higher when both seasons were considered in the statistical analysis (mean accuracy = 0.80 and mean heritability = 0.70). The mean accuracy in the second season (0.78) was somewhat higher than in the first season (0.75), whereas the heritability was similar between seasons (0.64). The CV_e was somewhat lower in the second season (0.13) than in the first season and the global analyses (both 0.15).

The type B genetic correlations between environments ranged from -0.98 to 0.99, with a mean of 0.11 (Figure 5). Note that high-correlated environments are not necessarily geographically close. Indeed, the same locations evaluated in different seasons and/or years

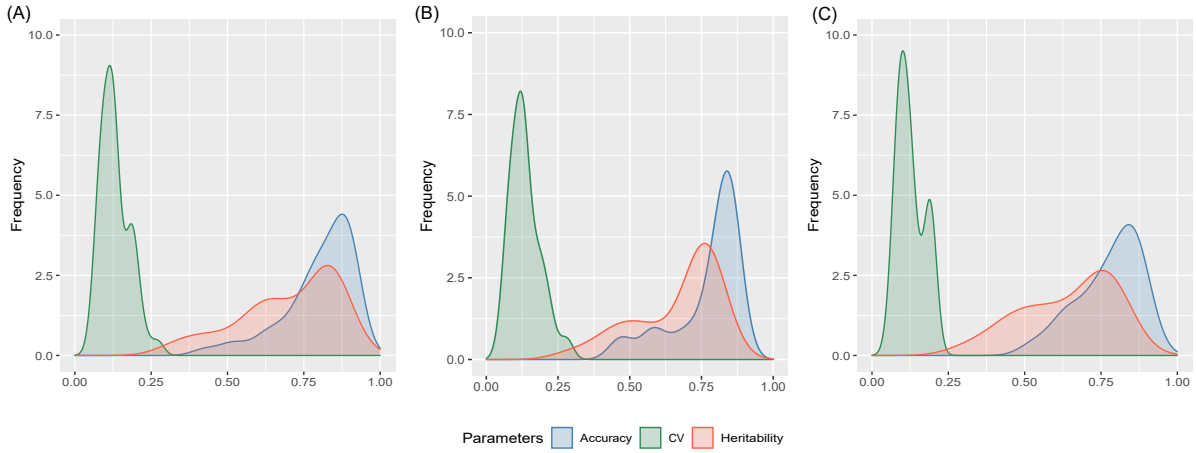


Figure 4: Distribution of accuracies, experimental coefficients of variation (CV_e) and generalized heritabilities across environments regarding both seasons (A, 48 environments), only the first season (B, 28 environments) and only the second season (C, 20 environments). The CV_e are at decimal scale

can have low type B genetic correlations (next to the nullity), indicating the environmental variability between years and seasons [see Janaúba's (JAN1) case in Figure 5 for an example].

We have identified the top eight most productive and stable hybrids according to the global analysis (Figure 6) and the season-wise analyses (Figure 7). The checks (C54 to C60) were not considered in the selection. We selected these hybrids using an $OP \times RMSD$ biplot. OP ranks the hybrids according to their production, so the higher the OP (or the higher the location of the hybrid in Figures 6 and 7), the more productive is the hybrid. $RMSD$ has the opposite interpretation, i.e. the lower the value (or the closer to zero in Figures 6 and 7), the more stable is the hybrid. Note that the criteria for selection are up to the breeder. Here, we highlighted only the top eight hybrids that had $RMSD$ lower than 0.5 with the higher OP.

Recall that $RMSD$ is a measure of general stability. For studying specific stability, one can substitute the $RMSD$ in the biplot for the responsiveness to all factors but the first (RE). Here, these plots were built only for season-wise analyses (Figure 8). According to the loadings sign and the degree of relation between the factor and the environment, one can visualize if the hybrids are more suitable for some environments rather than others. For example, H29 has higher performance for environments with positive loadings in the second factor and negative in the third factor in the first season (Figure 8A). On the other hand, the same hybrid is not influenced by the loadings' sign either in the second or third factor, indicating high stability in the second season (Figure 8B). Note in Figure 7B that H29 is the most stable hybrid. Thus, one can make a direct link between the hybrids' $RMSD$ (Figure 7) and their position regarding RE

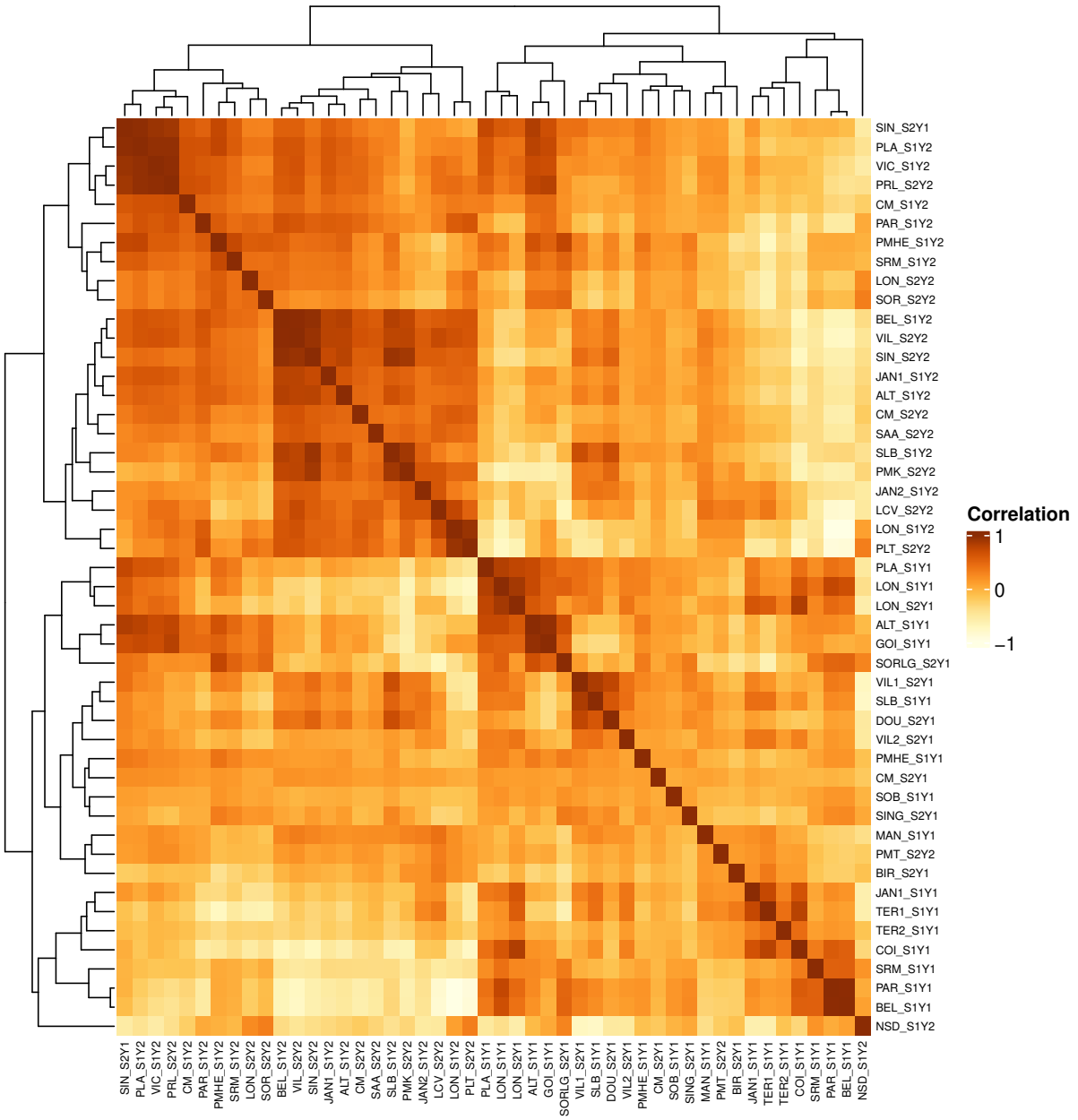


Figure 5: *Heatmap* of the type B genetic correlations between the 48 environments. The environments are clustered according to their genetic similarity. In the environments' names, "S1" and "S2" are for the first and second seasons, respectively; "Y1" and "Y2" are for the first and second years, respectively; and the name preceding the underscore symbol (_) is a code for the location

(Figure 8).

Only two (H29 and H51) and three (H28, H29 and H38) hybrids were selected considering both joint analysis and the individual analyses for the first and second seasons, respectively. Indeed, by selecting the putative best hybrids pointed out by the global analysis, the genetic gains would be jeopardized in the first and second seasons (Table 1). On the other hand, by particularizing the analysis for each season, the genetic gains are higher and more reliable (Table

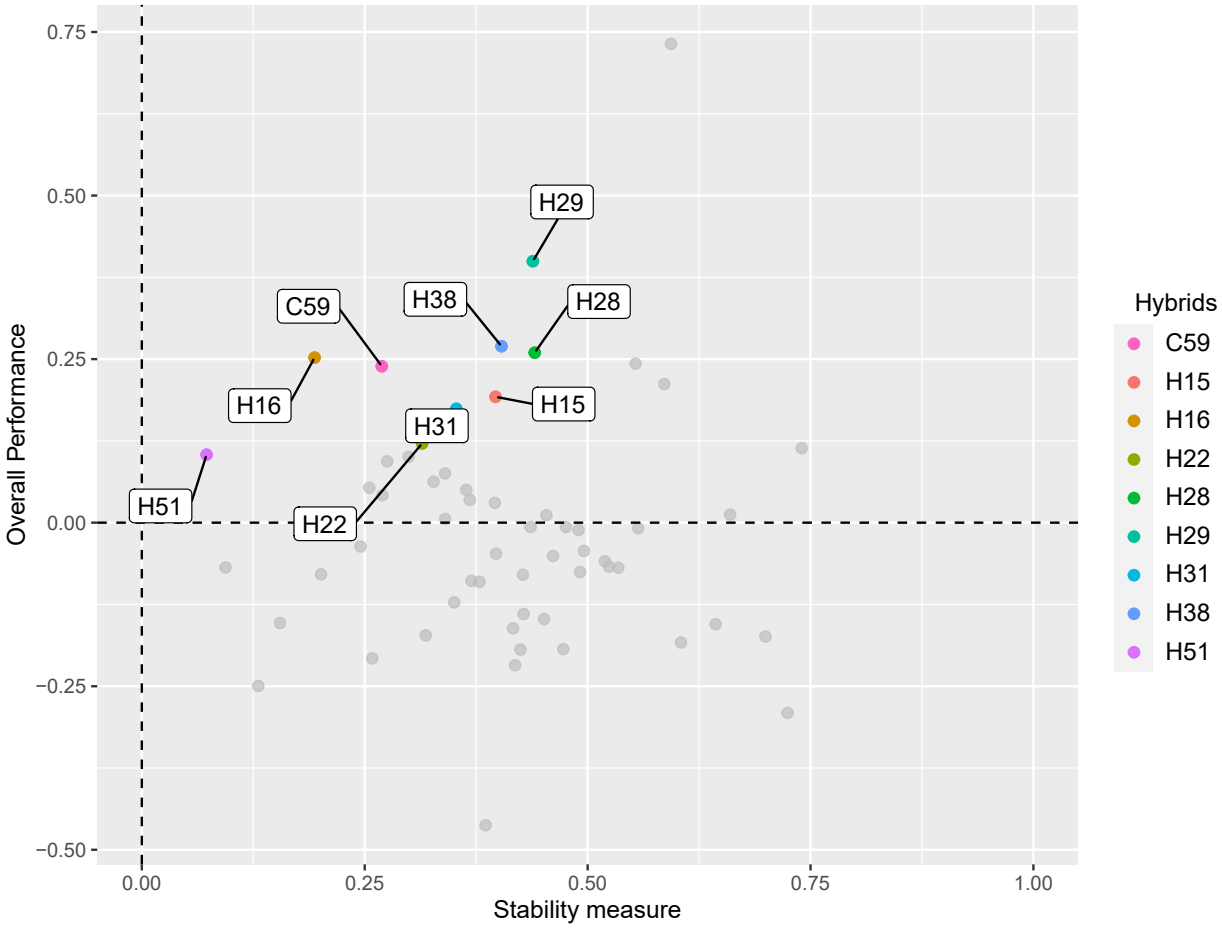


Figure 6: Overall performance (y-axis) and stability (x-axis) of the 53 maize hybrids and 7 checks regarding both seasons, i.e. 48 environments. The selected hybrids according to criteria of the $RMSD$ lower than 0.5 with the higher OP are highlighted

1).

Table 1: Genetic gains considering the selection of the top 13% maize hybrids according to the joint analysis and the season-wise analysis

Selection	Gain at		
	Joint	First season	Second season
Joint	23.66%	20.52%	14.99%
Only first season	15.11%	21.10%	1.40%
Only second season	3.37%	-0.26%	39.89%

4 Discussion

This study showed the efficiency of using the FAST proposed by Smith and Cullis (2018) on a tropical ongoing commercial maize breeding program. FAST summarized the complex FAMM outputs into three estimates (OP, $RMSD$ and RE), which composed the biplots that

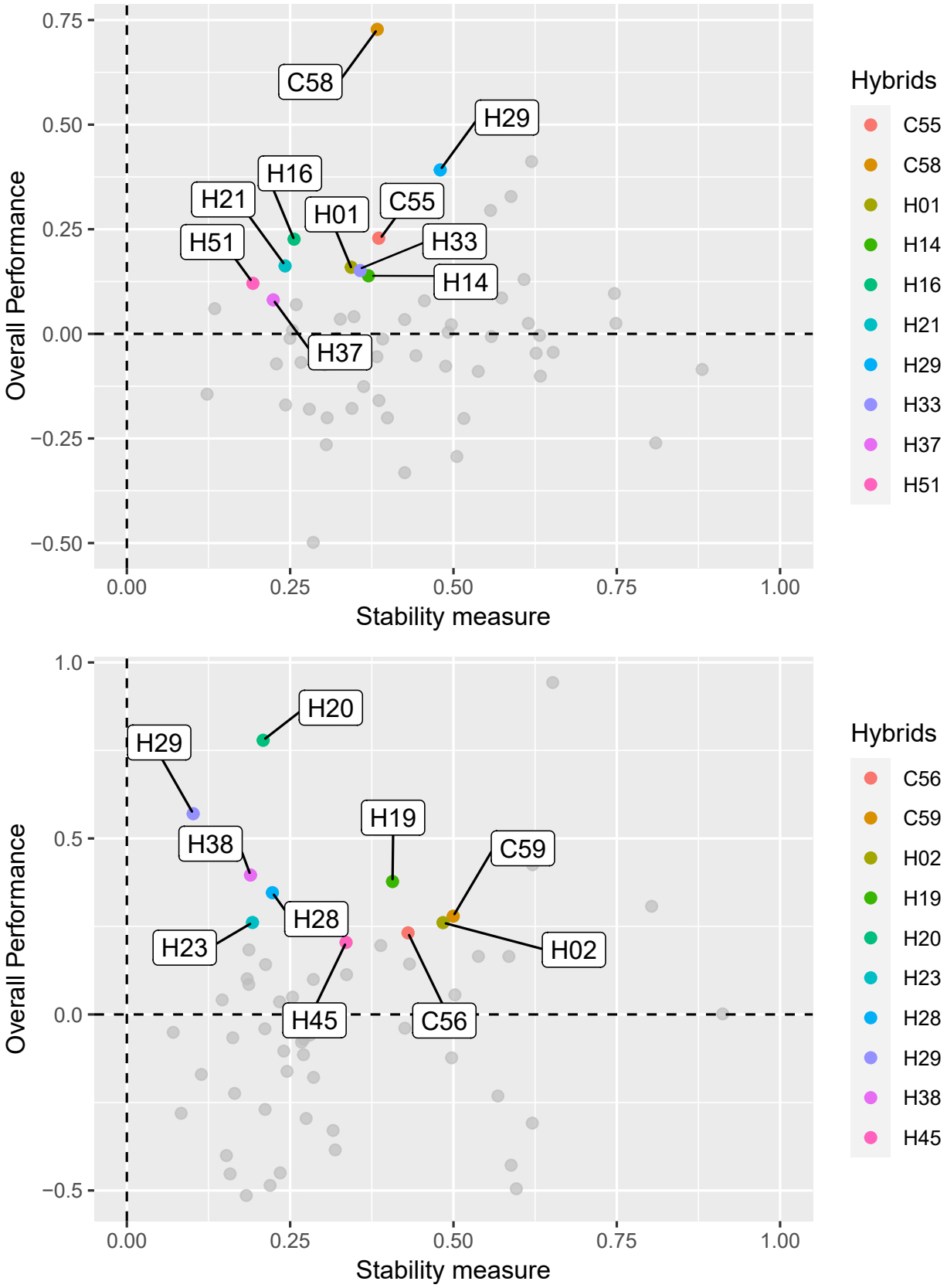


Figure 7: Overall performance (y-axis) and stability (x-axis) of the 60 maize hybrids regarding the first (A, 28 environments) and the second season (B, 20 environments). The selected hybrids according to criteria of the $RMSD$ lower than 0.5 with the higher OP are highlighted

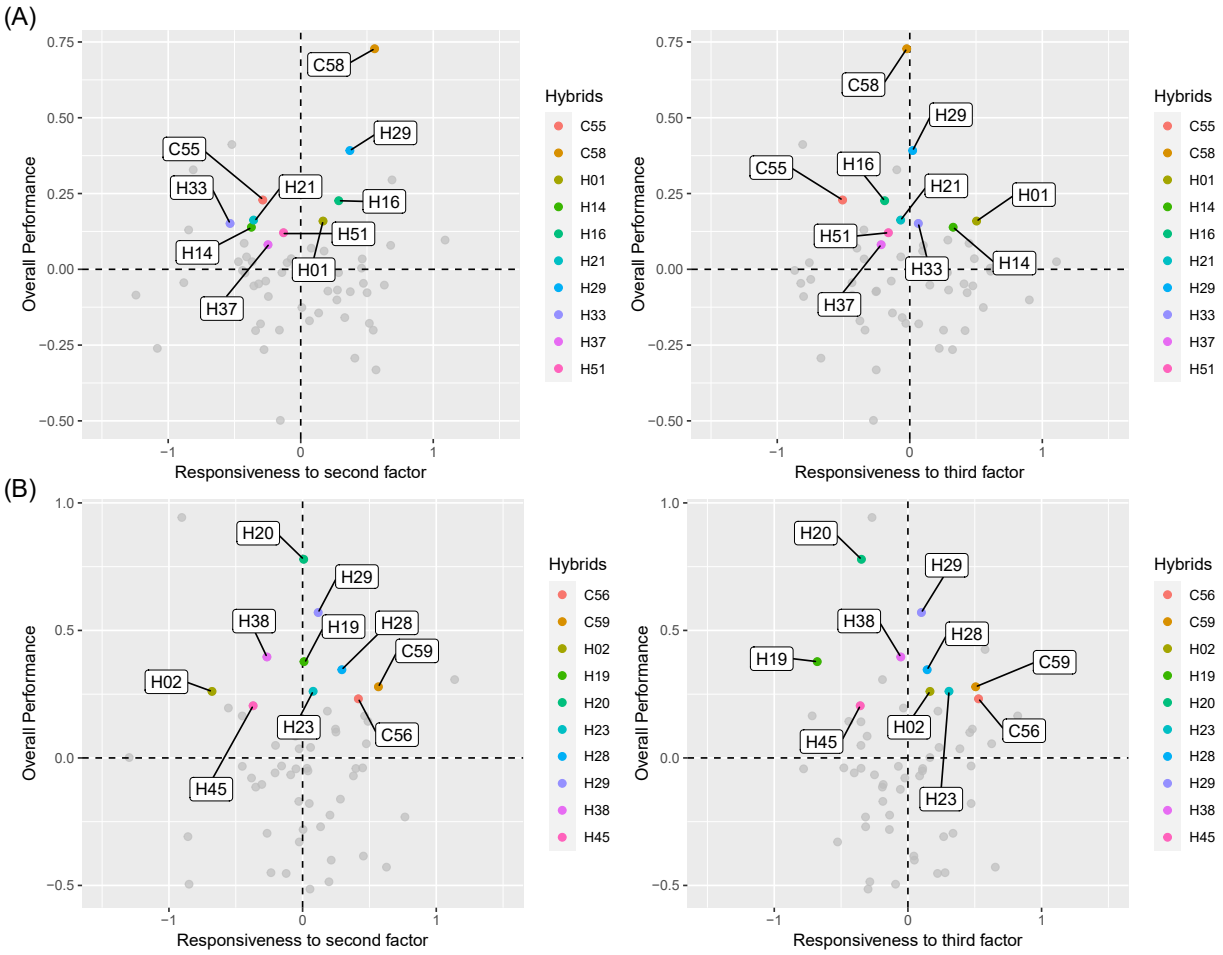


Figure 8: Overall performance and responsiveness to second and third factors of the 53 maize hybrids and 7 checks regarding the first season (A, 28 environments) and the second season (B, 20 environments)

supported the selection. In addition, particularizing the analysis for each season might lead to an improvement in the selection and higher genetic gains, especially for the second season. This proves that the different environmental conditions between seasons can influence phenotypic expression and mislead a global selection.

Using AIC as the sole criterion for model selection is not the best choice in the FAMM context (Figure 3). Breeders must seek models that keep the computational efficiency and parsimony, and, at the same time, account for a large portion of the dataset variance. This stresses the importance of using the overall explained variance as an auxiliary method for model selection (ISIK et al., 2017; SMITH et al., 2015; ZHANG et al., 2020). Note that in this study, the comparison was only between models with a different number of factors. This is because the residual and block heterogeneity along environments is a natural choice, given the great divergence of environmental conditions (MALOSETTI et al., 2016).

The slight reduction of the accuracy and generalized heritability from the joint analysis

to the season-wise analyses (Figure 4) is due to the smaller number of environments - thus, information - in the former. Nevertheless, the values for these parameters are consistent with the ones normally found in the literature for tropical maize (CANTELMO et al., 2017; DIAS et al., 2018; PEREIRA et al., 2022). Analysing concomitantly the accuracy and CV_e values, one can affirm that there was a high experimental precision, which grants credibility to the results.

Recall that the data comprises almost every Brazilian state, so it is natural that the GEI had an important part in the phenotype. Note from the type B correlation *heatmap* (Figure 5) that there were environments with highly similar or opposite performances and environments with no relevant relation whatsoever. This variation is clear evidence of crossover GEI (CULLIS et al., 2010), which justifies the concern of employing a robust yet straightforward method such as FAST to support the selection. Be aware that Figure 5 represents the GEI dynamics within the data set used herein and cannot be taken as a rule for Brazilian maize breeding. Observe from Figure 5 an explicit differentiation between the environments evaluated in the first year and in the second year. This proves that the yearly environmental variations provoke significant changes in the GEI pattern. Indeed, several studies showed that GEI in tropical maize breeding is highly affected by the annual effects (DIAS et al., 2020; PRASANNA et al., 2021). Thus, a reliable stratification should consider the climate variation over the year through a larger time series and the consideration of environmental covariates.

FAMM can be confidently used for tropical maize breeding data analysis, as performed by previous studies (DIAS et al., 2018; KRAUSE et al., 2020; MENGESHA et al., 2019). Note that Dias et al. (2018) and Krause et al. (2020) employed FAMM for a thorough study of the GEI, whereas Mengesha et al. (2019) selected the best genotypes using the FAMM in an AMMI fashion. Here, we also used the FAMM outputs to study the GEI dynamics, but based the selection on the FAST, a method that, to the best of our knowledge, was yet to be tested in tropical maize. FAST is one of the auxiliary methods that was meant to facilitate the comprehension of the FAMM outputs. In a timeline, Cullis et al. (2010) introduced the use of *heatmaps* to infer about GEI (see the previous paragraph). Then, Cullis et al. (2014) proposed the latent regressions to visualize a genotype's stability. In this method, the number of plots is the number of factors, depending on the selected model. In the first plot, the y-axis is $\mathbf{g}$ and the x-axis is the loadings for the first factor. From the second plot onwards, the y-axes are $\mathbf{g} - \sum_{k=1}^{\kappa-1} \lambda^*_{k_j} f^*_{k_i}$ and the x-axes are the loadings for the k-th factor. The latent regressions detail the hybrids' behaviour across the environments, exploring the differences

between them indicated by the loadings sign. Smith and Cullis (2018) proposed the FAST to facilitate even further the comprehension of the FAMM outputs. By summarizing the results into three estimates, breeders can use them to build selection indexes or biplots (Figures 6, 7 and 8), having the freedom to select using different criteria.

Regarding the selection in a global or season-wise way, particularizing the recommendations for each season may be the best option in tropical maize breeding. We proved this by comparing the genetic gains considering the selected hybrids according to the global analysis of each season (Table 1). Hence, if breeders aim at selecting for global adaptation, it is more reliable to perform the analysis for each season and select the genotypes that are at the top positions of both rankings, e.g. H19, H29 and H38. This happens because the environmental conditions may change drastically from one season to another (NÓIA JÚNIOR; SENTELHAS, 2019; PEREIRA et al., 2022). Consequently, there is a difference in the allele expression between seasons. For instance, second-season hybrids must hold alleles for early flowering and drought tolerance (PRASANNA et al., 2021), since the second season is more prone to drought and has a lower time interval of good climate conditions for cultivation. This is not necessarily true for first-season hybrids.

The interpretations go as far as the FAMM - and FAST - allow. We showed that this method has the flexibility and robustness required to be adopted as a standard procedure in tropical maize breeding. By analysing each season independently, breeders select the high-performance and stable hybrids for each season, resulting in a higher genetic gain. The selected genotypes were H01, H14, H16, H21, H29, H33, H37 and H38 for the first season; H02, H19, H20, H23, H28, H29, H38 and H45 for the second season; and only H29 for both seasons. Extra resources can improve the results even further, such as pedigree or genomic information, environmental covariates and optimization methods (LISLE et al., 2021; OLIVEIRA et al., 2020; TOLHURST et al., 2019).

References

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

BUTLER, D. G.; CULLIS, B. R.; GILMOUR, A. R.; GOGEL, B. J.; THOMPSON, R. **ASReml-R reference manual Version 4**. Hemel Hempstead, UK: VSN International, 2018.

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, n. 4, p. 51, 2017.

CROSSA, J. Statistical Analyses of Multilocation Trials. In: ADVANCES in Agronomy. Elsevier, 1990. v. 44. P. 55–85.

CROSSA, J. From Genotype × Environment Interaction to Gene × Environment Interaction. **Current Genomics**, v. 13, n. 3, p. 225–244, 2012.

CULLIS, B. R.; JEFFERSON, P.; THOMPSON, R.; SMITH, A. B. Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a *Pinus radiata* breeding programme. **Theoretical and Applied Genetics**, v. 127, n. 10, p. 2193–2210, 2014.

CULLIS, B. R.; SMITH, A. B.; BEECK, C. P.; COWLING, W. A. Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. **Genome**, v. 53, n. 11, p. 1002–1016, 2010.

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, n. 4, p. 381–393, 2006.

DIAS, F. S.; REZENDE, W. M.; ZUFFO, L. T.; CAIXETA, D. G.; MASSENSINI, M. A.; RIBEIRO JUNIOR, J. I.; DELIMA, R. O. Agronomic responses of maize hybrids to row spacing and plant population in the summer and winter seasons in Brazil. **Agronomy Journal**, v. 111, n. 6, p. 3119–3129, 2019.

DIAS, K. O. G.; PIEPHO, H. P.; GUIMARÃES, L. J. M.; GUIMARÃES, P. E. O.; PARENTONI, S. N.; PINTO, M. O.; NODA, R. W.; MAGALHÃES, J. V.; GUIMARÃES, C. T.; GARCIA, A. A. F.; PASTINA, M. M. Novel strategies for genomic prediction of untested single-cross maize hybrids using unbalanced historical data. **Theoretical and Applied Genetics**, v. 133, n. 2, p. 443–455, 2020.

DIAS, K. O. G.; SANTOS, J. P. R.; KRAUSE, M. D.; PIEPHO, H.-P.; GUIMARÃES, L. J. M.; PASTINA, M. M.; GARCIA, A. A. F. Leveraging probability concepts for cultivar recommendation in multi-environment trials. **Theoretical and Applied Genetics**, v. 135, n. 4, p. 1385–1399, 2022.

DIAS, K. O. G.; GEZAN, S. A.; GUIMARÃES, C. T.; NAZARIAN, A.; SILVA, L. C.; PARENTONI, S. N.; GUIMARÃES, P. E. O.; ANONI, C. O.; PÁDUA, J. M. V.; PINTO, M. O.; NODA, R. W.; RIBEIRO, C. A. G.; MAGALHÃES, J. V.; GARCIA, A. A. F.; SOUZA, J. C.; GUIMARÃES, L. J. M.; PASTINA, M. M. 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, n. 1, p. 24–37, 2018.

EEUWIJK, F. A. van; BUSTOS-KORTS, D. V.; MALOSETTI, M. What should students in plant breeding know about the statistical aspects of genotype × environment interactions? **Crop Science**, v. 56, n. 5, p. 2119–2140, 2016.

FINLAY, K. W.; WILKINSON, G. N. The analysis of adaptation in a plant-breeding programme. **Australian Journal of Agricultural Research**, v. 14, n. 6, p. 742, 1963.

GAUCH, H. G. Model selection and validation for yield trials with interaction. **Biometrics**, v. 44, n. 3, p. 705–715, 1988.

GOGEL, B.; SMITH, A.; CULLIS, B. R. Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. **Euphytica**, v. 214, n. 2, p. 44, 2018.

GU, Z.; EILS, R.; SCHLESNER, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. **Bioinformatics (Oxford, England)**, v. 32, n. 18, p. 2847–2849, 2016.

HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a selection model. **Biometrics**, v. 31, n. 2, p. 423, 1975.

ISIK, F.; HOLLAND, J.; MALTECCA, C. Multi Environmental Trials. In: ISIK, F.; HOLLAND, J.; MALTECCA, C. (Eds.). **Genetic Data Analysis for Plant and Animal Breeding**. Cham: Springer International Publishing, 2017. P. 227–262.

KRAUSE, M. D.; DIAS, K. O. G.; SANTOS, J.; OLIVEIRA, A. A.; GUIMARÃES, L. J. M.; PASTINA, M. M.; MARGARIDO, G. R. A.; GARCIA, A. A. F. Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models. **Crop Science**, v. 60, n. 6, p. 3049–3065, 2020.

LISLE, C.; SMITH, A. B.; BIRRELL, C. L.; CULLIS, B. R. Information based diagnostic for genetic variance parameter estimation in multi-environment trials. **Frontiers in Plant Science**, v. 12, 2021.

MALOSETTI, M.; BUSTOS-KORTS, D.; BOER, M. P.; EEUWIJK, F. A. van. Predicting responses in multiple environments: Issues in relation to Genotype × Environment Interactions. **Crop Science**, v. 56, n. 5, p. 2210–2222, 2016.

MALOSETTI, M.; RIBAUT, J.-M.; EEUWIJK, F. A. van. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. **Frontiers in Physiology**, v. 4, p. 44, 2013.

MENGESHA, W.; MENKIR, A.; MESEKA, S.; BOSSEY, B.; AFOLABI, A.; BURGUENO, J.; CROSSA, J. Factor analysis to investigate genotype and genotype × environment interaction effects on pro-vitamin A content and yield in maize synthetics. **Euphytica**, v. 215, n. 11, p. 180, 2019.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. 3rd ed. Boston, MA: CABI, 2014.

NÓIA JÚNIOR, R. d. S.; SENTELHAS, P. C. Soybean-maize succession in Brazil: Impacts of sowing dates on climate variability, yields and economic profitability. **European Journal of Agronomy**, v. 103, p. 140–151, 2019.

OLIVEIRA, I. C. M.; GUILHEN, J. H. S.; RIBEIRO, P. C. O.; GEZAN, S. A.; SCHAFFERT, R. E.; SIMEONE, M. L. F.; DAMASCENO, C. M. B.; CARNEIRO, J. E. S.; CARNEIRO, P. C. S.; PARRELLA, R. A. C.; PASTINA, M. M. Genotype-by-environment interaction and yield stability analysis of biomass sorghum hybrids using factor analytic models and environmental covariates. **Field Crops Research**, v. 257, p. 107929, 2020.

PATERNIANI, E. Maize breeding in the tropics. **Critical Reviews in Plant Sciences**, v. 9, n. 2, p. 125–154, 1990.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545–554, 1971.

PEREIRA, F. C.; RAMALHO, M. A. P.; RESENDE JUNIOR, M. F. R.; PINHO, R. G. V. Mega-environment analysis of maize breeding data from Brazil. **Scientia Agricola**, v. 79, n. 2, e20200314, 2022.

PIEPHO, H.-P. Analyzing genotype-environment data by mixed models with multiplicative terms. **Biometrics**, v. 53, n. 2, p. 761–766, 1997.

PIEPHO, H.-P. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures: **Theoretical and Applied Genetics**, v. 97, n. 1-2, p. 195–201, 1998.

PRASANNA, B. M.; CAIRNS, J. E.; ZAIDI, P. H.; BEYENE, Y.; MAKUMBI, D.; GOWDA, M.; MAGOROKOSHO, C.; ZAMAN-ALLAH, M.; OLSEN, M.; DAS, A.; WORKU, M.; GETHI, J.; VIVEK, B. S.; NAIR, S. K.; RASHID, Z.; VINAYAN, M. T.; ISSA, A. B.; SAN VICENTE, F.; DHLIWAYO, T.; ZHANG, X. Beat the stress: breeding for climate resilience in maize for the tropical rainfed environments. **Theoretical and Applied Genetics**, v. 134, n. 6, p. 1729–1752, 2021.

SANDHU, S.; DHILLON, B. S. Breeding plant type for adaptation to high plant density in tropical maize—A step towards productivity enhancement. **Plant Breeding**, v. 140, n. 4, p. 509–518, 2021.

SILVA, K. J.; GUIMARÃES, C. T.; TINOCO, S. M. S.; BERNARDINO, K. C.; TRINDADE, R. S.; QUEIROZ, V. A. V.; CONCEIÇÃO, R. R. P.; GUILHEN, J. H. S.; OLIVEIRA, N. T.; DAMASCENO, C. M. B.; NODA, R. W.; DIAS, L. A. S.; GUIMARÃES, L. J. M.; MELO, J. O.; PASTINA, M. M. A genome-wide association study investigating fumonisin contamination in a panel of tropical maize elite lines. **Euphytica**, v. 218, n. 9, p. 130, 2022.

SJOBERG, S. M.; CARTER, A. H.; STEBER, C. M.; GARLAND CAMPBELL, K. A. Application of the factor analytic model to assess wheat falling number performance and stability in multienvironment trials. **Crop Science**, v. 61, n. 1, p. 372–382, 2021.

SMITH, A. B.; CULLIS, B. R. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. **Euphytica**, v. 214, n. 8, p. 143, 2018.

SMITH, A. B.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. **The Journal of Agricultural Science**, v. 143, n. 6, p. 449–462, 2005.

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

SMITH, A. B.; GANESALINGAM, A.; KUCHEL, H.; CULLIS, B. R. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. **Theoretical and Applied Genetics**, v. 128, n. 1, p. 55–72, 2015.

TOLHURST, D. J.; MATHEWS, K. L.; SMITH, A. B.; CULLIS, B. R. Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model. **Journal of Animal Breeding and Genetics**, v. 136, n. 4, p. 279–300, 2019.

WALLACE, J. G.; RODGERS-MELNICK, E.; BUCKLER, E. S. On the road to breeding 4.0: Unraveling the good, the bad, and the boring of crop quantitative genomics. **Annual Review of Genetics**, v. 52, n. 1, p. 421–444, 2018.

WICKHAM, H. **ggplot2: Elegant graphics for data analysis**. 2. ed. Cham: Springer, 2016.

YAN, W.; HUNT, L.; SHENG, Q.; SZLAVNICS, Z. Cultivar evaluation and mega-environment investigation based on the GGE Biplot. **Crop Science**, v. 40, n. 3, p. 597–605, 2000.

ZHANG, R.; HAN, D.; HU, X. Analyzing the performance of corn in China using a factor-analytic variance-covariance structure with multiple factors. **Crop Science**, v. 60, n. 1, p. 190–201, 2020.

CHAPTER 4

REALIZED GENETIC GAIN WITH RECIPROCAL RECURRENT SELECTION IN A EUCALYPTUS BREEDING PROGRAM

There is no empirical validation of reciprocal recurrent selection (RRS) in eucalyptus breeding. Our study helps to fill this gap by quantifying the realized response to selection achieved after two cycles of RRS involving *Eucalyptus urophylla* and *E. grandis*. We also investigated the selection effects on the genetic parameters of the breeding populations. We evaluated 25 trials of the first cycle of RRS and 12 trials of the second cycle of RRS. These trials were established in two different regions, separated according to altitude. Fitting linear mixed models enabled the estimation of variance components and the prediction of mean components (general and specific hybridizing abilities). The realized response to selection was calculated as the difference between the mean of the predicted genotypic values of the first and second cycles of RRS. The RRS effectively improved the mean annual increment of wood volume by 28.5% in the high-altitude region and 12.3% in the low-altitude region from the first to the second cycle. The genetic variability also increased as a result of the new genotypes that arose through recombination. These findings provide insights for decision-making and reinforce that eucalyptus breeding can benefit from strategies that capture dominance effects.

1 Introduction

Eucalyptus (*Eucalyptus* L'Hér) breeding programs aim to release clones for forest-based industries while increasing the frequency of favourable alleles in the breeding population (REZENDE et al., 2014). Recurrent selection strategies allow breeders to simultaneously select high-performance candidates (testing/selection step) while keeping the decay of genetic variability between cycles at low levels (recombination step) (KERR et al., 2004; LAMKEY; EDWARDS, 1999). For successfully reaching these objectives, two factors play a major role:

the genetic architecture of the traits under selection and the effective population size required to initiate (founders) and maintain (selected genotypes from one cycle to another) the breeding program in the medium and long terms (ISIK; MCKEAND, 2019).

Recurrent selection in eucalyptus breeding programs can be conducted either within a single population or by employing two populations (ASSIS; RESENDE, 2011). The latter strategy is known as reciprocal recurrent selection (RRS) (COMSTOCK et al., 1949). RRS is particularly suitable for hybrid breeding, where the focus is not only on increasing the additive value of breeding populations but also on exploring the dominance effects via heterosis (COVARRUBIAS-PAZARAN et al., 2023; RESENDE; BARBOSA, 2005). The concept of heterosis was first defined by Shull (1908) as the increased vigour of the heterozygote compared to their homozygous parents. This concept was later outlined as "baseline heterosis" in the population level (LAMKEY; EDWARDS, 1999). However, in eucalyptus breeding, heterosis is achieved by crossing trees - which hardly are inbred - from two different species, deviating from the classical concept outlined by Shull (1908). Instead, eucalyptus breeding RRS explores the mid-parent panmictic heterosis, defined as the difference between the value of the interspecific hybrid population and the mean value of the two panmictic pure-species populations (concept adapted to eucalyptus) (LABROO et al., 2023; LAMKEY; EDWARDS, 1999).

Another source of vigour in interspecific crosses that can be observed in eucalyptus is complementarity (NIKLES; GRIFFING, 1992; POTTS; DUNGEY, 2004). A notable example is the *Eucalyptus grandis* W. Hill × *Eucalyptus urophylla* S. T. Blake hybrids, the so-called "urograndis" hybrids (CAMPINHOS JÚNIOR; IKEMORI, 1978), which corresponds to most of the eucalyptus' clones planted in Brazil and other tropical countries. These hybrids combine the rapid and vigorous growth traits of *E. grandis* with the pest and disease tolerance, high rooting ability, and wood quality of *E. urophylla* [see CABI (2019) for more information about "urograndis"]. In addition to the progress in vegetative propagation techniques, the establishment of "urograndis" clonal plantings has played a pivotal role in the successful history of eucalyptus cultivation in Brazil and other tropical countries (REZENDE et al., 2014). Most forest genetics literature takes complementarity and heterosis as two not-mutually-exclusive sources of hybrid vigour (MADHIBHA et al., 2013; NIKLES; GRIFFING, 1992; RETIEF; STANGER, 2009). In fact, they can act collectively to enhance the overall performance and phenotypic plasticity of hybrids (COVARRUBIAS-PAZARAN et al., 2023).

In the statistical genetics analysis of trials during the testing phase of RRS, full-sib

combining ability models (COMSTOCK et al., 1949) are commonly employed. Sprague and Tatum (1942) defined general combining ability (GCA) as the overall performance of a line in hybrid combinations, and specific combining ability (SCA) as the deviation from the expected performance of a hybrid based on the average performance of the parental lines. In the context of eucalyptus breeding, obtaining inbred lines is a challenging task due to factors such as the large juvenile period, outcrossing nature, and logistic issues (CORTÉS et al., 2020). Besides, as mentioned earlier, the populations involved in the RRS consist of two different species. Therefore, statistical models use data from interspecific hybrids to predict the general hybridizing ability (GHA) and the specific hybridizing ability (SHA) of the pure-species parentals. Nikles and Newton (1991) defined GHA as a quantitative measure of the "additive" values of pure-species parents, and SHA as the contribution of the interaction of interspecific parental alleles to the hybrid performance. The relationship between GCA and GHA is not always linear, implying that these parameters may be controlled by different alleles (IBARRA et al., 2023; RETIEF; STANGER, 2009; WENG et al., 2014). The relationships between GCA and GHA, and SCA and SHA, are vital for defining optimum breeding strategies.

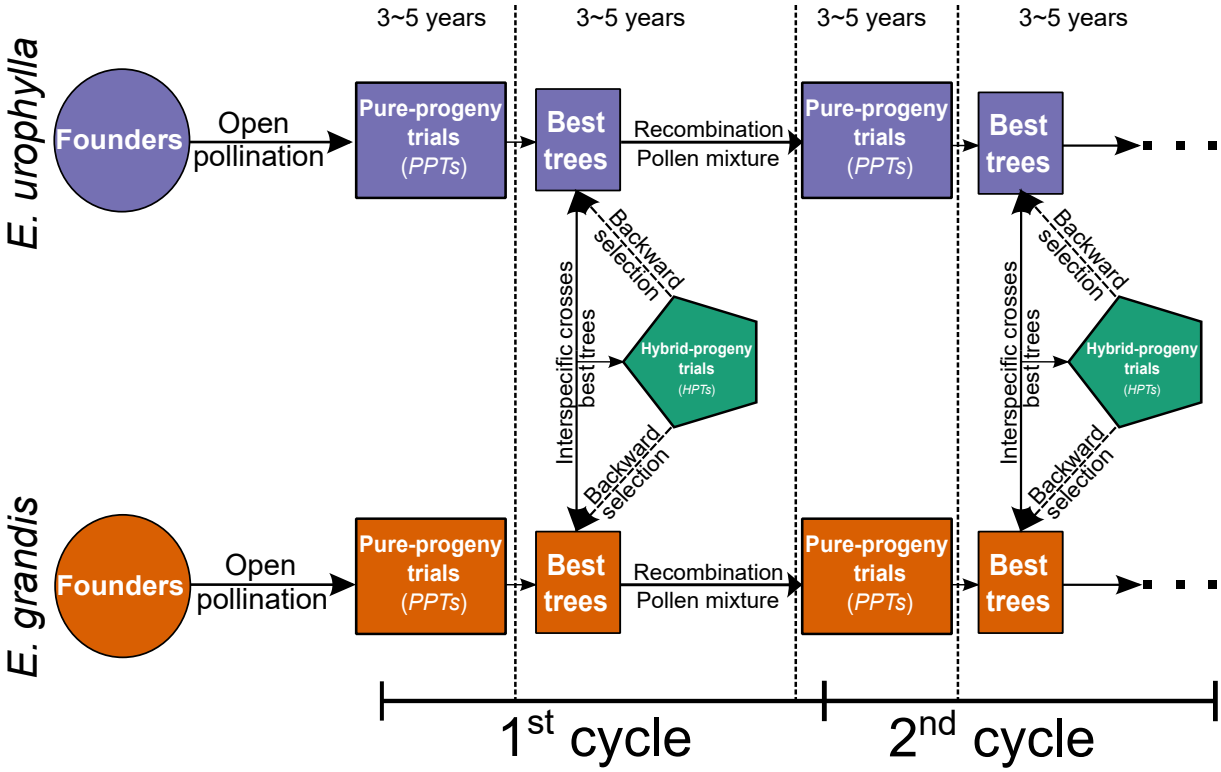
Studies that formally report RRS applied in eucalyptus breeding are scarce, and there is no empirical validation of RRS in eucalyptus breeding. Baudouin et al. (1997) succinctly describe the results of the RRS strategy applied to an "urograndis" breeding program in Congo, and other studies detail aspects of the genetic structure of the population, the mating design used and strategies for data analysis in this breeding program (BARIL et al., 1997; BOUVET; VIGNERON, 1996). However, these studies often lack a comprehensive assessment of the RRS efficacy in achieving the objectives of the breeding program, as they typically focus on single-cycle trials. The present study aims to address this gap by utilizing a historical dataset from an RRS strategy applied to "urograndis" breeding, providing empirical evidence of the efficiency of the RRS. To the best of our knowledge, this research study is the first to present this type of information. The objectives of this study were: i) to quantify the realized response to selection for the mean annual increment of wood volume achieved after two cycles of RRS and ii) to investigate the effects of the RRS on the variance components, genetic parameters, and mean components (GHA and SHA) through two cycles of RRS.

2 Materials and methods

2.1 Reciprocal recurrent selection

The RRS scheme (Figure 1) employed in the breeding program assessed in this study is an adaptation of the original proposal (COMSTOCK et al., 1949) adapted for eucalyptus breeding by Resende and Higa (1990). The scheme involves two populations (*E. grandis* and *E. urophylla*). The referred breeding program is divided into two partially independent sub-programs based on altitude: one for high-altitude regions (higher than 300 meters above sea level), and another for low-altitude regions. Although the RRS scheme was carried out independently for each region, some parents were used in both. Detailed information about the environmental conditions of each region is found in Table 1.

Figure 1: Reciprocal recurrent selection (RRS) scheme evaluated in this study. The legends on the bottom and top indicate the cycle and the time (in years) of each phase in the breeding pipeline, respectively. The program started with the evaluation of pure-species progenies in PPTs, where the best trees were selected for intercrossing. The resulting hybrids were evaluated in HPTs, which were used to select the best pure-species parents based on their GHA. The selected parents of each species were recombined to form the population of the next cycle. Then, the process repeats for each cycle of RRS.



The starting point was the evaluation of the progenies derived from the open pollination

Table 1: Climatic features of each location within the high- and low-altitude regions.

Region	Municipality	Temperature			Relative humidity	Rainfall	Deficit		Radiation		Wind speed
		Minimum	Maximum	Mean			Water	Vapour pressure	Global	Photosynthetically active	
		°C			%	mm	mm	hPa	MJ m^{-2} day^{-1}	mmol m^{-2} s^{-1}	m s^{-1}
High altitude	Sabinópolis	16.3	25.0	20.6	73.6	1209.1	205.0	6.0	15.7	33791.3	3.3
	Virginópolis	16.0	23.6	19.4	80.1	1092.6	187.3	5.1	15.6	32089.0	3.3
	Cocais	15.0	23.1	19.4	75.9	1258.9	121.8	4.1	14.1	30116.0	4.4
	Piracicaba	15.9	23.4	19.4	77.0	1390.1	140.0	5.1	14.1	31140.0	3.6
	Santa Bárbara	16.9	26.1	21.4	70.7	1451.7	184.7	6.1	14.9	33434.0	2.2
Low altitude	Belo Oriente	18.5	27.6	22.7	75.1	1259.7	309.3	6.7	16.1	33074.3	2.4
	Ipaba	19.0	29.3	23.8	73.2	1185.3	331.6	8.3	17.2	18145.0	1.8

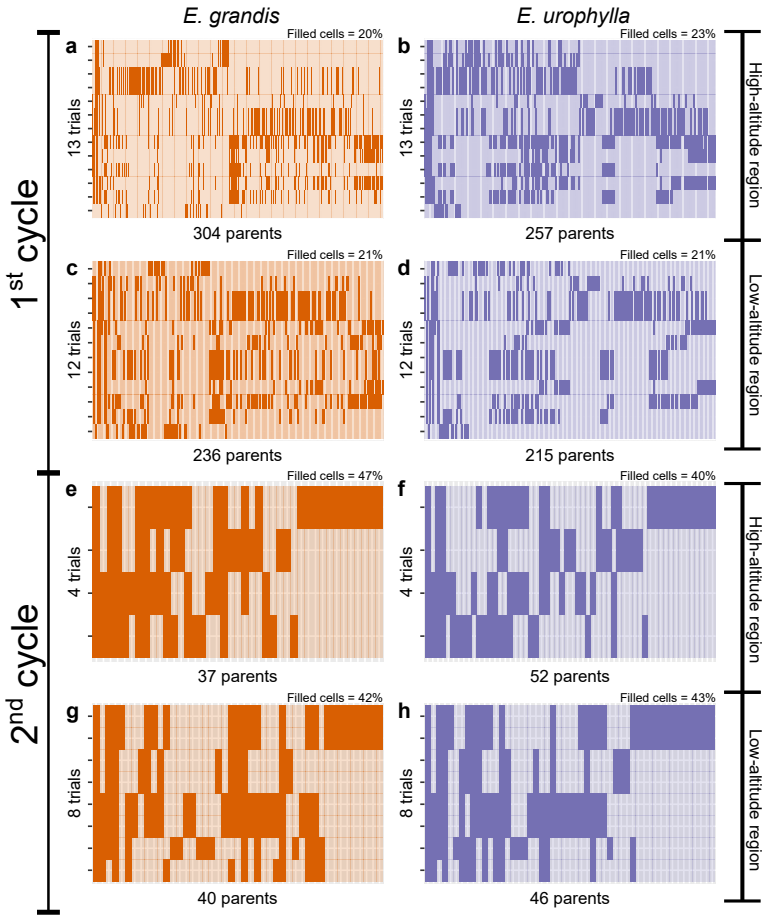
of the founder population within each species. This evaluation took place in pure-progeny trials (PPTs). Following the selection of trees with the highest genetic value, they were cloned, and planted in recombination orchards, where interspecific crosses were carried out. The "urograndis" hybrids resulting from the interspecific crosses were evaluated in hybrid-progeny trials (HPTs). The primary objectives of these HPTs were to *i*) select the best *E. grandis* and *E. urophylla* parents, based on GHA; and *ii*) select high-performing trees to clonal tests. The selected parents of each species were manually pollinated with a mixture of pollen from other selected parents of the same species. This was the recombination phase of the RRS scheme. The progenies resulting from these recombinations composed the second-cycle PPTs. The best trees acted as parents in interspecific crosses, originating the hybrids that composed second-cycle HPTs.

2.2 Phenotypic data

The datasets analyzed here consisted of HPTs from the first and second cycles of RRS in high-altitude and low-altitude regions. Thirteen first-cycle trials were conducted in the high-altitude region. In these trials, we evaluated 562 full-sib families from crosses between 304 *E. grandis* parents and 257 *E. urophylla* parents (Figure 2a and 2b). In the low-altitude region, there were 12 first-cycle trials, from which we assessed the performance of 427 full-sib families, obtained from crosses between 236 *E. grandis* parents and 215 *E. urophylla* parents (Figure 2c and 2d). From the second cycle, we evaluated four trials of the high-altitude region with 155 families from crosses between 37 *E. grandis* parents and 52 *E. urophylla* parents (Figure 2e and 2f) and 8 trials from the low-altitude region with 197 full-sib families from crosses between 40 *E. grandis* parents and 46 *E. urophylla* parents (Figure 2g and 2h).

All trials were laid out in randomized complete blocks design, and each trial had a minimum of two commercial checks for comparison purposes. These trials were conducted in different locations and had varying sowing and data collection dates. The trials differed regarding the layouts of plots (some trials had multiple trees per plot, while others had single-tree plots), and the number of replicates. Trials installed before 2005 typically had multiple trees per plot and fewer replicates, while trials installed after 2005 had single-tree plots with more replicates. Specific details of each trial, including these variations and other relevant information are provided in Table 2. Growth traits (height and diameter at breast height, DBH) were collected when the trees reached an age of 2.8 to 3.5 years. We used these traits to calculate the mean

Figure 2: Heatmaps depicting the distribution of *Eucalyptus grandis* and *Eucalyptus urophylla* parents across trials of the first RRS cycle (a and b, represent the high-altitude region and c and d illustrate the low-altitude region, respectively), and of the second RRS cycle (e and f illustrate the high-altitude region and g and h represent the low-altitude region). Intense-colored cells indicate the presence of the corresponding parent in the *x*-axis in the trial of the *y*-axis, and light-colored cells, the absence. The numbers in the *x*- and *y*-axes represent the number of trials and parents in that particular region and cycle, respectively. The percentage of filled cells, in the top right of each heatmap, represents the degree of connectivity between trials.



annual increment of wood volume (MAI, in m^3 ha^{-1} $year^{-1}$), given as follows (FERREIRA et al., 2023):

$$MAI = \frac{\exp[V1 + V2 \times \ln(DBH) + V3 \times \ln(H)] \times (\frac{10000}{AP}) \times (\frac{SRV}{100})}{A} \quad (1)$$

where $V1$, $V2$ and $V3$ are coefficients determined based on the forest inventory of each test, AP is the area per plant, SRV is the percentage of survival, and A is the age (in years).

Table 2: Hybrid-progeny trials (HPT) analysed in this study, and their respective region, cycle, sowing date, data collection date and location

HPT	Region	Cycle	Sowing	Data collection	Location	# Blocks	# Plants/ plot
HPT088	High altitude	C1	01/12/1996	24/05/1999	Santa Bárbara	3	10
HPT140	High altitude	C1	12/12/2003	31/10/2006	Cocais	8	5
HPT141	High altitude	C1	18/12/2003	04/10/2006	Cocais	8	5
HPT146	High altitude	C1	21/11/2003	19/12/2006	Sabinópolis	8	5
HPT147	High altitude	C1	21/11/2003	21/12/2006	Sabinópolis	8	5
HPT148	High altitude	C1	23/12/2003	30/11/2006	Santa Bárbara	8	5
HPT165	High altitude	C1	18/05/2006	03/07/2009	Santa Bárbara	36	1
HPT168	High altitude	C1	08/06/2006	13/07/2009	Cocais	36	1
HPT170	High altitude	C1	03/07/2006	07/07/2009	Sabinópolis	36	1
HPT192E1E2	High altitude	C1	12/06/2008	27/05/2011	Santa Bárbara	24	1
HPT192E3	High altitude	C1	12/06/2008	27/05/2011	Santa Bárbara	12	1
HPT216	High altitude	C1	20/01/2011	03/04/2014	Sabinópolis	36	1
HPT217	High altitude	C1	24/02/2011	03/04/2014	Santa Bárbara	36	1
HPT269	High altitude	C2	31/05/2016	26/03/2019	Virginópolis	35	1
HPT275	High altitude	C2	04/04/2017	20/04/2020	Sabinópolis	35	1
HPT283	High altitude	C2	14/06/2018	13/07/2021	Sabinópolis	35	1
HPT294	High altitude	C2	27/06/2019	08/06/2022	Sabinópolis	35	1
HPT081	Low altitude	C1	19/11/1996	13/08/1999	Belo Oriente	3	10
HPT142A1	Low altitude	C1	09/12/2003	09/10/2006	Belo Oriente	8	5
HPT142A2	Low altitude	C1	09/12/2003	09/10/2006	Belo Oriente	8	5
HPT142A3	Low altitude	C1	09/12/2003	09/10/2006	Belo Oriente	8	5
HPT143P1	Low altitude	C1	22/12/2003	27/10/2006	Belo Oriente	8	5
HPT143P2	Low altitude	C1	22/12/2003	27/10/2006	Belo Oriente	8	5
HPT144	Low altitude	C1	23/12/2003	30/10/2006	Belo Oriente	8	5
HPT145	Low altitude	C1	23/12/2003	20/11/2006	Belo Oriente	6	5
HPT166	Low altitude	C1	22/06/2006	30/06/2009	Belo Oriente	36	1
HPT167	Low altitude	C1	26/06/2006	17/08/2009	Belo Oriente	24	1
HPT169	Low altitude	C1	13/06/2006	15/06/2009	Belo Oriente	36	1
HPT215	Low altitude	C1	23/12/2010	04/12/2013	Belo Oriente	50	1
HPT267	Low altitude	C2	25/05/2016	23/01/2019	Belo Oriente	40	1
HPT268	Low altitude	C2	07/06/2016	12/03/2019	Belo Oriente	40	1
HPT276BA	Low altitude	C2	24/04/2017	05/03/2020	Belo Oriente	20	1
HPT276EN	Low altitude	C2	24/04/2017	05/03/2020	Belo Oriente	20	1
HPT282BA	Low altitude	C2	06/06/2018	21/05/2021	Ipaba	20	1
HPT282EN	Low altitude	C2	06/06/2018	21/05/2021	Ipaba	20	1
HPT293BA	Low altitude	C2	04/10/2019	27/09/2022	Belo Oriente	20	1
HPT293EN	Low altitude	C2	04/10/2019	27/09/2022	Belo Oriente	20	1

2.3 Statistical analyses

We used the residual maximum likelihood (PATTERSON; THOMPSON, 1971) to estimate the variance components required for the empirical best linear unbiased predictions (HENDERSON, 1975) of the GHA and SHA. We used the `ASReml-R` package (THE VSNI TEAM, 2023), version 4.2, for this purpose.

Individual analyses

We analysed each trial individually using the following linear mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{r} + \mathbf{Z}_2\mathbf{g} + \mathbf{Z}_3\mathbf{u} + \mathbf{Z}_4\mathbf{s} + \mathbf{e} \tag{2}$$

where $\mathbf{y}$ is the vector of the hybrids' phenotypic records, $\boldsymbol{\beta}$ is the vector of fixed effects (intercept and checks), $\mathbf{r}$ is the vector of block effects, assumed to be random [$\mathbf{r} \sim N(\mathbf{0}, \sigma_r^2 \mathbf{I}_B)$, where σ_r^2 is the variance of the block effects, and $\mathbf{I}$ is an identity matrix whose dimension is B, the number of blocks], $\mathbf{g}$ is the vector of random GHA effects of *E. grandis* parents [$\mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 \mathbf{I}_A)$, where σ_g^2 is the variance of the GHA effects of *E. grandis*, and $\mathbf{I}$ is an identity matrix whose dimension is A, the number of *E. grandis* parents], $\mathbf{u}$ is the vector of random GHA effects of *E. urophylla* parents [$\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_O)$, where σ_u^2 is the variance of the GHA effects of *E. urophylla*, and $\mathbf{I}$ is an identity matrix whose dimension is O, the number of *E. urophylla* parents], $\mathbf{s}$ is the vector of random SHA effects [$\mathbf{s} \sim N(\mathbf{0}, \sigma_s^2 \mathbf{I}_F)$, where σ_s^2 is the variance of the SHA effects, and $\mathbf{I}$ is an identity matrix whose dimension is F, the number of realized full-sib families], and $\mathbf{e}$ is the random residual effects [$\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_N)$, where σ_e^2 is the residual variance and $\mathbf{I}$ is an identity matrix whose dimension is N, the number of phenotypic records]. The capital letters accompanying the described vectors are the incidence matrices of the corresponding effects.

The analyses per trial had a double purpose: *i*) to filter out trials without significant genetic effects (GHAs and SHA), and *ii*) to provide an overview of the genetic effects relative to the total phenotypic variance. For the first objective, we conducted a Likelihood Ratio Test and excluded trials that $\sigma_g^2 = \sigma_u^2 = \sigma_s^2 = 0$ to further analyses. After filtering, we investigated how much of the total variance was due to the genetic effects (GHA effects and the SHA effects) for each component of the total genetic variance:

$$H^2 = \frac{\sigma_g^2 + \sigma_u^2 + \sigma_s^2}{\sigma_p^2}, \quad h_g^2 = \frac{\sigma_g^2}{\sigma_p^2}, \quad h_u^2 = \frac{\sigma_u^2}{\sigma_p^2}, \text{ and } \quad h_s^2 = \frac{\sigma_s^2}{\sigma_p^2} \tag{3}$$

where σ_p^2 is the phenotypic variance, given by $\sigma_p^2 = \sigma_g^2 + \sigma_u^2 + \sigma_s^2 + \sigma_r^2 + \sigma_e^2$. These ratios are proportional to the broad-sense heritability (H^2), individual narrow-sense heritabilities of the effects of *E. grandis* (h_g^2) and *E. urophylla* (h_u^2), and the ratio of dominance to phenotypic variance (h_s^2).

Multi-environment and cycle analysis

We analysed trials from both cycles in each region using the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{r} + \mathbf{Z}_2\mathbf{g} + \mathbf{Z}_3\mathbf{u} + \mathbf{Z}_4\mathbf{s} + \mathbf{Z}_5\mathbf{gt} + \mathbf{Z}_6\mathbf{ut} + \mathbf{Z}_7\mathbf{st} + \mathbf{e} \tag{4}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects (intercept, cycles, checks and trials), **g**, **u** and **s** are the vector of *E. grandis* GHA, *E. urophylla* GHA and SHA main effects; and **gt**, **ut**, **st** are the vectors of the genotype-by-environment interactions (GHA-by-environment and SHA-by-environment). These effects follow the same distribution described in Equation 2, i.e., a multivariate Gaussian with mean centered in zero and homogeneous variance. Nevertheless, in the multi-environment multi-cycle model, we considered cycle-specific variances for both the main effects and interaction effects. The two other (non-genetic) random effects, **r** and **e**, respectively, represent the block and residual effects. We considered that these effects follow a multivariate Gaussian, with the mean centered in zero and heterogeneous (trial-wise) variances. It is important to bear in mind that other effects such as differences in tree ages at the time of data collection or environmental variations between trials due to geographical distance or temporal dissimilarity (trials conducted in different years) are confounded within the effects of cycles and trials, considered in $\boldsymbol{\beta}$.

In the joint analysis, we also estimated h_g^2, h_u^2, h_s^2 and H^2 (Equation 3). The sole difference is that for the joint model, $\sigma^2_{p_m} = \sigma^2_g + \sigma^2_u + \sigma^2_s + \sigma^2_{tg} + \sigma^2_{tu} + \sigma^2_{ts} + \sigma^2_{r_m} + \sigma^2_{e_m}$, i.e., there are variances for the genotype-by-environment interactions, and there are particular block and error variances for each trial. Consequently, the genetic parameters were computed per trial (H^2_m, $h^2_{g_m}$, $h^2_{u_m}$, and $h^2_{s_m}$).

Further investigations on the dominance effects

The success of the RRS depends on the existence of heterosis acting at the gene loci that control the trait. Dominance is one of the factors that control heterosis. Therefore, a useful metric to infer the success of the RRS is the average degree of dominance (ADD), or other equivalent or related statistics. In quantitative genetic theory, ADD is given by the ratio of the heterozygote genotypic value (d) to the homozygote genotypic value (a) (COMSTOCK; ROBINSON, 1952; FALCONER, 1989). In populations with unknown genetic structures, one can only access these values using molecular data. Using phenotypic data from field trials, we can approximately compute the ADD leveraging the estimated dominance variance (σ^2_d) and additive variance (σ^2_a). Recall from quantitative genetics that $\sigma^2_d = (2pqd)^2$ and $\sigma^2_a = 2pq[a + d(q-p)]^2$, where p and q are allele frequencies. Considering $q = 1 - p$, $ADD = \frac{[2p(1-p)d]^2}{[2p(1-p)][a+d(1-2p)]^2}$. This formula represents a generalization to calculate the ADD considering any mean allelic frequency p in the population. In variance component terms, this

formula is given by:

$$ADD = \frac{\sqrt{\frac{\sigma_d^2/\sigma_a^2}{2p(1-p)}}}{1 - \sqrt{\frac{\sigma_d^2/\sigma_a^2}{2p(1-p)}} \times (1-2p)} \quad (5)$$

Be aware that the portion of σ_a^2 that is due to d - which is related to the dominance effects - is only annulled when $p = 0.5$. Indeed, when assuming $p = 0.5$, Equation 5 becomes the formula that is usually employed to compute the *ADD* from phenotypic data (COMSTOCK; ROBINSON, 1952; ROBINSON et al., 1955):

$$ADD = \sqrt{\frac{2\sigma_d^2}{\sigma_a^2}} \quad (6)$$

In this study, assuming $p = 0.5$ is reasonable given the genetic proximity of the breeding population from the wild founders (only two generations apart). We used genetic population theory to obtain the additive variance as $\sigma_a^2 = 2(\sigma_g^2 + \sigma_u^2)$, and the dominance variance as $\sigma_d^2 = 4 \times \sigma_s^2$. After computing the *ADD*, we classified the dominance status according to the scales proposed by Comstock and Robinson (1952).

We also investigated the relative contribution of dominance effects in determining the performance of eucalyptus hybrids. To assess this, we regressed the ranking considering the genotypic values, which take into account both the GHA of the parents and the SHA of the crosses, with the ranking considering the genetic values, which consider only the GHA.

Realized response to selection

To visualize the response to selection, we plotted the genotypic values of the full-sib families from each RRS cycle as a density plot. To quantify the response to selection, we calculated the difference between the mean genotypic value of the first cycle and the mean genotypic value of the second cycle (LYNCH; WALSH, 1998).

We used the software `R` (R CORE TEAM, 2023), version 4.3.1, to perform all the analyses. The plots were built using the `ggplot2` package (WICKHAM, 2016), with add-ins of the package `ggpubr` (KASSAMBARA, 2023).

3 Results

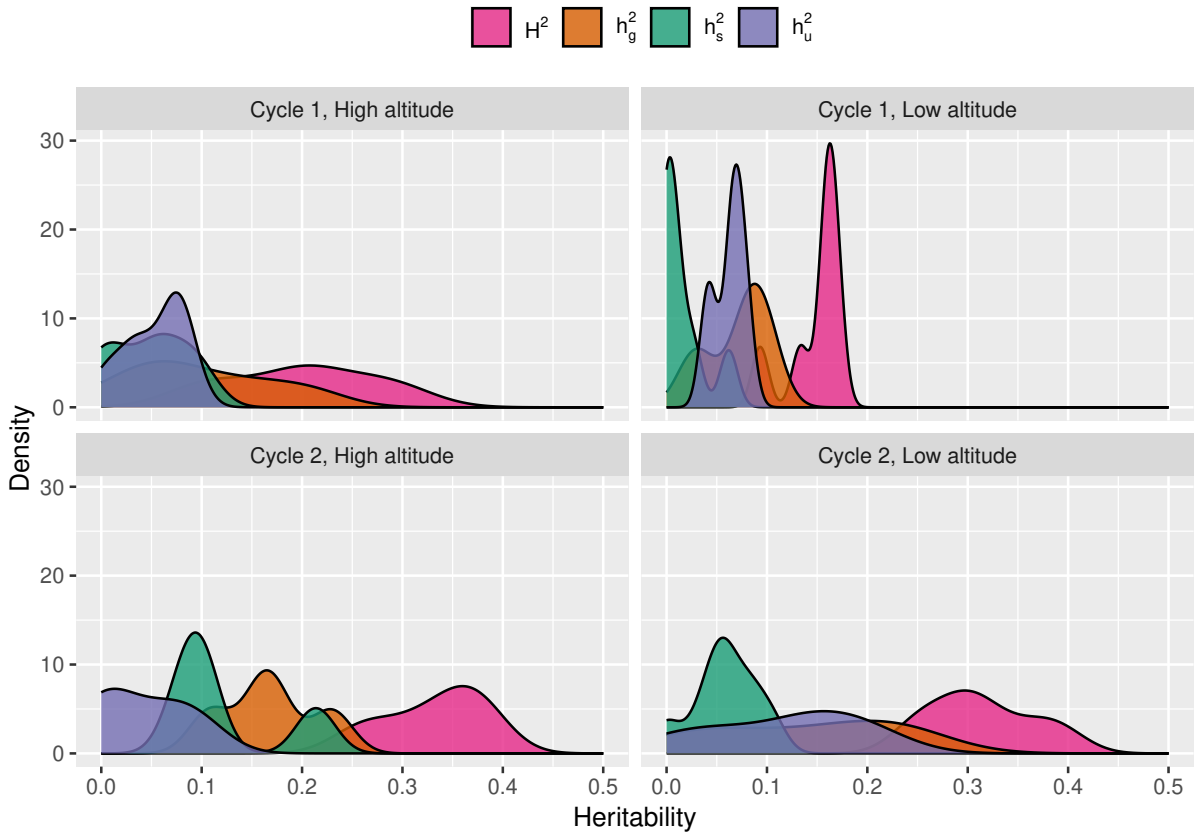
3.1 Individual analyses

The genetic parameters differed among trials (Figure 3 and Table 3). We discarded eight trials due to the non-significance of any genetic effect. Henceforth, the results presented refer to the remaining trials. In the first cycle, there was a greater variation observed among *E. grandis* parents than *E. urophylla* in both the high-altitude and low-altitude regions. Additionally, the dominance effects played a more important role in the phenotypic expression in the high-altitude region than in the low-altitude region (see the *ADD* values of each trial in Table 3). The broad-sense heritabilities had values around 0.21 in the high-altitude region and 0.15 in the low-altitude region. Note how the broad-sense heritability values were less variable across trials in the low-altitude region. In the second cycle, the scenario slightly changed. The dominance effects seem to play an important role in the phenotypic expression, and the broad-sense heritabilities increased compared to the first-cycle trials: from 0.21 to 0.33 in the high-altitude region and from 0.15 to 0.31 in the low-altitude region. In the high-altitude region, genetic variation between *E. grandis* parents were always higher than of *E. urophylla* parents. This pattern was not observed in the low-altitude region.

3.2 Multi-environment and cycle analysis

The joint analysis, which considered genotype-by-environment interactions, provided a less biased estimation of genetic parameters (Figure 4, Table 4 and Table 5). Overall, the RRS did not reduce the genetic variability. Instead, the available variability increased from the first to the second cycle in both high-altitude and low-altitude regions. As Figure 3 indicated, the broad-sense heritability was higher in the second-cycle hybrid population. The other parameters followed the same pattern, i.e., the parents involved in producing the second-cycle hybrids contributed a wide diversity of alleles that influenced the phenotypic expression of the studied trait. The importance of dominance effects also became more evident in the second cycle. For instance, in the low-altitude region, the dominance effects transitioned from being almost negligible in the first cycle to becoming preponderant in the second cycle (Table 5).

Figure 3: Distribution of the estimates of the broad-sense heritabilities (H^2) and the general hybridizing ability (GHA) variance of *Eucalyptus grandis* to total variance ratio (h_g^2), GHA variance of *E. urophylla* to total variance ratio (h_u^2), and specific hybridizing ability variance to total variance ratio (h_s^2) in individual trials of the first and second reciprocal recurrent selection cycles in both high- and low-altitude regions for the mean annual increment of wood volume.



3.3 Realized response to selection

The mean annual increment of wood volume was improved from the first to the second cycle in both the high-altitude and low-altitude regions (Figure 5). In the high-altitude region, the realized response to selection was 11.04 m^3 ha^{-1} $year^{-1}$, indicating a substantial improvement in the trait over the cycles. This gain was more pronounced compared to the low-altitude region, where the realized response to selection was 4.36 m^3 ha^{-1} $year^{-1}$. The genotypic values in the second cycle were more widely distributed compared to the first cycle in the low-altitude region. This is probably related to the increase in the dominance effects depicted in Figure 4.

Table 3: Values of the GHA variance of *E. grandis* to total variance ratio (h_g^2), GHA variance of *E. urophylla* to total variance ratio (h_u^2), SHA variance to total variance ratio (h_s^2), average degree of dominance (*ADD*) and dominance classification of each Hybrid-progeny trial (HPT) obtained from individual analyses.

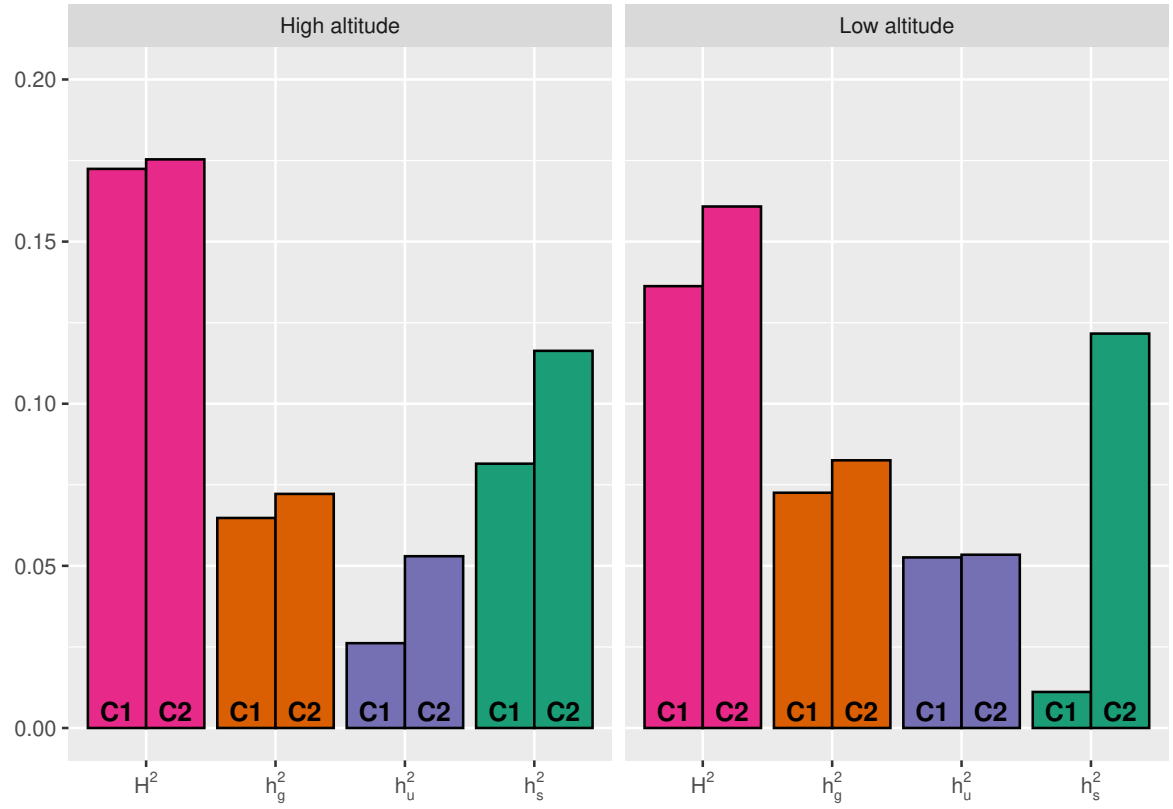
HPT	Cycle	Region	h_g^2	h_u^2	h_s^2	H^2	*ADD*	Classification
HPT088*	C1	High altitude	-	-	-	-	-	-
HPT140	C1	High altitude	0.106	0.086	0.017	0.209	0.589	Partial dominance
HPT141	C1	High altitude	0.061	0.012	0.045	0.118	1.585	Overdominance
HPT146	C1	High altitude	0.034	0.075	0.087	0.196	1.785	Overdominance
HPT147	C1	High altitude	0.038	0.042	0.094	0.175	2.163	Overdominance
HPT148	C1	High altitude	0.079	0.071	0.050	0.200	1.161	Overdominance
HPT165*	C1	High altitude	-	-	-	-	-	-
HPT168	C1	High altitude	0.211	0.082	0.005	0.298	0.255	Partial dominance
HPT170*	C1	High altitude	-	-	-	-	-	-
HPT192E1E2	C1	High altitude	0.137	0.033	0.060	0.230	1.183	Overdominance
HPT192E3	C1	High altitude	0.154	0.076	0.063	0.293	1.052	Overdominance
HPT216	C1	High altitude	0.069	0.037	0.000	0.105	0.002	No dominance
HPT217	C1	High altitude	0.198	0.070	0.000	0.267	0.002	No dominance
HPT269	C2	High altitude	0.172	0.096	0.103	0.371	1.240	Overdominance
HPT275	C2	High altitude	0.109	0.000	0.214	0.322	2.806	Overdominance
HPT283	C2	High altitude	0.230	0.060	0.080	0.370	1.051	Overdominance
HPT294	C2	High altitude	0.159	0.008	0.096	0.263	1.518	Overdominance
HPT081	C1	Low altitude	0.025	0.041	0.027	0.093	1.283	Overdominance
HPT142A1	C1	Low altitude	0.072	0.074	0.012	0.158	0.567	Partial dominance
HPT142A2	C1	Low altitude	0.098	0.065	0.000	0.163	0.004	No dominance
HPT143A3*	C1	Low altitude	-	-	-	-	-	-
HPT143P1	C1	Low altitude	0.033	0.064	0.062	0.160	1.592	Overdominance
HPT143P2	C1	Low altitude	0.084	0.081	0.000	0.165	0.003	No dominance
HPT144	C1	Low altitude	0.101	0.070	0.000	0.171	0.003	No dominance
HPT145	C1	Low altitude	0.083	0.043	0.007	0.133	0.468	Partial dominance
HPT166*	C1	Low altitude	-	-	-	-	-	-
HPT167*	C1	Low altitude	-	-	-	-	-	-
HPT169*	C1	Low altitude	-	-	-	-	-	-
HPT215*	C1	Low altitude	-	-	-	-	-	-
HPT267	C2	Low altitude	0.164	0.038	0.055	0.256	1.041	Overdominance
HPT268	C2	Low altitude	0.217	0.072	0.105	0.394	1.206	Overdominance
HPT276BA	C2	Low altitude	0.026	0.165	0.059	0.250	1.107	Overdominance
HPT276EN	C2	Low altitude	0.020	0.186	0.086	0.292	1.292	Overdominance
HPT282BA	C2	Low altitude	0.240	0.133	0.000	0.373	0.003	No dominance
HPT282EN	C2	Low altitude	0.229	0.000	0.076	0.305	1.150	Overdominance
HPT293BA	C2	Low altitude	0.135	0.133	0.034	0.302	0.713	Partial dominance
HPT293EN	C2	Low altitude	0.076	0.209	0.051	0.336	0.843	Partial dominance

*Trials where all genetic variance components estimates were non-significant ($p > 0.05$)

3.4 Dominance effects

The dominance effects impact the ranking of hybrids (Figure 6). In the high-altitude region, considering only the additive genetic values (only breeding values) would lead to a ranking almost 40% different from considering the genotypic values (additive + dominance effects). In the low-altitude region, the dominance effects are even more relevant: would only the genetic values be considered, the ranking would be 57% different.

Figure 4: Broad-sense heritability (H^2), general hybridizing ability variance of *Eucalyptus grandis* to total variance ratio (h_g^2), GHA variance of *E. urophylla* to total variance ratio (h_u^2), and specific hybridizing variance to total variance ratio (h_s^2) across trials in the same reciprocal recurrent selection cycle, in the high- and low-altitude regions for the mean annual increment of wood volume.



4 Discussion

As previously theorized (COMSTOCK et al., 1949; RESENDE; HIGA, 1990), RRS has proven to be a successful strategy to explore the heterosis to increase wood volume in eucalyptus. We empirically validated this fact by quantifying the realized response to selection after two cycles of RRS. In both high- and low-altitude regions, the average genotypic value of the "urograndis" hybrids increased by 28.5% and 12.3% between the first and second cycles, respectively. This happened without harming the genetic variability, as it also exhibited an increase from one cycle to the next. We also showed the importance of the dominance effects for the selection of the real top performers. Studies of this kind are scarce in forest breeding, probably due to the length of each breeding cycle. Such empirical validation may aid the decision-making in eucalyptus breeding.

Table 4: Values of the GHA variance of *E. grandis* to total variance ratio (h_g^2), GHA variance of *E. urophylla* to total variance ratio (h_u^2), SHA variance to total variance ratio (h_s^2), average degree of dominance (*ADD*) and dominance classification of each Hybrid-progeny trial (HPT) obtained from the joint analysis. Only trials that had at least one significant genetic variance component estimate were considered for the joint analysis.

HPT	Cycle	Region	h_g^2	h_u^2	h_s^2	H^2
HPT140	C1	High altitude	0.102	0.041	0.128	0.272
HPT141	C1	High altitude	0.103	0.041	0.129	0.273
HPT146	C1	High altitude	0.053	0.021	0.066	0.140
HPT147	C1	High altitude	0.050	0.020	0.063	0.133
HPT148	C1	High altitude	0.043	0.017	0.054	0.115
HPT168	C1	High altitude	0.080	0.032	0.100	0.212
HPT192E1E2	C1	High altitude	0.056	0.022	0.070	0.148
HPT192E3	C1	High altitude	0.072	0.029	0.090	0.191
HPT216	C1	High altitude	0.050	0.020	0.062	0.132
HPT217	C1	High altitude	0.047	0.019	0.060	0.126
HPT269	C2	High altitude	0.072	0.053	0.115	0.174
HPT275	C2	High altitude	0.116	0.085	0.186	0.281
HPT283	C2	High altitude	0.058	0.042	0.093	0.140
HPT294	C2	High altitude	0.044	0.032	0.071	0.106
HPT081	C1	Low altitude	0.154	0.111	0.024	0.289
HPT142A1	C1	Low altitude	0.068	0.049	0.010	0.127
HPT142A2	C1	Low altitude	0.089	0.065	0.014	0.167
HPT143P1	C1	Low altitude	0.048	0.035	0.007	0.090
HPT143P2	C1	Low altitude	0.058	0.042	0.009	0.108
HPT144	C1	Low altitude	0.039	0.028	0.006	0.073
HPT145	C1	Low altitude	0.053	0.039	0.008	0.100
HPT267	C2	Low altitude	0.070	0.045	0.103	0.137
HPT268	C2	Low altitude	0.092	0.060	0.136	0.180
HPT276BA	C2	Low altitude	0.080	0.052	0.118	0.156
HPT276EN	C2	Low altitude	0.142	0.092	0.209	0.277
HPT282BA	C2	Low altitude	0.085	0.055	0.125	0.165
HPT282EN	C2	Low altitude	0.080	0.052	0.117	0.155
HPT293BA	C2	Low altitude	0.049	0.032	0.073	0.096

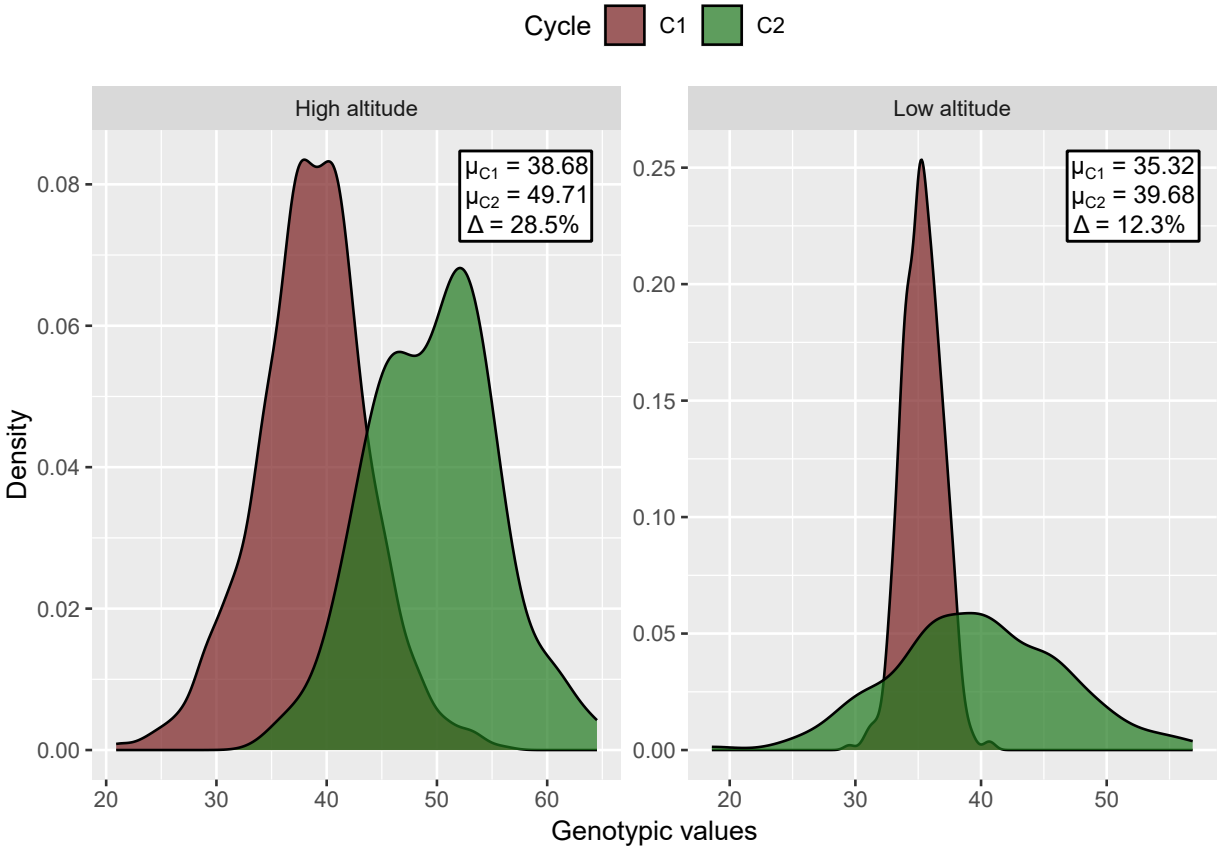
Table 5: Variance components estimates of the genetic effects, and average degree of dominance (*ADD*) obtained from the joint analysis of the Hybrid-progeny trials (HPT): σ_g^2 is the variance of the general hybridizing ability (GHA) of *E. grandis* parents, σ_u^2 is the variance of the GHA of *E. urophylla* parents, σ_s^2 is the variance of the specific hybridizing ability (SHA) effects, σ_{tg}^2 is the variance of the interaction of trials and the GHA of *E. grandis* parents, σ_{tu}^2 is the variance of the interaction of trials and the GHA of *E. urophylla* parents, and σ_{ts}^2 is the variance of the interaction of trials and the SHA

Region	Cycle	σ_g^2	σ_u^2	σ_s^2	σ_{tg}^2	σ_{tu}^2	σ_{ts}^2	*ADD*	Classification
High altitude	C1	37.999	15.346	47.820	5.113	12.152	8.166	1.894	Overdominance
High altitude	C2	41.645	30.561	67.100	77.358	0.000	0.000	2.243	Overdominance
Low altitude	C1	54.603	39.592	8.363	0.931	8.285	6.758	0.596	Partial dominance
Low altitude	C2	52.645	34.074	77.564	27.551	11.612	0.000	1.815	Overdominance

4.1 Genetic parameters

Forest breeding populations tend to have a close genetic relationship with their wild counterparts (ISIK; MCKEAND, 2019; WU et al., 2016). The populations examined in this study are no exception. Since few selections were performed from the gene pool provided by the founders, alleles that enhance commercial performance, such as rapid growth or superior

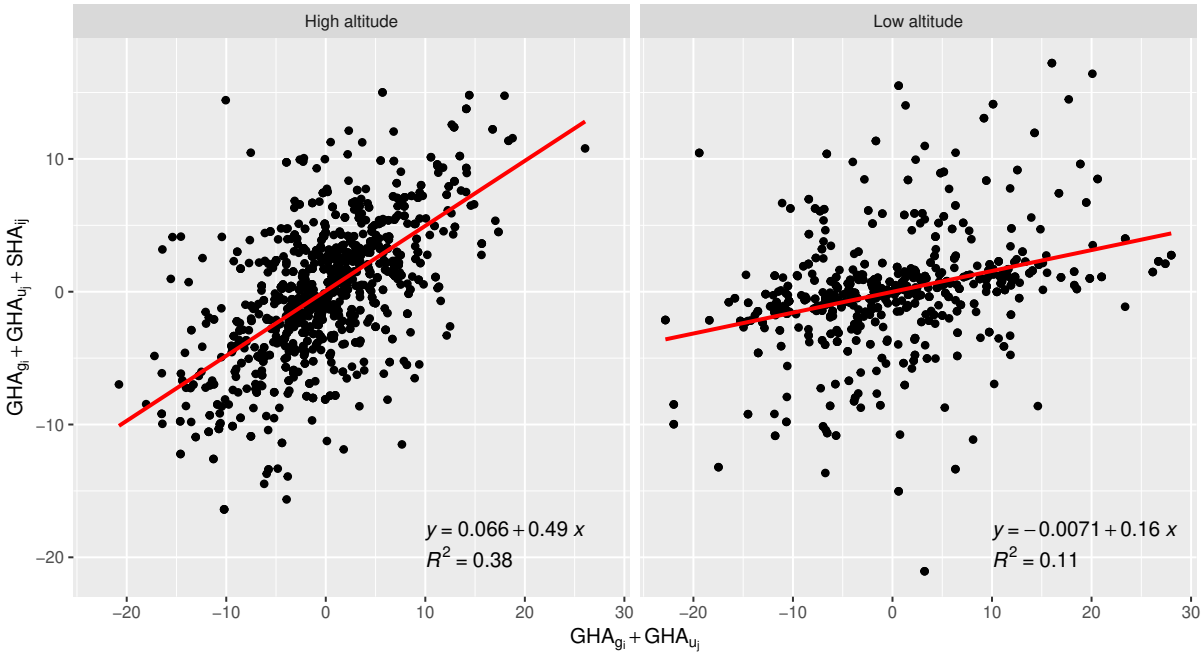
Figure 5: Distribution of the genotypic values of the mean annual increment of wood volume of the "urograndis" hybrids in each cycle in the high-altitude (a) and in the low-altitude regions (b). The caption in the upper right of each plot has the overall mean of the genotypic values in each cycle and the difference between cycles (Δ), in percentage.



wood quality, may have not yet become fixed within the population. This implies that the breeding process is occurring simultaneously with the domestication process (ISIK et al., 2015; JONES et al., 2006).

The parents chosen from the first to the second cycle are heterozygous individuals with presumed high hybridizing ability for improving growth traits. This selection process, combined with the high genetic diversity associated with the close relation to the wild founders and the heterozygosity of the parents, may have led to an intense production of new allelic combinations during the crossings and recombination step of the RRS. This should explain the observed increase in genetic variances in the second cycle. Similar short-term increases in genetic variance have been observed in other eucalyptus breeding programs (JONES et al., 2006; LEFÈVRE, 2004; LU et al., 2018). While it is uncertain how many RRS cycles can be sustained without compromising genetic diversity, the current structure of the breeding program, with an average time interval of 10 years between cycles, suggests that the decay of genetic

Figure 6: Relation between the additive genetic values (y-axis), built using only the general hybridizing ability (GHA) of the parents, and the genotypic values (x-axis), estimated using both the GHA and the specific hybridizing ability (SHA) of the parental combination in the high-altitude region and the low-altitude region. The caption in the lower right of each plot has the correlation between the values in the y-axis and the x-axis. GHA_{g_i} is the GHA of *Eucalyptus grandis* parents, GHA_{u_j} is the GHA of *E. urophylla* parents, and SHA_{ij} is the specific combining ability of the hybrid. The red line depicts the linear regression of additive genetic values over genotypic values. The regression equation and the coefficient of determination ($\hat{R}$) are on the lower left of each plot.



variability should not be a significant concern. However, changes in selection intensity, effective population size, and usage of new technologies (for instance, genomic selection/prediction) can influence these dynamics (HINZE et al., 2005; LABROO et al., 2023). Therefore, periodic re-evaluation of the available diversity is necessary to ensure the continued preservation and utilization of genetic resources in future cycles (RESENDE, 2002).

4.2 Dominance is important for eucalyptus hybrid breeding

The ultimate objective of eucalyptus breeding programs is to develop elite clonal cultivars by selecting the best trees for growth and wood quality traits. The results of this study suggest that the hybridizing ability of the pure-species parents is not a reliable indicator of the performance of interspecific hybrids. Other studies evaluating different eucalyptus breeding populations have reached similar conclusions (BERG et al., 2015; BISON et al., 2006; RETIEF; STANGER,

2009). Thus, to effectively address the outlined goal, the breeding program needs to capture the effects of both additive (GHA) and dominance (SHA) genetic effects.

In this study, both heterosis and complementarity appear to be important factors driving the phenotypic expression of the studied trait. Heterosis can be attributed to differences in allelic frequency, dominance effects and additive-additive epistatic effects (HILL, 1982; WILLHAM; POLLAK, 1985). The dominance effects played a major role in phenotypic expression, contributing to heterosis. Complementarity, on the other hand, is often considered separately from heterosis in tree breeding and is attributed to additive gene actions (POTTS; DUNGEY, 2004; RETIEF; STANGER, 2009). Nevertheless, we believe that complementarity and heterosis are closely related in additive-dominance traits. Genetically, complementarity can be achieved by summing the effects of favourable alleles, or when dominant alleles mask the expression of deleterious alleles in the genome. Multiple dominant alleles masking deleterious alleles and contributing to the increase (or decrease) of the trait expression characterize directional dominance, one of the causes of heterosis. In the eucalyptus RRS, two populations of different species are bred simultaneously to complement each other, allowing for the exploration of heterosis in interspecific hybrids. Furthermore, the RRS also potentializes heterosis by increasing the genetic divergence between populations, leading to the increase in genetic variance we observed in this study.

The results also provide a practical implication, highlighting the need for proper consideration of the dominance effects in the breeding pipeline. As plant breeding progresses and integrates predictive approaches using various "omics" frameworks, it becomes crucial for statistical-genetic models to account for the effects caused by dominance accurately. This action can significantly enhance the efficiency and success of the breeding program in selecting top-performing trees and achieving the desired genetic progress (COSTA E SILVA et al., 2004; DENIS; BOUVET, 2013; RESENDE et al., 2017).

4.3 Realized response to selection

Our results suggest that the selection of parents in the first cycle led to an increased performance in the second-cycle hybrids. In fact, considering a cycle interval of 10 years, the realized response to selection per year was approximately 1.10 m^3 ha^{-1} $year^{-1}$ and 0.42 m^3 ha^{-1} $year^{-1}$ in the high-altitude and low-altitude regions, respectively. A factor that may have influenced the genetic gains is that the genitors of the first-cycle hybrids were selected solely

for growth traits, whereas the parents of the second-cycle hybrids were selected for both growth and wood quality traits. Hence, the genetic gain exists not only for the mean annual increment of wood volume but for other traits related to wood quality, such as wood density and cellulose yield. It is important to acknowledge that other methods can be used to quantify the realized response to selection, and each method has its benefits and limitations [see Rutkoski (2019) and Krause et al. (2023) for more details about these methods].

In other perennial plants, the RRS was also successful. In oil palm (*Elaeis guineensis* Jacq.), the choice of SRR as the ideal breeding scheme was defined in the 1950s, shortly after the original proposition by Comstock et al. (1949). Meunier and Gascon (1972) describe the scheme used. The first selection cycle was completed in Côte d'Ivoire and Cameroon and the second cycle was conducted in Côte d'Ivoire and Indonesia. The third cycle is under development. The genetic gain obtained for productivity was 18% in the first cycle and, approximately, the same rate in the second cycle (COCHARD; NOIRET, 1993; GASCON et al., 1988). In Robusta coffee (*Coffea canephora* Pierre), RRS was initiated in 1984 in Côte d'Ivoire (LEROY et al., 1993). The first cycle provided a 60% increase in productivity (LEROY et al., 1997). The second cycle is in progress.

4.4 Suggestions to improve eucalyptus reciprocal recurrent selection

In annual species, it is well-established that the main challenge of the RRS strategy is time (BERNARDO, 2020; COVARRUBIAS-PAZARAN et al., 2023; REMBE et al., 2021). This issue is further exacerbated in perennial species, such as eucalyptus. Kerr et al. (2004) simulated four tree breeding strategies: RRS, RRS-SF (RRS with forward selection), pure species selection and synthetic selection. The last two are intraspecific recurrent selection strategies. The authors concluded that RRS provides the lowest annual gains, and is a suitable strategy only when dominance effects are pronounced. In fact, in the presence of dominance, the most suitable RRS is the RRS-SF, which eliminates the backward selection step. Another option to shorten the cycle interval is to employ genomic selection/prediction models (GRATTAPAGLIA, 2022; MELCHINGER et al., 2023; SIMIQUELI et al., 2023). Recent studies have proposed several strategies to leverage molecular data for accelerating the breeding pipeline (COVARRUBIAS-PAZARAN et al., 2023; LABROO et al., 2023). Although not directly proposed for eucalyptus, these strategies can be adapted to fit the species' unique characteristics. For instance, we observed that the dominance effects play a major role, and

growth traits probably rely on complementarity and heterosis simultaneously. Given these facts, an additive-dominance genomic prediction model may be more suitable for eucalyptus RRS (DIAS et al., 2018; RESENDE et al., 2017). Furthermore, strategies for multi-population genomic evaluation, as performed in animal breeding (CESARANI et al., 2022; LEGARRA et al., 2022), are still a gap in eucalyptus breeding.

Any modifications in the breeding pipeline must undergo thorough scrutiny. This process includes but is not limited to careful assessment of genetic gains, genetic variability decay, and the potential increase of inbreeding. To accurately predict genetic gains, breeders should properly estimate heritabilities, considering the mixed-mating systems (TAMBARUSSI et al., 2018), and using linear mixed models with adequate covariance structures (CHAVES et al., 2022). To prevent a rapid decay in genetic variability and inbreeding, breeders should employ appropriate mating strategies and effectively manage the effective population size in each cycle (ISIK; MCKEAND, 2019; KIMURA; CROW, 1963; WU et al., 2016). This is accurately performed using molecular data. By implementing these measures, the RRS may achieve its objectives more efficiently, increasing the probability of long-term success and sustainability of the breeding program.

5 Concluding remarks

The results of this study provide compelling evidence of the effectiveness of the RRS strategy in improving heterosis and productivity in wood volume in eucalyptus. Furthermore, the study revealed an increase in the genetic variability of both *E. grandis* and *E. urophylla* parents over the two RRS cycles. This led to a greater genetic variability among the generated hybrids, allowing for a wider range of genetic combinations and potential for further selection. We also stressed the importance of dominance genetic variance for improving the mean annual increment of wood volume in eucalyptus. These outcomes point out the efficiency of the RRS scheme in deploying superior hybrids regarding wood volume while keeping the genetic variability necessary for continued progress in the breeding program.

References

ASSIS, T. F.; RESENDE, M. D. V. d. Genetic improvement of forest tree species. **Crop Breeding and Applied Biotechnology**, v. 11, p. 44–49, 2011.

BARIL, C. P.; VERHAEGEN, D.; VIGNERON, P.; BOUVET, J. M.; KREMER, A. Structure of the specific combining ability between two species of *Eucalyptus*.I. RAPD data. **Theoretical and Applied Genetics**, v. 94, n. 6, p. 796–803, 1997.

BAUDOUIN, L.; BARIL, C.; CLÉMENT-DEMANGE, A.; LEROY, T.; PAULIN, D. Recurrent selection of tropical tree crops. **Euphytica**, v. 96, n. 1, p. 101–114, 1997.

BERG, G. J. van den; VERRYN, S. D.; CHIRWA, P. W.; DEVENTER, F. V. Genetic parameters of interspecific hybrids of *Eucalyptus grandis* and *E. urophylla* seedlings and cuttings. **Silvae Genetica**, v. 64, n. 1-6, p. 291–308, 2015.

BERNARDO, R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. **Heredity**, v. 125, n. 6, p. 375–385, 2020.

BISON, O.; RAMALHO, M. a. P.; REZENDE, G. D. S. P.; AGUIAR, A. M.; RESENDE, M. D. V. D. Comparison between open pollinated progenies and hybrids performance in *Eucalyptus grandis* and *Eucalyptus urophylla*. **Silvae Genetica**, v. 55, n. 1-6, p. 192–196, 2006.

BOUVET, J.-M.; VIGNERON, P. Age trends in variances and heritabilities in *Eucalyptus* factorial mating designs. In: DIETERS, M. J.; MATHESON, A. C.; NIKLES, D. G.; HARWOOD, C. E.; WALKER, S. M. (Eds.). **Tree lmprovement for Sustainable Tropical Forestry. Proceedings of QFRI-IUFRO Conference**. Brisbane: QFRI, 1996. P. 127–186.

CABI. *Eucalyptus urograndis* (urograndis hybrid). **CABI Compendium**, CABI Compendium, p. 22893, 2019.

CAMPINHOS JÚNIOR, E.; IKEMORI, Y. K. Tree improvement program for *Eucalyptus* spp.: preliminary results. In: BROWN, A. G.; PALMBERG, C. M. (Eds.). **Documents FAO Third World Consultation on Forest Tree Breeding**. FAO, 1978. v. 2. (Session 3. Population improvement). P. 717–738.

CESARANI, A.; LOURENCO, D.; TSURUTA, S.; LEGARRA, A.; NICOLAZZI, E.; VANRADEN, P.; MISZTAL, I. Multibreed genomic evaluation for production traits of dairy cattle in the United States using single-step genomic best linear unbiased predictor. **Journal of Dairy Science**, v. 105, n. 6, p. 5141–5152, 2022.

CHAVES, S. F. S.; EVANGELISTA, J. S. P. C.; ALVES, R. S.; FERREIRA, F. M.; DIAS, L. A. S.; ALVES, R. M.; DIAS, K. O. G.; BHERING, L. L. Application of linear mixed models for multiple harvest/site trial analyses in perennial plant breeding. **Tree Genetics & Genomes**, v. 18, n. 6, p. 44, 2022.

COCHARD, B.; NOIRET, J. M. Second cycle reciprocal recurrent selection in oil palm, *Elaeis guineensis* Jacq. Results of Deli x La Mé hybrid tests. **Oléagineux**, v. 48, n. 11, p. 442–451, 1993.

COMSTOCK, R. E.; ROBINSON, H. F. Estimation of average dominance of genes. In: GOWEN, J. W. (Ed.). **Heterosis: A record of researches directed toward explaining and utilizing the vigor of hybrids**. Ames: Iowa State College Press, 1952. P. 494–516.

COMSTOCK, R. E.; ROBINSON, H. F.; HARVEY, P. H. A breeding procedure designed to make maximum use of both general and specific combining ability. **Agronomy Journal**, v. 41, n. 8, p. 360–367, 1949.

CORTÉS, A. J.; RESTREPO-MONTOYA, M.; BEDOYA-CANAS, L. E. Modern strategies to assess and breed forest tree adaptation to changing climate. **Frontiers in Plant Science**, v. 11, 2020.

COSTA E SILVA, J.; BORRALHO, N. M. G.; POTTS, B. M. Additive and non-additive genetic parameters from clonally replicated and seedling progenies of *Eucalyptus globulus*. **Theoretical and Applied Genetics**, v. 108, n. 6, p. 1113–1119, 2004.

COVARRUBIAS-PAZARAN, G.; WERNER, C.; GEMENET, D. Reciprocal recurrent selection based on genetic complementation: An efficient way to build heterosis in diploids due to directional dominance. **Crop Science**, v. 63, n. 4, p. 2205–2219, 2023.

DENIS, M.; BOUVET, J.-M. Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. **Tree Genetics & Genomes**, v. 9, n. 1, p. 37–51, 2013.

DIAS, K. O. G.; GEZAN, S. A.; GUIMARÃES, C. T.; NAZARIAN, A.; SILVA, L. C.; PARENTONI, S. N.; GUIMARÃES, P. E. O.; ANONI, C. O.; PÁDUA, J. M. V.; PINTO, M. O.; NODA, R. W.; RIBEIRO, C. A. G.; MAGALHÃES, J. V.; GARCIA, A. A. F.; SOUZA, J. C.; GUIMARÃES, L. J. M.; PASTINA, M. M. 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, n. 1, p. 24–37, 2018.

FALCONER, D. S. **Introduction to quantitative genetics**. 3. ed. New York, NY: Longman Scientific and Technical, 1989.

FERREIRA, F. M.; CHAVES, S. F. S.; BHERING, L. L.; ALVES, R. S.; TAKAHASHI, E. K.; SOUSA, J. E.; RESENDE, M. D. V.; LEITE, F. P.; GEZAN, S. A.; VIANA, J. M. S.; FERNANDES, S. B.; DIAS, K. O. G. A novel strategy to predict clonal composites by jointly modeling spatial variation and genetic competition. **Forest Ecology and Management**, v. 548, p. 121393, 2023.

GASCON, P.; GUEN, V. L.; NOUY, B. Résultats d'essais de second cycle de sélection récurrente réciproque chez le palmier à huile *Elaeis guineensis* Jacq. **Oléagineux**, v. 43, n. 1, p. 1–7, 1988.

GRATTAPAGLIA, D. Twelve years into genomic selection in forest trees: climbing the slope of enlightenment of marker assisted tree breeding. **Forests**, v. 13, n. 10, p. 1554, 2022.

HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a selection model. **Biometrics**, v. 31, n. 2, p. 423, 1975.

HILL, W. G. Dominance and epistasis as components of heterosis. **Zeitschrift für Tierzüchtung und Züchtungsbiologie**, v. 99, n. 1-4, p. 161–168, 1982.

HINZE, L. L.; KRESOVICH, S.; NASON, J. D.; LAMKEY, K. R. Population genetic diversity in a maize reciprocal recurrent selection program. **Crop Science**, v. 45, n. 6, p. 2435–2442, 2005.

IBARRA, L.; HODGE, G.; ACOSTA, J. J. Quantitative genetics of a hybrid population of *Eucalyptus nitens* × *Eucalyptus globulus*: estimation of genetic parameters and implications for breeding strategies. **Forests**, v. 14, n. 2, p. 381, 2023.

ISIK, F.; KUMAR, S.; MARTÍNEZ-GARCÍA, P. J.; IWATA, H.; YAMAMOTO, T. Acceleration of Forest and Fruit Tree Domestication by Genomic Selection. In: ADVANCES in Botanical Research. Elsevier, 2015. v. 74. P. 93–124.

ISIK, F.; MCKEAND, S. E. Fourth cycle breeding and testing strategy for *Pinus taeda* in the NC State University Cooperative Tree Improvement Program. **Tree Genetics & Genomes**, v. 15, n. 5, p. 70, 2019.

JONES, T. H.; STEANE, D. A.; JONES, R. C.; PILBEAM, D.; VAILLANCOURT, R. E.; POTTS, B. M. Effects of domestication on genetic diversity in *Eucalyptus globulus*. **Forest Ecology and Management**, v. 234, n. 1-3, p. 78–84, 2006.

KASSAMBARA, A. **ggpubr: ggplot2 Based Publication Ready Plots**. 2023.

KERR, R. J.; DIETERS, M. J.; TIER, B. Simulation of the comparative gains from four different hybrid tree breeding strategies. **Canadian Journal of Forest Research**, v. 34, n. 1, p. 209–220, 2004.

KIMURA, M.; CROW, J. F. On the maximum avoidance of inbreeding. **Genetical Research**, v. 4, n. 3, p. 399–415, 1963.

KRAUSE, M. D.; PIEPHO, H.-P.; DIAS, K. O. G.; SINGH, A. K.; BEAVIS, W. D. Models to estimate genetic gain of soybean seed yield from annual multi-environment field trials. **Theoretical and Applied Genetics**, v. 136, n. 12, p. 252, 2023.

LABROO, M. R.; ENDELMAN, J. B.; GEMENET, D. C.; WERNER, C. R.; GAYNOR, R. C.; COVARRUBIAS-PAZARAN, G. E. Clonal diploid and autopolyploid breeding strategies to harness heterosis: insights from stochastic simulation. **Theoretical and Applied Genetics**, v. 136, n. 7, p. 147, 2023.

LAMKEY, K. R.; EDWARDS, J. W. Quantitative Genetics of Heterosis. In: COORS, J. G.; PANDEY, S. (Eds.). **ASA, CSSA, and SSSA Books**. Madison, WI, USA: American Society of Agronomy, Crop Science Society of America, 1999. P. 31–48.

LEFÈVRE, F. Human impacts on forest genetic resources in the temperate zone: an updated review. **Forest Ecology and Management**, v. 197, n. 1-3, p. 257–271, 2004.

LEGARRA, A.; GONZÁLEZ-DIÉGUEZ, D.; VITEZICA, Z. G. Computing strategies for multi-population genomic evaluation. **Genetics Selection Evolution**, v. 54, n. 1, p. 10, 2022.

LEROY, T.; MONTAGNON, C.; CHARRIER, A.; ESKES, A. B. Reciprocal recurrent selection applied to *Coffea canephora* Pierre: II. Estimation of genetic parameters. **Euphytica**, v. 74, n. 1-2, p. 121–128, 1993.

LEROY, T.; MONTAGNON, C.; CILAS, C.; YAPO, A.; CHARMETANT, P.; ESKES, A. Reciprocal recurrent selection applied to *Coffea canephora* Pierre. III. Genetic gains and results of first cycle intergroup crosses. **Euphytica**, v. 95, n. 3, p. 347–354, 1997.

LU, W.; ARNOLD, R. J.; ZHANG, L.; LUO, J. Genetic diversity and structure through three cycles of a *Eucalyptus urophylla* S.T.Blake breeding program. **Forests**, v. 9, n. 7, p. 372, 2018.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. 1. ed. Sunderland: Sinauer Associates, 1998.

MADHIBHA, T.; MUREPA, R.; MUSOKONYI, C.; GAPARE, W. Genetic parameter estimates for interspecific *Eucalyptus* hybrids and implications for hybrid breeding strategy. **New Forests**, v. 44, n. 1, p. 63–84, 2013.

MELCHINGER, A. E.; FERNANDO, R.; STRICKER, C.; SCHÖN, C.-C.; AUINGER, H.-J. Genomic prediction in hybrid breeding: I. Optimizing the training set design. **Theoretical and Applied Genetics**, v. 136, n. 8, p. 176, 2023.

MEUNIER, J.; GASCON, J. P. Le schéma général d'amélioration du palmier a huile a l'i R. H. O. **Oléagineux**, v. 27, n. 1, p. 1–12, 1972.

NIKLES, D. G.; GRIFFING, A. R. Breeding hybrids of forest trees: Definitions, theory, some practical examples, and guidelines on strategy with tropical acacias. In: CARRON, L. T.; AKEN, K. M. (Eds.). **Breeding technologies for tropical acacias: Proceedings of an international workshop held in Twau, Sabah, Malaysia, 1-4 July 1991**. Canberra: ACIAR, 1992. (ACIAR proceedings, 37). P. 101–109.

NIKLES, D. G.; NEWTON, R. S. Correlations of breeding values in pure and hybrid populations of hoop pine and some southern pines in Queensland and relevance to breeding strategies. In: DEAN, C. A. (Ed.). **Proceedings of the 11th meeting of RWG 1 (Forest Genetics)**. Canberra: CSIRO, 1991. P. 192–196.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545–554, 1971.

POTTS, B. M.; DUNGEY, H. S. Interspecific hybridization of *Eucalyptus*: key issues for breeders and geneticists. **New Forests**, v. 27, n. 2, p. 115–138, 2004.

R CORE TEAM. **R: A Language and environment for statistical computing**. Viena, Áustria: R Foundation for Statistical Computing, 2023.

REMBE, M.; REIF, J. C.; EBMEYER, E.; THORWARTH, P.; KORZUN, V.; SCHACHT, J.; BOEVEN, P. H. G.; VARENNE, P.; KAZMAN, E.; PHILIPP, N.; KOLLERS, S.; PFEIFFER, N.; LONGIN, C. F. H.; HARTWIG, N.; GILS, M.; ZHAO, Y. Reciprocal recurrent genomic selection is impacted by genotype-by-environment interactions. **Frontiers in Plant Science**, v. 12, p. 703419, 2021.

RESENDE, M. D. V. **Genética biométrica e Estatística no melhoramento de plantas perenes**. 1. ed. Brasilia: Embrapa Informação Tecnológica, 2002.

RESENDE, M. D. V.; BARBOSA, M. H. P. **Melhoramento genético de plantas de propagação assexuada**. 1. ed. Colombo: Embrapa Florestas, 2005.

RESENDE, M. D. V.; HIGA, A. R. Estratégias de melhoramento para eucaliptos visando a seleção de híbridos. **Boletim de Pesquisa Florestal**, v. 21, p. 49–60, 1990.

RESENDE, R. T.; RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F.; TAKAHASHI, E. K.; SILVA-JUNIOR, O. B.; GRATTAPAGLIA, D. Assessing the expected response to genomic selection of individuals and families in *Eucalyptus* breeding with an additive-dominant model. **Heredity**, v. 119, n. 4, p. 245–255, 2017.

RETIEF, E. C. L.; STANGER, T. K. Genetic parameters of pure and hybrid populations of *Eucalyptus grandis* and *E. urophylla* and implications for hybrid breeding strategy. **Southern Forests: a Journal of Forest Science**, v. 71, n. 2, p. 133–140, 2009.

REZENDE, G. D. S. P.; RESENDE, M. D. V.; ASSIS, T. F. Eucalyptus breeding for clonal forestry. In: FENNING, T. (Ed.). **Challenges and Opportunities for the World's Forests in the 21st Century**. Dordrecht: Springer Netherlands, 2014. (Forestry Sciences). P. 393–424.

ROBINSON, H. F.; COMSTOCK, R. E.; HARVEY, P. H. Genetic variances in open pollinated varieties of corn. **Genetics**, v. 40, n. 1, p. 45–60, 1955.

RUTKOSKI, J. E. Estimation of realized rates of genetic gain and indicators for breeding program assessment. **Crop Science**, v. 59, n. 3, p. 981–993, 2019.

SHULL, G. H. The composition of a field of maize. **Journal of Heredity**, os-4, n. 1, p. 296–301, 1908.

SIMIQUELI, G. F.; RESENDE, R. T.; TAKAHASHI, E. K.; SOUSA, J. E.; GRATTAPAGLIA, D. Realized genomic selection across generations in a reciprocal recurrent selection breeding program of *Eucalyptus* hybrids. **Frontiers in Plant Science**, v. 14, p. 1252504, 2023.

SPRAGUE, G. F.; TATUM, L. A. General vs. specific combining ability in single crosses of corn. **Agronomy Journal**, v. 34, n. 10, p. 923–932, 1942.

TAMBARUSSI, E. V.; PEREIRA, F. B.; SILVA, P. H. M.; LEE, D.; BUSH, D. Are tree breeders properly predicting genetic gain? A case study involving *Corymbia* species. **Euphytica**, v. 214, n. 8, p. 150, 2018.

THE VSNI TEAM. **asreml: Fits Linear Mixed Models using REML**. 2023.

WENG, Q.; HE, X.; LI, F.; LI, M.; YU, X.; SHI, J.; GAN, S. Hybridizing ability and heterosis between *Eucalyptus urophylla* and *E. tereticornis* for growth and wood density over two environments. **Silvae Genetica**, v. 63, n. 1-6, p. 15–23, 2014.

WICKHAM, H. **ggplot2: Elegant graphics for data analysis**. 2. ed. Cham: Springer, 2016.

WILLHAM, R. L.; POLLAK, E. Theory of Heterosis. **Journal of Dairy Science**, v. 68, n. 9, p. 2411–2417, 1985.

WU, H. X.; HALLINGBÄCK, H. R.; SÁNCHEZ, L. Performance of seven tree breeding strategies under conditions of inbreeding depression. **G3 Genes|Genomes|Genetics**, v. 6, n. 3, p. 529–540, 2016.

CHAPTER 5

DATA ANALYSIS IN PERENNIAL PLANT BREEDING

1 Introduction

The primary goals of perennial plant breeding initiatives revolve around the development of varieties that exhibit increased yield, stability, and resilience to biotic factors such as pests and diseases, as well as abiotic factors like drought, wind, and frost. In this context, the capability to generate data through systematic experimentation and effectively analyse and interpret it becomes paramount. Experimentation serves the purpose of planning, executing, gathering data, analysing that data, and drawing meaningful interpretations from the results. It forms an intrinsic component of statistical methodologies, as the soundness of outcomes from statistical analyses is intricately tied to the quality of experimentation. Ronald Fisher, a renowned statistician at the English research institute Rothamsted, is credited with the development of crucial statistical methods such as analysis of variance and maximum likelihood. Fisher underscored the pivotal importance of repetition, randomization, and local control for the efficiency of experiments (BOX, 1980; FISHER, 1926).

Scientific research extensively employs experiments to assess hypotheses. The methodology for experimenting can vary based on the research objective. Nevertheless, adherence to fundamental principles of experimentation, including repetition, randomization, and local control, is essential for ensuring the validity of conclusions. Below, we briefly describe each principle:

- **Repetition:** consists of applying the same treatment to two or more experimental units (plots). It allows estimating the experimental error and evaluating, more precisely, the effect of each treatment. Experimental error is characterized by the variance between experimental units that received the same treatment.

- **Randomization:** consists of randomly distributing the experimental units in the field. It

is used to guarantee the independence of errors, that is, to prevent certain treatments from being favoured.

- **Local control:** consists of grouping homogeneous experimental units into subsets called blocks which received all (complete blocks) or only part (incomplete blocks) of the treatments. It allows reducing the experimental error by controlling the variation between the experimental units.

1.1 Statistical analysis

In a comprehensive statistical analysis, several activities are undertaken, including predicting mean components, estimating variance components, conducting hypothesis tests, and inferring accuracy, bias, and precision of the estimation/prediction. In the context of mixed models, these activities involve Best Linear Unbiased Prediction (BLUP) for prediction of means (HENDERSON, 1975), Residual Maximum Likelihood (REML) for variance component estimation (PATTERSON; THOMPSON, 1971), likelihood ratio test (LRT) for hypothesis testing (WILKS, 1938), and the calculation of prediction accuracy and prediction error variance (MRODE, 2014). In the REML/BLUP procedure, bias is assumed to be null, as these estimators/predictors belong to the class of the best unbiased linear estimators/predictors (RESENDE, 2016).

1.2 Random effects and fixed effects

In the formulation of a mixed model, a crucial decision is determining which factors should be treated as fixed (in addition to the overall mean) and which as random (in addition to the residuals). A fixed factor usually represents the effect of specific conditions selected for the study, while a random factor typically represents the effect of a sample of observed conditions from a broader population, with the variability of the population/sample being of interest. Genetic effects are often considered random. Components of experimental design, such as blocks and plots effects, can be either random or fixed, depending on the situation (PIEPHO et al., 2008; RESENDE; DUARTE, 2007; SMITH et al., 2005).

1.3 Observation units and experimental units

The structure of certain plant breeding trials can sometimes lead breeders to inadvertently conflate observation units with experimental units. Consider, for instance, a trial arranged in a randomized complete blocks design. In this particular trial setup, each plot consisted of 9 trees, arranged in three rows with three trees originating from a single biparental cross (full-sibs). Across 5 blocks, a total of 10 progenies were evaluated. In this context, the experimental unit refers to the randomized and replicated entity. Conversely, the observation units pertain to each tree within the trial, where formal data collection takes place. Therefore, in our example, we encounter 50 experimental units (10×5) and 450 observation units ($10 \times 5 \times 9$). It is important to note that while the observation units are all distinct, the trees comprising the experimental units share common ancestors. Confounding the two concepts can lead to serious statistical problems - like pseudo-replication - and invalidate the results of years of experimenting. This underscores the necessity of clearly distinguishing between observation and experimental units. Doing so not only aids in the planning and proper installation of new trials but also ensures the accurate construction of statistical models.

1.4 `Selegen REML/BLUP`

This software employs mixed models and was specifically designed for routine plant genetic improvement programs, accommodating various plant categories, including allogamous, autogamous, and those with mixed reproductive systems. It is versatile in handling different experimental designs, crossing designs, multi-location trials, repeated measurements, and progenies from diverse populations. `Selegen REML/BLUP` is capable of adjusting effects, estimating variance components, predicting additive, dominance, and genotypic genetic values, calculating genetic gain with selection, and determining the effective population size. It supports the testing of the significance of random effects through the likelihood ratio test (LRT). The software accommodates both continuous variables (linear models) and categorical variables (generalized linear models). With its user-friendly interface and interpretability, `Selegen REML/BLUP` efficiently addresses various situations encountered in plant breeding (RESENDE, 2016). Details on its usage can be found in Resende (2007) and Resende (2002).

1.5 ASReml-R

ASReml-R (THE VSNI TEAM, 2023) is an R (R CORE TEAM, 2023) software package designed for comprehensive mixed model analyses, capable of conducting all the analyses cited in the preceding section. Beyond basic analyses, ASReml-R facilitates the adjustment of intricate models. This includes the application of autoregressive models for spatial analysis, accommodating different covariance structures for the examination of data derived from various environments or repeated measurements, and incorporating information from kinship, whether probabilistic or genomic. Subsequent sections will elucidate the utilization and interpretation of ASReml-R, version 4.2, in the context of data analysis for perennial plant breeding, shedding light on specific functionalities of the package. For a more in-depth guide, refer to the ASReml-R manual available at https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/ASReml-R-Reference-Manual-4.2.pdf.

2 Individual analysis

In this section, we will provide a detailed, step-by-step guide on how to conduct an individual analysis utilizing the ASReml-R package. The dataset employed for this demonstration is accessible at https://github.com/saulo-chaves/MHT_MET_MM (CHAVES et al., 2022). The dataset encompasses the assessment of the fruit yield of 25 cupuassu tree genotypes [*Theobroma grandiflorum* (Willd. Ex. Spreng) Schum.] in a randomized complete block design featuring five replications throughout 12 harvests. We will use the tidyverse (WICKHAM et al., 2019) package for data management and the ASReml-R package for data analysis:

Box 1. Packages used

```
library(tidyverse)
library(asreml)
```

To perform individual analyses, we will first load the dataset. As we will carry out an individual analysis, we must choose only one among the 12 harvests. We will use the following command:

Box 2. Dataset and factors

```
data = read.csv("https://raw.githubusercontent.com/saulo-chaves/
                                        MHT_MET_MM/main/Data/D1.csv")


# Transform 0 into NA
data$fy[data$fy == 0] = NA


# Select only one harvest
data_h07 = data[data$yr == "H07", c('gen', 'block', 'fy')]


# Transformation into factor
data_h07 = transform(data_h07, gen = as.factor(gen),
                     block = as.factor(block))
```

This dataset has missing data (NA). This will not be a problem for executing the analysis, since linear mixed models can deal with unbalanced data (RESENDE; DUARTE, 2007).

The linear mixed model for the individual analysis is:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_1\mathbf{b} + \mathbf{Z}_1\mathbf{g} + \mathbf{e} \tag{1}$$

where: $\mathbf{y}$ is the vector of phenotypic data, μ is the intercept, $\mathbf{b}$ is the vector of fixed repetition effects, $\mathbf{g}$ is the vector of random genotype effects [$\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}\sigma_g^2)$] and $\mathbf{e}$ is the vector of (random) residual effects [$\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$]. $\mathbf{1}$ is a vector of ones; and $\mathbf{X}_1$ and $\mathbf{Z}_1$ are the incidence matrices for $\mathbf{b}$ and $\mathbf{g}$, respectively. In R, the model will be built using the following command:

Box 3. Fitting the model in ASReml-R

```
m = asreml(fixed = fy ~ block,
           random = ~ gen,
           data = data_h07)
```

2.1 Hyphotesis test

Once the complete model has been fitted, it is necessary to assess the significance of genotype effects. This is performed by the likelihood ratio test (LRT). To do this, we must

fit the reduced model, that is, without the to-be-tested effect. With the reduced model adjusted, simply use the `lrt` function:

Box 4. Hypothesis test of random effects using the `lrt` function

```
m = asreml(fixed = fy ~ block,
           random = ~ gen,
           data = data_h07)
```

The command above will generate the result of Table 1. Note that the p-value < 0.05 [`Pr(Chisq)` - Chi-square test (χ^2)]. Thus, the tested effect is significant.

Table 1: Likelihood ratio test result: LR statistic and p-value considering the χ^2 test with $\alpha = 0.05$ and one degree of freedom

LR statistic	p-value (χ^2 test)
7.55	0.003

2.2 Variance component

Once the significance is checked via LRT, we can extract the variance components using the following command:

Box 5. Variance components

```
summary(m)$varcomp
```

This function will generate the results of Table 2.

Table 2: Variance component estimates and their respective standard errors. σ^2_g is the genotypic variance and σ^2_e is the residual variance

Variance component	Estimate	Standard error
σ^2_g	20.074	10.573
σ^2_e	78.555	11.339

2.3 Accuracy and heritability

After obtaining the variance components, we can calculate some useful genetic parameters such as accuracy and heritability, using the following equations:

- Accuracy (RESENDE, 2002):

$$r = \sqrt{1 - \frac{PEV}{\sigma_g^2}} \tag{2}$$

- Heritability on individual level:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \tag{3}$$

- Heritability on mean level:

$$H_p^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{r}} \tag{4}$$

- Heritability on mean level (CULLIS et al., 2006):

$$H_c^2 = 1 - \frac{\overline{V(\Delta)}}{2\sigma_g^2} \tag{5}$$

where PEV is the prediction error variance, obtained from the inverse of the diagonal elements of the mixed model equation's coefficient matrix, r is the number of replicates, and $\overline{V(\Delta)}$ is the mean variance of the difference between two BLUPs. The equivalences, interrelations, and similarities between H_p^2, H_c^2 and r^2 are presented in Resende and Alves (2022).

To estimate the accuracy and heritabilities, we will use `ASReml-R`'s `predict` function, which also extracts the BLUPs:

Box 6. BLUPs; accuracy and heritability

```
predm = predict(object = m, classify = "gen", sed = TRUE)


# BLUPs
blup = predm$pvals


# Accuracy
PEV = mean(predm$pvals$std.error^2)
s2g = varcomp["gen","component"]
acc = sqrt(1-PEV/s2g)


# Heritability on the individual level
s2pheno = sum(varcomp[, 1])
her_ind = s2g/s2pheno


# Heritability on the mean level
s2e = varcomp[2, 1]
her_med_std = s2g/(s2g + s2e/10)


# Heritability on the mean level Cullis et al. (2006)
MVdelta = mean((predm$sed^2)[upper.tri(predm$sed^2, diag = F)])
her_med_cullis = 1-MVdelta/(2*s2g)
```

Using the codes above, we reach the results of Table 3:

Table 3: Mean accuracy, and heritabilities on the individual level, mean level and mean level proposed by Cullis et al. (2006)

Parameter	Value
r	0.740
H^2	0.204
H_p^2	0.719
H_c^2	0.749

Note that H_p^2 and H_c^2, though both indicating heritability at the mean level, are distinct. This discrepancy arises due to the imbalance, as illustrated in the aforementioned example, attributed

to the loss of plots. The H_p^2 formula is proposed for balanced data and may occasionally distort the actual heritability value in the presence of imbalances. Apart from H_c^2, several alternatives exist for situations involving unbalanced data, detailed in Schmidt et al. (2019a) and Schmidt et al. (2019b). An alternative suggested by Henderson (1984) and Resende (2002) is reliability, equivalent to the square of accuracy (r^2). When dealing with balanced data, it's noteworthy that $r^2 = H_p^2 = H_c^2$. Note that this is not true in Table 3, given the imbalance of the dataset used.

2.4 Genetic gains

With the BLUPs, we can predict the genetic gains after selection:

$$\text{Gains} = \frac{\overline{BLUP} - \mu}{\mu} \times 100 \tag{6}$$

where $\overline{BLUP}$ is the mean BLUPs of the selected genotypes and μ is the grand mean. In the example, we selected the top five candidates:

Box 7. Genetic gain

```
# BLUPs of the selected genotypes
sel_blup = blup[order(blup$predicted.value, decreasing = T),][1:5,]


# Computing the gain
mu.sel = mean(sel_blup$predicted.value)
mu.pop = mean(blup$predicted.value)
Gain = (mu.sel - mu.pop)/mu.pop * 100
```

$$\mu_{sel} = \frac{32.825 + 31.259 + 27.779 + 27.328 + 26.662}{5} = 29.170$$

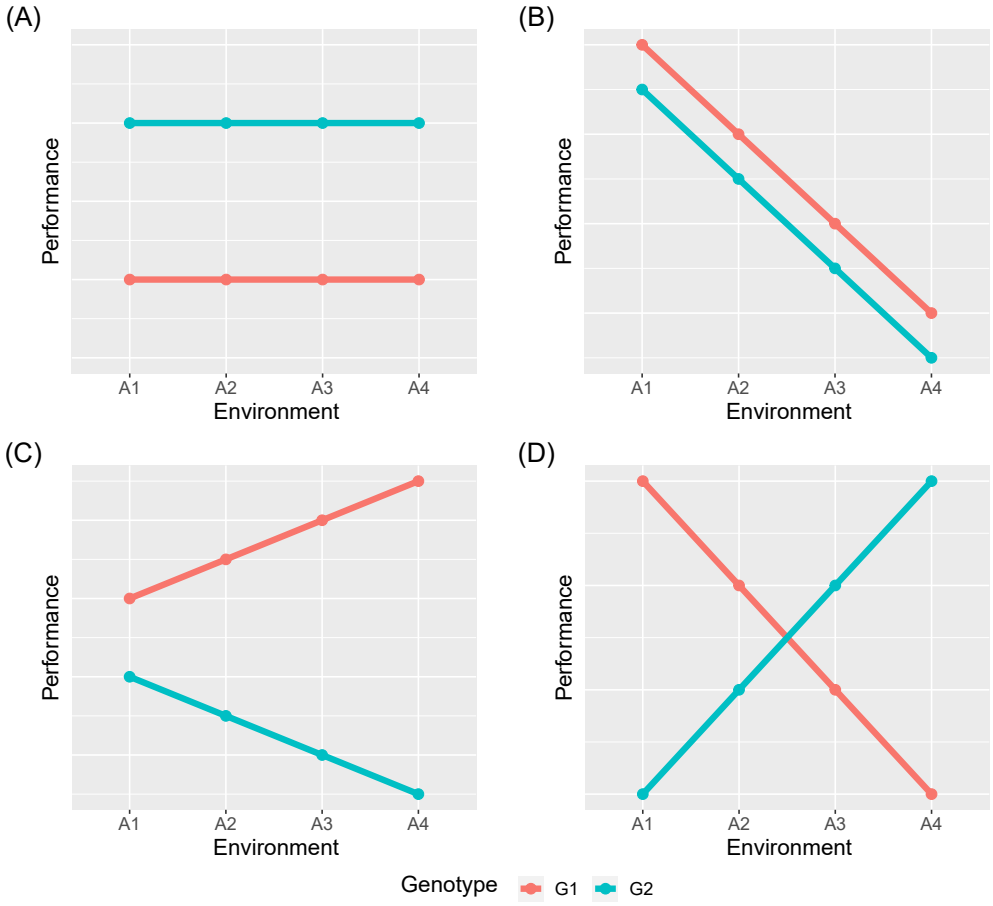
$$\text{Gain} = \frac{29.170 - 24.079}{24.079} \times 100 = 21.14\%$$

3 Multi-location or Multi-year analysis

The genotype-by-location or genotype-by-year interaction, hereinafter referred to as genotype-by-environment interaction (GEI) encompasses the inconsistent performance of two

or more genotypes across different environments. This interaction can manifest as simple (non-crossover or scale type) or complex (cross-over or rank type) (COOPER; DELACY, 1994; WATERS et al., 2023), as illustrated in Figure 1. In its simple form, the interaction does not pose challenges in cultivar recommendations and obviates the necessity for experiments across various environments. Conversely, in its complex manifestation, it becomes a challenge to recommend cultivars, necessitating experiments in diverse environments (MALOSETTI et al., 2013). In such complex scenarios, breeders may opt for specific recommendations, using environmental stratification based on groups of environments where the GA interaction is simple. Alternatively, general recommendations can be made through yield and stability analyses.

Figure 1: Possible types of genotype-by-environment interaction (GEI), considering four environments (A1, A2, A3 and A4) and two genotypes (G1 and G2). (A) There is no GEI (parallel lines; slope $\beta_1 = 0$. (B) There is no GEI (parallel lines; slope $\beta_1 \neq 0$; difference in phenotypic response attributed solely to differences between environments; additivity of environmental effects). (C) Simple GEI (non-parallel lines; scale effects due to heterogeneity of variances in the environments). (4) Complex GEI (intercept lines; rank inversion, absence of high and positive correlation between genotype responses in environments).



Several methods have been proposed for the evaluation of GA interaction (LI et al., 2017;

SMITH et al., 2005). Here, we will address the methods currently used to evaluate GA interaction in perennial plant breeding. The packages used will be the same as those used in the previous topic (Box 1).

3.1 Variance structures

To exemplify the usage of linear mixed models with different covariance structures, we will employ the same dataset used in the previous topic. However, this time, we will consider the 12 harvests:

Box 8. Declaring the factors in the complete dataset

```
data = transform(data, env = as.factor(yr), gen = as.factor(gen),
                 block = as.factor(block))
```

The linear mixed model for the joint analysis is:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_1\mathbf{b} + \mathbf{X}_2\mathbf{a} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{ga} + \mathbf{e} \tag{7}$$

where $\mathbf{y}$ is the vector of phenotypic data, μ is the intercept, $\mathbf{b}$ is the vector of fixed repetition effects, $\mathbf{a}$ is the vector of fixed environment (harvest) effects, $\mathbf{g}$ is the vector of random genotypic effects [$\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_g)$], $\mathbf{ga}$ is the vector of GEI effects [$\mathbf{ga} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_{ga}$], and $\mathbf{e}$ is the vector of random residual effects [$\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_e)$]. $\mathbf{1}$ is a vector of ones, and $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are the incidence matrices of $\mathbf{b}$, $\mathbf{a}$, $\mathbf{g}$ and $\mathbf{ga}$, respectively.

Beginning with the most straightforward and parsimonious model (`m1`), it is feasible to generate additional models by defining covariance structures. `m1` is a compound symmetry model, which assumes uniform variances for both genotypic and residual components, delivering an estimate for the GEI variance. In R, this model is implemented using the following command:

Box 9. Compound symmetry model

```
m1 = asreml(fixed = fy ~ env + block:env,
            random = ~gen + gen:env,
            maxit = 50,
            data = data)
```

Assuming that variances (genotypic and residual) are homogeneous is, in most cases, inappropriate. Modelling covariance structures can provide more credible results. For example, one can consider heterogeneous residual variances throughout environments (`m2`). In this case, a block diagonal covariance structure is adjusted for the residuals, using the following command:

Box 10. Compound symmetry model with heterogeneous residual variances

```
m2 = asreml(fixed = fy ~ env + block:env,
            random = ~gen + gen:env,
            residual = ~dsum(~id(units)|env),
            maxit = 50,
            data = data)
```

The same principle can be extended to modelling genotypic effects. A viable alternative is the compound symmetry model with heterogeneous variances (SCH) (`m3`). Similar to the block diagonal structure modelled for the error, SCH assumes that genotypic variances differ across environments. Unlike models estimating GEI variance, SCH focuses on parameterization where genotypic effects are nested within environments (DIAS et al., 2018; GEZAN et al., 2017). GEI is explored through the correlation between environments, a metric provided by the analysis itself. A high correlation (close to 1) indicates a simple interaction, while a lower correlation implies a complex interaction. This model is implemented in `R` using the following command:

Box 11. Heterogeneous compound symmetry model with heterogeneous residual variances

```
m3 = asreml(fixed = fy ~ env + block:env,
            random = ~corh(env):gen,
            residual = ~dsum(~id(units)|env),
            maxit = 50,
            data = data)
```

A second alternative for modelling genotypic effects is the unstructured covariance structure. In this scenario, each environment's response variable is treated as a distinct variable. Hence, in the fourth model (`m4`), variances for each environment and covariances between environments are estimated. This model can be implemented in `R` using the following command:

Box 12. Unstructured model with heterogeneous residual variances

```
m4 = asreml(fixed = fy ~ env + block:env,
            random = ~us(env):gen,
            residual = ~dsum(~id(units)|env),
            maxit = 50,
            data = data)
```

In this case, the GEI's complexity can be assessed using the pairwise genotypic correlations between environments (ρ):

$$\rho = \frac{\sigma_{g_{a,a'}}}{\sqrt{\sigma^2_{g_a}\sigma^2_{g_{a'}}}} \tag{8}$$

where $\sigma_{g_{a,a'}}$ is the genotypic covariance between the environments a and a'. Calculating the correlations using variances and covariances generates a more reliable estimate if compared to computing using environment-wise BLUPs (PIEPHO, 2018).

Due to overparameterization, particularly when dealing with a substantial number of environments (>10), convergence issues commonly arise, representing a limitation (KELLY et al., 2007). For instance, when running the command shown in Box 12, you might encounter the message: "`Log-likelihood not converged`", indicating that the model fit failed to reach convergence. To address this challenge, Factor Analytic Mixed Models (FAMM) can and should be employed, and these will be discussed in the next section.

Two criteria usually employed for model selection are: Akaike Information Criterion (AIC, AKAIKE, 1974) and Bayesian Information Criterion (BIC, SCHWARZ, 1978), given by:

$$AIC = -2LogL + 2p \tag{9}$$

$$BIC = -2LogL + pLog[n - r(x)] \tag{10}$$

where p is the number of estimated parameters, related to the number of random effects in the model and the covariance structures used in the model; L is the maximum point of the residual likelihood function, n is the number of observations and $r(x)$ is the rank of the fixed effects' incidence matrix. The lower the AIC or BIC values, the better the model fit. According to Resende and Alves (2020), the best model is the one that presents an AIC value with at least two units of difference lower than the other models in comparison. It is worth mentioning

that, frequently, AIC and BIC do not converge when it comes to choosing the best model (DRTON; PLUMMER, 2017), as can be seen in Table 4. AIC is an efficiency criterion and tends to prioritize more parameterized models than the BIC, which is a consistency criterion (CAVANAUGH; NEATH, 2019; ISIK et al., 2017; NEATH; CAVANAUGH, 2012). Therefore, it is important to choose just one when choosing the model. The information criteria can be obtained using the commands below:

Box 13. Information criteria

```
data.frame(
 "Model" = seq(1, 3),
 "AIC" = c(summary(m1)$aic, summary(m2)$aic, summary(m3)$aic),
 "BIC" = c(summary(m1)$bic, summary(m2)$bic, summary(m3)$bic)
           )
```

Table 4: Information criteria values of each tested model: AIC and BIC

Model	Genotypic effects	Residual effects	AIC	BIC
1	Compound symmetry	Homoscedastic	7531.32	7547.10
2	Compound symmetry	Heteroscedastic	7337.27	7310.88
3	Heterogeneous compound symmetry	Heteroscedastic	7226.57	7358.02

According to the AIC, the best-fitting model is Model 3 (`m3`), which has heterogeneous genotypic and residual variances. In turn, the BIC indicated Model 2 (`m2`), which has homogeneous genotypic variances and heterogeneous residuals. For this example, we will follow the result indicated by the AIC. Estimates of the variance components (Table 5) of the selected model are extracted using the following command:

Box 14. Variance components of the best-fitted model

```
summary(m3)$varcomp
```

A strong genotypic correlation between environments is evident, suggesting a simple GEI. Both genotypic and residual variances show variability across environments. Consequently, the genotypic values (BLUPs) predicted by Model 3 exhibit higher accuracy than those predicted by Models 1 and 2.

Table 5: Variance component estimates of Model 3 (heterogeneous compound symmetry for the genotypic effects and block diagonal for the residual effects). ρ_g is the genotypic correlation between environments, $\sigma^2_{g_a}$ is the genotypic variance in the environment a, and $\sigma^2_{e_a}$ is the residual variance in the environment a

Component	Estimate	Std.error
ρ_g	0.904	0.091
$\sigma^2_{g_1}$	0.975	0.645
$\sigma^2_{g_2}$	4.593	1.987
$\sigma^2_{g_3}$	19.778	10.832
$\sigma^2_{g_4}$	20.418	9.698
$\sigma^2_{g_5}$	15.734	7.555
$\sigma^2_{g_6}$	13.759	8.524
$\sigma^2_{g_7}$	7.656	5.441
$\sigma^2_{g_8}$	3.726	4.956
$\sigma^2_{g_9}$	2.088	3.036
$\sigma^2_{g_{10}}$	12.435	7.545
$\sigma^2_{g_{11}}$	8.926	4.965
$\sigma^2_{g_{12}}$	6.738	0.976
$\sigma^2_{e_1}$	10.175	1.459
$\sigma^2_{e_2}$	29.783	4.078
$\sigma^2_{e_3}$	81.303	11.011
$\sigma^2_{e_4}$	46.079	6.234
$\sigma^2_{e_5}$	84.933	11.456
$\sigma^2_{e_6}$	69.583	9.242
$\sigma^2_{e_7}$	114.371	14.855
$\sigma^2_{e_8}$	103.496	13.436
$\sigma^2_{e_9}$	73.584	9.896
$\sigma^2_{e_{10}}$	45.290	6.108
$\sigma^2_{e_{11}}$	0.904	0.091
$\sigma^2_{e_{12}}$	0.975	0.645

3.2 Factor analytic mixed models

Factor Analytic Mixed Models (FAMM) have the advantages of models with unstructured covariance structures, while maintaining parsimony (PIEPHO, 1997; PIEPHO, 1998). This method reduces dimensionality through the creation of a set of latent variables, or factors (K), capturing common genetic effects across environments. By estimating variances and covariances through these factors, FAMM approximates the unstructured covariance structure (RESENDE; THOMPSON, 2004; SMITH et al., 2001). Additionally, incorporating kinship information allows the decomposition of the GA interaction into additive and dominance effects across environments. This provides a more nuanced understanding of a trait in the population

and the environmental influence on each component of genetic variance (DIAS et al., 2018).

The FAMM associated with the example dataset's analysis has the same description as the one of Equation 7, except for the explicit GEI effect ($\mathbf{Z}_2\mathbf{ga}$). This adaptation merges the genotypic effects ($\mathbf{g}$) with the GEI effects, as previously described in the heterogeneous compound symmetry and unstructured models. In FAMM, $\mathbf{g}$ can be represented by a multiple regression:

$$\mathbf{g} = (\mathbf{\Lambda} \otimes \mathbf{I}_T)\mathbf{f} + \boldsymbol{\delta} \tag{11}$$

where $\mathbf{\Lambda}$ is a $A \times K$ matrix of loadings [$\mathbf{\Lambda} = \{\lambda_{k_a}\}$, in which λ_{k_a} is the loading of the k-th factor in the a-th environment], in which A is the number of environments, $\otimes$ is the Kronecker product, $\mathbf{I}_T$ is an identity matrix of order T, which represents the number of genotypes; $\mathbf{f}$ is the $TK \times 1$ vector of scores, and $\boldsymbol{\delta}$ is the $AT \times 1$ vector of lack of fit effects. The joint distribution of $\mathbf{f}$ and $\boldsymbol{\delta}$ is:

$$\begin{bmatrix} f \\ \delta \end{bmatrix} \sim NMV\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D} \otimes \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi} \otimes \mathbf{P} \end{bmatrix}\right) \tag{12}$$

where $\mathbf{P}$ can be a kinship matrix (probabilistic or genomic, additive or dominant) or an identity matrix if the kinship information is unavailable, $\mathbf{D}$ is a $K \times K$ symmetric positive (semi)-definite matrix of factor score variance; and $\mathbf{\Psi}$ is $A \times A$ a diagonal matrix of specific variances ($\mathbf{\Psi} = \{\psi_a\}$, in which ψ_a is the specific variance of the a-th environment). The specific variances are the information not captured by any factor. Thus, $\boldsymbol{\delta}$ in Equation 11 represents effects particular to each environment. We can separate BLUP into two parts: the common genotypic effects $(\mathbf{\Lambda} \otimes \mathbf{I}_T)\mathbf{f}$ and the isolated effects ($\boldsymbol{\delta}$). Stefanova and Buirchell (2010) and Smith and Cullis (2018) stress that using marginal BLUPs is more adequate for the prediction of genotypic values.

The factor analytic covariance structure is described as:

$$\mathbf{\Sigma}_g = (\mathbf{\Lambda}\mathbf{D}\mathbf{\Lambda}' + \mathbf{\Psi}) \otimes \mathbf{P} \tag{13}$$

where $\mathbf{\Sigma}_g$ is the matrix of genotypic covariances, and $\mathbf{\Lambda}'$ is the transpose of the loadings matrix. The diagonal elements of $\mathbf{\Sigma}_g$ represent the environment-wise genotypic variance, and off-diagonal elements are the covariances.

To fit a FAMM with two factors using `ASReml-R`, we use the following command:

Box 15. FAMM with two factors

```
m5 = asreml(fixed = fy ~ env + block:env,
            random = ~fa(env, 2):gen,
            residual = ~dsum(~id(units)|env),
            data = data,
            maxit = 50)
```

We built a function to easily extract the results of the FAMM. This function is stored in GitHub, and can be accessed and employed using the following command:

Box 16. `fa.outs` function

```
# Loading the "fa.outs" function
source("https://raw.githubusercontent.com/saulo-chaves/May_b_useful
       /main/fa.outs.R")

# Using the "fa.outs" function
m5.res = fa.outs(model = m5, name.env = "env", name.gen = "gen")
```

Rotation

In FAMM, the number of factors corresponds to one minus the number of environments in the dataset. However, it's crucial to note that a higher number of factors implies increased complexity and model parameterization. When considering more than one factor ($K > 1$), $\mathbf{\Lambda}$ becomes non-unique, impacting its identifiability and necessitating the imposition of constraints. `ASReml-R` imposes two constraints: making $\mathbf{D} = \mathbf{I}_K$, and setting the $K(K-1)/2$ elements above the diagonal of $\mathbf{\Lambda}$ to be 0, implying an increasing number of null elements as more factors are considered. For instance, in a hypothetical model with four factors, the loading

matrix takes the following form:

$$\mathbf{\Lambda} = \begin{bmatrix} 424.191 & 0 & 0 & 0 \\ 183.58 & 390.664 & 0 & 0 \\ 262.917 & 112.777 & 70.908 & 0 \\ 61.022 & 224.59 & -27.492 & 81.848 \\ 310.049 & -53.288 & 47.619 & -103.561 \\ 248.826 & -21.621 & 147.042 & -99.481 \\ -70.48 & 340.896 & -34.743 & 173.096 \\ 278.531 & 97.429 & -51.619 & 37.303 \\ 76.908 & 182.156 & 109.951 & 174.963 \\ 258.95 & 3.865 & 218.588 & 135.483 \end{bmatrix} \tag{14}$$

Note that this constraint has no biological meaning and can harm interpretations since some environments will have nil loadings. For this reason, $\mathbf{\Lambda}$ must be rotated. A relatively simple method is to perform the singular value decomposition of $\mathbf{\Lambda}$ (SMITH et al., 2021), so that:

$$\mathbf{\Lambda} = \mathbf{ULV}' \tag{15}$$

where $\mathbf{L}$ is a diagonal matrix with elements equal to the square root of $\mathbf{\Lambda\Lambda}'$ eigenvalues, in increasing order; and $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices, with columns equal to $\mathbf{\Lambda\Lambda}'$ and $\mathbf{\Lambda}'\mathbf{\Lambda}$, respectively. A quick reminder: an orthonormal matrix holds the properties of an orthogonal matrix and a normal matrix: the inner-product of its columns is nil, and the norm of each column is one ($\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = 1$). After decomposing, $\mathbf{U}$ will be the matrix of rotated loadings, multiplied by 1 or -1 (i.e., $\mathbf{\Lambda}^{\star} = \mathbf{U}c$, with $c = 1$ or $c = -1$). The objective is to ensure that most of the first factor loadings are positive (SMITH; CULLIS, 2018). $\mathbf{L}$ will be the matrix of factor score variances (i.e., $\mathbf{D} = \mathbf{L}$). This procedure provides two properties that benefit interpretability (SMITH et al., 2021): $\mathbf{D}$ is a diagonal matrix with elements written in decreasing order and $\mathbf{\Lambda\Lambda}'$ is an identity matrix, i.e., $\mathbf{\Lambda}$ is composed of orthonormal columns. The matrix below is the same as Equation 14 after rotation:

$$\boldsymbol{\Lambda}^{\star} = \begin{bmatrix} 0.206 & 0.034 & -0.232 & 0.018 \\ 0.089 & -0.314 & -0.169 & 0.043 \\ 0.149 & -0.084 & -0.104 & 0.004 \\ 0.025 & -0.178 & -0.078 & 0.121 \\ 0.16 & 0.059 & -0.144 & -0.118 \\ 0.161 & 0.015 & -0.031 & -0.138 \\ -0.037 & -0.282 & -0.012 & 0.229 \\ 0.122 & -0.052 & -0.205 & 0.075 \\ 0.079 & -0.156 & 0.057 & 0.187 \\ 0.199 & -0.007 & 0.072 & 0.108 \end{bmatrix} \tag{16}$$

In addition, we have the **D** matrix written as:

$$\mathbf{D} = \begin{bmatrix} 3016402.75 & 0 & 0 & 0 \\ 0 & 1360897.08 & 0 & 0 \\ 0 & 0 & 916670.40 & 0 \\ 0 & 0 & 0 & 764247.87 \end{bmatrix} \tag{17}$$

Once we rotated the loadings, we must also rotate the scores. This is done as follows:

$$\mathbf{f}^{\star} = (\mathbf{D}\mathbf{V}' \otimes \mathbf{I}_T)\mathbf{f} \tag{18}$$

note that **V** must also be multiplied by -1 if most of the first factor loadings (first column of **U**) are negative. The function `fa.outs` is programmed to automatically rotate the loadings and scores. We can use the following commands to access the results:

Box 17. Matrix of rotated loadings and vector (matrix) of rotated scores

```
# Rotated loadings
m5.res$rot.loads


# Rotated scores
m5.res$rot.scores
```

The scores are presented in matrix form, instead of vector, for the sake of visualization.

The rotated loadings are used in Equation 11 for obtaining the BLUPs. `fa.outs` generates

two types of BLUPs: conditional and marginal. Conditional BLUPs contain the lack of fit effects, i.e., those related to the specific variances. Marginal BLUPs are composed of the common genetic effects only. BLUPs are accessed using the following code:

Box 18. Marginal and conditional BLUPs

```
m5.res$blups
```

Model selection: number of factors

Comparing two FAMMs for adequacy using information criteria like AIC or BIC may not always reflect the ideal model, especially in FA models. This is because the relationship between explained parsimony and variance is not linear. Therefore, the models suggested by information criteria might not be sufficiently informative (ISIK et al., 2017; SMITH et al., 2015). In this context, two additional criteria can aid in the selection process. One of these is the percentage of variance explained by all K factors in the model ($\bar{v}$), given by (SMITH et al., 2015):

$$\bar{v} = \frac{\text{Trace}\left(\boldsymbol{\Lambda}^{\star}\boldsymbol{\Lambda}^{\star\prime}\right)}{\text{Trace}(\boldsymbol{\Sigma}_g)} \tag{19}$$

where Trace is the sum of diagonal elements. The second criterion is the average semivariance ratio (ASR), given as follows (CHAVES et al., 2023a; PIEPHO, 2019):

$$ASR = \frac{\frac{2}{A(A-1)}\sum_{a=1}^{A-1}\sum_{a'=a+1}^{A}\frac{1}{2}(\sum_{k=1}^{K}\lambda_{k_a}^{\star^2}d_k + \sum_{k=1}^{K}\lambda_{k_{a'}}^{\star^2}d_k) - \sum_{k=1}^{K}\lambda_{k_a}^{\star}\lambda_{k_{a'}}^{\star}d_k}{\frac{2}{A(A-1)}\sum_{a=1}^{A-1}\sum_{a'=a+1}^{A}\frac{1}{2}[(\sum_{k=1}^{K}\lambda_{k_a}^{\star^2}d_k + \psi_a) + (\sum_{k=1}^{K}\lambda_{k_{a'}}^{\star^2}d_k + \psi_{a'})] - \sum_{k=1}^{K}\lambda_{k_a}^{\star}\lambda_{k_{a'}}^{\star}d_k} \times 100 \tag{20}$$

where d_k is the k^{th} element of the diagonal of $\mathbf{D}$. The main difference between $\bar{v}$ and ASR is that the latter considers covariances in its computation, an important part of FAMMs.

Keep in mind that factors are latent variables designed to capture common variations between environments. Therefore, the number of factors required depends on the number of environments. Ideally, the goal is to select the most parsimonious model that explains the highest amount of variance. Diagnostics to assist in choosing the best FAMM model can be obtained using the following command:

Box 19. Diagnostics

```
m5.res$diagnostics
```

Fitting FAMMs with one, two and three factors, we achieved the results presented in Table 6. The model with two factors was selected, for it allies parsimony and explicative power.

Table 6: Information criteria values of each tested model: AIC and BIC

No. factors	AIC	BIC	$\bar{v}(\%)$	$ASR(\%)$
1	7235.70	7425.23	76.66	32.69
2	7229.94	7477.05	95.16	87.66
3	7231.00	7530.69	98.74	97.13

Pairwise genotypic correlations

The matrix of genotypic covariances can be converted into a matrix of genotypic correlations. This can aid in investigating the GEI dynamics in the dataset. This can be done as follows (CULLIS et al., 2010):

$$\Upsilon = \mathbf{\Phi}\mathbf{\Sigma}_g\mathbf{\Phi} \tag{21}$$

where $\mathbf{\Phi}$ is a diagonal matrix whose elements are the inverse of the square root of the diagonal elements of $\mathbf{\Sigma}_g$. The following commands can be used to both $\mathbf{\Sigma}_g$ and Υ:

Box 20. Matrix of genotypic covariances and genotypic correlations

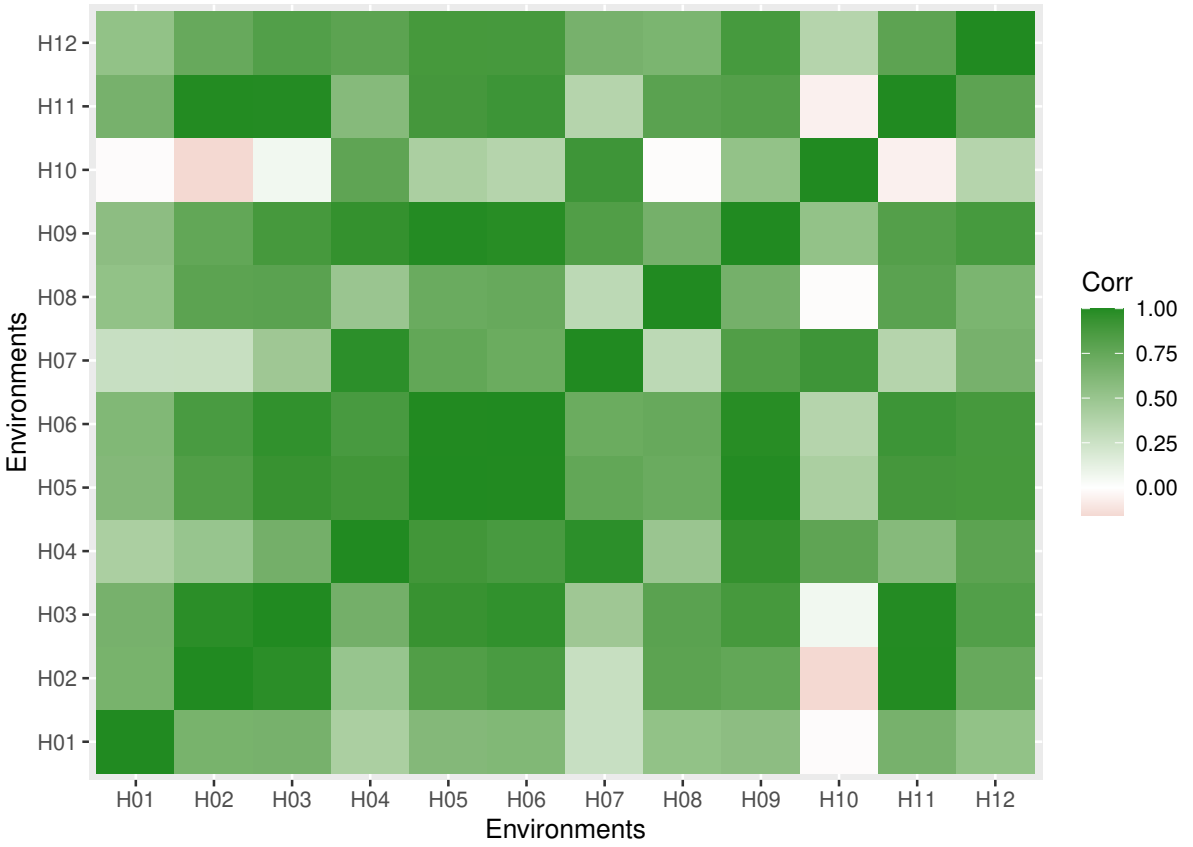
```
# Covariances
m5.res$Gvcov


# Correlations
m5.res$Gcor
```

Heatmaps are an interesting way of representing genotypic correlations (Figure 2). The following commands can be used to build a heatmap:

Box 20. Heatmap of genotypic correlations

```
as.data.frame(m5.res$Gcor) |> rownames_to_column('env_x') |>
  pivot_longer(H01:H12, names_to = "env_y", values_to = "corr") |>
  ggplot(aes(x = env_x, y = env_y, fill = corr)) +
  geom_tile() +
  scale_fill_gradient2(low = 'darkred', mid = 'white',
                       high = 'forestgreen') +
  labs(x = "Environments", y = "Environments", fill = "Corr")
```

Figure 2: Genotypic correlation between environments.



Latent regressions

Cullis et al. (2014) introduced latent regression plots to investigate genotype stability. Recall from Equation 11 that genotypic values are derived from a multiple regression model: $g_{at} = \lambda^{\star}_{1_a} f^{\star}_{1_t} + \lambda^{\star}_{2_a} f^{\star}_{2_t} + \cdots + \lambda^{\star}_{K_a} f^{\star}_{K_t} + \delta_{at}$. In this model, factor scores are analogous to regression coefficients (CULLIS et al., 2014). Therefore, if a hypothetical genotype has scores

close to 0 for all factors, it is insensitive to changes in factor loadings. As these changes signify variation between environments, the genotype is considered stable. To derive latent regressions, only the marginal genotypic values are used, without the lack-of-fit effect (δ_{at}). The number of possible plots corresponds to the number of factors considered, as each graph represents a factor. The y-axis of the graphs comprises the marginal genotypic values, while the x-axis represents the environmental variation contained in each factor, depicted by the loadings. Note that, despite representing the same parameters, the axes of the graphs differ:

Plot 1: y-axis = $\lambda^{\star}_{1_a} f^{\star}_{1_t} + \lambda^{\star}_{2_a} f^{\star}_{2_t} + \lambda^{\star}_{3_a} f^{\star}_{3_t} + \cdots + \lambda^{\star}_{K_a} f^{\star}_{K_t}$; x-axis = $\lambda^{\star}_{1_a}$

Plot 2: y-axis = $\lambda^{\star}_{2_a} f^{\star}_{2_t} + \lambda^{\star}_{3_a} f^{\star}_{3_t} + \cdots + \lambda^{\star}_{K_a} f^{\star}_{K_t}$; x-axis = $\lambda^{\star}_{2_a}$

Plot 3: y-axis = $\lambda^{\star}_{3_a} f^{\star}_{3_t} + \cdots + \lambda^{\star}_{K_a} f^{\star}_{K_t}$; x-axis = $\lambda^{\star}_{3_a}$

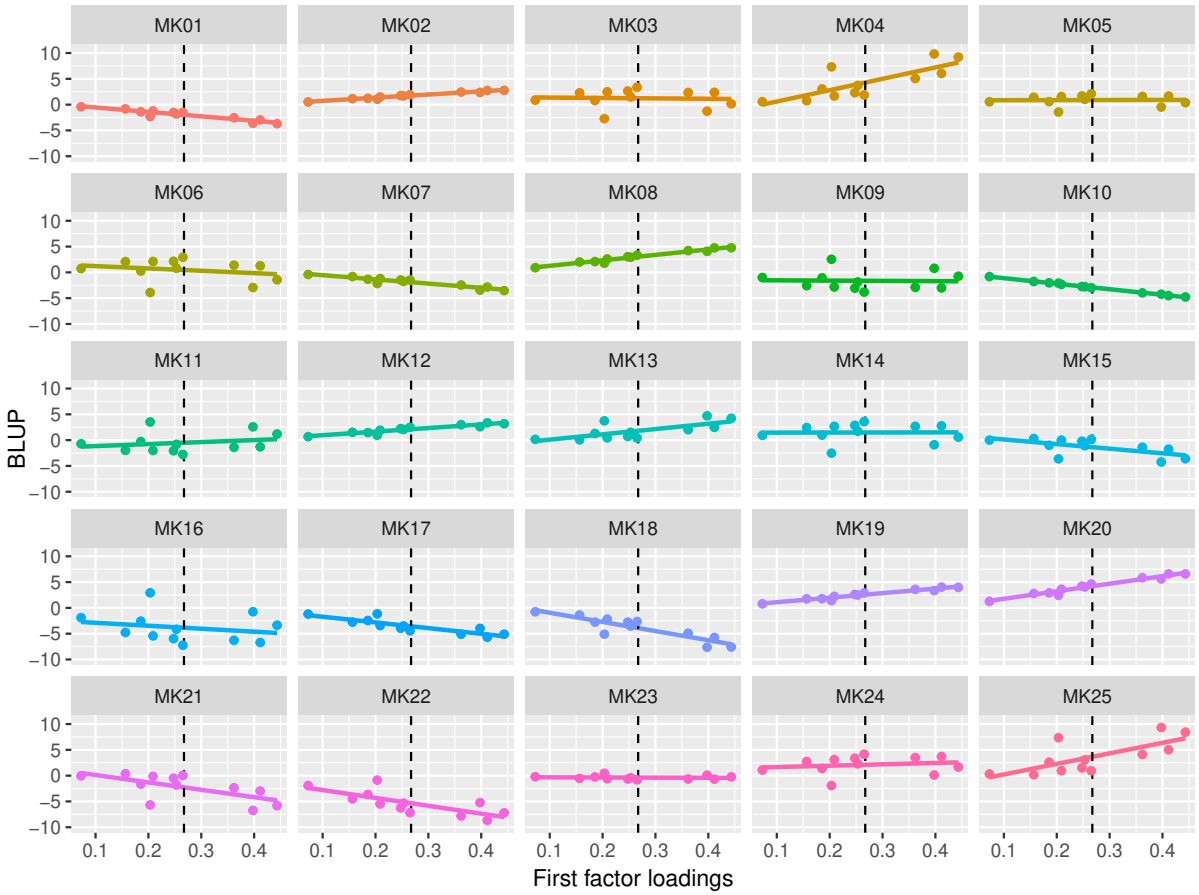
Plot K: y-axis = $\sum_{k=1}^{K-1} \lambda^{\star}_{K_a} f^{\star}_{K_t}$; x-axis = $\lambda^{\star}_{K_a}$

In the example, we obtained the plots shown in Figures 3 and 4 using the following commands:

Box 21. Latent regression plots for the first factor

```
m5.res$blups |>
  left_join(as.data.frame(m5.res$rot.loads) |>
                           rownames_to_column("env")) |>
  ggplot(aes(x = fa1, y = marginal, colour = gen)) +
  geom_smooth(method = lm, fill = NA) +
  facet_wrap(~gen) +
  geom_vline(xintercept = mean(m5.res$rot.loads[,1]),
             linetype = "dashed") +
  geom_point() +
  labs(x = "First factor loadings", y = "BLUP") +
  theme(legend.position = 'none')
```

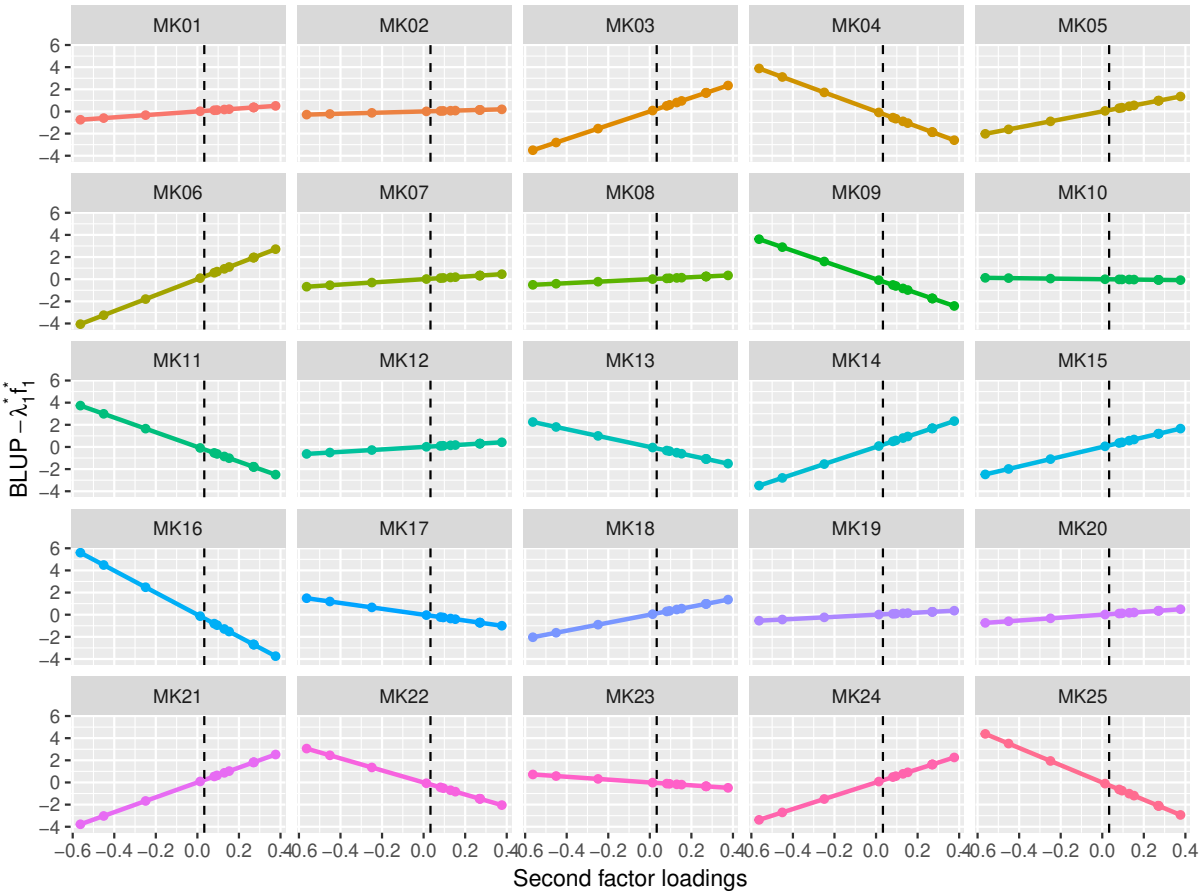
Figure 3: Latent regressions for the first factor



Box 22. Latent regression plots for the second factor

```
m5.res$blups |>
  left_join(as.data.frame(m5.res$rot.loads) |>
                          rownames_to_column("env")) |>
  mutate(marginal = marginal - kronecker(m5.res$rot.loads[,1],
         diag(nlevels(data$gen))) %*% m5.res$rot.scores[,1]) |>
  ggplot(aes(x = fa2, y = marginal, colour = gen)) +
  geom_smooth(method = lm, fill = NA) +
  facet_wrap(~gen) +
  geom_vline(xintercept = mean(m5.res$rot.loads[,2]),
             linetype = "dashed") +
  geom_point() + theme(legend.position = 'none')
  labs(x = "Second factor loadings",
       y = expression(BLUP - lambda[1]^'*' * f[1]^'*'))
```

Figure 4: Latent regressions for the second factor



Remember that the majority of genetic variation is explained by the first factor, followed by the second, and so forth. Given that most loadings on the first factor are positive, it is beneficial for genotypes to exhibit high scores for this factor, as observed with MK04 in Figure 3. Starting from the second factor, variations in the sign of the loads emerge, influencing the patterns of variation between environments. In such cases, breeders should compare latent regression plots with the loading matrix to ascertain in which environment a genotype performs optimally (CULLIS et al., 2014). Stable genotypes will exhibit less steep straight lines, as illustrated by MK10 in Figure 4. When employing kinship information—whether probabilistic or genomic—this concept can be extended to additive and dominance effects, facilitating detailed genetic analysis of how components of genetic variation behave in the population under different environmental conditions (DIAS et al., 2018).

Factor Analytic selection tools

While latent regressions offer insight into genotype stability, they may not be ideal for practical selection due to the increasing complexity as the number of genotypes and factors in the model rises, resulting in a multitude of graphs and information to interpret. Addressing this concern, Smith and Cullis (2018) introduced the "Factor Analytic selection tools" (FAST). Considering the marginal values, we can express the multiple regression model for obtaining genotypic values as $g_{at} = \lambda^{\star}_{1_a} f^{\star}_{1_t} + \epsilon_{at}$, where $\epsilon_{at} = \lambda^{\star}_{2_a} f^{\star}_{2_t} + \cdots + \lambda^{\star}_{K_a} f^{\star}_{K_t}$. Given that the first factor captures the majority of genetic variation, it can be presumed to represent the bulk of genotype performance - or the genotypic main effects (STEFANOVA; BUIRCHELL, 2010). Consequently, other factors play a role in representing the stability of the genetic materials under scrutiny. Bearing this relationship in mind, the authors proposed estimators for overall performance (OP) and stability (root mean squared deviation - $RMSD$), respectively:

$$OP_t = \frac{1}{A} \sum_{a=1}^{A} \lambda^{\star}_{1_a} f^{\star}_{1_t} \tag{22}$$

$$RMSD_t = \sqrt{\frac{1}{A} \sum_{a=1}^{A} \epsilon_{at}} \tag{23}$$

An ideotype will have a high OP and a low $RMSD$. We can associate these metrics with BLUPs' reliabilities (the squared accuracies). These three pieces of information can be illustrated in a scatter plot (Figure 5). The following commands can be used to do so:

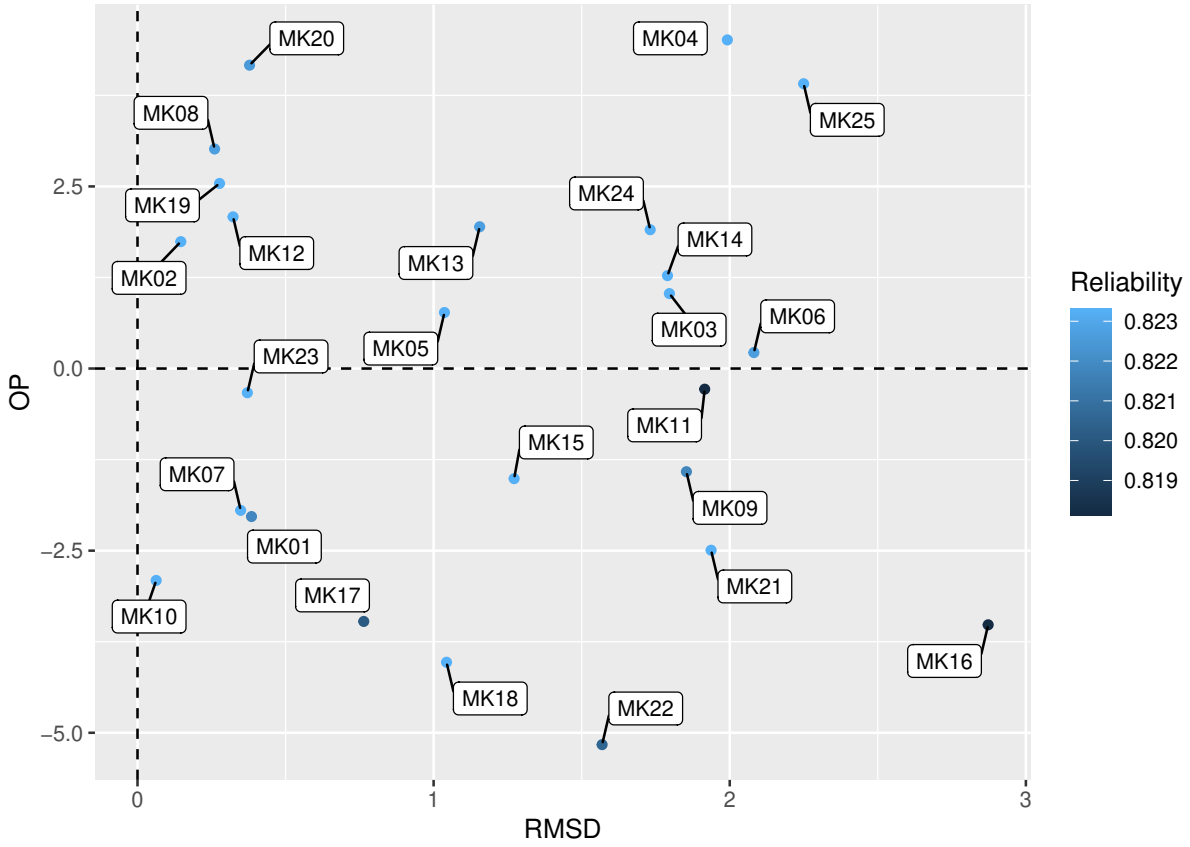
Box 23. Factor Analytic selection tools

```
FAST = data.frame(
  amb = m5.res$blups$amb,
  gen = m5.res$blups$gen,
  RMSD = (m5.res$blups$marginal -
    kronecker(m5.res$rot.loads[,1], diag(nlevels(data$gen)) %*%
                m5.res$rot.scores[,1]))^2) |>
  reframe(RMSD = sqrt(mean(RMSD)), .by = gen) |>
  mutate(OP = mean(m5.res$rot.loads[,1]) * m5.res$rot.scores[,1],
         m5.res$blups |>
           reframe(rel = 1-(mean(std.error^2)/
                    mean(diag(m5.res$Gvcov))),
         .by = gen))

FAST |> ggplot(aes(x = RMSD, y = OP)) +
  geom_point(aes(color = rel)) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "RMSD", y = "OP", color = "Reliability") +
  # For using the command below, install the package "ggrepel"
  ggrepel::geom_label_repel(aes(label = gen), size = 3,
                            box.padding = .5)
```

In this graph, genotypes positioned close to zero horizontally and distant from zero vertically are considered the most favourable. The selection thresholds can be customized by the breeder, such as defining a specific area on the graph where only genotypes within this region are chosen. Alternatively, an index incorporating both OP and RMSD can be constructed for selection purposes. For a more in-depth exploration, refer to Chaves et al. (2023a) and Chaves et al. (2023b).

Research on the use and enhancement of FAMM is continually appearing in the scientific literature. These models can be seamlessly integrated with genomic information (DIAS et al., 2018; TOLHURST et al., 2019) and/or environmental covariates (ARAÚJO et al., 2024;

Figure 5: Scatter plot illustrating the overall performance (y-axis), stability (root mean squared deviation - RMSD, x-axis) and the reliability of the assessed candidates



CALLISTER et al., 2024; TOLHURST et al., 2022), providing a comprehensive exploration of the genotypic-environment interaction.

4 Spatial Analysis

The purpose of randomization is to prevent favouritism towards certain treatments, ensuring that errors are independent and resulting in a residual covariance matrix of $\sigma_e^2\mathbf{I}$. However, in perennial plant breeding, experiments often take place in heterogeneous areas, leading to potential spatial dependence and an increase in residual variation (GEZAN et al., 2010).

Once trials are established, they are susceptible to various uncontrollable effects, whether environmental or anthropogenic. Consequently, two types of residual effects are identified: natural effects related to soil fertility gradients, topography, humidity, etc.; and global effects caused by experiment management (GILMOUR et al., 1997). Ideally, natural residual effects should be preemptively addressed through randomization. However, if identification occurs

during the experiment, *a priori* actions may not fully nullify these effects (RESENDE et al., 2014). In such cases, it becomes essential to merge the fundamental principles of experimentation with *a posteriori* statistical methods capable of mitigating the impact of spatial trends, thereby reducing biases in model-based inferences. This section explores some of these methods.

To identify spatial trends, breeders need the trial grid. Two columns denoting the position (row and column) of each plot are added to the dataset using this grid information. Spatial trends can then be visualized through maps and dispersion diagrams of residuals based on plot positions, commonly referred to as variograms.

In the context of a spatial trend, a variable is no longer considered random but regionalized (RESENDE et al., 2014). While random variables exhibit well-distributed values across the field, the values of regionalized variables depend on the distance between plots. In other words, the correlation between two plots diminishes as the distance between them increases (e.g., $p1$ is more related to $p2$ than to $p20$). This behaviour characterizes spatial autocorrelation, which can manifest in the row, column, or both directions. The presence of autocorrelation heightens the risk of type I and II errors (DUARTE; VENCOVSKY, 2005).

Various *a posteriori* methods have been proposed to address spatial trends, including neighbourhood analyses using the method of Papadakis (1937) or moving averages (HAINING, 1978), spline curve fitting, incorporating row and column effects as random components in the model, and employing methods rooted in time series (GEZAN et al., 2010; GILMOUR et al., 1997; VELAZCO et al., 2017; VERBYLA et al., 1999). This section will predominantly focus on first-order autoregressive models (AR1), as popularized by Gilmour et al. (1997). Subsequently, we will walk through the step-by-step implementation of the analysis proposed by these authors.

4.1 Linear mixed models with autoregressive residual adjustment

As briefly described above, residual independence implies, in biometric terms:

$$\mathbf{R} = \sigma_e^2 \mathbf{I} = \sigma_e^2 \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \tag{24}$$

considering that each row/column of the identity matrix above represents a plot, it is assumed that there is no residual covariance between them.

Autoregressive models consider the presence of spatial autocorrelation in the trial. This autocorrelation can be present in one or both directions (row or column). The separation of directions when fitting spatial trends allows autocorrelations of different magnitudes to exist between rows and columns. In this case, the residual covariance matrix will be:

$$\mathbf{R} = \sigma_e^2 \mathbf{\Sigma}_r(\rho_r) \otimes \mathbf{\Sigma}_c(\rho_c) \tag{25}$$

where:

$$\mathbf{\Sigma}_r(\rho_r) = \begin{bmatrix} 1 & \rho_r & \rho_r^2 & \dots & \rho_r^{R-1} \\ \rho_r & 1 & \rho_r & \dots & \rho_r^{R-2} \\ \rho_r^2 & \rho_r & 1 & \dots & \rho_r^{R-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_r^{R-1} & \rho_r^{R-2} & \rho_r^{R-3} & \dots & 1 \end{bmatrix} \tag{26}$$

$$\mathbf{\Sigma}_c(\rho_c) = \begin{bmatrix} 1 & \rho_c & \rho_c^2 & \dots & \rho_c^{C-1} \\ \rho_c & 1 & \rho_c & \dots & \rho_c^{C-2} \\ \rho_c^2 & \rho_c & 1 & \dots & \rho_c^{C-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_c^{C-1} & \rho_c^{C-2} & \rho_c^{C-3} & \dots & 1 \end{bmatrix} \tag{27}$$

where *R* and *C* are the number of rows and columns, respectively; and *rho* is the autocorrelation coefficient. This coefficient decreases as plots grow apart.

Using AR1 models, Burgueño (2018) proposed a series of steps for spatial analysis in plant breeding, based on Gilmour et al. (1997):

1. Begin with data analysis using models without spatial adjustment to test assumptions like normality, homoscedasticity, and independence of residues. If there's no evident dependence, proceed to the second step.

2. Adjust residual effects using the AR1 structure. In this stage, breeders should initially adjust in one direction (row or column) and then in both directions $[\mathbf{\Sigma}_r(\rho_r) \otimes \mathbf{\Sigma}_c(\rho_c)]$.

3. Move on to the third step, which involves analysing the variograms of each model. In these graphs, a flat surface represents homogeneity in the experimental field. The curves on the surfaces indicate variations, and the greater the number and size of these curves, the more pronounced the spatial trend. If this is observed, proceed to the fourth step.

4. Check for the presence of outliers and/or add more effects to the model. Options include incorporating row and column effects as either fixed or random effects, introducing covariates, or including the nugget residual effect, which is independent of spatial trends. In this step, breeders can leverage practical experience, considering aspects like machinery direction in the field, irrigation practices, fertilization, etc. After incorporating these effects, revisit the third step to visualize the variogram.

Information criteria such as AIC and BIC should not be used to ascertain model fit, as models with different fixed effects are not readily comparable. To overcome this problem, Verbyla (2019) proposed an adjustment to both AIC (AIC_c) and BIC (BIC_c) that allows direct comparison between models with different fixed or random effects:

$$AIC_c = -2LogL_c + 2(p+q) \tag{28}$$

$$BIC_c = -2LogL_c + (p+q)Log(n) \tag{29}$$

where $LogL_c$ is the maximum point of the full likelihood function, p is the number of fixed effects, q is the number of random effects, and n is the number of observations. The idea of Verbyla (2019) allows AIC_c and BIC_c to be used for comparing models with different fixed effects. The author ably built a function in `R` to compute these metrics. This function was slightly modified by us and is available at `https://raw.githubusercontent.com/saulo-chaves/May_b_useful/main/icreml.R`.

We will do the spatial analysis using the environment "H05" of the dataset. Note that both row and column information are present in the dataset. Without it, it would be impossible to fit

a spatial model.

Box 24. Dataset for spatial analysis and data ordering

```
data_h05 = data[data$yr == "H05",
                c('gen', 'block', 'fy', 'row', "col")]
data_h05 = transform(data_h05, gen = as.factor(gen),
                               block = as.factor(block),
                               row = as.factor(row),
                               col = as.factor(col))
data_h05 = data_h05[order(data_h05$col, data_h05$row),]
```

Two details are essential for spatial analysis: the declaration of row and column effects as factors, and the reorganization of the data set in ascending order of columns and rows (or rows and columns, depending on how it is declared in the model). Model 1 is the same as shown in Equation 1. From Model 1, 5 spatial models were fitted (Table 7).

Table 7: Fixed and random effects of the spatial models

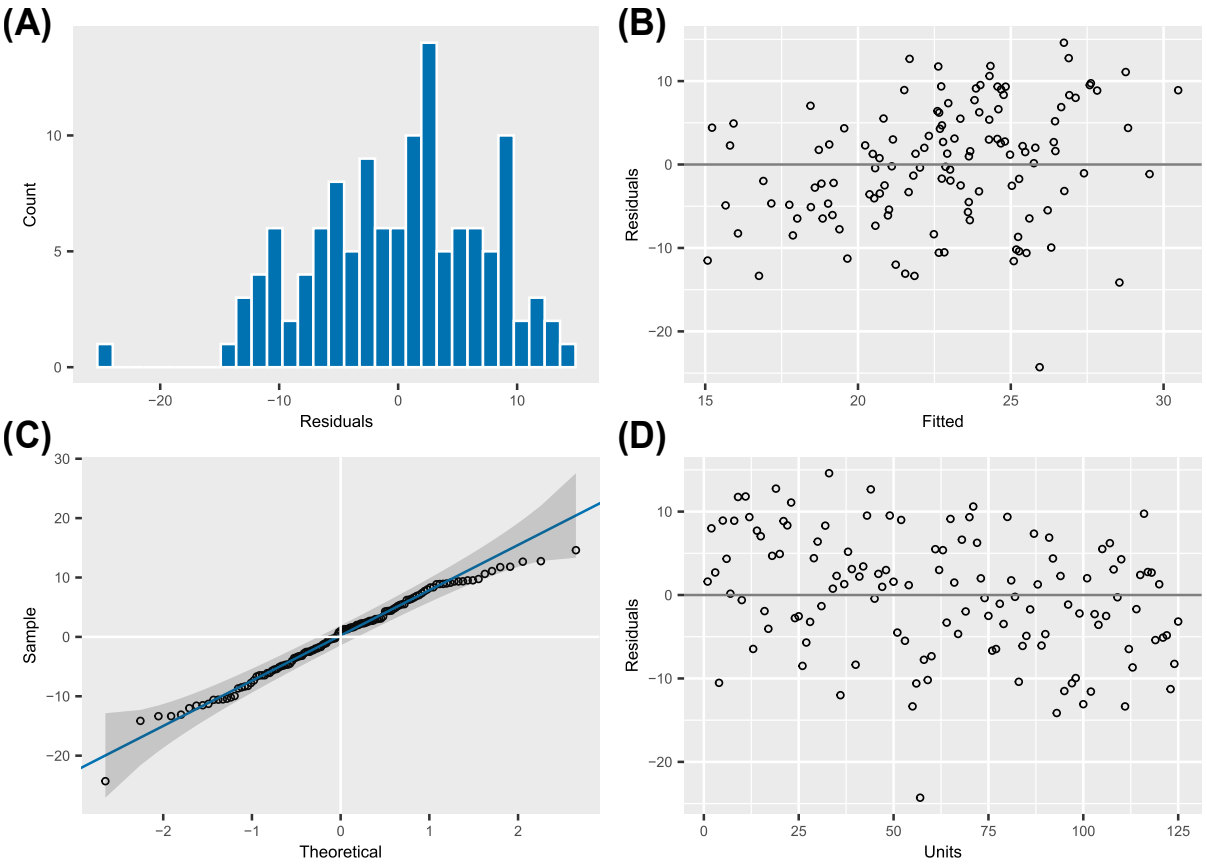
Model	Fixed terms	Random terms	
		Non-residual	Residual
1	Block	Genotype	$\sigma_e^2\mathbf{I}$
2	Block	Genotype	$\sigma_e^2\mathbf{I}_r \otimes \mathbf{\Sigma}_c(\rho_c)$
3	Block	Genotype	$\sigma_e^2\mathbf{\Sigma}_r(\rho_r) \otimes \mathbf{\Sigma}_c(\rho_c)$
4	Block	Genotype	$\sigma_e^2\mathbf{\Sigma}_r(\rho_r) \otimes \mathbf{\Sigma}_c(\rho_c)$ $\xi =$ Nugget effect
5	Block	Genotype Column	$\sigma_e^2\mathbf{\Sigma}_r(\rho_r) \otimes \mathbf{\Sigma}_c(\rho_c)$ $\xi =$ Nugget effect
6	Block Column (linear)	Genotype Column	$\sigma_e^2\mathbf{\Sigma}_r(\rho_r) \otimes \mathbf{\Sigma}_c(\rho_c)$ $\xi =$ Nugget effect

Model 1 is fitted using the codes shown in Box 25. From this model, we can use graphical tools to investigate if spatial trends exist (Figure 6)

Box 25. Visual residual analysis

```
m1 = asreml(fixed = fy ~ block,
            random = ~ gen,
            data = data_h05)
plot(m1)
```

Figure 6: Model 1 residual analysis: Histogram depicting residual distribution (A), error vs fitted value relationship to observe homoscedasticity (B), QQ-plot to observe adherence to normality (C), and error vs plots relationship for observation of spatial trends (D).

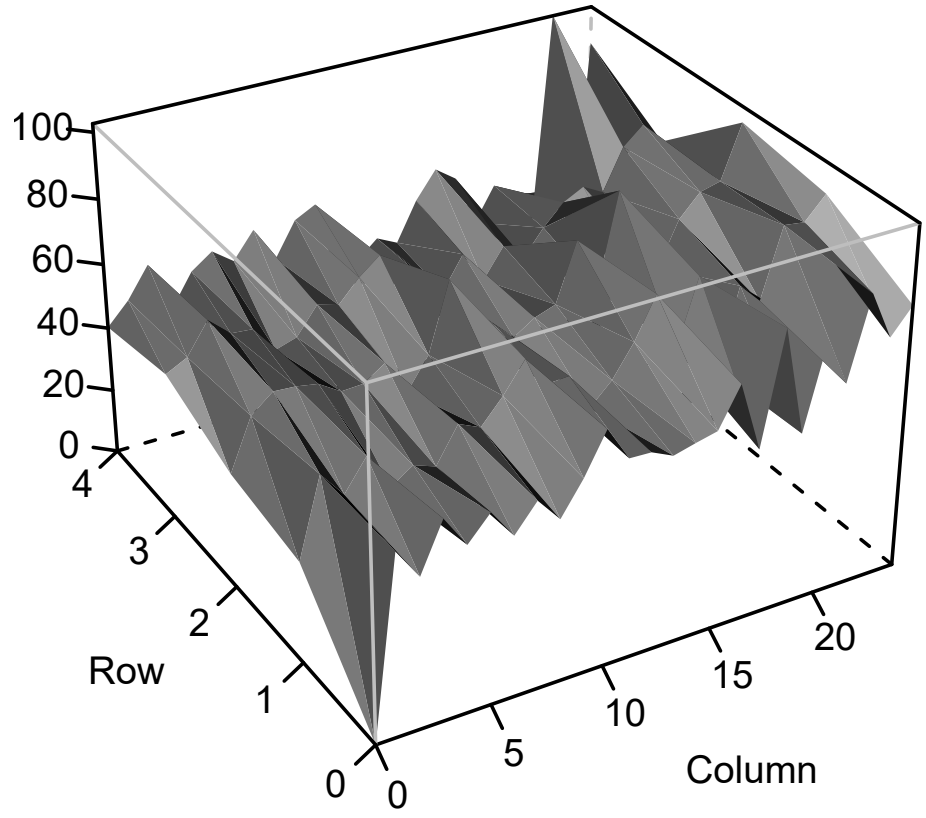


Graphical analyses of residuals offer insights into normality, homoscedasticity, and residual independence. Figure 6D helps identify potential spatial trends. In an ideal scenario of independent residuals, points should be uniformly distributed around the central line. However, a noticeable decrease in error as plots change suggests a need to investigate the presence of spatial trends. Consequently, we apply a second model with autoregressive adjustment in the column direction. The codes for adjusting this model, along with obtaining the corresponding variogram (Figure 7), are:

Box 26. Model 2 with autoregressive adjustment in one direction (column)

```
m2 = asreml(fixed = fy ~ block,
            random = ~ gen,
            residual = ~ar1(col):id(row),
            data = data_h05)
plot(varioGram(m2))
```

Figure 7: Variogram obtained after fitting model 2



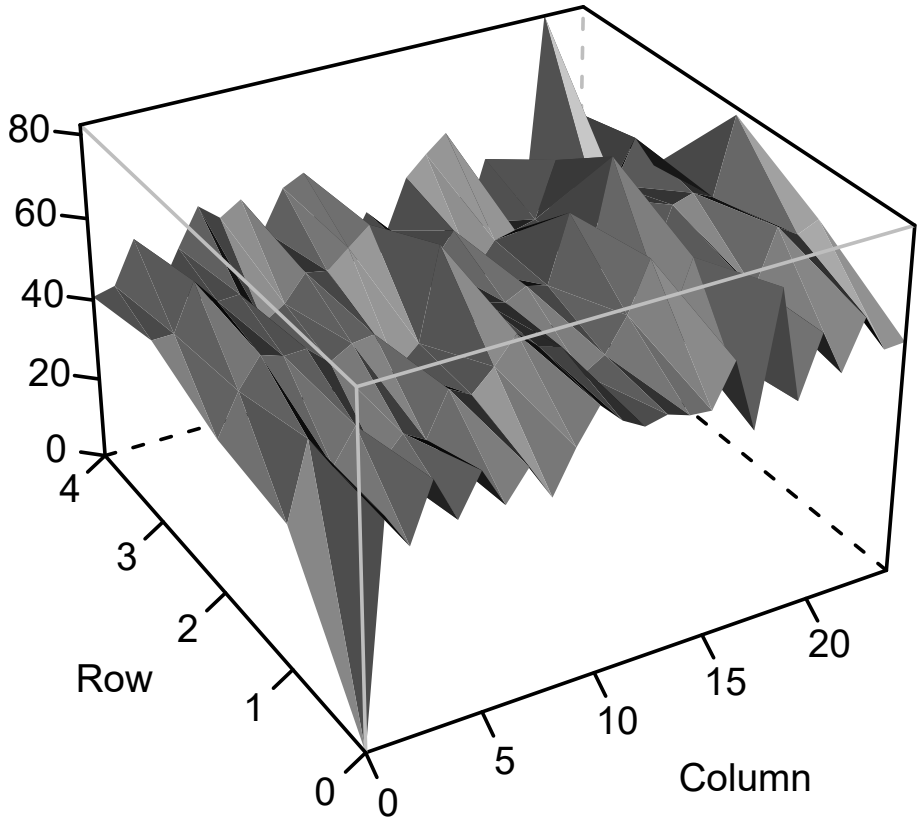
From Figure 7, we observe that the variogram surface is not flat, indicating non-uniform residual effects across the experimental area. Consequently, let's test an autoregressive model in both directions and visualize its variogram (Figure 8):

Box 27. Model 3 with autoregressive adjustment in both directions

```
m3 = asreml(fixed = fy ~ block,
            random = ~ gen,
            residual = ~ar1(col):ar1(row),
            data = data_h05, maxit = 50)
plot(varioGram(m3))
```

In the fourth model, we aim to smooth spatial trends using the nugget residual effect, which is independent of correlated residual effects. The following commands are used to adjust the model and obtain the variogram (Figure 9):

Figure 8: Variogram obtained after fitting model 3

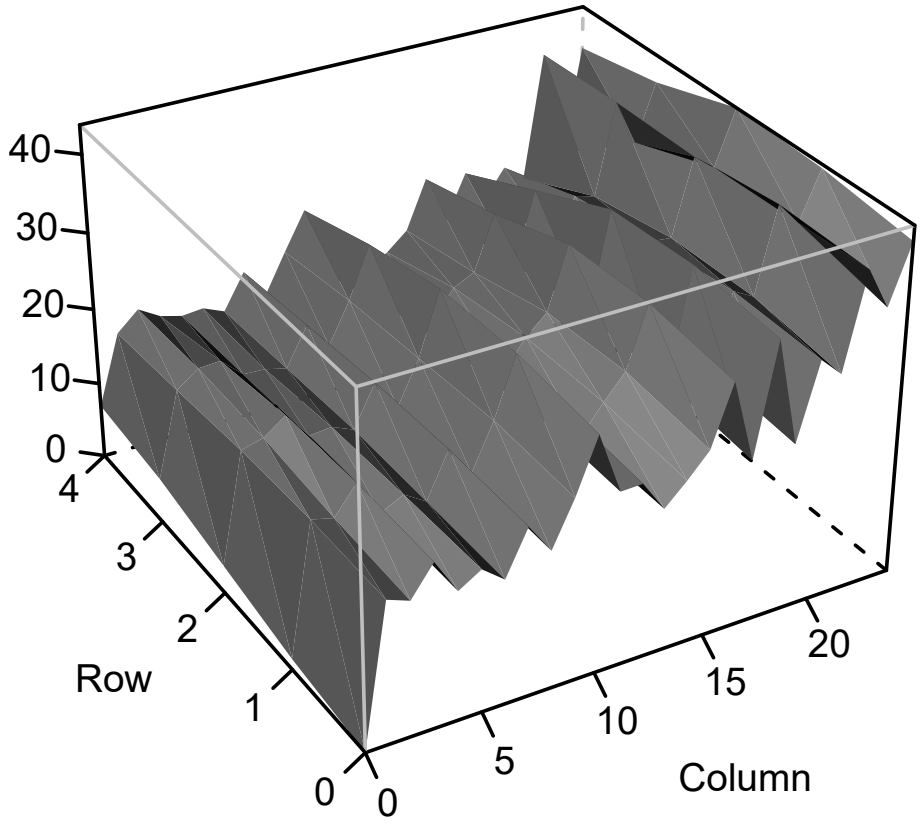


Box 28. Model 4 with autoregressive adjustment in both directions and nugget effect

```
m4 = asreml(fixed = fy ~ block,
            random = ~ gen + units,
            residual = ~ar1(col):ar1(row),
            data = data_h05, maxit = 50)
plot(varioGram(m4))
```

Note that, indeed, there is some smoothing on the surface of the variogram in Figure 9. However, it is possible to observe a strong tendency towards the columns. We will try to account for these trends by adding random column effects in model 5:

Figure 9: Variogram obtained after fitting model 4

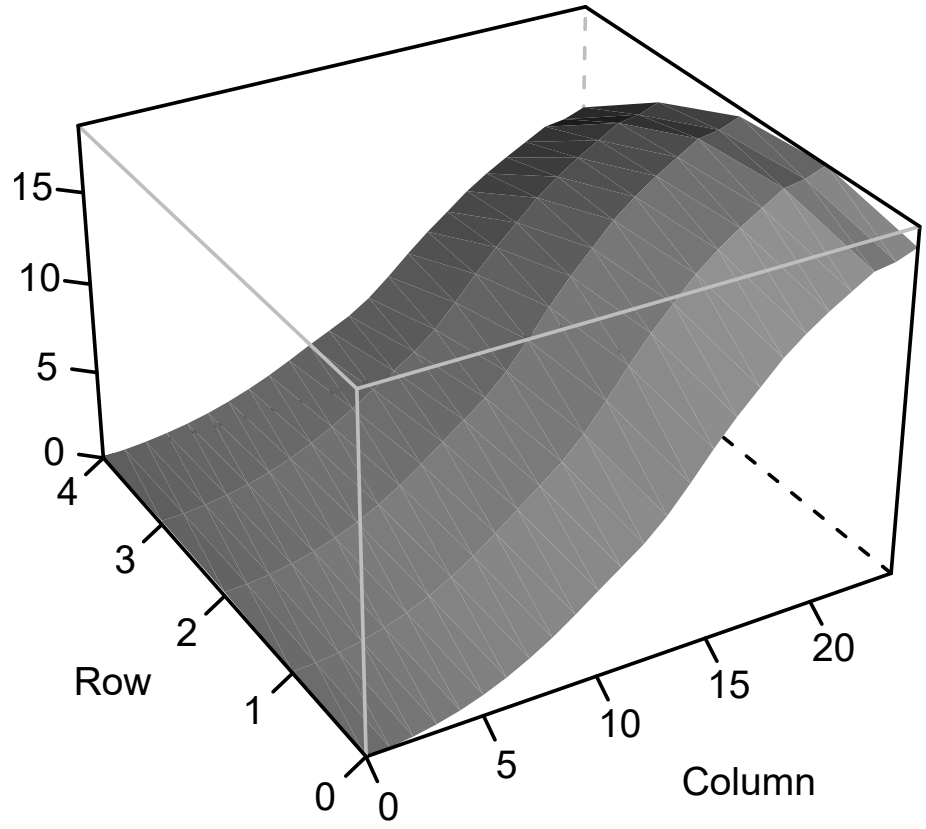


Box 29. Model 5 with autoregressive adjustment in both directions; nugget effect and random column effect

```
m5 = asreml(fixed = fy ~ block,
            random = ~ gen + units + col,
            residual = ~ar1(col):ar1(row),
            data = data_h05, maxit = 50)
plot(varioGram(m5))
```

The smoothing of the variogram surface in Figure 10 indicates progress. There seem to be no more issues with spatial trends in the row direction. However, the residuals still show an increase towards the columns. We will attempt to mitigate this effect by adding a linear column effect to the fixed part of the model:

Figure 10: Variogram obtained after fitting model 5

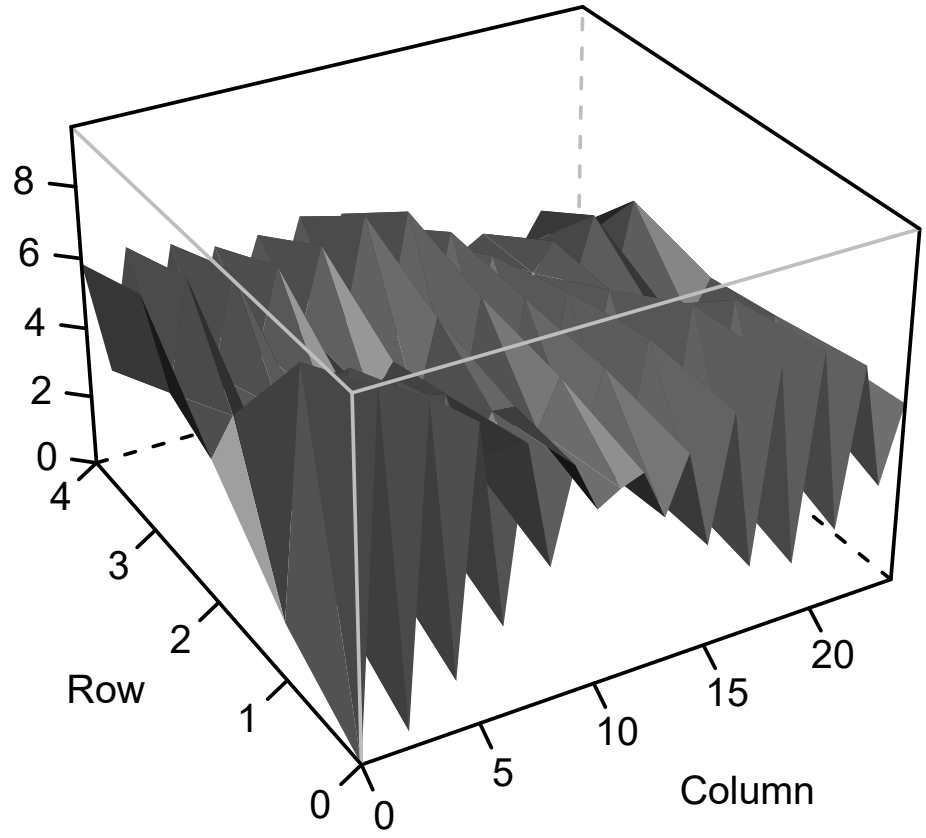


Box 30. Model 6 with autoregressive adjustment in both directions; nugget effect; random column effect and fixed linear column effect

```
m6 = asreml(fixed = fy ~ block + lin(col),
            random = ~ gen + units + col,
            residual = ~ar1(col):ar1(row),
            data = data_h05, maxit = 50)
plot(varioGram(m6))
```

Note the differences in the scales of the vertical axes of the variograms of model 2 (Figure 7, where the maximum axis limit is 80) and model 6 (Figure 11, where the maximum axis limit is 8). This reduction indicates that the adjustments made were effective in correcting the spatial trends of the experimental field. This is further supported by comparing the values of the corrected information criteria for each model (Table 8). Model 6 exhibits the lowest values for both AIC_c and BIC_c among the adjusted models, underscoring its suitability. Corrected information criteria are obtained using the following command:

Figure 11: Variogram obtained after fitting model 6



Box 31. Corrected information criteria

```
source("https://raw.githubusercontent.com/saulo-chaves/May_b_useful/
        main/icreml.R")
Cic = icREML(fm = list(m1 = m1, m2 = m2, m3 = m3,
                       m4 = m4, m5 = m5, m6 = m6))
```

Table 8: Model selection for spatial trends correction: LogL is the logarithm of the maximum point of the full likelihood function, p is the number of fixed effects, q is the number of random effects, AIC_c is the conditional Akaike information criterion and BIC_c is the conditional Bayesian information criterion. Both criteria were derived by Verbyla (2019)

Modelo	LogL	p	q	AIC_c	BIC_c
1	-326.90	5	2	667.80	687.60
2	-326.95	5	3	669.90	692.53
3	-319.98	5	4	657.95	683.41
4	-318.46	5	5	656.92	685.21
5	-321.84	5	5	663.67	691.96
6	-312.08	6	6	648.16	682.10

Let's see the variance component estimates of the selected model (Table 9) using the

following command:

Box 32. Variance components of the best-fitted model

```
summary(m6)$varcomp
```

Table 9: Estimates of variance components by model 6. σ^2_g is the genotypic variance, σ^2_c is the variance of column effects, σ^2_ξ is the variance of uncorrelated residual effects, σ^2_η is the variance of correlated residual effects, ρ_c is the column-wise autocorrelation coefficient, andρ_r is the row-wise autocorrelation coefficient.

Component	Estimate	Standard error
σ^2_g	15.52	7.13
σ^2_c	11.19	7.18
σ^2_ξ	32.26	7.36
σ^2_η	10.56	8.30
ρ_c	-0.83	0.22
ρ_r	0.56	0.5

Several studies demonstrated the effectiveness of spatial analysis (BERNARDELI et al., 2021; GOGEL et al., 2018; SMITH et al., 2001). Nevertheless, spatial analysis has not been universally adopted as a standard procedure in most perennial plant breeding programs.

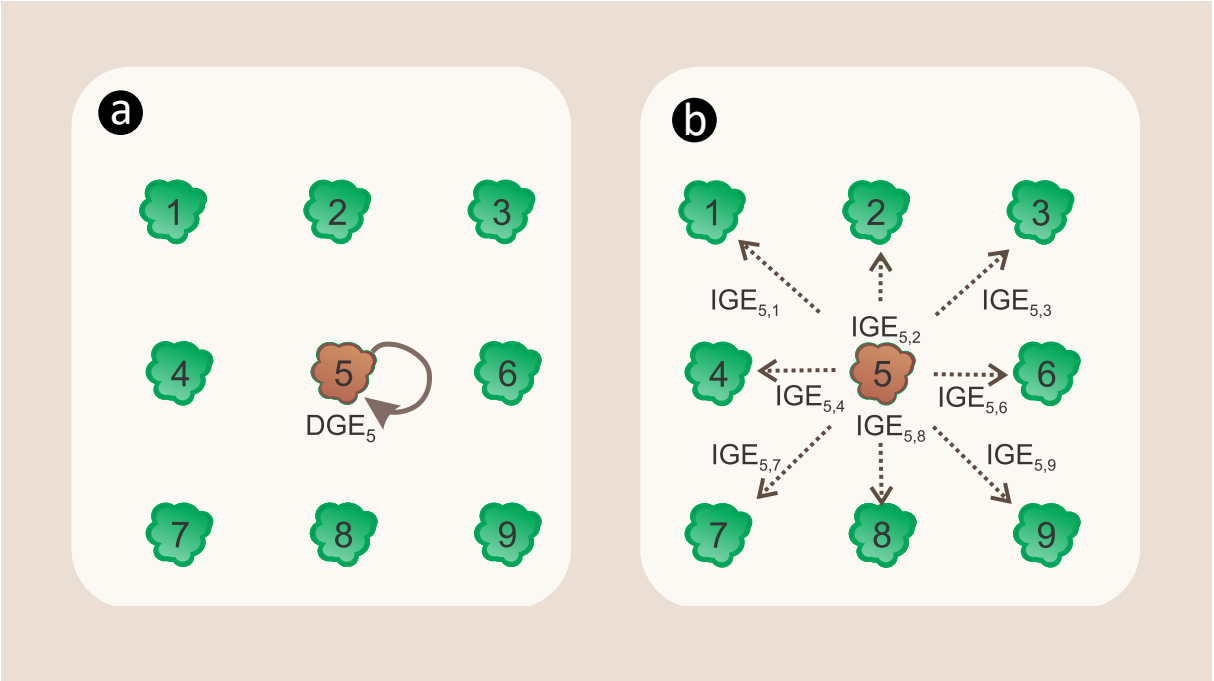
5 Competition analysis

The variogram in Figure 11 and the auto-correlation coefficient in the column direction in Table 9 highlight an antagonistic effect between plots along the columns, suggesting competition among plots. Competition, a common interaction between plants in experiments, becomes more pronounced when resources like water, nutrients, or light are limited (POMMERENING; MEADOR, 2018). In forestry, competition effects are often addressed at the individual level through indices aiming to quantify and correct phenotypic values (BURKHART; TOMÉ, 2012; LEDERMANN; STAGE, 2001). However, this approach overlooks the genetic factor influencing the ability to compete, which could be explored in breeding programs (GRIFFING, 1967).

In genetic-statistical models considering the competition, the genotypic value can be dissected into direct and indirect components (Figure 12). The direct genotypic effect (DGE) pertains to the inherent performance of the focal tree (tree 5 in Figure 12), expressed in its own

phenotype. On the other hand, the indirect genotypic effects (IGE) are linked to the impact of the focal tree on the genotypic value of neighbours, reflected in their phenotypes (BIJMA, 2014; RESENDE et al., 2014; WALSH; LYNCH, 2018). For growth traits, a negative covariance typically exists between DGE and IGE. Taking diameter at breast height (DBH) as an example, a focal tree with high DGE and a high, negative IGE is likely to have a high DBH and be surrounded by neighbours with low DBH. Consequently, the more negative the IGE of a tree, the more aggressive it is toward its neighbours. IGE can be used to categorize genotypes as aggressive, homeostatic (unresponsive to competition), or sensitive (FERREIRA et al., 2023). The direction and magnitude of covariance between DGE and IGE depend on the trait's nature, genetic relationships among neighbours, resource availability, and planting density (BIJMA, 2014).

Figure 12: Partition of total genotypic effects considering competition: direct genotypic effect (DGE, in a) taken as the performance per se of focal tree 5; and indirect genotypic effect (IGE, in b), given by the influence of focal tree 5 on its neighbours. This figure was adapted from Ferreira et al. (2023).



Relying solely on DGE for selection might lead to errors that can compromise the realized genetic gain. This is because aggressive trees may have an inflated genotypic value due to their competition with less aggressive trees. To address this, Muir (2005) proposes, in the context of a linear mixed model, that selection should be based on an index considering both DGE and IGE. Cappa and Cantet (2008) extended this idea and formulated an index as the sum of DGE

and IGE, weighted by the competition intensity factor (CIF). CIF is a function of the distance from the focal tree to its neighbours and the number of neighbours surrounding it. However, their theory assumes equal distances between rows and columns, which may not reflect actual field conditions. Costa e Silva and Kerr (2013) adjusted the estimators to accommodate different distances. These estimators are:

$$f_{D_v} = \frac{p}{\sqrt{(n_{R_v}p^4) + (n_{R_v}p^2) + (n_{C_v}p^2) + (n_{D_v}p^2) + n_{C_v}}} \tag{30}$$

$$f_{R_v} = f_{D_v}\sqrt{1+p^2} \tag{31}$$

$$f_{C_v} = \frac{f_{D_v}\sqrt{1+p^2}}{p} \tag{32}$$

where p is the inter-row and inter-column distances ratio, n_{R_v}, n_{C_v} and n_{D_v} are the number of neighbours of the focal tree v in the row, column and diagonal directions, respectively; and f_{R_v}, f_{C_v} and f_{D_v} are the CIFs in the row, column and diagonal directions, respectively. The CIFs are used to compute the total genotypic value (TGV). This metric considers both DGE and IGE, and is the most appropriate value to base the selection when competition is significant. For the t-th genotype, the TGV is given by:

$$TGV_t = DGE_t + (\overline{n_D}\overline{f_D} + \overline{n_R}\overline{f_R} + \overline{n_C}\overline{f_C}) \times IGE_t \tag{33}$$

where $(\overline{n_D}\overline{f_D} + \overline{n_R}\overline{f_R} + \overline{n_C}\overline{f_C})$ is the overall competition intensity factor. The genetic relationship between neighbours can alter the competition dynamics, as demonstrated by Costa e Silva et al. (2013) and Cappa et al. (2017)

From a practical point of view, how to fit a genetic competition model? We demonstrate it using a toy example adapted from Ferreira et al. (2023). Our starting point will be Equation 1. From this model, we can divide **g** into DGE and IGE. By doing this, we come up with a new model:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_1 + \mathbf{Z}_1\mathbf{d} + \mathbf{Z}_2\mathbf{c} + \mathbf{e} \tag{34}$$

where **d** is the DGE and **c** is the vector of IGE. Now, consider the experiment in Table 10. It is an experiment designed in randomized complete blocks, with six treatments, two replications and one plant per plot. The spacing between lines is 2.5 m and between columns, 3 m. The incidence matrices $\mathbf{X}_1$ and $\mathbf{Z}_1$ of Equation 34 are built as follows:

Table 10: Grid of a fictitious experiment to illustrate the construction of incidence matrices for a genetic competition model. The number between parentheses represents the plot number, and dictates the order in which data was collected

Rows	Column			
	1	2	3	4
1	G4 (1)	G5 (2)	G5 (3)	G7 (4)
2	G6 (8)	G7 (7)	G9 (6)	G8 (5)
3	G8 (9)	G9 (10)	G4 (11)	G6 (12)
	Block 1		Block 2	

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}; \quad \mathbf{Z}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{35}$$

The competition matrix, $\mathbf{Z}_2$, is built using the competition intensity factors previously described. For example, for genotype G7, located in the second row and second column:

$$p = \frac{3}{2.5} = 1.2 \tag{36}$$

$$f_{D_{G7}} = \frac{1.2}{\sqrt{(2\times 1.2^4) + (2\times 1.2^2) + (2\times 1.2^2) + (4\times 1.2^2) + 2}} = 0.29 \tag{37}$$

$$f_{R_{G7}} = 0.29\sqrt{1 + 1.2^2} = 0.45 \tag{38}$$

$$f_{C_{G7}} = \frac{0.29\sqrt{1 + 1.2^2}}{1.2} = 0.38 \tag{39}$$

$$\mathbf{Z}_2 = \begin{bmatrix} 0 & 0.69 & 0.57 & 0.44 & 0 & 0 \\ 0.53 & 0.53 & 0.34 & 0.44 & 0 & 0.34 \\ 0 & 0.53 & 0 & 0.53 & 0.34 & 0.44 \\ 0 & 0.69 & 0 & 0 & 0.57 & 0.44 \\ 0.36 & 0.36 & 0.47 & 0.47 & 0 & 0.56 \\ \mathbf{0.38} & \mathbf{0.38} & \mathbf{0.29} & \mathbf{0.45} & \mathbf{0.45} & \mathbf{0.29} \\ 0.29 & 0.38 & 0.45 & 0 & 0.29 & 0.45 \\ 0.47 & 0.36 & 0 & 0.56 & 0.47 & 0.36 \\ 0 & 0 & 0.57 & 0.44 & 0 & 0.69 \\ 0.53 & 0 & 0.34 & 044 & 0.53 & 0.34 \\ 0 & 0 & 0.53 & 0.34 & 0.34 & 0.53 \\ 0.69 & 0 & 0 & 0 & 0.57 & 0.44 \end{bmatrix} \tag{40}$$

The line corresponding to the calculated portion is in bold in the matrix. Once the matrices are constructed, the model is solved conventionally, using REML/BLUP. In `ASReml-R`, the model can be fitted using the following code:

Box 33. Genetic competition model

```
data = cbind(Z2, data)
m7 = asreml(fixed = y ~ block,
            random = ~ gen + grp(g1),
            group = list(g1 = 1:ncol(Z2)),
            data = data, maxit = 50)
```

Note that, before fitting the model, $\mathbf{Z}_2$ must be incorporated into the data frame (`data = cbind(Z2, data)`). Both the DGE and IGE are obtained from this model, and they are used to compute the TGV (Table 11)

More information on genetic competition models can be found in Walsh and Lynch (2018, chapter 22), Resende et al. (2014, chapter 11), Ferreira et al. (2023), Bijma (2014) and Muir (2005).

Table 11: Hypothetical results of the analysis of the example trial using a genetic competition model. The TGV was calculated considering a CIF of 1.5

Genotype	DGE	IGE	TGV
G4	0.824	-0.254	0.442
G5	0.240	-0.500	-0.510
G6	-0.548	0.825	0.690
G7	-0.047	-0.167	-0.298
G8	0.048	0.208	0.361
G9	0.911	-0.914	-0.460

6 The evolution of linear mixed models

Linear mixed models (LMMs) for normally distributed variables were pioneered by Henderson (1975), employing BLUP for random effects and estimating variance components through REML (PATTERSON; THOMPSON, 1971). When dealing with unknown variance components, the estimators are inserted into the Henderson mixed model equations, yielding a non-linear estimator for the mean parameter. Generalized linear models (GLMs), introduced by Nelder and Wedderburn (1972) for handling discontinuous variables, were extended to generalized linear mixed models (GLMMs) by Thompson and Baker (1981). Lee and Nelder (1996) broadened the BLUP approach to hierarchical generalized linear mixed models (HGLMMs), a class of statistical models incorporating random effects based on the quasi-likelihood methods of Nelder and Pregibon (1987) (RESENDE et al., 2018).

GLMMs assume that residuals might not follow a normal distribution, yet other random effects in the model are assumed to be normally distributed. However, this assumption may not always be suitable. Hierarchical generalized linear mixed models (HGLMMs), extensively described by Lee et al. (2006), allow the specification of a probability distribution and a link function for each random factor in the model. Lee and Ha (2010) presented a BLUP predictor for HGLMMs. For non-normally distributed data and HGLMMs, the linear BLUP might not be as efficient. In addressing this, the authors proposed a combination of BLUP with Tweedie dispersion models based on the Exponential distribution (RESENDE et al., 2018).

Gianola and Fernando (1986) proposed Bayesian estimation for random genetic evaluation models. In this approach, aside from the distributions commonly adopted for random effects in the classical linear model and the likelihood of the observation vector (y), assigning a priori distributions for systematic effects (fixed effects under a frequentist approach) and variance components becomes necessary. Utilizing non-informative or uniform *a priori* distributions for

systematic effects and variance components characterizes vague *a priori* knowledge about these effects and components (RESENDE et al., 2018).

The estimation of fixed and random effects in the frequentist model using the Bayesian approach can proceed as long as non-informative prior distributions are assigned for systematic effects, a Normal prior distribution is employed for random effects, and a Normal likelihood is assumed for the observation vector (RESENDE et al., 2018). Employing non-informative prior distributions for fixed effects and variance components, the modes of the posterior marginal distributions of variance components correspond to estimates obtained via REML (RESENDE et al., 2018). On the other hand, non-Bayesian estimation of random models, yielding results comparable to those obtained via Bayesian methods, can be achieved through hierarchical likelihoods under HGLMMs, sometimes offering computational advantages such as shorter processing time and a straightforward convergence criterion (RESENDE et al., 2018). HGLMMs can be fitted using hierarchical likelihoods, an extension of the joint likelihood used by Henderson. These likelihoods constitute a joint density for observations and random effects. Estimates for fixed and random effects are derived from hierarchical likelihood maximization, providing direct extensions of the Henderson mixed model equations. Variance components are estimated by maximizing the adjusted hierarchical likelihood profile, a direct extension of REML. Lee and Nelder (1996) thus extended the family theory of BLUP used in genetic improvement to a broader class of models (RESENDE et al., 2018).

By leveraging the hierarchical structure of the hierarchical likelihood, models for variance components and dispersion parameters can be sequentially incorporated. This flexibility in modelling is heightened by the ability to use a wide range of probability distributions for both the response variable and random effects (RESENDE et al., 2018). Examples include employing a Poisson distribution for the response variable and a Gamma distribution for random effects. Frailty models for survival analysis offer another instance, enabling the handling of heterogeneity by introducing random effects for residual variance. Random effects can also be integrated into models for data smoothing.

The hierarchical likelihood framework is not only useful for estimation but can also facilitate model selection. The corrected Akaike information criterion (AIC_c) derived from hierarchical likelihood is equivalent to the deviance information criterion (DIC) applied in Bayesian Statistics (LEE; NOH, 2012).

The estimation of random effects under various probability distribution assumptions is of

great interest across multiple domains and can be accomplished through HGLMMs. More information and practical examples can be found in Resende et al. (2018).

References

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

ARAÚJO, M. S.; CHAVES, S. F. S.; DIAS, L. A. S.; FERREIRA, F. M.; PEREIRA, G. R.; BEZERRA, A. R. G.; ALVES, R. S.; HEINEMANN, A. B.; BRESEGHELLO, F.; CARNEIRO, P. C. S.; KRAUSE, M. D.; COSTA-NETO, G.; DIAS, K. O. G. GIS-FA: an approach to integrating thematic maps, factor-analytic, and envirotyping for cultivar targeting. **Theoretical and Applied Genetics**, v. 137, n. 4, p. 80, 2024.

BERNARDELI, A.; ROCHA, J. R. A. S. C.; BORÉM, A.; LORENZONI, R.; AGUIAR, R.; SILVA, J. N. B.; BUENO, R. D.; ALVES, R. S.; JARQUIN, D.; RIBEIRO, C.; LAMAS COSTA, M. D.-B. Modeling spatial trends and enhancing genetic selection: An approach to soybean seed composition breeding. **Crop Science**, v. 61, n. 2, p. 976–988, 2021.

BIJMA, P. The quantitative genetics of indirect genetic effects: a selective review of modelling issues. **Heredity**, v. 112, n. 1, p. 61–69, 2014.

BOX, J. F. R. A. Fisher and the Design of Experiments, 1922-1926. **The American Statistician**, v. 34, n. 1, p. 1–7, 1980.

BURGUEÑO, J. Spatial Analysis of Field Experiments. In: GLAZ, B.; YEATER, K. M. (Eds.). **ASA, CSSA, and SSSA Books**. Madison, WI, USA: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America, Inc., 2018. P. 319–344.

BURKHART, H. E.; TOMÉ, M. Indices of Individual-Tree Competition. In: MODELING Forest Trees and Stands. Dordrecht: Springer Netherlands, 2012. P. 201–232.

CALLISTER, A. N.; COSTA-NETO, G.; BRADSHAW, B. P.; ELMS, S.; CROSSA, J.; BRAWNER, J. T. Enviromic prediction enables the characterization and mapping of *Eucalyptus globulus* Labill breeding zones. **Tree Genetics & Genomes**, v. 20, n. 1, p. 3, 2024.

CAPPA, E. P.; CANTET, R. J. C. Direct and competition additive effects in tree breeding: bayesian estimation from an individual tree mixed model. **Silvae Genetica**, v. 57, n. 2, p. 45–56, 2008.

CAPPA, E. P.; EL-KASSABY, Y. A.; MUÑOZ, F.; GARCIA, M. N.; VILLALBA, P. V.; KLÁPŠTĚ, J.; POLTRI, S. N. M. Improving accuracy of breeding values by incorporating genomic information in spatial-competition mixed models. **Molecular Breeding**, v. 37, n. 10, p. 125, 2017.

CAVANAUGH, J. E.; NEATH, A. A. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. **WIREs Computational Statistics**, v. 11, n. 3, e1460, 2019.

CHAVES, S. F. S.; ALVES, R. S.; DIAS, L. A. S.; ALVES, R. M.; DIAS, K. O. G.; EVANGELISTA, J. S. P. C. Analysis of repeated measures data through mixed models: An application in *Theobroma grandiflorum* breeding. **Crop Science**, v. 63, n. 4, p. 2131–2144, 2023.

CHAVES, S. F. S.; EVANGELISTA, J. S. P. C.; ALVES, R. S.; FERREIRA, F. M.; DIAS, L. A. S.; ALVES, R. M.; DIAS, K. O. G.; BHERING, L. L. Application of linear mixed models for multiple harvest/site trial analyses in perennial plant breeding. **Tree Genetics & Genomes**, v. 18, n. 6, p. 44, 2022.

CHAVES, S. F. S.; EVANGELISTA, J. S. P. C.; TRINDADE, R. S.; DIAS, L. A. S.; GUIMARÃES, P. E.; GUIMARÃES, L. J. M.; ALVES, R. S.; BHERING, L. L.; DIAS, K. O. G. Employing factor analytic tools for selecting high-performance and stable tropical maize hybrids. **Crop Science**, v. 63, n. 3, p. 1114–1125, 2023.

COOPER, M.; DELACY, I. H. Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. **Theoretical and Applied Genetics**, v. 88, n. 5, p. 561–572, 1994.

COSTA E SILVA, J.; KERR, R. J. Accounting for competition in genetic analysis, with particular emphasis on forest genetic trials. **Tree Genetics & Genomes**, v. 9, n. 1, p. 1–17, 2013.

COSTA E SILVA, J.; POTTS, B. M.; BIJMA, P.; KERR, R. J.; PILBEAM, D. J. Genetic control of interactions among individuals: contrasting outcomes of indirect genetic effects arising from neighbour disease infection and competition in a forest tree. **New Phytologist**, v. 197, n. 2, p. 631–641, 2013.

CULLIS, B. R.; JEFFERSON, P.; THOMPSON, R.; SMITH, A. B. Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a *Pinus radiata* breeding programme. **Theoretical and Applied Genetics**, v. 127, n. 10, p. 2193–2210, 2014.

CULLIS, B. R.; SMITH, A. B.; BEECK, C. P.; COWLING, W. A. Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. **Genome**, v. 53, n. 11, p. 1002–1016, 2010.

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, n. 4, p. 381–393, 2006.

DIAS, K. O. G.; GEZAN, S. A.; GUIMARÃES, C. T.; NAZARIAN, A.; SILVA, L. C.; PARENTONI, S. N.; GUIMARÃES, P. E. O.; ANONI, C. O.; PÁDUA, J. M. V.; PINTO, M. O.; NODA, R. W.; RIBEIRO, C. A. G.; MAGALHÃES, J. V.; GARCIA, A. A. F.; SOUZA, J. C.; GUIMARÃES, L. J. M.; PASTINA, M. M. 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, n. 1, p. 24–37, 2018.

DRTON, M.; PLUMMER, M. A Bayesian information criterion for singular models. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 79, n. 2, p. 323–380, 2017.

DUARTE, J. B.; VENCOVSKY, R. Spatial statistical analysis and selection of genotypes in plant breeding. **Pesquisa Agropecuária Brasileira**, v. 40, p. 107–114, 2005.

FERREIRA, F. M.; CHAVES, S. F. S.; BHERING, L. L.; ALVES, R. S.; TAKAHASHI, E. K.; SOUSA, J. E.; RESENDE, M. D. V.; LEITE, F. P.; GEZAN, S. A.; VIANA, J. M. S.; FERNANDES, S. B.; DIAS, K. O. G. A novel strategy to predict clonal composites by jointly modeling spatial variation and genetic competition. **Forest Ecology and Management**, v. 548, p. 121393, 2023.

FISHER, R. A. The arrangement of field experiments. **Journal of the Ministry of Agriculture**, v. 33, p. 503–515, 1926.

GEZAN, S. A.; WHITE, T. L.; HUBER, D. A. Accounting for spatial variability in breeding trials: a simulation study. **Agronomy Journal**, v. 102, n. 6, p. 1562–1571, 2010.

GEZAN, S. A.; CARVALHO, M. P.; SHERRILL, J. Statistical methods to explore genotype-by-environment interaction for loblolly pine clonal trials. **Tree Genetics & Genomes**, v. 13, n. 1, p. 1, 2017.

GIANOLA, D.; FERNANDO, R. L. Bayesian methods in animal breeding theory. **Journal of Animal Science**, v. 63, n. 1, p. 217–244, 1986.

GILMOUR, A. R.; CULLIS, B. R.; VERBYLA, A. P. Accounting for natural and extraneous variation in the analysis of field experiments. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 2, n. 3, p. 269–293, 1997.

GOGEL, B.; SMITH, A.; CULLIS, B. R. Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. **Euphytica**, v. 214, n. 2, p. 44, 2018.

GRIFFING, B. Selection in Reference to Biological Groups I. Individual and Group Selection Applied to Populations of Unordered Groups. **Australian Journal of Biological Sciences**, v. 20, n. 1, p. 127–140, 1967.

HAINING, R. P. The moving average model for spatial interaction. **Transactions of the Institute of British Geographers**, v. 3, n. 2, p. 202–225, 1978.

HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a selection model. **Biometrics**, v. 31, n. 2, p. 423, 1975.

HENDERSON, C. R. **Applications of linear models in animal breeding**. University of Guelph, 1984.

ISIK, F.; HOLLAND, J.; MALTECCA, C. Multi Environmental Trials. In: ISIK, F.; HOLLAND, J.; MALTECCA, C. (Eds.). **Genetic Data Analysis for Plant and Animal Breeding**. Cham: Springer International Publishing, 2017. P. 227–262.

KELLY, A. M.; SMITH, A. B.; ECCLESTON, J. A.; CULLIS, B. R. The accuracy of varietal selection using Factor Analytic Models for Multi-Environment plant breeding trials. **Crop Science**, v. 47, n. 3, p. 1063–1070, 2007.

LEDERMANN, T.; STAGE, A. R. Effects of competitor spacing in individual-tree indices of competition. **Canadian Journal of Forest Research**, v. 31, n. 12, p. 2143–2150, 2001.

LEE, Y.; HA, I. D. Orthodox BLUP versus h-likelihood methods for inferences about random effects in Tweedie mixed models. **Statistics and Computing**, v. 20, n. 3, p. 295–303, 2010.

LEE, Y.; NELDER, J. A. Hierarchical Generalized Linear Models. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 58, n. 4, p. 619–678, 1996.

LEE, Y.; NELDER, J. A.; PAWITAN, Y. **Generalized Linear Models with Random Effects**. Boca Raton, Fla: Chapman and Hall/CRC, 2006.

LEE, Y.; NOH, M. Modelling random effect variance with double hierarchical generalized linear models. **Statistical Modelling**, v. 12, n. 6, p. 487–502, 2012.

LI, Y.; SUONTAMA, M.; BURDON, R. D.; DUNGEY, H. S. Genotype by environment interactions in forest tree breeding: review of methodology and perspectives on research and application. **Tree Genetics & Genomes**, v. 13, n. 3, p. 60, 2017.

MALOSETTI, M.; RIBAUT, J.-M.; EEUWIJK, F. A. van. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. **Frontiers in Physiology**, v. 4, p. 44, 2013.

MRODE, R. A. **Linear models for the prediction of animal breeding values**. 3rd ed. Boston, MA: CABI, 2014.

MUIR, W. M. Incorporation of competitive effects in forest tree or animal breeding programs. **Genetics**, v. 170, n. 3, p. 1247–1259, 2005.

NEATH, A. A.; CAVANAUGH, J. E. The Bayesian information criterion: background, derivation, and applications. **WIREs Computational Statistics**, v. 4, n. 2, p. 199–203, 2012.

NELDER, J. A.; PREGIBON, D. An Extended Quasi-Likelihood Function. **Biometrika**, v. 74, n. 2, p. 221–232, 1987.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society. Series A (General)**, v. 135, n. 3, p. 370–384, 1972.

PAPADAKIS, J. Méthode statistique pour des experiences sur champ. **Bull. Inst. Amel. Plantes a Salonique**, v. 23, p. 13–29, 1937.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545–554, 1971.

PIEPHO, H.-P. Analyzing genotype-environment data by mixed models with multiplicative terms. **Biometrics**, v. 53, n. 2, p. 761–766, 1997.

PIEPHO, H.-P. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures: **Theoretical and Applied Genetics**, v. 97, n. 1-2, p. 195–201, 1998.

PIEPHO, H.-P. Allowing for the structure of a designed experiment when estimating and testing trait correlations. **The Journal of Agricultural Science**, v. 156, n. 1, p. 59–70, 2018.

PIEPHO, H.-P. A coefficient of determination (R^2) for generalized linear mixed models. **Biometrical Journal**, v. 61, n. 4, p. 860–872, 2019.

PIEPHO, H.-P.; MÖHRING, J.; MELCHINGER, A. E.; BÜCHSE, A. BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, v. 161, n. 1-2, p. 209–228, 2008.

POMMERENING, A.; MEADOR, A. J. S. Tamm review: Tree interactions between myth and reality. **Forest Ecology and Management**, v. 424, p. 164–176, 2018.

R CORE TEAM. **R: A Language and environment for statistical computing**. Viena, Áustria: R Foundation for Statistical Computing, 2023.

RESENDE, M. D. V. **Genética biométrica e Estatística no melhoramento de plantas perenes**. 1. ed. Brasilia: Embrapa Informação Tecnológica, 2002.

RESENDE, M. D. V. **SELEGEN-REML/BLUP: Sistema estatístico e seleção genética computadorizada via Modelos Lineares Mistos**. 1. ed. Colombo: Embrapa Florestas, 2007.

RESENDE, M. D. V. Software Selegen-REML/BLUP: a useful tool for plant breeding. **Crop Breeding and Applied Biotechnology**, v. 16, n. 4, p. 330–339, 2016.

RESENDE, M. D. V.; ALVES, R. S. Linear, generalized, hierarchical, bayesian and random regression mixed models in genetic/genomics in plant breeding. **Functional Plant Breeding Journal**, v. 2, n. 2, p. 1–31, 2020.

RESENDE, M. D. V.; ALVES, R. S. Statistical significance, selection accuracy, and experimental precision in plant breeding. **Crop Breeding and Applied Biotechnology**, v. 22, n. 3, e42712238, 2022.

RESENDE, M. D. V.; AZEVEDO, C. F.; SILVA, F. F.; NASCIMENTO, M.; GOIS, I. B.; ALVES, R. S. **Modelos Hierárquicos Generalizados Lineares Mistos (HGLMM), Máxima Verossimilhança Hierárquica (HIML) e HG-BLUP**. Visconde do Rio Branco, MG: Suprema, 2018.

RESENDE, M. D. V.; DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, v. 37, n. 3, p. 182–194, 2007.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão, aleatória, seleção genômica, QTL, GWAS, estatística espacial e temporal, competição, sobrevivência.** Viçosa, MG, Brazil: UFV, 2014.

RESENDE, M. D. V.; THOMPSON, R. Factor analytic multiplicative mixed models in the analysis of multiple experiments. **Revista de Matemática e Estatística**, v. 22, n. 2, p. 31–52, 2004.

SCHMIDT, P.; HARTUNG, J.; BENNEWITZ, J.; PIEPHO, H.-P. Heritability in plant breeding on a genotype-difference basis. **Genetics**, v. 212, n. 4, p. 991–1008, 2019.

SCHMIDT, P.; HARTUNG, J.; RATH, J.; PIEPHO, H.-P. Estimating Broad-Sense Heritability with unbalanced data from agricultural cultivar trials. **Crop Science**, v. 59, n. 2, p. 525–536, 2019.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, v. 6, n. 2, p. 461–464, 1978.

SMITH, A. B.; CULLIS, B. R. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. **Euphytica**, v. 214, n. 8, p. 143, 2018.

SMITH, A. B.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. **The Journal of Agricultural Science**, v. 143, n. 6, p. 449–462, 2005.

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

SMITH, A. B.; GANESALINGAM, A.; KUCHEL, H.; CULLIS, B. R. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. **Theoretical and Applied Genetics**, v. 128, n. 1, p. 55–72, 2015.

SMITH, A. B.; NORMAN, A.; KUCHEL, H.; CULLIS, B. R. Plant variety selection using interaction classes derived from factor analytic linear mixed models: models with independent variety effects. **Frontiers in Plant Science**, v. 12, p. 1857, 2021.

STEFANOVA, K. T.; BUIRCHELL, B. Multiplicative mixed models for genetic gain assessment in lupin breeding. **Crop Science**, v. 50, n. 3, p. 880–891, 2010.

THE VSNI TEAM. **asreml: Fits Linear Mixed Models using REML**. 2023.

THOMPSON, R.; BAKER, R. J. Composite Link Functions in Generalized Linear Models. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 30, n. 2, p. 125–131, 1981.

TOLHURST, D. J.; GAYNOR, R. C.; GARDUNIA, B.; HICKEY, J. M.; GORJANC, G. Genomic selection using random regressions on known and latent environmental covariates. **Theoretical and Applied Genetics**, v. 135, n. 10, p. 3393–3415, 2022.

TOLHURST, D. J.; MATHEWS, K. L.; SMITH, A. B.; CULLIS, B. R. Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model. **Journal of Animal Breeding and Genetics**, v. 136, n. 4, p. 279–300, 2019.

VELAZCO, J. G.; RODRÍGUEZ-ÁLVAREZ, M. X.; BOER, M. P.; JORDAN, D. R.; EILERS, P. H. C.; MALOSETTI, M.; EEUWIJK, F. A. van. Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. **Theoretical and Applied Genetics**, v. 130, n. 7, p. 1375–1392, 2017.

VERBYLA, A. P.; CULLIS, B. R.; KENWARD, M. G.; WELHAM, S. J. The analysis of designed experiments and longitudinal data by using smoothing splines. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 48, n. 3, p. 269–311, 1999.

VERBYLA, A. P. A note on model selection using information criteria for general linear models estimated using REML. **Australian & New Zealand Journal of Statistics**, v. 61, n. 1, p. 39–50, 2019.

WALSH, B.; LYNCH, M. **Evolution and Selection of Quantitative Traits**. Oxford University Press, 2018. v. 1.

WATERS, D. L.; WERF, J. H. J. van der; ROBINSON, H.; HICKEY, L. T.; CLARK, S. A. Partitioning the forms of genotype-by-environment interaction in the reaction norm analysis of stability. **Theoretical and Applied Genetics**, v. 136, n. 5, p. 99, 2023.

WICKHAM, H.; AVERICK, M.; BRYAN, J.; CHANG, W.; MCGOWAN, L. D.; FRANÇOIS, R.; GROLEMUND, G.; HAYES, A.; HENRY, L.; HESTER, J.; KUHN, M.; PEDERSEN, T. L.; MILLER, E.; BACHE, S. M.; MÜLLER, K.; OOMS, J.; ROBINSON, D.; SEIDEL, D. P.; SPINU, V.; TAKAHASHI, K.; VAUGHAN, D.; WILKE, C.; WOO, K.; YUTANI, H. Welcome to the Tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.

WILKS, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. **The Annals of Mathematical Statistics**, v. 9, n. 1, p. 60–62, 1938.

CHAPTER 6

PROBBREED: A NOVEL TOOL FOR CALCULATING THE RISK OF CULTIVAR RECOMMENDATION IN MULTI-ENVIRONMENT TRIALS

Published article: CHAVES, S. F. S.; KRAUSE, M. D.; DIAS, L. A. S.; GARCIA, A. A. F.; DIAS, K. O. G. ProbBreed: A novel tool for calculating the risk of cultivar recommendation in multi-environment trials. **G3 Genes|Genomes|Genetics**, v. 14, n. 3, jkae013, 2024.

Neglecting genotype-by-environment interactions in multi-environment trials (MET) increases the risk of flawed cultivar recommendations for growers. Recent advancements in probability theory coupled with cutting-edge software offer a more streamlined decision-making process for selecting suitable candidates across diverse environments. Here, we present the user-friendly `ProbBreed` package in `R`, which allows breeders to calculate the probability of a given genotype outperforming competitors under a Bayesian framework. This article outlines the package's basic workflow and highlights its key features, ranging from MET model fitting to estimating the *per se* and pairwise probabilities of superior performance and stability for selection candidates. Remarkably, only the selection intensity is required to compute these probabilities. By democratizing this complex yet efficient methodology, `ProbBreed` aims to enhance decision-making and ultimately contribute to more accurate cultivar recommendations in breeding programs.

1 Introduction

Plant breeding programs routinely evaluate experimental genotypes in multi-environmental trials (MET). The phenotypic manifestation in MET for quantitative traits is shaped by genotype-by-environment interactions (GEI), which complicates selection due to crossover (complex) interactions (COOPER; DELACY, 1994; LYNCH; WALSH, 1998). Neglecting GEI increases the risk of selecting a genotype that performs poorly in specific environments or

mega-environments (regions). Thus, exploring GEI is critical for cultivar recommendation in the target population of environments (TPE).

Several studies used the frequentist framework to compute a measure of risk to rank genotypes in MET (ANNICCHIARICO, 1992; BARAH et al., 1981; ESKRIDGE et al., 1991; MEAD et al., 1986). More recently, Dias et al. (2022) proposed a novel Bayesian method that employs the posterior distribution to get Hamiltonian Monte Carlo estimates of performance and stability probabilities. Their core ideas are to assess the predictability of an experimental genotype's performance through its probability of being amongst the selected genotypes in a global (marginal, across-environments) or specific context (conditional, within environments or mega-environments); and the probability of a selection candidate having an invariant performance across environments. The method also provides pairwise probabilities, useful for direct comparison of experimental genotypes, or experimental genotypes versus check cultivars.

The package `ProbBreed` was built upon the method of Dias et al. (2022) to allow the application of probability theory to cultivar recommendation in MET. Its underlying method is intuitive for plant breeders for two main reasons. First, it emulates a situation usually faced by growers: choosing cultivar(s) that are likely to perform well for the next cropping season. Second, probabilities (marginal or conditional) are calculated according to the intensity of selection, which is also part of plant breeders' routine. Furthermore, and in contrast with biplot-based methods (YAN et al., 2000), our method is straightforward given the sole metric used for selection is the calculated probabilities.

Plant breeding programs can benefit from using our probabilistic approach to perform a more rapid and effective decision-making process toward cultivar recommendation for a TPE. Thus, we present the open-source `R` (R CORE TEAM, 2023) package `ProbBreed`, a user-friendly tool that democratizes the method from Dias et al. (2022), regardless of the user's programming abilities. We first provide an overview of the **Theory** behind `ProbBreed` and describe the **Motivating example** contained within the package. Finally, at **Results and discussion**, we illustrate a workflow of the package's usage, employing the described dataset. We also provide further novelties regarding the paper of Dias et al. (2022) in this section, such as the multi-location-year model and a comparison with a frequentist linear mixed model.

2 Methods

2.1 Theory

When analyzing data from MET, the main goal is to select high-performance genotypes with stable phenotypic responses across environments, given a selection intensity. Naturally, selecting and recommending experimental genotypes to a TPE encompasses latent risks that plant breeders assume. The probabilities proposed by Dias et al. (2022) allow considering these risks when performing the selection. We detail these probabilities below.

Probability of superior performance

Consider a dataset in which J genotypes ($j = 1, 2, \ldots, J$) were evaluated at K environments ($k = 1, 2, \ldots, K$) with y observed phenotypes. Let Ω be a subset of the high-performance selected genotypes according to the intensity of selection. A given genotype j will belong to Ω if its genotypic marginal value ($\hat{g}_j$) is high (or low) enough compared to its peers. We can emulate the occurrence of S trials ($s = 1, 2, \ldots, S$) with Bayesian models by leveraging Monte Carlo discretized samples from the posterior distributions of the fitted Bayesian models. Then, the probability of the j^{th} genotype belonging to Ω is its ratio of success ($\hat{g}_j \in \Omega$) events over the total number of sampled events [$S = (\hat{g}_j \in \Omega) + (\hat{g}_j \notin \Omega)$], defined as follows:

$$Pr(\hat{g}_j \in \Omega|y) = \frac{1}{S}\sum_{s=1}^{S} I(\hat{g}_j^{(s)} \in \Omega|y) \tag{1}$$

where $I(\hat{g}_j^{(s)} \in \Omega|y)$ is an indicator variable that can assume two values: (1) if $\hat{g}_j \in \Omega$ in the s^{th} sample, and (0) otherwise.

Similarly, the conditional probability of superior performance can be applied to individual environments. Let Ω_k represent the subset of superior genotypes in the k^{th} environment, so that the probability of the $j^{\text{th}} \in \Omega_k$ can calculated as follows:

$$Pr(\hat{g}_{jk} \in \Omega_k|y) = \frac{1}{S}\sum_{s=1}^{S} I(\hat{g}_{jk}^{(s)} \in \Omega_k|y) \tag{2}$$

where $I(\hat{g}_{jk}^{(s)} \in \Omega_k|y)$ is an indicator variable mapping success (1) if $\hat{g}_{jk}^{(s)}$ exists in Ω_k, failure (0) otherwise, and $\hat{g}_{jk}^{(s)} = \hat{g}_j^{(s)} + \widehat{ge}_{jk}^{(s)}$. Note that computing conditional probabilities (i.e.,

conditional to the k^{th} environment or mega-environment), the interaction of the j^{th} genotype with the k^{th} environment is accounted.

The pairwise probabilities of superior performance can also be calculated across or within environments. This metric assesses the probability of the j^{th} genotype being superior to another experimental genotype or a commercial check. The calculations are as follows:

$$Pr(\hat{g}_j > \hat{g}_{j'}|y) = \frac{1}{S}\sum_{s=1}^{S} I(\hat{g}_j^{(s)} > \hat{g}_{j'}^{(s)}|y) \tag{3}$$

$$Pr(\hat{g}_{jk} > \hat{g}_{j'k}|y) = \frac{1}{S}\sum_{s=1}^{S} I(\hat{g}_{jk}^{(s)} > \hat{g}_{j'k}^{(s)}|y) \tag{4}$$

Note that equations 3 and 4 are set for when the selection direction is positive (i.e., the aim is to increase the trait value). If the selection is negative, > can simply be switched by <. Equation 3 computes the pairwise probabilities across environments, while equation 4 within environments.

Probability of superior stability

Probabilities of superior performance highlight high-performance genotypes. For stability, the probability of superior stability is more adequate. This metric can be directly compared to the method of Shukla (1972): a stable genotype is the one that has a low variance of the GEI effects [$var(\widehat{ge})$]. Using the same probability principles previously described, the probability of superior stability is given as follows:

$$Pr[var(\widehat{ge}_{jk}) \in \Omega|y] = \frac{1}{S}\sum_{s=1}^{S} I[var(\widehat{ge}_{jk}^{(s)}) \in \Omega|y] \tag{5}$$

where $I[var(\widehat{ge}_{jk}^{(s)}) \in \Omega|y]$ indicates if $var(\widehat{ge}_{jk}^{(s)})$ exists in Ω (1) or not (0). Note that this probability can only be computed across environments since it depends on $var(\widehat{ge}_{jk})$. Pairwise probabilities of superior stability are also computed in the context of stability:

$$Pr[var(\widehat{ge}_{jk}) < var(\widehat{ge}_{j'k})|y] = \frac{1}{S}\sum_{s=1}^{S} I[var(\widehat{ge}_{jk})^{(s)} < var(\widehat{ge}_{ik})^{(s)}|y] \tag{6}$$

Joint probability of superior performance and stability

The joint probability of the occurrence of independent events is the product of the individual probabilities. The estimated genotypic main effects and the variances of GEI effects are independent due to the design of linear models, thus the joint probability of superior performance and stability are given as follows:

$$Pr[\hat{g}_j \in \Omega \cap var(\widehat{ge}_{jk}) \in \Omega] = Pr(\hat{g}_j \in \Omega) \times Pr[var(\widehat{ge}_{jk}) \in \Omega] \tag{7}$$

The estimation of probabilities in this section is closely related to some key questions that are part of plant breeding programs' daily routine, such as "what is the risk of recommending a selection candidate for a TPE?", or "how probable is it that a given selection candidate perform similarly across environments?", or even "what is the probability of a selection candidate having better performance (or more stable performance) than a check cultivar in the TPE and in specific environments?".

2.2 Motivating example

We demonstrate the application of `ProbBreed` using a dataset (named `soy`) from the USDA Northern Region Uniform Soybean Tests, which is a subset of the data used by Krause et al. (2023). It contains the empirical best linear unbiased estimates (column named "Y" in the data frame) of genotypic means of the seed yield from 39 experimental genotypes ("G01" to "G39" in the column named "Gen" in the data frame) evaluated in 14 environments ("E1" to "E14" in the column named "Loc" in the data frame) across three mega-environments ("R1", "R2", and "R3" in the columns named "Reg" in the data frame). In this dataset, we are considering an environment as the combination of locations and years. The analysis was performed on a computer with 8 GB of RAM and a 12th Gen Intel® Core™ i7-1255U processor, with a base frequency of 1.70 GHz. Computational time was recorded with the `get_elapsed_time` function from the `rstan` package.

3 Results and discussion

3.1 Bayesian MET models

The first step is to fit the Bayesian MET model using the `bayes_met` function. Internally, the Bayesian models are fitted using `rstan`, a package that links `Stan` to `R` (STAN DEVELOPMENT TEAM, 2023a,b). `Stan` is a probabilistic library set in `C++` language that uses the No-U-turn Sampler (HOFFMAN; GELMAN, 2014) to automatically tune up the Hamiltonian Monte Carlo algorithm by eliminating the need to specify the number of leapfrog updates. This avoids the random-walk behaviour and improves computational efficiency. For more details about the No-U-turn Sampler and its advantage over the regular Hamiltonian Monte Carlo algorithm, see Hoffman and Gelman (2014), and Nishio and Arakawa (2019).

Currently, there are twelve models implemented in `ProbBreed`. These models differ according to the considered information regarding locations, years and breeding regions (Figure 1). Additionally, one might consider the collective information from a combination of environmental factors, such as the location-year combination for instance, as constituting an "environment". Models that consider the information of years are a novelty concerning Dias et al. (2022), see **Multi-location-year model** for further information. These models also differ regarding the experimental design: entry-mean (i.e., adjusted means), randomized complete block design (RCBD), and incomplete block design (IBD). For example, the `soy` dataset has information on breeding regions (or mega-environments) and the reported phenotypes are empirical best linear unbiased estimates of genotypic means (i.e., entry-mean basis). The function `bayes_met` is detailed in Box 1:

Box 1. Usage of function `bayes_met`

```
mod = bayes_met(data = soy,
                gen = "Gen",
                loc = "Loc",
                repl = NULL,
                reg = "Reg",
                year = NULL,
                res.het = FALSE,
                trait = "Y",
                iter = 40000,
                cores = 4,
                chains = 4)
```

In summary, users may choose which model to use based on their dataset and by changing the arguments `year`, `reg`, and `repl` (Figure 1). Users might consider an "environment" as a composite of multiple environmental factors rather than differentiating individual components, as demonstrated by (`loc = "Loc"` and `Reg = "Reg"`). In this case, only the `loc` argument would be employed, such as `loc = "Environment"`, while `year = NULL` and `reg = NULL`. Note from Box 1 that `bayes_met` has an additional argument that controls if residual variances should be considered homogeneous (`res.het = TRUE`) or heterogeneous (`res.het = FALSE`) across locations (or environments). It is noteworthy that even when breeding regions are accounted for in the model and `res.het = TRUE`, the residual variances are still considered heterogeneous only across locations. Users may also control the number of iterations (`iter`) and Markov chains (`chains`). The argument `cores` determines whether Markov chains run in parallel (`cores` > 1) or not (`cores` = 1). Each Markov chain runs the specified number of iterations set by the user independently, and, by default, half of them are reserved for the burn-in process. The function supports additional arguments passed to the `sampling` function of `rstan`. This allows advanced users to modify parameters such as the number of burn-in iterations, the frequency of saving samples, and other default settings that influence the behaviour of the sampler. Users can also define initial values, specify parameters of interest, and select the preferred sampling algorithm. Changing these parameters can aid in fixing convergence and mixing issues (See **On warnings about mixing and convergence issues** sub-

section). `bayes_met` documentation has more details on these arguments.

The assumptions of the models implemented in `bayes_met` have some presets described in detail by Dias et al. (2022). In summary, $y \sim N(E[y], \sigma)$, where the expectation of $E[y]$ depends on the models' choice. The prior probability distributions of the model effects are $x \sim N(0, S^{[x]})$, where x can be any effect but the error, with hyperprior $S^{[x/\sigma]} \sim HalfCauchy(0, \phi)$. N and $HalfCauchy$ represent the Normal and Half-Cauchy distributions, respectively, where the former is constrained to be positive (GELMAN et al., 2013). The global hyperparameter ϕ is defined as $\phi = \max(y) \times 10$. The error term has the sampling variance $\sigma \sim HalfCauchy(0, S^{[\sigma]})$ for homogeneous residual variances, and $\sigma_k \sim HalfCauchy(0, S^{[\sigma_k]})$ for heterogeneous residual variances. The weakly informative prior distributions with their respective hyperpriors allow the model to take full advantage of the data to infer the posterior distribution.

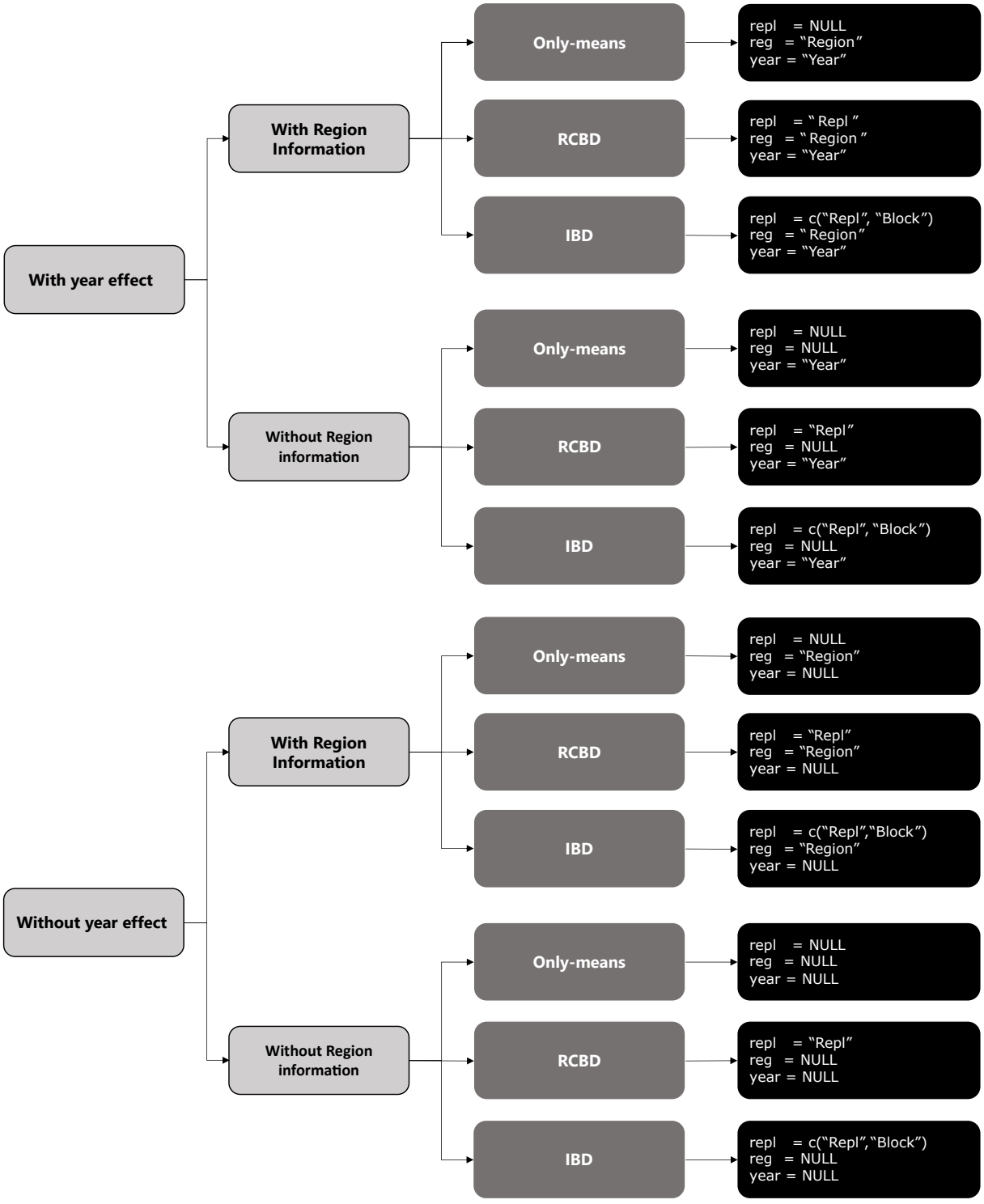
On warnings about mixing and convergence issues

By default, `rstan` detects and warns users of potential mixing and convergence issues on the fitted model. Usual problems are - but are not limited to - divergent transitions after warm-up, large $\hat{R}$, and insufficient bulk and tail effective sample size. A detailed tutorial on these problems and how to deal with them is available in `https://mc-stan.org/misc/warnings.html`. We recommend other tools to explore the model's output and easily detect and solve complications, namely the packages `posterior` (BÜRKNER et al., 2023; VEHTARI et al., 2021), `bayesplot` (GABRY et al., 2019) and `shinystan` (GABRY; VEEN, 2022). It is worth mentioning that even though `rstan` is conservative in identifying abnormalities in model fitting, models with alleged imperfect mixing and convergence can still yield acceptable results. We recommend examining the goodness-of-fit diagnostics of the `extr_outs` (as described in the next section) before making any adjustments to the model or default parameters. If `bayes_met` shows warnings, but the diagnostics of `extr_outs` does not indicate grave issues, one may carry on with the analysis.

3.2 Posterior effects and goodness-of-fit diagnostics

After fitting a Bayesian model, the information from the posterior distribution is accessed with the `extr_outs` function as follows:

Figure 1: Options to declare replications and/or blocks (`repl`), years (`year`) and regions (`reg`) effects in the `bayes_met` function. Users must substitute `Repl`, `Block`, `Year`, and `Region` with the name of the column that contains the information about replicates, block nested in replicates (if applicable), year (if available) and region (if available). RCDB and IBD are acronyms for randomized complete block design and incomplete block design, respectively.



Box 2. Usage of function `extr_outs`

```
outs = extr_outs(data = soy,
                 trait = "Y",
                 model = mod,
                 probs = c(0.05, 0.95),
                 check.stan.diag = TRUE,
                 verbose = TRUE)
```

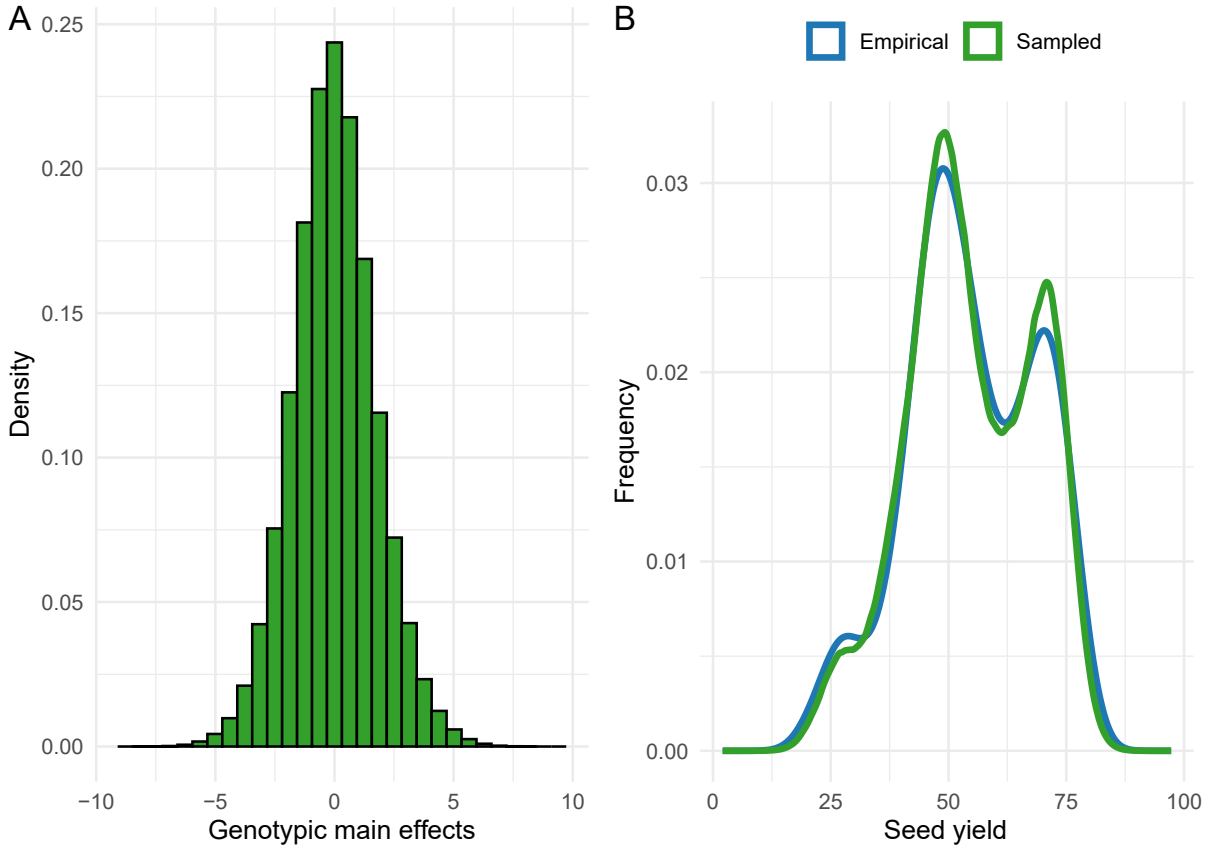
This function extracts the posterior distributions, the maximum values *a posteriori*, and the data generated by the model. `probs` is a vector with two probabilities in the decimal scale used to calculate the highest posterior density interval (HPD) of the variance components (Table 1). `mod` is the model fitted using `bayes_met`. `extr_outs` uses the posterior distributions and the data generated by the model to build plots that allow an overview of the model's goodness-of-fit (Figure 2). The function builds histograms (Figure 2A) and density plots (Figure 2B), which provide a visualization of the posterior effects' distribution; and trace plots, useful for detecting problems with the convergence of each chain. Figure 2B is particularly handy to assess if the model was able to generate data with a similar distribution to the real data. `extr_outs` provides further diagnostic plots when the argument `check.stan.diag` is set to `TRUE`. Internally, these plots are built using the `stan_diag` function. Further options are available in **`stan_diag` help page**.

Table 1: Estimates of variance components of the declared effects, and their respective standard deviation (SD), naive standard error (Naive SE), and inferior and superior high posterior density interval [HPD (0.05) and HPD (0.95), respectively]

Components	**Variance**	**SD**	**Naive SE**	**HPD (0.05)**	**HPD (0.95)**
Genotype (G)	3.314	1.392	0.005	1.39	5.822
Location (L)	251.972	138.611	0.49	107.984	502.227
G×L	6.861	5.47	0.019	0.187	16.328
Region (R)	3181.473	42228.65	149.301	0.772	8138.598
G×R	1.139	1.036	0.004	0.014	3.145
Residual	11.179	5.447	0.019	2.045	18.558

In addition to the referred plots, goodness-of-fit parameters such as the Bayesian "p-values" of test statistics, the Watanabe-Akaike information criterion (WAIC2, WATANABE, 2013), and the mean potential scale reduction factor ($\hat{R}$, GELMAN; RUBIN, 1992) are provided by `extr_outs`. The WAIC2 has a similar interpretation as AIC (the lower, the better), and it is

Figure 2: Histogram of the posterior genotypic main effects (A) and density plot of the data generated in comparison to the distribution of the real data (B). All plots were built with `ggplot2` (WICKHAM, 2016).



useful to compare different models. The $\hat{R}$ evaluates the equilibrium among chains, i.e., if all chains converged to a common distribution. In fact, it is the ratio between the average variance of samples within chains to the variance across chains, so values closer to one indicate that these variances are similar, which is desirable (FABRETI; HÖHNA, 2022). The Bayesian "p-values" are computed as the probability of a given test statistic (the mean, for example) being higher in the generated data than in the real data. If the generated data resembles the observed data, Bayesian p-values are expected to be far from the extremes (0.99 or 0.01) (GELMAN et al., 2013). A Bayesian p-value closer to 0.5 is desirable (DIAS et al., 2022). When `check.stan.diag = TRUE`, `extr_outs` provides specific diagnostics on possible divergent transitions, tree depth problems and the Bayesian Fraction of Missing Information (BFMI) values of each chain.

Table 2: Goodness-of-fit parameters: Bayesian "p-values" of test statistics [maximum, minimum, median, mean and standard deviation], the effective number of parameters, Watanabe-Akaike information criterion (WAIC2), potential scale reduction factor ($\hat{R}$), and effective sample size

Parameter	**Value**
p-value of the maximum	0.9689
p-value of the minimum	0.2614
p-value of the median	0.6710
p-value of the mean	0.5029
p-value of the std. deviation	0.5256
Effective number of parameters	134.068
WAIC2	2550.87
$\hat{R}$	1.0192
Effective sample size	0.05

3.3 Probabilities

The pipeline finishes with the `prob_sup` function, which computes probabilities of superior performance and superior stability of the selection candidates. For the `soy` dataset, the following command line was used:

Box 3. Usage of function `prob_sup`

```
results = prob_sup(data = soy,
                   trait = "Y",
                   gen = "Gen",
                   loc = "Loc",
                   mod.output = outs,
                   reg = "Reg",
                   year = NULL,
                   int = .2,
                   increase = TRUE,
                   save.df = FALSE,
                   interactive = FALSE,
                   verbose = TRUE)
```

In this example, we applied a 20% selection intensity (`int = .2`) and our goal was to increase the average seed yield (`increase = TRUE`) in the selected panel. These two pieces of information dictate how probabilities are computed in `prob_sup`. The argument

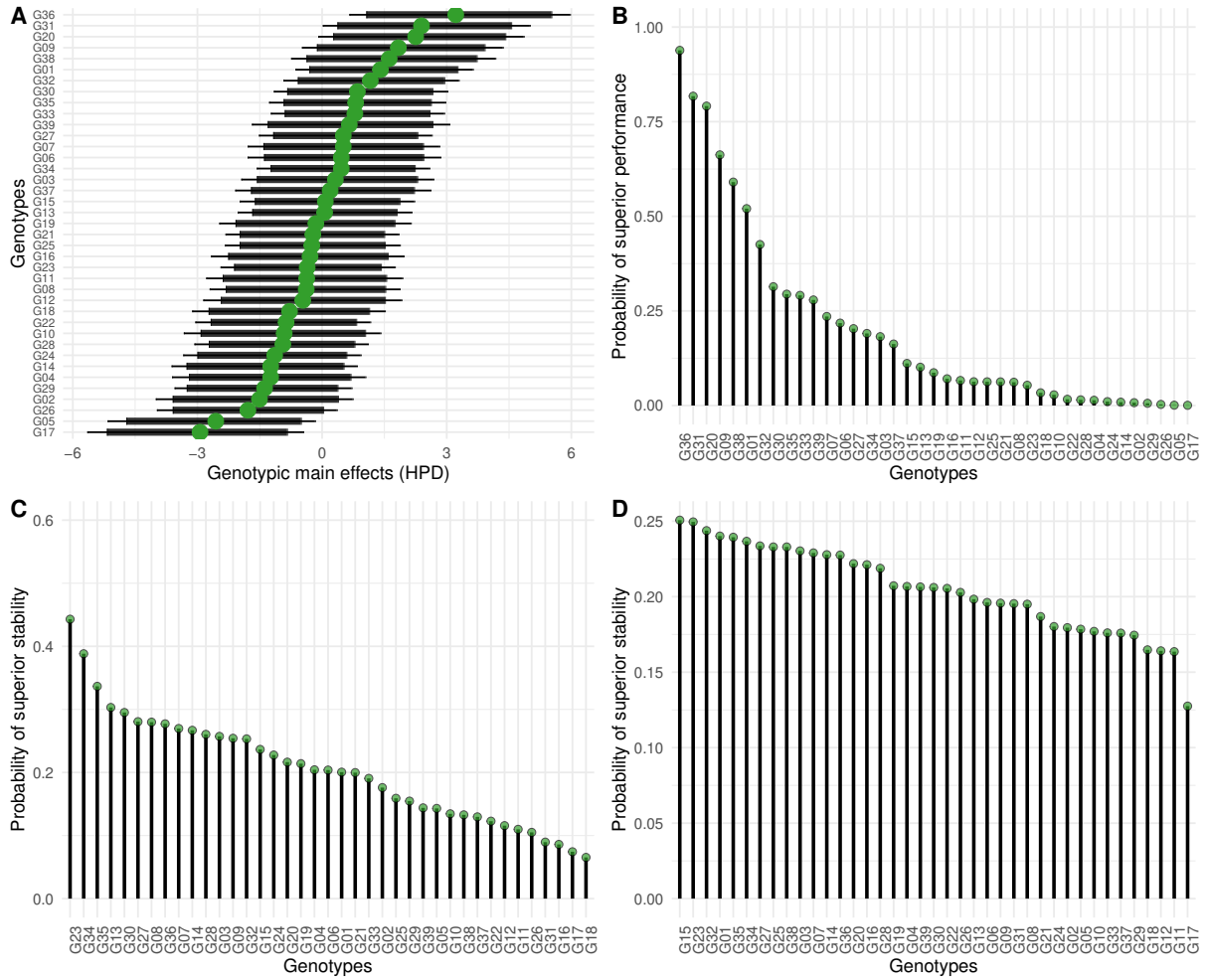
`mod.output` receives the object that stores the outcomes of the `extr_outs` function. `save.df` and `interactive` receive logical values, and determine if data frames with probabilities should be saved in the working directory (in `.csv` format) and if static plots should be converted into interactive plots using `plotly` (SIEVERT, 2020), respectively.

`prob_sup` provides an overview of the selection candidates' performance across environments, represented in a caterpillar plot containing the posterior genotypic main effects and their respective HPD intervals (Figure 3A). The maximum a posteriori values are equivalent to marginal empirical BLUPs of Frequentist linear mixed models, assuming independent genotypic effects (See the **Simulations** section). Then, it represents probabilities of superior performance and stability in lollipop plots as in Figures 3B, 3C, and 3D. For example, G36 was the candidate with the highest probability of superior performance (about 94%, Figure 3B). In other words, there is only a 6% risk of poor performance, conditioned to the intensity of selection $[Pr(\hat{g}_j \in \Omega|y)]$. The same interpretation is valid for the probability of superior stability: across locations (Figure 3C), G23 has the greatest chance to perform equally (44%), while across regions (Figure 3D), G15 has the most invariant performance (25%). Note how the results change across probability metrics. This illustrates the reason why plant breeders must have clear objective criteria before performing the analyses. If performance is preferred, Figure 3B is the one to follow. Otherwise, if stability is the final goal, Figures 3C and 3D must be prioritized.

In addition to the *per se* probabilities, we can compute pairwise probabilities for comparisons among genotypes (Figures 4A, 4B and 4C). Suppose that G35 is a promising experimental genotype and that we want to investigate if it performs better than the commercial check G11. Across locations, G35 performs better than G11 at 80% of the times (Figure 4A), and it has a more stable performance than G11 at 78% of the times (Figure 4B). Then, there is evidence to hypothesize that genotype G35 is better than the commercial check. Finally, if breeders want to identify genotypes that simultaneously have high performance and stability, 4D is the one to analyse, as it contains the joint probability of superior performance and stability (circles). Note that the same genotype will hardly be the best in all probability metrics. Probabilities of superior performance and pairwise probabilities of superior performance are also available within locations and regions (Figure 5), which is useful for specific recommendations.

With four Markov chains running into four cores in parallel, the analysis of the Bayesian

Figure 3: Highest posterior density (HPD) of the posterior genotypic main effects (A), probability of superior performance across environments (B), and probability of superior stability across locations (C) and regions (D). The dots at (A) are the maximum posterior, and the thick and thin lines at (A) represent the 95% and 97.5% HPD intervals, respectively. The *x*-axis of (B), (C) and (D) are sorted in decreasing order considering the computed probabilities. All plots were built with `ggplot2` (WICKHAM, 2016).

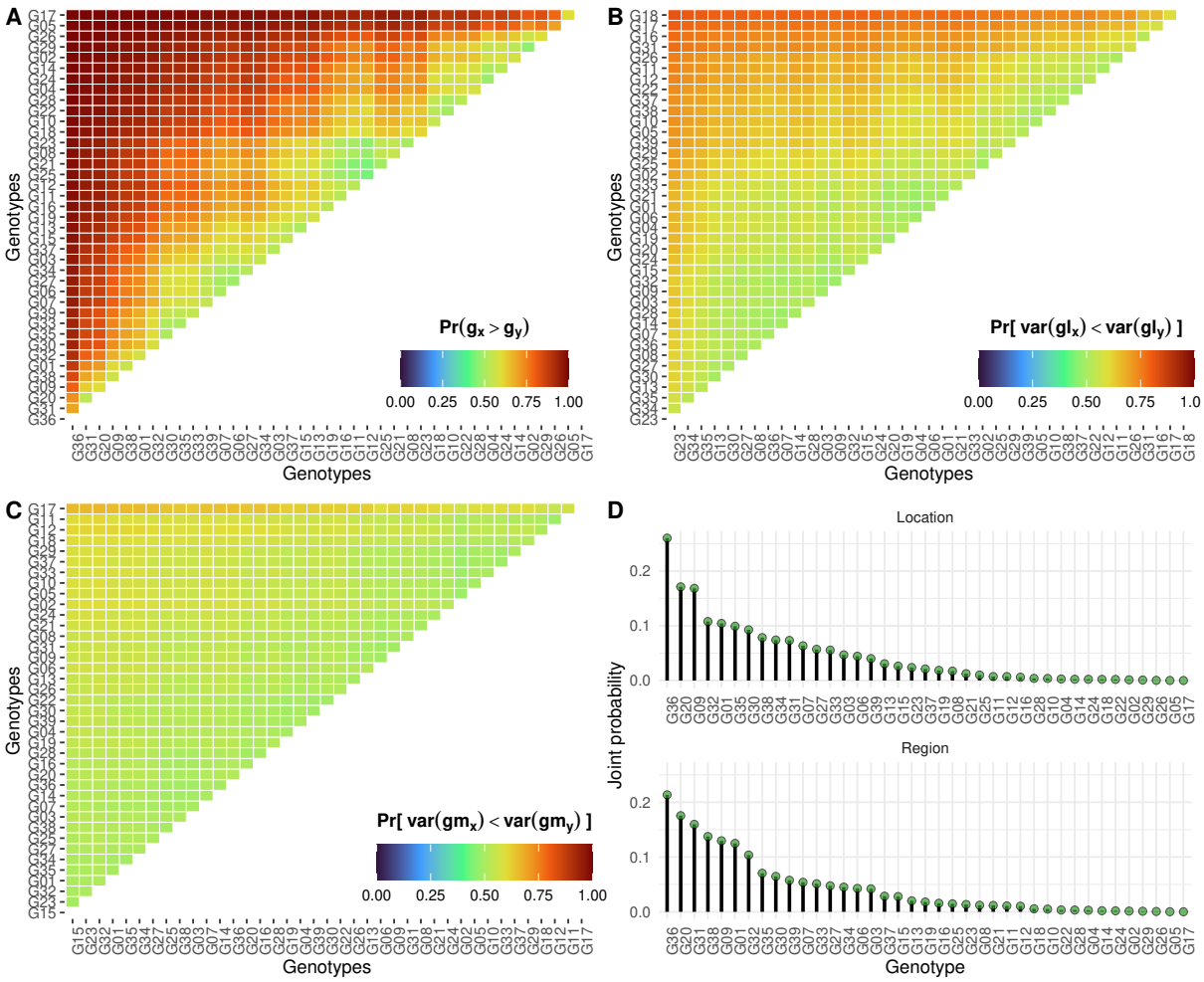


model fitted for the `soy` dataset took about 7.5 hours to run with 40,000 iterations (3.6 hours warming up and 3.9 hours sampling).

3.4 Multi-location-year model

The models described by Dias et al. (2022) consider information on locations or locations and breeding regions. As a novelty, we implemented in `ProbBreed` models that also consider the effect of years. To exemplify the usage of this model, we used four years (2000 to 2003) of the USDA Northern Region Uniform Soybean Test (KRAUSE et al., 2023). The dataset has the empirical best linear unbiased estimates (eBLUEs) of 20 genotypes evaluated at 29

Figure 4: Pairwise probabilities of superior performance across locations (A), superior stability across locations (B), superior stability across regions (C), and joint probability of superior performance and stability (D). The heatmaps at (A), (B) and (C) illustrate the probability of genotypes at the *x*-axis being superior to those on the *y*-axis. All plots were built with `ggplot2` (WICKHAM, 2016).



locations. Six genotypes were evaluated in all four years, 4 were evaluated in three years, and the remainder were evaluated in only two years. In this situation, the conditional normal likelihood of the model is as follows:

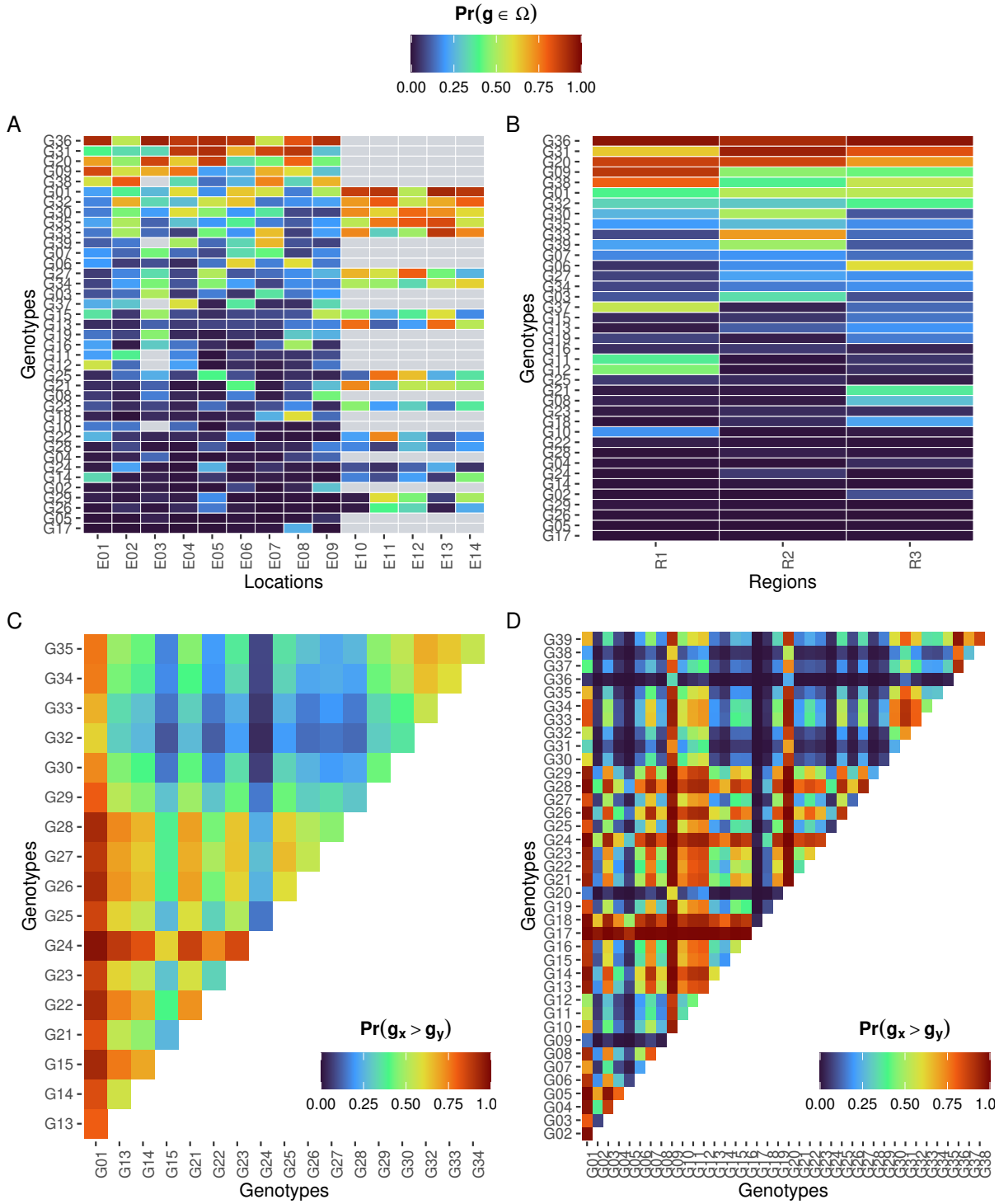
$$y_{jkh} \sim N(E[y_{jkh}], \sigma) \tag{8}$$

with

$$E[y_{jkh}] = \mu + g_j + l_k + t_h + gl_{jk} + gt_{jh} \tag{9}$$

where y_{jkh} is the eBLUE of the j^{th} genotype in the k^{th} location and in the h^{th} year, μ is the

Figure 5: Heatmaps representing the specific probabilities of superior performance within locations (A) and within regions (B), and the pairwise probabilities of superior performance between genotypes evaluated in locations “E14” (C), and in the region “R2” (D). At (A), the grey cells are locations where the genotype specified in the row was not evaluated. At (C) and (D), the probability of genotypes on the *x*-axis being superior to those on the *y*-axis are represented. All plots were built with `ggplot2` (WICKHAM, 2016).



intercept, g_j is the genotypic effect, l_k is the effect of the location, t_h is the effect of the year, gl_{jk} is the genotype-by-location interaction effect, and gt_{jh} is the genotype-by-year interaction effect. The prior and hyperprior probability distributions of this model follow the standard for all models in `ProbBreed` (as previously described). We ran the Bayesian model using 2000 iterations and 4 Markov chains. Leveraging the posterior distribution of this model, we computed the probabilities described in the Methods section considering a selection intensity of 20%.

The chains had an adequate mixing ($\hat{R} = 1.002$), meaning that the generated data is a good representation of the empirical phenotypes (Figure 6A). The histograms of the genotypic (Figure 6B), genotype-by-year (Figure 6C) and genotype-by-location effects (Figure 6D) show the distribution of the posterior values of these effects.

Figure 6: Density plot of the data generated in comparison to the distribution of the real data (A), and histograms of the genotypic effect (B), genotype-by-location effect (C) and genotype-by-year effect (D)

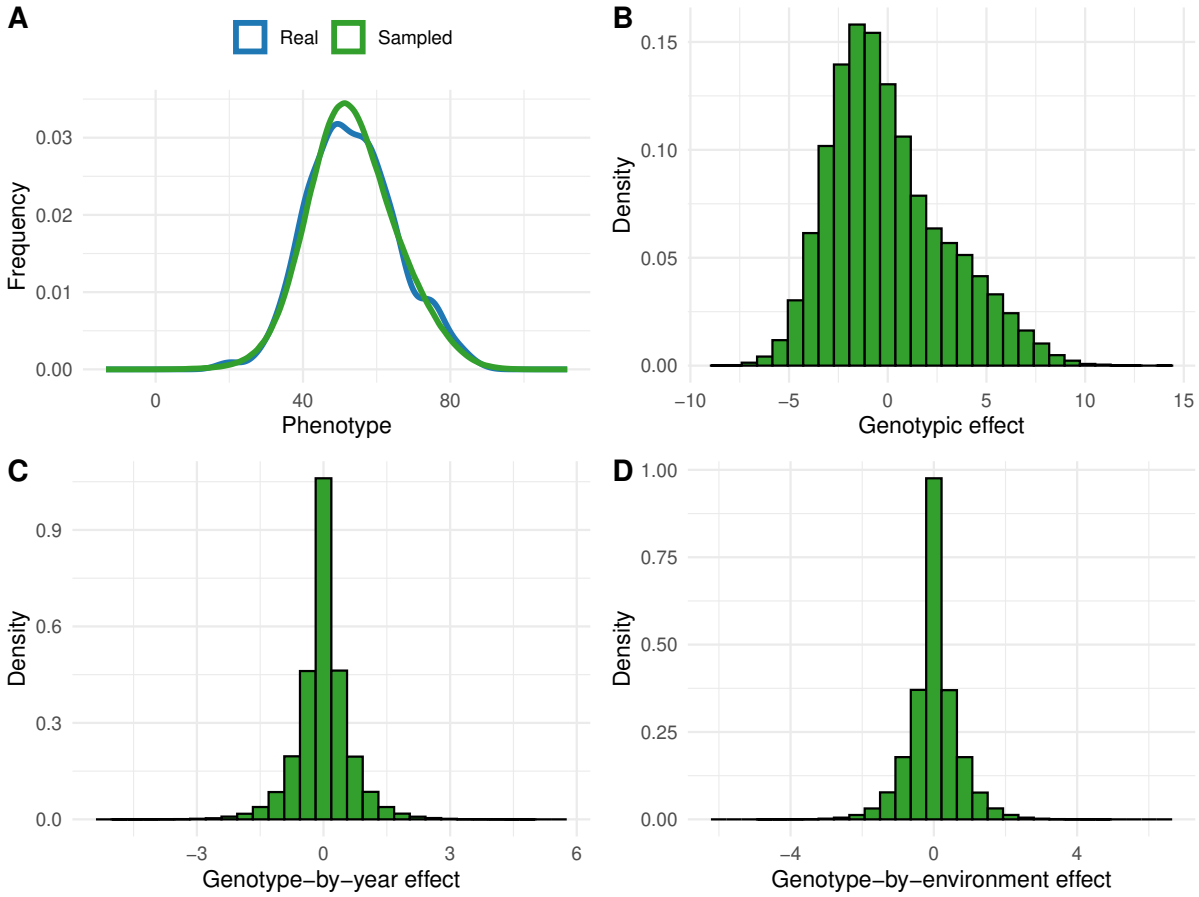
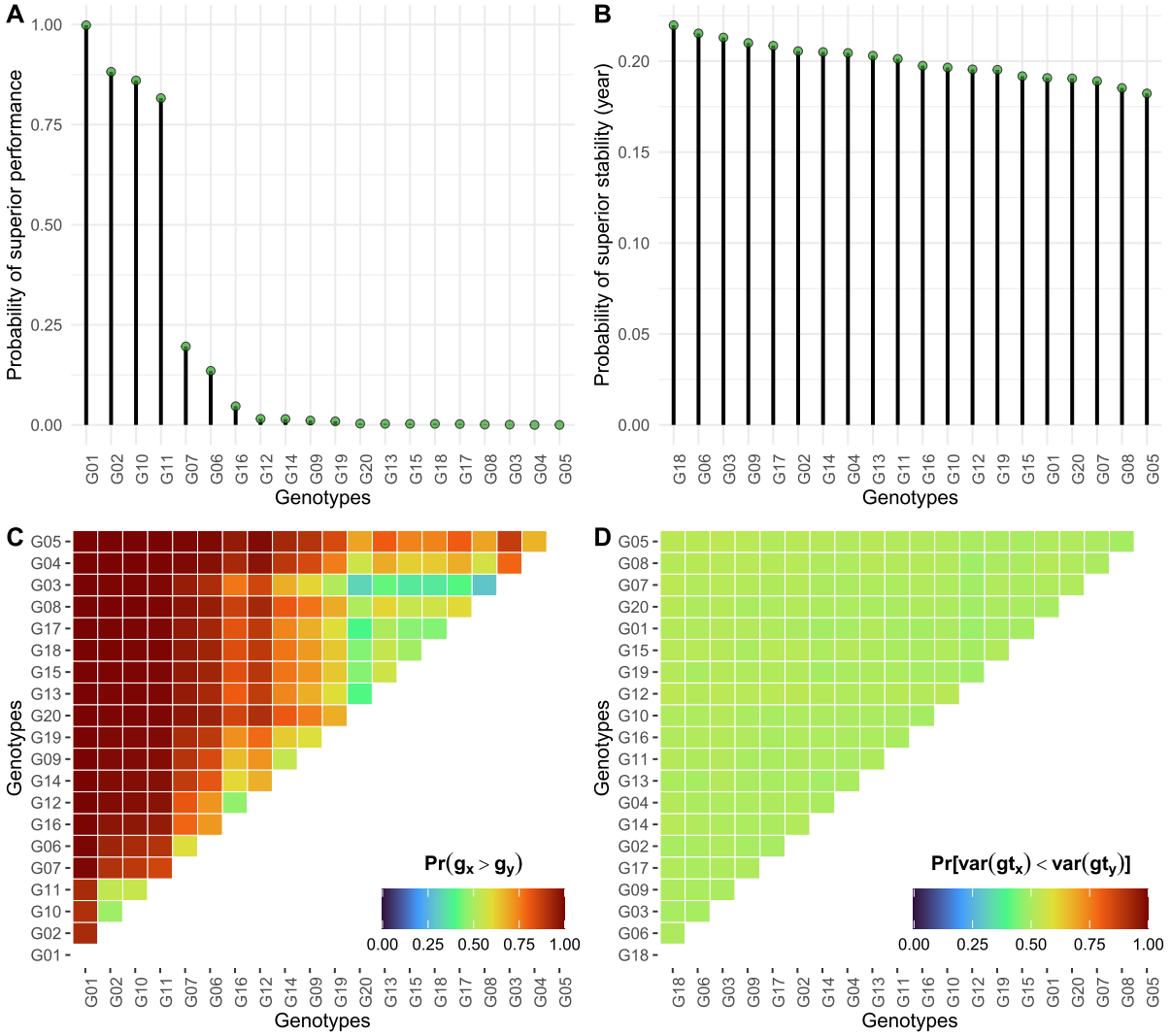


Figure 7A shows that genotypes “G01”, “G02”, “G10”, and “G11” have a higher probability of performing across all locations. Supporting information is given by the pairwise probability

of superior performance (Figure 7C). The probability of superior stability between years has little variation among genotypes (Figure 7D), meaning that they tend to have similar behaviour in terms of stability across years. This is evidenced by the colour uniformity in Figure 7D. `ProbBreed` also provides other plots, such as the probability of superior stability and pairwise of superior stability across locations, and the joint probability of superior performance and stability.

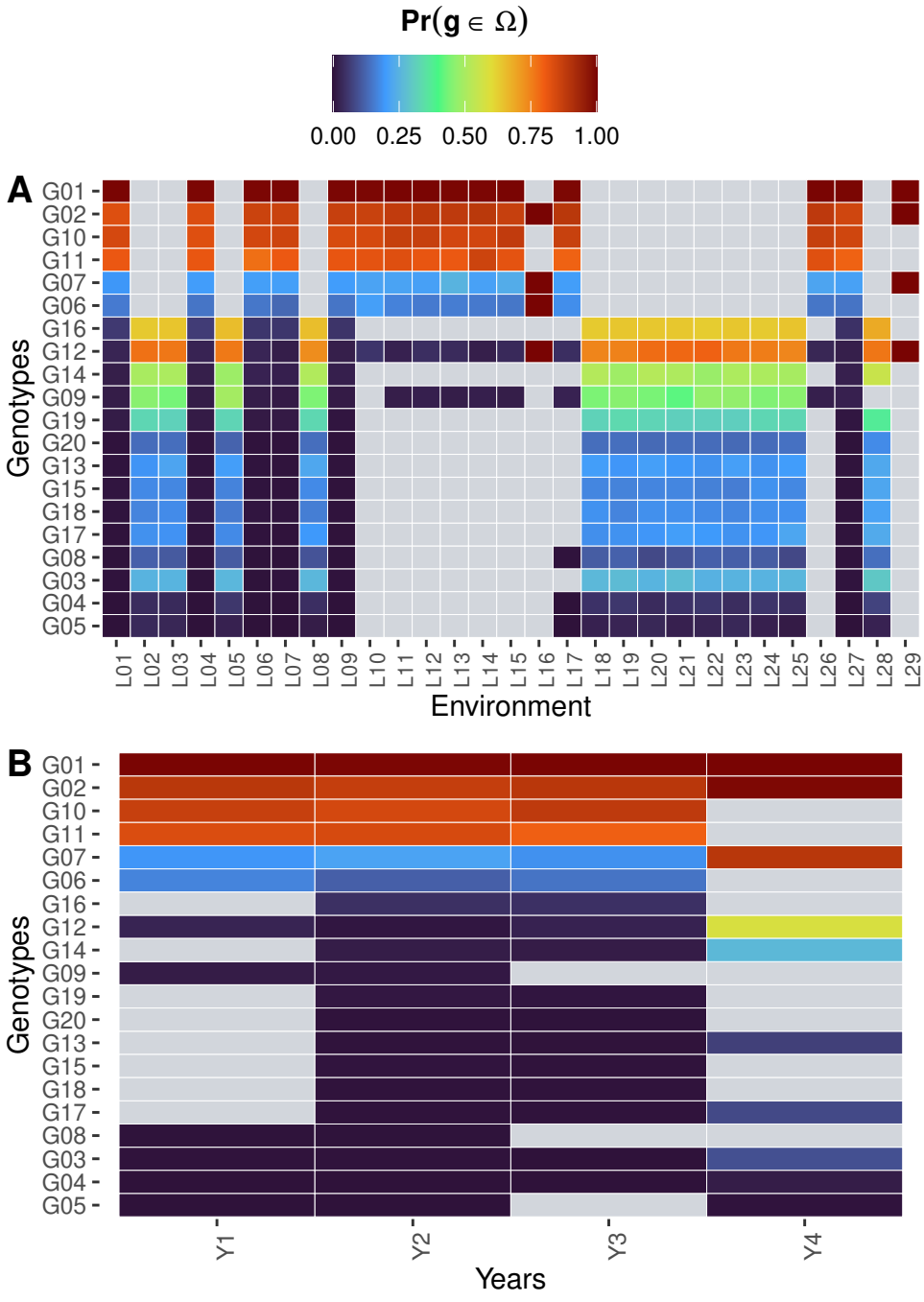
Figure 7: Probability of superior performance (A) and stability considering the genotype-by-year interaction (B), and pairwise probabilities of superior performance (C) and stability (D) considering the genotype-by-year interaction. At (C) and (D), the probability of genotypes on the *x*-axis being superior to those on the *y*-axis are represented



Finally, we can investigate the performance of genotypes within locations and years (Figure 8). Note that we are dealing with highly unbalanced data. In these situations, conclusions should be drawn with caution. Take the trials where the good performers “G01”, “G02”, “G10”

and “G11” were tested as examples (Figure 8A). In trials with few genotypes (see “L10” to “L17”), depending on the selection intensity, “G01”, “G02”, “G10” and “G11” will always appear among the selected candidates. Thus, we have to consider trials with more genotypes to determine if “G01”, “G02”, “G10” and “G11” are, indeed, better performers than their peers. The same criteria should be applied when analysing the performance of genotypes within each year (Figure 8B).

Figure 8: Probabilities of superior performance within locations (A) and years (B). The grey cells represent genotypes that were not evaluated in that specific location/year



Computational requirements

Using a laptop with 10 threads, 8 GB of RAM, and a 12th Gen Intel® Core™ i7-1255U processor with 1.70 GHz, it took 17 minutes to fit the Bayesian model. Ten of these minutes were dedicated to the warm-up iterations and the rest to the sampling iterations. We tracked the computational time to fit the Bayesian model using the `get_elapsed_time` function of `rstan`.

3.5 Simulations

Stochastic simulations compared `ProbBreed` with traditional linear mixed models. Twenty datasets, for each scenario, were simulated according to the following hypothetical plans:

1. 100 genotypes evaluated in 20 environments, trait with high heritability ($H^2 = 0.6$).
2. 100 genotypes evaluated in 20 environments, trait with low heritability ($H^2 = 0.3$).
3. 20 genotypes evaluated in 30 environments, trait with high heritability ($H^2 = 0.6$).
4. 20 genotypes evaluated in 30 environments, trait with low heritability ($H^2 = 0.3$)

Scenarios 1 and 2 emulate the intermediate stages of a breeding program when several genotypes are tested in fewer environments, whereas scenarios 3 and 4 represent the final stage of a breeding program when fewer genotypes are being tested in several environments. Phenotypes were simulated with the following model:

$$\mathbf{y} = \mathbf{Z}_1\mathbf{e} + \mathbf{Z}_2\mathbf{b} + \mathbf{Z}_3\mathbf{g} + \boldsymbol{\varepsilon} \tag{10}$$

where $\mathbf{y}$ is the vector of simulated phenotypes, $\mathbf{e}$ is the vector of environmental effects, $\mathbf{b}$ is a vector of blocks within environments, $\mathbf{g}$ is the vector of genotypic effects and $\boldsymbol{\varepsilon}$ is the vector of errors. The capital letters represent the incidence matrices for their respective effects. $\mathbf{e}$, $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ were simulated to follow a normal distribution with mean zero and variance σ^2_e, σ^2_b and σ^2_ε, respectively. We simulated $\mathbf{g}$ from a multivariate normal distribution with mean zero and variance-covariance $(\sigma^2_g\mathbf{J} + \sigma^2_{ge}\mathbf{I})$, where σ^2_g and σ^2_{ge} are the variances of genotype and genotype-by-environment interaction effects, $\mathbf{J}$ is a matrix of ones, and $\mathbf{I}$ is an identity matrix. The dimensions of both matrices depend on the number of environments. Empirical estimates of variance components were obtained from Chaves et al. (2023) (Table 3).

Table 3: Estimates of the variance components (kg^2) used for simulation. σ^2_e is the variance of the environmental effects, σ^2_b is the variance of block effects, σ^2_g is the genotypic variance, σ^2_{ge} is the variance of the genotype-by-environment interaction effects, and σ^2_ε is the residual variance.

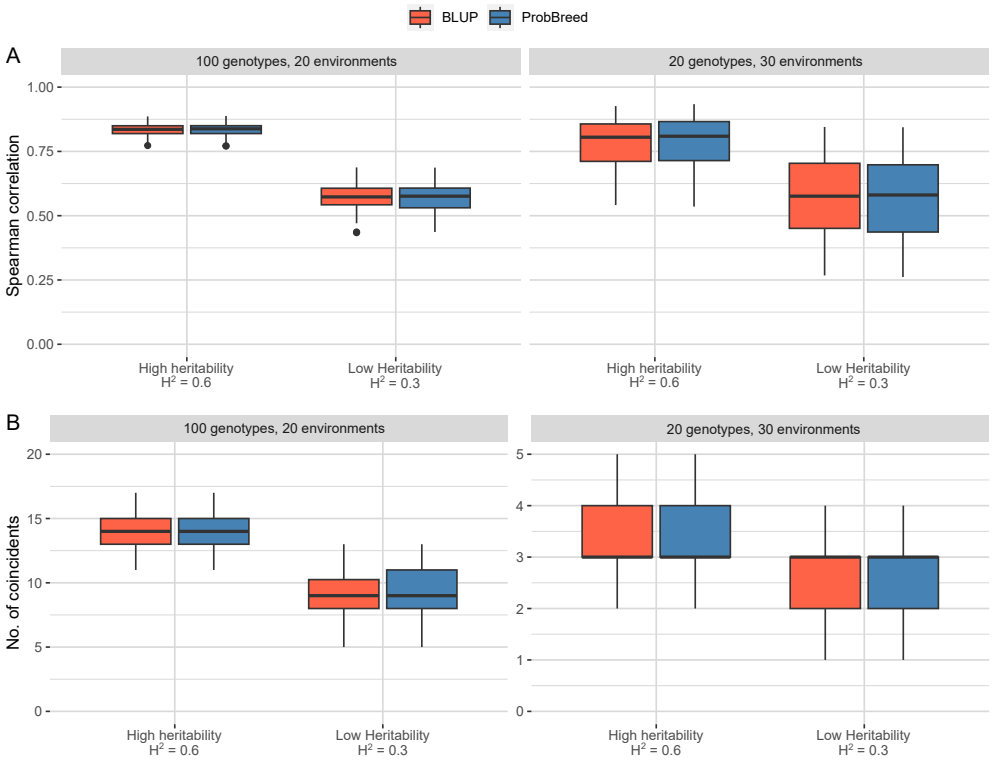
Variance component	**Estimate**
σ^2_e	5.3137
σ^2_b	0.1075
σ^2_g	0.1215
σ^2_{ge}	0.3907
σ^2_ε	1.1420

In *ProbBreed*, we set Bayesian models with a default of 2000 iterations and 4 Markov chains. The simulated data was analysed on a computer with 67 GB DDR4 of memory and a 12th Gen Intel® Core™ i7-12100 processor with a base frequency of 2.10 GHz. We registered the average time to fit the models using the `get_elapsed_time` function of `rstan`. Once the models were fitted, we followed the pipeline previously to obtain the probability metrics. Then, we performed the ranking correlation between estimated probabilities and simulated genotypic values. We also verified the number of coincident genotypes when selecting the top 20 genotypes via probabilities and simulated genotypic values for the datasets with 100 genotypes, whereas the top 5 genotypes were considered for the datasets with 20 genotypes. For comparison, the datasets were analysed with a modified version of Model 10 using a Frequentist linear mixed model with `ASReml-R` (version 4.2, THE VSNI TEAM, 2023), where **e** and **b** were considered as fixed effects. Using residual maximum likelihood, the Frequentist model predicted the best linear unbiased predictions (eBLUPs) of each genotype (HENDERSON, 1975; PATTERSON; THOMPSON, 1971). These eBLUPs were also compared with the simulated genotypic values using ranking correlations and detecting the coincidence among the selected genotypes, as previously described.

The results refer to the average across simulations. Both `ProbBreed` and BLUP yield more reliable results when analysing traits with high heritability. The difference between the rank correlations in the high- and low-heritability scenarios was 0.27 units and 0.22 units in datasets with 100 genotypes and 20 environments, and 20 genotypes and 30 environments, respectively (Figure 9A). Similarly, the difference was evident in the number of coincident selected genotypes (Figure 9B).

The results indicated no difference in selection/ranking between the probability of superior performance via `ProbBreed` and BLUP via the frequentist linear model regardless of the dataset

Figure 9: Comparisons between the simulated genotypic value and both the marginal probabilities of superior performance via `ProbBreed` (in blue) and the BLUPs via `ASReml-R` (in red): Spearman (rank) correlations (A) and number of coincident selected genotypes (B). The name of each facet and the text in the *x*-axes describe the simulated scenario. In (B), we considered the top 20 and the top 5 genotypes in the scenarios with 100 genotypes assessed in 20 environments and 20 genotypes evaluated in 30 environments, respectively.



size and the heritability level (Figure 9). In fact, it is well-known that when weakly informative priors are used, Bayesian and Frequentist models are equivalent (SORENSEN; GIANOLA, 2002). Thus, results from `ProbBreed` are comparable to the ones of other Frequentist methods in the context of multi-environmental trials. Despite the similarity, using probability metrics as criteria for selection has benefits that can aid in decision-making, such as the pairwise and joint probabilities of superior performance and stability. Future versions of the package may allow the consideration of different priors, which can increase the differences between `ProbBreed` and Frequentist models.

Computational requirements

The models used to analyse the simulated datasets with 100 genotypes and 20 environments took about 10 hours to fit, six dedicated to the warm-up iterations and four to the sampling iterations. Conversely, in the smaller simulated dataset needed, the computational time was reduced to 30 minutes (20 minutes warming up and 10 minutes sampling).

4 Concluding remarks

`ProbBreed` is a work in progress. The functionalities described in this paper can and will likely be improved, as well as other resources introduced in the future. Recommendations and suggestions from users are welcome. The computational time required to fit the Bayesian model is currently a limiting factor that should be emphasized. This time depends mainly on the processing capacity of the machine, the number of iterations, cores and chains set in `bayes_met`, and on the number of genotypes, locations, years, and regions.

In summary, `ProbBreed` is a user-friendly package for employing the risk/probability method proposed by Dias et al. (2022) for selecting genotypes in MET. We believe the package's accessibility combined with the advantages of the Bayesian approach will encourage its adoption in the plant breeding community. The main advantage of using `ProbBreed` is effective decision-making for cultivar recommendation in MET. We recommend its usage mainly in late-stage breeding trials when a few dozen genotypes are evaluated in several environments.

References

ANNICCHIARICO, P. Cultivar adaptation and recommendation from alfalfa trials in Northern Italy. **Journal of Genetics and Breeding (Italy)**, v. 46, p. 269–278, 1992.

BARAH, B. C.; BINSWANGER, H. P.; RANA, B. S.; RAO, N. G. P. The use of risk aversion in plant breeding; Concept and application. **Euphytica**, v. 30, n. 2, p. 451–458, 1981.

BÜRKNER, P.-C.; GABRY, J.; KAY, M.; VEHTARI, A. **posterior: Tools for Working with Posterior Distributions**. 2023.

CHAVES, S. F. S.; EVANGELISTA, J. S. P. C.; TRINDADE, R. S.; DIAS, L. A. S.; GUIMARÃES, P. E.; GUIMARÃES, L. J. M.; ALVES, R. S.; BHERING, L. L.; DIAS, K. O. G. Employing factor analytic tools for selecting high-performance and stable tropical maize hybrids. **Crop Science**, v. 63, n. 3, p. 1114–1125, 2023.

COOPER, M.; DELACY, I. H. Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. **Theoretical and Applied Genetics**, v. 88, n. 5, p. 561–572, 1994.

DIAS, K. O. G.; SANTOS, J. P. R.; KRAUSE, M. D.; PIEPHO, H.-P.; GUIMARÃES, L. J. M.; PASTINA, M. M.; GARCIA, A. A. F. Leveraging probability concepts for cultivar recommendation in multi-environment trials. **Theoretical and Applied Genetics**, v. 135, n. 4, p. 1385–1399, 2022.

ESKRIDGE, K.; BYRNE, P.; CROSSA, J. Selection of stable varieties by minimizing the probability of disaster. **Field Crops Research**, v. 27, n. 1-2, p. 169–181, 1991.

FABRETI, L. G.; HÖHNA, S. Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation. **Methods in Ecology and Evolution**, v. 13, n. 1, p. 77–90, 2022.

GABRY, J.; SIMPSON, D.; VEHTARI, A.; BETANCOURT, M.; GELMAN, A. Visualization in Bayesian Workflow. **Journal of the Royal Statistical Society Series A: Statistics in Society**, v. 182, n. 2, p. 389–402, 2019.

GABRY, J.; VEEN, D. **shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models**. 2022.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian Data Analysis**. 3. ed.: Chapman and Hall/CRC, 2013.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, v. 7, n. 4, p. 457–472, 1992.

HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a selection model. **Biometrics**, v. 31, n. 2, p. 423, 1975.

HOFFMAN, M. D.; GELMAN, A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. **Journal of Machine Learning Research**, v. 15, p. 1593–1623, 2014.

KRAUSE, M. D.; DIAS, K. O. G.; SINGH, A. K.; BEAVIS, W. D. Using soybean historical field trial data to study genotype by environment variation and identify mega-environments with the integration of genetic and non-genetic factors. **bioRxiv : the preprint server for biology**, 2023.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. 1. ed. Sunderland: Sinauer Associates, 1998.

MEAD, R.; RILEY, J.; DEAR, K.; SINGH, S. P. Stability comparison of intercropping and monocropping systems. **Biometrics**, v. 42, n. 2, p. 253–266, 1986.

NISHIO, M.; ARAKAWA, A. Performance of Hamiltonian Monte Carlo and No-U-Turn Sampler for estimating genetic parameters and breeding values. **Genetics Selection Evolution**, v. 51, n. 1, p. 73, 2019.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545–554, 1971.

R CORE TEAM. **R: A Language and environment for statistical computing**. Viena, Áustria: R Foundation for Statistical Computing, 2023.

SHUKLA, G. K. Some statistical aspects of partitioning genotype-environmental components of variability. **Heredity**, v. 29, n. 2, p. 237–245, 1972.

SIEVERT, C. **Interactive Web-Based Data Visualization with R, plotly, and shiny**. Chapman and Hall/CRC, 2020.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. New York: Springer-Verlag, 2002. (Statistics for biology and health).

STAN DEVELOPMENT TEAM. **RStan: the R interface to Stan**. 2023.

STAN DEVELOPMENT TEAM. **Stan Modeling Language Users Guide and Reference Manual**. 2023.

THE VSNI TEAM. **asreml: Fits Linear Mixed Models using REML**. 2023.

VEHTARI, A.; GELMAN, A.; SIMPSON, D.; CARPENTER, B.; BÜRKNER, P.-C. Rank-Normalization, Folding, and Localization: An Improved R^ for Assessing Convergence of MCMC (with Discussion). **Bayesian Analysis**, v. 16, n. 2, p. 667–718, 2021.

WATANABE, S. A widely applicable bayesian information criterion. v. 14, p. 867–897, 2013.

WICKHAM, H. **ggplot2: Elegant graphics for data analysis**. 2. ed. Cham: Springer, 2016.

YAN, W.; HUNT, L.; SHENG, Q.; SZLAVNICS, Z. Cultivar evaluation and mega-environment investigation based on the GGE Biplot. **Crop Science**, v. 40, n. 3, p. 597–605, 2000.