MAURÍCIO DOS SANTOS ARAÚJO

INTEGRATING STATISTICAL GENETICS, GEOGRAPHICAL INFORMATION SYSTEMS AND ENVIROTYPING: A NOVEL APPROACH FOR PREDICTIVE BREEDING AND DECISION-MAKING

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

Advisor: Luiz Antônio dos Santos Dias

Co-advisors: Kaio Olimpio das Graças Dias and Rodrigo Silva Alves

VIÇOSA - MINAS GERAIS
2024

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

T
A663i
2024

Araújo, Maurício dos Santos, 1995-
Integrating statistical genetics, geographical information systems and envirotyping: a novel approach for predictive breeding and decision-making / Maurício dos Santos Araújo. – Viçosa, MG, 2024.
1 tese eletrônica (79 f.): il. (algumas color.).

Texto em inglês.
Inclui apêndice.
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.2024.091
Modo de acesso: World Wide Web.

1. Interação genótipo-ambiente. 2. Genética - Métodos estatísticos. 3. Sistemas de informação geográfica. 4. Processo decisório. I. Dias, Luiz Antônio dos Santos, 1957-. II. Universidade Federal de Viçosa. Departamento de Agronomia. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 631.5233

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

MAURÍCIO DOS SANTOS ARAÚJO

INTEGRATING STATISTICAL GENETICS, GEOGRAPHICAL INFORMATION SYSTEMS AND ENVIROTYPING: A NOVEL APPROACH FOR PREDICTIVE BREEDING AND DECISION-MAKING

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

APPROVED: February 27th, 2024.

Assent:

Documento assinado digitalmente
gov.br
MAURICIO DOS SANTOS ARAUJO
Data: 01/03/2024 11:10:06-0300
Verifique em https://validar.iti.gov.br

Maurício dos Santos Araújo
Author

Documento assinado digitalmente
gov.br
LUIZ ANTONIO DOS SANTOS DIAS
Data: 01/03/2024 12:06:25-0300
Verifique em https://validar.iti.gov.br

Luiz Antônio dos Santos Dias
Advisor

To my mother, Maria do Espírito Santo, and to my father, Francisco Carvalho de Araújo, in acknowledgment of the unconditional love and unwavering support they have always provided me.

DEDICATION

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude, especially to my mother, Maria do Espírito Santo, for her unconditional love and friendship. To my father, Francisco Carvalho de Araújo, for providing comprehensive support in all aspects of my life. To my sisters, Danielle dos Santos Araújo, Verônica dos Santos Araújo, and Glayciane dos Santos Carvalho, to whom I dedicate my deep affection.

My thanks also extend to my advisor, Prof. Dr. Luiz Antônio dos Santos Dias, whose knowledge, conversations, and advice were crucial over the four years of my doctoral journey. To my co-advisor, Prof. Dr. Kaio Olimpio das Graças Dias, who played a direct role in this thesis, I am grateful for the enriching conversations and discussions that significantly contributed to my professional and personal development. I also appreciate Dr. Rodrigo Silva Alves for the initial ideas that drove the development of this thesis.

My gratitude extends to the professors of the Graduate Program in Genetics and Breeding (PPGGM/UFV), whose teachings were of paramount importance to my academic formation. To the members of the qualifying committee, especially Dr. Germano Costa-Neto, I am thankful for the valuable advice and insights throughout the entire process of this thesis. I thank André Ricardo Gomes Bezerra for providing essential soybean data for this work, as well as Dr. Matheus Dalsente Krause for assistance and support in constructing this thesis. To Dr. Alexandre Bryan Heinemann, I express gratitude for providing rice data and contributions that enhanced the work.

I acknowledge all friends and colleagues at the Universidade Federal de Viçosa (UFV), where I learned and thrived in an enriching environment. I highlight the collaboration and friendship of Saulo Fabrício da Silva Chaves, whose assistance was fundamental in the elaboration of this thesis.

I express my thanks to the members of the Grupo de Estudos em Genética e Melhoramento (GenMelhor), where I participated for three terms, for providing valuable opportunities in my career and allowing me to interact with a variety of people. To the members of the Agroenergy Laboratory, I extend thankfulness for friendship and knowledge exchange throughout this doctorate.

Finally, I acknowledge UFV and, especially, PPGGM/UFV, the Agroenergia Laboratory, the Biometria Laboratory, Statistical Genetics and Computational Biology Laboratory, Fundação MS, Embrapa Rice and Beans, and the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for the scholarship that enabled the completion of my doctorate.

"The act of discovery consists of looking at what everyone else is seeing and thinking something different." - Albert Szent-Györgyi

ABSTRACT

ARAÚJO, Mauricio dos Santos, D.Sc., Universidade Federal de Viçosa, February, 2024. **Integrating statistical genetics, geographical information systems and envirotyping: a novel approach for predictive breeding and decision-making.** Advisor: Luiz Antônio dos Santos Dias. Co-adivisors: Kaio Olimpio das Graças Dias and Rodrigo Silva Alves.

The crossover genotype-by-environments (G × E) interaction is responsible for the variation in genotype performance across different environments. Disregarding the effect of this interaction means neglecting the specific adaptations of genotypes in the target population of environments. Environmental characterization enables an understanding of the specificities and similarities among different environments. In this context, enviromics has emerged as a new area that integrates information from data analysis, quantitative genetics, geographic information systems (GIS), and principles of ecophysiology. Incorporating environmental features into Statistical Genetics models contributes to enhancing the predictive capability of these models and a better understanding of the cultivation environment. Thus, this study aimed to propose a new predictive breeding method (GIS-FA) that integrates GIS information, factor analytic (FA) models, partial least squares regression (PLS), and enviromics to predict purelines of rice and soybean in untested environments. Two databases were used: one for rice, with 80 purelines cultivated in the years 2009/10 and 2010/11 in 21 environments in eight Brazilian states; and the second composed of 195 soybean purelines evaluated in 49 environments (in years 2019/20, 2020/21, and 2021/22) in the state of Mato Grosso do Sul. The term “environment” refers to the location-year combination. For both datasets, FA models were adjusted, with FA4 being selected based on the average semivariance ratio. A total of 32 environmental features (EF) were collected, including three geographical, 16 climatic, and 13 soil-related features. To make predictions, 50 points were randomly chosen within each municipality in the evaluated states, and, for each point, EFs data were obtained from a historical series (2000-2021). Leveraging the FA model outcomes, we used the PLS method to predict the overall performance and stability of both crops in untested environments. Cross-validation was performed using the leave-one-out method, and subsequently, the GIS-FA method was compared with the GGE-GIS approach, which uses directly the within-environment eBLUPs to perform the prediction. After the spatial prediction, performance and stability parameters were represented in thematic maps. For predicting eBLUEs, GIS-FA was 10% and 1% superior to GIS-GGE in the rice and soybean datasets, respectively. For predicting eBLUPs, GIS-FA was 9% and 5% more effective than GIS-GGE. Three types of maps were created: (i) zones of genotype adaptation; (ii) pairwise comparison between pureline `vs.` check and pureline `vs.` pureline; (iii) which-won-where. The GIS-FA approach proved to be efficient in predicting genotypes for untested environments, allowing the evaluation of the G × E interaction throughout the experimental network.

Keywords: Environmental features. Factor analytic. Predictive models. Partial Least Squares.

RESUMO

ARAÚJO, Mauricio dos Santos, D.Sc., Universidade Federal de Viçosa, fevereiro, 2024. **Integrando genética estatística, sistemas de informação geográfica e tipagem ambiental: uma nova abordagem para melhoramento preditivo e tomada de decisão.** Orientador: Luiz Antônio dos Santos Dias. Coorientadores: Kaio Olimpio das Graças Dias e Rodrigo Silva Alves.

A interação genótipos por ambientes (G $\times$ A) complexa é responsável pela variação no desempenho dos genótipos em diferentes ambientes. Desconsiderar o efeito dessa interação significa, pode negligenciar as adaptações específicas dos genótipos em população de ambientes alvo. A caracterização ambiental possibilita a compreensão das especificidades e semelhanças entre diferentes ambientes. Nesse contexto, a *enviromics* surgiu como uma nova área que integra informações de análise de dados, genética quantitativa, sistemas de informação geográfica (SIG) e princípios de ecofisiologia. Incorporar características ambientais (CA) em modelos de genético estatístico contribui para aprimorar a capacidade preditiva desses modelos e para uma melhor compreensão do ambiente de cultivo. Assim, este estudo teve como objetivo propor um novo método de melhoramento preditivo (SIG-FA) que integra informações SIG, modelos de análise fator analítico (FA), regressão de mínimos quadrados parciais (PLS) e *enviromics* para predizer linhas puras de arroz e soja em ambientes não testados. Foram utilizados dois bancos de dados: um para arroz, com 80 linhas puras cultivadas nos anos 2009/10 e 2010/11 em 21 ambientes em oito estados brasileiros; e o segundo composto por 195 linhas puras de soja avaliadas em 49 ambientes (nos anos 2019/20, 2020/21 e 2021/22) no estado de Mato Grosso do Sul. O termo "ambiente" refere-se à combinação local-ano. Para ambos os conjuntos de dados, foram ajustados modelos FA, sendo selecionado o FA4 com base na razão média de semivariância. Um total de 32 CA foram coletadas, incluindo três geográficas, 16 climáticas e 13 relacionadas ao solo. Para fazer a predição, 50 pontos foram coletados aleatoriamente dentro de cada município nos estados avaliados, e, para cada ponto, os dados de CA foram obtidos de uma série histórica (2000-2021). Utilizando os resultados do modelo FA, empregamos o método PLS para predizer o desempenho geral e a estabilidade de ambas as culturas em ambientes não testados. A validação cruzada foi realizada usando o método *leave-one-out*, e posteriormente, o método SIG-FA foi comparado com a abordagem GGE-GIS, que utiliza diretamente os eBLUPs dentro do ambiente para realizar a predição. Após a predição espacial, os parâmetros de desempenho e estabilidade foram representados em mapas temáticos. Para predizer eBLUEs, o SIG-FA foi 10% e 1% superior ao GIS-GGE nos conjuntos de dados de arroz e soja, respectivamente. Para predizer eBLUPs, o SIG-FA foi 9% e 5% mais eficaz do que o GIS-GGE. Três tipos de mapas foram criados: (*i*) zonas de adaptação do genótipo; (*ii*) comparação par a par entre linhas puras `vs.` testemunha e linha pura `vs.` linha pura; (*iii*) quem-ganhou-onde. A abordagem SIG-FA mostrou-se eficiente na predição de genótipos para ambientes não testados, permitindo a avaliação da interação G $\times$ A em toda a rede experimental.

Palavras-chave: Características ambientais. Análise fator analítico. Modelos preditivos. Mínimos quadrados parciais.

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AMMI	additive main effects and multiplicative interaction
ASR	average semivariance ratio
CFSR	climate forecast system reanalysis
CRU TS	climate research unit time-series
CHELSA	climatologies at high resolution for the earth's land surface areas
CV	cross-validation
EOSDIS	earth observing system data and information system
eBLUPs	empirical best linear unbiased predictions
ECMWF	European Centre for medium-range weather forecasts
FA	factor analytic
FAST	factor analytic selection tools
FR	factorial regression
GEI	genotype-by-environment interaction
GIS	geographic information system
GHCN	global historical climatology network
IDW	inverse distance weighting
LogL	logarithm of the likelihood function
METs	multi-environment trials
MVN	multivariate normal distribution
NCEI	National Centers for environmental information
no. par.	number of parameters
n	observations
OP	overall performance
PLS	partial least square

PEV	prediction error variance
p	predictor variables
pp	percentage points
QTL	quantitative trait loci
RMSD	root mean square deviation
SI	selection index
TPE	target population of environments

LIST OF SYMBOLS

β	Lower case Greek letter beta
ϵ	Lower case Greek letter epsilon
η	Lower case Greek letter eta
Γ	Upper case Greek letter Gamma
Λ	Upper case Greek letter Lambda
λ	Lower case Greek letter lambda
Ω	Upper case Greek letter Omega
$\oplus$	Plus-minus symbol
$\otimes$	Math symbol Kronecker product
Ψ	Upper case Greek letter Psi
ρ	Lower case Greek letter rho
σ	Lower case Greek letter sigma
$\sim$	Math symbol tilde
$\sum$	Summation symbol
ζ	Lower case Greek letter zeta

SUMMARY

I

CHAPTER I

1 GENERAL INTRODUCTION

1.1 Genotype-by-environment interaction and Statistical Genetics Models

Plant breeders use multi-environment trials (METs) to elucidate the effect of genotype-by-environment interaction (GEI) on the selection of superior genotypes (SMITH; CULLIS; THOMPSON, 2005). These trials are important to (i) represent the target population of environments (TPE) or future cultivation environments; (ii) assessing genotype performance under multi-environmental conditions; and (iii) providing insights into genotype adaptation to specific or multi-environments (ISIK; HOLLAND; MALTECCA, 2017; MALOSETTI et al., 2016; COSTA-NETO et al., 2023). In the presence of crossover interaction, the ranking of genotypes across different environments changes (FEHR, 1987; COOPER; DELACY, 1994). Ignoring GEI interaction in the improvement process can lead to biased results and underestimate the effectiveness of selection practices (EEUWIJK; BUSTOS-KORTS; MALOSETTI, 2016).

Over the years, several methods have been proposed to measure the GEI. Each approach has its peculiarities and assumptions. Methods based on analysis of variance (PLAISTED; PETERSON, 1959; SHUKLA, 1972), regression models (FINLAY; WILKINSON, 1963; EBERHART; RUSSELL, 1966), non-parametric methods (LIN; BINNS, 1998), multiplicative approaches such as the genotype main effect plus genotype-by-environment interaction (GGE) Biplot (YAN et al., 2000), and additive main effect and multiplicative interaction (AMMI) (GAUCH; ZOBEL, 1997; GAUCH-JR, 2006), based on Linear mixed models (HENDERSON, 1949; HENDERSON, 1950), factor analytic (FA) linear mixed models (PIEPHO, 1997; PIEPHO, 1998; SMITH; CULLIS; THOMPSON, 2001), and Bayesian approaches (DIAS et al., 2022), have been explored in plant breeding.

Particularly, the FA models demonstrate robustness in dealing with various data structures, especially unbalanced ones. They are a parsimonious approximation of the unstructured model, as they indirectly build the complete genetic covariance structure (heterogeneous variances and covariances). This resource enables the exploration of the genetic covariance between environments or traits, providing a good fit for data of METs. This is possible due to the dimensionality reduction provided by the latent variable, namely factors (SMITH; CULLIS; THOMPSON, 2001; PIEPHO, 1998). Furthermore, since they are linear mixed models, they enable the inclusion of relatedness information, whether genomic (marker) or ancestral (pedigree) (SMITH; CULLIS; THOMPSON, 2005). Smith e Cullis (2018) introduced the Factor Analytic Selection Tools (FAST), which incorporate parameters for easily assessing the overall performance (OP) and stability (via RMSD, Root Mean Square Deviation). These parameters are valuable in assisting breeders in decision-making, offering a more informed and statistically grounded approach to data analysis. The FA models are currently the benchmark for dealing with unbalanced METs data using linear mixed models (CHAVES et al., 2023b; BAKARE et al., 2022; TOLHURST et al., 2022; CALLISTER et al., 2024).

1.2 Enviromics and Spatial Prediction

Apart from choosing the most appropriate statistical method, modern plant breeding requires further tools to address the challenge of improving the predictive ability of models. In the last decade, environmental features proved themselves to be valuable for this purpose when dealing with METs (RESENDE et al., 2021; XU, 2016). The idea of using environmental features in genetic analyses is no novelty (EEUWIJK; ELGERSMA, 1993; WOOD, 1976). Advances in hardware technology enabled the usage of big datasets, facilitating the integration of environmental features into statistical genetics models. The nowadays challenges are regarding improving the predictive ability of models by incorporating environmental features and drawing thematic maps that facilitate the recommendation process (MILLET et al., 2019; COSTA-NETO et al., 2020; BUSTOS-KORTS et al., 2022). Furthermore, strategies in defining breeding zones allow the selection of the best genotypes and maximization of the genetic gain from selection (RESENDE et al., 2021; CALLISTER et al., 2024).

Enviromics is a specialized field that combines environmental databases with statistics and quantitative genetics, utilizing knowledge of plant ecophysiology to better understand how the environment can impact plant development and the plasticity of important agronomic traits (COSTA-NETO; FRITSCHE-NETO, 2021). Accordingly, envirotypes are all sources of environmental variations related to the development of the plants that can act as environmental markers in statistical genetics models to predict genotypic effects in non-evaluated environments (XU, 2016; RESENDE et al., 2021).

The addition of information derived from Geographic Information System (GIS) techniques into predictive models has been encouraged to improve the efficiency of breeding programs (GUARINO et al., 2002). An initial effort was made by Booth (1990) aiming to indicate climatically suitable regions for the introduction of tree species at a global scale based on the environmental conditions where they were collected. Other studies employed soil maps to overlap breeding trial information aiming to match potential candidates from extensive maize germplasm that could adapt to alkaline soils (CHAPMAN; BARRETO et al., 1996), and to identify phosphorus-efficient common bean genotypes (BEEBE et al., 1997). Annicchiarico, Bellah e Chiari (2006) assessed how GIS-based methodologies could aid the recommendation of durum wheat genotypes in MET, as compared to traditional methodologies. Recently, sophisticated statistical approaches are being developed by using information from breeding programs and GIS tools to predict the phenotypic performance of unobserved genotypes in untested environments (COSTA-NETO; FRITSCHE-NETO, 2021; COSTA-NETO et al., 2020). For example, High-dimensional genomic and environmental data were tested in the genomic selection framework by applying reaction norm models (JARQUIN et al., 2014) and random regression models (TOLHURST et al., 2022).

The importance of using environmental information to exploit GEI was further evidenced by Piepho, Denis e Eeuwijk (1998) by comparing a few methodologies, and the benefits of combining environmental information and molecular markers' information to identify QTLs that are specifically associated with environmental conditions was discussed by Crossa et al. (1999). Environmental features were also incorporated with crop growth modelling into the genomic selection framework, improving the accuracy of genotype predictions across a range of environments (HESLOT et al., 2014). The increasing interest in using environmental features is linked to advances in computational technology and the availability of high throughput environmental data at low cost through GIS techniques. Many recent studies incorporated information obtained via GIS into prediction models (COOPER; MESSINA, 2021; ROGERS et al., 2021; DIEPENBROCK et al., 2022), which improved the model's predictive ability.

Resende et al. (2021) discuss a concept called "envirotypic-assisted selection" which integrates genomic data with environmental information to improve selection accuracy in plant breeding using interpolation techniques. A GIS-based tool methodology applied factorial regression (FR) to model spatial trends and draw thematic maps of yield adaptability to upland rice in Brazil (COSTA-NETO et al., 2020). Another method was proposed using non-linear kernels to model additive, dominance, and GEI effects, using geographically referenced information (COSTA-NETO; FRITSCHE-NETO; CROSSA, 2021). Due to the importance of these works, among others, emerges a new field of study called enviromics.

The use of several sources of information from breeding programs (phenotypic and genomic data) and from GIS tools (envirotypes) can generate two statistical problems. The first is multicollinearity due to the high relationship between the predictor variables (p), and the second is that the number p can be greater than the number of observations (n). In these situations, methods such as partial least square (PLS) regression can be applied, since it can identify linear combinations of environmental features strongly correlated with the phenotypic traits by using latent variables, known as components (MONTESINOS-LOPEZ et al., 2022b). The components are linear combinations of the original variables that capture the underlying structure in the data that is most relevant to explain the observed variation (MONTESINOS-LOPEZ et al., 2022a; CALLISTER et al., 2024). Chapter II contains the proposal of a new enviromics prediction model for recommending cultivars based on FA models, PLS and thematic maps.

References

ANNICCHIARICO, P.; BELLAH, F.; CHIARI, T. Repeatable genotype × location interaction and its exploitation by conventional and gis-based cultivar recommendation for durum wheat in algeria. *European Journal of Agronomy*, v. 24, p. 70–81, 2006. 22, 32

BAKARE, M. A. et al. Parsimonious genotype by environment interaction covariance models for cassava *Manihot esculenta*. *Frontiers in Plant Science*, v. 13, p. 978248, 2022. 22

BEEBE, S. et al. A geographical approach to identify phosphorus-efficient genotypes among landraces and wild ancestors of common bean. *Euphytica*, v. 95, p. 325–338, 1997. 22, 32

BOOTH, T. H. Mapping regions climatically suitable for particular tree species at the global scale. *Forest Ecology and Management*, v. 36, n. 1, p. 47–60, 1990. 22

BUSTOS-KORTS, D. et al. Identification of environment types and adaptation zones with self-organizing maps: applications to sunflower multi-environment data in Europe. *Theoretical and Applied Genetics*, v. 135, p. 2059–2082, 2022. 22, 60

CALLISTER, A. N. et al. Enviromic prediction enables the characterization and mapping of *Eucalyptus globulus* Labill breeding zones. *Tree Genetics & Genomes*, v. 20, p. 3, 2024. 22, 23, 59, 60

CHAPMAN, S.; BARRETO, H. et al. Using simulation models and spatial databases to improve the efficiency of plant breeding programs. *Plant Adaptation and Crop Improvemententent*, p. 563–587, 1996. 22, 32

CHAVES, S. F. S. et al. Employing factor analytic tools for selecting high-performance and stable tropical maize hybrids. *Crop Science*, v. 63, n. 3, p. 1114–1125, 2023. 22, 43, 57, 58

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, p. 561–572, 1994. 21, 30, 42, 57

COOPER, M.; MESSINA, C. D. Can we harness "enviromics" to accelerate crop improvement by integrating breeding and agronomy? *Frontiers in Plant Science*, v. 12, p. 735143, 2021. 23, 31

COSTA-NETO, G. et al. Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. *G3 Genes|Genomes|Genetics*, v. 13, n. 2, p. jkac313, 2023. 21

COSTA-NETO, G.; FRITSCHE-NETO, R. Enviromics: bridging different sources of data, building one framework. *Crop Breeding and Applied Biotechnology*, v. 21, n. spe, p. e393521S12, 2021. 22, 23, 32

COSTA-NETO, G.; FRITSCHE-NETO, R.; CROSSA, J. Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity*, v. 126, n. 1, p. 92–106, 2021. 23, 32

COSTA-NETO, G. et al. A novel GIS-based tool to reveal spatial trends in reaction norm: upland rice case study. *Euphytica*, v. 216, p. 37, 2020. 22, 23, 32, 45, 60, 61

CROSSA, J. et al. Interpreting genotype × environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theoretical and Applied Genetics*, v. 99, p. 611–625, 1999. 23, 58, 59

DIAS, K. O. G. et al. Leveraging probability concepts for cultivar recommendation in multi-environment trials. *Theoretical and Applied Genetics*, v. 135, p. 1385–1399, 2022. 21, 31, 61

DIEPENBROCK, C. H. et al. Can we harness digital technologies and physiology to hasten genetic gain in us maize breeding? *Plant Physiology*, v. 188, n. 2, p. 1141–1157, 2022. 23, 31

EBERHART, S. A.; RUSSELL, W. A. Stability parameters for comparing varieties. *Crop Science*, v. 6, p. 36–40, 1966. 21, 31

EEUWIJK, F. A. V.; ELGERSMA, A. Incorporating environmental information in an analysis of genotype by environment interaction for seed yield in perennial ryegrass. *Heredity*, v. 70, n. 5, p. 447–457, 1993. 22, 31, 58

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. 21, 31

FEHR, W. R. (Ed.). *Principles of Cultivars Development*. [S.l.]: Macmillan, 1987. (Contents v. I. Theory and technique; v.2. Crop species). Includes bibliographies and indexes. 21

FINLAY, K.; WILKINSON, G. The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, v. 14, p. 742, 1963. 21, 31

GAUCH, H. G.; ZOBEL, R. Identifying mega-environments and targeting genotypes. *Crop Science*, v. 37, p. 311–326, 1997. 21, 31

GAUCH-JR, H. G. Statistical analysis of yield trials by AMMI and GGE. *Crop Science*, v. 46, p. 1488–1500, 2006. 21

GUARINO, L. et al. Geographic information systems (GIS) and the conservation and use of plant genetic resources. In: CABI PUBLISHING WALLINGFORD UK. *Managing plant genetic diversity. Proceedings of an international conference, Kuala Lumpur, Malaysia, 12-16 June 2000.* [S.l.], 2002. p. 387–404. 22, 32

HENDERSON, C. R. Estimates of changes in herd environment. *Journal of Dairy Science*, v. 61, p. 294–300, 1949. 21, 31, 40

HENDERSON, C. R. Estimation of genetic parameters. *Annals of Mathematical Statistics*, v. 21, p. 309–310, 1950. 21, 31, 40

HESLOT, N. et al. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, v. 127, p. 463–480, 2014. 23

ISIK, F.; HOLLAND, J.; MALTECCA, C. *Genetic data analysis for plant and animal breeding.* [S.l.]: Cham: Springer International Publishing, 2017. 21

JARQUIN, D. et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, v. 127, n. 3, p. 595–607, 2014. 23, 32, 61

LIN, C. S.; BINNS, M. R. A superiority measure of cultivar performance for cultivar × location data. *Canadian Journal of Plant Science*, v. 68, p. 193–198, 1998. 21

MALOSETTI, M. et al. Predicting responses in multiple environments: issues in relation to genotype × environment interactions. *Crop Sci.*, v. 56, p. 2210–2222, 2016. 21

MILLET, E. J. et al. Genomic prediction of maize yield across european environmental conditions. *Nature Genetics*, v. 51, n. 6, p. 952–956, 2019. 22, 31

MONTESINOS-LOPEZ, O. A. et al. Partial least squares enhances genomic prediction of new environments. *Frontiers in Genetics*, v. 13, p. 920689, 2022. 23, 32, 59

MONTESINOS-LOPEZ, O. A. et al. Multi-trait genome prediction of new environments with partial least squares. *Frontiers in Genetics*, v. 13, p. 966775, 2022. 23, 32, 59

PIEPHO, H. P. Analysis of a randomized block design with unequal subclass numbers. *Agronomy Journal*, v. 89, p. 718–723, 1997. 21, 31, 40, 57

PIEPHO, H. P. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics*, v. 95, p. 195–201, 1998. 21

PIEPHO, H. P.; DENIS, J.-B.; EEUWIJK, F. A. van. Predicting cultivar differences using covariates. *Journal of Agricultural, Biological, and Environmental Statistics*, p. 151–162, 1998. 23

PLAISTED, R. L.; PETERSON, L. C. A technique for evaluating the ability of selections to yield consistently in different locations or seasons. *American Potato Journal*, v. 36, n. 11, p. 381–385, 1959. 21

RESENDE, R. T. et al. Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theoretical and Applied Genetics*, v. 134, p. 95–121, 2021. 22, 23, 32

ROGERS, A. R. et al. The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3 Genes|Genomes|Genetics*, v. 11, n. 2, p. jkaa050, 2021. 23, 31

SHUKLA, G. K. Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity*, v. 29, p. 237–245, 1972. 21

SMITH, A. B.; CULLIS, B.; THOMPSON, R. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, v. 57, p. 1138–1147, 2001. 21, 31, 40, 41, 57

SMITH, A. B.; CULLIS, B. R. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, v. 214, p. 143, 2018. 21, 31, 36, 43, 56

SMITH, A. B.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *J. Agric. Sci.*, v. 143, p. 449–462, 2005. 21

TOLHURST, D. J. et al. Genomic selection using random regressions on known and latent environmental covariates. *Theoretical and Applied Genetics*, v. 135, p. 3393–3415, 2022. 22, 23, 31, 41, 60, 61

WOOD, J. T. The use of environmental variables in the interpretation of genotype-environment interaction. *Heredity*, v. 37, n. 1, p. 1–7, 1976. 22, 31

XU, Y. Envirotyping for deciphering environmental impacts on crop plants. *Theoretical and Applied Genetics*, v. 129, p. 653–673, 2016. 22, 32, 35

YAN, W. et al. Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science*, v. 40, p. 597–605, 2000. 21, 31

II

CHAPTER II

2 ARTICLE

GIS-FA: AN APPROACH TO INTEGRATING THEMATIC MAPS, FACTOR-ANALYTIC, AND ENVIROTYPING FOR CULTIVAR TARGETING

2.1 Abstract

Parsimonious methods that capture genotype-by-environment interaction (GEI) in multi-environment trials (MET) are important in breeding programs. Understanding the causes and factors of GEI allows the utilization of genotype adaptations in the target population of environments through environmental features and factor-analytic (FA) models. Here, we present a novel predictive breeding approach called GIS-FA, which integrates geographic information systems (GIS) techniques, FA models, partial least squares (PLS) regression, and enviromics to predict phenotypic performance in untested environments. The GIS-FA approach enables: (i) the prediction of the phenotypic performance of tested genotypes in untested environments, (ii) the selection of the best-ranking genotypes based on their overall performance and stability using the FA selection tools, and (iii) the creation of thematic maps showing overall or pairwise performance and stability for decision-making. We exemplify the usage of the GIS-FA approach using two datasets of rice [*Oryza sativa* (L.)] and soybean [*Glycine max* (L.) Merr.] in MET spread over tropical areas. In summary, our novel predictive method allows the identification of new breeding scenarios by pinpointing groups of environments where genotypes demonstrate superior predicted performance. It also facilitates and optimizes cultivar recommendations by utilizing thematic maps.

Keywords: Enviromics. Genotype-by-environment interaction. Environmental feature. Soybean. Rice.

2.2 Introduction

Crossover interaction refers to changes in the ranking of genotypes caused by the lack of genotypic correlation and negative correlations between environments, which is the most critical source of genotype-by-environment interaction (GEI) for plant breeders (COOPER; DELACY, 1994; CROSSA; YANG; CORNELIUS, 2004). Cultivar development programs for crops evaluate experimental genotypes (i.e., prior to release) in multi-environmental trials (MET) to (*i*) depict GEI patterns for future cultivar placement and (*ii*) increase the accuracy of selection. Therefore, analytical methods that fully explore the GEI patterns from MET are needed for decision-making (MALOSETTI; RIBAUT; EEUWIJK, 2013; EEUWIJK; BUSTOS-KORTS; MALOSETTI, 2016; DIAS et al., 2022; TOLHURST et al., 2022).

The first attempt to consider the GEI in plant breeding was proposed by Yates e Cochran (1938), who decomposed the part due to the interaction from the total phenotypic variation. Later, Finlay e Wilkinson (1963) used marginal environmental means as independent variables in the regression analysis to depict GEI, and several approaches were developed within that framework (EBERHART; RUSSELL, 1966; LI et al., 2018). Multivariate techniques such as additive main effects and multiplicative interaction (AMMI) (GAUCH; ZOBEL, 1997) and the genotype plus GEI (GGE) biplot (YAN et al., 2000) have also been extensively used (YAN et al., 2007; BALESTRE et al., 2009; SILVA et al., 2021). Further model expansions were made possible by the development of the linear mixed model equations (HENDERSON, 1949; HENDERSON, 1950), which allowed for the incorporation of covariance between relatives and environments and the relaxation of assumptions such as homogeneous residual variances (PIEPHO et al., 2008). Factor-analytic (FA) mixed models (PIEPHO, 1997; SMITH; CULLIS; THOMPSON, 2001) can be employed to explore the covariance between environments. These models offer the flexibility to account for heterogeneous genotypic (or genetic) covariances between environments using a few latent variables known as factors (K). In addition to the overall (i.e., across environments) and conditional (i.e., within environments) performance, metrics such as stability and sensitivity can also be computed from FA models to facilitate the decision-making process (STEFANOVA; BUIRCHELL, 2010; CULLIS et al., 2014; DIAS et al., 2018; SMITH; CULLIS, 2018; SMITH et al., 2021).

An extension to statistical models that address GEI involves incorporating environmental information, such as physical and chemical soil properties, as well as environmental features like temperature and rainfall precipitation (TOLHURST et al., 2022). The advantages of integrating environmental features into a prediction model include (i) the capability to untangle environmental determinants and the crossover GEI main drivers and (ii) the ability to predict phenotypic performance in yet-to-be-seen environments (SAE-LIM et al., 2014; OLIVEIRA et al., 2020; TOLHURST et al., 2022). Furthermore, categorizing similar environments into homogeneous groups facilitates resource optimization and the identification of mega-environments (WOOD, 1976; DENIS, 1988; EEUWIJK; ELGERSMA, 1993; MILLET et al., 2019; COSTA-NETO et al., 2021; KRAUSE et al., 2022). Therefore, advances in computational resources, along with the development of geographic information systems (GIS) techniques, are essential for designing novel prediction strategies in MET (COOPER; MESSINA, 2021; ROGERS et al., 2021; COOPER et al., 2022; DIEPENBROCK et al., 2022).

GIS techniques have been defined as computer-based systems used for analyzing and interpreting spatially referenced information (COPPOCK; RHIND, 1991) and are powerful tools in the integration of genetics and environmental information (CHAPMAN; BARRETO et al., 1996; BEEBE et al., 1997; GUARINO et al., 2002; JARQUIN et al., 2014; HERNANDEZ et al., 2019; COSTA-NETO; FRITSCHE-NETO, 2021). For example, Annicchiarico, Bellah e Chiari (2006) identified consistent genotype-by-location interactions using GIS-based models to recommend cultivars for durum wheat in Algeria. Costa-Neto et al. (2020) applied a GIS-based tool with factorial regression to analyze spatial trends and create thematic maps of yield performance for upland rice in Brazil. In addition, Costa-Neto, Fritsche-Neto e Crossa (2021) integrated GIS techniques with non-linear kernels to model additive, dominance, and GEI effects. All the mentioned techniques fall under the umbrella of "envirotypic-assisted selection," which integrates genomic and environmental data to improve the accuracy of selection in plant breeding programs (RESENDE et al., 2021).

The combination of statistics, quantitative genetics, and GIS techniques enabled the introduction of the field of enviromics in the plant breeding community (COOPER et al., 2014; XU, 2016; COSTA-NETO; FRITSCHE-NETO, 2021). Coupled with knowledge from plant ecophysiology, this field aims to describe how the environment impacts plant development and the phenotypic plasticity of important agronomic traits (COSTA-NETO; FRITSCHE-NETO, 2021). Accordingly, envirotypes are all sources of environmental variations related to plant development that can act as environmental markers in statistical genetics models to predict genotypic effects in non-evaluated environments (XU, 2016; RESENDE et al., 2021). However, integrating phenotypic and genomic data with environmental features can generate two statistical problems: high correlation among predictors resulting in multicollinearity and the curse of dimensionality when the number of observations is smaller than the predictors. In these situations, methods such as partial least squares (PLS), which combine features from principal components analysis and multiple regression (WOLD; SJöSTRöM; ERIKSSON, 2001), and Bayesian factor analytic models (NUVUNGA et al., 2019), can be applied to identify linear combinations of predictors that capture the underlying structure of the data (MONTESINOS-LOPEZ et al., 2022a; MONTESINOS-LOPEZ et al., 2022b).

Here, we present a novel predictive breeding approach called GIS-FA that combines FA, PLS, and enviromics to predict the phenotypic performance of experimental genotypes in untested environments. The GIS-FA uses environmental information collected from GIS tools to predict the factor loadings of untested environments via PLS, where the estimated factor loadings from the observed environments are used as the training set. The empirical best linear unbiased predicted values (eBLUPs) of genotypic means in untested environments are then calculated as the linear combination of the predicted loadings via

PLS and genotypic scores from the FA models. We hypothesize that the GIS-FA model has higher prediction accuracy compared to a PLS model trained with eBLUPs within observed environments (henceforth called GIS-GGE). We tested this hypothesis using two MET datasets from Brazil: rice trials located in the Brazilian Savanna (Cerrado) and the Amazon rainforest, as well as soybean trials located in the state of Mato Grosso do Sul. Thus, this study aims to: (*i*) propose the GIS-FA methodology for predicting genotypes' performance in untested environments and compare its predictive ability with the GIS-GGE methodology; (*ii*) apply GIS-FA to select the best-ranking genotypes based on their overall performance (OP) and stability using the FA selection tools; and (*iii*) create thematic maps that illustrate the genotypes' performance across environments in the breeding zone.

2.3 Material and Methods

2.3.1 Phenotypic data

We exemplify the GIS-FA model using two datasets from MET covering tropical areas in Brazil. These trials have been used to make decisions regarding the release of cultivars by both public and proprietary breeding organizations. The soybean dataset contains three years of field trials conducted in the state of Mato Grosso do Sul (represented by triangles in Figure 1), whereas the rice dataset includes two years of field trials conducted across eight states (represented by circles in Figure 1). It is important to note that the variation in elevation varies across the studied area (Figure 1a). This factor, along with latitude and longitude, influences changes in both weather and soil conditions, as indicated by the Köppen-Geiger classification (ALVARES et al., 2013) in Figure 1b and the Brazilian Soil Classification System (SANTOS, 2018) in Figure 1c. Both datasets include field trials planted in the same location and year but during different planting seasons. Thus, henceforth, the term "environment" refers to the combination of location, year, and planting season. Another common characteristic shared by both datasets is that not all genotypes were evaluated in all environments (Figure S10 and S11). This has three main reasons: i) seed availability, ii) discarding low-performing lines at the end of each agricultural year, and iii) including cultivars/genotypes from partner breeding programs for evaluation in the target population of environments (TPE). It is expected that the inclusion/exclusion of selected candidates in the MET does not yield relevant bias in the variance component estimates (PIEPHO; MöHRING, 2006; HARTUNG; PIEPHO, 2021).

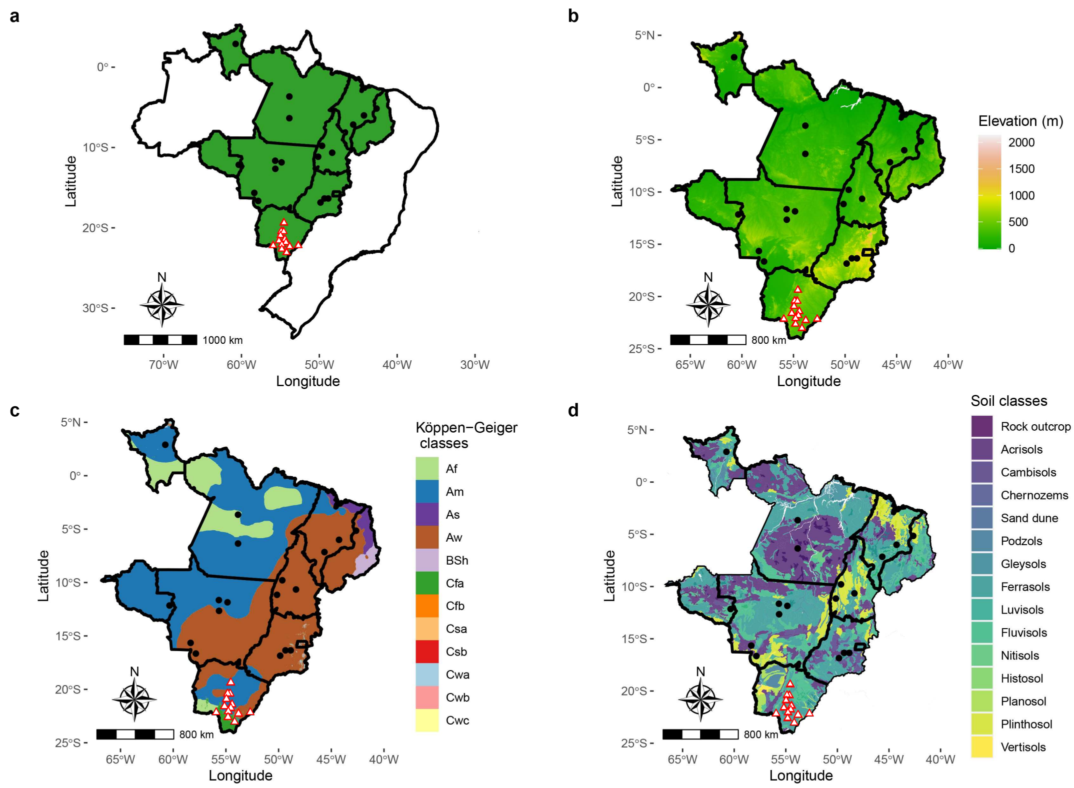


Figure 1 – Maps of the studied area. Figure "a" shows the map of Brazil, highlighting the states where the rice (dots) and soybean (triangles) trials were conducted. We subset these states in Figures "b", "c" and "d." Figure "b" depicts the elevation in meters, Figure "c" displays the Köppen-Geiger classification Alvares et al. (2013), and Figure "d" highlights the Brazilian soil classification adapted to FAO classification (SANTOS, 2018; FAO, 2014).

2.3.1.1 Rice dataset

The rice dataset is composed of 80 pure lines developed by the Brazilian Agricultural Research Corporation (Embrapa Rice and Beans). These pure lines plus three commercial cultivars were evaluated for their value of cultivation and use (VCU) in 21 environments during the cropping seasons of 2009/2010 and 2010/2011. Candidate cultivars that demonstrate high yield and agronomic stability in the TPE will be registered for commercial use. The TPE of the Upland Rice Breeding Program is located within the geographical coordinates of 1° North to 17° South and 42° West to 70° West. It includes eight states from the Mid-West (Mato Grosso and Goiás), the Northeast (Maranhão and Piauí), and the North (Pará, Rondônia, Roraima, and Tocantins). Further details are presented in Supplementary Table 1. Eighteen locations were sampled in the TPE (Figure 1), where trials were arranged in randomized complete blocks with four replications. Experimental plots consisted of four 5 m rows spaced 0.3 m apart, totaling an area of 6 m^2, with 60 seeds sown per meter. Seed yield (kg ha^{-1}) was measured in the two central rows. Management practices in these regions followed the technical recommendations adopted for upland rice.

2.3.1.2 Soybean dataset

The soybean dataset comprises 195 pure lines that were evaluated over three cropping seasons (2019/20, 2020/21, and 2021/22) at 13 locations in the state of Mato Grosso do Sul and the Central-West region of Brazil (Figure 1). Trials were conducted under rainfed conditions and overseen by the Mato Grosso do Sul Foundation (Fundação MS) in 49 different environments. The experimental design involved randomized, complete blocks with three replications. The plots consisted of five 12 m-long rows spaced 0.5 m apart, with a total area of 30 m^2. Seed yield (kg ha^{-1}) was measured in the three central rows and corrected for 13% moisture. Weed and pest control were carried out following the recommendations for the region.

2.3.2 GIS-FA workflow

Here, we will summarize the procedures for applying the GIS-FA methodology. The method was created to evaluate the OP and stability of genotypes in untested environments and to plot the spatial prediction on thematic maps. This enables breeders to define strategies for recommending adaptable cultivars, prospect new target environments that maximize genetic gain through selection, and define breeding zones based on the pattern of environmental features. The procedures to apply the GIS-FA are:

- **Step 1 – Geographic data collection from tested and untested environments**: To implement the GIS-FA method, it is imperative to acquire geographic information. This includes but is not limited to, latitude and longitude. For the tested environments, such data can be obtained *in situ* in the experimental area or via GIS tools. For the untested environments, one can sample pixels (coordinates) of the breeding region (or the area under consideration for prediction). These pixels must be representative of the different environmental conditions found in the breeding region. We detail the sampling process adopted in this study in Section 2.3.3.

- **Step 2 – Environmental data collection**: This step requires information on the sowing and harvest times for each trial. More detailed results can be achieved by using genotype-specific harvest dates. The process of envirotyping (data collection and processing) is crucial for understanding the environmental factors that drive the GEI and shape the development of the plant (COOPER et al., 2014; XU, 2016; COSTA-NETO; CROSSA; FRITSCHE-NETO, 2021). Environmental features can be obtained in the form of *in situ* data (e.g., from sensors attached to drones or high-throughput phenotyping stations) or in raster format (e.g., historical series for a given geographic point stored on online platforms as rasters). Other methods of obtaining this data include meteorological stations, the National Centers for Environmental Information (NCEI) (NOAA, 2023), the Climate Forecast System Reanalysis (CFSR) (NCEI, 2018), the European Centre for Medium-Range Weather Forecasts (ECMWF) (ECMWF, 2023), the Global Historical Climatology Network (GHCN) (GHCND, 2023), the NASA Earth Observing System Data and Information System (EOSDIS) (EOSDIS, 2023), WorldClim (FICK; HIJMANS, 2017), Climatologies at High Resolution for the Earth's Land Surface Areas (CHELSA) (CHELSA, 2023), and the Climate Research Unit Time-Series (CRU TS). Soil data can be collected through analysis conducted in the experiment itself or obtained from databases such as SoilGrids (SOILGRIDS, 2022). We detail the collection of environmental features in both datasets analyzed in Section 2.3.3. The incorporation of environmental features in statistical-genetic models is based on Shelford's Law (SHELFORD, 1911), which states that the growth of a species is regulated by environmental factors (within a range of maximum and minimum values). The environmental features can serve as environmental markers, enabling a deeper understanding of phenotypic expression. This concept was introduced in the context of G $\times$ E analysis for plant breeding by (COSTA-NETO; CROSSA; FRITSCHE-NETO, 2021), in which more details of its theoretical application are provided in the text. In this case, there is an association between the environmental marker and the evaluated genotype. Environmental features can also be used to characterize both tested and untested environments, allowing for the determination of the similarity of the sampled points to the TPE (see Section 2.3.4 for details).

- **Step 3 – Phenotypic data analysis**: In this step, we fit FA models with different numbers of factors and choose one based on parsimony and/or explanatory ability (as detailed in Section 2.3.5.2). After choosing the model, we use the FA selection tools (STEFANOVA; BUIRCHELL, 2010; SMITH; CULLIS, 2018) to build a selection index and select the best-ranking genotypes across different environments (further details in Section 2.3.5.4).

- **Step 4 – Prediction for the untested environments**: The matrix of rotated loadings of the chosen FA model is used to train a PLS regression model with the gathered environmental features. The goal is to predict the factor loading of untested environments only by providing the model with environmental information about these locations. Once the loadings are predicted, they are used in linear combinations with the experimental genotypes' factor scores to predict the eBLUPs in untested environments. This process is thoroughly detailed in Section 2.3.6.

- **Step 5 – Map-based recommendation**: The prediction phase provides the performance of each genotype in the new locations that were sampled in the first step. To extrapolate to the whole breeding region, an interpolation process is required (detailed in Section 2.3.4). We proposed three types of thematic maps, considering interpolation:

(i) adaptation zones, which allow for the identification of adaptation areas for each genotype, i.e., areas where genotypes are expected to have better responses to the local environmental effects; (ii) pairwise comparisons, which compare the performance of two genotypes (or a genotype and a commercial check) in untested environments; and (iii) which-won-where, used to identify the most promising experimental genotypes in the breeding region. At i and ii, one can make a pre-selection of which genotypes to evaluate using the FA selection tools and perform a detailed study about these selection candidates' adaptation throughout the breeding region.

2.3.3 Environmental information

We used 32 environmental features in this study, including three geographical coordinates (altitude, latitude, and longitude), 16 related to weather conditions, and 13 soil traits (Table 1). The weather variables for each environment were obtained as daily averages for the growing season (i.e., between sowing and harvest dates) and processed using the `R` (R Core Team, 2023) [version 4.2.3] package `EnvRtype` (COSTA-NETO et al., 2021), which retrieves raw data from the NASA database (SPARKS, 2018; NASAPOWER, 2022). Most of the soil variables for each location (i.e., latitude/longitude combination) were acquired using the `geodata` package (HIJMANS et al., 2023), which downloads rasters from the SoilGrids platform (SOILGRIDS, 2022). Only the raster data for soil temperature, isothermality, temperature seasonality, and mean diurnal range were manually downloaded from the platform of Lembrechts et al. (2022). Soil rasters were downloaded for a depth interval of 5–15 cm with a resolution of 30 arcseconds. Each pixel represents an area of approximately 1 km^2 and was processed using the `raster` package (HIJMANS, 2020).

In this study, we aimed to perform spatial predictions using environmental information in a three-step procedure as follows: (i) defining the scope of the prediction area based on the political borders of the Brazilian states where trials were conducted; (ii) implementing a sampling approach to generate a cloud of geographical points (latitude/longitude) for collecting environmental data. Fifty points were randomly sampled from each municipality within states, ensuring an unbiased sampling of possible environmental conditions in the states; and (iii) using the data from ii, performed a spatial interpolation to cover the entire area of the state(s) and computed the spatial predictions. In ii, the soil-related environmental features were obtained as previously described for the tested environments. Monthly averages for the weather-related environmental features were obtained from 2000 to 2021. Further details will be provided in the following sections.

Table 1 – Summary statistics of the 32 environmental features classified into three groups: geographical, climatic, and soil-related. Climatic covariates were obtained from 2000 to 2021.

Group	Environmental information	ID	Unit	Rice data			Soybean data		
				Min	Mean	Max	Min	Mean	Max
Geographical	Altitude	alt	meters (m)	70.00	410.19	1033.00	234.00	410.35	661.00
	Latitude	lat	graus (°)	-16.85	-11.68	2.90	-23.10	-21.66	-19.38
	Longitude	lon	graus (°)	-60.75	-52.05	-42.65	-56.55	-54.65	-52.72
Climatic	All sky insolation incidents on a horizontal surface	sw	MJ/m²/day	16.43	18.90	20.83	21.38	22.65	24.69
	Clear sky insolation incident on a horizontal surface	lw	MJ/m²/day	380.72	405.20	424.89	401.17	408.72	417.74
	Total precipitation	prec	mm/day	4.58	7.23	9.77	1.87	3.65	6.32
	Relative humidity	rh	%	79.15	85.47	91.92	57.30	66.27	77.09
	Slope of saturation vapour pressure curve	spv	kPa°C	0.16	0.19	0.22	0.19	0.22	0.24
	Potential evapotranspiration	etp	mm.day	7.53	8.64	9.36	9.88	10.51	11.47
	Deficit by precipitation	pept	mm.day	-4.49	-1.40	2.21	-9.60	-6.86	-4.01
	Vapor pressure deficit	vpd	kPa	0.36	0.60	0.96	0.94	1.70	2.22
	Wind speed at 2 m above ground	ws	m/s	0.06	1.04	1.82	0.62	1.69	2.14
	Dew/Frost Point Temperature	tdew	°C	18.50	22.03	24.07	17.90	19.50	21.02
	Daily temperature range	trange	°C day	5.65	7.22	8.87	9.36	12.16	14.00
	Temperature at 2 m above ground	tmean	°C	21.99	24.84	27.40	25.06	27.48	28.96
	Maximum temperature at 2 m above ground	tmax	°C	26.69	28.69	31.71	29.94	33.91	36.08
	Minimum temperature at 2 m above ground	tmin	°C	17.82	21.47	23.58	20.53	21.75	22.80
	Growing degree days	gdd	°C d^{-1}	14.26	17.08	19.65	17.26	19.83	21.19
	Effect of temperature on radiation use efficiency	frue	-	0.65	0.78	0.89	0.78	0.89	0.94
Soil	Bulk density of the fine earth fraction	bdod	kg dm^{-3}	1.10	1.27	1.40	1.20	1.28	1.40
	Clay (<0.002 mm) in fine earth	clay	%	18.00	28.58	42.00	16.00	36.22	52.00
	Silt (0.002-0.05 mm) in fine earth	silt	%	11.00	19.88	32.00	10.00	17.76	23.00
	Sand (>0.05 mm) in fine earth	sand	%	39.00	51.65	69.00	25.00	45.94	66.00
	Volume fraction of coarse fragments (>2 mm)	cfvo	%	1.00	4.38	11.00	2.00	3.59	5.00
	Nitrogen content	nit	g kg^{-1}	0.80	1.48	2.40	1.30	1.69	2.10
	Organic carbon density	ocd	kg m^{-3}	1.90	2.46	3.20	2.00	2.45	3.00
	pH (H_2O)	phh2o	-	4.30	5.27	5.80	5.20	5.39	5.80
	Soil organic carbon in fine earth	soc	g kg^{-1}	8.80	18.63	35.40	14.10	17.85	24.20
	Soil temperature	tsoil	K	226.17	253.46	292.83	237.17	257.94	270.00
	Temperature seasonality	sts	°C	86.10	155.20	255.70	242.70	349.49	401.90
	Isothermality	iso	-	-84.60	13.67	30.70	15.30	19.74	23.30
	Mean diurnal range	mdr	-	-2.00	1.17	2.40	1.90	2.47	3.00

2.3.4 Environmental similarity and interpolation grid

The package `pdist` (WONG, 2022) was used to quantify the environmental similarity by calculating the Euclidean distances between the observed and unobserved (i.e., sampled points) environments. Let $\mathbf{W}$ be a $J \times P$ matrix of scaled values representing P environmental features in J observed environments, and let $\mathbf{\Omega}$ be a matrix containing the same information but for U unobserved environments. The environmental features were scaled to variance 1. Then, the Euclidean distance between an observed environment j and an unobserved environment u (D_{ju}) is given by the distances between the rows of $\mathbf{W}$ and $\mathbf{\Omega}$ that correspond to j and u, respectively:

$$D_{ju} = \sqrt{\sum_{p=1}^{P}(w_{jp} - \omega_{up})^2} \tag{2.1}$$

where w_{jp} and ω_{up} are entries of $\mathbf{W}$ and $\mathbf{\Omega}$ that represent the value of the p^{th} environmental feature for the j^{th} tested environment and the u^{th} untested environment, respectively.

After calculating the distances between all J and U environments, we expanded these results to include all possible environments within the delimited prediction area using the inverse distance weighting (IDW) interpolation method. The IDW was performed using the `Spatstat` package (BADDELEY; RUBAK; TURNER, 2015). Let $u^\star$ represent an untested and unsampled environment ($u^\star = 1, 2, \ldots, U^\star$, with $U^\star >> U$). The Euclidean distance between a given j and $u^\star$ is defined as:

$$D_{u^\star j} = \frac{\sum_{u=1}^{U} \frac{1}{||u^\star - x_u||^\tau} D_{uj}}{\sum_{u=1}^{U} \frac{1}{||u^\star - x_u||^\tau}} \tag{2.2}$$

where $||u^\star - x_u||$ represents the Euclidean distance between $u^\star$ and a given sampled point x_u within the observation window, and τ is a power of the multiplication determined through cross-validation (CV). Values of τ ranging from 0.1 to 5.0, with an increment of 0.1, were tested in the CV. The value that yielded the lowest mean squared error between the predicted and observed values at the sampled points was selected.

Once we have performed the interpolation and obtained the Euclidean distances between all tested and untested environments, we consider the environmental similarity between the u^{th} (or u^{*th}) untested environment and the observed environments of the TPE to be the minimum distance of u (or $u^\star$) to any j:

$$S_u = min(D_{uj}) \quad \& \quad S_u^\star = min(D_{u^\star j}) \tag{2.3}$$

2.3.5 Phenotypic analysis

The phenotypic analyses across environments for both data sets were performed using the following linear mixed model (HENDERSON, 1949; HENDERSON, 1950) in the ASReml-R package [version 4.1.2] (BUTLER, 2021). Variance components were estimated using residual maximum likelihood (PATTERSON; THOMPSON, 1971).

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_1\mathbf{s} + \mathbf{X}_2\mathbf{r} + \mathbf{Z}_1\mathbf{g} + \epsilon \tag{2.4}$$

where $\mathbf{y}$ is the vector of phenotypic records, $\mu\mathbf{1}$ is the intercept, $\mathbf{s}$ is the vector of fixed effects of environments with design matrix $\mathbf{X}_1$, $\mathbf{r}$ is the fixed vector of within-environment block effects with design matrix $\mathbf{X}_2$, $\mathbf{g}$ is the vector of random genotypic effects nested within environments with incidence matrix $\mathbf{Z}_1$, and ϵ is the residual term. The distributional assumptions for $\mathbf{g}$ and ϵ are detailed below.

Using the available information on the coordinates (row and column) of each plot in the soybean dataset, we implemented a strategy to control the spatial trends in a single step, following the approach proposed by Gogel, Smith e Cullis (2018). In summary, we conducted model testing in each environment, considering spatial analysis. These adjustments included incorporating autoregressive processes in the error term as well as linear and non-linear effects as fixed or random terms, as previously demonstrated by Gilmour, Cullis e Verbyla (1997). We identified the best-fitting model for each specific environment. Once we determined the optimal model for each environment, we incorporated the factors from these models into Equation 2.4. Each additional factor followed a block diagonal covariance structure, with non-nil effects only for environments where these factors were present in the best within-environment model. Detailed information about this procedure can be found in Supplementary Table 2. For spatially-adjusted trials, the residual effects are distributed as $\epsilon \sim MVN(\mathbf{0}, \oplus_{j=1}^{J}\sigma^2_{\epsilon_j}[\mathbf{\Gamma}_{C_j} \otimes \mathbf{\Gamma}_{R_j}])$, where $\mathbf{\Gamma}_{C_j}$ and $\mathbf{\Gamma}_{R_j}$ are autocorrelation matrices of dimensions $C_j \times C_j$ and $R_j \times R_j$, respectively. Here, C_j represents the number of columns, and R_j represents the number of rows in the j^{th} trial. These matrices have a value of 1 on the diagonal, and the off-diagonal elements represent the autocorrelation coefficients that quantify the spatial trends in the column or row directions. For environments where no spatial adjustment was necessary, $\epsilon \sim MVN(\mathbf{0}, \oplus_{j=1}^{J}\sigma^2_{\epsilon_j}\mathbf{I}_{N_j})$, where $\mathbf{I}_{N_j}$ is an identity matrix of order N_j, which corresponds to the number of phenotypic records per environment. $\oplus$ represents the direct sum, which generates a block diagonal matrix, and $\otimes$ denotes the Kronecker product. For the rice dataset, since we did not have access to spatial information, $\epsilon \sim MVN(\mathbf{0}, \oplus_{j=1}^{J}\sigma^2_{\epsilon_j}\mathbf{I}_{N_j})$.

Genotypic effects were modeled using the FA covariance structure (PIEPHO, 1997; SMITH; CULLIS; THOMPSON, 2001):

$$\mathbf{g} = (\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I}_V)\tilde{\mathbf{f}} + \tilde{\boldsymbol{\delta}} \tag{2.5}$$

where $\hat{\boldsymbol{\Lambda}}$ is the $J \times K$ matrix of K loadings for the J environments ($\hat{\boldsymbol{\Lambda}}^\star = \{\hat{\lambda}^\star_{k_j}\}$), $\tilde{\mathbf{f}}^\star$ is a vector of K scores for the V genotypes ($\tilde{\mathbf{f}}^\star = \{f^\star_{k_v}\}$), and $\tilde{\boldsymbol{\delta}}$ is the vector of the VJ lack of fit effects ($\tilde{\boldsymbol{\delta}} = \{\hat{\delta}_{v_j}\}$). $\mathbf{I}_V$ is an identity matrix of order V. $\tilde{\mathbf{f}}$ and $\tilde{\boldsymbol{\delta}}$ are independent and distributed as multivariate Gaussian with zero means and variances given by $\mathbf{D} \otimes \mathbf{I}_V$ and $\boldsymbol{\Psi} \otimes \mathbf{I}_V$, respectively. $\mathbf{D}$ is a $K \times K$ symmetric positive (semi)-definite factor score variance matrix, and $\boldsymbol{\Psi}$ is a $J \times J$ diagonal matrix of environment-wise variances that were not captured by any factor ($\hat{\boldsymbol{\Psi}} = \{\hat{\psi}_j\}$). For more information about the estimation process of $\hat{\boldsymbol{\Lambda}}$, $\tilde{\mathbf{f}}$, and $\tilde{\boldsymbol{\delta}}$, refer to Smith, Cullis e Thompson (2001), Thompson et al. (2003), and Tolhurst et al. (2022).

2.3.5.1 Rotation

We followed the rotation process recommended by Smith et al. (2021), where two constraints are imposed for the sake of interpretability: $\mathbf{D}$ is a diagonal matrix with elements arranged in decreasing order, and $\boldsymbol{\Lambda\Lambda}'$ is an identity matrix, i.e., $\boldsymbol{\Lambda}$ is composed of orthonormal columns. To address these conditions, we performed the singular value decomposition of $\hat{\boldsymbol{\Lambda}}$:

$$\hat{\boldsymbol{\Lambda}} = \mathbf{U}\mathbf{L}^{\frac{1}{2}}\mathbf{V}' \tag{2.6}$$

where $\mathbf{U}$ is an $M \times K$ orthonormal matrix whose columns are the eigenvectors of $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$, M represents the number of environments, $\mathbf{L}$ is a $K \times K$ diagonal matrix with elements given by the eigenvalues of $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$ in decreasing order, and $\mathbf{V}$ is a $K \times K$ orthonormal matrix whose columns are the eigenvectors of $\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Lambda}}$. Note that $\mathbf{U}$ meets the conditions of the second constraint, so $\hat{\boldsymbol{\Lambda}}^\star = \mathbf{U}$, in which $\hat{\boldsymbol{\Lambda}}^\star$ is the matrix of rotated loadings. By considering $\mathbf{D} = \mathbf{L}$, we fulfill the condition of the first constraint. The rotated scores were obtained as $\tilde{\mathbf{f}}^\star = (\mathbf{D}\mathbf{V}' \otimes \mathbf{I}_V)\mathbf{f}$, where $\tilde{\mathbf{f}}^\star$ is the vector of rotated scores. After rotation, the conditional distribution of the genotypic effects is $\mathbf{g} \sim MVN[\mathbf{0}, (\hat{\boldsymbol{\Lambda}}^\star\mathbf{D}\hat{\boldsymbol{\Lambda}}^{\star\prime} + \hat{\boldsymbol{\Psi}}) \otimes \mathbf{I}_V]$.

2.3.5.2 FA model selection

FA models with different numbers of factors were fitted and compared in terms of their explanatory ability. We used the average semivariance ratio (ASR) (PIEPHO, 2019; CHAVES et al., 2023a) as a selection criterion. By calculating the ratio between the average semivariance of $\hat{\boldsymbol{\Lambda}}^\star\mathbf{D}\hat{\boldsymbol{\Lambda}}^{\star\prime}$ and the average semivariance of $\hat{\boldsymbol{\Lambda}}^\star\mathbf{D}\hat{\boldsymbol{\Lambda}}^{\star\prime} + \boldsymbol{\Psi}$, it is possible to investigate the amount of total covariance that is being captured by the factors of the FA model. The ASR is given as follows:

$$ASR = \frac{\frac{2}{J(J-1)}\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\frac{1}{2}(\sum_{k=1}^{K}\hat{\lambda}_{k_j}^{\star 2}d_k + \sum_{k=1}^{K}\hat{\lambda}_{k_{j'}}^{\star 2}d_k) - \sum_{k=1}^{K}\hat{\lambda}_{k_j}^{\star}\hat{\lambda}_{k_{j'}}^{\star}d_k}{\frac{2}{J(J-1)}\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\frac{1}{2}[(\sum_{k=1}^{K}\lambda_{k_j}^{\star 2}d_k + \hat{\psi}_j) + (\sum_{k=1}^{K}\hat{\lambda}_{k_{j'}}^{\star 2}d_k + \hat{\psi}_{j'})] - \sum_{k=1}^{K}\hat{\lambda}_{k_j}^{\star}\hat{\lambda}_{k_{j'}}^{\star}d_k} \times 100 \tag{2.7}$$

where d_k is the k^{th} element of the diagonal of $\mathbf{D}$.

We defined an ad hoc threshold of 75% for explanatory ability. As complementary information, we also estimated the proportion of genetic variance explained by the k^{th} factor in the j^{th} environment [v_{k_j},] (SMITH et al., 2015):

$$v_{k_j} = \frac{\sum_{k=1}^{K}\hat{\lambda}_{k_j}^{\star 2}d_k}{\sum_{k=1}^{K}\hat{\lambda}_{k_j}^{\star 2}d_k + \hat{\psi}_j} \times 100 \tag{2.8}$$

From the best-fit model, we estimated some useful parameters to investigate the experimental precision, such as the environment-wise generalized heritabilities (CULLIS; SMITH; COOMBES, 2006) and coefficients of experimental variation (CV), which are given by the following equations, respectively:

$$H_j^2 = 1 - (\frac{\overline{v}_{\Delta}^{BLUP}}{2\sigma_{g_j}^2}) \tag{2.9}$$

$$CV_j = \frac{\sigma_{e_j}}{\mu_j} \tag{2.10}$$

where $\overline{v}_{\Delta}^{BLUP}$ is the average pairwise prediction error variance, $\sigma_{g_j}^2$ is the genotypic variance for the j^{th} environment, taken from the diagonal elements of $\hat{\mathbf{\Lambda}}^{\star}\mathbf{D}\hat{\mathbf{\Lambda}}^{\star'} + \hat{\mathbf{\Psi}}$; σ_{e_j} is the estimated residual standard deviation for the j^{th} environment, and μ_j is the mean of the trait for the j^{th} environment.

2.3.5.3 Genotype-by-environment interaction investigation tools

We investigated the GEI dynamics in the datasets by examining the pairwise genetic correlations between environments and the partitioning of GEI variance into crossover and non-crossover patterns. The pairwise genetic correlation between environments ($\rho_{jj'}$) is given as follows Cullis, Beeck e Cowling (2010):

$$\mathbf{\Upsilon} = \mathbf{\Delta}(\hat{\mathbf{\Lambda}}^{\star}\mathbf{D}\hat{\mathbf{\Lambda}}^{\star'} + \hat{\mathbf{\Psi}})\mathbf{\Delta} \tag{2.11}$$

where $\mathbf{\Upsilon}$ is a $J \times J$ matrix of genetic correlations, and $\mathbf{\Delta}$ is a diagonal matrix whose elements are the inverse of the square roots of the diagonal values of $\hat{\mathbf{\Lambda}}^{\star}\mathbf{D}\hat{\mathbf{\Lambda}}^{\star'} + \hat{\mathbf{\Psi}}$.

The decomposition of the GEI variance was performed using the following equation, adapted from Cooper e Delacy (1994):

$$\sigma_{ge_{rank}}^2 = 1 - \frac{Var(\sqrt{\sigma_{g_j}^2})}{\sigma_{ge}^2} \tag{2.12}$$

where σ^2_{ge} is the variance attributed to the GEI, which is determined by fitting a compound symmetry model. This model has the same structure as Eq. 2.4, but the variance-covariance matrix of genetic effects has the form $\sigma^2_g\mathbf{J} + \sigma^2_{ge}\mathbf{I}_J$, where $\mathbf{J}$ is a $J \times J$ matrix of ones.

2.3.5.4 Selection tools for overall performance and stability

The target features of most breeding programs are to achieve high performance and stability across the TPE. Using the best-fit FA model, we estimated metrics to assess the performance and stability of genotypes regarding seed yield. The performance was measured using the OP metric (OP_v), which was obtained as follows (STEFANOVA; BUIRCHELL, 2010; SMITH; CULLIS, 2018):

$$OP_v = \frac{1}{J}\sum_{j=1}^{J}\hat{\lambda}^{\star}_{1_j} f^{\star}_{1_v} \tag{2.13}$$

Note that only the first factor is used to compute the OP_v. This factor captures the largest portion of the total variance. Thus, it provides a generalized measure of the genetic main effects (Figure S12) (STEFANOVA; BUIRCHELL, 2010). According to empirical observations by Smith e Cullis (2018), this is valid when the majority of loadings in the first factor are positive, indicating the absence (or insignificance) of crossover GEI in the first factor. Using this principle, the other factors are used to represent stability. Considering that the genetic effect of a given genotype v at the j^{th} environment, disregarding the lack of fit effect, is $g_{vj} = \hat{\lambda}^{\star}_{1_j} f^{\star}_{1_v} + \hat{\lambda}^{\star}_{2_j} f^{\star}_{2_v} + \cdots + \hat{\lambda}^{\star}_{K_j} f^{\star}_{K_v}$, which is equivalent to $g_{vj} = \hat{\lambda}^{\star}_{1_j} f^{\star}_{1_v} + \epsilon_{vj}$, the stability of v is given by:

$$RMSD_v = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\epsilon^2_{vj}} \tag{2.14}$$

in which $RMSD_v$ is the root-mean-square deviation of v, representing the distance between the point and the slope in a latent regression given by $g_{vj} = \hat{\lambda}^{\star}_{1_j} f^{\star}_{1_v} + \epsilon_{vj}$ (SMITH; CULLIS, 2018).

A desirable genotype i has a high OP_i and a low $RMSD_i$. Following these principles, we applied a selection index (SI_v) with these metrics (CHAVES et al., 2023b; COWLING et al., 2023), given as follows:

$$SI_v = 2 \times \frac{OP_v - \overline{OP}}{\sqrt{V(OP)}} - \frac{RMSD_v - \overline{RMSD}}{\sqrt{V(RMSD)}} \tag{2.15}$$

In addition to the selection index, the reliability of the v^{th} genotype (MRODE, 2014) was calculated as follows:

$$r_v = 1 - \frac{PEV_v}{\overline{\sigma_g^2}} \tag{2.16}$$

where PEV_v represents the prediction error variance of the v^{th} genotype, and $\overline{\sigma_g^2}$ is the average genotypic variance across environments. The reliability metric associated with the selection index is useful for improving the accuracy of selection, especially when dealing with unbalanced data sets. We adopted a selection intensity of 15% for both datasets.

2.3.6 Spatial predictions in the breeding zone

In this study, GIS tools were used to: (1) collect georeferenced data from the evaluated trials, (2) build environmental markers, and (3) perform spatial predictions for a larger area. Here, we used PLS regression (WOLD, 1966; AASTVEIT; MARTENS, 1986) to make the predictions. This method is useful when the number of predictors is much larger than the number of observations and when these predictors are correlated. When PLS is used to predict genotypic performances in untested environments, the response variable is the genotypic performance in the testing set. In this situation, the response variable is a $J \times 1$ vector ($\mathbf{y}$) of phenotypic records if a genotype-wise PLS model is fitted or a $J \times V$ matrix ($\mathbf{Y}$) when a multivariate PLS model is fitted considering all genotypes at once (MONTEVERDE et al., 2019; COSTA-NETO et al., 2022). We refer to the multivariate model as GIS-GGE.

We modified GIS-GGE by using the rotated loadings of the tested environments ($\hat{\lambda}^{\star}_{k_j}$) as response variables instead of the within-environment phenotypic records of the genotypes. This procedure, we called GIS-FA. We obtained these loadings from the previously chosen FA model (Section 2.3.5.2). With the predicted loadings and the previously estimated scores for each genotype from the FA model, we can predict the empirical BLUPs of the genotypes in untested environments. The PLS regression model was trained using the rotated loadings and environmental features of the tested environments:

$$\hat{\mathbf{\Lambda}}^{\star} = \mathbf{W}\mathbf{B}^{\star} + \mathbf{E} \tag{2.17}$$

where $\mathbf{B}^{\star}$ is a $P \times K$ vector of coefficients, $\mathbf{E}$ is a $J \times K$ matrix of lack-of-fit effects, and $\hat{\mathbf{\Lambda}}^{\star}$ and $\mathbf{W}$ were previously described in Sections 2.3.5 and 2.3.4, respectively. We obtained $\mathbf{B}^{\star}$ using a kernel PLS algorithm (LINDGREN; GELADI; WOLD, 1993; DAYAL; MACGREGOR, 1997) implemented in the `pls` package (LILAND; MEVIK; WEHRENS, 2022). This algorithm is detailed in Appendix 2.7.

After training the model, we substituted $\mathbf{W}$ with $\mathbf{\Omega}$ to predict the K loadings of the U untested environments:

$$\hat{\mathbf{\Lambda}}^{\star}_U = \mathbf{\Omega}\mathbf{B}^{\star} + \mathbf{E} \tag{2.18}$$

recall from Section 2.3.3 that $\mathbf{\Omega}$ was built using historical weather data from 2000 to 2021, as well as soil environmental features. Once we predicted the loadings of untested environments, we used them in linear combinations with the previously predicted scores of each genotype (see Section 2.3.5) to estimate their eBLUPs within untested environments:

$$\mathbf{g}_U = (\hat{\mathbf{\Lambda}}_U^{\star} \otimes \mathbf{I}_V)\tilde{\mathbf{f}}^{\star} \tag{2.19}$$

Note that we use the same scores to predict the eBLUPs of both tested and untested environments. Nevertheless, the scores are predicted based solely on the data collected from the tested environments. In other words, the environments in the data set must accurately reflect the TPE so that the loadings of the untested environments closely match the loadings of the tested ones.

A CV process is required to obtain $\mathbf{B}^{\star}$. We employed a leave-one-out scheme, where data from a single environment was removed (the testing set), and predictions were made using the information provided by the remaining environments (the training set). The predicted eBLUPs were then correlated with the actual eBLUPs and eBLUEs of each environment to determine the predictive ability of the PLS regression model. The model with the highest number of components demonstrating predictive ability was chosen. We leveraged the same CV scheme to compare the predictive ability of GIS-FA and GIS-GGE. In this study, the PLS regression of GIS-GGE was trained with the within-environment empirical eBLUPs of each genotype as response variables.

2.3.6.1 Thematic maps

Thematic maps combine cartographic principles and GIS tools to represent and analyze spatial and geographic phenomena. The incorporation of spatial interpolation methods enables the estimation of values in untested locations, resulting in a seamless representation of the phenomenon (DABROWSKI et al., 2021). This facilitates the identification of patterns and trends, aiding decision-making across various fields of study (COSTA-NETO et al., 2020).

Recall that $\mathbf{\Omega}$ has U rows, and the predictions must be extrapolated to all $U^{\star}$ untested environments within the targeted area. For this purpose, we used an interpolation process similar to the one described in Section 2.3.4. The difference is that for the environmental similarity maps, we interpolated Euclidean distances, while for the thematic maps described in this section, we interpolated eBLUPs. Once the spatial prediction was interpolated across the whole breeding region, we built thematic maps to aid in the visualization and interpretation of the results. We created maps with three themes:

- Adaptation zones: These maps depict the expected spatial prediction of each selection candidate across the breeding zone. The adaptation of a genotype to an environment is assessed by the expected response of that genotype when it is planted in that environment. Thus, in this context, "adaptation" is used as a synonym for specific performance. For improved visualization, we divided the predicted eBLUPs into eight categories (from expected yield lower than 2500 kg ha^{-1} to expected yield higher than 4000 kg ha^{-1}), and each category was then assigned a specific color.

- Pairwise comparisons: These maps allow for a direct comparison of the expected responses of different genotypes in specific environments. Two distinct colors, one for each candidate, were used to indicate that the superior selection candidate was superior in each location on the map. This visual representation helps to quickly identify which selection candidate outperforms the other in each pixel, facilitating the interpretation of competitive advantages among genotypes in specific environments.
- Which-won-where: The genotype that achieved the best performance in each location on the map is highlighted. This map provides a clear depiction of the winning genotype for each specific location, enabling a comprehensive understanding of the distribution of high-performing genotypes across the breeding zone.

These maps, like all the other plots, were built using the `ggplot2` package (WICKHAM, 2016), with the addition of the `ggspatial` (DUNNINGTON, 2023) and `sf` (PEBESMA; BIVAND, 2023) packages. The shapefiles we used are freely available at the Brazilian Institute of Geography and Statistics (IBGE in the Portuguese acronym) website (https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html), or they can be downloaded using the `geodata` package.

2.4 Results

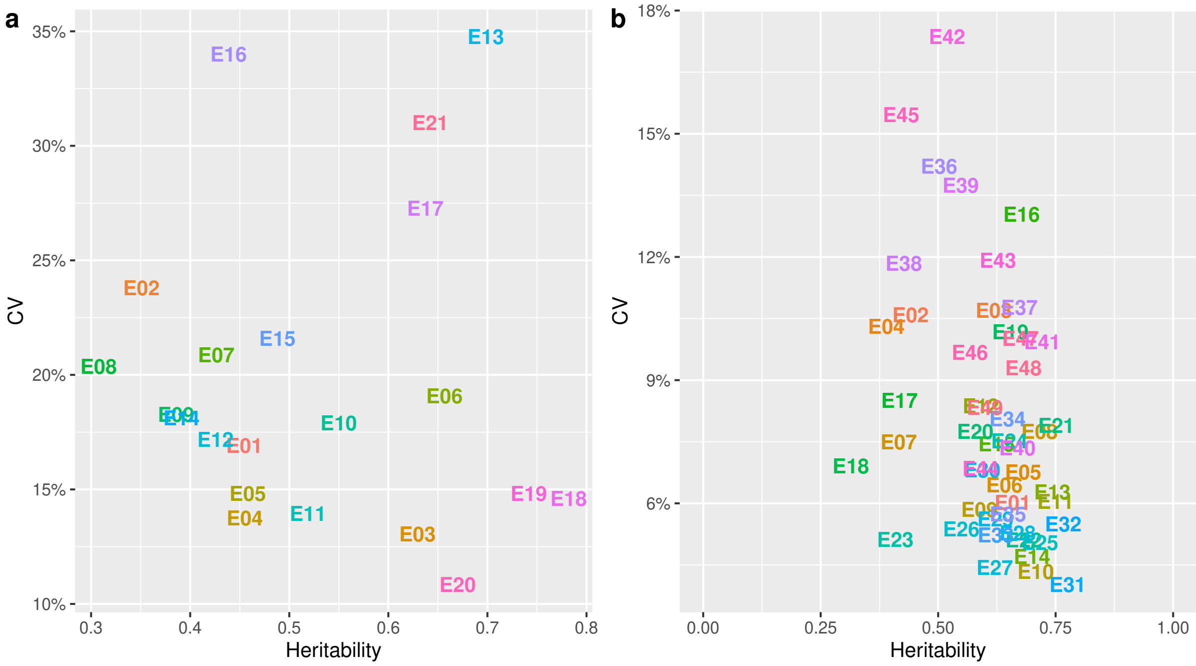
2.4.1 Experimental accuracy

In the rice dataset, CV_j ranged from 0.11 (E20) to 0.34 (E13), and H_j^2 ranged from 0.31 (E08) to 0.78 (E18) (Figure 2a). In the soybean dataset, CV_j ranged from 0.04 (E31) to 0.17 (E42), and H_j^2 ranged from 0.31 (E18) to 0.77 (E31)(Figure 2b). Spatial trends were modeled in 37 out of 49 soybean trials (Table S2).

2.4.2 Genotype recommendations for tested environments

The FA model with four factors (FA4) met our criteria for both datasets. It explained more than 75% of the variance (Table 2). This model captured most of the within-environment variance in both datasets (Figure S13).

Figure 2 – Scatter plot representing the experimental coefficient of variation (CV, on a decimal scale) in the y-axis and the generalized heritability in the x-axis for grain yield (kg ha^{-1}) of rice (a) and seed yield (kg ha^{-1}) of soybean (b) trials.

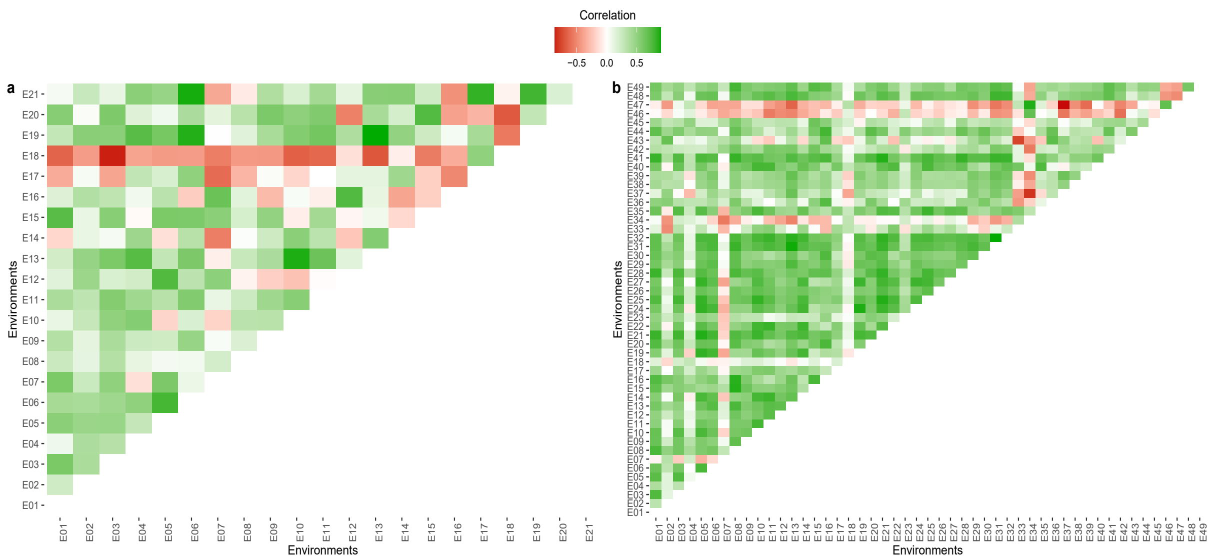


The genotypic correlations ranged from -0.0031 (E07 *vs.* E19) to 0.8936 (E13 *vs.* E19) for the rice dataset (Figure 3a) and from -0.0010 (E07 *vs.* E41) to 0.9753 (E031 *vs.* E32) for the soybean dataset (Figure 3b). In the rice dataset, environments E17 and E18 exhibited the most contrasting patterns compared to the other environments. Their correlations with the remaining environments were predominantly negative or close to zero. Similarly, in the soybean dataset, negative or negligible correlations were observed for contrasts involving environments E18, E33, E34, E43, E46, and E47. These findings indicate substantial differences between these specific environments and the rest of the dataset. The wide range of correlation magnitudes is reflected in the percentage of crossover GEI in the datasets: 76% and 81% of the total GEI were due to crossover interactions in the rice and soybean datasets, respectively.

Table 2 – Fitted factor-analytic mixed models for each dataset (rice and soybean) and their respective logarithm of the likelihood function (LogL), number of parameters (no. par.), and average semivariance ratio (ASR). In the rice dataset, models with five factors onward had singularity issues. The selected models are in bold.

Model	LogL	no. par.	ASR
Rice data			
FA1	-10482.19	61	21.38
FA2	-10470.01	76	51.60
FA3	-10456.57	92	66.85
FA4	**-10444.36**	**108**	**78.70**
Soybean data			
FA1	-37722.05	227	22.58
FA2	-37627.68	276	54.08
FA3	-37578.47	332	69.08
FA4	**-37528.83**	**336**	**76.63**
FA5	-37462.72	400	83.52
FA6	-37418.20	433	91.40
FA7	-37374.42	468	93.92

Figure 3 – Heatmaps representing the genetic correlation between pairs of environments in the rice (**a**) and soybean (**b**) datasets. The color gradient depicts the direction of the correlation: red designates a negative correlation, whereas green represents a positive correlation.



The selected candidates based on the selection index are highlighted in Figure 4. Despite the low reliability of the rice dataset, genotypes G23, G18, G29, G31, and G26 stand out for their high stability. Genotypes G10, G09, G03, and G01 presented high OP and reliability. The check treatment (C83) had the highest OP, but it exhibited low stability and reliability compared to the other selected genotypes (Figure 4a). Among the soybean genotypes, G178, G031, G101, G052, and G035 exhibited the highest stability. On the other hand, G177, G100, G144, G088, and G016 were notable for their high OP. Genotype G16 showed high OP, stability, and reliability (Figure 4b). The reliability of the selected candidates was higher in the soybean dataset, likely due to the stronger correlation between environments in the soybean dataset compared to the rice dataset.

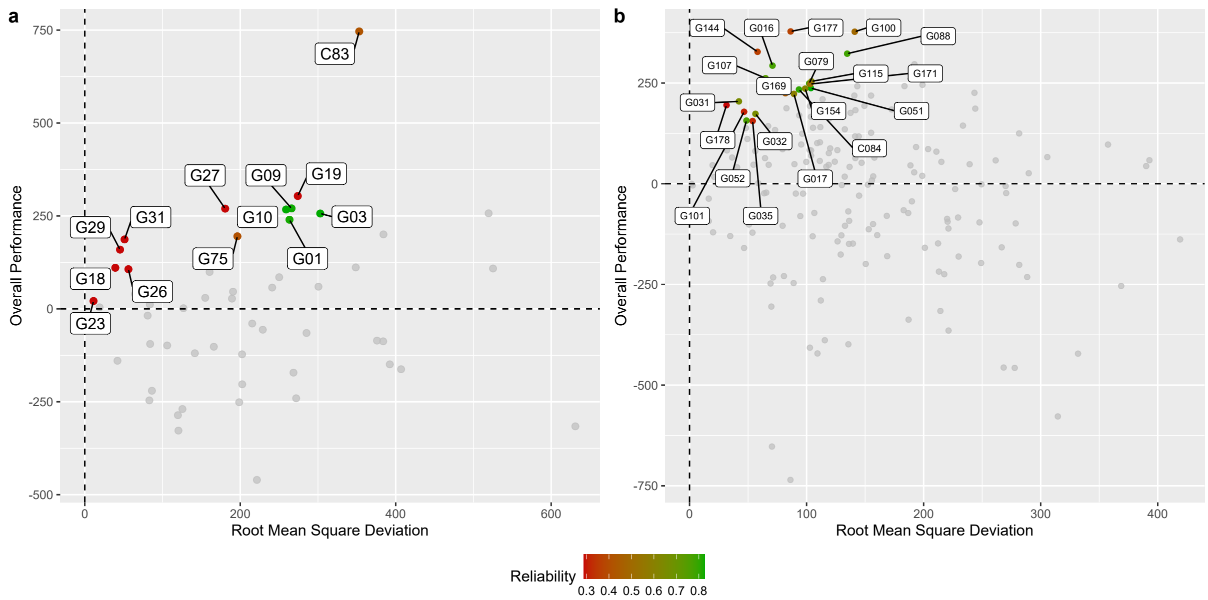


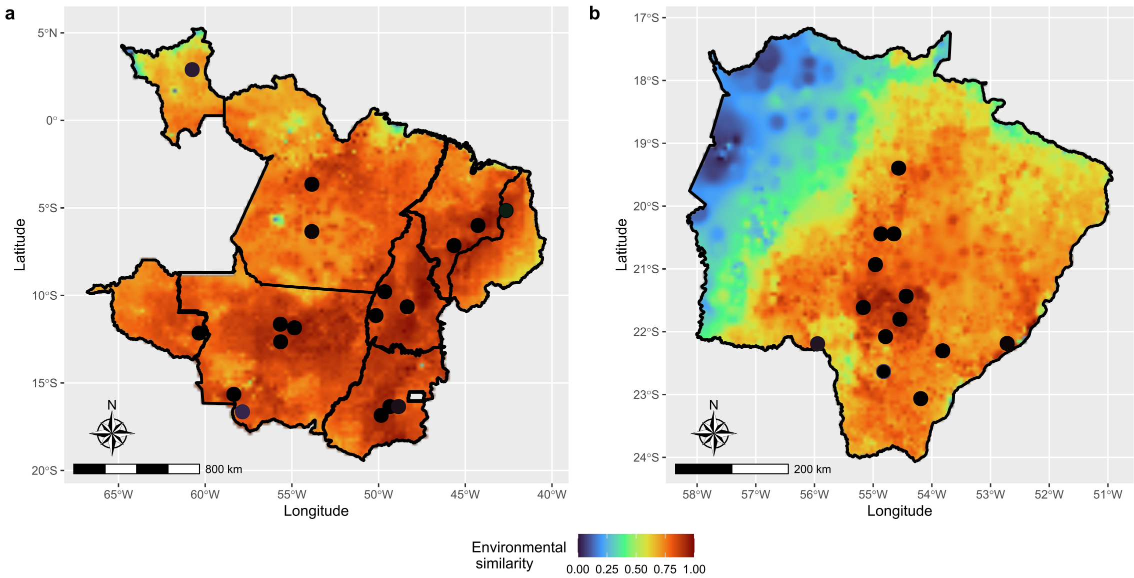
Figure 4 – Overall performance (y-axis) and root-mean-square deviation (x-axis) of the experimental genotypes in the rice (**a**) and soybean (**b**) datasets. The most productive genotypes are oriented towards the upper part on the y-axis, and the most stable ones are toward the left in the x-axis.

2.4.3 Predictions using environmental markers in untested environments

2.4.3.1 Environmental similarity

The rice trials are spread throughout the breeding region and effectively capture the environmental conditions of the areabeing studied (Figure 5a). On the other hand, the trials in the soybean dataset are concentrated in the central part of the state, while there is a region to the west that exhibits low similarity. This area corresponds exactly to the Pantanal biome, which is a protected area with legal restrictions on soybean planting (Figure 5b). This is probably the reason why there is no trial in this region.

Figure 5 – Environmental similarity between tested and untested environments in the target population of environments in the rice (**a**) dataset and in the soybean (**b**) dataset. The warmer the color, the higher the similarity, and consequently, the higher the prediction reliability. Colored circles represent the trials' locations.



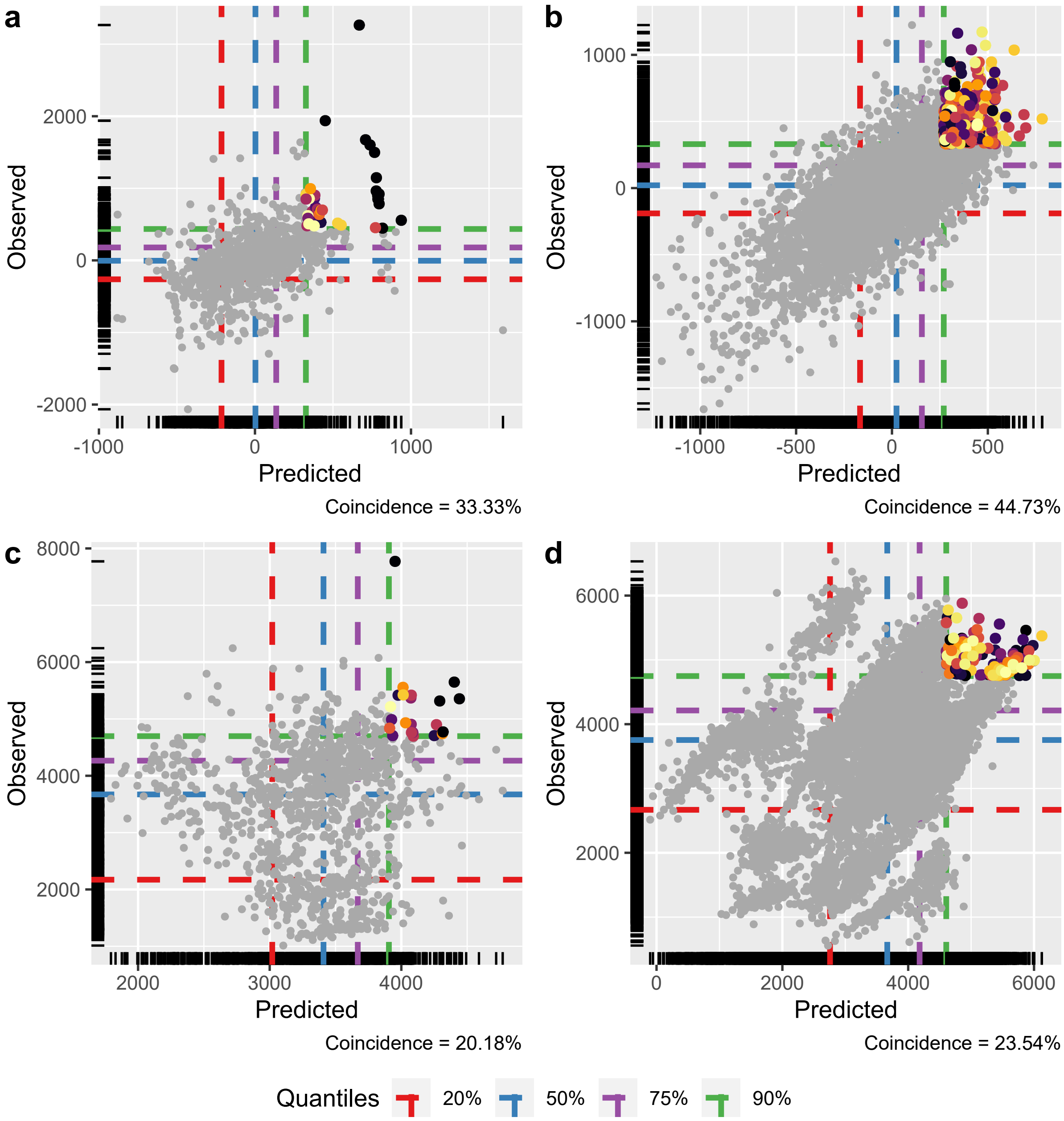
2.4.3.2 GIS-FA validation

In comparison to GIS-GGE, our proposal yields a higher prediction accuracy (as measured by the simple correlation between predicted and observed values) for both datasets. For predicting eBLUEs, GIS-FA is 10% and 1% better than GIS-GGE in the rice and soybean datasets, respectively (Table 3). For predicting eBLUPs, GIS-FA is 9% and 5% more effective than GIS-GGE. A second way to assess the predictive ability of the methods is to check the coincidence between the top 10% of observed and predicted values (Figure 6). GIS-FA provides more assertive results (Figures 6a and 6b) than GIS-GGE (Figures 6c and 6d). In other words, when recommending elite candidates based on predicted values, it is more probable that the true top performers will be recommended using GIS-FA than using GIS-GGE. In the rice dataset (Figures 6a and 6c), GIS-FA has an accuracy that is 13.15 percentage points higher than GIS-GGE. In the soybean dataset (Figures 6b and 6d), GIS-FA is 21.19 percentage points more advantageous than GIS-GGE.

Table 3 – Prediction accuracy of eBLUEs and eBLUPs using the proposed method GIS-FA and the conventional method GIS-GGE. For more information about these methods, see the Material and Methods section.

Model	Prediction	Prediction accuracy	
		Rice data	Soybean data
GIS-GGE	BLUE	0.40	0.53
GIS-GGE	BLUP	0.55	0.71
GIS-FA	BLUE	0.44	0.55
GIS-FA	BLUP	0.60	0.74

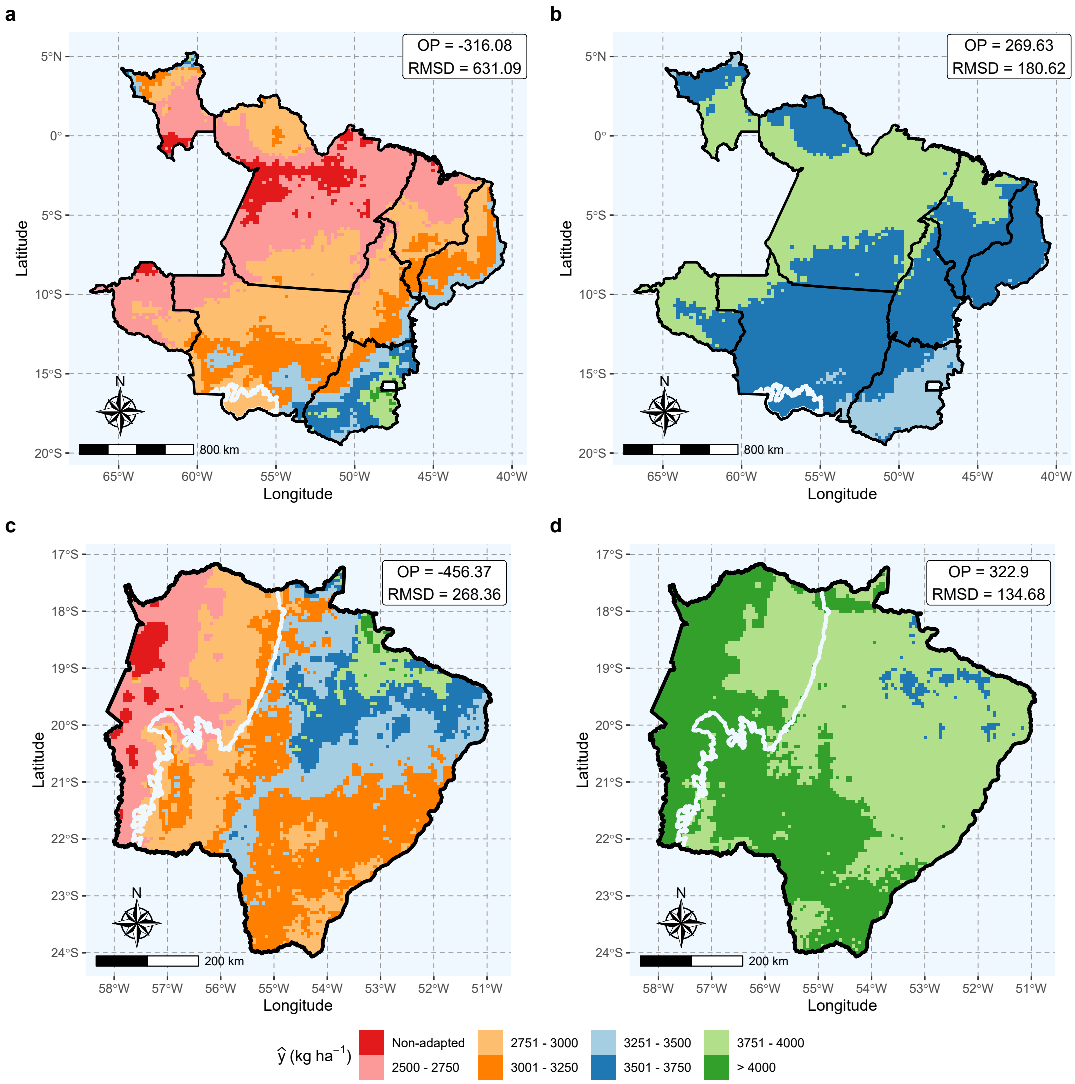
Figure 6 – Scatter plot of all predicted values (x-axis) in the leave-one-out cross-validation scheme against observed values (y-axis). The dashed lines represent the empirical percentiles (20, 50, 75, and 90%) associated with the trait value. The colored dots represent the coincident selection candidates when selecting the top 10% performers using observed and predicted values. Each color represents a different genotype. "Coincidence" in the lower left corner of each plot depicts the accuracy of selecting the top 10% using the predicted values. Figures **a** and **b** illustrate the results for the GIS-FA method in the rice and soybean datasets, respectively. Figures **c** and **d** represent the results for the GIS-GGE method in the rice and soybean datasets, respectively.



2.4.3.3 Thematic maps of adaptation zones

The spatial prediction done by GIS-FA was useful in assessing the expected performance of the experimental genotypes in untested environments. This helps to define adaptation zones for each genotype, which are the theme of the maps in Figure 7. For example, G16 of the rice dataset, shown in Figure 7a, seems to be well adapted only in a small portion of Goiás State (green region), and it responds poorly to the environmental effects of other locations within the breeding region. Conversely, G27 of the rice dataset, shown in Figure 7b, exhibits a broader spectrum in terms of adaptation in the breeding region. The same interpretation applies to the genotypes in the soybean dataset. G064 (Figure 7c) is an unstable candidate, with a very restricted area where it is better adapted (in the northern part of the breeding region). On the other hand, G088 (Figure 7d) is a stable genotype, meaning it possesses alleles that respond favorably to the environmental effects of different locations across the state. In each map, we provide the OP and $RMSD$ of the corresponding genotype. We have deliberately chosen two promising candidates (Rice's G27 and soybean's G088, which are among those selected in Figure 4), as well as two low-yielding genotypes (Rice's G16 and soybean's G064), to be included in Figure 7. Nevertheless, we recommend using OP and $RMSD$ as criteria to choose the genotype for which an adaptation map should be created.

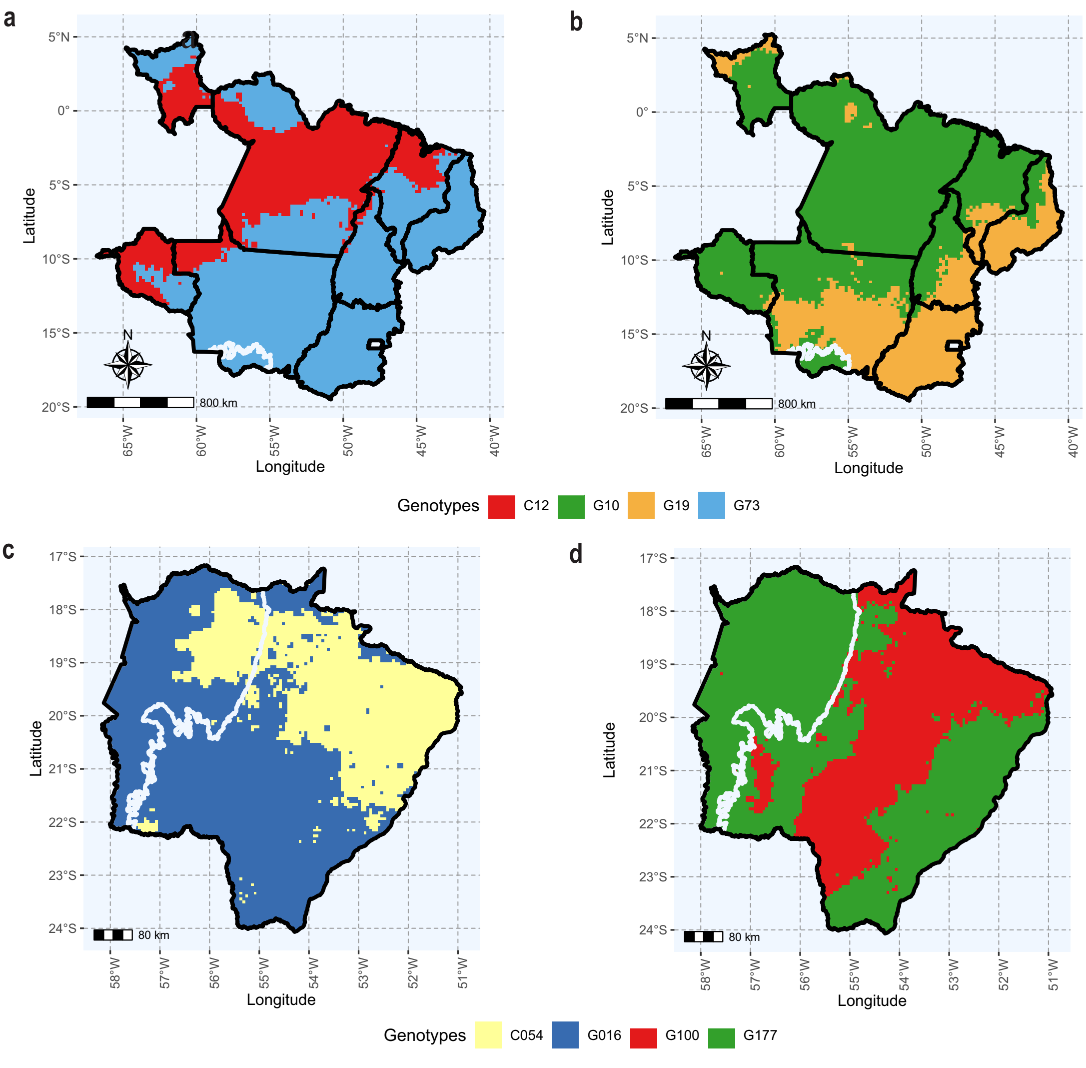
Figure 7 – Genotype-wise adaptation map showing the adaptation zones of the genotypes G16 (rice dataset, **a**), G27 (rice dataset, **b**), G064 (soybean dataset, **c**), and G088 (soybean dataset, **d**). The color scale represents the expected yield classes, from non-adapted (intense red) to more than 4000 $kg\ ha^{-1}$ (intense green). The white contour delimits the Pantanal biome. On the upper right of each map, we provide the overall performance (OP) and root-mean-square deviation (RMSD) of each genotype.



2.4.3.4 Thematic maps of pairwise comparison

To support the decision-making process, we developed a second thematic map: the pairwise comparison maps (Figure 8), which facilitate the comparison of two candidates. Take, for example, G10 and G19 in Figure 4a and G100 and G177 in Figure 4b. These candidates have somewhat similar performances, according to their OP and $RMSD$. However, they are clearly adapted to different zones within the breeding region. G10 shows better responses at lower latitudes, while G19 is more suitable for higher latitudes (Figure 8b). G100 is better adapted to the central region of the soybean's breeding region, and G177 is more compatible with the environmental conditions at the breeding region's horizontal extremes (Figure 8d).

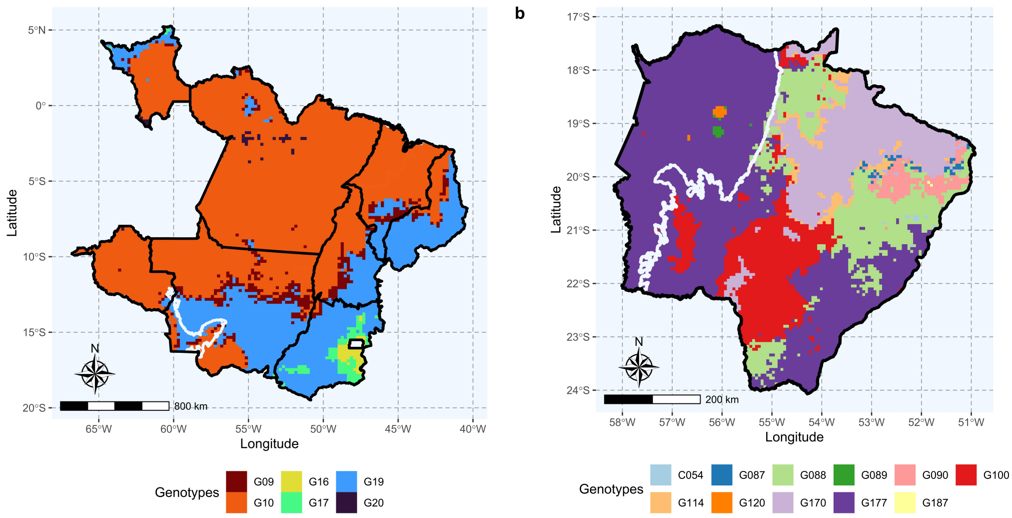
Figure 8 – Pairwise comparison map showing the regions within the rice (**a** and **b**) and soybean (**c** and **d**) target populations of environments where a selection candidate outperforms a given peer. The colors across the map represent the winning genotype. **a** and **c** are examples of pairwise comparisons between an experimental genotype and a commercial check, while **b** and **d** contrast the performance of two promising experimental genotypes along the breeding region. The white contour delimits the Pantanal biome.



2.4.3.5 Thematic maps of which-won-where

The which-won-where map (Figure 9) shows the experimental genotype that is most suitable for a specific environment within the breeding zone. In the rice dataset (Figure 9a), G10 emerges as the most promising experimental genotype in almost all environments in the central and northern portions of the breeding zone, while G19 prevails in the southern and eastern regions. G09, G16, G17, and G20 are the most suitable for specific environments. The breeding region of the soybean dataset is more diverse, with G177, G100, G170, and G088 being the most important experimental genotypes, as they have emerged as the winners in the widest area. The other selection candidates, including a cultivar check (C054), are the top performers in only a few restricted environments (Figure 9b).

Figure 9 – Which-won-where map depicting the most promising genotype at each location across the target population of environments of the rice dataset (**a**) and the soybean dataset (**b**) Each color represents the experimental genotype that wins in a specific environment within the breeding region. The white contour delimits the Pantanal biome.



2.5 Discussion

The GIS-FA method represents the integration of modern statistical genetics with GIS principles. We showed how GIS-FA can aid plant breeders in making decisions by considering the observed performance in tested environments and spatial predictions in untested environments. For observed environments, GIS-FA leverages the resources of FA models to provide useful inferences about the dynamics of the GEI and to select candidates with high performance and stability using customized selection tools (STEFANOVA; BUIRCHELL, 2010; SMITH; CULLIS, 2018). In untested environments, GIS-FA allows for a data-driven recommendation of cultivars based on spatial predictions derived from soil characteristics, climatic conditions, and empirical data parameters (i.e., factor loadings for genotypes). The GIS-FA method allows for data-driven decision-making with the aid of graphical tools such as thematic maps. These maps include: *i*) adaptation zone maps, which depict the expected spatial prediction of each genotype within the entire breeding zone; *ii*) pairwise comparison maps, which facilitate the comparison of performance between two selection candidates (or a candidate and a commercial check); and *iii*) which-won-where maps, which show the most promising experimental genotype (the winner) in each location within the breeding zone.

2.5.1 Genotype-by-environment interaction and selection in tested environments

Increasing crop yield and adapting to different growing conditions are important goals in plant breeding. These traits are the outcomes of a plethora of small quantitative trait loci (QTLs) effects that are highly influenced by the environment (LYNCH; WALSH, 1998; CROSSA, 2012). In terms of cultivar recommendation in the TPE, the most concerning source of the GEI is the lack of genotypic correlation between environments (COOPER; DELACY, 1994), as observed in both data sets (Figure 3). As a consequence, it is unlikely that the same set of experimental genotypes will exhibit similar performance across uncorrelated environments. In this case, if a global (i.e., across environments) recommendation is needed, metrics such as the selection index, which combines performance and stability, might be employed. The weight of each metric in the selection index is determined by the breeder (CHAVES et al., 2023b). Here, we prioritized performance over stability.

In the GIS-FA method, we leverage the resources of FA mixed models (PIEPHO, 1997; SMITH; CULLIS; THOMPSON, 2001) that explore the complexity of the GEI while handling highly unbalanced data sets. Furthermore, FA models allow for a parsimonious estimation of environment-wise genotypic variances and pairwise covariances. These covariances can be used to investigate the dynamics of GEI, as fully described in this study. The efficiency of the GIS-FA method depends on the choice of the number of factor loadings in the FA model, i.e., a poor choice will provide erroneous results. In GIS-FA, it is important to note that the factor loadings of observed environments are used as the training set. This allows for the prediction of the loadings of untested environments in the testing set. Thus, when selecting the best-fit FA model, selection criteria such as the AIC and ASR should always be considered. Naturally, using more factors will provide greater explanatory ability. Nevertheless, it will hinder parsimony and computational efficiency, especially in large data sets.

Assuming that the observed environments accurately represent the expected environmental conditions throughout the breeding zone, the most promising genotypes in the tested environments are probably the best ones in the untested environments. Thereby, the idea is to prioritize selected experimental genotypes when drawing the thematic maps "genotype-wise adaptation" and "pairwise comparisons."

2.5.2 Spatial interpolations in untested environments

Like molecular markers, environmental feature similarity can be used for both inference and prediction purposes. Inference models aim to determine the effect of each environmental feature on phenotypic expression and the GEI, which is analogous to QTL mapping models (DENIS, 1988; EEUWIJK; ELGERSMA, 1993; CROSSA et al., 1999; COSTA-NETO et al., 2021; HEINEMANN et al., 2022). In this work, we focused on environmental-wide predictions, regardless of the particular effect of each EF on phenotypic expression and GEI. As polygenic models are used to perform whole-genome regressions (MEUWISSEN; HAYES; GODDARD, 2001), we assumed that the core of ecophysiological effects captured by the environmental feature could be sufficient to generate genotype-wise predictions across the spatial grid. The benefits of incorporating environmental features into predictive breeding are advantageous in most cases, whether integrated with genomic information or not (de los Campos et al., 2020; BUNTARAN; FORKMAN; PIEPHO, 2021; JARQUIN et al., 2021; COSTA-NETO et al., 2022). However, recent work from Crossa et al. (2023) demonstrated that the inclusion of environmental covariates could either increase or decrease prediction accuracy, depending on the specific case. Techniques such as feature selection (CROSSA et al., 2023) and exhaustive search (LI et al., 2018) can be considered when selecting environmental features.

2.5.2.1 Environmental similarity

Environmental similarity maps revealed a need to perform an adequate sampling of the different environmental types within a given target breeding region (Figure 5). This entails including samples from various climatic conditions and soil traits that may be encountered in future predictive environments. Essentially, these maps illustrate a metric of reliability for spatial predictions by benchmarking the similarity between observed and unobserved environments. They demonstrate the environmental similarity between tested and untested environments. In other words, the more similar an untested environment is to a tested environment, the higher the chances of making an assertive prediction. The results depicted in the maps of Figure 5 can be attributed to the geographical distribution of trials in relation to the Brazilian biomes [refer to Figure 1 of Chaves et al. (2023b) for a map with the Brazilian biomes]. The soybean breeding region comprises two biomes, namely the Pantanal (wet lowlands) and the Cerrado (highland savanna conditions). All trials were conducted in the Cerrado, which explains the lack of similarity between the TPE and the environments in the Pantanal biome. Consequently, the prediction for this particular region is likely to be compromised. The rice breeding region also includes two biomes: Amazonia (a wet tropical rainforest) and Cerrado. Unlike the soybean dataset, there are representative trials from both biomes, providing comprehensive coverage of the relevant environmental conditions.

2.5.2.2 Predicting using partial least squares regression

The association between PLS regression, GEI, and environmental features was introduced by Aastveit e Martens (1986) for inference purposes. Their aim was to address challenges related to the curse of dimensionality and multicollinearity in explaining the dynamics of GEI using two datasets. Their model was later expanded to include information on molecular markers to investigate QTL-by-environment interactions (CROSSA et al., 1999; VARGAS et al., 2006). Nevertheless, employing environmental features in statistical models to explain and predict GEI has not gained significant popularity among plant breeders (VARGAS et al., 2001; ORTIZ et al., 2007; RAMBURAN; ZHOU; LABUSCHAGNE, 2012; PORKER et al., 2020). With the advancement of computational technology and the democratization of "enviromics" resources, PLS has emerged as a suitable method for exploring big data and performing spatial predictions of experimental genotypes in new environments (MONTEVERDE et al., 2019; RINCENT et al., 2019; GUO et al., 2021; COSTA-NETO et al., 2022). In fact, PLS has emerged as a relevant alternative for prediction purposes, even when breeders do not specifically incorporate environmental data into the model (MONTESINOS-LOPEZ et al., 2022b; MONTESINOS-LOPEZ et al., 2022a; ORTIZ et al., 2023).

In most studies that employed PLS regression for prediction purposes, the training set typically consisted of the performance *per se* of genotypes and environmental features from the tested environments (MONTEVERDE et al., 2019; COSTA-NETO et al., 2022). Our study demonstrated that associating environmental features with the rotated factor loadings of the tested environment yields superior results. Through GIS-FA, we achieved higher prediction accuracy (Table 3) and enhanced the ability to distinguish high-performance experimental genotypes when relying solely on predicted values (Figure 6). By predicting the factor loadings for untested environments, we establish a connection between the observed environmental feature values and the underlying causes of GEI, as well as the genetic covariance that exists between environments. A prior study by Rincent et al. (2019) also utilized PLS models to predict latent factors of the AMMI components for untested environments. This approach enabled them to construct an appropriate covariance structure that improved the accuracy of their predictions. Recently, Callister et al. (2024) used FA and PLS models to define breeding zones for *Eucalyptus globulus* in Australia. The findings of Rincent et al. (2019), Callister et al. (2024) and the results of this work provide evidence of the potential of using PLS models to indirectly perform spatial predictions by initially predicting the latent elements that contribute to a particular performance. Spatial prediction can also be performed in a single step, as proposed by Tolhurst et al. (2022), who demonstrated the efficiency of combining known and latent environmental features to predict both tested and untested environments.

2.5.2.3 Thematic maps

An important feature of GIS-FA is the illustration of the spatial predictions from selection candidates using thematic maps (Figures 7, 8, and 9). Figure 7 offers information on the areas within the breeding zone where the experimental genotypes are expected to thrive. Figure 7 allows the evaluation of the merit of a certain candidate cultivar based on its ability to outperform a commercial cultivar used as a reference or another promising experimental genotype. Figure 9 provides a straightforward solution for genotype recommendations across the breeding region, indicating which candidate is more suitable for a specific environment within the breeding zone. Thematic maps serve as valuable tools in decision-making, assisting in the allocation of genotypes in the breeding region (COSTA-NETO et al., 2020; BUSTOS-KORTS et al., 2022). In addition, the thematic maps provide information on the genotypes' stability and adaptation from a geographic perspective. Costa-Neto et al. (2020) suggested that, in a GIS context, "stability" means lower variability in spatial patterns, while "adaptation" refers to the expected performance in a specific environment in the breeding region.

One advantage of this approach is the possibility of integrating high-quality satellite images from diverse platforms. Here, we used freely available geographic databases on online platforms to achieve an efficient prediction method without incurring any additional costs. Furthermore, implementing partial geographic visualizations can optimize resource allocation when defining the experimental network of trials. The higher resolution of the satellite-based data could enable the delivery of spatial predictions at the farmer's level. This could benefit the product development and placement stages by extending this methodology to accommodate satellite-based enviromics while also accounting for historical agronomic records.

2.5.2.4 Future directions

The statistical models of GIS-FA can be improved by integrating molecular information to leverage the covariance between relatives and employing more informative environmental features in the PLS model (DIAS et al., 2018; MONTEVERDE et al., 2019; CROSSA et al., 2023). The utilization of ecophysiological environmental features in crop growth models could enhance our understanding of the link between phenotypic expression and environmental factors (RINCENT et al., 2019; COSTA-NETO; CROSSA; FRITSCHE-NETO, 2021). GIS-FA can also be benchmarked with other enviromic-based approaches fit for predicting genotypes in untested environments (JARQUIN et al., 2014; TOLHURST et al., 2022; COSTA-NETO et al., 2020). Other statistical resources and even artificial intelligence methods can replace the PLS in the prediction step (GUO et al., 2021; HEINEMANN et al., 2022). Finally, future research can explore the potential risks associated with assigning genotypes to specific environments using GIS-FA. This can be done through the application of probabilistic methods (DIAS et al., 2022).

Declarations

Funding

This research was supported by the Minas Gerais State Agency for Research and Development (FAPEMIG), the Coordination for the Improvement of Higher Education Personnel (CAPES), and the Brazilian National Council for Scientific and Technological Development (CNPq).

Author contribution statement

M.S.A., S.F.S.C., and K.O.G.D. conceived the research. M.S.A. and S.F.S.C. executed the statistical analyses and drafted the initial manuscript. M.D.K. and G.C.N. provided insights into the methodology. L.A.S.D., F.M.F., G.R.P., R.S.A., P.C.S.C., M.D.K., and G.C.N. provided critical revisions of the paper drafts. A.R.G.B. provided knowledge on the structure of the soybean dataset, while A.B.H. and F.B. provided information about the rice dataset. M.S.A., S.F.S.C., and M.D.K. built the tutorial available in the Supplementary Material. All authors approved the final version of the manuscript.

Data availability

The dataset used and/or analysed during the current study is available from the corresponding author upon reasonable request.

Acknowledgements

This work was supported by the Minas Gerais State Agency for Research and Development (FAPEMIG), the Brazilian National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher Education Personnel

(CAPES), the Mato Grosso do Sul Foundation (Fundação MS), the Brazilian Agricultural Research Corporation (Embrapa Rice and Beans), and the Federal University of Viçosa (UFV).

2.6 References

AASTVEIT, A. H.; MARTENS, H. ANOVA interactions interpreted by partial least squares regression. *Biometrics*, v. 42, n. 4, p. 829–844, 1986. 44, 59

ALVARES, C. A. et al. Köppen's climate classification map for brazil. *Meteorologische Zeitschrift*, v. 22, p. 711–728, 2013. 33, 34

ANNICCHIARICO, P.; BELLAH, F.; CHIARI, T. Repeatable genotype × location interaction and its exploitation by conventional and gis-based cultivar recommendation for durum wheat in algeria. *European Journal of Agronomy*, v. 24, p. 70–81, 2006. 22, 32

BADDELEY, A.; RUBAK, E.; TURNER, R. Spatial point patterns: methodology and applications with R. *Journal of Statistical Software*, v. 75, p. 1–6, 2015. 39

BALESTRE, M. et al. Genotypic stability and adaptability in tropical maize based on AMMI and GGE biplot analysis. *Genetics and Molecular Research*, v. 8, n. 4, p. 1311–1322, 2009. 31

BEEBE, S. et al. A geographical approach to identify phosphorus-efficient genotypes among landraces and wild ancestors of common bean. *Euphytica*, v. 95, p. 325–338, 1997. 22, 32

BUNTARAN, H.; FORKMAN, J.; PIEPHO, H. P. Projecting results of zoned multi-environment trials to new locations using environmental covariates with random coefficient models: accuracy and precision. *Theoretical and Applied Genetics*, v. 134, p. 1513–1530, 2021. 58

BUSTOS-KORTS, D. et al. Identification of environment types and adaptation zones with self-organizing maps: applications to sunflower multi-environment data in Europe. *Theoretical and Applied Genetics*, v. 135, p. 2059–2082, 2022. 22, 60

BUTLER, D. *Asreml: fits the linear mixed model*. [S.l.], 2021. R package version 4.1.0.160. Disponível em: <www.vsni.co.uk>. 40

CALLISTER, A. N. et al. Enviromic prediction enables the characterization and mapping of *Eucalyptus globulus* Labill breeding zones. *Tree Genetics & Genomes*, v. 20, p. 3, 2024. 22, 23, 59, 60

CHAPMAN, S.; BARRETO, H. et al. Using simulation models and spatial databases to improve the efficiency of plant breeding programs. *Plant Adaptation and Crop Improvementent*, p. 563–587, 1996. 22, 32

CHAVES, S. F. S. et al. Analysis of repeated measures data through mixed models: An application in *Theobroma grandiflorum* breeding. *Crop Science*, v. 63, n. 4, p. 2131–2144, 2023. 41

CHAVES, S. F. S. et al. Employing factor analytic tools for selecting high-performance and stable tropical maize hybrids. *Crop Science*, v. 63, n. 3, p. 1114–1125, 2023. 22, 43, 57, 58

CHELSA. *Glimatologies at high resolution for the earth's land surface areas*. 2023. Disponível em: <https://chelsa-climate.org/>. 36

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, p. 561–572, 1994. 21, 30, 42, 57

COOPER, M.; MESSINA, C. D. Can we harness "enviromics" to accelerate crop improvement by integrating breeding and agronomy? *Frontiers in Plant Science*, v. 12, p. 735143, 2021. 23, 31

COOPER, M. et al. Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop and Pasture Science*, v. 65, p. 311, 2014. 32, 35

COOPER, M. et al. Predicting genotype × environment × management (G × E × M) interactions for the design of crop improvement strategies. In: _____. *Plant Breeding Reviews*. [S.l.]: John Wiley Sons, Ltd, 2022. cap. 8, p. 467–585. 31

COPPOCK, J. T.; RHIND, D. W. In: _____. [S.l.]: Longman Scientific Technical, 1991. p. 21–43. 32

COSTA-NETO, G. et al. Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. *G3 Genes|Genomes|Genetics*, v. 13, n. 2, p. jkac313, 2022. 44, 58, 59

COSTA-NETO, G.; CROSSA, J.; FRITSCHE-NETO, R. Enviromic assembly increases accuracy and reduces costs of the genomic prediction for yield plasticity in maize. *Frontiers in Plant Science*, v. 12, p. 717552, 2021. 35, 36, 61

COSTA-NETO, G.; FRITSCHE-NETO, R. Enviromics: bridging different sources of data, building one framework. *Crop Breeding and Applied Biotechnology*, v. 21, n. spe, p. e393521S12, 2021. 22, 23, 32

COSTA-NETO, G.; FRITSCHE-NETO, R.; CROSSA, J. Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity*, v. 126, n. 1, p. 92–106, 2021. 23, 32

COSTA-NETO, G. et al. EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. *G3 Genes|Genomes|Genetics*, v. 11, n. 4, p. jkab040, abr. 2021. 31, 37, 58

COSTA-NETO, G. et al. A novel GIS-based tool to reveal spatial trends in reaction norm: upland rice case study. *Euphytica*, v. 216, p. 37, 2020. 22, 23, 32, 45, 60, 61

COWLING, W. A. et al. Optimal contribution selection improves the rate of genetic gain in grain yield and yield stability in spring canola in Australia and Canada. *Plants*, v. 12, p. 383, 2023. 43

CROSSA, J. From genotype × environment interaction to gene × environment interaction. *Current Genomics*, v. 13, n. 3, p. 225–244, 2012. 57

CROSSA, J. et al. Do feature selection methods for selecting environmental covariables enhance genomic prediction accuracy? *Frontiers in Genetics*, v. 14, p. 7016, 2023. 58, 60

CROSSA, J. et al. Interpreting genotype × environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theoretical and Applied Genetics*, v. 99, p. 611–625, 1999. 23, 58, 59

CROSSA, J.; YANG, R.-C.; CORNELIUS, P. L. Studying crossover genotype × environment interaction using linear-bilinear models and mixed models. *Journal of Agricultural, Biological, and Environmental Statistics*, v. 9, n. 3, p. 362–380, 2004. 30

CULLIS, B.; BEECK, C. P.; COWLING, W. A. Analysis of yield and oil from a series of canola breeding trials. Part II: exploring V × E using factor analysis. *Genome*, v. 53, p. 1002–1016, 2010. 42

CULLIS, B. R. et al. 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* breed-ing programme. *Theoretical and Applied Genetics*, v. 127, p. 2193–2210, 2014. 31

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, p. 381, 2006. 42

DABROWSKI, P. S. et al. Three-dimensional thematic map imaging of the yacht port on the example of the polish national sailing Centre Marina in Gdańsk. *Applied Sciences*, v. 11, n. 15, p. 7016, 2021. 45

DAYAL, B. S.; MACGREGOR, J. F. Improved PLS algorithms. *Journal of Chemometrics*, v. 11, n. 1, p. 73–85, 1997. 44, 71

de los Campos, G. et al. A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. *Nature Communications*, v. 11, p. 4876, 2020. 58

DENIS, J. B. Two way analysis using covarites1. *Statistics*, v. 19, n. 1, p. 123–132, 1988. 31, 58

DIAS, K. O. G. et al. Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. *Heredity*, v. 121, p. 24–37, 2018. 31, 60

DIAS, K. O. G. et al. Leveraging probability concepts for cultivar recommendation in multi-environment trials. *Theoretical and Applied Genetics*, v. 135, p. 1385–1399, 2022. 21, 31, 61

DIEPENBROCK, C. H. et al. Can we harness digital technologies and physiology to hasten genetic gain in us maize breeding? *Plant Physiology*, v. 188, n. 2, p. 1141–1157, 2022. 23, 31

DUNNINGTON, D. *Ggspatial: spatial data framework for ggplot2*. [s.n.], 2023. R package version 1.1.8. Disponível em: <https://CRAN.R-project.org/package=ggspatial>. 46

EBERHART, S. A.; RUSSELL, W. A. Stability parameters for comparing varieties. *Crop Science*, v. 6, p. 36–40, 1966. 21, 31

ECMWF. *European Centre for Medium-Range Weather Forecasts*. 2023. Disponível em: <https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00765/>. 36

EEUWIJK, F. A. V.; ELGERSMA, A. Incorporating environmental information in an analysis of genotype by environment interaction for seed yield in perennial ryegrass. *Heredity*, v. 70, n. 5, p. 447–457, 1993. 22, 31, 58

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. 21, 31

EOSDIS. *NASA Earth Observing System Data and Information System*. 2023. Disponível em: <https://worldview.earthdata.nasa.gov>. 36

FAO. *World reference base for soil resources 2014*. 2014. 1-192 p. Disponível em: <www.fao.org/3/i3794en/I3794en.pdf>. 34

FICK, S. E.; HIJMANS, R. J. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, v. 32, p. 4302–4315, 2017. 36

FINLAY, K.; WILKINSON, G. The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, v. 14, p. 742, 1963. 21, 31

GAUCH, H. G.; ZOBEL, R. Identifying mega-environments and targeting genotypes. *Crop Science*, v. 37, p. 311–326, 1997. 21, 31

GHCND. *Global Historical Climatology Network daily*. 2023. Disponível em: <https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily/>. 36

GILMOUR, A. R.; CULLIS, B.; VERBYLA, A. p. Accounting for natural and extraneous variation in the analysis of field experiment. *Journal of Agricultural, Biological and Environmental Statistics*, v. 2, p. 269–293, 1997. 40

GOGEL, B.; SMITH, A.; CULLIS, B. Comparison of a one- and two-stage mixed model analysis of australia's national variety trial southern region wheat data. *Euphytica*, v. 214, p. 44, 2018. 40

GUARINO, L. et al. Geographic information systems (GIS) and the conservation and use of plant genetic resources. In: CABI PUBLISHING WALLINGFORD UK. *Managing plant genetic diversity. Proceedings of an international conference, Kuala Lumpur, Malaysia, 12-16 June 2000*. [S.l.], 2002. p. 387–404. 22, 32

GUO, Y. et al. Prediction of rice yield in East China based on climate and agronomic traits data using artificial neural networks and partial least squares regression. *Agronomy*, v. 11, n. 2, p. 282, 2021. 59, 61

HARTUNG, J.; PIEPHO, H. P. Effect of missing values in multi-environmental trials on variance component estimates. *Crop Science*, v. 61, n. 6, p. 4087–4097, 2021. 33

HEINEMANN, A. B. et al. Enviromic prediction is useful to define the limits of climate adaptation: a case study of common bean in brazil. *Field Crops Research*, v. 286, p. 108628, 2022. 58, 61

HENDERSON, C. R. Estimates of changes in herd environment. *Journal of Dairy Science*, v. 61, p. 294–300, 1949. 21, 31, 40

HENDERSON, C. R. Estimation of genetic parameters. *Annals of Mathematical Statistics*, v. 21, p. 309–310, 1950. 21, 31, 40

HERNANDEZ, M. V. et al. Modeling genotype × environment interaction using a factor analytic model of on-farm wheat trials in the Yaqui Valley of Mexico. *Agronomy Journal*, v. 111, n. 6, p. 2647–2657, 2019. 32

HIJMANS, R. *Raster: Geographic Data Analysis and Modeling. R package version 3.6-3.* 2020. Disponível em: <https://CRAN.R-project.org/package=raster>. 37

HIJMANS, R. J. et al. *geodata: download geographic data.* [S.l.], 2023. R package version 0.5-8. Disponível em: <https://CRAN.R-project.org/package=geodata>. 37

JARQUIN, D. et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, v. 127, n. 3, p. 595–607, 2014. 23, 32, 61

JARQUIN, D. et al. Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. *Frontiers in Genetics*, v. 11, p. 592769, 2021. 58

KRAUSE, M. D. et al. Using large soybean historical data to study genotype by environment variation and identify mega-environments with the integration of genetic and non-genetic factors. *bioRxiv*, v. 4, p. 487885, 2022. 31

LEMBRECHTS, J. J. et al. Global maps of soil temperature. *Global Change Biology*, v. 28, n. 9, p. 3110–3144, 2022. 37

LI, X. et al. Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proceedings of the National Academy of Sciences*, v. 115, n. 26, p. 6679–6684, 2018. 31, 58

LILAND, K. H.; MEVIK, B.-H.; WEHRENS, R. *PLS: partial least squares and principal component regression.* [S.l.], 2022. R package version 2.8-1. Disponível em: <https://CRAN.R-project.org/package=pls>. 44

LINDGREN, F.; GELADI, P.; WOLD, S. The kernel algorithm for PLS. *Journal of Chemometrics*, v. 7, n. 1, p. 45–59, 1993. 44, 71

LYNCH, M.; WALSH, B. *Genetics and analysis of quantitative traits.* 1. ed. Sunderland: Sinauer Associates, 1998. 57

MALOSETTI, M.; RIBAUT, J. M.; EEUWIJK, F. A. V. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Genetics Selection Evolution*, v. 4, p. 44, 2013. 31

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v. 157, p. 1819–1829, 2001. 58

MILLET, E. J. et al. Genomic prediction of maize yield across european environmental conditions. *Nature Genetics*, v. 51, n. 6, p. 952–956, 2019. 22, 31

MONTESINOS-LOPEZ, O. A. et al. Partial least squares enhances genomic prediction of new environments. *Frontiers in Genetics*, v. 13, p. 920689, 2022. 23, 32, 59

MONTESINOS-LOPEZ, O. A. et al. Multi-trait genome prediction of new environments with partial least squares. *Frontiers in Genetics*, v. 13, p. 966775, 2022. 23, 32, 59

MONTEVERDE, E. et al. Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (*Oryza sativa* L.) grown in subtropical areas. *G3 Genes|Genomes|Genetics*, v. 9, n. 5, p. 1519–1531, 2019. 44, 59, 60

MRODE, R. A. *Linear models for the prediction of animal breeding values*. 3rd ed. ed. [S.l.]: CABI, 2014. 43

NASAPOWER. *Prediction of worldwide energy resource*. 2022. Disponível em: <https://power.larc.nasa.gov/data-access-viewer>. 37

NCEI. *Climate Forecast System Reanalysis (CFSR), for 1979 to 2011*. 2018. Disponível em: <https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00765/>. 36

NOAA. *Climate Data Online*. 2023. Disponível em: <https://www.ncei.noaa.gov/cdo-web>. 36

NUVUNGA, J. J. et al. Bayesian factor analytic model: An approach in multiple environment trials. *PLoS ONE*, v. 14, n. 8, p. e0220290, 2019. 32

OLIVEIRA, I. C. et al. 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. 31

ORTIZ, R. et al. Studying the effect of environmental variables on the genotype× environment interaction of tomato. *Euphytica*, v. 153, p. 119–134, 2007. 59

ORTIZ, R. et al. Partial least squares enhance multi-trait genomic prediction of potato cultivars in new environments. *Scientific Reports*, v. 13, n. 1, p. 9947, 2023. 59

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. *Biometrika*, v. 58, p. 545–554, 1971. 40

PEBESMA, E.; BIVAND, R. *Spatial data science: with applications in R*. [S.l.], 2023. 352 p. Disponível em: <https://r-spatial.org/book/>. 46

PIEPHO, H.; MöHRING, J. Selection in cultivar trials—Is it ignorable? *Crop Science*, v. 46, n. 1, p. 192–201, jan. 2006. 33

PIEPHO, H. P. Analysis of a randomized block design with unequal subclass numbers. *Agronomy Journal*, v. 89, p. 718–723, 1997. 21, 31, 40, 57

PIEPHO, H. P. A coefficient of determination (r^2) for generalized linear mixed models. *Biometrical Journal*, v. 61, n. 4, p. 860–872, 2019. 41

PIEPHO, H. P. et al. Blup for phenotypic selection in plant breeding and variety testing. *Euphytica*, v. 161, p. 209–228, 2008. ISSN 00142336. 31

PORKER, K. et al. Using a novel PLS approach for envirotyping of barley phenology and adaptation. *Field Crops Research*, v. 246, p. 107697, 2020. 59

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. Disponível em: <https://www.R-project.org/>. 37

RAMBURAN, S.; ZHOU, M.; LABUSCHAGNE, M. Integrating empirical and analytical approaches to investigate genotype × environment interactions in sugarcane. *Crop Science*, v. 52, n. 5, p. 2153–2165, 2012. 59

RESENDE, R. T. et al. Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theoretical and Applied Genetics*, v. 134, p. 95–121, 2021. 22, 23, 32

RINCENT, R. et al. Using crop growth model stress covariates and AMMI decomposition to better predict genotype-by-environment interactions. *Theoretical and Applied Genetics*, v. 132, n. 12, p. 3399–3411, 2019. 59, 60, 61

ROGERS, A. R. et al. The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3 Genes|Genomes|Genetics*, v. 11, n. 2, p. jkaa050, 2021. 23, 31

SAE-LIM, P. et al. Identifying environmental variables explaining genotype-by-environment interaction for body weight of rainbow trout (*Onchorynchus mykiss*): reaction norm and factor analytic models. *Genetics Selection Evolution*, v. 46, n. 16, p. 1–11, 2014. 31

SANTOS, H. G. *Sistema brasileiro de classificação de solos (in Portuguese)*. 5ª edição revista e ampliada. ed. Brasília, DF: Embrapa, 2018. Disponível em: <https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/1094003/sistema-brasileiro-de-classificacao-de-solos>. 33, 34

SHELFORD, V. E. Animal communities in temperate america as illustrated in the chicago region. *Biological Bulletin*, v. 21, p. 95–167, 1911. 36

SILVA, K. J. et al. Identification of mega-environments for grain sorghum in Brazil using GGE biplot methodology. *Agronomy Journal*, v. 113, p. 1–12, 2021. 31

SMITH, A. et al. 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. 978248, 2021. 31, 41

SMITH, A. B.; CULLIS, B.; THOMPSON, R. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, v. 57, p. 1138–1147, 2001. 21, 31, 40, 41, 57

SMITH, A. B.; CULLIS, B. R. Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, v. 214, p. 143, 2018. 21, 31, 36, 43, 56

SMITH, A. B. et al. Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theoretical and Applied Genetics*, v. 128, p. 55–72, 2015. 42

SOILGRIDS. *SoilGrids — global gridded soil information*. 2022. Disponível em: <https://www.isric.org/explore/soilgrids/>. 36, 37

SPARKS, A. H. Nasapower: a nasa power global meteorology, surface solar energy and climatology data client for R. *Journal of Open Source Software*, v. 3, n. 30, p. 1035, 2018. 37

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. 31, 36, 43, 56

THOMPSON, R. et al. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. *Australian & New Zealand Journal of Statistics*, v. 45, n. 4, p. 445–459, 2003. 41

TOLHURST, D. J. et al. Genomic selection using random regressions on known and latent environmental covariates. *Theoretical and Applied Genetics*, v. 135, p. 3393–3415, 2022. 22, 23, 31, 41, 60, 61

VARGAS, M. et al. Interpreting treatment × environment interaction in agronomy trials. *Agronomy Journal*, v. 93, n. 4, p. 949–960, 2001. 59

VARGAS, M. et al. Mapping QTLs and QTL × environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. *Theoretical and Applied Genetics*, v. 112, n. 6, p. 1009–1023, 2006. 59

VERBYLA, A. P. A note on model selection using information criteria for general linear models estimated using reml. *Aust. N. Z. J. Stat.*, v. 61, p. 39–50, 2019. 74

WICKHAM, H. *Ggplot2: elegant graphics for data analysis.* [S.l.]: Springer. Springer. Cham, 2 editions, 2016. 46

WOLD, H. O. A. *Estimation of principal components and related models by iterative least squares.* New York: Academic Press, 1966. 391–420 p. 44

WOLD, S.; SJöSTRöM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, v. 58, p. 109–130, 2001. 32

WONG, J. *Pdist: partitioned distance function.* [S.l.], 2022. R package version 1.2.1. Disponível em: <https://CRAN.R-project.org/package=pdist>. 39

WOOD, J. T. The use of environmental variables in the interpretation of genotype-environment interaction. *Heredity*, v. 37, n. 1, p. 1–7, 1976. 22, 31

XU, Y. Envirotyping for deciphering environmental impacts on crop plants. *Theoretical and Applied Genetics*, v. 129, p. 653–673, 2016. 22, 32, 35

YAN, W. et al. Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science*, v. 40, p. 597–605, 2000. 21, 31

YAN, W. et al. GGE biplot vs. AMMI analysis of genotype-by-environment data. *Crop Science*, v. 47, p. 643–653, 3 2007. 31

YATES, F.; COCHRAN, W. G. The analysis of groups of experiments. *Journal of Agricultural Science*, v. 28, p. 556–580, 1938. 31

2.7 Appendix

Partial least squares regression

Here, we employed the kernel PLS algorithm (LINDGREN; GELADI; WOLD, 1993; DAYAL; MACGREGOR, 1997) to predict the factor loadings of untested environments. Details about this algorithm are presented below:

Take the following multiple regressions as a starting point:

$$\hat{\mathbf{\Lambda}}^{\star} = \mathbf{WB} + \mathbf{E} \tag{2.20}$$

where $\hat{\mathbf{\Lambda}}^{\star}$ is the $J \times K$ matrix of K rotated loadings for the J observed environments, $\mathbf{W}$ is a $J \times P$ matrix of scaled values for P environmental features in the J observed environments, $\mathbf{B}$ is a $P \times K$ vector of coefficients, and $\mathbf{E}$ is a $J \times K$ matrix of lack of fit effects. Note that most of the environmental features are correlated (Supplementary Figure 4), so $\mathbf{W}$ has multicollinearity problems, and $\mathbf{B} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\hat{\mathbf{\Lambda}}^{\star}$ does not yield a proper solution. To overcome this issue, we employed kernel PLS regression to transform $\mathbf{B}$ into $\mathbf{B}^*$, using the following equation:

$$\mathbf{B}^* = \mathbf{\Phi}(\mathbf{\Theta}'\mathbf{\Phi})^{-1}\mathbf{\Xi}' \tag{2.21}$$

where $\mathbf{\Phi}$ is a $P \times C$ matrix of weights for $\mathbf{W}$ ($\mathbf{\Phi} = \{\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2\ \ldots \boldsymbol{\phi}_C\}$), with C being the number of PLS components; $\mathbf{\Theta}$ is a matrix of loadings for $\mathbf{W}$ ($\mathbf{\Theta} = \{\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2\ \ldots \boldsymbol{\theta}_C\}$) and has the same dimension as $\mathbf{\Phi}$, and $\mathbf{\Xi}$ is a $K \times C$ matrix of weights for $\mathbf{\Lambda}$ ($\mathbf{\Xi} = \{\boldsymbol{\xi}_1\ \boldsymbol{\xi}_2\ \ldots \boldsymbol{\xi}_C\}$). We describe the CV procedure that defined the number of components ($c = 1, 2, \ldots, C$) in Section 2.3.6. $\mathbf{\Phi}$, $\mathbf{\Theta}$, and $\mathbf{\Xi}$ were defined using an iterative process that leveraged the kernel functions of $\mathbf{W}$ and $\mathbf{\Lambda}$. First, $\boldsymbol{\phi}_c$ is estimated as the eigenvector that is equivalent to the largest eigenvalue of the kernel $\mathbf{W}'\hat{\mathbf{\Lambda}}^\star\hat{\mathbf{\Lambda}}^{\star\prime}\mathbf{W}$. We used this vector to initialize an iterative process whose number of repetitions is equivalent to C. Let $\mathbf{R} = \mathbf{\Phi}(\mathbf{\Theta}'\mathbf{\Phi})^{-1}$, with $\mathbf{R} = \{\mathbf{r}_1\ \mathbf{r}_2\ \ldots\ \mathbf{r}_C\}$. In the first iteration, $\mathbf{r}_1 = \boldsymbol{\phi}_1$. Subsequently, $\mathbf{r}_c = \boldsymbol{\phi}_c - \boldsymbol{\theta}'_{c-1}\boldsymbol{\phi}_c\boldsymbol{\xi}'_{c-1}$. On each iteration, $\boldsymbol{\theta}_c$ and $\boldsymbol{\xi}_c$ are estimated as follows:

$$\boldsymbol{\theta}_c = \frac{\mathbf{r}'_c(\mathbf{W}'\mathbf{W})}{\mathbf{r}'_c(\mathbf{W}'\mathbf{W})\mathbf{r}_c} \quad \boldsymbol{\xi}_c = \frac{\mathbf{r}'_c(\mathbf{W}'\hat{\mathbf{\Lambda}}^\star)}{\mathbf{r}'_c(\mathbf{W}'\mathbf{W})\mathbf{r}_c} \tag{2.22}$$

The solutions of these equations are stored in $\mathbf{\Theta}$ and $\mathbf{\Xi}$, respectively, and are used to update the covariance matrix for the next iteration as follows:

$$(\mathbf{W}'\hat{\mathbf{\Lambda}}^\star)_{c+1} = (\mathbf{W}'\hat{\mathbf{\Lambda}}^\star)_c - \boldsymbol{\theta}_c\boldsymbol{\xi}'_c[(\mathbf{W}\mathbf{r}_c)'\mathbf{W}\mathbf{r}_c] \tag{2.23}$$

When the iteration process is finished, $\mathbf{B}^*$ provides a proper solution to Equation 2.20 and can be used for prediction purposes. We used $\mathbf{B}^*$ in Equation 2.17 to train the PLS model and in Equation 2.18 to make predictions.

2.8 Supplementary Material

Figure 10 – Figure S11A: Information about data connectivity for rice and soybean cultivation. (a) corresponds to the Rice data evaluated during crop seasons 2009 and 2010

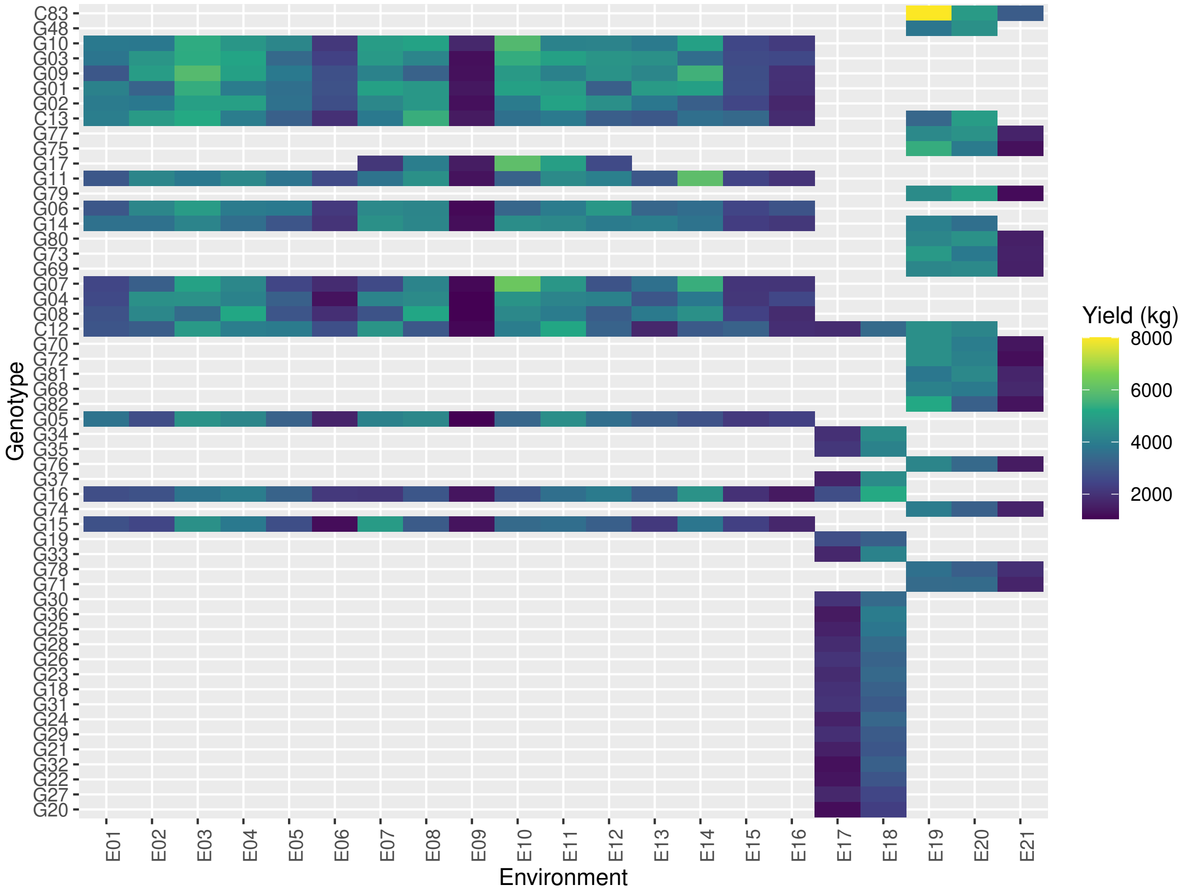


Table S1: Evaluation locations of rice and soybean genotypes conducted in the state of Mato Grosso do Sul and in eight Brazilian states, respectively

Rice data				
Location	**State**	**2009/10**		**2010/11**
Altamira	PA	E07		-
Anápolis	GO	E14		-
Boa Vista	RR	E01		-
Caceres	MT	E11		E20
Colinas	MA	E04		-
Goianira	GO	-		E17
Marianópolis	TO	E05		-
Palmeiras de Goiás	GO	E15		-
Porto Nacional	TO	E06		-
Santa Carmem	MT	E10		-
Santo Antônio de Goiás	GO	E13		
São José dos Quatro Marcos	MT	-		E21
São Raimundo das Mangabeiras	MA	E02		E18
Sinop	MT	E16		-
Sorriso	MT	E09		E19
Teresina	PI	E03		-
Uruará	PA	E08		-
Vilhena	RO	E12		-
Soybean data				
Location	**State**	**2019/20**	**2020/21**	**2021/22**
Anaurilândia		E1, E2*	E20	-
Antônio João		E03	E21	E36
Bela Vista		-	-	E37
Caarapó		E04	E22	-
Campo Grande		E05	-	-
Itaporã		E06	E23	E38
Ivinhema	MS	E07	E24	E39, E40
Maracaju		E08, E09, E10, E11	E25, E26, E27, E28	E41
Naviraí		E12, E13	E29, E30	E42
Rio Brilhante		E14, E15, E16	E31, E32	E43, E44, E45
São Gabriel do Oeste		E17, E18	E33, E34	E46, E47
Sidrolândia		E19	E35	E48
Terenos		-	-	E49

*Trials conducted at the same location at different seasons.

For the soybean dataset, we fitted thirteen statistical models to test if there are spatial trends within the plots (Table S2). The standard model (M0) does not include any spatial adjustments. In the M1 model, the residual variance matrix was modeled by applying a first-order autoregressive structure ($AR1 \otimes AR1$) in row and column directions. Other structures were tested in the row direction, such as linear and spline. Because blocking was made in the column direction, we did not include the column effect in the model. Since the tested models have different fixed effects, the residual maximum likelihood cannot be compared using classical information criteria such as the AIC. Therefore, we used the AIC corrected by Verbyla (2019) (AIC_c), which can be calculated as follows:

$$AIC_c = -2logL + 2(p + q) \tag{2.24}$$

where $logL$ is the logarithm of the maximized value of the likelihood function for the fitted model, p and q are the number of fixed and random parameters in the model, respectively.

Table S2: Description of the 13 models adjusted for each productivity trial for the soybean dataset

Model	Fixed effect[a]	Random effect[b]	Residual[c]
M0	block	gen	$\mathbf{I}\sigma_e^2\mathbf{I}_r \otimes \mathbf{I}_c$
M1	block	gen	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M2	block	gen + nugget	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M3	block	gen + nugget + row	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M4	block + lin(row)	gen + nugget + row	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M5	block + spl(row)	gen + nugget + row	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M6	block + lin(row)	gen + nugget	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M7	block + spl(row)	gen + nugget	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M8	block	gen + row	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M9	block + lin(row)	gen + row	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M10	block + spl(row)	gen + row	$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M11	block + lin(row)		$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$
M12	block + spl(row)		$\sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$

[a] $lin(row)$ and $spl(row)$ are the effects of linear regression for the fixed effect of "row" and the spline component of "row" for the fixed effect, respectively. [b] row: random effect of "row"; nugget $= \sigma_\eta^2(I_r \otimes I_c)$ is the variance-covariance structure of the independent part of the residual, assuming a first-order autoregressive correlation between rows and columns. [c] $AR1 \otimes AR1 = \sigma_\xi^2[\sum_r(\rho_r) \otimes \sum_c(\rho_c)]$ variance-covariance structure of the residual considers the first-order autoregressive correlation in the row-column direction, that is, the dependent portion of the residual; $\mathbf{I}$ is an identity matrix; σ_e^2 is the residual variance; σ_η^2 is the residual variance of the independent part; and σ_ξ^2 is the residual variance of the correlated part; *block* represents block repetition; *gen* represents the genotype. Although tested, the nugget effect η was disregarded in the joint model due to convergence difficulties.

Figure 11 – Figure S11B: Information about data connectivity for rice and soybean cultivation. (b) Soybean data evaluated in 2019, 2020, and 2021. The colored cells indicate that the genotype was evaluated in that environment. The color corresponds to the productivity range. Cells without color represent the location where the genotype was not evaluated.

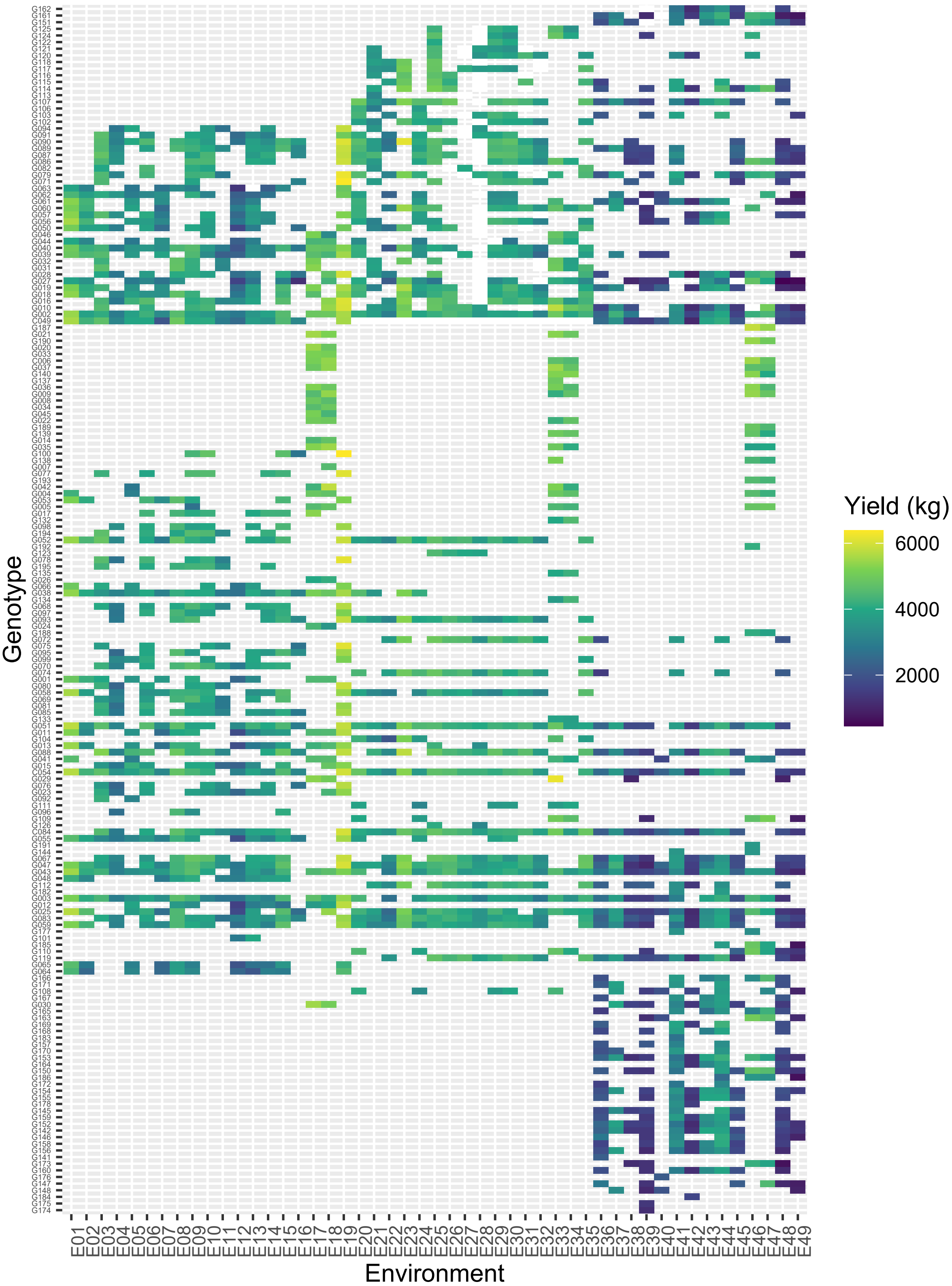


Figure 12 – Figure S12: Adjusted scores from the factor-analytic (FA) model displaying high correspondence plots between FA_1 scores and empirical Best Linear Unbiased Prediction (eBLUP) for (a) the Rice dataset and (b) the Soybean dataset

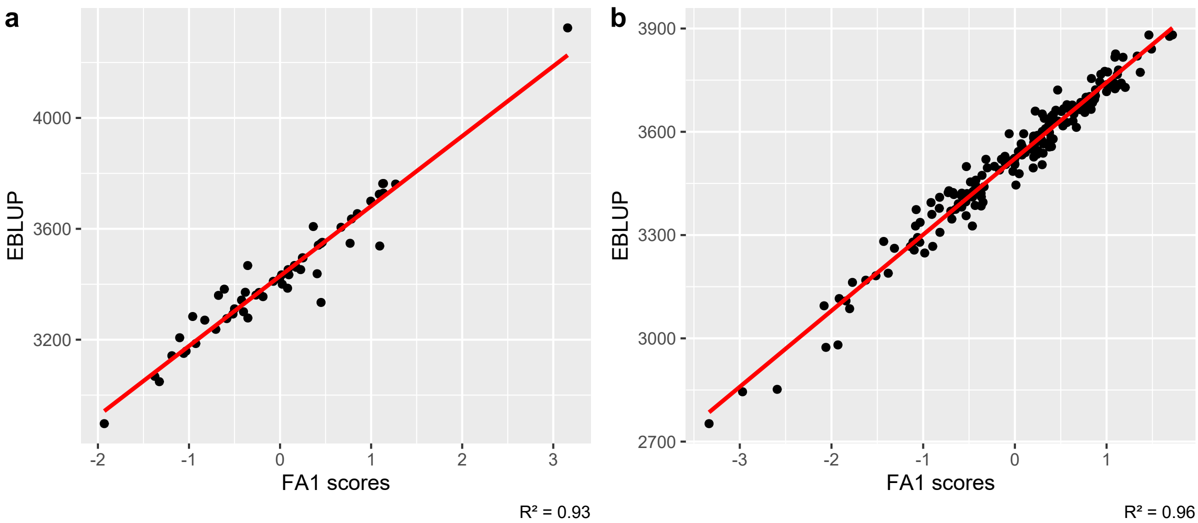


Figure 13 – Figure S13: Fitted factor analytic mixed models for the rice (a) and soybean (b) datasets and their total variance to four factors

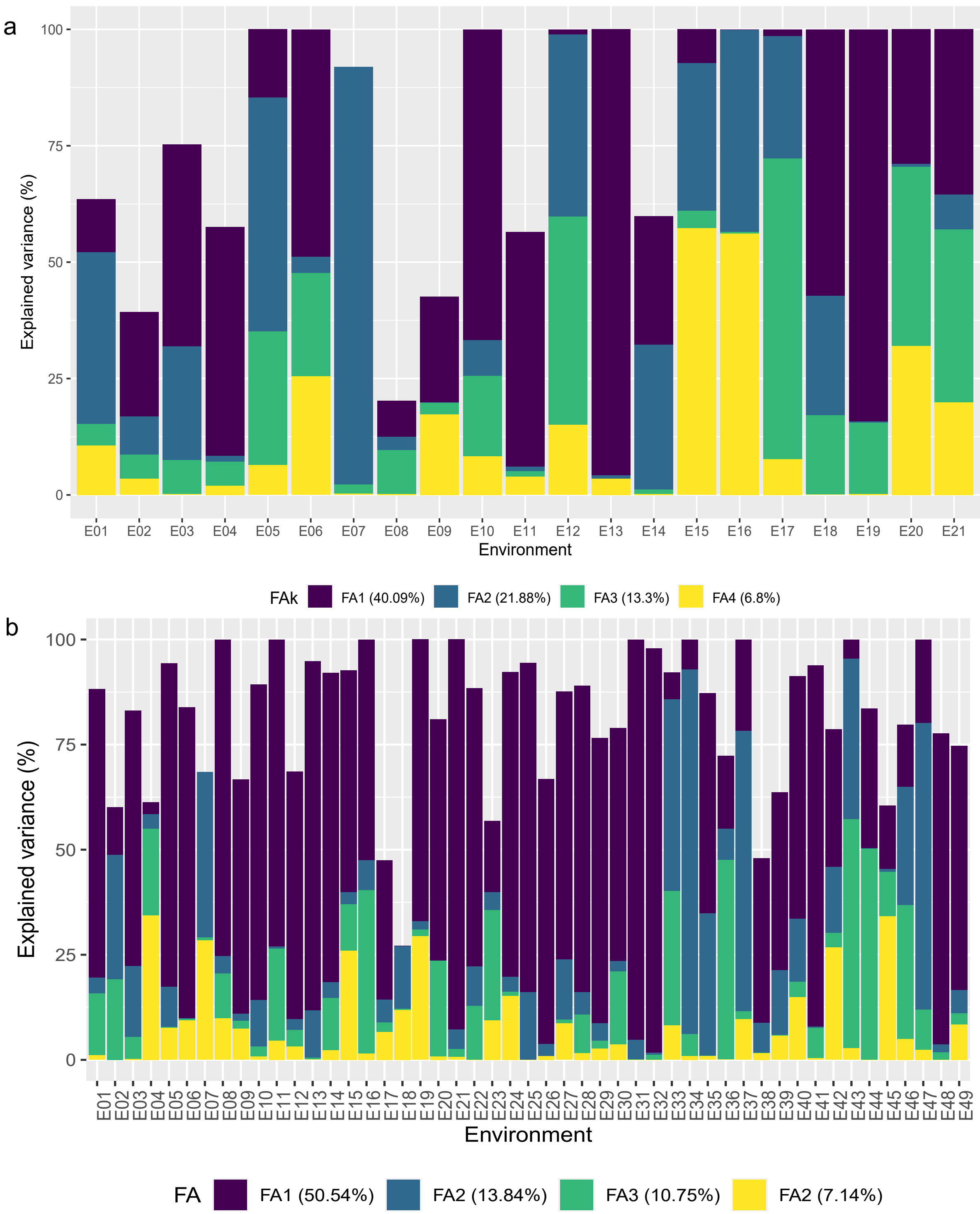
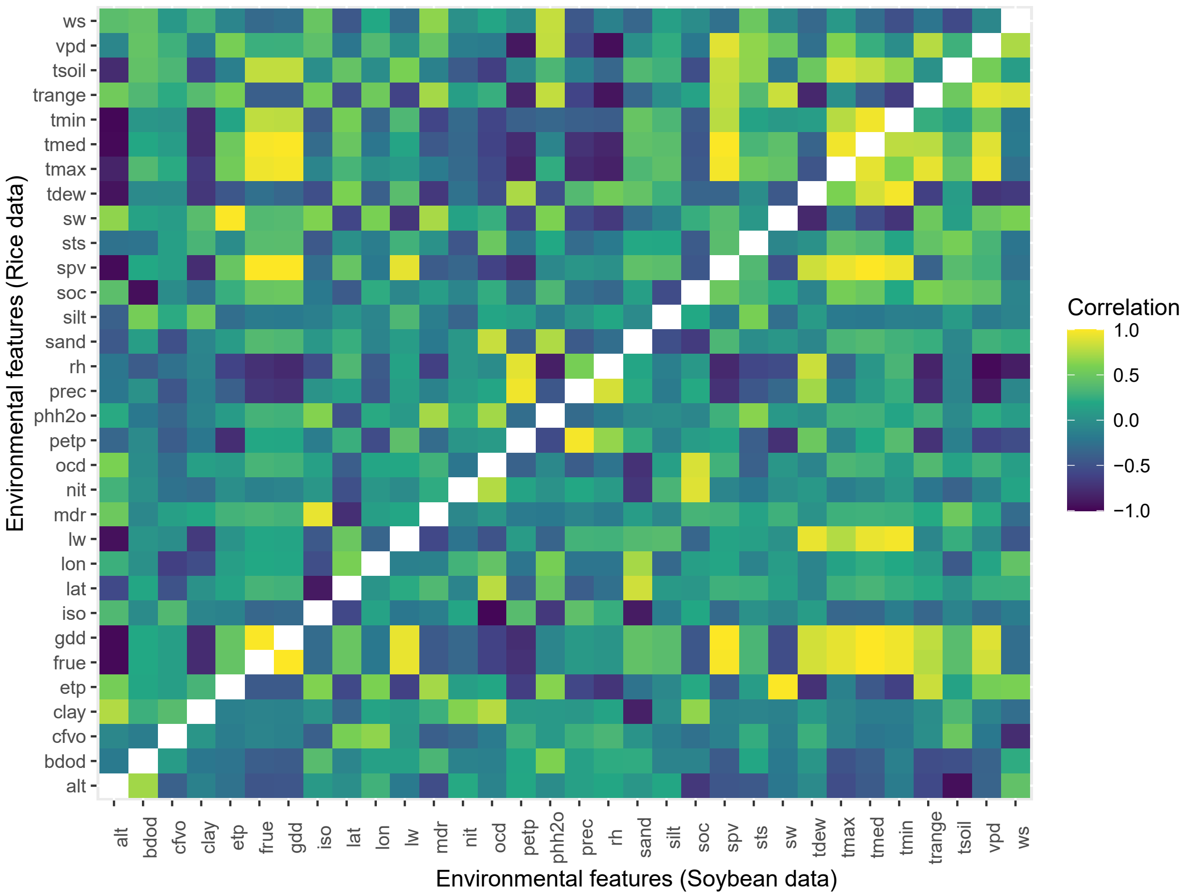


Figure 14 – Figure S14: Pearson correlation among 32 environmental features. On the left diagonal are all the features for the rice dataset, and on the right side are the soybean data



For the rice dataset, check C83 beated all pure lines, so it was not plotted in the "which-won-where" thematic map.

Figure 15 – Figure S15: Map which-won-where with check C83 from the rice dataset. This genotype was not compared with the others due to its superior productivity across the TPE. The white line delimits the Pantanal biome