**UNIVERSIDADE FEDERAL DE VIÇOSA**

**IZABELA CLARA FIALHO**

# ANÁLISE DE FATORES APLICADA A PREDIÇÃO GENÔMICA CONSIDERANDO SELEÇÃO DE MARCADORES EM *Oryza sativa*

**VIÇOSA - MINAS GERAIS**
**2021**

**IZABELA CLARA FIALHO**

# ANÁLISE DE FATORES APLICADA A PREDIÇÃO GENÔMICA CONSIDERANDO SELEÇÃO DE MARCADORES EM *Oryza sativa*

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Camila Ferreira Azevedo

Coorientadores: Moysés Nascimento
Ana Carolina Campana Nascimento

**VIÇOSA - MINAS GERAIS**
**2021**

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

T
F438a
2021

Fialho, Izabela Clara, 1994-
Análise de fatores aplicada a predição genômica considerando seleção de marcadores em *Oryza sativa* / Izabela Clara Fialho. – Viçosa, MG, 2021.
49 f. : il. (algumas color.) ; 29 cm.

Orientador: Camila Ferreira Azevedo.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 46-49.

1. Arroz - Seleção - Métodos estatísticos. 2. Genômica . 3. BLUP. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 633.182

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB6 2523

**IZABELA CLARA FIALHO**

# ANÁLISE DE FATORES APLICADA A PREDIÇÃO GENÔMICA CONSIDERANDO SELEÇÃO DE MARCADORES EM *Oryza sativa*

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 19 de fevereiro de 2021.

Assentimento:

Izabela Clara Fialho
Izabela Clara Fialho
Autora

Camila Ferreira Azevedo
Camila Ferreira Azevedo
Orientadora

*In memoriam* ao meu pai, José Antônio Fialho,
com todo meu amor e gratidão.

## AGRADECIMENTOS

A Deus, o criador infinito e eterno, preservador do Universo. O qual me fez sua imagem e semelhança, para que eu pudesse ser capaz de realizar todos os meus sonhos ao ter fé.

A mim mesma, porque tudo que acontece em nossas vidas começa por nós mesmo. Sou capaz de vivenciar uma vida mais leve ou com dificuldades, simplesmente por escolher como quero viver. Então, agradeço por não ter permitido que a dúvida permanecesse em meu coração e acreditado fielmente que esse momento chegaria. Também, por ter me dedicado e por ter me permitido vivenciar essa experiência e principalmente, por eu possuir a consciência de ser uma pessoa que foi criada com uma porção da própria energia de Deus e por isso, possuo a energia que nasce dele.

Ao meu pai José Antônio Fialho (*in memoriam*) que foi o primeiro a acreditar que esse momento chegaria. Sua certeza era grande, maior que de qualquer outra pessoa, até mesmo da minha que não tinha ideia do que viria pela frente ao finalizar a graduação. Acredito que tive sua ajuda para ingressar no mestrado. Afinal, não perdemos uma pessoa, mas ganhamos um mentor espiritual.

A minha mãe, Maria Aparecida Lopes Fialho, fonte inesgotável de amor, que tamanha é sua dedicação por querer me proteger e me fazer feliz, faz dos meus sonhos os seus.

Ao meu irmão Welinton pelo companheirismo e torcida. Pelas conversas de desabafos e motivações. Por estar presente em todos os momentos.

Ao meu amado, Carlos Ayallas por ser luz no meu caminho. Nessa jornada, clareou meus conhecimentos ao me ajudar nas disciplinas e em programação. Também, iluminou meu coração e o encheu de paz quando surgiram os momentos difíceis. Além, de intensificar as energias positivas existente em mim.

Ao meu amigo Lucas Coelho por ter sido o meu anjo. Não mediu esforços para me ajudar a ingressar no mestrado. Foram muitos materiais emprestados e dúvidas tiradas. Além disso, me abençoou com sua torcida e se permite ser feliz com a minha felicidade.

A toda minha família e amigos, por torcerem pela minha conquista e pelas orações que me fortaleceram.

A minha orientadora Camila Ferreira Azevedo, pelos ensinamentos, pela paciência, preocupação, confiança e dedicação a pesquisa. Sou muito grata por Deus ter colocado uma

pessoa abençoada para me orientar. Seu jeito cativante e amoroso me transmitiu boas energias que foi capaz de tornar essa jornada mais leve.

Aos meus coorientadores Ana Carolina Campana Nascimento e Moysés Nascimento, por contribuírem diretamente no meu aprendizado e pelas sugestões nos trabalhos até aqui realizados.

Aos professores e funcionários do departamento de estatística da Universidade Federal de Viçosa, por se mostrarem acessíveis, dispostos a compartilharem conhecimento e suporte para com os alunos.

A Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria pela oportunidade.

A Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudo. O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) ou Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG). Também, o presente trabalho foi realizado com apoio da CAPES - Brasil - Código de Financiamento 001.

Por fim, a todos os que contribuíram direta ou indiretamente para a concretização desta etapa da minha vida. Gratidão!

## BIOGRAFIA

IZABELA CLARA FIALHO, filha de Maria Aparecida Lopes Fialho e José Antônio Fialho, nasceu em Viçosa, Minas Gerais, em 4 de abril de 1994.

Em maio de 2012, ingressou no curso de Matemática na Universidade Federal de Viçosa, Viçosa-MG. No ano de 2016 mudou-se para o curso de Licenciatura em Matemática pela mesma universidade, graduando em janeiro de 2018.

Em maio de 2019, iniciou o curso de Mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 19 de fevereiro de 2021.

## RESUMO

FIALHO, Izabela Clara, M.Sc., Universidade Federal de Viçosa, fevereiro de 2021. **Análise de fatores aplicada a predição genômica considerando seleção de marcadores em *Oryza sativa*.** Orientadora: Camila Ferreira Azevedo. Coorientadores: Ana Carolina Campana Nascimento e Moysés Nascimento.

O arroz asiático *Oryza sativa* é um dos alimentos mais consumidos em grande parte do mundo. Assim, o crescimento populacional justifica o interesse dos pesquisadores em tornar as variedades deste arroz altamente produtivas. Para alcançar esse objetivo, a Seleção Genômica Ampla (*Genome Wide Selection* - GWS) é uma ferramenta utilizada pelos programas de melhoramento. A GWS emprega dados fenotípicos e genotípicos por meio de marcadores SNPs (*Single Nucleotide Polymorphism*) amplamente distribuídos no genoma. Porém, nem todos estes SNPs estão associados as características de relevância para o pesquisador, o que torna necessário o uso de métodos estatísticos para a seleção de marcadores. O BLUP genômico (*Genomic best linear unbiased prediction* – G-BLUP) é um método amplamente aplicado a predição genômica, e quando este está associado a seleção de marcadores dá origem ao chamado G-BLUP supervisionado. Dessa forma, o presente trabalho objetivou-se avaliar a eficiência na predição genômica ao combinar a análise de fatores e a seleção de marcadores via G-BLUP supervisionado para grupos de características de interesse. O conjunto de dados de arroz utilizado é público e faz parte de dois projetos, o Projeto OryzaSNP e o Projeto OMAP e está disponível no site http://ricediversity.org/data/. O arquivo contém informações de 28 características fenotípicas e 36.901 marcadores SNPs de 413 indivíduos. Os resultados obtidos indicam que a aplicação da análise de fatores combinada à seleção de marcadores SNPs para a predição genômica se mostrou eficiente, visto que apresentaram valores semelhantes para capacidade preditiva em relação aos encontrados considerando as análises individuais das variáveis (em ambas as análises obteve-se variação entre 0,6 a 0,8 de capacidade preditiva) e alta concordância (acima de 50% de concordância para todos os grupos de marcadores) entre os indivíduos selecionados considerando o fator e as variáveis individuais.



**Palavras-chave:** Arroz. Seleção genômica. G-BLUP.

## ABSTRACT

FIALHO, Izabela Clara, M.Sc.,Universidade Federal de Viçosa, February 2021, **Factor analysis applied to genomic prediction considering selection of markers in Oryza Sativa**. Advisor: Camila Ferreira Azevedo. Co-advisors: Ana Carolina Campana Nascimento and Moysés Nascimento.

Asian rice, Oryza sativa, is one of the most consumed foods in much of the world. Thus, the population growth justifies the researchers' interest in making the rice varieties highly productive. For this purpose, Genome-Wide Selection (GWS) is a tool used by breeding programs. GWS employs phenotypic and genotypic data through SNPs (Single Nucleotide Polymorphism) markers widely distributed in the genome. However, not all of these SNPs are associated with traits of relevance to the researcher, making it necessary to use statistical methods for selecting markers. The genomic BLUP (Genomic best linear unbiased prediction - G-BLUP) is widely applied to genomic prediction. The G-BLUP with selecting markers called supervised G-BLUP (SG-BLUP). Thus, the present study aimed to evaluate the efficiency in genomic prediction by combining factor analysis and selecting markers via supervised G-BLUP for groups of traits of interest. The rice dataset used is public, and it is part of two projects, the OryzaSNP Project and the OMAP Project, and is available at http://ricediversity.org/data/. The file contains 28 phenotypic traits and 36,901 SNPs from 413 individuals. The results indicate that the application of factor analysis combined with the selection of SNPs for genomic prediction proved to be efficient. They presented similar values for predictive ability concerning those found considering the individual analyzes of the variables (in both analyzes obtained if the variation between 0.6 to 0.8 of predictive ability) and high agreement (above 50% agreement for all groups of markers) between the selected individuals considering the factor and the individual variables.



**Keywords:** Rice. Genomic selection. G-BLUP.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## SUMÁRIO

# INTRODUÇÃO GERAL

As metodologias estatísticas da Seleção Genômica Ampla (*Genome Wide Selection* - GWS) são ferramentas utilizadas pelos programas de melhoramento, tanto animal como vegetal. Meuwissen et al. (2001) a propuseram visando utilizar as informações diretas do DNA para selecionar indivíduos geneticamente superiores de forma rápida e mais acurada. Segundo Resende et al. (2012), a GWS tem como fim a seleção e a predição do mérito genético dos indivíduos, fundamentado na análise simultânea entre os fenótipos e um grande número de marcadores, como por exemplo os SNPs (*Single Nucleotide Polymorphism*). No entanto, apesar dos SNPs serem amplamente distribuídos no genoma, nem todos eles estão associados a uma determinada característica de interesse. Dessa forma, a existência desse vasto número de marcadores SNPs no genoma e a demanda dos programas de melhoramento em contemplar uma grande quantidade características de interesse, faz-se necessário o uso de métodos estatísticos multivariados que possam reduzir o conjunto de variáveis originais em variáveis latentes e que considerem a seleção desses marcadores.

Visando essa necessidade, um método que objetiva transformar as variáveis fenotípicas originais em variáveis latentes, denominadas fatores, é a análise de fatores (AF). Desse modo, espera-se que um grupo de fenótipos esteja representado por um fator e este seja usado para estimar o mérito genético dos indivíduos. Assim, a seleção dos indivíduos geneticamente superiores por meio do fator contemplaria mais de uma característica. Teixeira et al. (2015, 2016) propuseram a aplicação desse método em características fenotípicas de suínos a fim de reduzir o número de variáveis resposta para a GWS e obtiveram sucesso. Apesar da AF ter sido usada anteriormente, ela ainda não foi empregada conjuntamente a metodologias de seleção de marcadores.

Os métodos estatísticos aplicados a predição genômica são divididos em métodos baseados em efeitos de marcadores e métodos baseados em parentesco genômico entre os indivíduos (Resende et al. 2012). Os principais métodos baseados em marcadores são RR-BLUP (*Randon Regression Best Linear Unbiased Predictor*) (Whittaker et al., 2000), BayesA e BayesB (Meuwissen et al. 2001) e o BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*) (de los Campos et al. 2009). O principal método baseado em matriz de parentesco genômico é o BLUP genômico (*Genomic best linear unbiased prediction* – G-BLUP) e este é amplamente aplicado na predição genômica, como exemplos podemos citar os estudos realizados por Wang et al. (2018), Alvarenga et al. (2020) e Karaman et al. (2020).

A seleção de marcadores é importante para a predição genômica que faz uso de métodos baseados em efeito de marcadores pois nestes casos o processo de estimação apresenta dois desafios: i) alta dimensionalidade: o número de marcadores é muito maior que o número de indivíduos genotipados e fenotipados, ii) multicolinearidade: os marcadores são altamente correlacionados. A alta dimensionalidade impossibilita a obtenção das estimativas do efeito dos marcadores no fenótipo por meio de métodos fundamentados em quadrados mínimos ordinários. Ademais, caso não haja a alta dimensionalidade, a multicolinearidade impossibilitaria a obtenção de estimativas estáveis dos efeitos dos marcadores sobre os fenótipos por meio dos quadrados mínimos ordinários.

Ademais, os métodos baseados em parentesco genômico não sofrem dos problemas advindos da alta dimensionalidade e da multicolinearidade. No entanto, dois indivíduos são geneticamente idênticos, para uma característica, se carregam o mesmo genótipo nos locos de características quantitativas (*Quantitative trait locus* - QTL) causais ou genes, ou indiretamente os mesmos marcadores em desequilíbrio de ligação (*Linkage disequilibrium* - LD) com eles. Dessa forma, a seleção de marcadores se torna importante neste contexto também, uma vez que a matriz de parentesco genômica deve ser idealmente construída contendo apenas os marcadores relevantes para determinada característica (Resende et al., 2012).

Desse modo, o presente estudo teve como objetivo aplicar a análise de fatores sob o enfoque de seleção de marcadores associados a grupos de características de interesse (fator) e avaliar sua eficiência na predição genômica via BLUP.

Esta dissertação está dividida em dois capítulos. No capítulo 1 foi realizado uma revisão de literatura apresentando a definição e importância da GWS, bem como a descrição e desenvolvimento dos métodos que serão utilizados posteriormente no capítulo seguinte: Análise de fatores; RR-BLUP e G-BLUP. Já no capítulo 2 aplicou-se a análise de fatores combinada à seleção de marcadores SNPs associados a grupos de características de interesse na predição genômica via BLUP genômico e comparou-se os resultados com os obtidos via análises individuais considerando um único fenótipo por análise.

# CAPÍTULO 1

# REVISÃO DE LITERATURA

## 1.1 Seleção Genômica Ampla

A Seleção Genômica Ampla (*Genome Wide Selection* - GWS) foi proposta por Meuwissen et al. (2001), e tem como objetivo principal predizer o valor genético futuro de uma população baseando-se apenas em informações de marcadores moleculares. Os marcadores moleculares mais utilizados para a predição genômica são os SNPs (*Single Nucleotide Polymorphism* - SNPs) devido a sua genotipagem em larga escala, sua abundância no genoma e sua codominância (Resende et al., 2010). A GWS utiliza os marcadores SNPs e as características de interesse visando estimar os efeitos destes marcadores nos fenótipos, e posteriormente, selecionar e predizer o mérito genético dos indivíduos com maior rapidez e eficiência. De posse dos efeitos dos marcadores é possível que os indivíduos geneticamente superiores sejam rapidamente identificados, não havendo necessidade de esperar a manifestação da característica desejada pelo indivíduo para a seleção.

No entanto, nem todos os marcadores SNPs estão em desequilíbrio de ligação (*Linkage disequilibrium* - LD) com os locos de características quantitativas (*Quantitative trait locus* - QTL), ou seja, nem todos os marcadores SNPs estão associados a uma determinada característica, fazendo-se necessário o uso de métodos estatísticos para a seleção destes marcadores. Além disso, os programas de melhoramento contemplam uma infinidade de características de interesse e enfrentam alguns desafios, como a alta dimensionalidade e multicolinearidade. Neste contexto, métodos estatísticos que considera a seleção de variáveis e que permita analisar um grupo de características conjuntamente são desejáveis. Dessa forma, a análise de fatores aplicada a matriz de fenótipos juntamente com um método estatístico de predição genômica e de seleção de marcadores podem contornar estes desafios.

## 1.2 Análise de fatores

### 1.2.1 Definição

A análise de fatores foi desenvolvida com o objetivo de descrever a variabilidade original dos dados através de um conjunto menor de variáveis denominadas fatores, de tal forma que tenhamos uma perda mínima de informações. Assim, segundo Mingoti (2007) os fatores estão relacionados a partir de um modelo linear com a variabilidade original do vetor aleatório

**X**. Tabachinick e Fidell (2007) relatam que essa técnica pode ser dividida em análise de fatores confirmatória (AFC) e exploratória (AFE).

A AFC pretende testar uma hipótese, em que o pesquisador parte de uma teoria pré-concebida a respeito do número de fatores ou variáveis latentes, que são combinações lineares das variáveis originais. Já a AFE, que é a mais utilizada, e a qual trabalharemos no próximo capítulo, não é necessário o conhecimento prévio, pelo pesquisador, da relação de dependência entre as variáveis. Mas, deve fornecer ao pesquisador, o número de fatores necessários para melhor representar os dados explorados.

### 1.2.2 Autovalores e Autovetores

Um dos métodos utilizados para realizar a análise de fatores é a partir do conhecimento de autovalores e autovetores. Segundo Anton e Rorres (2012), se tivermos uma matriz $\boldsymbol{A}$ quadrática (dimensão $n \times n$), um escalar $\lambda$ é denominado um autovalor de $\boldsymbol{A}$ se existe um vetor não nulo $\boldsymbol{x}$ em $R^n$, tal que $\mathbf{A}\boldsymbol{x} = \lambda\boldsymbol{x}$. Dessa forma, o vetor $\boldsymbol{x}$ é um autovetor de $\boldsymbol{A}$ associado a $\lambda$. Assim, as seguintes afirmações são equivalentes:

(i) $\lambda$ é um autovalor de $\boldsymbol{A}$;

(ii) O sistema de equações $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \mathbf{0}$ tem solução não trivial;

(iii) $\boldsymbol{A} - \lambda\boldsymbol{I}$ é singular;

(iv) $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$ sendo $det(\ )$ a notação para o cálculo do determinante de uma matriz quadrada e $\lambda$ uma solução dessa equação característica.

### 1.2.3 Decomposição Espectral

A partir do conhecimento de autovalores e autovetores, pode-se realizar a Decomposição Espectral. Esse resultado é de suma importância na Análise de fatores, pois possibilita a identificação das variáveis latentes.

Considere uma matriz $\boldsymbol{A}$ quadrática (dimensão $n \times n$) e simétrica, ou seja, uma matriz que satisfaz a condição $\boldsymbol{A} = \boldsymbol{A}'$ (sendo $\boldsymbol{A}'$ a matriz transposta de $\boldsymbol{A}$). Logo, a decomposição espectral da matriz $\boldsymbol{A}$ é dada por (FERREIRA, 2011):

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}' \qquad (1.1)$$

em que $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$, $\boldsymbol{P} = [\boldsymbol{x_1} \quad \boldsymbol{x_2} \quad \cdots \quad \boldsymbol{x_n}]$, são respectivamente, a matriz diagonal dos autovalores $(\lambda_1, \lambda_2, \dots, \lambda_n)$ e a matriz ortogonal formada pelo autovetores $(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})$ da matriz $\boldsymbol{A}$.

### 1.2.4 Adequabilidade da matriz de correlação

#### 1.2.4.1 Critério de Kaiser-Meyer-Olkin (KMO)

Antes, porém, de se realizar a análise de fatores devemos averiguar a adequabilidade da matriz de correlação. Para isso, destacamos o Critério de Kaiser-Meyer-Olkin (KMO). Este critério compara os valores dos coeficientes de correlação parcial com os coeficientes de correlação lineares simples verificando se a matriz de correlação inversa e a matriz de correlação diagonal se aproximam. Dessa forma, identificam se o modelo da análise de fatores utilizado é adequadamente ajustado aos dados. Este critério baseia-se em um índice dado por (MINGOTI, 2007):

$$KMO = \frac{\sum_{i \neq j} R_{ij}^2}{\sum_{i \neq j} R_{ij}^2 + \sum_{i \neq j} Q_{ij}^2} \qquad (1.2)$$

em que $R_{ij}$ e $Q_{ij}$ são, respectivamente, as correlações amostrais e as parciais das variáveis. $Q_{ij}$ representa a correlação entre a i-ésima e a j-ésima variáveis quando todas as outras $p - 2$ variáveis são consideradas como constantes.

A correlação amostral $(R_{ij})$ e a correlação parcial $(Q_{ij})$ entre a i-ésima e a j-ésima variáveis são definidas a seguir (RENCHER, 2008):

$$R_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

$$Q_{ij} = Q_{ij.rs\dots q} = \frac{\sigma_{ij.rs\dots q}}{\sigma_{ii.rs\dots q}\sigma_{ii.rs\dots q}}$$

em que $\sigma_{ij}$ é a covariância entre a i-ésima e a j-ésima variáveis, $\sigma_i$ é o desvio-padrão da i-ésima variável, $\sigma_j$ é o desvio-padrão da j-ésima variável, $\sigma_{ij.rs\dots q}$ é a covariância entre a i-ésima e a j-ésima variáveis sem o efeito das demais $p - 2$ variáveis $(r, s, \dots, q)$, $\sigma_{ii.rs\dots q}$ é a variância da i-ésima variável sem o efeito das demais e $\sigma_{jj.rs\dots q}$ é a variância da j-ésima variável sem o efeito das demais.

Este índice pode apresentar valores entre zero e um. Sendo que, quanto maior o valor do índice KMO, melhor será a adequabilidade da matriz de correlação. A classificação de adequabilidade da matriz, segundo Rencher (2002), é apresentada na Tabela 1 a seguir.

Tabela 1 – Classificação do Índice Kaiser-Meyer-Olkin (KMO)

| **KMO** | **Adequabilidade** |
|---|---|
| 0,9 – 1,0 | Excelente |
| 0,8 – 0,9 | Ótima |
| 0,7 – 0,8 | Boa |
| 0,6 – 0,7 | Regular |
| 0,5 – 0,6 | Ruim |
| Abaixo de 0,5 | Inadequada |

1.2.4.2 Teste de esfericidade de Bartlett

O teste de esfericidade de Bartlett tem por objetivo testar a hipótese nula de que a matriz de correlação da população ($\boldsymbol{\rho_{pxp}}$) é estatisticamente igual a uma matriz identidade ($\boldsymbol{I_{pxp}}$).

A estatística do teste é dada pela seguinte expressão (MINGOTI, 2007):

$$T = -\left[n - \left(\frac{2p+11}{6}\right)\right]\left[\sum_{i=1}^{p} ln\left(\widehat{\lambda_i}\right)\right] \qquad (1.3)$$

em que $\widehat{\lambda_i}$ é o i-ésimo autovalor da matriz de correlação amostral; $p$ é o número de variáveis e $n$ é o tamanho da amostra. Quando $n$ é suficientemente grande a estatística $T$ se aproxima da distribuição qui-quadrado com $\frac{1}{2}p(p-1)$ graus de liberdade. Assim, a hipótese de nulidade testada ($H_0$: $\boldsymbol{\rho_{p\times p} = I_{p\times p}}$) que admite ausência de associação linear entre as variáveis em estudo, deve ser rejeitada para que possamos aplicar a Análise de Fatores.

1.2.5 Modelo da análise de fatores

Analisando a correlação entre inúmeras variáveis e identificando a capacidade dessas variáveis serem associadas à um número menor de variáveis que não são diretamente observáveis, denominadas variáveis latentes, podemos construir o modelo fatorial ortogonal que é dado por:

$$Y_i - \mu_i = \sum_{i=1}^{p}\sum_{j=1}^{m} l_{ij}F_j + \varepsilon_i \qquad (1.4)$$

em que $p$ é o número de variáveis observáveis; $m$ é o número do fatores comum; $Y_i$ é a variável observável ($i = 1, \dots, p$); $\mu_i$ é a média da variável observável $Y_i$; $l_{ij}$ é a carga fatorial, ou seja, a correlação entre a variável observável $i$ e o fator $j$ ($j = 1, \dots, m$); $F_j$ é o j-ésimo fator comum e $\varepsilon_i$ é erro aleatório que está associado a i-ésima variável $Y_i$.

Esse modelo pode ser escrito, matricialmente, da seguinte forma (FERREIRA, 2011):

$$\boldsymbol{Y} - \boldsymbol{\mu} = \boldsymbol{\Gamma F} + \boldsymbol{\varepsilon} \qquad (1.5)$$

em que $\boldsymbol{Y} = [Y_1 \quad Y_2 \quad \dots \quad Y_p]'$ é o vetor de variável observável; $\boldsymbol{\mu} = [\mu_1 \quad \mu_2 \quad \dots \quad \mu_p]'$ é o vetor de médias de $\boldsymbol{Y}$; $\boldsymbol{\Gamma}$ é a matriz de cargas fatoriais com dimensão $p \times m$ de posto $m \leq p$; $\boldsymbol{F}$ é o vetor de fatores comuns com dimensão $m \times 1$ e $\boldsymbol{\varepsilon}$ é o vetor de erros aleatórios ou fatores específicos com dimensão $p \times 1$.

Para proceder à estimação dos parâmetros desse modelo, são necessárias algumas suposições, sendo elas:

i) O valor esperado do vetor de variáveis observáveis é igual a média populacional, ou seja, $E(\boldsymbol{Y}) = \boldsymbol{\mu}$;

ii) O valor esperado do vetor de fatores comuns é igual ao valor esperado do vetor de erros aleatórios, que é igual a um vetor cujos elementos são iguais a zero, ou seja, $E(\boldsymbol{F}) = E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$;

iii) $\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$;

iv) O vetor de fatores comuns e o vetor de erros aleatórios são não correlacionados $\mathrm{Cov}(\boldsymbol{F}, \boldsymbol{\varepsilon}) = 0$;

v) A variância dos fatores é igual a um e a covariância entre os fatores é igual a zero, ou seja, $\mathrm{Cov}(\boldsymbol{F}) = \boldsymbol{I_m}$ sendo $\boldsymbol{I_m}$ uma matriz identidade com dimensão $m \times m$;

vi) $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\psi}$ em que $\boldsymbol{\psi}$ é uma matriz diagonal dada por:

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p \end{bmatrix}$$

sendo $\psi_i > 0$ para todo $i$.

### 1.2.6 Método de extração dos fatores

Um dos métodos para a extração de fatores é a Análise de Componentes Principais. Este método maximiza a variância das variáveis originais, a partir de combinações lineares. A sucessiva busca por combinações lineares entre as variáveis que explique a variância, resulta

em fatores não correlacionados entre si, denominados fatores ortogonais (CORRAR, L. J. et al, 2014).

Utilizando-se desse método, podemos estimar as cargas fatoriais $\left(\hat{\boldsymbol{\Gamma}}\right)$ e as unicidades $\left(\widehat{\boldsymbol{\psi}}\right)$, utilizando a Decomposição Espectral da matriz de correlação ($\boldsymbol{R}$), ou seja, a decomposição formada por autovalores $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{i}}$ e autovetores $\hat{\boldsymbol{e}}_{\boldsymbol{i}}$. Assim, temos:

$$\hat{\boldsymbol{\Gamma}} = \prod_{i=1}^{m} \sqrt{\widehat{\boldsymbol{\lambda}}_{\boldsymbol{i}}}\, \hat{\boldsymbol{e}}_{\boldsymbol{i}} \qquad (1.6)$$

e

$$\widehat{\boldsymbol{\psi}}_{\boldsymbol{p}\times\boldsymbol{p}} = \boldsymbol{diag}\left(\boldsymbol{R}_{\boldsymbol{p}\times\boldsymbol{p}} - \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}^{\boldsymbol{T}}\right). \qquad (1.7)$$

Estimados $\hat{\boldsymbol{\Gamma}}$ e $\widehat{\boldsymbol{\psi}}$, podemos obter os valores atribuídos às novas variáveis não observáveis, que são os escores fatoriais $\hat{\mathbf{F}}_{\boldsymbol{j}}$, dados por (FERREIRA, 2011):

$$\hat{\mathbf{F}}_{\boldsymbol{j}} = \hat{\boldsymbol{\Gamma}}^{\boldsymbol{T}}\left(\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}^{\boldsymbol{T}} + \widehat{\boldsymbol{\psi}}\right)^{-\mathbf{1}}\left(\boldsymbol{Y}_{\boldsymbol{j}} - \overline{\boldsymbol{Y}}\right) \qquad (1.8)$$

em que $\boldsymbol{Y}_{\boldsymbol{j}}$ é o vetor referente aos valores assumidos pelo conjunto de variáveis do j-ésimo indivíduo e $\overline{\boldsymbol{Y}}$ é o vetor de médias referente as variáveis avaliadas.

A associação das variáveis originais em cada fator é feita por meio das cargas fatoriais $l_{ij}$, também chamadas de *loadings*. Quanto maior a carga fatorial, em módulo, mais relacionada a variável original estará com o respectivo fator. Uma medida importante para a análise de fatores é a comunalidade, que consiste na proporção de variação de cada variável que é explicada pelo fator a qual ela pertence. A comunalidade é dada pela seguinte expressão (FERREIRA, 2011):

$$c^2 = \sum_{j=1}^{m} l_{ij}^2 \qquad (1.9)$$

em que $l_{ij}$ é a carga fatorial da i-ésima variável e do j-ésimo fator.

### 1.2.7 Escolha do número de fatores

Após a verificação da adequabilidade da matriz de correlação, o próximo passo é determinar o número de fatores a serem utilizados na análise. Para isso, podemos aplicar os processos descritivos que fundamentam na obtenção dos autovalores da matriz de correlação, já descritos. Assim, o procedimento para escolha do número de fatores inicia-se ordenando os autovalores de forma crescente, e assim, verifica-se quais são os autovalores mais importantes

de acordo com o critério da análise da proporção da variância total explicada por $K$ fatores. Esta proporção é calculada por meio da divisão entre os $K$ primeiros autovalores e a soma de todos os autovalores, ou seja, $\frac{\sum_{k=1}^{K}\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ sendo $i = 1, \dots, K, \dots, p$. O objetivo é escolher o menor número de fatores que explica o máximo da variação presente nos dados originais amostrais. Sendo que o primeiro fator explica a maior porcentagem de variação, depois o segundo fator e, assim, sucessivamente. Logo, utiliza-se a quantidade de fatores necessárias para explicar uma porcentagem satisfatória desta variabilidade. Normalmente, a literatura recomenda explicar 70% ou mais da variação dos dados (FERREIRA, 2011).

### 1.2.8 Rotação *varimax*

Muitas vezes, alocar as variáveis originais em cada fator não é um procedimento fácil. Logo, é útil utilizar uma transformação ortogonal nos fatores a fim de buscar uma estrutura mais simples para a matriz de cargas fatoriais. Esta transformação tem o intuito de simplificar a identificação das variáveis latentes. Assim, podemos aplicar a rotação *varimax* que tem como objetivo buscar fatores com grandes variabilidades nas cargas fatoriais a partir de uma rotação ortogonal dos fatores. Para isso, utilizamos a maximização da variação ($v$) dos quadrados das cargas fatorais dada pela seguinte expressão (FERREIRA, 2011):

$$v = \frac{1}{p^2}\sum_{j=1}^{r}\left[p\sum_{i=1}^{p} z_{ij}^4 - \left(\sum_{i=1}^{p} z_{ij}^2\right)^2\right], \qquad (1.10)$$

em que $z_{ij} = \frac{l_{ij}}{\sqrt{\sum_{j=1}^{r} l_{ij}^2}}$, $l_{ij}$ é a carga fatorial da i-ésima variável e do j-ésimo fator e $r$ é o número de indivíduos observados.

## 1.3 RR-BLUP

O RR-BLUP (*Randon Regression Best Linear Unbiased Predictor*) é um método estatístico utilizado para a predição genômica e assume que os marcadores são variáveis de efeito aleatório. A estimação dos efeitos dos marcadores sobre o fenótipo é feita com base no seguinte modelo linear misto:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{W}\boldsymbol{m} + \boldsymbol{e} \qquad (1.11)$$

em que $\boldsymbol{y}$ é o vetor de fenótipos observados (com dimensão $r \times 1$ sendo $r$ igual ao número de indivíduos observados); $\mathbf{1}$ é um vetor com todos os elementos iguais a um e de mesma dimensão de $\boldsymbol{y}$; μ é a média geral da característica; $\boldsymbol{m}$ ($n \times 1$ com $n$ igual ao número de marcadores SNPs)

é o vetor dos efeitos aleatórios dos marcadores sendo $\boldsymbol{m}\sim\boldsymbol{N}(\boldsymbol{0},\boldsymbol{I}\sigma_m^2)$ e $\sigma_m^2$ a variância de cada marcador; $\boldsymbol{W}$ $(r \times n)$ é a matriz de incidência para $\boldsymbol{m}$, onde para um indivíduo diploide, contém os valores $0{,}1$ e $2$ para o número de alelos do marcador; $\boldsymbol{e}$ corresponde ao vetor de erros aleatórios sendo $\boldsymbol{e}\sim N(\boldsymbol{0},\boldsymbol{I}\sigma_e^2)$ $(r \times 1)$ e $\sigma_e^2$ a variância residual.

Assim, os efeitos genéticos dos marcadores podem ser preditos baseando nas seguintes equações de modelos mistos (RESENDE et al., 2007;2008;2012):

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'W} \\ \boldsymbol{W'X} & \boldsymbol{W'W} + \boldsymbol{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\boldsymbol{m}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{W'y} \end{bmatrix} \quad (1.12)$$

sendo pressuposto *a priori* que todos os marcadores SNPs explicam quantidades iguais da variância genética aditiva $(\sigma_g^2)$. Assim, o parâmetro de penalização (*shrinkage*) $\lambda = \frac{\sigma_e^2}{(\sigma_g^2/n_q)}$, em que $\sigma_e^2$ é a variância residual e $n_q$ é desconhecida *a priori*, podendo ser inferida a partir de $n_q = \sum_{i=1}^{n} 2p_i(1-p_i)$, em que $p_i$ é a frequência alélica do marcador $i$. Os componentes de variância são estimados via método da Máxima Verossimilhança Restrita (*Restricted maximum Likelihood* - REML)

Portanto, considerando o indivíduo $j$ $(j = 1, \dots, r)$, a predição do seu valor genético genômico (*Genomic Breeding Value* - GBV) é obtida pela seguinte expressão (RESENDE et al., 2012):

$$\hat{g}_j = \sum_i W_{ji}\, \hat{m}_i \qquad (1.13)$$

em que $W_{ji}$ é a incidência do i-ésimo marcador no j-ésimo indivíduo e $\hat{m}_i$ é o efeito estimado do i-ésimo marcador com $i = 1{,}2, \dots, n$. A expressão acima pode ser escrita matricialmente como sendo $\hat{\boldsymbol{g}} = \boldsymbol{W}\hat{\boldsymbol{m}}$.

## 1.4 G-BLUP

A predição genômica via BLUP (*Best Linear Unbiased Predictor*) genômico utiliza as relações de parentesco entre os indivíduos, provenientes das informações dos marcadores, a fim de estimar os méritos genéticos dos mesmos. Esse método possui um modelo simples e requer um baixo esforço computacional, mesmo diante de uma infinidade de marcadores, tornando-se na prática, uma abordagem tradicional para a predição genômica (GAO et al., 2012).

Segundo Resende et al. (2012), para que os efeitos genéticos aditivos sejam estimados, usamos um modelo linear misto, ajustado em um grupo de indivíduos genotipados e fenotipados e dado por:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zg} + \boldsymbol{e} \qquad (1.14)$$

em que $\boldsymbol{y}$ corresponde ao vetor de observações fenotípicas ($r \times 1$, onde $r$ é o número de indivíduos observados); $\mathbf{1}$ é um vetor com todos os elementos iguais a um e de mesma dimensão de $\boldsymbol{y}$; μ é a média geral da característica; $\boldsymbol{g}$ $(r \times 1)$ é o vetor dos efeitos genéticos aditivos individuais (aleatórios) sendo $\boldsymbol{g} \sim N(\mathbf{0}, \boldsymbol{G}\sigma_g^2)$ e $\sigma_g^2$ a variância genética aditiva e $\boldsymbol{G}$ a matriz de parentesco aditiva; $\boldsymbol{Z}$ $(r \times r)$ a matriz de incidência dos efeitos genéticos aditivos individuais e $\boldsymbol{e}$ corresponde ao vetor de erros aleatórios sendo $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$ $(r \times 1)$ e $\sigma_e^2$ a variância residual.

Ainda, segundo Resende et al. (2012), o modelo linear misto (1.14), pode ser descrito, utilizando informações fenotípicas e dos marcadores conduzindo ao método RR-BLUP, sendo:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{ZWm} + \boldsymbol{e} \qquad (1.15)$$

em que $\boldsymbol{m}$ ($n \times 1$ com $n$ igual ao número de marcadores SNPs) é o vetor dos efeitos aleatórios dos marcadores sendo $\boldsymbol{m} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{I}\sigma_m^2)$ e $\sigma_m^2$ a variância de cada marcador; $\boldsymbol{W}$ (r× $n$) é a matriz de incidência para $\boldsymbol{m}$, onde para um indivíduo diploide, contém os valores 0,1 e 2 para o número de alelos do marcador. Portanto, podemos fazer a analogia dos dois modelos por meio de $\boldsymbol{g} = \boldsymbol{Wm}$. Podemos perceber, então, que o método G-BLUP pode ser considerado uma reparametrização do RR-BLUP e vice-versa.

Assim, utilizar as equações de modelo misto para predizer $\boldsymbol{g}$ equivale a:

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{G^{-1}}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{g}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'Y} \\ \boldsymbol{Z'Y} \end{bmatrix} \qquad (1.16)$$

em que a variância dos erros aleatórios ($\sigma_e^2$ ) e dos efeitos genéticos aditivos ($\sigma_g^2$) são estimadas pelo método da Máxima Verossimilhança Restrita. A matriz de parentesco genômico dos efeitos aditivos é dada por (RESENDE et al., 2012):

$$\boldsymbol{G} = \frac{\boldsymbol{WW'}}{\sum_{i=1}^{n} \boldsymbol{2p_i q_i}} \qquad (1.17)$$

em que $n$ é número de marcadores SNPs; $p_i$ e $q_i$ são frequências alélicas de A e a, respectivamente.

Já as estimativas dos efeitos dos marcadores ($\hat{\boldsymbol{m}}$) podem ser obtidas por meio da estimação dos valores genéticos ($\hat{\boldsymbol{g}}$) via G-BLUP conforme o seguinte desenvolvimento $\hat{\boldsymbol{g}} = \boldsymbol{W}\hat{\boldsymbol{m}} \implies \boldsymbol{W}'\hat{\boldsymbol{g}} = \boldsymbol{W}'\boldsymbol{W}\hat{\boldsymbol{m}} \implies (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\hat{\boldsymbol{g}} = \hat{\boldsymbol{m}}$.

## REFERÊNCIAS BIBLIOGRÁFICAS

ALVARENGA, A.B. et al. Comparing Alternative Single-Step GBLUP Approaches and Training Population Designs for Genomic Evaluation of Crossbred Animals. **Frontiers in Genetics**, v.11, p.1-19, 2020.Disponível em: https://doi.org/10.3389/fgene.2020.00263. Acesso em: 2 nov.2020.

AMMIRAJU, J.S.S. et al. The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. **Genome Research**, v.16, n.1, p.140-147, 2006. Disponível em: http://www.genome.org/cgi/doi/10.1101/gr.3766306. Acesso em: 9 nov.2019.

ANTON, H. ; RORRES, C. **Álgebra linear com aplicações**.10.ed. Porto Alegre: Bookman. 768 p., 2012.

CATTELL, R. B. The scree test for the number of factors. **Multivariate Behavioral Research**, v. 1, p. 245-276, 1966.

CORRAR, L. J. et al. **Análise Multivariada:** para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas. 568p.,2014.

DE LOS CAMPOS, G. et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, p. 375-385, 2009.

FERREIRA, D. F. **Estatística Multivariada**. 2.ed. Lavras: UFLA. 675p., 2011.

GAO, H. et al. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. **Genetics Selection Evolution**, v. 44, p. 2-8, 2012.

KARAMAN, E.; LUND, M.S.; SU, G. Multi-trait single-step genomic prediction accounting for heterogeneous (co)variances over the genome. **Heredity,** v.124, p**.**274–287, 2020. Disponível em: https://doi.org/10.1038/s41437-019-0273-4. Acesso em: 29 nov.2020.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** Uma abordagem aplicada. Belo Horizonte: UFMG. 295p.,2007.

REIFSCHNEIDER, F. J. B. et al. **Uma pitada de biodiversidade na mesa dos brasileiros**. 1. ed. Brasília: DF. 156p., 2015.

RESENDE, M. D.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana(MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial.** Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291p. Disponível em: http://www.det.ufv.br/ppestbio/corpo_docente.php. Acesso em: 4 mar. 2019.

TABACHNICK, B.; FIDELL, L. **Using multivariate analysis**. Needham Heights: Allyn & Bacon, 2007.

TEIXEIRA, F. R. F. et al. Determinação de fatores em características de suínos. **Revista Brasileira de Biometria**, v.33, n.2, p.130-138, 2015.

TEIXEIRA, F.R.F. et al. Factor analysis applied to genome prediction for high-dimensional phenotypes in pigs. **Genetics and Molecular Research,** v. 15, p. 1-16, 2016.

WANG, J. et al. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. **Heredity,** v.121, p.648–662, 2018. Disponível em: https://doi.org/10.1038/s41437-018-0075-0. Acesso em: 15 nov. 2019.

WHITTAKER, J. C.; THOMPSON, R.; DENHAM, M. C. Marker-assisted selection using ridge regression. **Genetics Research**, v. 75, p. 249-252, 2000.

ZHAO, K. et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. **Nature Communication**, v.2, p.1-10, 2011. Disponível em: https://doi.org/10.1371/journal.pgen.1002221. Acesso em: 9 nov. 2019.

# CAPÍTULO 2

## ANÁLISE DE FATORES APLICADA A PREDIÇÃO GENÔMICA CONSIDERANDO SELEÇÃO DE MARCADORES EM *ORYZA SATIVA*.


### RESUMO

O objetivo do presente trabalho foi avaliar a combinação da análise de fatores e da seleção de marcadores na predição genômica via BLUP genômico (*Genomic best linear unbiased prediction* – G-BLUP) ao selecionar marcadores associados a grupos de características de interesse de arroz *Oryza Sativa*. Utilizou-se o conjunto de dados público de arroz que está disponível no site http://ricediversity.org/data/. O arquivo contém informações de 28 características fenotípicas e 36.901 marcadores SNPs (*Single Nucleotide Polymorphism*) de 413 indivíduos. Os resultados obtidos indicam que a aplicação da análise de fatores combinada à seleção de marcadores SNPs para a predição genômica se mostrou eficiente, visto que apresentaram valores semelhantes para capacidade preditiva encontrados considerando as análises individuais das variáveis (em ambas as análises obteve-se variação entre 0,6 a 0,8) e alta concordância (acima de 50% de concordância para todos grupo de marcadores) entre os indivíduos selecionados considerando o fator e as variáveis individuais.


**Palavras-chave:** Arroz. Seleção genômica. G-BLUP.


### ABSTRACT

The study objective was to evaluate the combination of factor analysis and marker selection in genomic prediction via genomic BLUP (Genomic best linear unbiased prediction - G-BLUP) when selecting markers associated with groups of traits Oryza Sativa. The public rice dataset was used and is available at http://ricediversity.org/data/. The dataset contains 28 phenotypic traits and 36,901 SNPs markers from 413 individuals. The results obtained indicate that the application of the factor analysis with the selection of SNPs (*Single Nucleotide Polymorphism*) for the genomic prediction was to be efficient. They presented similar values for predictive ability found considering the individual analyzes of the variables (in both analyzes, there was variation between 0.6 to 0.8) and high agreement (above 50% agreement for all group of markers) between the selected individuals considering the factor and the individual variables.


**Keywords:** Rice. Genomic selection. G-BLUP.

## 1. INTRODUÇÃO

O principal arroz cultivado mundialmente é da espécie *Oryza sativa,* popularmente conhecido como arroz asiático. Esta espécie possui uma alta capacidade de se adaptar a diversas regiões e à armazenagem de água por períodos longos (CONAB, 2015). Além disso, segundo Reifschneider et. al. (2015), este arroz tem grande importância econômica e social, sendo considerado uma das principais fontes alimentares, capaz de suprir metade da caloria energética necessária para milhões de pessoas. Assim, há a necessidade de desenvolvimento de novas linhas considerando melhorias na produtividade em relação as variedades existentes. Segundo o Conab (2020), estima-se que a safra brasileira 2020/21 de arroz será 2,1% menor que a da safra 2019/20. Em contrapartida, Spindel et al. (2015) menciona que o processo de desenvolvimento de novas linhagens de arroz é extremamente demorado quando se usa métodos convencionais de melhoramento e de seleção, levando cerca de dez anos em média para as variedades de elite serem desenvolvidas e identificadas.

Visando essas demandas, a melhoria de fenótipos desejáveis, por meio dos programas de melhoramento genético se torna imprescindível, pois possibilita o aumento da produtividade juntamente com outras características importantes para os produtores e consumidores sem que haja aumento das áreas agricultáveis. Para alcançar esse propósito, uma ferramenta empregada nesses programas é a Seleção Genômica Ampla (*Genome Wide Selection* - GWS), o qual utiliza as informações diretas do DNA. Meuwissen et al. (2001) propuseram esses procedimentos tendo como objetivo a estimação de ganhos genéticos de forma mais eficiente e rápida, em comparação com a tradicional, que se baseia apenas em dados fenotípicos. Assim, fundamentado na análise simultânea entre os fenótipos e um grande número de marcadores SNPs (*Single Nucleotide Polymorphism*) amplamente distribuídos no genoma, a GWS tem como fim a predição do mérito genético dos indivíduos e a seleção de indivíduos geneticamente superiores (RESENDE et al., 2012). No entanto, nem todos os marcadores SNPs estão em desequilíbrio de ligação (*Linkage disequilibrium* - LD) com os locos de características quantitativas (*Quantitative trait locus* - QTL), ou seja, nem todos os marcadores SNPs estão associados a uma determinada característica. Além disso, tendo em vista que os programas de melhoramento contemplam uma infinidade de características de interesse e enfrentam alguns desafios, como a alta dimensionalidade e multicolinearidade faz-se necessário o uso de métodos estatísticos multivariados e que consideram a seleção de marcadores.

O BLUP genômico (*Genomic best linear unbiased prediction* – G-BLUP) e este é amplamente aplicado na predição genômica e baseado na matriz de parentesco genômico, como

exemplos podemos citar os estudos realizados por Wang et al. (2018), Alvarenga et al. (2020) e Karaman et al. (2020). O método G-BLUP com seleção de marcadores (G-BLUP supervisionado) proposto por Resende et al. (2010, 2012) pode ser utilizado para reduzir a dimensionalidade e fazer com que a matriz de parentesco genômica seja construída contendo apenas os marcadores relevantes para determinada característica (Resende et al., 2012). Isso é importante pois dois indivíduos são geneticamente idênticos, para uma característica, se carregam o mesmo genótipo nos QTLs causais ou genes, ou indiretamente os mesmos marcadores em LD com eles. Além disso, estudos usando esse método com dados simulados (Zhang et al. 2010; Zhang et al. 2011) e em pinheiros (Resende Jr et al., 2012) tem mostrado resultados equivalentes aos obtidos por métodos bayesianos. Porém, segundo Resende Jr et al. (2012), a vantagem do G-BLUP supervisionado é a menor demanda computacional comparado aos métodos bayesianos.

Já a análise de fatores (AF) realizada em dados fenotípicos tem sido utilizada na recomendação de genótipos, como por exemplo genótipos de *Coffea arábica* (BARBOSA et al., 2019) e de frangos (PAIVA et al., 2020). Já Teixeira et al. (2015, 2016), considerando a predição genômica, propuseram a aplicação da análise fatores em características fenotípicas de suínos a fim de reduzir o número de variáveis resposta para a GWS. Esta redução se deve a transformação das variáveis fenotípicas originais em variáveis latentes, denominadas fatores, sendo assim, um grupo de fenótipos estaria representado por um fator e este seria usado para estimar o mérito genético dos indivíduos. Dessa forma, a seleção dos indivíduos geneticamente superiores por meio do fator contemplaria mais de uma característica. Apesar da AF ter sido usada anteriormente na predição genômica, ela ainda não foi empregada conjuntamente a metodologias de seleção de marcadores.

Desse modo, o objetivo desse estudo será aplicar a análise de fatores associados a grupos de características de interesse (fator) e avaliar sua eficiência na predição genômica via G-BLUP considerando seleção de marcadores. Estes resultados serão comparados aos da análise individual de cada característica. Para tanto, utilizou-se 28 características fenotípicas de arroz e 36.901 marcadores SNPs associados a 413 indivíduos.

## 2. MATERIAIS E MÉTODOS

### 2.1 Conjuntos de dados

O conjunto de dados utilizado é referente ao arroz asiático *Oryza sativa*. Os dados são públicos e fazem parte de dois projetos, o Projeto OryzaSNP e o Projeto OMAP (Ammiraju et al., 2006; Zhao et al., 2011) e estão disponíveis no site http://www.ricediversity.org/data/sets/44kgwas/. O arquivo contém 413 indivíduos, 28 variáveis fenotípicas e 36.901 marcadores SNPs após o controle de qualidade considerando $call\ rate < 70\%$ e baixa frequência do alelo mais raro (*Minor Allele Frequency* - MAF) < 1%.

As variáveis que foram analisadas estão descritas a seguir: i) Tempo de floração no Arkansas; ii) Tempo de floração no Aberdeen; iii) Razão entre o tempo de floração do Arkansas e do Aberdeen; iv) Razão entre o tempo de floração do Faridpur e do Aberdeen; v) Ângulo da base do colmo principal; vi) Pubescência foliar; vii) Comprimento da folha da bandeira; viii) Presença de arista; ix) Número de panículas por planta; x) Altura da planta; xi) Comprimento da panícula; xii) Número de ramos da panícula primária; xiii) Número de sementes por panícula; xiv) Floretes por panícula; xv) Comprimento da semente; xvi) Largura da semente; xvii) Volume da semente; xviii) Área da Superfície da semente; xix) Comprimento da semente de arroz integral; xx) Largura da semente de arroz integral; xxi) Superfície de arroz integral; xxii) Volume de arroz integral; xxiii) Razão entre o comprimento da semente e a largura; xxiv) Razão entre o comprimento e a largura do arroz integral; xxv) Cor da semente; xxvi) Cor do pericarpo; xxvii) Tempo de floração no Arkansas considerando o ano 07; xxviii) Tempo de floração no Arkansas considerando o ano 06.

### 2.2 Análise de fatores

A análise de fatores foi aplicada na matriz de fenótipos afim de reduzir a dimensionalidade do conjunto de variáveis originais por meio das variáveis latentes, denominadas fatores. No entanto, primeiramente, realizou-se uma análise da adequabilidade da matriz de correlação a partir de dois métodos estatísticos: Critério de Kaiser-Meyer-Olkin (KMO) e o teste de esfericidade de Barlett.

O Critério de Kaiser-Meyer-Olkin (KMO), foi utilizado para verificar se o modelo da análise de fatores utilizado era adequadamente ajustado aos dados. Quanto mais próximo de 1

for o índice, melhor a amostra adequou-se à aplicação da análise de fatores. O índice dado por (MINGOTI, 2007):

$$KMO = \frac{\sum_{i \neq j} R_{ij}^2}{\sum_{i \neq j} R_{ij}^2 + \sum_{i \neq j} Q_{ij}^2} \qquad (2.1)$$

em que $R_{ij}$ e $Q_{ij}$ são, respectivamente, as correlações amostrais e as parciais das variáveis. Segundo Rencher (2002), 0,70 é o valor mínimo para considerar boa adequabilidade da matriz de correlação. Já o teste de esfericidade de Barlett, foi utilizado, com o nível de 5% de significância, para averiguar a ausência de associação linear entre as variáveis em estudo (variáveis não correlacionadas). Assim, a hipótese de nulidade ($H_0$: $\boldsymbol{\rho_{28\times28} = I_{28\times28}}$), deveria ser rejeitada ($p - valor < 0{,}05$), para que pudéssemos aplicar a análise de fatores.

Confirmada a adequabilidade da matriz de correlação, foi utilizado a análise de fatores de forma exploratória. Assim, foi possível construir o modelo fatorial ortogonal dado por (FERREIRA, 2011):

$$\boldsymbol{Y_i - \mu_i = \sum_{i=1}^{p} \sum_{j=1}^{m} l_{ij} F_j + \varepsilon_i} \qquad (2.2)$$

em que $p = 28$ é o número de variáveis fenotípicas originais observáveis; $m \leq 28$ é o número do fatores comum; $Y_i$ é a variável fenotípica observável; $\mu_i$ é a média da variável fenotípica observável $Y_i$; $l_{ij}$ é a carga fatorial, ou seja, a correlação entre a variável fenotípica observável $i = 1,2, \dots p$ e o fator $j = 1,2, \dots, m$; $F_j$ é o j-ésimo fator comum e $\varepsilon_i$ é erro aleatório que está associado a i-ésima variável fenotípica $Y_i$.

A matriz de cargas fatoriais $\left(\hat{\Gamma}_{28\times m}\right)$ e a matriz de unicidades $\left(\hat{\psi}_{28\times 28}\right)$ são estimadas utilizando a Decomposição Espectral da matriz de correlação entre as variáveis fenotípicas ($R_{28\times28}$). Assim, temos:

$$\boldsymbol{\hat{\Gamma} = \prod_{i=1}^{m} \sqrt{\hat{\lambda}_i}\, \hat{e}_i} \qquad e \qquad \boldsymbol{\hat{\psi}_{28\times28} = diag\left(R_{28\times28} - \hat{\Gamma}\hat{\Gamma}^T\right)} \qquad (2.3)$$

sendo $\hat{\lambda}_i$ o i-ésimo autovalor e $\hat{e}_i$ o i-ésimo autovetor. Posteriormente, os escores fatoriais $\hat{\mathrm{F}}_j$ foram estimados por meio de (FERREIRA, 2011):

$$\boldsymbol{\hat{F}_j = \hat{\Gamma}^T \left(\hat{\Gamma}\hat{\Gamma}^T + \hat{\psi}\right)^{-1} \left(Y_j - \overline{Y}\right)} \qquad (2.4)$$

em que $Y_j$ é o vetor referente aos valores assumidos pelo conjunto de variáveis fenotípicas do j-ésimo ($j = 1,2, \dots, 413$) indivíduo e $\overline{Y}$ é o vetor de médias referente as 28 variáveis fenotípicas avaliadas. Desse modo, foi determinado o número de fatores que explicassem 70% ou mais da

variabilidade dos dados, que segundo Ferreira (2011), é a porcentagem de explicação satisfatória para a redução da dimensionalidade dos dados sem que haja a perda de informações relevantes para o estudo. Além disso, considerou-se o número de fatores que melhor alocava as variáveis originais.

A destinação das variáveis fenotípicas originais em cada fator foi feita por meio das cargas fatoriais $l_{ij}$, também chamadas de *loadings*. Para facilitar essa locação utilizamos a rotação por meio do procedimento *varimax*. Portanto, quanto maior o valor, em módulo, das cargas, mais relacionada a variável fenotípica original é com o respectivo fator. Em seguida, analisamos a proporção de cada variável explicada pelo fator a qual ela pertence e a proporção explicada pelo erro aleatório. Para isso, utilizou-se a comunalidade dada pela seguinte expressão (FERREIRA, 2011):

$$c^2 = \sum_{j=1}^{m} l_{ij}^2 \qquad (2.5)$$

em que $l_{ij}$ é a carga fatorial da i-ésima variável e do j-ésimo fator. Segundo Figueiredo Filho (2010) os valores das comunalidades devem ser superiores a 0,50.

## 2.3 Predição Genômica

Posteriormente, esses fatores foram utilizados como variáveis dependentes nos modelos mistos genômicos. O modelo linear misto associado ao método G-BLUP comtempla um grupo de indivíduos genotipados e fenotipados visando estimar os efeitos genéticos aditivos dos indivíduos (RESENDE et al., 2012) conforme a seguir:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zg} + \boldsymbol{e} \qquad (2.6)$$

em que $\boldsymbol{y}$ corresponde ao vetor de observações fenotípicas $(413 \times 1)$; $\mathbf{1}$ é um vetor com todos os elementos iguais a um e de mesma dimensão de $\boldsymbol{y}$; μ é a média geral da característica; $\boldsymbol{g}$ $(413 \times 1)$ é o vetor dos efeitos genéticos aditivos individuais (aleatórios) sendo $\boldsymbol{g} \sim N\left(\mathbf{0}, \boldsymbol{G}\sigma_g^2\right)$ e $\sigma_g^2$ a variância genética aditiva e $\boldsymbol{G}$ a matriz de parentesco aditiva; $\boldsymbol{Z}$ $(413 \times 413)$ a matriz de incidência dos efeitos genéticos aditivos individuais e $\boldsymbol{e}$ corresponde ao vetor de erros aleatórios sendo $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$ $(413 \times 1)$ e $\sigma_e^2$ a variância residual.

Assim, utilizar as equações de modelo misto para predizer $\boldsymbol{g}$ equivale a:

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \boldsymbol{G^{-1}}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{g}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'Y} \\ \boldsymbol{Z'Y} \end{bmatrix} \qquad (2.7)$$

em que a variância dos erros aleatórios ($\sigma_e^2$ ) e dos efeitos genéticos aditivos individuais ($\sigma_g^2$) são estimadas pelo método da Máxima Verossimilhança Restrita. A matriz de parentesco genômico dos efeitos aditivos é dada por (RESENDE et al., 2012):

$$\boldsymbol{G} = \frac{\boldsymbol{WW'}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{36901}} 2p_i q_i} \qquad (2.8)$$

em que $\boldsymbol{W}$ $(413 \times 36901)$ é a matriz de incidência para $\boldsymbol{m}$**,** onde para um indivíduo diploide, contém os valores 0,1 e 2 para o número de alelos do marcador, $p_i$ e $q_i$ são frequências alélicas de A e a, respectivamente. As estimativas dos efeitos dos marcadores ($\widehat{\boldsymbol{m}}$) foram obtidas por meio de $\widehat{\boldsymbol{m}} = (\boldsymbol{W'W})^{-1}\boldsymbol{W'}\widehat{\boldsymbol{g}}$**.**

Dessa forma, medimos o grau de correspondência entre o fenótipo e o genótipo a partir do valor da herdabilidade ($h^2$) de cada variável fenotípica dada por (RESENDE et al., 2012):

$$h^2 = \frac{\sigma_g^2}{\sigma_e^2}. \qquad (2.9)$$

## 2.4 Seleção de marcadores

A seleção de marcadores foi empregada a partir do método G-BLUP supervisionado, o qual é semelhante ao método RR-BLUP_B (RESENDE et al., 2010; RESENDE JR et al., 2012). O método G-BLUP é aplicado a cada fenótipo (ou a cada fator) considerando todos os 36.901 marcadores. Após a estimação dos efeitos, os módulos dos mesmos foram ordenados e agrupados considerando os 100, 1.000, 2.000, ..., 36.901 marcadores com maior magnitude de efeito. A re-estimação dos efeitos desses grupos demarcadores foi utilizando novamente o método G-BLUP por meio do procedimento de validação cruzada. Dessa forma, selecionou-se os marcadores de maior efeito absoluto para o fator (denominaremos de análise conjunta) e para um único fenótipo (denominaremos de análise individual).

O processo da validação cruzada *k-fold* foi aplicado para avaliarmos a capacidade preditiva associada as análises utilizando os fenótipos diretamente na predição genômica e as análises utilizando os fatores. A partir de 413 indivíduos, escolhemos o número de *folds* igual a $k = 7$, que é um divisor do número total de indivíduos. Deste modo, foi considerando 7 subconjuntos da população contendo 59 indivíduos cada. Na primeira iteração, utilizou-se um subconjunto como população de validação e os restantes foram utilizados como população de estimação. Esse processo foi repetido 7 vezes, de forma que cada um dos subconjuntos fosse utilizado uma única vez na validação do modelo. A cada iteração, na população de estimação

foram estimados os efeitos de marcadores que foram usados para predizer os valores genéticos da população de validação.

2.5 Avaliação do método

A eficiência da análise de fatores combinada com a seleção de marcadores foi mensurada por meio de duas medidas: i) capacidade preditiva que é a correlação entre o valor fenotípico ($\boldsymbol{y}$) e o valor genômico estimado via fator ou via análise individual ($\widehat{\boldsymbol{g}}$), ou seja, $Cor(\boldsymbol{y}, \widehat{\boldsymbol{g}})$; ii) coeficiente de regressão entre $\boldsymbol{y}$ e $\widehat{\boldsymbol{g}}$, $b_{y\widehat{g}} = \frac{Cov(\boldsymbol{y},\widehat{\boldsymbol{g}})}{Var(\widehat{\boldsymbol{g}})}$, sendo $Cov$ o operador da covariância e $Var$ da variância. O esperado é que esse coeficiente de regressão seja próximo de um, para que as avaliações do fator e do fenótipo sejam não viesadas. Ademais, verificou-se a coincidência dos marcadores selecionados por meio do fator e do fenótipo individualmente.

Todos os métodos estatísticos foram realizados através do *software* livre R (R Development Core Team, 2019) versão 3.5.1, utilizando os pacotes: "psych"( REVELLE WILLIAM, 2019); "rrBLUP" (ENDELMAN J. B.,2011) e "matrixStats" (BENGTSSON HENRIK, 2020).

## 3. RESULTADOS E DISCUSSÃO

De acordo com o Critério de Kaiser-Meyer-Olkin (KMO), foi constatado um índice de 0,705, que segundo Rencher (2002) é um bom valor para a adequabilidade da matriz de correlação. Já o teste de esfericidade de Bartlett, apresentou significância estatística (p-valor < 0,05), admitindo associação linear entre as variáveis fenotípicas. Logo, verificou-se que há adequabilidade dos dados para a aplicação da análise de fatores.

Com as 28 variáveis fenotípicas, constatou-se (Tabela 2) que 5 fatores captaram 72,56% da variabilidade dos dados, 6 fatores que captaram 77,42%, 7 fatores captaram 81,12% e 11 fatores captaram 91,85%. Assim, foi possível verificar que o número de fatores que melhor representava as características fenotípicas foi igual a 6 fatores pois a partir deles conseguimos formar melhor grupos de variáveis associadas a Tempo de floração, Morfologia, Produtividade e Morfologia das sementes.

Tabela 2 – Características do arroz Oryza sativa, siglas, número de fatores, alocação de cada fator e a variação explicada por cada análise de fatores.

| **Características** | **Siglas** | **Número de Fatores** | | | |
|---|---|---|---|---|---|
| | | **5** | **6** | **7** | **11** |
| Tempo de floração no Arkansas | T.F.Ark | 4 | 6 | 5 | 5 |
| Tempo de floração no Aberdeen | T.F.Ab | 4 | 4 | 4 | 4 |
| Razão entre o tempo de floração do Arkansas e do Aberdeen | R.T.F.Ark.Ab | 4 | 4 | 4 | 4 |
| Razão entre o tempo de floração do Faridpur e do Aberdeen | R.T.F.Far.Ab | 4 | 4 | 4 | 4 |
| Tempo de floração no Arkansas considerando o ano 07 | T.F.Ark.07 | 5 | 6 | 5 | 5 |
| Tempo de floração no Arkansas considerando o ano 06 | T.F.Ark.06 | 3 | 4 | 4 | 5 |
| Ângulo da base do colmo principal | A.B.C.P | 5 | 5 | 6 | 7 |
| Pubescência foliar | P.F | 5 | 5 | 6 | 11 |
| Comprimento da folha da bandeira | C.F.B | 3 | 5 | 3 | 6 |
| Presença arista | P.A | 3 | 3 | 7 | 9 |
| Número de panículas por planta | N.P.P | 1 | 3 | 3 | 7 |
| Altura da planta | A.P | 5 | 5 | 6 | 6 |
| Comprimento da panícula | C.P | 3 | 5 | 6 | 6 |
| Número de ramos da panícula primária | N.R.P.P | 3 | 3 | 3 | 3 |
| Número de sementes por panícula | N.S.P | 3 | 3 | 3 | 3 |
| Floretes por panícula | FL.P | 3 | 3 | 3 | 3 |
| Comprimento da semente | C.S | 2 | 2 | 2 | 2 |
| Largura da semente | L.S | 1 | 1 | 1 | 1 |
| Volume da semente | V.S | 1 | 1 | 1 | 1 |
| Área da Superfície da semente | A.S.S | 1 | 1 | 1 | 1 |
| Comprimento da semente de arroz integral | C.S.A.I | 2 | 2 | 2 | 2 |
| Largura da semente de arroz integral | L.S.A.I | 1 | 1 | 1 | 1 |
| Superfície de arroz integral | S.A.I | 1 | 1 | 1 | 1 |
| Volume de arroz integral | V.A.I | 1 | 1 | 1 | 1 |
| Razão entre o comprimento da semente e a largura | R.C.S.L | 2 | 2 | 2 | 2 |
| Razão entre o comprimento e a largura do arroz integral | R.C.L.A.I | 2 | 2 | 2 | 2 |
| Cor da semente | Cor.S | 5 | 6 | 7 | 10 |
| Cor do pericarpo | Cor.P | 5 | 6 | 5 | 8 |
| **Variância Explicada Acumulada (%)** | | 72.56 | 77.42 | 81.12 | 91.85 |

As comunalidades foram encontradas (Tabela 3) e apenas três variáveis não apresentaram comunalidade maior que 0,50, são elas: Presença arista (0,07), Pubescência foliar (0,10) e Cor da semente (0,20). Desse modo, desconsiderou-se estas variáveis da predição genômica, uma vez que estas variáveis não possuem alto poder de explicação pelo fator.

Os seis fatores apresentaram interpretação prática. No entanto, perceba que algumas variáveis, mesmo com comunalidade acima de 0,50 apresentam *loadings* negativos, ou seja, correlações negativas com o fator. São elas, Número de panículas por planta e as Razões do tempo de Florescimento. Segundo Ramão et al. (2019) e Dalchiavon et al. (2012) o número de panículas por plantas demostrou correlação positiva com a produtividade de grãos. Dessa forma, não é eficiente fazer a seleção dos indivíduos para estas duas variáveis por meio do fator.

Tabela 3 – Fatores e suas respectivas variáveis associadas, *loadings* para cada variável em relação a todos os fatores, a variação explicada de cada um dos fatores e as comunalidades ($c^2$).

(Continua)

| **Fator** | **Variáveis** | **F1** | **F2** | **F3** | **F4** | **F5** | **F6** | $c^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | L. S | **0,77** | -0,62 | -0,07 | -0,01 | -0,07 | -0,02 | 0,98 |
| | V. S. | **0,97** | -0,16 | -0,08 | -0,02 | -0,07 | -0,02 | 0,98 |
| | A.S.S | **0,95** | 0,27 | -0,07 | -0,02 | -0,05 | -0,02 | 0,98 |
| | L.S.A.I | **0,72** | -0,67 | -0,07 | -0,02 | -0,08 | -0,02 | 0,98 |
| | S.A.I | **0,95** | 0,26 | -0,04 | -0,03 | -0,05 | -0,01 | 0,97 |
| | V.A.I | **0,97** | -0,19 | -0,06 | -0,03 | -0,06 | -0,02 | 0,98 |
| 2 | C.S | 0,37 | **0,92** | -0,02 | -0,01 | 0,01 | -0,02 | 0,98 |
| | C.S.A.I | 0,31 | **0,94** | 0,03 | 0,01 | -0,01 | -0,01 | 0,98 |
| | R.C.S.L | -0,30 | **0,94** | 0,04 | 0,01 | 0,03 | -0,03 | 0,98 |
| | R.C.L.A.I | -0,26 | **0,95** | 0,06 | 0,03 | 0,01 | -0,02 | 0,98 |
| 3 | P.A | -0,06 | -0,11 | **-0,19** | -0,07 | 0,06 | -0,10 | 0,07 |
| | N.P.P | -0,45 | -0,12 | **-0,54** | -0,02 | 0,38 | 0,08 | 0,67 |
| | N.R.P.P | -0,06 | 0,03 | **0,79** | 0,26 | -0,10 | -0,15 | 0,73 |
| | N.S.P | -0,16 | -0,10 | **0,86** | 0,07 | 0,11 | -0,06 | 0,79 |
| | FL.P | -0,16 | -0,01 | **0,91** | 0,16 | 0,10 | -0,10 | 0,90 |
| 4 | T.F.Ab | -0,07 | -0,08 | 0,15 | **0,88** | 0,00 | -0,01 | 0,80 |
| | R.T.F.Ark.Ab | -0,02 | 0,08 | -0,13 | **-0,83** | 0,05 | 0,34 | 0,84 |
| | R.T.F.Far.Ab | -0,03 | -0,11 | -0,15 | **-0,76** | -0,07 | -0,10 | 0,62 |
| | T.F.Ark.06 | -0,10 | 0,15 | 0,50 | **0,53** | 0,30 | 0,20 | 0,70 |
| 5 | A.B.C.P | -0,26 | -0,15 | -0,31 | -0,30 | **0,56** | 0,14 | 0,61 |
| | P.F | 0,09 | -0,27 | -0,26 | 0,03 | **0,49** | -0,12 | 0,10 |
| | C.F.B | 0,01 | 0,12 | 0,46 | 0,24 | **0,51** | -0,01 | 0,55 |
| | A.P | -0,07 | 0,10 | 0,09 | 0,07 | **0,81** | 0,17 | 0,71 |
| | C.P | -0,20 | 0,39 | 0,38 | 0,13 | **0,62** | 0,06 | 0,74 |

Tabela 3 – Fatores e suas respectivas variáveis associadas, *loadings* para cada variável em relação a todos os fatores, a variação explicada de cada um dos fatores e as comunalidades ($c^2$).

(Continuação)

| Fator | Variáveis | F1 | F2 | F3 | F4 | F5 | F6 | $c^2$ |
|---|---|---|---|---|---|---|---|---|
| 6 | T.F.Ark | -0,16 | 0,08 | 0,43 | 0,52 | 0,32 | **0,56** | 0,90 |
| | T.F.Ark.07 | -0,17 | 0,05 | 0,28 | 0,39 | 0,28 | **0,74** | 0,89 |
| | Cor.S | 0,03 | -0,02 | -0,12 | -0,08 | -0,07 | **0,41** | 0,20 |
| | Cor.P | -0,02 | -0,07 | -0,15 | -0,12 | 0,13 | **0,83** | 0,75 |
| **Variância Explicada Acumulada (%)** | | 28,29 | 42,07 | 57,58 | 66,79 | 72,56 | 77,42 | |

Após a exclusão das variáveis devido a comunalidade baixa e correlação negativa com o fator, o KMO foi igual a 0,700, o teste de esfericidade de Bartlett também apresentou significância estatística (p-valor < 0,05) e todas as comunalidades fora, acima de 0,65 (Tabela 4). As variáveis associadas a cada um dos seis fatores não sofreram alterações e os fatores são descritos a seguir, por ordem de maior explicação dos dados.

O primeiro fator foi formado por variáveis relacionadas a característica de morfologia das sementes (Tabela 4); o segundo fator também associou apenas variáveis relacionadas a característica de morfologia das sementes; o terceiro fator agrupou características de produtividade; o quarto fator agrupou as variáveis associadas a tempo de floração. Estes resultados indicam que as variáveis relacionadas a morfologia das sementes, produtividade e tempo de floração estão altamente correlacionadas entre si. Além disso, observa-se pelos valores de *loadings* que estas variáveis apresentam altos valores positivos de correlação. Dessa forma, quanto maior o valor destas variáveis, maior será o valor dos escores das novas variáveis "Morfologia das sementes", "Produtividade" e "Tempo de Floração". Portanto, ao analisar os escores do fator, analisa-se conjuntamente todas as variáveis pertencentes a ele.

Já o quinto fator associou as variáveis: ângulo da base do colmo principal (A.B.C.P); pubescência foliar (P.F); comprimento da folha da bandeira (C.F.B) relacionadas a morfologia juntamente com as variáveis: altura da planta (A.P) e comprimento da panícula (C.P) relacionadas a produtividade. O sexto fator agrupou as variáveis: tempo de floração no Arkansas (T.F.Ark), tempo de floração no Arkansas considerando 07 anos (T.F.Ark.07) associados a tempo de floração com a variável cor do pericarpo (Cor.P).

Tabela 4 – Fatores suas respectivas variáveis associadas e *loadings*.

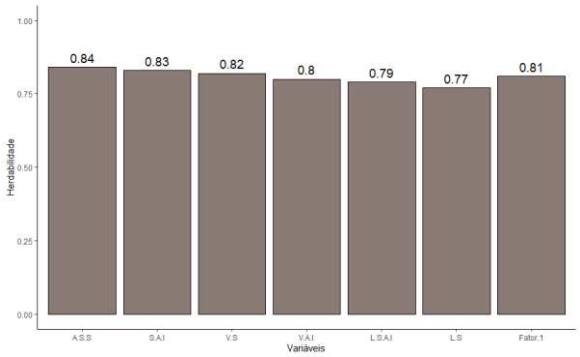
| Fator | Variáveis | F1 | F2 | F3 | F4 | F5 | F6 | $c^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | L. S | **0,77** | -0,61 | -0,06 | -0,03 | -0,12 | -0,04 | 0,98 |
| | V. S. | **0,97** | -0,14 | -0,09 | -0,03 | -0,09 | -0,05 | 0,99 |
| | A.S.S | **0,95** | 0,28 | -0,09 | -0,03 | -0,05 | -0,05 | 0,99 |
| | L.S.A.I | **0,72** | -0,66 | -0,07 | -0,04 | -0,12 | -0,04 | 0,98 |
| | S.A.I | **0,95** | 0,27 | -0,06 | -0,04 | -0,04 | -0,03 | 0,98 |
| | V.A.I | **0,97** | -0,18 | -0,06 | -0,04 | -0,04 | -0,03 | 0,98 |
| 2 | C.S | 0,36 | **0,92** | -0,05 | -0,01 | 0,06 | -0,03 | 0,98 |
| | C.S.A.I | 0,30 | **0,94** | 0,00 | 0,01 | 0,04 | -0,02 | 0,99 |
| | R.C.S.L | -0,31 | **0,94** | 0,02 | 0,02 | 0,08 | -0,01 | 0,99 |
| | R.C.L.A.I | -0,27 | **0,94** | 0,02 | 0,02 | 0,08 | -0,01 | 0,99 |
| 3 | N.R.P.P | -0,03 | 0,06 | **0,81** | 0,31 | -0,03 | -0,13 | 0,78 |
| | N.S.P | -0,12 | -0,07 | **0,92** | 0,03 | 0,12 | 0,08 | 0,89 |
| | FL.P | -0,12 | -0,02 | **0,93** | 0,18 | 0,16 | -0,02 | 0,94 |
| 4 | T.F.Ab | -0,07 | -0,09 | 0,11 | **0,84** | 0,02 | -0,01 | 0,74 |
| | T.F.Ark.06 | -0,09 | 0,13 | 0,39 | **0,65** | 0,36 | 0,17 | 0,77 |
| 5 | A.B.C.P | -0,24 | -0,19 | -0,23 | -0,48 | **0,50** | 0,23 | 0,68 |
| | C.F.B | 0,02 | 0,07 | 0,29 | 0,40 | **0,63** | -0,11 | 0,65 |
| | A.P | -0,07 | 0,04 | 0,00 | 0,02 | **0,86** | 0,21 | 0,78 |
| | C.P | -0,19 | 0,35 | 0,28 | 0,17 | **0,70** | 0,07 | 0,76 |
| 6 | T.F.Ark | -0,13 | 0,07 | 0,31 | 0,64 | 0,33 | **0,56** | 0,94 |
| | T.F.Ark.07 | -0,14 | 0,04 | 0,18 | 0,48 | 0,25 | **0,76** | 0,93 |
| | Cor.P | 0,00 | -0,06 | -0,18 | -0,13 | 0,05 | **0,90** | 0,86 |
| **Variância Explicada Acumulada (%)** | | 31,37 | 51,10 | 68,34 | 78,59 | 84,38 | 88,83 | |

Após a determinação dos 6 fatores, utilizou-se os métodos G-BLUP/REML para estimar a herdabilidade genômica das variáveis fenotípicas (Figura 1) e dos fatores (Figura 2). Tendo como base os estudos realizados por Akinwale et al. (2011) que classificou a magnitude da herdabilidade em uma população de arroz asiático *Oryza sativa* como: alta (> 70%); média (31 -70%) e baixa (<30%). Foi verificado resultados semelhantes considerando o fator e as variáveis individuais. Constatando-se alta herdabilidade para o Fator 1 (0,81); Fator 2 (0,80); Fator 5 (0,71); Fator 6 (0,88) e para todas as variáveis do primeiro e segundo fator, com variação de 0,77 à 0,84 para o primeiro fator; 0,78 à 0,82 para o segundo fator. Herdabilidade média para

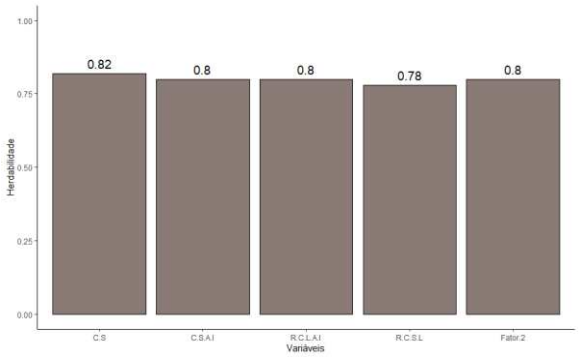
o Fator 3 (0,56); Fator 4 (0,53) e para as variáveis do terceiro e quarto fator com variações de 0,46 a 0,67 e 0,58 a 0,61, respectivamente. As variáveis do quinto e sexto fator, demostraram herdabilidade média e alta, sendo constatado estimativas de 0,61 à 0,71 e 0,66 à 0,99, respectivamente. Desse modo, apurou-se média e alta correspondência entre fenótipo e genótipo, indicando sucesso de seleção genômica, como reportaram Bisne et al (2009). Guo et al. (2014), ao aplicar o método G-BLUP no mesmo banco de dados, obtiveram estimativas de herdabilidade genômica semelhantes aos encontrados.

Figura 1 – Gráficos da herdabilidade dos fatores e de suas variáveis associadas.

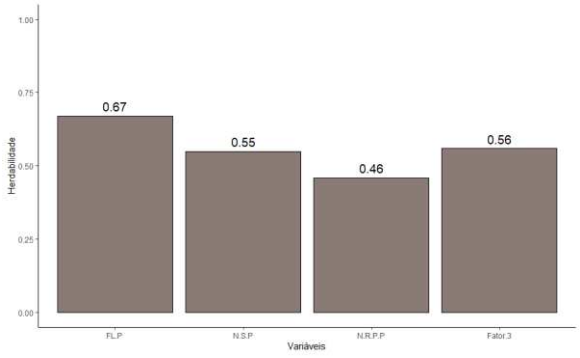
a) Fator 1



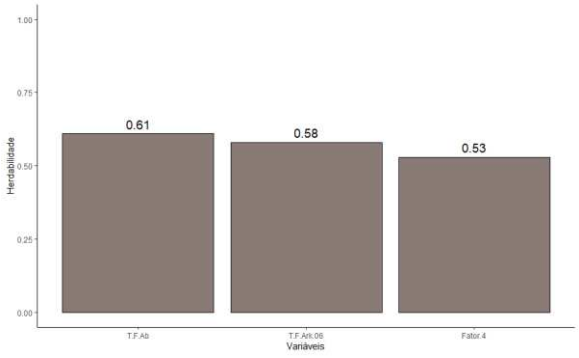

b) Fator 2



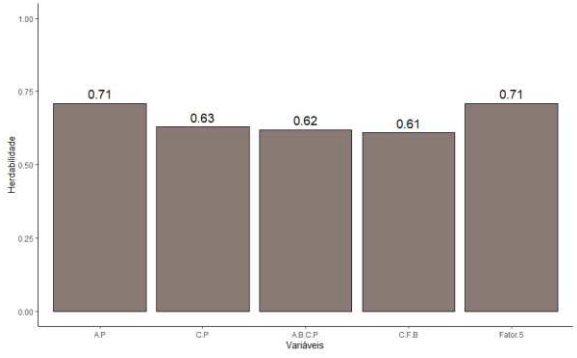

c) Fator 3



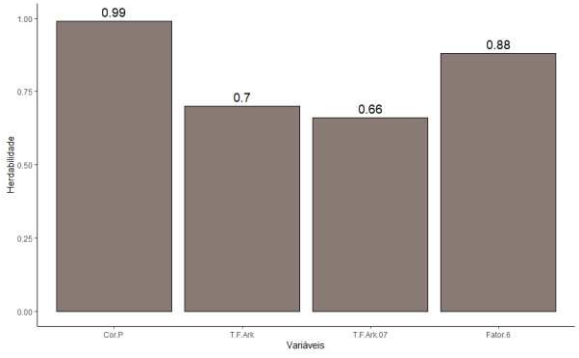

d) Fator 4




e) Fator 5



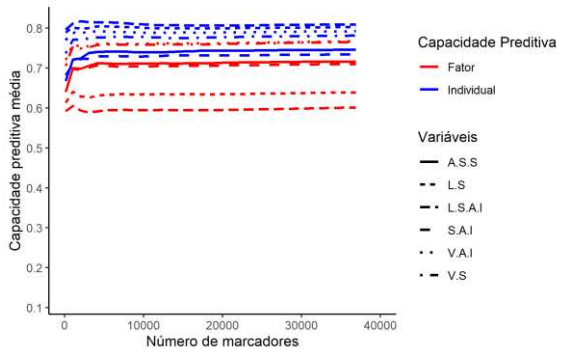

f) Fator 6



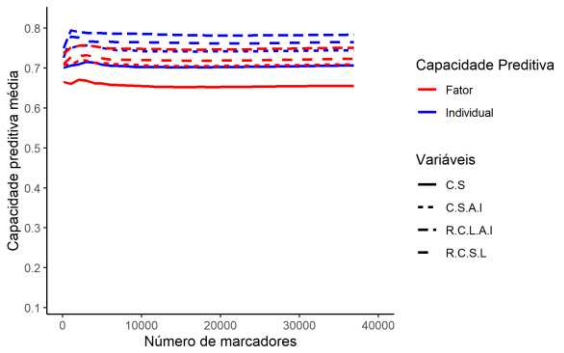
De acordo com de los Campos et al. (2009), os métodos estatísticos em GWS podem ser avaliados a partir da estimação acurada da herdabilidade. Conforme Resende et al.(2012), a capacidade preditiva é uma medida da capacidade do método predizer de forma acurada. Desse modo, avaliou-se (Figura 2) a predição genômica tanto na análise individual como na com fator. Apesar da variável fenotípica comprimento da folha da bandeira (C.F.B), pertencente ao quinto fator, obter baixa predição considerando apenas a análise com o fator. De modo geral, os resultados foram semelhantes para ambas as análises. Destacando-se o primeiro; segundo e terceiro fator como melhores em predizer de forma acurada, o que condiz com as altas comunalidades encontradas (Tabela 4).

Figura 2 – Gráficos de linha com a capacidade preditiva média das variáveis individuais e do fator com suas respectivas variáveis.

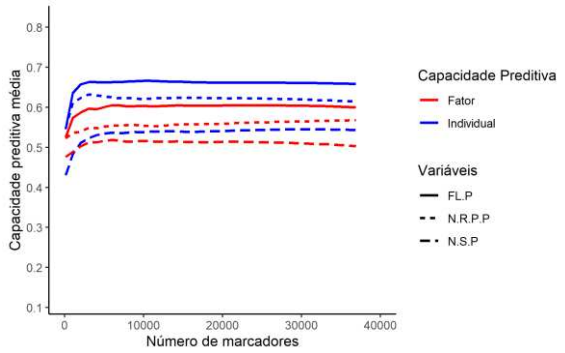
a) Fator 1



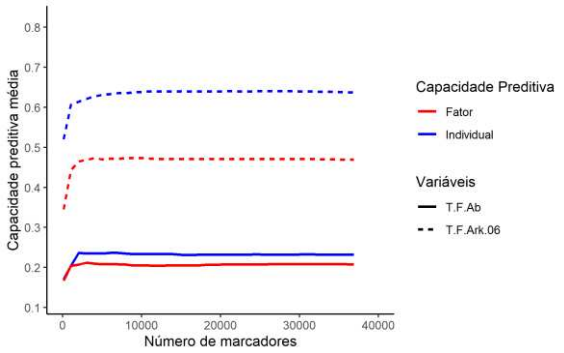

b) Fator 2



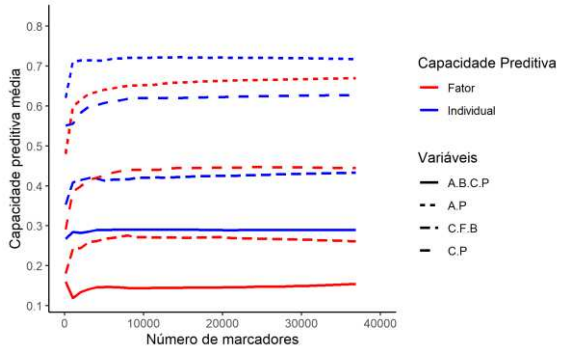

c) Fator 3



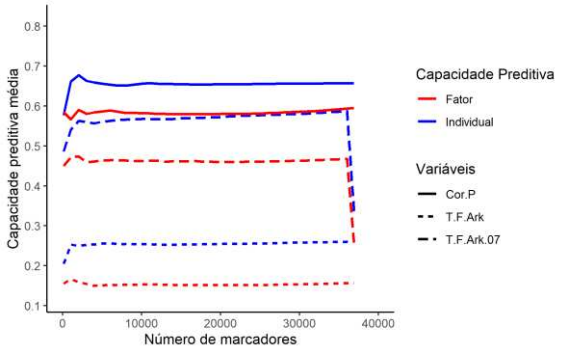

d) Fator 4




e) Fator 5

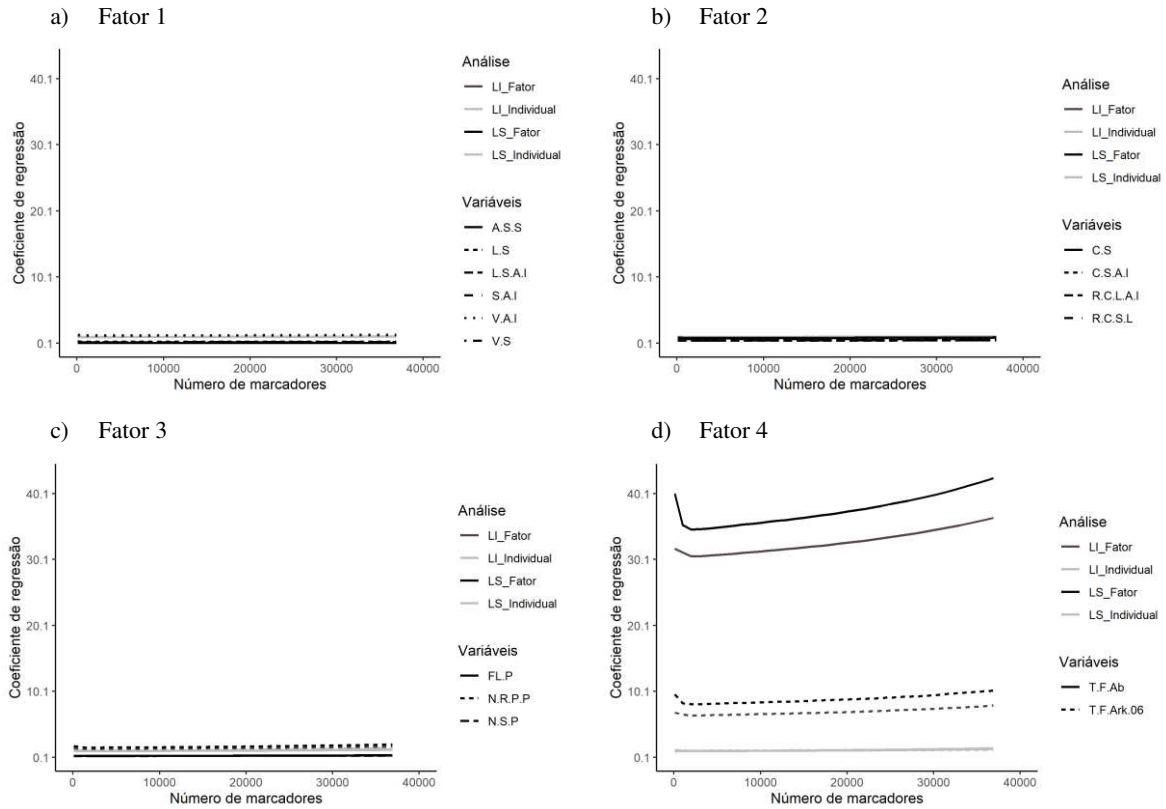



f) Fator 6

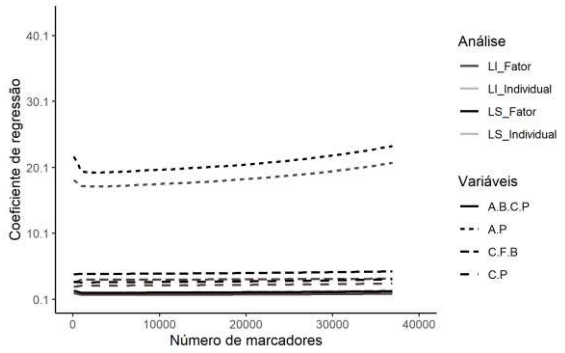


Além disso, segundo Resende et al. (2012), quando se analisa vários valores de herdabilidade, deve-se escolher, como melhor estimativa, aquele que fornece coeficiente de regressão igual a um. De acordo com o mesmo, esse coeficiente mede a capacidade do método predizer de forma não viesada. Assim, para os valores de coeficientes de regressão, entende-se que: Abaixo de um (< 1), os valores genéticos genômicos (*Genomic estimated breeding values* - GEBVs) estão superestimados, para valores acima de um ( > 1), os GEBVs estão subestimados e iguais a um (= 1), os GEBVs são não viesados. Dessa maneira, verificou-se (Figura 3) ausência de fatores que apresentaram GEBVs não viesados. Porém, na análise individual Cor do pericarpo (Cor.P) e Cor da semente (Cor.S) apresentaram GEBVs não viesados para marcadores acima de 100.

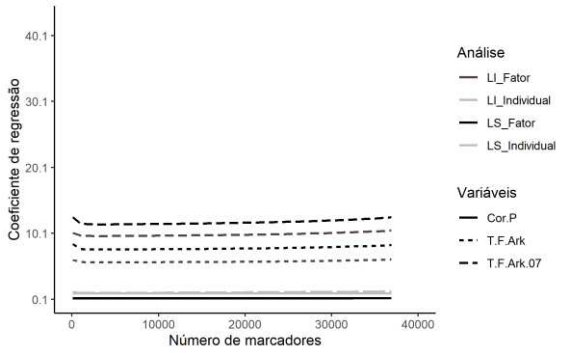
Figura 3 – Gráficos de linha com o limite inferior (LI) e superior (LS) do coeficiente de regressão da análise individual e com fator.



e) Fator 5



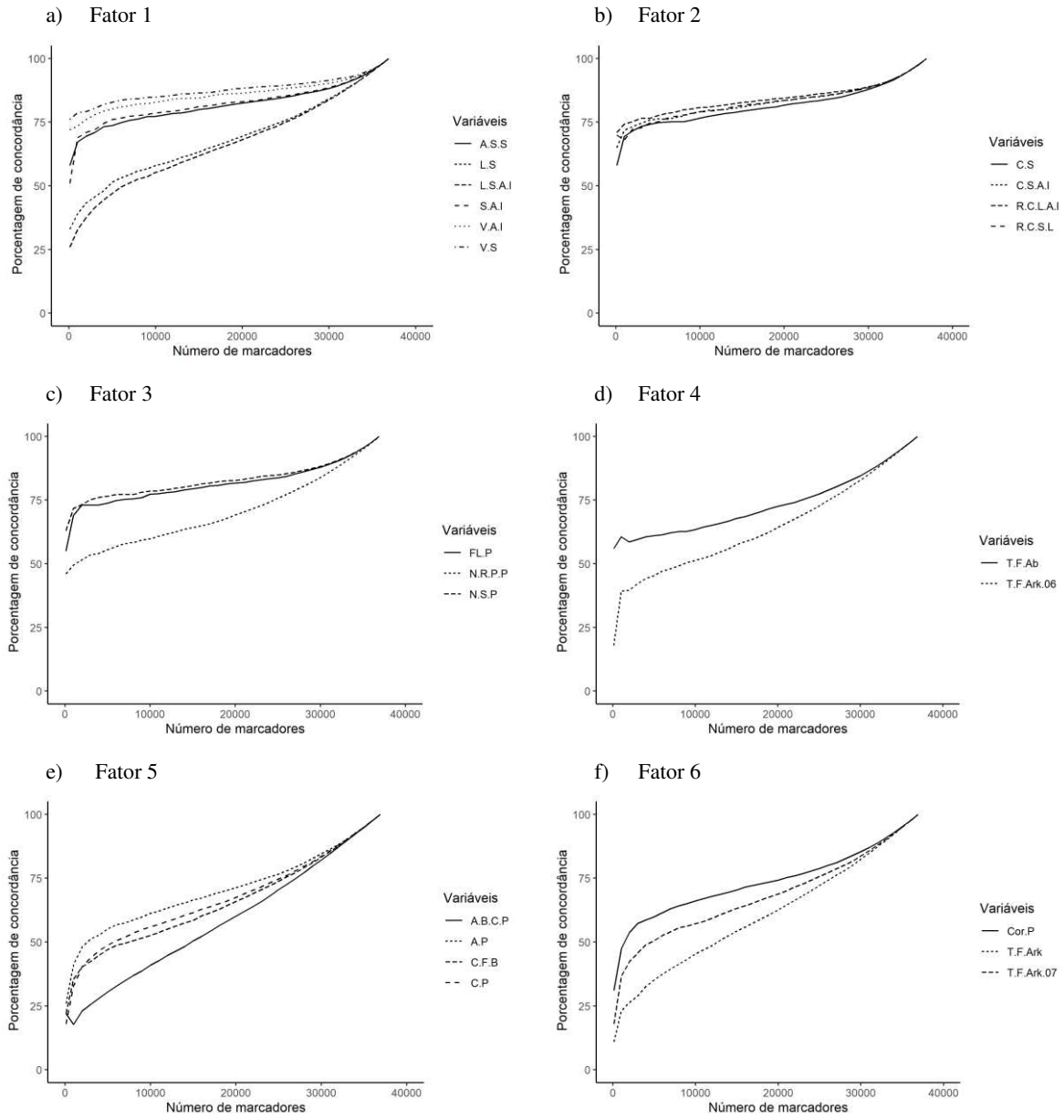

f) Fator 6




Considerando a análise individual, Suela et al. (2019) reportaram que as características do arroz (*Oryza sativa*): Comprimento da folha bandeira (C.F.B); número de panículas por planta (N.P.P); número de ramos da panícula primária (N.R.P.P); comprimento da semente (C.S); largura da semente (L.S), apresentaram coeficiente de regressão próximo de 1 ao utilizar um índice genômico via G-BLUP na predição genômica. Ainda, Suela et al.(2019) apresentaram valores semelhantes ao encontrado para capacidade preditiva para essas variáveis ao utilizar o índice genômico. Já GUO et al. (2014) reportaram valores da capacidade preditiva semelhantes para todas as variáveis do presente estudo.

Por fim, analisou-se a concordância entre os marcadores selecionados considerando o fator e as variáveis individuais. Acredita-se que ao escolher fatores com alta concordância reduz-se o tempo de computação, pois seleciona-se um grupo de variáveis e não uma única variável fenotípica. Desse modo, foi apresentada na Figura 4 a concordância entre os marcadores selecionados na análise individual e na análise com o fator. O primeiro fator obteve quatro variáveis com porcentagem acima de 50% de concordância para todos os grupos de marcadores, sendo elas: Área da superfície da semente (A.S.S); superfície de arroz integral (S.A.I); volume de arroz integral (V.A.I); volume da semente (V.S). O segundo fator apresentou o melhor resultado, com todas as variáveis acima de 50%. O terceiro fator obteve duas variáveis com porcentagem acima de 50% de concordância para todos os grupos de marcadores: Floretes por panícula (FL.P); número de sementes por panícula (N.S.P). O quarto fator apresentou apenas a variável tempo de floração no Aberdeen (T.F.Ab) com concordância acima de 50% entre os marcadores selecionados na análise individual e na análise com o fator para todos os grupos de marcadores na. Já o quinto e o sexto fator não apresentaram variáveis fenotípicas acima dessa porcentagem para todas as marcas.

Figura 4 – Gráficos de linha com a concordância entre as marcas selecionadas na análise individual e com fator.

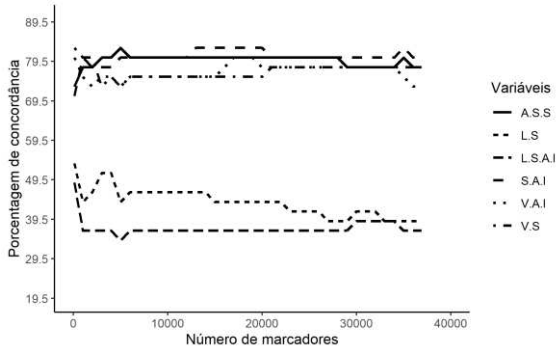



Para concordância entre os 10% melhores (Figura 5) e 10% piores (Figura 6) indivíduos na análise individual e com fator, considerando acima de 50% de concordância para todos grupo de marcadores. O primeiro fator obteve 4 variáveis fenotípicas: Volume da semente (V.S); área da superfície da semente (A.S.S); superfície de arroz integral (S.A.I) e volume de arroz integral (V.A.I). O segundo e terceiro fator obtiveram todas as variáveis acima de 50%. O quarto fator não obteve variável para os 10% piores, mas obteve uma única variável para concordância entre os 10% melhores: Tempo de floração no Aberdeen (T.F.Ab). O quinto fator não obteve nenhuma variável. O sexto fator apresentou para a concordância entre os 10% melhores indivíduos na análise individual e com fator, 2 variáveis: Tempo de floração no Arkansas

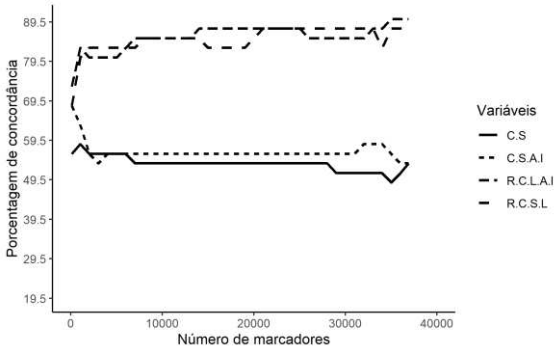
considerando o ano 07 (T.F.Ark.07) e cor do pericarpo (Cor.P). Porém, para os 10% piores, não obteve variáveis acima de 50% de concordância para todos os grupos de marcadores.

Figura 5 – Gráficos de linha com a concordância entre os 10% melhores indivíduos na análise individual e com fator.

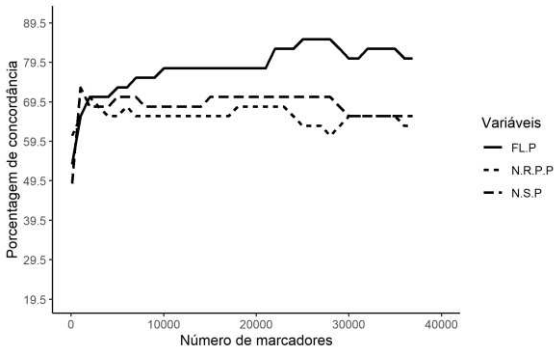
a) Fator 1



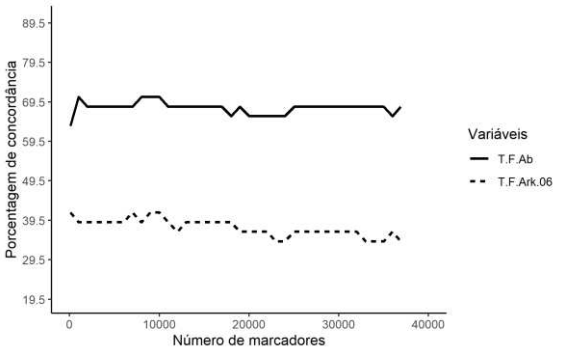

b) Fator 2



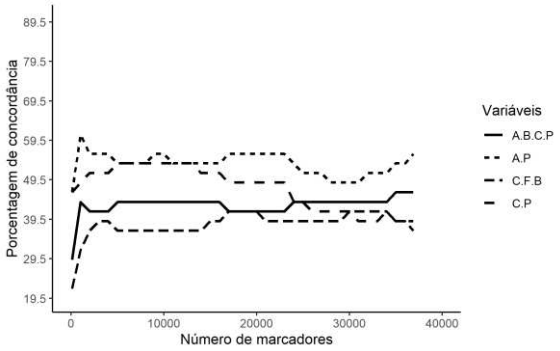

c) Fator 3




d) Fator 4




e) Fator 5




f) Fator 6

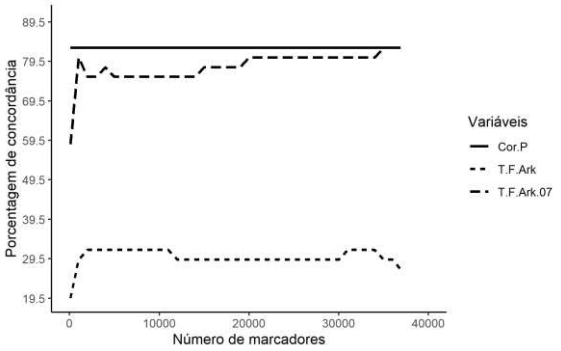
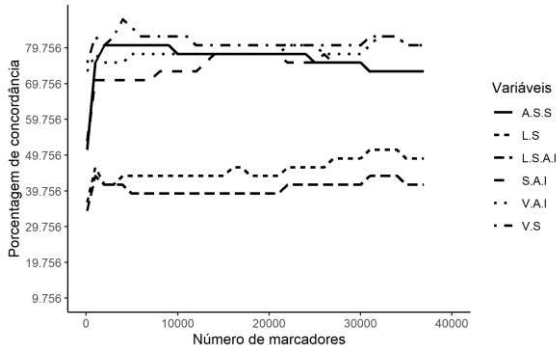
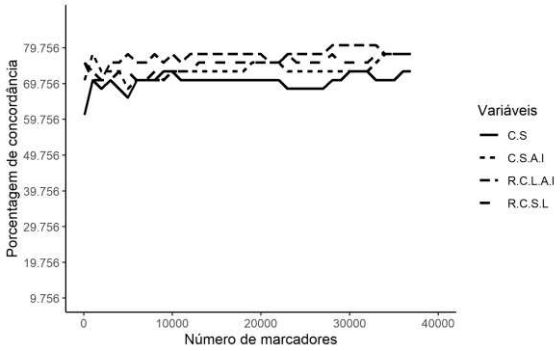


Figura 6 – Gráficos de linha com a concordância entre os 10% piores indivíduos na análise individual e com fator.

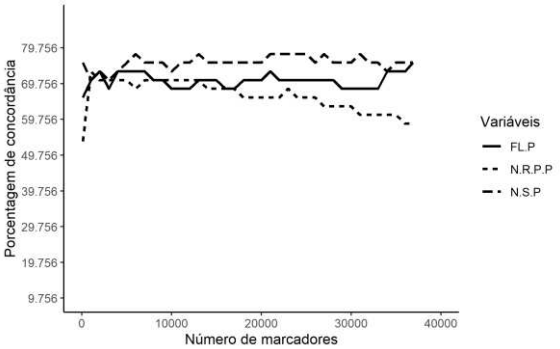
a) Fator 1



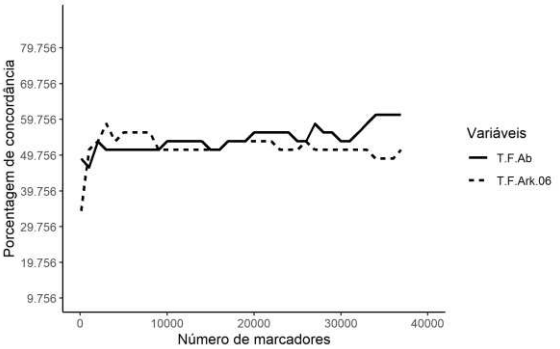

b) Fator 2



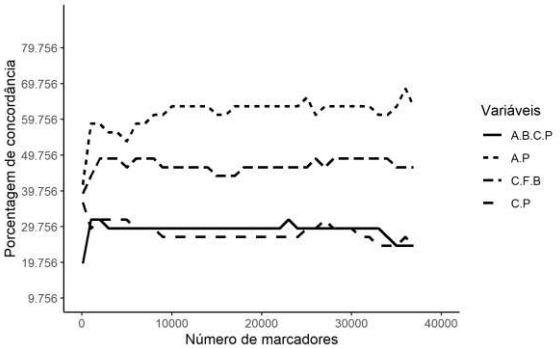

c) Fator 3



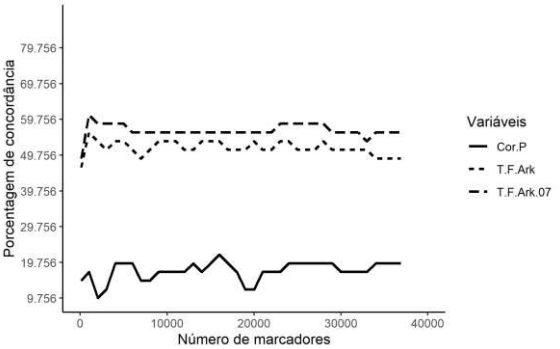

d) Fator 4




e) Fator 5




f) Fator 6



# CONCLUSÃO

A eficiência da análise de fatores ao selecionar marcadores associados a grupos de características de interesse e na predição genômica via BLUP genômico, apresentou resultados satisfatórios, uma vez que ao comparar a análise individual apresentou resultados semelhantes. Destacando, o segundo fator composto por variáveis relacionadas a característica de morfologia das sementes, visto que os resultados foram satisfatórios para todas as variáveis fenotípicas em termos de capacidade preditiva e concordância com as marcas selecionadas na análise individual.

## REFERÊNCIAS BIBLIOGRÁFICAS

ALVARENGA, A.B. et al. Comparing Alternative Single-Step GBLUP Approaches and Training Population Designs for Genomic Evaluation of Crossbred Animals. **Frontiers in Genetics**, v.11, p.1-19, 2020.Disponível em: https://doi.org/10.3389/fgene.2020.00263. Acesso em: 2 nov.2020.

AMMIRAJU, J.S.S. et al. The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. **Genome Research**, v.16, n.1, p.140-147, 2006. Disponível em: http://www.genome.org/cgi/doi/10.1101/gr.3766306. Acesso em: 9 nov.2019.

AKINWALE, M.G.et al . Heritability and correlation coeficiente analysis for yield and its components in rice (*Oryza sativa L.*). **African Journal of Plant Science**, v.5, p. 207-212, 2011.

BARBOSA, I. P. et al. Recommendation of Coffea arabica genotypes by factor analysis. **Euphytica**, v. 215, p.178, 2019. Disponível em: https://doi.org/10.1007/s10681-019-2499-x. Acesso em: 9 abr. 2020.

BISNE, R; SARAWGI, A. K.; VERULKAR, S. B. Study of heritability, genetic advance and variability for yield contributing characters in rice. **Bangladesh Journal of Agricultural Research**, v. 34, p. 175-179, 2009.

CATTELL, R. B. The scree test for the number of factors. **Multivariate Behavioral Research**, n. 1, p. 245-276, 1966.

CONAB. COMPANHIA NACIONAL DE ABASTECIMENTO. **A cultura do arroz**. Brasília: Companhia Nacional de Abastecimento. Disponível em: http://www.conab.gov.br. Acesso em: 22 de dez. 2020.

CONAB. COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da safra brasileira de grãos**. Brasília: Companhia Nacional de Abastecimento. Disponível em: http://www.conab.gov.br. Acesso em: 22 de dez. 2020.

CORRAR, L. J. et al. **Análise Multivariada:** para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas. 568p.,2014.

DALCHIAVON, F.C. et al. Correlação linear entre componentes da produção e produtividade do arroz de terras altas em sistema plantio direto**. Semina**: Ciências Agrárias**,** v. 33, n. 5, p. 1629-1642, 2012.

DE LOS CAMPOS, G.; GIANOLA, D.; ROSA, G.J.M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal Animal Science**, v. 87, p. 1883–1887, 2009.

FERREIRA, D. F. **Estatística Multivariada**. 2.ed. Lavras: UFLA. 675p., 2011.

FIGUEIREDO FILHO, D. B; JÚNIOR, J. A. S. Visão além do alcance: uma introdução à análise fatorial. **Opinião Pública**, v. 16, p. 160-185, 2010.

GAO, H. et al. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. **Genetics Selection Evolution**, v. 44, p. 2-8, 2012.

GUO, Z. et al. The impact of population structure on genomic prediction in stratified populations. **Theoretical and applied genetics**, v.127, p.749-762, 2014. Disponível em: https://doi.org/10.1007/s00122-013-2255-x. Acesso em: 9 set.2020.

KARAMAN, E.; LUND, M.S.; SU, G. Multi-trait single-step genomic prediction accounting for heterogeneous (co)variances over the genome. **Heredity,** v.124, p**.**274–287, 2020. Disponível em: https://doi.org/10.1038/s41437-019-0273-4. Acesso em: 29 nov.2020.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** Uma abordagem aplicada. Belo Horizonte: UFMG. 295p.,2007.

PAIVA, J. T. et al. Genetic evaluation for latent variables derived from factor analysis in broilers. **British Poultry Science**, v. 61, p.3-9,2020. Disponível em: https://doi.org/10.1080/00071668.2019.1680801. Acesso em: 18 set.2020.

RAMÃO, C.J. et al. Efeito do número de operações mecanizadas de nivelamento de solo sobre os componentes de rendimento e altura da lâmina de água na cultura do arroz irrigado. **Tecnológica**, v. 23, n. 1, p. 14-21, 2019. Disponível em: https://doi.org/10.17058/tecnolog.v23i1.12246. Acesso em: 18 mai.2020.

RENCHER, A. C. **Methods of multivariate analysis**. New York: John Wiley, 2002

RESENDE, M. D.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana(MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial.** Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291p. Disponível em: http://www.det.ufv.br/ppestbio/corpo_docente.php. Acesso em: 4 mar. 2019.

RESENDE JR, et al. Accuray of Genomic Selection Methods in a Standard Dataset of Loblolly Pine (Pinus taeda L.). **Genetics**, v. 190, p. 1503 – 1510, 2012.

REIFSCHNEIDER, F. J. B. et al. **Uma pitada de biodiversidade na mesa dos brasileiros**. 1. ed. Brasília: DF. 156p., 2015.

SPINDEL, J. et al. Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. **Plos genetics**, v.11, p. 1-25, 2015. Disponível em: https://doi.org/10.1371/ journal.pgen.1004982. Acesso em: 22 dez. 2020.

SUELA, M.M. et al. Combined index of genomic prediction methods applied to productivity traits in rice. **Ciência Rural**, Santa Maria, v. 49, n. 6, p.1-9, 2019. Disponível em: https://doi.org/10.1590/0103-8478cr20181008. Acesso em: 2 dez. 2020.

TEIXEIRA, F. R. F. et al. Determinação de fatores em características de suínos. **Revista Brasileira de Biometria**, v.33, p.130-138, 2015.

TEIXEIRA, F.R.F. et al. Factor analysis applied to genome prediction for high-dimensional phenotypes in pigs. **Genetics and Molecular Research**, v. 15, p. 1-10, 2016.

ZHANG, Z. et al. Best linear unbiased prediction of genomic breeding values using a trait-specifc marker-derided relationship matrix. **Plos one**, v.5, p.423–447., 2010.

ZHANG, Z. et al. Accuracy of genomic prediction using low-density marker panels. **Journal Dairy Science**, v.94, n. 7, p. 3642-3650, 2011.

ZHAO, K. et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. **Nature Communication**, v.2, p.1-10, 2011. Disponível em: https://doi.org/10.1371/journal.pgen.1002221. Acesso em: 9 nov. 2019.