UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS RURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

Murilo Vieira Loro

QUESTÕES PARA O EXAME DE QUALIFICAÇÃO DE DOUTORADO

Santa Maria, RS
2024

SUMÁRIO

QUESTÕES PROFESSOR DR. ALBERTO CARGNELUTTI FILHO

Questão 01. A Análise Fatorial (AF) é uma técnica estatística utilizada para identificar estruturas subjacentes em um conjunto de dados, ou seja, para reduzir um grande número de variáveis observadas em um número menor de fatores latentes. Esses fatores são variáveis hipotéticas que explicam os padrões de correlação entre as variáveis observadas. Considerando as dimensões do banco de dados para a sua tese que envolvem genótipos, datas de semeadura e variáveis avaliadas, elabore possíveis problemas/objetivos que poderiam ser resolvidos com base em aplicações da análise fatorial. **Resposta: 1. INTRODUÇÃO A ANÁLISE FATORIAL e 2. APLICAÇÕES DA ANÁLISE FATORIAL EM UM CONJUNTO DE DADOS DA TESE.**

1. INTRODUÇÃO A ANÁLISE FATORIAL

1.1 O que é análise fatorial?

Dentre os caracteres avaliados em um programa de melhoramento de milho, os melhoristas têm maior interesse na produtividade de grãos, que é controlada por vários genes e é altamente influenciada pelo ambiente (NARDINO, BARETTA, et al., 2016). Portanto, é comum avaliar outros caracteres como altura da planta, massa da espiga, dias para o florescimento, e outros componentes da produtividade. Esses caracteres estão correlacionados à produtividade de grãos e podem ser levadas em consideração na fase de seleção para obter genótipos produtivos que apresentem outros caracteres desejáveis para o melhoramento (MOHAMMADI, PRASANNA, & SINGH, 2003). Além disso, utilizam-se variáveis meteorológicas, afim de identificar como os caracteres agronômicos respondem às variações ambientais (LORO et al., 2024).

Os dados de experimentos agronômicos frequentemente apresentam um volume grande e complexamente interligado, tornando a análise direta um desafio. Com isso, identificar quais caracteres estão associados ao desempenho agronômico ou quais variáveis meteorológicas influenciam a produtividade se torna uma tarefa complexa. A análise fatorial é uma técnica de análise multivariada que facilita a identificação de grupos de caracteres inter-relacionados, permitindo ao pesquisador destacar dimensões ocultas nos dados, como um fator nutricional ou meteorológico, que auxiliam na tomada de decisão para o melhoramento das cultivares de milho (HAIR et al., 2009).

Por exemplo, em um estudo com dados de 78 genótipos de milho, foram avaliados os seguintes caracteres: altura de planta (AP, em cm), altura da inserção da espiga (AE, em cm), comprimento da espiga (CE, cm), diâmetro da espiga (DE, g), massa da espiga (ME, g), massa de grãos da espiga (MGE, g), produtividade de grãos (PROD, em Mg ha^{-1}), proteína bruta (CP, g/100 g), lisina (LYS, g/100 g), metionina (MET, g/100 g), cistina (CYS, g/100 g), treonina (THR, g/100 g), triptofano

(TRP, g/100 g), valina (VAL, g/100 g), isoleucina (ILE, g/100 g), leucina (LEU, g/100 g), fenilalanina (PHE, g/100 g), histidina (HIS, g/100 g), arginina (ARG, g/100 g), radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), e soma térmica do florescimento feminino à colheita (SFFC, °C dia).

Observar a estrutura de relações entre os caracteres por meio da análise de correlação linear é uma tarefa complexa. Assim, a aplicação da análise fatorial permite agrupar essas variáveis em fatores significativos que explicam grande parte da variabilidade dos dados. Assim, pode ser possível identificar, por exemplo, um (i) fator de caracteres morfológicos (AP e AE), (ii) fator de caracteres produtivos (CE, DE, ME, MGE e PROD), (iii) fator de caracteres nutricionais (LYS, MET, CYS, THR, TRP, VAL, ILE, LEU, PHE, HIS e ARG), e (iv) fator de variáveis meteorológicas (RSFM, RSFF, RFMC, RFFC, SSFM, SSFF, SFMC e SFFC). Logo, pode-se reduzir o conjunto de 26 caracteres em apenas quatro fatores (variáveis latentes) que explicam uma determinada porcentagem da variância total do conjunto de dados.

A análise fatorial é uma técnica que explora as inter-relações entre caracteres, organizando-os em grupos chamados fatores, que representam dimensões latentes dos dados. Esses fatores, criados a partir de combinações lineares dos caracteres originais, tem sido eficiente em simplificar um conjunto extenso de caracteres em um número menor de fatores que ainda preservam a variabilidade e as características essenciais dos dados em experimentos agronômicos (GRANATE et al., 2001; BHARATHIVEERAMANI, PRAKASH, 2012; WADE et al., 2020; DHALIWAL; WILLIAMS, 2020).

A análise fatorial tem dois usos principais: resumir e reduzir o número de caracteres, facilitando a interpretação e permitindo que conceitos complexos, como a qualidade nutricional do milho (dependente de teor de proteína e aminoácidos), sejam representados de forma concisa em outras análises multivariadas. Diferente de métodos de dependência, como a regressão múltipla, onde um caractere é usado como dependente e os outros como independentes, a análise fatorial é uma técnica de interdependência. Logo, todos os caracteres são analisados simultaneamente, sem distinção de dependência, ou seja, cada um é influenciado por todos os demais. Enquanto as técnicas de dependência têm como objetivo a previsão e explicação de resultados, as técnicas de interdependência, como a análise fatorial, buscam identificar a estrutura subjacente entre os caracteres (HAIR et al., 2009).

1.2 Objetivos da aplicação da análise fatorial

O propósito da análise fatorial é condensar a informação contida em diversos caracteres originais em um conjunto menor de novas dimensões compostas ou fatores, com perda mínima de informação. Ao atingir esse objetivo, a análise fatorial é ajustada com as seguintes questões: obtenção do resumo de dados; seleção de caracteres e uso de resultados da análise fatorial com outras técnicas multivariadas.

A análise fatorial geralmente é aplicada a partir de uma matriz de correlação entre os caracteres do estudo. Nesta situação, o objetivo da pesquisa é resumir os caracteres, ou seja, analisar o conjunto de variáveis para identificar as dimensões latentes que não são fáceis de observar. A análise fatorial fornece para o pesquisador duas saídas distintas: resumo de dados e redução de dados. No resumo de dados, a análise fatorial obtém fatores que, quando interpretados e compreendidos, descrevem os dados em um número muito menor de conceitos do que as variáveis individuais originais. Redução de dados estende esse processo derivando um valor empírico (escore fatorial) para cada dimensão (fator) e então substituindo o valor original por esse novo valor. Embora a analise fatorial seja utilizada para reduzir as dimensões dos dados, é necessário definir quais variáveis serão utilizadas nesta análise. A inclusão indiscriminada de um grande número de variáveis pode aumentar a possibilidade de resultados insatisfatórios. A qualidade e o significado dos fatores obtidos refletem aas bases conceituais das variáveis incluídas na análise.

A partir dos resultados da análise fatorial, é possível obter uma visão direta das inter-relações entre variáveis e a estrutura subjacente dos dados, sendo um excelente ponto de partida para outras análises multivariadas. Logo, pode-se compreender quais variáveis podem atuar juntas e quantas variáveis podem realmente ser consideradas como tendo impacto na análise. Assim, problemas associados com grandes números de variáveis ou altas intercorrelações entre variáveis podem ser reduzidas pela substituição das novas variáveis.

1.3 Processo para aplicação da análise fatorial

Para melhor compreender o processo de aplicação da análise fatorial será utilizado o seguinte exemplo: Um pesquisador avaliou 80 genótipos de milho quanto a caracteres fenológicos, morfológicos, produtivos, nutricionais e variáveis meteorológicas. Em cada grupo foram avaliados quatro caracteres, somando 20 caracteres. A partir deste conjunto de caracteres é possível verificar as relações lineares por meio de uma matriz de correlação que contém os coeficientes de correlação de Pearson.

É comum verificar em uma matriz de correlação a ocorrência de coeficientes altos entre subconjuntos de caracteres, por exemplo, entre os caracteres fenológicos ou entre as variáveis meteorológicas, conforme observado por Loro et al. (2024). Isso indica que esses caracteres podem

estar medindo uma mesma dimensão subjacente, que é denominada de fator ou variável latente. Portanto, na análise fatorial, o objetivo é reduzir essa matriz de correlação à sua dimensão subjacente verificando quais caracteres parecem se agrupar de maneira significativa (HAIR et al., 2009).

Assim, a análise fatorial busca identificar os caracteres que apresentam correlações altas com um grupo de caracteres específicos, mas que não se correlacionam, ou apresentam baixas correlações com os caracteres fora daquele grupo. Já na exploração inicial dos dados, na matriz de correlação, é possível identificar os caracteres que possivelmente não irão se comportar bem na análise fatorial, especialmente se exibirem baixas correlações com os grupos de caracteres. Assim, os caracteres que não se correlacionam com nenhum grupo de caracteres já podem ser removidos da análise fatorial (HAIR et al., 2009).

1.3.1 Teste de Kaiser-Meyer-Olkin (KMO) e de esfericidade de Bartlett

Para iniciar a análise fatorial, considerando o exemplo de 20 caracteres avaliados em 80 genótipos de milho, primeiro é preciso verificar se as relações existentes entre os caracteres estão adequadas para serem agrupados em fatores. Para isso, utilizam-se basicamente dois testes: teste de Kaiser-Meyer-Olkin (KMO) e o teste de esfericidade de Bartlett. O teste KMO, indica se os fatores identificados na análise fatorial podem descrever adequadamente as variações nos dados originais. Essa medida, varia entre 0 e 1, e indica a proporção da variância dos caracteres explicada pelos fatores. Quanto mais próxima de 1, mais adequado é o ajuste dos dados para uma análise fatorial. Hair et al. (2009) sugere 0,50 como o mínimo aceitável, valores abaixo disso indicam a possível necessidade de ajustar os caracteres incluídos.

Quando não há correlação entre os caracteres, a matriz de correlação se torna uma matriz identidade, na qual apenas os valores da diagonal principal são diferentes de zero. Nessa situação de independência total, não é possível construir fatores, pois não há associação entre os caracteres. Logo, o teste de Bartlett é utilizado para verificar se os dados são independentes (hipótese nula) (HAIR et al., 2009). Se o teste for significativo ($p < 0,05$), rejeita-se a hipótese de nula de independência, indicando que há relação entre os caracteres, permitindo que sejam agrupadas em fatores. No entanto, como relatado por Field, Miles, Field (2012), o teste de Bartlett é sensível ao tamanho da amostra, geralmente resulta em significância estatística em amostras grandes. Assim, mesmo que o teste indique associações, recomenda-se descartar caracteres com correlações muito baixas, pois tendem a não se agrupar distintamente nos fatores.

1.3.2 Métodos de extração, números e rotação de fatores

A análise fatorial é uma técnica que agrupa caracteres inter-relacionados em fatores, que representam conceitos gerais nos dados. No exemplo, ao avaliar 20 caracteres em 80 genótipos de

milho, o pesquisador espera que os fatores identificados façam sentido em relação ao estudo. Na fase final, é importante nomear cada fator com base nos caracteres de maior carga, garantindo que esses grupos tenham coerência teórica. A análise fatorial, ainda que exploratória, exige que os caracteres escolhidos tenham algum fundamento, pois todos afetam o resultado final (MATOS; RODRIGUES, 2019).

Uma vez que os caracteres foram definidos e a matriz de coeficientes de correlação foi adequada pode-se aplicar a análise fatorial. A partir disso, deve-se tomar decisões sobre qual método de extração dos fatores utilizar e o número de fatores a serem selecionados para explicar a variabilidade dos dados. Existem alguns métodos de extração dos fatores como: o de componente principal, fator principal, máxima verossimilhança, mínimos quadrados ordinários e mínimos quadrados generalizados (HAIR et al., 2009; HONGYU, 2018; MATOS; RODRIGUES, 2019). Em geral, os métodos de componente principal e máxima verossimilhança fornecem os melhores resultados e mais recomendados para análise fatorial quando as amostras apresentam distribuição normal e não-normal, respectivamente (COSTELLO; OSBORNE, 2005). O método do componente principal tem sido relatado com um dos melhores métodos de extração de fatores, embora muito similar aos métodos de máxima verossimilhança e fator principal (HONGYU, 2018).

A partir dos fatores extraídos, é essencial determinar o número de fatores que serão utilizados para explicar a variabilidade dos dados. O número de fatores extraídos é igual ao número de caracteres utilizados. Por exemplo, nos dados do exemplo, foram considerados 20 caracteres, portanto, serão extraídos 20 fatores. Se o objetivo da análise fatorial foi reduzir a dimensionalidade dos dados, não faz sentido utilizar todos os 20 fatores. Assim, busca-se utilizar procedimentos que orientam selecionar um número menor de fatores que captam a maior parte da variação dos dados. Para isso existem alguns critérios que ajudam a definir um número adequado de fatores; critério do autovalor, critério do diagrama de inclinação e critério da porcentagem de variância explicada. Se um dos fatores extraídos apresentar um autovalor de elevada magnitude, indica que este fator contribui muito para a explicação das variâncias nos caracteres e pode ser selecionado como um fator importante. Assim, somente fatores com autovalores de alta magnitude deve ser selecionados. O critério de Kaiser-Guttman (GUTTMAN; 1954; KAISER, 1960) é utilizado para definir o número de fatores a serem retidos. Esse critério estabelece que são retidos os fatores com autovalor maior que 1,0.

O critério do diagrama de inclinação, ou teste *scree*, ajuda a decidir quantos fatores devem ser usados na análise fatorial. Esse método utiliza um gráfico onde o eixo y representa os autovalores e o eixo x representa os fatores, ordenados por tamanho dos autovalores. Nos dados do exemplo, com 20 caracteres de milho, o gráfico terá 20 pontos, cada um representando o autovalor de um fator. Ao traçar uma linha passando por esses pontos, observa-se onde a curva começa a se estabilizar, ou seja,

a ficar horizontal. Esse ponto indica o número máximo de fatores significativos, pois os fatores após esse ponto contribuem muito pouco para explicar a variabilidade dos dados.

O critério da porcentagem da variância observada utiliza o percentual acumulado de variância explicada por cada fator para determinar quantos fatores manter na análise. O objetivo é escolher fatores que representem uma parte significativa da variabilidade dos dados. Segundo Hair et al. (2009), é recomendado que os fatores selecionados expliquem, no mínimo, 60% da variância total. Assim, se no exemplo dos dados de milho três fatores já capturam 80% da variância, os demais 17 fatores podem ser desconsiderados, pois contribuem pouco para a explicação dos dados.

Após a extração dos fatores, as cargas fatoriais são usadas para avaliar o quanto os caracteres estão associados a cada fator. Normalmente, muitos caracteres têm cargas altas no fator principal (primeiro fator) e baixas nos outros fatores, o que pode dificultar a interpretação. Para melhorar essa distinção entre os fatores, utiliza-se a técnica de rotação de fatores, que tem por objetivo melhorar a interpretação dos resultados (COSTELLO; OSBORNE, 2005; HAIR et al., 2009). A partir da rotação, é possível obter uma separação mais evidente entre os fatores, permitindo uma interpretação precisa.

Existem dois tipos de rotação dos fatores: ortogonal e obliqua. Na rotação fatorial ortogonal, os fatores extraídos permanecem independentes entre si, ou seja, não há correlação entre fatores, com os eixos sendo mantidos perpendiculares. Já na rotação fatorial oblíqua, os fatores podem ser correlacionados entre si, ou seja, seus eixos não permanecem perpendiculares, permitindo que os fatores sejam interdependentes.

Na literatura são apresentados diversos tipos de rotação ortogonal como: *quartimax*, *equimax* e *varimax*. A rotação *varimax*, sendo sido comumente utilizada e preferida dentre as rotações ortogonais, quando se espera que os fatores sejam independentes (GRANATE et al., 2001; HONGYU, 2018; DHALIWAL, WILLIAMS, 2020). No entanto, as rotações ortogonais estipulam, a priori, que não há correlação entre os fatores ($r = 0$), gerando, portanto, fatores totalmente independentes uns dos outros. Entretanto, esse pressuposto é raramente obtido nas pesquisas das ciências agrárias.

No exemplo dos dados de milho, espera-se que haja uma relação entre caracteres morfológicos, produtivos, nutricionais proteicos e variáveis meteorológicas, uma vez que o desenvolvimento das plantas é influenciado pelas condições ambientais. Métodos ortogonais, ao assumirem que os fatores são independentes, podem perder confiabilidade quando esses fatores estão de fato correlacionados, além de superestimar a variância explicada, já que não consideram as interações entre os fatores (HAIR et al. 2009). Nesse contexto, a rotação oblíqua permite que os fatores sejam correlacionados, refletindo melhor a realidade dos dados. Existem diferentes métodos de rotação oblíqua, como *oblimin*, *quartimin* e *promax*, que apresentam resultados similares, sem um consenso sobre qual é o mais adequado, conforme indicado por Hair et al. (2009).

A escolha entre rotação ortogonal ou oblíqua depende de algumas considerações importantes. Primeiro, se houver uma base teórica sólida para acreditar que os fatores são independentes, a rotação ortogonal pode ser uma opção. No entanto, Field (2009) destaca que em áreas como as Ciências Humanas e Sociais, onde as variáveis são frequentemente correlacionadas, a rotação ortogonal não é adequada, pois não faz sentido assumir que os fatores não têm relação entre si. Embora o autor tenha relacionado a área de Ciências Humanas e Sociais, pode-se fazer uma analogia a área de Ciências Agrárias em estudo com plantas, como o exemplo com a cultura do milho, onde os caracteres são altamente relacionados entre si. Para isso, a rotação oblíqua é mais indicada, já que permite que os fatores sejam correlacionados.

A partir da rotação, os resultados podem ser interpretados. O objetivo é verificar se é possível identificar fatores que podem ser nomeados com base nos caracteres do fator. Por exemplo, pode-se esperar que um fator seja constituído pelos seguintes caracteres: LYS, MET, CYS, THR, TRP, VAL, ILE, LEU, PHE, HIS e ARG. Assim, esse fator pode ser nomeado como o fator "nutricional proteico", pois está associado a caracteres proteicos (proteína e aminoácidos) que determinam a qualidade nutricional dos grãos.

Após a rotação dos fatores, é possível calcular os escores fatoriais para cada genótipo avaliado. Esses escores representam a combinação dos caracteres em fatores, e podem ser utilizados em análises subsequentes, como análise de trilha, regressão linear ou distância euclidiana. Por exemplo, se no estudo com os dados de milho for identificado um fator relacionado às variáveis meteorológicas (RSFM, RSFF, RFMC, RFFC, SSFM, SSFF, SFMC e SFFC), é possível extrair os escores desse fator e analisar a correlação entre o fator e a produtividade de grãos. Isso permite investigar a relação entre o fator e a produtividade, usando uma variável latente (o fator) ao invés de realizar correlações entre cada variável meteorológica e a produtividade separadamente. Existem dois métodos principais para a obtenção dos escores individuais: método de regressão e o método dos mínimos quadrados ponderados.

1.4 Possíveis aplicações práticas da análise fatorial em dados agronômicos

No melhoramento de plantas, a análise fatorial é uma técnica estatística que pode ser usada para simplificar e interpretar a variabilidade em conjuntos de dados complexos, como os que envolvem diversos caracteres. No caso dos 80 genótipos de milho, por exemplo, pode-se avaliar caracteres morfológicos, produtivos, nutricionais proteicos e variáveis meteorológicas. A análise fatorial permite agrupar esses caracteres em fatores mais simples e compreensíveis, facilitando a interpretação dos dados e o processo de seleção.

Uma vez definidos os fatores e identificados os caracteres associados, essas informações podem ser aplicadas de diversas maneiras no melhoramento de plantas. A análise fatorial agrupa

características correlacionadas em um número reduzido de fatores, permitindo que os melhoristas se concentrem nas variáveis mais relevantes de cada fator. Isso simplifica o processo de seleção, possibilitando a avaliação de múltiplos caracteres simultaneamente de maneira mais eficiente, sem a necessidade de analisar cada característica individualmente. Ao invés de selecionar plantas observando o desemprenho em cada caractere individual, os fatores podem ser utilizados como critérios de seleção agregados. Por exemplo, se um fator estiver fortemente relacionado aos caracteres produtivos (CE, DE, ME, MGE e PROD), os genótipos que apresentarem escores elevados nesse fator podem ser selecionados. Isso torna o processo de seleção mais direto e alinhado aos objetivos do programa de melhoramento.

Além disso, os fatores podem ser interpretados como representações de diferentes ideótipos. No contexto dos 80 genótipos de milho, um fator pode agrupar caracteres associadas produtividade, enquanto outro pode agrupar caracteres relacionadas qualidade nutricional dos grãos. Isso permite que os melhoristas identifiquem plantas que atendem melhor ao ideótipo desejado para um determinado ambiente ou objetivo de cultivo, como, por exemplo, genótipos de maior produtividade e qualidade nutricional dos grãos. A análise de fatores permite visualizar a estrutura de correlação entre os caracteres. Isso auxilia a identificação dos caracteres redundantes (ou seja, que estão fortemente correlacionadas) ou caracteres que têm efeitos opostos dentro de um fator. Com essas informações, o melhorista pode tomar decisões sobre quais caracteres priorizar na avaliação e seleção, afim de otimizar o melhoramento.

Os fatores extraídos na análise fatorial podem ser usados em várias análises complementares, como a correlação, regressão e a análise de agrupamento. Esses escores de fatores permitem correlacionar grupos de caracteres com outros caracteres importantes. Por exemplo, ao identificar um fator que representa variáveis meteorológicas, pode-se extrair os escores de cada genótipo para esse fator e, em seguida, analisar a correlação ou fazer uma regressão desses escores com a produtividade de grãos ou o teor de proteína. Esse processo ajuda a entender como as condições meteorológicas determinam caracteres importantes, como produtividade e qualidade.

A análise de fatores também pode orientar a identificação de genótipos que tenham escores altos em fatores complementares, afim de formar uma população base para iniciar um programa de melhoramento de milho. Isso aumenta a frequência de alelos favoráveis na população original, aumentando a probabilidade de desenvolver variedade de polinização aberta, linhagens ou híbridos que apresentam combinação adequada dos caracteres desejado. Dessa maneira, a análise fatorial é uma ferramenta que pode promover uma seleção eficiente, baseada em dados e adaptada à complexidade dos caracteres e variáveis meteorológicas que determinam o desenvolvimento das plantas.

2. APLICAÇÕES DA ANÁLISE FATORIAL EM UM CONJUNTO DE DADOS DA TESE

Nessa aplicação utilizou-se a análise fatorial, afim de verificar o agrupamento dos caracteres em fatores e a relação entre os fatores. Além disso, com os escores individuais dos genótipos em cada fator buscou-se identificar os genótipos que maximizam simultaneamente os caracteres de interesse. Em estudos futuros, há possibilidade de verificar quais fatores determinam, por exemplo, a produtividade de grãos por meio de análise de correlação ou regressão linear múltipla. Ainda, essas análises poderão ser realizadas por base genética.

Título: Agrupamento de caracteres agronômicos, nutricionais e variáveis meteorológicas para avaliação de genótipos de milho via análise fatorial

Resumo: Os objetivos deste estudo foram verificar se é possível agrupar caracteres e analisar as relações lineares entre grupos de caracteres e identificar genótipos de milho com base em multi-caracteres em datas de semeadura. Foram avaliados 78 genótipos de milho em dez datas de semeadura nas safras 2021/2022 e 2022/2023, em Santa Maria, RS, Brasil. Em cada de semeadura, avaliaram-se os genótipos em relação aos caracteres agronômicos, qualidade nutricional dos grãos e acumulo de radiação solar global e soma térmica nos estádios vegetativo e reprodutivo, totalizando 27 caracteres avaliados. Calcularam-se os coeficientes de correlação entre pares de caracteres e, na matriz de correlação, verificou-se a adequação das relações para análise fatorial. Em seguida, aplicou-se a análise fatorial com extração dos fatores pelo método do componente principal e utilizou-se a rotação oblíqua *oblimin* para a obtenção final dos fatores. Obteve-se os escores individuais dos genótipos em cada fator pelo método de regressão. Em cada data de semeadura, esses escores individuais foram utilizados para identificar os genótipos em relação a um ideótipo teórico, estabelecido para maximizar os escores de todos os fatores selecionados. Os 25 caracteres são agrupados nos seguintes grupos: morfológicos, produtivos, nutricionais proteicos e meteorológicos. Os caracteres produtivos e nutricionais proteicos correlacionam-se negativamente, enquanto os caracteres morfológicos e meteorológicos correlacionam-se positivamente entre si e com os caracteres produtivos e nutricionais proteicos. É possível identificar genótipos que incrementam simultaneamente caracteres morfológicos, produtivos, nutricionais proteicos e meteorológicos.

2.1 Introdução

O milho (*Zea mays* L.) é amplamente utilizado na alimentação animal devido ao alto teor de amido (72%) nos grãos (BUTTS-WILMSMEYER et al., 2019), e teores de proteína, fibras, lipídios e aminoácidos, que também compõem a estrutura dos grãos (RODRIGUEZ et al., 2020). O desenvolvimento e seleção de genótipos superiores tem sido eficiente para aumentar o desempenho produtivo (CRISPIM-FILHO et al., 2020) e a qualidade nutricional dos grãos, afim de atender a demanda de grãos de trigo. É importante que os grãos de milho apresentem qualidade suficiente para utilização na produção de rações para alimentação animal.

Estudos sobre a composição média de proteína e aminoácidos em grãos de cereais mostram um perfil variado, essencial para a avaliação nutricional (ALVES et al., 2016; ALVES; CARGNELUTTI FILHO, 2017; LORO et al., 2023; SIMÕES et al., 2023). Os grãos de milho contêm aminoácidos

essenciais, como lisina, metionina e triptofano, que são fundamentais para a dieta de aves e suínos, embora em concentrações menores que outras fontes proteicas, como a soja (SRIPERM; PESTI; TILLMAN, 2010). O perfil de aminoácidos é um indicador da qualidade nutricional, relevante na formulação de rações. Portanto, é importante desenvolver genótipos de milho que combinem alta produtividade e qualidade de grãos.

Entender as relações entre componentes de produtividade, qualidade nutricional dos grãos e variáveis meteorológicas permite desenvolver estratégias de seleção de genótipos com base em múltiplas características e adequar as datas de cultivo. Relações negativas entre caracteres produtivos e nutricionais proteicos tem sido observada em milho (ALVES et al., 2016; ALVES; CARGNELUTTI FILHO, 2017; GUO et al., 2022), o que pode dificultar a seleção de plantas para incremento simultâneo dos caracteres. Já as relações das variáveis meteorológicas como radiação solar global e soma térmica com caracteres agronômicos dependem do estádio de desenvolvimento das plantas (LORO et al., 2024). Nessas pesquisas, a análise de correlação linear tem sido útil para identificar as relações entre os caracteres. Entretanto, quando vários caracteres são avaliados em cada genótipo, torna-se complexo identificar padrões de relação e caracterizar os genótipos considerando um conjunto amplo de características.

Logo, a análise fatorial é uma técnica estatística que estuda correlações entre um grande número de variáveis agrupando-as em fatores. Essa técnica permite a redução de dados criando um novo conjunto de variáveis, menor que o original (HAIR et al., 2009). Isso facilita a caracterização dos genótipos e identificação da inter-relação desses caracteres (HAIR et al., 2009). Estudos tem utilizado a análise fatorial para verificar estruturas adjacentes de vários caracteres agronômicos e ambientais na cultura do milho (GRANATE et al., 2001; BHARATHIVEERAMANI, PRAKASH, 2012; WADE et al., 2020; DHALIWAL; WILLIAMS, 2020).

Os escores da análise fatorial são amplamente usados em análises multivariadas, como a regressão linear múltipla, e podem ser usados para calcular a distância euclidiana entre genótipos e um ideótipo predefinido. Esse ideótipo é uma referência ideal, com características agronômicas desejáveis para o cultivo. O uso da distância euclidiana fornece uma medida objetiva para selecionar genótipos que atendam aos critérios agronômicos, facilitando a escolha de variedades com alto desempenho e adaptação ao ambiente. Isso contribui para uma seleção eficiente de genótipos com melhoria simultânea em vários caracteres (CRISPIM-FILHO et al., 2020). Neste sentido, os objetivos deste estudo foram avaliar a possibilidade de agrupar e analisar as relações lineares entre grupos de caracteres e genótipos de milho com base em multi-caracteres em datas de semeadura.

2.2 Material e Métodos

Os experimentos foram conduzidos na área do Departamento de Fitotecnia da Universidade Federal de Santa Maria, situada a 29º42'S de latitude, 53º49'O de longitude e 95 m de altitude. O clima local, segundo a classificação de Köppen, é Cfa, subtropical úmido, com verões quentes e ausência de estação seca definida (ALVARES et al., 2013). O solo é classificado como Argissolo vermelho distrófico arênico (SANTOS et al., 2018a).

Na safra 2021/2022 foram avaliados genótipos de milho em cinco datas de semeadura: 21 de setembro de 2021, 20 de outubro de 2021, 20 de novembro de 2021, 20 de dezembro de 2021 e 30 de janeiro de 2022. Com exceção da semeadura em 21 de setembro de 2021 que foi composta por 71 genótipos, em todas as demais datas foram semeados 78 genótipos de milho de diferentes bases genéticas (híbridos simples, híbridos triplos, híbridos duplos e variedades). Na safra de 2022/2023 foram avaliados os mesmos 78 genótipos de milho em cinco datas de semeadura: 06 de setembro de 2022, 14 de outubro de 2022, 24 de novembro de 2022, 30 de dezembro de 2022 e 06 de fevereiro de 2023. Os experimentos foram conduzidos sem a utilização de irrigação suplementar.

Em cada data de semeadura, os genótipos de milho foram semeados em parcelas de uma fileira, lado a lado. Cada parcela foi constituída por uma fileira de 5 m de comprimento espaçada em 0,80 m entre fileiras e 0,20 m entre plantas na fileira, totalizando 4 m^2. A densidade foi ajustada por meio de desbaste para 62.500 plantas ha^{-1} (25 plantas por parcela). Foram feitas bordaduras nas laterais e extremidades dos blocos, com plantas de milho. Realizou-se a gradagem da área e a adubação com 415 kg ha^{-1} de adubo químico da fórmula (NPK) 05-20-20. A adubação nitrogenada com ureia (N - 46%) foi fracionada, sendo a primeira aplicação de 250 kg ha^{-1} no estádio V4 e a segunda de 150 kg ha^{-1} no estádio V6 da cultura. Os demais manejos culturais, como controle de plantas daninhas, pragas e doenças, foram realizados de acordo com as indicações técnicas para a cultura de milho, de forma homogênea em todos os genótipos (FANCELLI; DOURADO NETO, 2009).

Os genótipos foram caracterizados fenologicamente nas respectivas datas de semeadura. Para isso, registraram-se as datas de florescimento masculino (50% das plantas em cada parcela apresentavam a última ramificação do pendão visível); florescimento feminino (50% das plantas de cada parcela apresentavam estigmas visíveis na espiga); e o ponto de colheita (palha da espiga e a folha da base da espiga estavam 100% senescentes). Assim, foram obtidos os seguintes caracteres fenológicos: dias da semeadura ao florescimento masculino (SFM, dias), dias da semeadura ao florescimento feminino (SFF, dias), dias do florescimento masculino à colheita (FMC, dias) e dias do florescimento feminino à colheita (FFC, dias).

Após o florescimento masculino e feminino, mensurou-se os caracteres morfológicos de altura de planta (AP, em cm), sendo considerada a média das distâncias entre a superfície do solo e a inserção

da folha bandeira, de cinco plantas por genótipo; e de altura da inserção da espiga (AE, em cm), sendo considerada a média das distâncias entre a superfície do solo e a inserção da primeira espiga, de cinco plantas por genótipo. Posteriormente, foram mensurados os seguintes caracteres produtivos: comprimento da espiga (CE, cm), diâmetro da espiga (DE, g); massa da espiga (ME, g); e massa de grãos da espiga (MGE, g). A partir de todas as plantas da parcela foi avaliada a produtividade de grãos (PROD, em Mg ha^{-1}), corrigida a 13% de umidade.

Uma amostra de 100 g de grãos de milho foi retirada de cada genótipo em cada data de semeadura (773 amostras), acondicionada em sacos de papel e levada à estufa de circulação forçada de ar para atingir 13% de umidade. Posteriormente, os grãos foram moídos em Moinho Retsch (modelo ZM 200) acoplado com peneira de 1mm para obtenção de amostras. Em seguida, amostras moídas foram escaneadas em equipamento NIRS FOSS DS2500 por meio de Espectroscopia de Refletância no Infravermelho Próximo (*Near Infrared Reflectance Spectroscopy*) da empresa Adisseo Brasil Nutrição Animal. Com os espectros de absorbância, gerados para cada amostra, as predições dos valores nutricionais foram determinadas utilizando calibração multivariada para milho moído na plataforma Precision Nutrition Evaluation (PNE). Assim, para cada amostra foram determinados os seguintes caracteres nutricionais proteicos em g/100 g de matéria seca: proteína bruta (CP, g/100 g), lisina (LYS, g/100 g), metionina (MET, g/100 g), cistina (CYS, g/100 g), treonina (THR, g/100 g), triptofano (TRP, g/100 g), valina (VAL, g/100 g), isoleucina (ILE, g/100 g), leucina (LEU, g/100 g), fenilalanina (PHE, g/100 g), histidina (HIS, g/100 g) e arginina (ARG, g/100 g).

As variáveis meteorológicas foram obtidas na estação do Instituto Nacional de Meteorologia (INMET), localizada a 100 metros da área experimental. Foi obtido, para cada dia, a radiação solar global horária, em MJ m^{-2} $hora^{-1}$, as temperaturas máximas e mínimas horárias, em ºC e a precipitação pluviométrica, em mm, para o período entre a semeadura e o ponto de colheita dos genótipos, referente a cada data de semeadura. Para cada dia, foi obtida a radiação solar global diária, em MJ m^{-2} dia^{-1} por meio do somatório da radiação solar global horária, em MJ m^{-2} $hora^{-1}$. Após, para cada genótipo, em cada data de semeadura, a partir da radiação solar global diária, foi obtido a radiação solar global acumulada nos subperíodos: semeadura ao florescimento masculino (RSFM, MJ m^{-2}); semeadura ao florescimento feminino (RSFF, MJ m^{-2}); florescimento masculino à colheita (RFMC, MJ m^{-2}); e florescimento feminino à colheita (RFFC, MJ m^{-2}).

Para cada dia, a partir das temperaturas máximas (Tmax) e mínimas (Tmin) horárias, foram calculados os 24 valores de temperatura média horária. Em seguida, foi calculada a temperatura média (Tmed) diária do ar, em ºC, pela média dos 24 valores de temperatura média horária. Para cada genótipo, em cada data de semeadura, a soma térmica foi obtida somando-se os valores de GD de cada um dos seguintes subperíodos: semeadura ao florescimento masculino (SSFM, °C dia);

semeadura ao florescimento feminino (SSFF, °C dia); florescimento masculino à colheita (SFMC, °C dia); e florescimento feminino à colheita (SFFC, °C dia).

Foram calculados os coeficientes de correlação linear de Pearson entre os pares de caracteres, com a significância avaliada pelo teste t de *Student* a 5% de probabilidade. Como os caracteres estavam em escalas diferentes e para evitar que aqueles com altas variâncias distorcessem a análise, utilizou-se a matriz de correlação para realizar a análise fatorial exploratória. A adequação da matriz de correlação para a análise fatorial foi verificada pelo teste de Kaiser-Meyer-Olkin (KMO), que indica se os fatores identificados na análise fatorial podem descrever adequadamente as variações nos dados originais. Também foi aplicado o teste de esfericidade de Bartlett, o qual testa se a matriz de correlação é uma matriz identidade, ou seja, se as variáveis não apresentam correlação significativa na população (hipótese nula).

A análise fatorial foi aplicada a partir da matriz de coeficientes de correlação. As cargas dos fatores foram estimadas pelo método de componentes principal. O número de fatores a serem retidos foi determinado pelo critério de Kaiser-Guttman (GUTTMAN; 1954; KAISER, 1960), ou seja, são retidos os fatores com autovalor maior que 1,0. Simultaneamente a esse critério, verificou-se a porcentagem de explicação acumulada dos fatores, tendo por objetivo reter fatores de modo que a explicação fosse superior a 60% (HAIR et al., 2009).

A escolha do tipo de rotação de fatores a ser realizada considerou a possibilidade de verificar as relações entre os fatores. A utilização de rotação ortogonal promove a independência entre os fatores, algo que em ciências agrárias é difícil de ocorrer. Já a rotação oblíqua mantém as relações entre os fatores, o que possibilita verificar o sentido das relações entre os fatores retidos. De acordo com Hair et al. (2009), em geral, as duas formas de rotação produzem resultados similares. Logo, a rotação *oblimin* (obliqua) foi utilizada, uma vez que é esperado uma relação de dependência entre os dados agronômicos, pois fatores como produtividade, caracteres morfológicos e resposta a condições meteorológicas frequentemente estão correlacionados. Em milho, por exemplo, fatores como altura da planta, produtividade, e radiação podem estar interligados, pois o desenvolvimento e o desempenho das plantas geralmente são determinados por condições ambientais.

Os escores dos fatores rotacionados para cada um dos genótipos foram estimados pelo método de regressão. Com base nesses escores dos genótipos em cada fator, e em cada data de semeadura, calculou-se a distância euclidiana média dos genótipos em relação a um ideótipo definido com o objetivo de maximizar os escores de todos os fatores retidos (quatro fatores). A partir disso, em cada data de semeadura, selecionaram-se os cinco genótipos de milho que apresentaram a menor distância euclidiana média em relação ao ideótipo. Assim, os genótipos selecionados foram caracterizados por apresentar os melhores valores dos caracteres avaliados. Todas as análises foram realizadas por meio do software R (R CORE TEAM, 2024).

2.3 Resultados e Discussão

As variáveis meteorológicas SSFM e SSFF apresentaram cargas cruzadas, ou seja, cargas fatoriais elevadas em dois fatores, por isso foram eliminadas da análise fatorial, conforme estabelecido por Hair et al. (2009). A análise de correlação de Pearson revelou quatro grupos distintos de correlações entre os caracteres avaliados (Figura 1). Os caracteres de altura, AP e AE, apresentaram correlação positiva entre si. Os caracteres produtivos, como DE, CE, ME e MG, também apresentaram correlações positivas. Outro grupo foi formado por proteína e aminoácidos (CP, LYS, MET, CYS, THR, TRP, VAL, ILE, PHE, HIS e ARG), que mostraram correlações entre si, o que sugere uma interação que pode influenciar a qualidade proteica. As variáveis meteorológicas (RSFM, RSFF, RFMC, RFFC, SFMC e SFFC) também apresentaram correlações positivas.

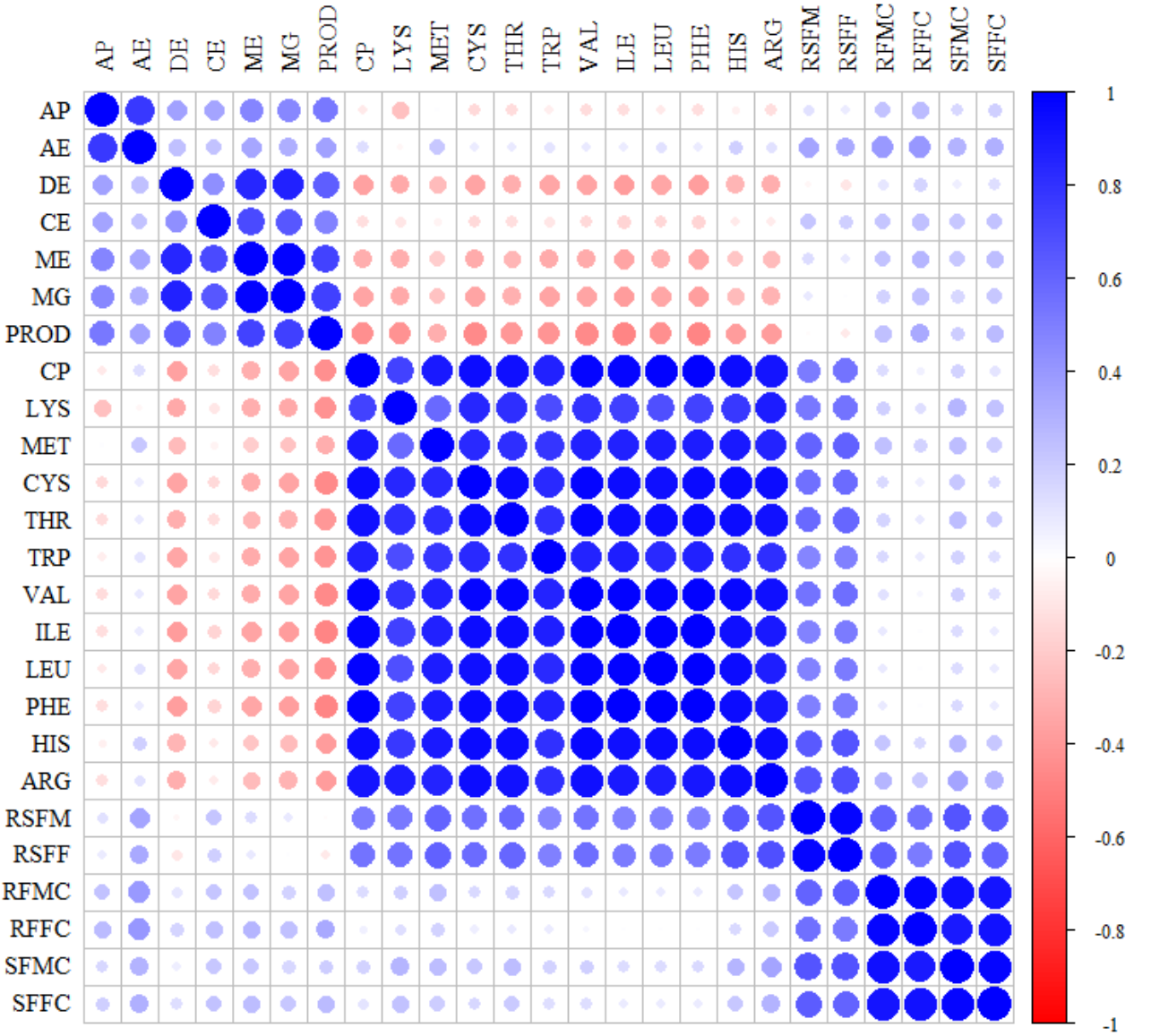

Figura 1. Matriz de coeficientes de correlação entre os caracteres fenológicos, morfológicos, produtivos, nutricionais e meteorológicos de 78 genótipos de milho avaliados em 10 datas de

semeadura nas safras 2021/2022 e 2022/2023, Santa Maria, RS, Brasil. Caracteres: altura de planta (AP, em cm), altura da inserção da espiga (AE, em cm), comprimento da espiga (CE, cm), diâmetro da espiga (DE, g), massa da espiga (ME, g), massa de grãos da espiga (MGE, g), produtividade de grãos (PROD, em Mg ha^{-1}), proteína bruta (CP, g/100 g), lisina (LYS, g/100 g), metionina (MET, g/100 g), cistina (CYS, g/100 g), treonina (THR, g/100 g), triptofano (TRP, g/100 g), valina (VAL, g/100 g), isoleucina (ILE, g/100 g), leucina (LEU, g/100 g), fenilalanina (PHE, g/100 g), histidina (HIS, g/100 g), arginina (ARG, g/100 g), radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), e soma térmica do florescimento feminino à colheita (SFFC, °C dia).

Coeficientes de correlação positivos foram observados entre os caracteres morfológicos e produtivos, morfológicos e meteorológicos, produtivos e meteorológicos e meteorológicos e nutricionais. Enquanto relações negativas foram observadas entre os caracteres morfológicos e nutricionais e produtivos e nutricionais. Essa estrutura de correlações indica que há interdependência entre caracteres que atuam em conjuntos funcionais específicos, sugerindo que o melhoramento de um grupo pode impactar diretamente os caracteres correlacionados. Todos os caracteres estão correlacionados com os demais, ou seja, não houve evidências de independência de caracteres. Isso indica que os caracteres utilizados podem ser mantidos para posterior análise fatorial.

Para a análise fatorial exploratória (AF), quanto maior o tamanho da amostra, mais adequada se torna para a realização de análises precisas. Segundo Hair et al. (2009), a amostra deve conter mais de 50 observações por caractere. Além disso, a relação entre o número de observações e a quantidade de caracteres deve ser superior a cinco para um, o que ajuda a fortalecer a confiabilidade das inferências feitas a partir dos dados. Neste estudo, a relação entre o número de observações e o número de caracteres foi de (773 observações/25 caracteres) 30,92 para 1, revelando a adequação do tamanho da amostra para a análise fatorial. O teste de Kaiser-Meyer-Olkin (KMO) apresentou um valor de 0,87, indicando que os fatores identificados na análise fatorial conseguem descrever de forma satisfatória as variações dos dados originais. Segundo Hair et al. (2009), valores acima de 0,50 já são considerados aceitáveis, e um valor de 0,87 é altamente desejável, pois sugere que a estrutura de fatores é adequada para representar as relações entre os caracteres. O teste de esfericidade de Bartlett foi significativo ($p < 0,05$, qui-quadrado = 56.460,15), rejeitando a hipótese nula que a matriz de correlação é uma matriz identidade, ou seja, as variáveis não são correlacionadas na população. Logo, existe relação suficiente entre os caracteres para aplicação da análise fatorial.

Os quatro primeiros fatores apresentaram autovalores acima de 1,00 e, juntos, explicaram 88% da variabilidade total dos dados (Tabela 1 e Figura 2). Logo, de acordo com Hair et al. (2009) fatores com autovalores acima de 1,00 podem ser utilizados para a análise fatorial. Portanto, foram utilizados os quatro primeiros fatores para representar a variabilidade dos dados. Bharathiveeramani, Prakash (2012) realizaram uma análise fatorial exploratória em 17 caracteres de milho, identificando cinco fatores significativos que representaram 78,3% da variância total.

Tabela 1. Cargas fatoriais associadas aos fatores extraídos pelo método de componentes principais, com rotação *oblimin*, a partir de variáveis avaliados em 78 genótipos de milho semeados em 10 datas de semeadura nas safras 2021/2022 e 2022/2023.

Caracteres	FA1	FA2	FA3	FA4	h^2
AP	-0,02	-0,03	0,13	**0,89**	0,89
AE	0,12	0,15	-0,02	**0,90**	0,88
DE	-0,07	-0,11	**0,87**	0,00	0,77
CE	0,12	0,05	**0,81**	-0,07	0,60
ME	-0,02	0,01	**0,96**	0,04	0,96
MG	-0,04	-0,04	**0,96**	0,03	0,95
PROD	-0,29	0,11	**0,56**	0,26	0,72
CP	**0,98**	-0,06	-0,05	0,07	0,96
LYS	**0,76**	0,17	0,00	-0,27	0,74
MET	**0,89**	0,04	0,00	0,14	0,83
CYS	**0,97**	0,00	-0,02	-0,05	0,95
THR	**0,97**	0,02	0,03	-0,06	0,94
TRP	**0,86**	-0,02	-0,09	0,06	0,78
VAL	**1,00**	-0,05	-0,01	-0,02	0,98
ILE	**0,98**	-0,09	-0,06	0,03	0,97
LEU	**0,99**	-0,12	-0,03	0,07	0,96
PHE	**0,99**	-0,09	-0,05	0,03	0,97
HIS	**0,97**	0,05	0,03	0,04	0,96
ARG	**0,92**	0,16	0,01	-0,08	0,94
RSFM	0,54	**0,58**	0,18	-0,01	0,78
RSFF	0,54	**0,59**	0,11	-0,02	0,77
RFMC	-0,06	**0,97**	-0,06	0,11	0,94
RFFC	-0,12	**0,95**	-0,02	0,12	0,92
SFMC	0,02	**0,98**	0,00	-0,04	0,95
SFFC	-0,04	**0,97**	0,04	-0,03	0,94
Var (%)	46,00	19,00	15,00	8,00	-
Var Ac (%)	46,00	65,00	80,00	88,00	-

h^2: comunalidade; FA1: fator 1, FA2: fator 2; FA3: fator 3; FA4: fator 4; Var: variância explicada; Var Ac: variância explicada acumulada. Caracteres: altura de planta (AP, em cm), altura da inserção da espiga (AE, em cm), comprimento da espiga (CE, cm), diâmetro da espiga (DE, g), massa da espiga (ME, g), massa de grãos da espiga (MGE, g), produtividade de grãos (PROD, em Mg ha^{-1}), proteína bruta (CP, g/100 g), lisina (LYS, g/100 g), metionina (MET, g/100 g), cistina (CYS, g/100 g), treonina (THR, g/100 g), triptofano (TRP, g/100 g), valina (VAL, g/100 g), isoleucina (ILE, g/100 g), leucina (LEU, g/100 g), fenilalanina (PHE, g/100 g), histidina (HIS, g/100 g), arginina (ARG, g/100 g), radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), e soma térmica do florescimento feminino à colheita (SFFC, °C dia).

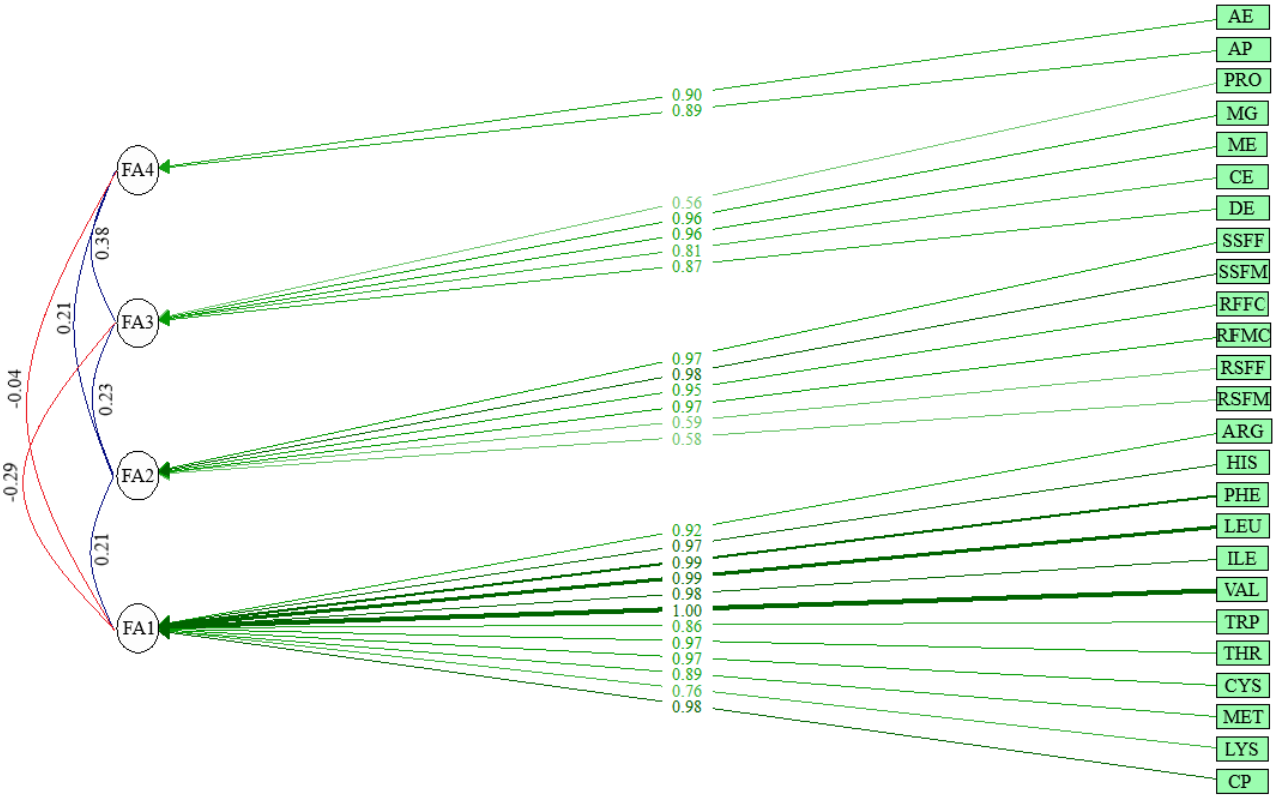

Figura 2. Cargas fatoriais das variáveis em cada fator e coeficientes de correlação linear de Pearson entre os fatores. FA1: fator 1, FA2: fator 2; FA3: fator 3; FA4: fator 4. Caracteres: altura de planta (AP, em cm), altura da inserção da espiga (AE, em cm), comprimento da espiga (CE, cm), diâmetro da espiga (DE, g), massa da espiga (ME, g), massa de grãos da espiga (MGE, g), produtividade de grãos (PROD, em Mg ha^{-1}), proteína bruta (CP, g/100 g), lisina (LYS, g/100 g), metionina (MET, g/100 g), cistina (CYS, g/100 g), treonina (THR, g/100 g), triptofano (TRP, g/100 g), valina (VAL, g/100 g), isoleucina (ILE, g/100 g), leucina (LEU, g/100 g), fenilalanina (PHE, g/100 g), histidina (HIS, g/100 g), arginina (ARG, g/100 g), radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), e soma térmica do florescimento feminino à colheita (SFFC, °C dia).

As cargas fatoriais dos caracteres indicam sua contribuição para cada fator (Tabela 1 e Figura 2). O fator FA1 foi relacionado ao teor de proteína e aminoácidos, pois apresenta altas cargas positivas em caracteres como CP, LYS, MET, CYS, THR, VAL, ILE, LEU e PHE. Assim, FA1 pode ser interpretado como uma dimensão ligada ao perfil nutricional e qualidade de proteínas/aminoácidos das amostras. O FA2 captou a influência das condições meteorológicas (radiação solar e soma

térmica) sobre o desenvolvimento das plantas nos subperíodos avaliados. Em termos práticos, FA2 pode indicar como as variações em radiação e temperatura ao longo do ciclo de desenvolvimento afetam o desempenho fenológico das plantas, podendo ser útil para prever a produtividade ou ajustar o manejo de acordo com as condições climáticas de cada subperíodo. Já o fator FA3 está relacionado a características produtivas, devido às altas cargas nas variáveis DE, CE, ME e MG. Isso permite interpretá-lo como um fator que reflete os aspectos produtivos das plantas. O fator FA4 associou-se à morfologia da planta, pois apresentou altas cargas em caracteres como AP e AE. Em estudo com 144 linhagens de milho, Bharathiveeramani, Prakash (2012) e Granate et al. (2001) revelaram que os caracteres morfológicos e produtivos também foram agrupados em diferentes fatores.

Cada fator agrupou caracteres que possuem um sentido comum, facilitando a interpretação: FA1 foi o fator mais relevante para a qualidade nutricional, FA2 para variáveis meteorológicas, FA3 para caracteres produtivos e FA4 para caracteres morfológicos. Esses fatores podem ser usados para classificar os genótipos em termos de adequação para diferentes fins, como aumento da qualidade nutricional, seleção para caracteres produtivos e avaliação do desenvolvimento. Os valores de comunalidade foram altos para a maioria dos caracteres, indicando que o modelo de análise fatorial explicou satisfatoriamente a variabilidade total dos caracteres.

Os coeficientes de correlação entre os fatores (FA1, FA2, FA3 e FA4) indicam como esses fatores relacionam-se entre si (Figura 2). O fator FA1 correlacionou-se negativamente com o fator FA3 (r = -0,29). Isso indica que plantas de milho com maior desempenho produtivo apresentam redução dos teores de proteína e aminoácidos nos grãos. Relações negativas entre caracteres produtivos e nutricionais proteicos tem sido observada em milho (ALVES et al., 2016; ALVES; CARGNELUTTI FILHO, 2017; GUO et al., 2022). Essa relação negativa implica que o aumento da produtividade está associado a uma redução na qualidade nutricional proteica dos grãos, o que representa um desafio na seleção de genótipos que equilibrem esses caracteres.

O fator FA2 correlacionou-se positivamente com os fatores FA1 (r = 0,21), FA3 (r = 0,23) e FA4 (r = 0,21), ou seja, o maior acúmulo de radiação solar global e soma térmica potencializam a expressão dos caracteres nutricionais proteicos, produtivos e morfológicos dos genótipos de milho. Coeficiente de correlação positivo (r = 0,38) também foi observado entre os fatores FA3 e FA4, indicando a dependência positiva entre caracteres produtivos e morfológicos. Loro et al. (2024) também observaram que o maior acúmulo de radiação e soma térmica promoveram incrementos no desempenho produtivo de milho, principalmente em semeaduras de outubro, novembro e dezembro. Os coeficientes de correlação encontrados entre os fatores evidenciam que a escolha por uma rotação oblíqua foi adequada (MATOS; RODRIGUES, 2019).

Em cada data de semeadura, cinco genótipos foram selecionados com base na menor distância euclidiana média em relação ao ideótipo agronômico estabelecido, que maximiza os escores dos

quatro fatores (FA1, FA2, FA3 e FA4). Esses fatores representam caracteres morfológicos, produtivos, nutricionais proteicos e meteorológicos, como o acúmulo de radiação solar global e a soma térmica (Tabela 2). Assim, os genótipos selecionados se destacam por apresentarem maior estatura de planta, melhores componentes produtivos, qualidade nutricional proteica dos grãos e maior acúmulo de radiação solar global nos estádios vegetativo e reprodutivo, além de maior soma térmica no estádio reprodutivo. A seleção baseada em vários caracteres tem sido aplicada para desenvolver genótipos de milho que apresentam um melhor equilíbrio entre os caracteres agronômicos. A seleção baseada apenas na produtividade de grãos apresenta os melhores ganhos, como esperado, mas promove fenótipos indesejáveis para as outros caracteres (CRISPIM-FILHO et al., 2020).

O genótipo Alvaré apresentou maior proximidade com o ideótipo em quatro das dez datas de semeadura, destacando-se por seu desempenho consistente nos fatores avaliados. A presença recorrente de Alvaré sugere uma ampla adaptação e estabilidade às variações ambientais entre as datas de semeadura. Outros genótipos, como BM3066, 36799 e Feroz, também se destacaram em várias datas, o que pode indicar uma resposta favorável a determinadas condições específicas de cultivo. Esses genótipos podem ser utilizados como base para novos programas de melhoramento, visando o aumento do desempenho produtivo e nutricional em novas cultivares.

Tabela 2. Classificação de genótipos com menor distância euclidiana média em relação ao ideótipo agronômico para cada data de semeadura, considerando os escores individuais dos genótipos nos quatro fatores da análise fatorial.

Ordem	21/09/2021	20/10/2021	20/11/2021	20/12/2021	30/01/2022
1º	BRS PLANALTO	BM3069	BM3063	B2401	FEROZ
2º	NS80	DKB290	P3565	BM3066	B2401
3º	ALVARÉ	36799	AS1633	B2620	36799
4º	AS1633	LG3040	NTX303	MG593	B2620
5º	IPR164	30A95	FS670	P3565	FS533
Ordem	06/09/2022	14/10/2022	24/11/2022	30/12/2022	06/02/2023
1º	ALVARÉ	MG652	SHSUPER	MG593	BM3066
2º	SHSUPER	CODIGO	M274	ALVARÉ	BM3063
3º	LG3040	MG699	DKB290	SHSUPER	36790
4º	MG699	P3565	MG618	DKB177	ALVARÉ
5º	FEROZ	BM270	BM270	AG1051	MG618

O diferencial de seleção (Ds, %) reflete o quanto os genótipos mais próximos do ideótipo superaram a média geral dos genótipos do experimento (Xo) em cada caractere avaliado (Tabelas 3 e 4). Em geral, observou-se ganhos para os caracteres morfológicos, produtivos, nutricionais proteicos e meteorológicos que podem beneficiar a produção e a qualidade do milho. Por exemplo, no caractere AP, o diferencial de seleção foi positivo em todas as datas, variando de 5,23% a 14,08% ao longo das datas de semeadura. Para PROD observou-se diferencial de seleção máximo de até 91,25%. Esses ganhos são significativos para a melhoria do desempenho dos genótipos, revelando a eficácia do

processo seletivo em identificar genótipos com potencial produtivo superior. Além disso, outras variáveis, como a concentração de MET, também mostram aumentos significativos, sugerindo um enriquecimento nutricional importante, especialmente em aminoácidos essenciais, o que é altamente desejável para a qualidade nutricional dos grãos de milho.

Tabela 3. Estimativas do diferencial de seleção nos caracteres avaliados em 78 genótipos de milho, cultivados em cinco datas de semeadura na safra 2021/2022, com base na distância euclidiana calculada a partir dos escores fatoriais.

VAR	21/09/2021			20/10/2021			20/11/2021			20/12/2021			30/01/2022		
	Xs	Xo	Ds	Xs	Xo	Ds	Xs	Xo	Ds	Xs	Xo	Ds	Xs	Xo	Ds
AP	227,56	200,60	13,44	186,12	175,22	6,22	205,32	185,21	10,86	208,80	195,11	7,02	199,88	179,94	11,08
AE	150,00	122,97	21,98	110,24	100,63	9,55	118,84	107,02	11,04	132,64	120,60	9,98	116,96	100,01	16,95
DE	4,24	4,30	-1,48	4,09	3,82	7,10	4,50	4,22	6,63	4,75	4,15	14,41	4,39	3,88	12,99
CE	15,70	14,91	5,28	13,65	13,62	0,28	17,58	15,84	11,00	14,37	14,39	-0,13	13,93	12,68	9,85
ME	126,61	125,66	0,75	112,53	92,29	21,93	172,85	130,83	32,12	173,98	120,82	44,00	120,14	80,66	48,95
MG	99,80	102,92	-3,03	87,43	69,47	25,85	132,59	101,43	30,73	145,22	100,08	45,10	99,62	65,37	52,39
PROD	4,98	5,63	-11,43	4,20	2,93	43,31	5,84	4,87	19,81	7,05	4,56	54,52	5,06	2,64	91,25
CP	9,42	8,24	14,26	10,05	9,57	4,95	8,94	8,70	2,72	8,98	8,86	1,40	8,91	8,76	1,64
LYS	0,27	0,25	8,49	0,29	0,28	2,61	0,26	0,26	0,42	0,26	0,25	0,80	0,24	0,25	-5,26
MET	0,18	0,16	10,46	0,19	0,18	4,07	0,17	0,17	0,91	0,17	0,17	0,99	0,17	0,16	2,46
CYS	0,21	0,19	12,41	0,23	0,21	5,62	0,20	0,20	3,26	0,20	0,20	1,63	0,20	0,20	0,94
THR	0,33	0,29	14,89	0,35	0,33	6,09	0,31	0,30	3,43	0,31	0,30	1,98	0,30	0,29	2,84
TRP	0,08	0,07	13,60	0,08	0,08	4,00	0,07	0,07	-0,07	0,07	0,07	0,29	0,07	0,07	1,37
VAL	0,48	0,42	14,45	0,51	0,49	5,29	0,46	0,44	3,25	0,45	0,45	1,38	0,45	0,44	1,58
ILE	0,34	0,29	16,77	0,36	0,34	6,82	0,32	0,31	4,09	0,31	0,31	0,38	0,32	0,31	1,74
LEU	1,17	0,97	21,40	1,28	1,19	7,99	1,11	1,04	5,94	1,11	1,08	2,33	1,13	1,07	5,12
PHE	0,50	0,42	18,49	0,54	0,51	6,76	0,47	0,45	3,23	0,47	0,46	0,91	0,47	0,46	2,67
HIS	0,27	0,24	11,97	0,29	0,28	4,77	0,26	0,25	2,27	0,26	0,26	1,91	0,25	0,25	2,49
ARG	0,44	0,40	9,92	0,47	0,46	2,90	0,43	0,42	0,57	0,42	0,42	1,47	0,39	0,40	-1,62
RSFM	1678,06	1658,32	1,19	1816,96	1797,04	1,11	1702,52	1673,96	1,71	1576,64	1628,30	-3,17	1201,97	1238,59	-2,96
RSFF	1690,08	1668,47	1,30	1884,48	1925,88	-2,15	1790,71	1757,72	1,88	1664,46	1661,29	0,19	1228,23	1257,65	-2,34
RFMC	1633,47	1622,83	0,66	1551,05	1544,81	0,40	1222,46	1209,45	1,08	1168,71	1079,17	8,30	822,39	797,83	3,08
RFFC	1621,44	1612,67	0,54	1483,53	1415,96	4,77	1134,27	1125,69	0,76	1080,89	1046,18	3,32	796,13	778,77	2,23
SFMC	1006,43	1000,82	0,56	1040,62	1029,93	1,04	803,15	803,90	-0,09	799,05	734,57	8,78	429,30	411,04	4,44
SFFC	996,57	992,28	0,43	988,53	950,09	4,05	749,41	751,51	-0,28	735,33	711,22	3,39	421,25	400,29	5,23

Média de todos os genótipos (Xo), média dos genótipos mais próximos ao ideótipo (Xs) e diferencial de seleção (em %). Caracteres: altura de planta (AP, em cm), altura da inserção da espiga (AE, em cm), comprimento da espiga (CE, cm), diâmetro da espiga (DE, g), massa da espiga (ME, g), massa de grãos da espiga (MGE, g), produtividade de grãos (PROD, em Mg ha^{-1}), proteína bruta (CP, g/100 g), lisina (LYS, g/100 g), metionina (MET, g/100 g), cistina (CYS, g/100 g), treonina (THR, g/100 g), triptofano (TRP, g/100 g), valina (VAL, g/100 g), isoleucina (ILE, g/100 g), leucina (LEU, g/100 g), fenilalanina (PHE, g/100 g), histidina (HIS, g/100 g), arginina (ARG, g/100 g), radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), e soma térmica do florescimento feminino à colheita (SFFC, °C dia).

Tabela 4. Estimativas do diferencial de seleção nos caracteres avaliados em 78 genótipos de milho, cultivados em cinco datas de semeadura na safra 2021/2022, com base na distância euclidiana calculada a partir dos escores fatoriais.

VAR	6/9/2022			14/10/2022			24/11/2022			30/12/2022			6/2/2023		
	Xs	Xo	Ds	Xs	Xo	Ds	Xs	Xo	Ds	Xs	Xo	Ds	Xs	Xo	Ds
AP	208,80	196,62	6,19	181,12	172,12	5,23	186,24	163,26	14,08	200,68	182,73	9,82	180,36	160,57	12,33

AE	133,20	117,15	13,70	122,64	111,56	9,93	107,28	93,73	14,46	112,80	97,57	15,61	97,04	83,97	15,57
DE	4,34	4,19	3,54	4,19	3,84	9,14	4,11	3,95	4,02	4,50	4,00	12,44	4,47	4,05	10,34
CE	15,50	14,81	4,69	14,60	13,07	11,74	15,40	14,34	7,39	14,12	12,87	9,69	13,76	13,11	4,94
ME	133,02	112,17	18,58	112,16	80,61	39,13	111,30	95,53	16,50	126,80	88,19	43,78	118,32	87,26	35,60
MG	112,70	94,94	18,71	92,54	66,82	38,49	90,46	78,58	15,12	109,74	76,38	43,68	99,98	73,74	35,58
PROD	4,02	4,39	-8,45	2,44	2,61	-6,40	3,35	2,84	17,99	3,45	2,74	26,06	4,68	3,24	44,57
CP	10,29	9,58	7,38	10,06	9,62	4,60	10,49	9,98	5,18	9,71	9,15	6,08	8,52	7,97	6,99
LYS	0,27	0,26	1,87	0,30	0,30	1,65	0,29	0,30	-0,34	0,28	0,27	2,79	0,26	0,24	5,53
MET	0,20	0,19	6,78	0,19	0,18	3,19	0,19	0,18	2,05	0,18	0,17	5,48	0,16	0,15	6,16
CYS	0,22	0,21	5,91	0,23	0,22	2,63	0,24	0,23	4,52	0,22	0,21	6,08	0,20	0,18	8,20
THR	0,35	0,32	6,63	0,37	0,35	3,96	0,38	0,36	4,52	0,35	0,33	5,00	0,31	0,27	11,32
TRP	0,09	0,08	8,54	0,09	0,08	7,33	0,09	0,08	5,64	0,08	0,08	7,68	0,07	0,07	4,80
VAL	0,51	0,48	7,26	0,53	0,51	5,10	0,56	0,53	6,67	0,52	0,48	8,16	0,44	0,40	9,70
ILE	0,37	0,34	8,37	0,37	0,35	5,17	0,40	0,37	7,90	0,37	0,33	9,30	0,31	0,28	9,60
LEU	1,34	1,20	11,06	1,29	1,23	5,11	1,41	1,29	9,10	1,25	1,16	8,21	1,06	0,94	12,72
PHE	0,55	0,51	8,62	0,56	0,53	5,49	0,59	0,55	7,91	0,54	0,50	8,15	0,46	0,41	10,98
HIS	0,29	0,27	5,86	0,30	0,28	4,23	0,30	0,29	3,68	0,28	0,26	4,96	0,24	0,22	8,88
ARG	0,46	0,45	3,73	0,48	0,47	2,39	0,49	0,48	1,82	0,45	0,44	2,69	0,41	0,38	8,59
RSFM	1776,86	1766,50	0,59	1882,66	1820,88	3,39	1892,56	1821,65	3,89	1586,35	1569,08	1,10	1208,61	1219,21	-0,87
RSFF	1832,91	1789,63	2,42	1957,12	1896,66	3,19	1947,25	1860,00	4,69	1610,31	1585,25	1,58	1223,67	1234,24	-0,86
RFMC	1340,78	1185,37	13,11	1383,31	1341,32	3,13	1054,08	1034,52	1,89	954,64	929,31	2,73	879,65	802,06	9,67
RFFC	1284,72	1162,24	10,54	1308,84	1265,54	3,42	999,39	996,17	0,32	930,68	913,10	1,93	864,59	787,02	9,86
SFMC	774,31	674,59	14,78	861,64	837,77	2,85	761,44	741,55	2,68	726,73	705,76	2,97	597,01	534,69	11,66
SFFC	742,19	661,20	12,25	820,98	794,51	3,33	727,02	717,55	1,32	707,85	691,94	2,30	585,03	522,89	11,88

Média de todos os genótipos (Xo), média dos genótipos mais próximos ao ideótipo (Xs) e diferencial de seleção (em %). Caracteres: altura de planta (AP, em cm), altura da inserção da espiga (AE, em cm), comprimento da espiga (CE, cm), diâmetro da espiga (DE, g), massa da espiga (ME, g), massa de grãos da espiga (MGE, g), produtividade de grãos (PROD, em Mg ha^{-1}), proteína bruta (CP, g/100 g), lisina (LYS, g/100 g), metionina (MET, g/100 g), cistina (CYS, g/100 g), treonina (THR, g/100 g), triptofano (TRP, g/100 g), valina (VAL, g/100 g), isoleucina (ILE, g/100 g), leucina (LEU, g/100 g), fenilalanina (PHE, g/100 g), histidina (HIS, g/100 g), arginina (ARG, g/100 g), radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), e soma térmica do florescimento feminino à colheita (SFFC, °C dia).

A identificação de uma correlação negativa inicial entre os caracteres produtivos e nutricionais proteicos (Figura 2) indica um desafio comum em programas de melhoramento: a tendência de que o aumento do desempenho produtivo, por exemplo, esteja associado a uma menor concentração de proteína e aminoácidos. Os genótipos selecionados mostram que, apesar da relação negativa, foi possível obter incremento em ambos os atributos. Esse resultado é promissor para programas de melhoramento, pois demonstra a viabilidade de identificar e selecionar genótipos que minimizam a correlação negativa entre esses caracteres. A análise fatorial é vantajosa nesse contexto, pois permite a seleção simultânea em vários caracteres, a partir de poucos fatores que representam grupos de caracteres altamente correlacionados, tornando o processo de seleção mais eficiente (GRANATE et al., 2001). Castoldi (1997), ao aplicá-la na seleção de famílias de meios-irmãos de milho comum, destacou que a análise fatorial pode ser uma ferramenta útil para a seleção simultânea de múltiplas características. A utilização da análise fatorial e da distância euclidiana para identificar proximidade ao ideótipo permite uma seleção mais precisa e científica dos genótipos, orientando o melhoramento com base em um conjunto equilibrado de caracteres de interesse. Isso acelera o desenvolvimento de variedades de milho que maximizam o rendimento e qualidade em diferentes condições de cultivo.

2.5 Conclusão

Os caracteres são agrupados nos seguintes grupos: morfológicos, produtivos, nutricionais proteicos e meteorológicos.

Os caracteres produtivos e nutricionais proteicos correlacionam-se negativamente, enquanto os caracteres morfológicos e meteorológicos correlacionam-se positivamente entre si e com os caracteres produtivos e nutricionais proteicos.

É possível identificar genótipos que incrementam simultaneamente caracteres morfológicos, produtivos, nutricionais proteicos e meteorológicos.

2.6 Referências

ALVARES, C. A. *et al*. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, p. 711-728, 2013.

ALVES, B. M.; CARGNELUTTI FILHO, A. Linear relationships between agronomic and nutritional traits in transgenic genotypes of maize. **Journal of Cereal Science**, v. 76, p. 35-41, 2017.

ALVES, B. M.; CARGNELUTTI FILHO, A.; BURIN, C.; TOEBE, M. Correlações canônicas entre caracteres agronômicos e nutricionais proteicos e energéticos em genótipos de milho. **Revista Brasileira de Milho e Sorgo**, v. 15, n. 2, p. 171-185, 2016.

ARNOLD, C. Y. Maximum-minimum temperatures as a basis for computing heat units. **Journal of the American Society for Horticultural Sciences**, v. 76, n. 1, p. 682-692, 1960.

BHARATHIVEERAMANI, B.; PRAKASH, M. Factor analysis for yield contributing traits in maize (Zea mays L.). **Electronic Journal of Plant Breeding**, v. 3, n. 4, p. 998-1001, 2012.

BUTTS-WILMSMEYER, C. J. *et al*. Weather during key growth stages explains grain quality and yield of maize. **Agronomy**, v. 9, n. 1, p. 16, 2019.

CASTOLDI, F. L. Comparação de métodos multivariados aplicados na seleção em milho. 1997. Tese (Doutorado em Genética e Melhoramento) - Universidade Federal de Viçosa, Viçosa, 1997.

COSTELLO, A. B.; OSBORNE, J. W. Best practices in exploratory factor analysis: Fourre commendations for getting the most from your analysis. **Practical Assessment, Research & Evaluation**, v. 10, n.7, p.1-9, 2005.

CRISPIM-FILHO, A. J. *et al.* Dealing with multiple traits in maize: A new approach for selecting progenies. **Crop Science**, v. 60, n. 6, p. 3151-3165, 2020.

DHALIWAL, D. S.; WILLIAMS, M. M. Understanding variability in optimum plant density and recommendation domains for crowding stress tolerant processing sweet corn. **PloS One**, v. 15, n. 2, p. e0228809, 2020.

FANCELLI, A. L.; DOURADO NETO, D. **Milho: manejo e produtividade**. Piracicaba: ESALQ/USP. 2009, 181p.

FIELD, A.; MILES, J.; FIELD, Z. Discovering statistics using R. Sage Publications, 2012.

GUO, J.; QU, L.; WANG, L.; LU, W.; LU, D. Effects of post silking drought stress degree on grain yield and quality of waxy maize. **Journal of the Science of Food and Agriculture**, v. 103, n. 3, p. 1530-1540, 2023.

GUTTMAN, L. Some necessary conditions for common factor analysis. **Psychometrika**, v.19, p.149-162, 1954.

HONGYU, Kuang. Análise Fatorial Exploratória: resumo teórico, aplicação e interpretação. **E&S Engineering And Science**, v. 7, n. 4, p. 88-103, 2018.

KAISER, H.F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, v.23, p.187-200. 1958.

LORO, M. V.; CARGNELUTTI FILHO, A.; ORTIZ, V. M.; ANDRETTA, J. A. Relações lineares entre variáveis meteorológicas e caracteres fenológicos, morfológicos e produtivos em bases genéticas de milho. **Revista Vivências**, v. 20, n. 41, p. 95-111, 2023.

MATOS, D. A.; RODRIGUES, E. C. **Análise fatorial**. Brasília: Enap, 2019. 74 p.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria. 2024.

RODRIGUEZ, D. A. *et al*. Digestibility of amino acids, fiber, and energy by growing pigs, and concentrations of digestible and metabolizable energy in yellow dent corn, hard red winter wheat, and sorghum may be influenced by extrusion. **Animal Feed Science and Technology**, v. 268, n. 114602, p. 1-11, 2020.

SANTOS, H. G. *et al*. **Sistema Brasileiro de Classificação de Solos**. 5. ed. Brasília: Embrapa, 2018, 356p.

SIMÕES, C. T. et al. A. Assessment of field traits, nutrient composition and digestible amino acids of corns with different endosperm textures for poultry and swine. **Animal Feed Science and Technology**, v. 295, p. 115510, 2023.

SRIPERM, N.; PESTI, G. M.; TILLMAN, P. B. The distribution of crude protein and amino acid content in maize grain and soybean meal. **Animal Feed Science and Technology**, v. 159, n. 3-4, p. 131-137, 2010.

WADE, J. *et al.* Improved soil biological health increases corn grain yield in N fertilized systems across the Corn Belt. **Scientific Reports**, v. 10, n. 1, p. 3917, 2020.

2.7 Script em R

```
library(readxl)
dados <- read_excel("12-Milho_2021-2022-2023_Definitivo_Dissertacao_Murilo - An
alises Alberto + Conferencia Morfologicos-Definitivo-Tese.xlsx",
              sheet = "AnaFat", range = "A1:AN774")
attach(dados)

#-----------------------------------PASSO 2 - MATRIZ DE CORRELAÇÕES
library(RColorBrewer)
library(corrplot)
```

```r
library(ggcorrplot)

matcor <- cor(dados[-1:-15])
print(matcor, digits = 2)

#Figura de correlação
custom_colors <- colorRampPalette(c("red", "white", "blue"))(200)
par(family = "serif")
corrplot(matcor,
         method = "circle",
         col = custom_colors,
         tl.col = "black")


#com heat map
col <- colorRampPalette(c("red", "white", "blue"))(20)
heatmap(x = matcor, col = col, symm = TRUE)

#----------PASSO 3 - VERIFICAR A ADEQUAÇAO DA MATCOR PARA ANALISE FATORIAL


#Podem ser utilizados os seguintes testes:
#  * Teste de esfericidade de Bartlett
#  * Kaiser-Meyer-Olkin (KMO)

#Ho: A matriz de correlação da população é uma matriz identidade, ou seja as
#variáveis não são correlacionadas na população.

#H1: A matriz de correlação da população não é uma matriz identidade, ou seja
#as variáveis são correlacionadas na população.

psych::cortest.bartlett(dados[-1:-15])
psych::KMO(dados[-1:-15])

#----------PASSO 4 - VERIFICAR A VARIâNCIA EXPLICADA POR CADA FATOR
fit<-princomp(dados[-1:-15],cor=TRUE)
fit

summary(fit)

screeplot(fit)
plot(fit,type="lines")

#----------PASSO 4 - DETERMINAÇÃO DO NUMERO DE FATORES

# Sem rotação
FatPCA <- principal(dados[-1:-15], nfactors=4,
                    n.obs=773,rotate="none", scores=TRUE)
FatPCA

#Rotação ortogonal (varimax) - fatores são independentes
PCAvarimax <- principal(dados[-1:-15], nfactors=4,
                            n.obs=773,rotate="promax",scores=TRUE)
PCAvarimax
PCAvarimax$values #acessar autovalores
PCAvarimax$loadings #contribuição de cada variável (cargas fatoriais)
```

```r
biplot(PCAvarimax)  #grafico biplot

#Rotação obliqua (oblimin) - fatores são correlacionados
PCAoblimin <- principal(dados[-1:-15], nfactors=4,
                        n.obs=773,rotate="oblimin",scores=TRUE)
PCAoblimin
PCAoblimin$values #acessar autovalores
PCAoblimin$loadings #contribuição de cada variável (cargas fatoriais)
biplot(PCAoblimin)  #grafico biplot


# verificando cada genótipo
escores <- factor.scores(dados[-1:-15],PCAoblimin,
              Phi = TRUE,
              method = c("Thurstone"),
              rho=NULL)

gen <- escores[["scores"]]
#, "tenBerge", "Anderson", "Bartlett", "Harman","components"

library(psych)
fa.diagram(PCAoblimin, digits = 2, rsize = 5, cex = 0.2)

# Carregar o pacote semPlot
library(semPlot)

# Supondo que `PCAoblimin` é o modelo de análise fatorial ajustado
# Extrair a matriz de cargas do modelo
loadings_matrix <- PCAoblimin$loadings

# Criar uma nova matriz de cargas, zerando valores abaixo de 0.5
loadings_matrix[abs(loadings_matrix) < 0.55] <- 0

# Reatribuir a matriz de cargas ao objeto PCAoblimin
PCAoblimin$loadings <- loadings_matrix

# Agora, plotar com semPaths usando a nova matriz de cargas
semPaths(PCAoblimin, "par",edge.label.cex = 1.3,
         sizeMan = 4, sizeLat = 4, shapeInt = 5,
         rotation = 2, layout = "tree",
         color = list(
  lat = rgb(253, 253, 253, maxColorValue = 255),
  man = rgb(155, 253, 175, maxColorValue = 255)),
  mar = c(1, 1, 1, 1))

#------------------------GRAFICO DE CORRELAÇÃO ENTRE FATORES
dadoscor <- read_excel("12-Milho_2021-2022-2023_Definitivo_Dissertacao_Murilo -
Analises Alberto + Conferencia Morfologicos-Definitivo-Tese.xlsx",
                       sheet = "AnaFat", range = "AO1:AR774")

matcor1 <- cor(dadoscor)

custom_colors <- colorRampPalette(c("red", "white", "blue"))(200)
par(family = "serif") # Fonte times new roman

corrplot::corrplot(matcor1,
```

```
                    method = "circle",
                    col = custom_colors,
                    tl.col = "black",
                    number.digits = 2,
                    addCoef.col = "black", # Exibe os coeficientes na cor preta
                    number.cex = 0.8)
```

QUESTÕES PROFESSOR DR. FERNANDO MACHADO HAESBAERT

Questão 1. Em seu trabalho de doutorado foram abordados temas importantes relacionados ao estudo de relações entre variáveis. Assim solicito: 1) Considerando a sua experiência na pesquisa agrícola, proponho que prepare uma aula com tempo de 50 minutos, em nível de graduação em Agronomia, com o tema "Modelos de Predição". Nesta aula explores as diferentes possibilidades de uso dos modelos de predição para estudos agronômicos e explique a importância de cada modelo, de forma que os estudantes de graduação compreendam as diferenças fundamentais entre abordagens lineares, não lineares e métodos de ensemble. Pode utilizar os seus dados de caracteres agronômicos e nutricionais com variáveis meteorológicas de bases genéticas de milho para exemplificar os modelos. Explore modelos como regressão linear, classificação linear (SVM linear), regressão logística, máquinas de vetores de suporte, árvores de regressão, Random Forests, Gradient Boosting e redes neurais. Para isto solicito: a) Um plano de aula: considere o uso de metodologias ativas de ensino-aprendizagem. **Resposta: 3. PLANO DE AULA.** b) Uma apostila (escreva tudo o que falaria na aula) – incluir os recursos que utilizaria em sala de aula; justifique com fundamentações técnicas e científicas e referências bibliográficas. Inclua os códigos comentados das análises realizadas por meio do software R. **Resposta: 4. APOSTILA** c) Uma videoaula (poste no YouTube, pode ser no modo não listado ou no Google Drive) e disponibilize o link. Dê atenção a edição do vídeo. **Resposta: 5. LINK VIDEOAULA.** d) Exercícios (tema de casa para os alunos bem como o gabarito detalho de cada questão). **Resposta: 6. EXERCÍCIO COM GABARITO.**

Questão 2. A Análise de Componentes Principais (PCA) é utilizada para reduzir a dimensionalidade dos dados e facilitar a interpretação dos principais padrões de variação. No entanto, quando se trabalha com variáveis de diferentes naturezas, como agronômicas, nutricionais e meteorológicas, a análise de componentes principais tradicional (ajustada pelo método de Mínimos Quadrados Ordinários) podem ter interpretações complexas. Pergunto: Como interpretar corretamente os componentes principais quando as variáveis envolvem diferentes escalas, unidades de medidas, características de variáveis (qualitativas, quantitativas)? Quais estratégias podem ser adotadas para evitar vieses na interpretação dos componentes principais e garantir que as variáveis com maior variância não dominem os componentes, considerando a natureza física das variáveis? **Resposta: 7. ANÁLISE DE COMPONENTES PRINCIPAIS**. Discuta a possibilidade de usar a Análise de Componentes Principais ACP por Método Generalizado dos Momentos, Análise de Componentes Principais Robusta, Análise de Componentes Principais Não Linear (Kernel PCA), Análise de Componentes Principais Generalizada, *Minimum Classification Error* PCA. Crie variáveis para comparar os modelos, discuta as aplicações, vantagens e desvantagens de cada método e compare os componentes principais obtidos em cada método e explique como eles diferem. **Resposta 8. VARIAÇÕES DA ANÁLISE DE COMPONENTES PRINCIPAIS**

Questão 3. A análise de correlação entre variáveis agronômicas, nutricionais e meteorológicas é central para compreender as relações entre essas dimensões. No entanto, a correlação linear de Pearson, muitas vezes, pode ser insuficiente para capturar relações não lineares. Pergunto: - Em que situações a correlação de Pearson poderia fornecer uma visão incompleta das relações entre os caracteres agronômicos e nutricionais e as variáveis meteorológicas? Descreva pelo menos dois cenários em que Pearson não seria o método ideal, considerando relações não lineares ou a presença de outliers. **Resposta: 9. RELAÇÕES NÃO LINEARES E PRESENÇA DE OUTLIERS.** - Quais alternativas metodológicas poderiam ser adotar para capturar relações não lineares? Discuta as características e diferenças entre a correlação de Spearman, a correlação de Kendall e o Coeficiente de Máxima Informação (MIC), explicando como cada uma lida com padrões monotônicos e não lineares. **Resposta: 10. AVALIAÇÃO DE PADRÕES NÃO LINEARES E MONOTÔNICOS.** - Analise a possibilidade de uso da correlação canônica como uma forma de identificar padrões

multivariados. Compare-a com os métodos de correlação não linear (Spearman, Kendall, MIC), discutindo quando cada abordagem é mais adequada para explorar relações entre variáveis de diferentes dimensões. **Resposta: 11. CORRELAÇÃO CANÔNICA, SPEARMAN, KENDALL E MIC.** - Simule dados representativos de cada um dos cenários discutidos, por exemplo: Uma relação linear para a correlação de Pearson; uma relação monotônica não linear para a correlação de Spearman e Kendall; uma relação complexa não linear para o MIC. Análise de Resultados: Para cada método, calcule os coeficientes de correlação e interprete os resultados, destacando como cada método captura (ou falha em capturar) as diferentes relações entre as variáveis. Discuta as vantagens e limitações de cada método com base nos resultados obtidos. **Resposta: 12. AVALIAÇÃO DOS MÉTODOS EM CENÁRIOS LINEARES, MONOTÔNICA NÃO LINEAR E COMPLEXA.**

3. PLANO DE AULA

UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS RURAIS
DEPARTAMENTO DE FITOTECNIA
PROFESSOR: MURILO VIEIRA LORO

PLANO DE AULA

a) Identificação do tema
- Modelos de predição

b) Desenvolvimento do tema
1. Importânica dos modelos de predição 2. Teoria e aplicações dos modelos de predição 2.1 Regressão linear 2.2 Regressão Logística 2.3 Árvores de Regressão 2.4 Random Forest 2.5 Gradient Boosting 2.6 Máquinas de Vetores de Suporte 2.7 Redes Neurais 3. Considerações finais

c) Lista de exercícios
1) Utilizando o banco de dados "dados", que contém as variáveis MG, RSFM, RSFF, RFMC, RFFC, SFMC e SFFC, o objetivo é criar modelos preditivos que estimem a variável MG com base nas variáveis explicativas RSFM, RSFF, RFMC, RFFC, SFMC e SFFC. Para isso, utilize os algoritmos Árvore de Regressão, Random Forest e Gradient Boosting. Avalie o desempenho de cada modelo. Compare os resultados obtidos por cada abordagem, considerando métricas como o raiz do erro quadrático médio (RMSE) e o coeficiente de determinação (R^2).

d) Identificação dos pré requesitos
- Conceitos básicos de estatística descritiva - Conceitos de biometria e análise de dados. Experiência prática com a linguagem de programação R.

e) Modo de avaliar o aprendizado
Exercícios práticos no sofware para avaliar o processo de ensino-aprendizagem.

f) Objetivos
- Compreender o que são e a importância de modelos de predição; - Discutir a teoria e as diferenças entre os modelos de predição; - Compreender a aplicação dos modelos de predição;

g) Referências
Bibliografia básica: FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. 1. ed. Rio de Janeiro: LTC, 2011. IZBICKI, R.; SANTOS, T. M. Aprendizado de máquina: uma abordagem estatística. 2020. Disponível em http://www.rizbicki.ufscar. br/AME.pdf FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. The elements of statistical learning. New York, NY, USA: Springer Series in Statistics, 2001. Bibliografia complementar: MORETTIN, P. A.; SINGER, J. M. Estatística e ciência de dados. 2022. JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An introduction to statistical learning (with applications in R). New York: Springer, 2013.

4. APOSTILA

UNIVERSIDADE FEDERAL DE SANTA MARIA - UFSM
CENTRO DE CIÊNCIAS RURAIS
DEPARTAMENTO DE FITOTECNICA

MODELOS DE PREDIÇÃO

Autor: Murilo Vieira Loro

Santa Maria, RS.

4.1 Introdução aos modelos de predição

4.1.1 O que são modelos de predição?

A apostila Modelos de Predição teve como objetivo apresentar os principais modelos de predição aplicados em estudos agronômicos, proporcionando aos estudantes uma compreensão das técnicas disponíveis. Através da utilização do software R, foram fornecidos exemplos práticos e códigos comentados para análise de modelos, enriquecendo o aprendizado em sala de aula. Os modelos de predição em estudos agronômicos, permitem a predição de variáveis de interesse com base em dados coletados. A compreensão e aplicação desses modelos são importantes para a tomada de decisão e desenvolvimento de estratégias na agricultura. Por meio deste material, os estudantes terão a oportunidade de compreender a relevância e benefícios dos modelos de predição em contextos agronômicos, preparando-os para a aplicação prática dessas técnicas. O principal objetivo da apostila é promover a compreensão lógica e aplicação de modelos de predição em estudos agronômicos, utilizando o software R como ferramenta principal. Além disso, a apostila possibilitará a exploração de diferentes modelos, proporcionando uma visão ampla das possibilidades de aplicações e relevâncias de cada abordagem.

Em estudos agronômicos, principalmente em avaliações de plantas, é comum a análise de um grande número de variáveis. A produtividade de grãos costuma ser o principal interesse, pois é controlada por múltiplos genes e altamente influenciada por fatores ambientais. Por essa razão, outras variáveis, como altura da planta, massa da espiga e dias para o florescimento, também são frequentemente avaliadas, pois estão relacionadas à produtividade e podem ser utilizados no processo de seleção para identificar genótipos produtivos que possuam características adicionais desejáveis. Além disso, variáveis meteorológicas são utilizadas para compreender como os caracteres agronômicos respondem às variações ambientais (LORO et al., 2024b).

Os dados gerados em experimentos agronômicos são geralmente volumosos e apresentam relações complexas, o que dificulta uma análise direta. Identificar os caracteres associados ao desempenho agronômico ou as variáveis meteorológicas que influenciam a produtividade é um dos principais desafios. Para isso, os modelos de predição são amplamente utilizados para predizer uma variável dependente com base em variáveis independentes. Por exemplo, é possível utilizar a massa da espiga de milho para predizer a massa de grãos (LORO et al., 2024a). Essa abordagem pode reduzir significativamente o tempo e o esforço necessários para o trabalho do melhorista, permitindo a seleção de plantas de alto desempenho apenas com a pesagem das espigas, sem a necessidade de debulhar os grãos. Esse exemplo ilustra a relevância dos modelos de predição na área agronômica.

Em outro exemplo, o pesquisador pode estar interessado em verificar se é possível predizer a composição de proteína nos grãos de milho por meio de variáveis meteorológicas. Para isso, podem ser utilizados alguns modelos de predição e identificar quais são as variáveis meteorológicas que determinam a proteína nos grãos de milho. Portanto, vários problemas podem ser solucionados por meio da modelagem preditiva. Em situações em que há uma relação de linearidade das variáveis, os modelos de regressão linear simples e múltipla, por exemplo, são eficientes para realizar predições com alta precisão (HAIR et al., 2009; BUSSAB; MORETTIN, 2017). Modelos lineares são aqueles em que a relação entre as variáveis dependente e independente é modelada como uma combinação linear dos preditores (HAIR et al., 2009).

No entanto, em algumas situações, principalmente em experimentos agronômicos, as variáveis apresentam relações não lineares. Por exemplo, o aumento da radiação solar global pode inicialmente elevar a produtividade de grãos até atingir um platô, após o qual a produtividade pode diminuir. Nesses casos, os modelos lineares, como a regressão linear, apresentam limitações e baixa capacidade preditiva. Modelos não lineares, por sua vez, são mais adequados para capturar essas relações complexas. Exemplos incluem Árvores de Regressão, *Random Forest*, *Gradient Boosting*, Máquinas de Vetores de Suporte (SVM) e Redes Neurais, que oferecem maior capacidade preditiva em conjuntos de dados complexos (BREIMAN, 2001). Entre os modelos não lineares, *Random Forest* e *Gradient Boosting* são conhecidos como métodos ensemble. Esses métodos, combinam várias predições de diferentes modelos base para melhorar o desempenho preditivo final. O *Random Forest* combina múltiplas árvores de decisão para desenvolver uma predição robusta, enquanto no *Gradient Boosting*, as árvores de decisão são construídas sequencialmente, com cada árvore corrigindo os erros de predição anteriores.

Além de predizer valores contínuos, há situações em que o objetivo é identificar, por exemplo, se um genótipo tem baixo ou alto valor nutricional. Para isso, o modelo como a Regressão Logística, calcula probabilidade de um genótipo pertencer a uma classe de baixa ou alta qualidade nutricional. Portanto, vários são os modelos de predição que podem ser utilizados na modelagem de dados. O objetivo dessa apostila é fornecer uma teoria da lógica de funcionamento de cada modelo, associado com a demonstração da aplicação no software R.

4.2 Teoria dos modelos de predição

Os dados utilizados nos exemplos práticos de cada modelo preditivo podem ser obtidos no repositório do GitHub, da seguinte forma:

```
# URL do arquivo no GitHub
url <- "https://github.com/muriloloro/Modelos-de-Predicao/raw/main/dados.xlsx"
```

```
library(httr)
library(readxl)


# Definir o caminho para salvar o arquivo
file <- tempfile(fileext = ".xlsx")
GET(url, write_disk(file, overwrite = TRUE))
```

Os dados referem-se a variáveis agronômicas, nutricionais proteicas e meteorológicas avaliadas em 78 genótipos de milho. As variáveis são as seguintes: altura de planta (AP, cm), diâmetro da espiga (DE, cm), comprimento da espiga (CE, cm), massa da espiga (ME, g), massa de grãos da espiga (MG, g), produtividade de grãos (PRO, Mg ha^{-1}), proteína bruta (CP, g/100g), radiação solar global acumulada entre a semeadura e o florescimento masculino (RSFM, MJ m^{-2}), radiação solar global acumulada entre a semeadura e o florescimento feminino (RSFF, MJ m^{-2}), radiação solar global acumulada entre o florescimento masculino e a colheita (RFMC, MJ m^{-2}), radiação solar global acumulada entre o florescimento feminino e a colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino a colheita (SFMC, ºC dia), soma térmica do florescimento feminino a colheita (SFFC, º dia) e qualidade nutricional proteica (NUT, alta qualidade nutricional = 1, baixa qualidade nutricional = 0).

4.2.1 Regressão linear

4.2.1.1 Teoria

A regressão linear é uma técnica de modelagem que analisa a relação entre uma variável dependente contínua e uma ou mais variáveis independentes (HAIR et al., 2009). Para introduzir o conceito de regressão linear, será considerado um exemplo prático de um programa de melhoramento genético de milho que enfrenta um desafio: selecionar as progênies com maior MG. Para avaliar a MG é necessário a colheita, debulha e pesagem dos grãos de cada espiga. Geralmente, há um grande número de progênies a serem avaliadas, isso demanda muito tempo e mão de obra. Logo, o pesquisador tem por objetivo verificar se é possível desenvolver um modelo para predizer a MG com base em uma variável de fácil mensuração, como a ME. Para isso, o melhorista decide usar a regressão linear simples e avalia a MG e a ME de 78 genótipos de milho, para verificar se há uma relação entre essas duas variáveis. Se uma relação linear consistente for encontrada, o modelo poderá facilitar o trabalho, permitindo que a MG seja estimada apenas pela ME (Exemplo 1).

Portanto, percebe-se que no modelo de regressão linear simples, assume-se uma relação linear direta entre a variável de interesse (MG) e uma variável preditora (ME). Chama-se de variável dependente y, aquela cujo resposta será explicada pela variável x, chamada de variável independente ou explicativa (BUSSAB; MORETTIN, 2017). A ideia é bastante simples, é estimar a equação de

uma reta. Essa equação é descrita como $y_i = a + bx_i + e_i$. O objetivo é calcular os valores de a e b. Existem alguns métodos para ajustar uma reta entre as variáveis X e Y, e o mais utilizado é denominado método dos mínimos quadrados (MMQ) (BUSSAB; MORETTIN, 2017). Esse método exige que os estimadores a e b sejam escolhidos de tal forma que a soma dos quadrados dos desvios dos mesmos em relação à reta de regressão ajusta seja mínima. Os valores de b e a são obtidos por meio das seguintes expressões:

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} \quad a = \frac{\sum Y - b(\sum X)}{n}$$

ou

$$b = \frac{Cov\,(X,Y)}{V(X)} \quad a = \bar{Y}i - b\bar{X}i$$

Agora, para a regressão linear múltipla, será considerado um cenário em que o melhorista deseja melhorar ainda mais essa predição. Além da ME, inclui outras variáveis de fácil avaliação, como o CE e o DE. Assim, é possível ajustar um modelo de regressão linear múltipla que utiliza essas três variáveis (ME, CE e DE) para predizer a MG com maior precisão (Exemplo 2). Esse modelo pode ajudar a identificar progênies promissoras de maneira ainda mais eficiente, sem a necessidade de debulhar as espigas.

Portanto, percebe-se que a regressão linear múltipla é uma técnica em que se utiliza mais de uma variável independente para tentar predizer uma variável dependente (HAIR et al., 2009). A equação da regressão linear múltipla é similar com a da regressão linear simples, mas inclui outras variáveis preditoras, resultando na seguinte equação: $y_i = a + b_1x_1 + b_2x_2 + ... + b_nx_n$, onde cada b é um coeficiente associado a uma variável independente x. Após calcular os coeficientes na regressão linear (simples ou múltipla), aplica-se um teste de hipótese para verificar se esses coeficientes são estatisticamente significativos, ou seja, se realmente contribuem para a predição da variável dependente. Além disso, avalia-se a qualidade do ajuste do modelo usando o coeficiente de determinação (R^2), que indica a porcentagem da variação nos dados explicada pelas variáveis independentes. Logo, quanto maior o R^2, melhor o modelo explica os dados (BUSSAB; MORETTIN, 2017).

A partir da análise de regressão é possível realizar predições sobre uma variável dependente com base em uma ou mais variáveis independentes. Contudo, é importante ressaltar que a regressão linear parte do princípio de que as relações entre as variáveis são lineares. Assim, quando não há uma relação linear, o modelo não consegue identificar a relação entre as variáveis, o que não significa que elas não estejam relacionadas, apenas que essa relação não é linear.

4.2.1.2 Aplicação em R

Serão realizadas aplicações da análise de regressão linear simples e múltipla no software R considerando os dados dos exemplos 1 e 2, respectivamente. Considera-se o exemplo 1, no qual o melhorista pretende verificar se é possível predizer a MG por ME. A análise de regressão linear simples pode ser realizada por meio da função *lm*, nativa do software R. Observa-se que o objetivo foi predizer a MG por meio da ME.

```
summary(lm(MG~ME, dados))

##
## Call:
## lm(formula = MG ~ ME, data = dados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.488  -3.345   2.178   3.859  11.667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.95059    2.64410  -7.545 8.13e-11 ***
## ME            0.96894    0.02732  35.470  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.043 on 76 degrees of freedom
## Multiple R-squared:  0.943,  Adjusted R-squared:  0.9423
## F-statistic:  1258 on 1 and 76 DF,  p-value: < 2.2e-16
```

Neste exemplo, o intercepto (-19,95059) e o coeficiente angular (0,96894) foram significativos a 5% de probabilidade pelo teste t de *Student*. Logo, o aumento de 1 grama na ME promove o aumento de 0,96894 gramas da MG, ou seja, há uma relação linear positiva entre as duas variáveis. O coeficiente de determinação foi de 0,943. Isso indica que 94,30% da variação da MG foi explicada pela ME, ou seja, revela um ótimo ajuste do modelo.

Além de analisar a significância dos coeficientes e o coeficiente de determinação, o pesquisador pode avaliar a validade do modelo utilizando a validação cruzada. A validação cruzada é um método que divide o conjunto de dados em várias partes. O modelo é treinado em uma parte dos dados e testado em outra, repetindo o processo várias vezes com diferentes divisões dos dados. Esse procedimento ajuda a avaliar a capacidade do modelo de fazer predições confiáveis em dados novos, ou seja, em situações fora do conjunto de dados original.

Por exemplo, o pesquisador pode utilizar a validação cruzada do tipo *k-fold*, onde o conjunto de dados é dividido em *k* partes (ou *folds*). O modelo é treinado em k-1 *folds* e usa o *fold* restante para

testar o modelo, repetindo o processo *k* vezes, cada vez com uma parte diferente dos dados sendo usada para teste. No final, os erros de predição em cada rodada são calculados e avaliados, indicando se o modelo é robusto e capaz de generalizar. Para avaliar a validade do modelo de forma ainda mais rigorosa, o pesquisador pode utilizar a validação cruzada *leave-one-out*. Nesse método, cada observação é utilizada como um teste individual, enquanto as demais são usadas para treinar o modelo. Assim, o conjunto de dados com n observações é dividido em n partes, e em cada rodada o modelo é treinado em n-1 observações e testado na observação restante. A validação *leave-one-out* é especialmente útil quando o conjunto de dados é pequeno, pois maximiza o uso de cada observação para o treinamento.

Para realizar a análise de regressão linear e aplicar a validação cruzada *leave-one-out*, podemos utilizar o pacote *caret* no R. Nesse caso, configura-se o parâmetro *method* como "LOOCV" (*Leave-One-Out Cross-Validation*), o que indica que cada observação será testada individualmente enquanto o modelo é treinado com as observações restantes. Após realizar essa validação, o coeficiente de determinação da validação cruzada de 0,9389551, o que demonstra que o modelo possui uma alta capacidade de fazer predições em dados novos, reforçando sua confiabilidade para futuras predições.

```
modelols <- caret::train(MG ~ ME,
           data = dados,
           method = "lm",
           trControl = trainControl(method = "LOOCV"))

summary(modelols)
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.488  -3.345   2.178   3.859  11.667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.95059    2.64410  -7.545 8.13e-11 ***
## ME            0.96894    0.02732  35.470  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.043 on 76 degrees of freedom
## Multiple R-squared:  0.943,  Adjusted R-squared:  0.9423
## F-statistic:  1258 on 1 and 76 DF,  p-value: < 2.2e-16

m

## Linear Regression
##
```

```
## 78 samples
##  1 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   7.197741  0.9389551  5.48031
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
y_predls <- predict(modelols, dados)
#Gráfico de predição
library(ggplot2)
ggplot(dados, aes(x = MG, y = y_predls)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2,
              linetype = "dashed") +
  labs(title = "Predicao do Modelo de Regressao Linear Simples",
       x = "Valores Reais",
       y = "Valores Preditos")+
  theme_bw()
```

Para visualizar a capacidade preditiva do modelo, pode-se utilizar a função *ggplot()* do pacote *ggplot2* para criar um gráfico, no qual utiliza os dados preditos e originais. A linha vermelha tracejada indica a linha de identidade, ou seja, onde os valores reais (eixo x) são exatamente iguais aos valores previstos (eixo y). Em um modelo ideal, todos os pontos do gráfico estariam alinhados sobre essa linha. A distância entre os pontos e a linha indica o erro do modelo.

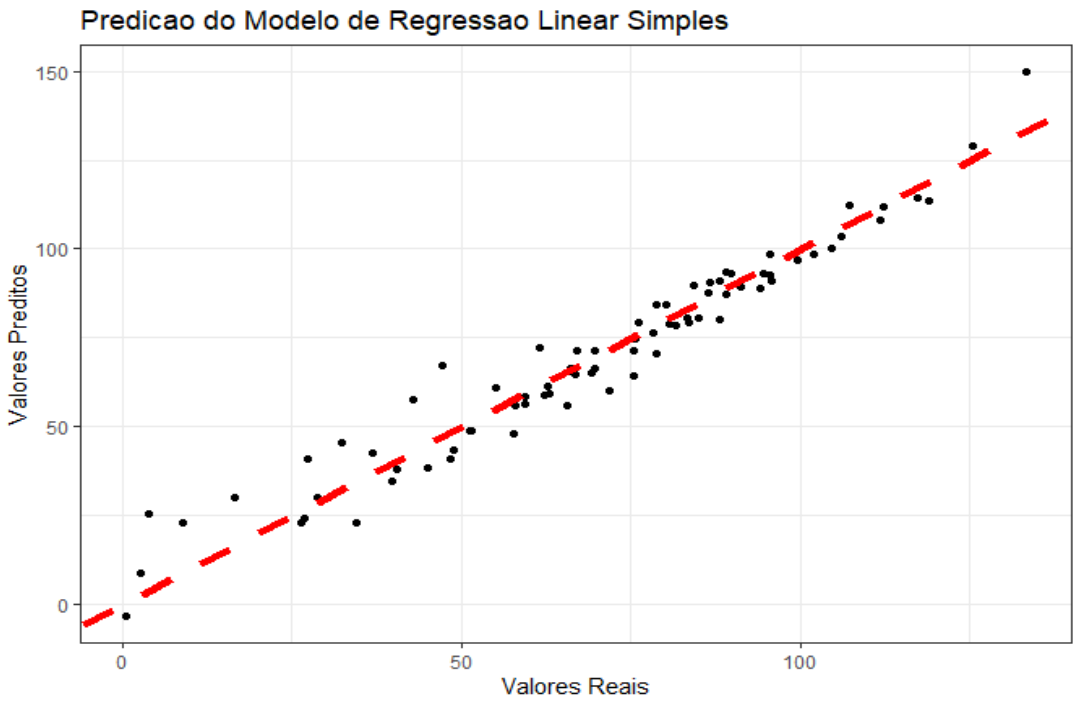

Figura 1. Relação entre os valores originais e os valores preditos por meio de Regressão Linear Simples.

Agora, considera-se o exemplo 2, no qual o pesquisador pretende verificar se é possível predizer a MG com base nas variáveis ME, CE e DE. Realiza-se a análise de regressão linear múltipla utilizando a função *lm*, nativa do software R. Dessa forma, busca-se predizer a MG a partir das variáveis ME, CE e DE.

```
summary(lm(MG~ME+CE+DE, dados))

##
## Call:
## lm(formula = MG ~ ME + CE + DE, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1416  -3.3377   0.4642   4.7237  11.5655
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.36193   14.66701  -2.411  0.01839 *
## ME            0.82533    0.06054  13.634  < 2e-16 ***
## CE           -0.99787    0.66634  -1.498  0.13851
## DE           11.05895    3.45036   3.205  0.00199 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.233 on 74 degrees of freedom
## Multiple R-squared:  0.9566, Adjusted R-squared:  0.9548
## F-statistic: 543.1 on 3 and 74 DF,  p-value: < 2.2e-16
```

Neste exemplo, o intercepto (-35,36193) e os coeficientes das variáveis ME, CE e DE foram significativos 5% de probabilidade pelo teste t de *Student*. Logo, a ME pode ser predita pela seguinte equação: MG = -35,36 + 0,82ME – 0,99CE + 11,05DE. O coeficiente de determinação foi de 0,9566. Isso indica que 95,66% da variação da MG foi explicada pela ME, CE e DE, ou seja, revela um ótimo ajuste do modelo.

A validação cruzada da análise de regressão linear múltipla pode ser realizada pelo mesmo pacote *caret*. Para isso define-se o método LOOCV, que indica que a validação cruzada *leave-one-out* será realizada. Observa-se que após validação cruzada, o coeficiente de determinação foi de 0,9505413, ou seja, indica elevada capacidade do modelo de fazer boas predições em novos dados.

```
modelolm <- caret::train(MG ~ ME+CE+DE,
           data = dados,
           method = "lm",
           trControl = trainControl(method = "LOOCV"))

summary(modelolm)

##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1416  -3.3377   0.4642   4.7237  11.5655
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.36193   14.66701  -2.411  0.01839 *
## ME            0.82533    0.06054  13.634  < 2e-16 ***
## CE           -0.99787    0.66634  -1.498  0.13851
## DE           11.05895    3.45036   3.205  0.00199 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.233 on 74 degrees of freedom
## Multiple R-squared:  0.9566, Adjusted R-squared:  0.9548
## F-statistic: 543.1 on 3 and 74 DF,  p-value: < 2.2e-16

m

## Linear Regression
##
## 78 samples
##  3 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results:
```

```
##
##   RMSE      Rsquared   MAE
##   6.481262  0.9505413  5.037965
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
y_predlm <- predict(modelolm, dados)
#Gráfico de predição SVM Polinomial
library(ggplot2)
ggplot(dados, aes(x = MG, y = y_predlm)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2,
              linetype = "dashed") +
  labs(title = "Predicao do Modelo de Regressao Linear Multipla",
       x = "Valores Reais",
       y = "Valores Preditos")+
  theme_bw()
```

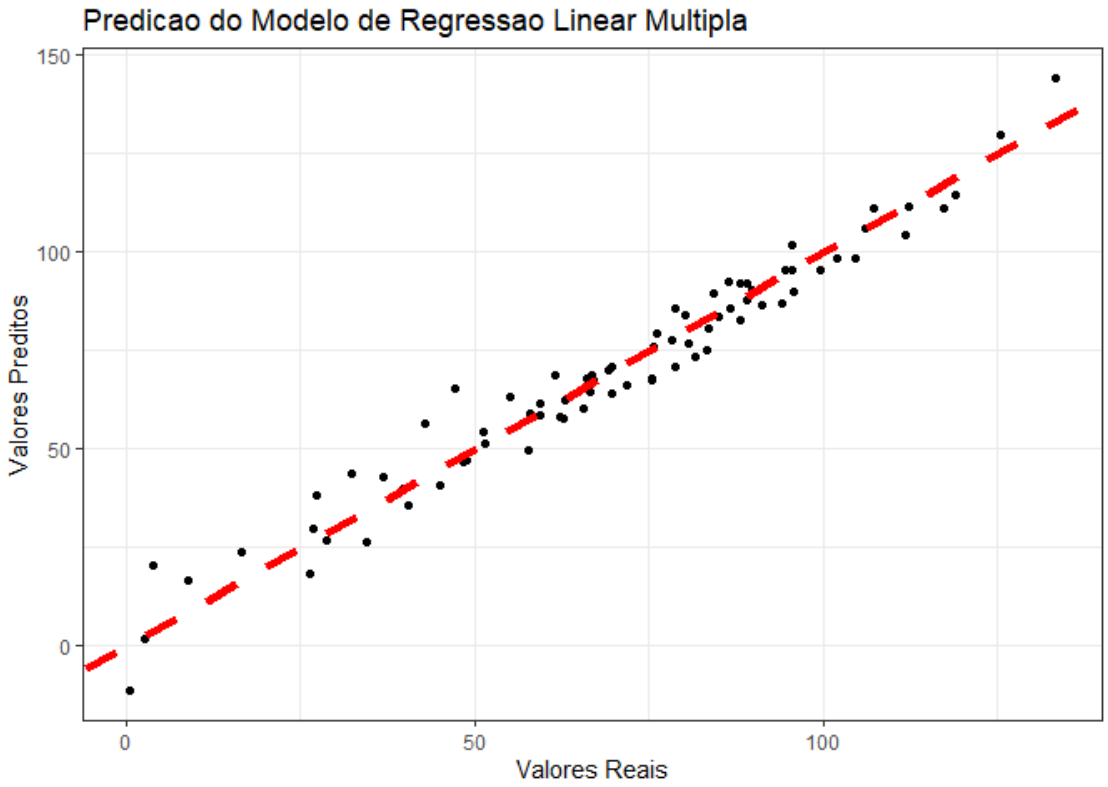

Figura 2. Relação entre os valores originais e os valores preditos por meio de Regressão Linear Múltipla.

4.2.2 Regressão logística

4.2.2.1 Teoria

A regressão logística é utilizada para modelar dados categóricos binários (NICK; CAMPBELL, 2007). Em vez de prever valores contínuos, estima a probabilidade de uma observação pertencer a uma categoria específica. Considerando o banco de dados de milho, um melhorista deseja verificar se é possível desenvolver um modelo para predizer se os genótipos podem ser classificados em alta e baixa qualidade nutricional dos grãos (NUT) com base em caracteres agronômicos e variáveis meteorológicas. Para isso, realizou uma análise fatorial e verificou o agrupamento das variáveis em quatro fatores que explicaram 88% da variação total dos dados. O primeiro fator foi caracterizado por representar a qualidade nutricional proteica dos grãos, uma vez que foi formado por aminoácidos e proteína bruta. A partir dos escores individuais de cada genótipo nesse fator, os genótipos foram classificados em baixa (0) e alta (1) qualidade nutricional proteica. Para isso, os genótipos com escore individual $< 0{,}30$ foram classificados como de baixa qualidade nutricional proteica nos grãos (0), enquanto os genótipos com escores individuais $\geq 0{,}30$ foram classificados como de alta qualidade nutricional proteica nos grãos (1) (HAIR et al., 2009). A partir desses dados, busca-se verificar se é possível classificar os genótipos nos grupos: baixa (0) e alta (1) qualidade nutricional (NUT), por meio da avaliação de variáveis agronômicas e meteorológicas.

Para isso, a análise de regressão logística pode ser utilizada. Modelos de regressão logística são modelos de regressão, uma vez que fornece uma saída numérica: probabilidades de classe (NICK; CAMPBELL, 2007). Essas probabilidades, no entanto, podem ser usadas para classificação. Pode-se, por exemplo, classificar um genótipo como de alta qualidade nutricional se a probabilidade prevista de que seja de alta qualidade for de pelo menos 0,5. Pode-se, portanto, usar a regressão logística como um classificador.

4.2.2.2 Aplicação em R

O pacote *caret* pode ser utilizado para ajustar o modelo de regressão logística e usar validação cruzada para avaliá-lo. É necessário fornecer os argumentos *method = "glm"* para a função *train()* especificar que se busca um modelo de regressão logística. A coluna NUT dos *dados* apresenta as classificações 0 e 1 para o desempenho nutricional dos grãos. As variáveis AP, MG, RSFM, RSFF, RFMC, RFFC, SFM e SFFC foram utilizadas para predizer a classe de NUT dos genótipos.

```
dados$NUT <- as.factor(dados$NUT)
set.seed(1)
modelolog <- caret::train(NUT~AP+MG+RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
                  data = dados,
                  method = "glmnet",
                  trControl = trainControl(method = "LOOCV"))
```

```
modelolog
```

```
## glmnet
##
## 78 samples
##  8 predictor
##  2 classes: '0', '1'
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   alpha  lambda        Accuracy   Kappa
##   0.10   0.0005541062  0.7692308  0.4717833
##   0.10   0.0055410620  0.7692308  0.4717833
##   0.10   0.0554106205  0.7692308  0.4620690
##   0.55   0.0005541062  0.7692308  0.4717833
##   0.55   0.0055410620  0.7692308  0.4717833
##   0.55   0.0554106205  0.7692308  0.4620690
##   1.00   0.0005541062  0.7564103  0.4474273
##   1.00   0.0055410620  0.7692308  0.4717833
##   1.00   0.0554106205  0.7948718  0.5035800
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were alpha = 1 and lambda = 0.05541062.
```

O modelo final foi otimizado para maximizar a acurácia de predição, com um valor de $\alpha = 1$ e $\lambda = 0{,}05541062$, que teve a melhor performance entre os parâmetros testados. A acurácia de 76,92% indica que o modelo possui uma boa capacidade preditiva, com um desempenho confiável na classificação dos genótipos de milho nas classes de qualidade nutricional (0 e 1) (NICK; CAMPBELL, 2007).

Para usar as informações do modelo ajustado com *glmnet* e fazer predições, é preciso aplicar a função *predict()* no modelo treinado. O *data.frame* com as novas observações (genótipos de milho) para os quais busca fazer a predição deve ter as mesmas variáveis que foram usadas para treinar o modelo. A função *predict()* faz a predição com o modelo ajustado. O parâmetro *s* recebe o valor de **λ** (0,05541062) utilizado para regularização. O parâmetro *type = "prob"* indica que se busca a probabilidade de cada classe, ou seja, a probabilidade de cada amostra (genótipo) pertencer a classe 0 ou 1. Para converter essas probabilidades em classes (0 ou 1), pode-se utilizar um limiar de 0,5. Se a probabilidade for maior que 0,5, a classe será *"1"*, caso contrário, será *"0"*.

```
#Pode-se utilizar o modelo ajustado para predizer classes de novos dados
predicoes <- predict(modelolog, dados, s = 0.05541062, type = "prob")
```

```r
# As predições vão retornar a probabilidade da classe "1"
# Pode-se arredondar os valores de probabilidade para apenas as classes 0 ou 1.
predicoes_classe <- ifelse(predicoes > 0.5, 1, 0)
```

Com o modelo ajustado, é possível realizar predições em novos conjuntos de dados que incluam as variáveis agronômicas e meteorológicas (AP, MG, RSFM, RSFF, RFMC, RFFC, SFMC, SFFC). As saídas do modelo, representadas por *predicoes_classe*, indicarão se a qualidade nutricional proteica dos grãos (NUT) de cada genótipo de milho está classificada como alta (1) ou baixa (0). Essa classificação permitirá identificar quais genótipos apresentam melhor qualidade nutricional com base nas variáveis analisadas, contribuindo para decisões sobre seleção e manejo dos genótipos.

4.2.3 Árvores de regressão

4.2.3.1 Teoria

Uma árvore de decisão é um modelo preditivo organizado em forma de uma estrutura hierárquica que representa decisões e seus possíveis resultados (BREIMAN, 1984). No contexto de dos dados de milho, onde variáveis agronômicas como AP e MG, variáveis meteorológicas como RSFM, RFMC, SSFM e SFMC e variáveis nutricionais dos grãos como CP foram analisadas, uma árvore de regressão pode ser utilizada para predizer o teor de CP com base nas demais variáveis. Esse modelo é útil quando o objetivo é predizer valores numéricos contínuos, como o CP, ao contrário das árvores de classificação, que trabalham com categorias.

Uma árvore de regressão funciona dividindo os dados em diferentes ramificações baseadas nas variáveis preditoras, de forma que cada divisão reduz a variabilidade do valor a ser predito, neste caso o CP (BREIMAN, 1984). Inicialmente, todos os dados estão agrupados e a árvore identifica a variável que melhor separa os genótipos de milho em relação ao CP. Por exemplo, a primeira divisão pode ser: a RSFM é maior que 350 MJ m^{-2}?. A partir dessa separação, a árvore continua ramificando, analisando outras variáveis, como MG ou AP, até que cada ramo termine em uma folha. Cada folha contém um valor médio de CP, representando a predição do modelo para os dados que chegaram até aquela folha.

Suponha que, para o experimento, a árvore gerada tenha a seguinte estrutura: no primeiro nó, *RSFM > 350 MJ/m^{-2}?*. Se a resposta for *Sim*, a próxima pergunta é *MG > 250 g?*. Caso a resposta seja novamente *Sim*, a CP média predita é de 12%. Caso seja *Não*, a CP média é de 10%. Se no primeiro nó a resposta for *Não*, a próxima pergunta será *AP > 150 cm?*. Se a resposta for *Sim*, o CP média será 8%, e se *Não*, será 6%. Assim, para um genótipo com RSFM = 360 MJ/m², MG = 260 g e AP = 140 cm, o modelo prediz que o CP será 12%. A árvore de regressão tem sido amplamente utilizada em estudos agronômicos para predição de variáveis dependentes em milho (LORO et al.,

2024a), teosinto (KONRAD et al., 2023; REIS et al., 2023), alfafa (CHEROBINI et al., 2024), arroz (MEUS et al., 2024) e soja (SCARTON et al., 2023).

4.2.3.2 Aplicação em R

Pode-se usar o pacote *rpart* para ajustar o modelo de árvore de regressão. É necessário fornecer os argumentos *method = "anova"* para a função *rpart* especificar que se busca um modelo de árvore de regressão, ao invés de um modelo de árvore de classificação (*method = "class"*). O argumento *cp* é o parâmetro de complexidade que controla o processo de poda da árvore de decisão. É utilizado para evitar o *overfitting* (ajuste excessivo), simplificando a árvore ao remover divisões que não contribuem significativamente para a melhoria do modelo. Cada divisão na árvore aumenta sua complexidade. O *cp* define o ganho mínimo de ajuste necessário para justificar uma divisão adicional. Divisões que não reduzem o erro relativo em pelo menos o valor de *cp* são descartadas. Um valor mais alto de *cp* resulta em árvores mais simples, enquanto valores menores permitem árvores mais complexas.

Para esse exemplo, o objetivo será predizer a variável CP em função das variáveis RSFM, RSFF, RFMC, RFFC, SFM e SFFC, ou seja, busca-se predizer a CP por meio de variáveis meteorológicas. A partir dos resultados será possível identificar como condições de radiação solar global e soma térmica determinam a expressão de proteína bruta nos grãos de milho. Inicialmente, será utilizado o pacote *caret* para definir o melhor valor de *cp*.

```
#------------------------AJUSTANDO PARAMETRO cp
set.seed(123)
caret::train(CP~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
               data = dados,
               method = "rpart",
               tuneGrid = data.frame(cp = c(0, 0.01, 0.05, 0.1, 0.15, 0.2)),
               trControl = trainControl(method = "LOOCV"))

## CART
##
## 78 samples
##  6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   cp    RMSE       Rsquared   MAE
##   0.00  0.6236357  0.3950318  0.4984328
##   0.01  0.6150569  0.4079631  0.4877489
##   0.05  0.6362455  0.3629763  0.4982286
##   0.10  0.6692520  0.3039632  0.5145622
##   0.15  0.6641189  0.2956279  0.5144362
```

```
##   0.20  0.6641189  0.2956279  0.5144362
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.01.
```

O valor de *cp* foi definido como 0,01, ou seja, o modelo aceitará divisões na árvore somente se a redução no erro relativo for maior ou igual a 1%. Esse valor é importante para equilibrar a simplicidade e o desempenho preditivo do modelo, prevenindo tanto *overfitting* (modelo muito complexo) quanto *underfitting* (modelo muito simples). A partir do valor de *cp* definido, utiliza-se a função *rpart* para extrair o modelo de árvore de regressão.

```
#----------------------MODELO DE ÁRVORE DE REGRESSÃO
modeloar <- rpart::rpart(CP~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
             data = dados,
                    method = "anova",
                    cp = 0.01)
```

O modelo de árvore de regressão foi ajustado, agora busca-se verificar quais as variáveis apresentaram maior contribuição para o modelo. A partir da função *barplot*, é possível plotar um gráfico com a importância de cada variável para a predição da CP.

```
#contribuição das variáveis
barplot(modeloar$variable.importance)
```

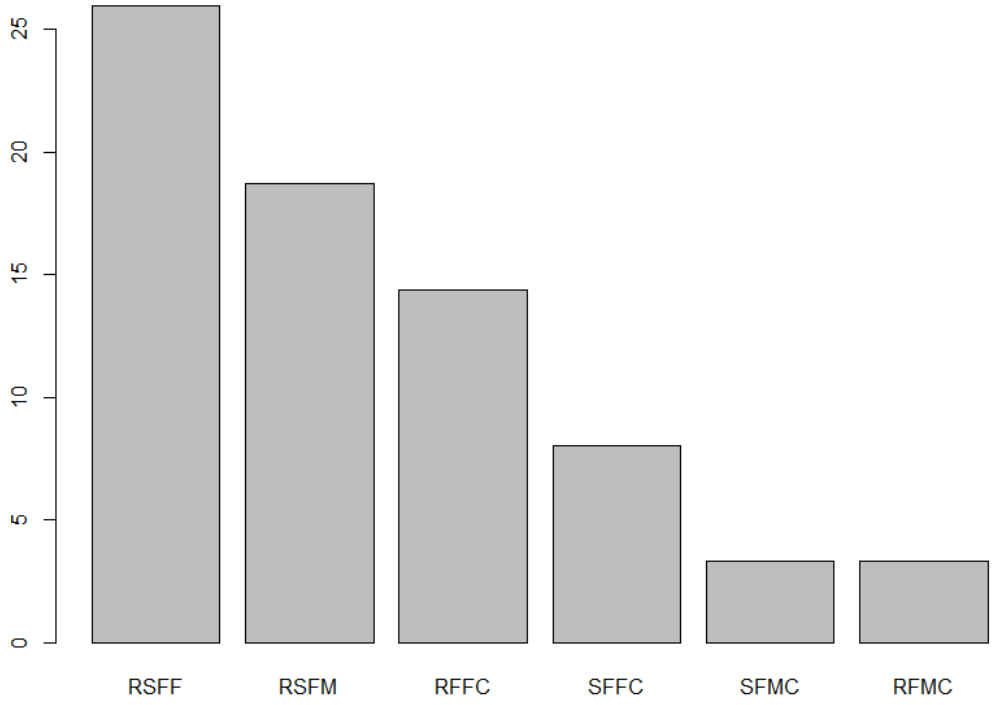

Figura 3. Contribuição das variáveis meteorológicas para predição CP por meio da árvore de decisão baseada em regressão.

Verifica-se que as variáveis RSFF e RSFM são as que mais contribuem para o modelo de predição. Assim, pode-se inferir que a CP é determinada principalmente por essas duas variáveis meteorológicas. A partir da compreensão das principais variáveis, pode-se plotar a árvore de regressão por meio da função *rpart.plot()*.

```
# plotar a arvore

rpart.plot::rpart.plot(modeloar,
            type = 0,
            extra = 101,
            box.palette = c("yellow","green"),
            branch.lty=2,
            shadow.col = "black",
            nn=TRUE,
            cex = 1.2)
```

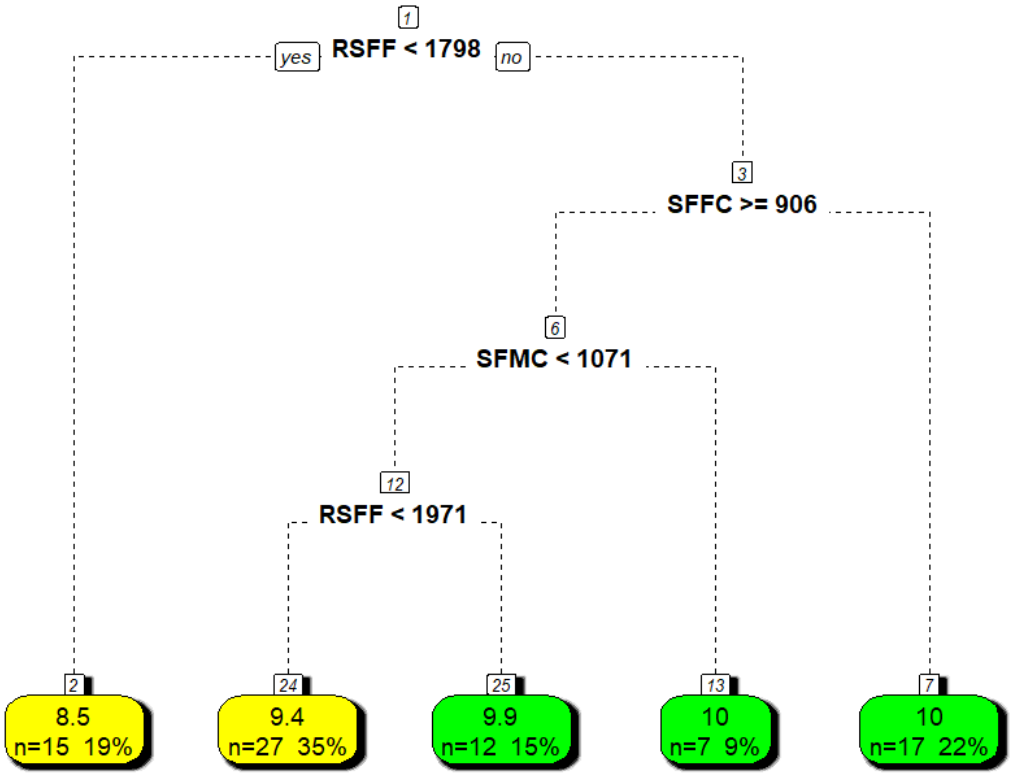

Figura 4. Árvore de decisão baseada em regressão para predição da CP em função de variáveis meteorológicas.

A maior média de CP nos grãos de milho foi 10%, que constituiu 31% das observações do banco de dados. Logo, verificou-se que os genótipos que apresentaram a maior expressão de CP nos grãos (10%) caracterizaram-se por apresentar RSFF ≥ 1798 MJ m^{-2} e SFFC < 906 ºC. Os genótipos com RSFF ≥ 1798 MJ m^{-2}, SFFC ≥ 906 ºC e SFMC ≥ 1071 °C também promovem maior média de CP (10%). A menor média de CP nos grãos foi de 8,50%, e foi observada em genótipos que apresentaram RSFF < 1798 MJ m^{-2}.

As estatísticas de precisão do modelo podem ser obtidas de duas formas (modo 1 e modo 2), conforme o script abaixo. O coeficiente de determinação (R^2) do modelo de árvore de regressão foi de 0,601648, indicando que o modelo foi capaz de explicar 60,16% da variabilidade da CP a partir das variáveis meteorológicas. Embora o R^2 seja relativamente baixo, a principal vantagem da árvore de regressão reside na sua facilidade de interpretação. É possível identificar, de forma prática, os limites das variáveis preditoras que contribuem para o aumento médio da variável resposta.

```
# Estatísticas de precisão do modelo - MODO 1
y_predito <- predict(modeloar, newdata = dados)
y_original <- dados$CP
n <- length(y_original)
```

```r
#R²
1-(sum((y_original-y_predito)^2)/
  sum((y_original-mean(y_original))^2))
```

```
## [1] 0.6016448
```

```r
#MAPE
sum(abs(y_original-y_predito)/y_original)/n*100
```

```
## [1] 4.131537
```

```r
#MAE
sum(abs(y_original-y_predito))/n
```

```
## [1] 0.3928098
```

```r
#RMSE
sqrt((sum((y_original-y_predito)^2))/n)
```

```
## [1] 0.4972892
```

```r
# Estatísticas de precisaõ do modelo - MODO 2
teste <- data.frame(obs = dados$CP, pred=y_predito)
caret::defaultSummary(teste)
```

```
##      RMSE  Rsquared       MAE
## 0.4972892 0.6016448 0.3928098
```

```r
#Gráfico de predição SVM Polinomial
library(ggplot2)
ggplot(dados, aes(x = CP, y = y_predito)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2,
              linetype = "dashed") +
  labs(title = "Predicao do Modelo de Arvore de Regressao",
       x = "Valores Reais",
       y = "Valores Preditos")+
  theme_bw()
```

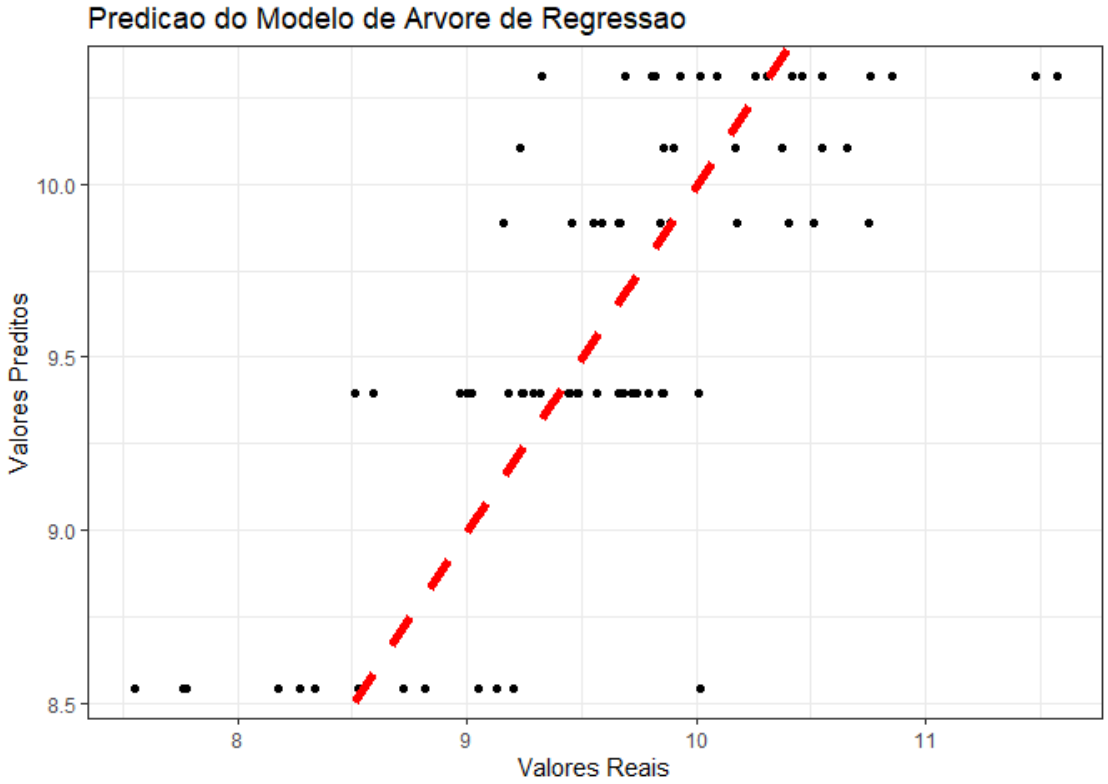

Figura 5. Relação entre os valores originais e os valores preditos por meio do modelo de Árvore de Regressão.

4.2.4 Random Forest

4.2.4.1 Teoria

Random Forest é um método de ensemble que constrói múltiplas árvores de decisão (floresta) e combina suas predições para melhorar a precisão da predição (BREIMAN, 2001). Esse método reduz o risco de *overfitting* ao combinar várias árvores, cada uma construída com um subconjunto aleatório dos dados e variáveis. O primeiro passo do algoritmo é selecionar aleatoriamente um conjunto de observações (por exemplo, genótipos de milho) a partir dos dados originais. Essa seleção é realizada por meio de reamostragem *bootstrap*, onde o algoritmo cria várias amostras com reposição, ou seja, algumas observações podem aparecer mais de uma vez em cada amostra (BREIMAN, 2001). Para cada uma dessas amostras, uma árvore de decisão é construída. No início, o algoritmo seleciona um subconjunto aleatório de variáveis para decidir qual delas será usada para dividir os dados no primeiro nó da árvore. O número de variáveis pode ser determinado pelo pesquisador, caso não seja, geralmente o número de variáveis é definido pela raiz quadrada do número total de variáveis. Em seguida, repete-se esse processo para o próximo nó, novamente escolhendo aleatoriamente um subconjunto de variáveis, e assim por diante, até que a árvore esteja completa.

Uma vez que a primeira árvore de decisão é construída, o processo é repetido. Uma nova amostra é criada aleatoriamente a partir dos dados originais (novamente com reposição) e uma nova

árvore de decisão é construída da mesma forma, usando um novo subconjunto de variáveis para cada nó. É possível indicar ao algoritmo o número de árvores individuais que se deseja criar. Após várias árvores serem construídas, o algoritmo usa o resultado combinado de todas para fazer uma predição final. No caso de classificação, a predição final é baseada na votação majoritária entre as árvores, ou seja, a classe mais votada entre todas as árvores será a predição final. Para regressão, a predição final é a média das predições feitas por cada árvore. Por exemplo, após o pesquisador desenvolver várias árvores de regressão, é possível predizer o valor de da variável dependente de uma nova observação. Assim, ao fornecer essa observação ao modelo, o algoritmo passa a observação por todas as árvores de regressão criadas e o resultado final predito é a média da predição de todas as árvores individuais. Esse algoritmo tem uma ampla aplicação em culturas como a soja (BATISTELLA et al., 2023; SMIDT et al., 2016) e milho (KHAN; LI; MAIMAITIJIANG, 2022).

4.2.4.2 Aplicação em R

Pode-se usar o pacote *caret* para utilizar a metodologia *random forest*. É necessário fornecer os argumentos *method = "rf"* para a função *train()* especificar a utilização de *random forest*. O argumento *ntree* indica o número de árvores de regressão que se deseja criar. Será utilizada a validação cruzada *leave-one-out*, para validar o modelo. Para esse exemplo, o objetivo será predizer a variável CP em função das variáveis RSFM, RSFF, RFMC, RFFC, SFM e SFFC, ou seja, busca-se predizer a CP por meio de variáveis meteorológicas. Esse é mesmo exemplo utilizado no método de árvore de regressão (Subitem 2.5.2), afim de verificar a diferença entre os métodos.

```
set.seed(1)
modelorf <- caret::train(CP~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
                  data = dados,
                  method = "rf",
                  ntree = 100,
                  trControl = trainControl(method = "LOOCV"))

modelorf

## Random Forest
##
## 78 samples
##  6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##   2     0.5903316  0.4491635  0.4497198
##   4     0.6016967  0.4387674  0.4624021
```

```
##   6     0.6135132  0.4326330  0.4747689
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
```

Os resultados obtidos para o modelo de *Random Forest* ajustado podem ser analisados sob duas perspectivas: a precisão estimada durante a validação cruzada *leave-one-out* e o desempenho avaliado diretamente nos dados originais. Durante a validação cruzada *leave-one-out*, o modelo apresentou um R^2 de 0,4491635, indicando que foi capaz de explicar cerca de 44,91% da variabilidade da CP, com base nas variáveis preditoras. Esse valor, apesar de moderado, demonstra que o modelo possui alguma capacidade de generalização ao ser aplicado a dados que não foram utilizados no treinamento.

Quando o desempenho do modelo foi avaliado diretamente nos dados originais, o R^2 atingiu 0,8768105, um aumento expressivo em relação à validação cruzada, o que é esperado, já que o modelo está familiarizado com os dados usados para ajustar seus parâmetros. Esse resultado, no entanto, reflete o ajuste do modelo aos dados de treinamento e não necessariamente sua capacidade de generalizar para novos conjuntos de dados. A diferença nos valores de R^2, MAE e RMSE entre a validação cruzada e a avaliação nos dados originais ressalta a importância de considerar métricas obtidas por métodos robustos de validação, como a *Leave-One-Out*, para evitar interpretações inadequadas do desempenho do modelo. Enquanto os resultados nos dados originais indicam um ajuste eficiente, apenas os resultados da validação cruzada são indicativos da capacidade real de generalização do modelo. Portanto, a análise sugere que, embora o modelo de *Random Forest* apresente boa capacidade de ajuste aos dados, sua aplicabilidade para predições deve ser avaliada com cuidado, especialmente considerando o desempenho mais moderado observado nos testes de validação cruzada.

```r
# Estatísticas de precisão do modelo - MODO 1
y_predit <- predict(modelorf, newdata = dados)
y_origina <- dados$CP
nrf <- length(y_origina)

#R²
1-(sum((y_origina-y_predit)^2)/
  sum((y_origina-mean(y_origina))^2))
```

```
## [1] 0.8612126
```

```r
#MAPE
sum(abs(y_origina-y_predit)/y_origina)/nrf*100
```

```
## [1] 2.346093
```

```r
#MAE
sum(abs(y_origina-y_predit))/nrf
```

```
## [1] 0.2252012

#RMSE
sqrt((sum((y_origina-y_predit)^2))/nrf)

## [1] 0.2935276

# Estatísticas de precisão do modelo - MODO 2
test <- data.frame(obs = dados$CP, pred=y_predit)
caret::defaultSummary(test)

##      RMSE  Rsquared       MAE
## 0.2935276 0.8768105 0.2252012
```

```
#Gráfico de predição Random Forest
library(ggplot2)
ggplot(dados, aes(x = CP, y = y_predit)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2,
              linetype = "dashed") +
  labs(title = "Predicao do Modelo Random Forest",
       x = "Valores Reais",
       y = "Valores Preditos")+
  theme_bw()
```

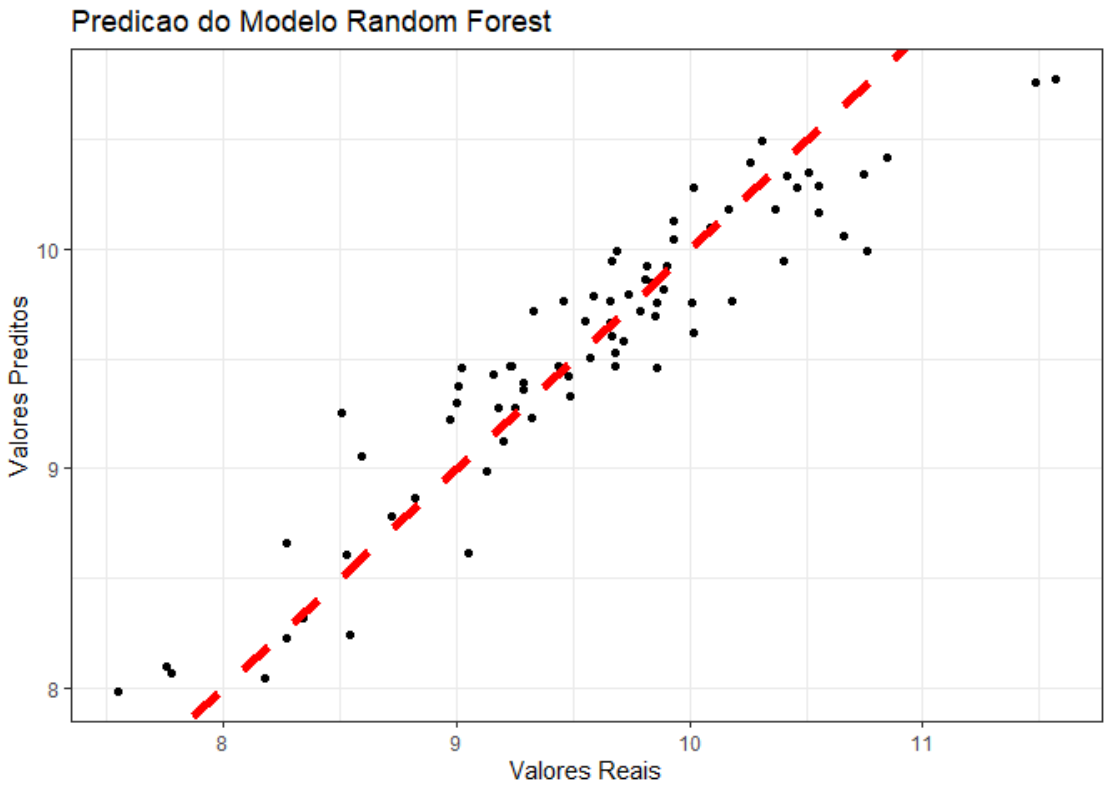

Figura 6. Relação entre os valores originais e os valores preditos por meio do modelo Random Forest.

Os resultados do R^2 para os modelos de *Random Forest* e árvore de regressão (*subitem 2.5.2*) mostram diferenças no desempenho. No caso da *Random Forest*, o R^2 nos dados originais foi de 0,87, indicando que o modelo conseguiu explicar 87% da variabilidade da CP nos dados de treinamento. Durante a validação cruzada, o melhor R^2 foi de 0,44, quando o parâmetro *mtry* foi igual a 2. Esses resultados refletem a maior capacidade da *Random Forest* em capturar padrões nos dados, tanto nos originais quanto nos testes de generalização, devido à sua estrutura robusta e ao uso de múltiplas árvores que reduzem o risco de *overfitting*.

Por outro lado, a árvore de regressão apresentou um R^2 nos dados originais de 0,60, inferior ao da *Random Forest*, sugerindo menor capacidade de ajuste aos dados de treinamento. Nos resultados da validação cruzada, o R^2 máximo alcançado foi de 0,40 com o parâmetro de complexidade (*cp*) igual a 0,01. Isso indica que, embora a árvore de regressão seja mais fácil de interpretar, possui menor poder preditivo e capacidade de generalização em comparação à *Random Forest*. Logo, a *Random Forest* demonstrou desempenho superior, tanto nos dados originais quanto na validação cruzada, com R^2 consideravelmente mais altos. Esses resultados refletem a capacidade da *Random Forest* de capturar interações complexas e padrões nos dados que uma única árvore de regressão não consegue modelar. A simplicidade interpretativa da árvore de regressão pode ser uma vantagem em situações em que a interpretabilidade é mais importante que o desempenho preditivo.

4.2.5 Gradient Boosting

4.2.5.1 Teoria

O *Gradient Boosting* é outro método de ensemble que gera modelos sequenciais, onde cada novo modelo corrige os erros do anterior. Caracteriza-se por ser um método flexível e capaz de ajustar modelos para dados complexos e não lineares. Adequado para predições precisas em experimentos com grandes volumes de dados e muitas variáveis preditoras. Enquanto as florestas aleatórias (*Random Florest*) criam várias árvores independentes e profundas, o *Gradient Boosting* constrói um conjunto de árvores simples de forma sequencial (NATEKIN; KNOLL, 2013). Em vez de criar todas as árvores de uma vez, cada árvore é construída para corrigir os erros que a árvore anterior cometeu. Embora as árvores simples, tenham baixa capacidade preditiva, quando combinadas apresentam alta capacidade de predição. Esse processo de aprendizagem vai ajustando o modelo aos poucos, e, quando bem treinado, o *Gradient Boosting*, geralmente, ter um desempenho elevado, superando outros tipos de algoritmos de aprendizado de máquina. A principal diferença está na forma como cada árvore vai aprendendo e melhorando o modelo, o que torna o *Gradient Boosting* uma ferramenta eficiente para verificar relações complexas em conjuntos de dados.

Portanto, a principal estratégia do *Gradient Boosting* é construir árvores forma sequencial, no qual cada árvore corrige os erros de predição da árvore anterior (FRIEDMAN, 2001). Embora uma

árvore simples seja um modelo de menor capacidade preditiva, quando várias árvores são combinadas, o modelo resultante se torna mais preciso. Para melhor compreender como o *Gradient Boosting* funciona, será utilizado um exemplo com dados de milho. Suponha que o objetivo seja predizer a CP de grãos de milho com base em variáveis meteorológicas. No início, o modelo começa com uma árvore simples, que tenta predizer a CP dos grãos do milho apenas com base em uma variável, como a RFMC. Essa primeira árvore pode apresentar uma baixa capacidade preditiva, pois a predição é baseada em apenas uma variável.

Após a construção da primeira árvore, o algoritmo avalia os erros de predição e os utiliza desenvolver a próxima árvore. Logo, a segunda árvore é construída com o objetivo de corrigir os erros de predição da primeira árvore. Para isso, é utilizada uma taxa de aprendizado que pode variar entre 0 e 1. Valores baixos indicam que o algoritmo aprende de uma forma lenta e gradual, enquanto valores altos indicam que o algoritmo vai aprender de forma acelerada. O valor da taxa de aprendizado depende do número de árvores que serão criadas. Se o número de árvores for baixo, é necessário um maior valor de taxa de aprendizado, se o número de árvores for alto, pode-se utilizar um menor valor de taxa de aprendizado.

Esse processo continua por várias iterações, com cada nova árvore tentando corrigir os erros das árvores anteriores. Ao final, o modelo combina todas as árvores para gerar uma predição final. A ideia é que, embora cada árvore individualmente seja um modelo simples, a combinação de várias árvores resulta em um modelo robusto e preciso. Logo, tem alta capacidade preditiva para dados complexos e variáveis interdependentes.

4.2.5.2 Aplicação em R

Pode-se usar o pacote *caret* para utilizar a metodologia *Gradient Boosting*. É necessário fornecer os argumentos *method = "gbm"* para a função *train()* especificar a utilização de *Gradiente Boosting*. Para esse exemplo, o objetivo será predizer a variável CP em função das variáveis RSFM, RSFF, RFMC, RFFC, SFM e SFFC, ou seja, busca-se predizer a CP por meio de variáveis meteorológicas. Inicialmente foram definidos alguns parâmetros para serem testados no ajuste do modelo. A função *expand.grid()* foi utilizada para realizar os ajustes. Os parâmetros *n.trees*, *interaction.deppth*, *shrinkage* e *n.minobsinnode* são usados para informar o número de árvores, profundidade das árvores, taxa de aprendizado e o número mínimo de observações por nó, respectivamente.

```
# Definindo a grade de parâmetros (taxa de aprendizagem e número de árvores)
tune_grid <- expand.grid(
  n.trees = c(50, 100, 150, 300),    # Número de árvores
  interaction.depth = c(1, 3, 5),    # Profundidade das árvores
```

```
  shrinkage = c(0.01, 0.05, 0.1),     # Taxa de aprendizado (learning rate)
  n.minobsinnode = c(5)               # Número mínimo de observações por nó term
inal
)
```

Em seguida foi realizado o ajuste do modelo com a função *train* e verificado os melhores hiper parâmetros do modelo. O parâmetro *n.minobsinnode* foi mantido fixo com o valor de 10. Isso ajuda a evitar a criação de nós com poucos dados, o que poderia tornar o modelo muito sensível aos dados de treinamento, levando a um sobreajuste (*overfitting*). A seleção do modelo ideal foi feita utilizando a métrica RMSE (*Root Mean Square Error*), que é um indicador comum de precisão em modelos de regressão. O objetivo foi minimizar esse valor, o que significa que o modelo foi ajustado para reduzir ao máximo a diferença entre as previsões e os valores reais observados. Após o ajuste, os parâmetros finais escolhidos para o modelo foram os seguintes: n.trees = 50, interaction.depth = 1 e shrinkage = 0,1.

```
#Treinamento do modelo
set.seed(1)
library(caret)
modelogb <- caret::train(CP~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
                  data = dados,
                  method = "gbm",#define gradient boosting
                  verbose = FALSE,
                  trControl = trainControl(method = "LOOCV"),
                  tuneGrid = tune_grid)# Passando a grade de parâmetros
modelogb

## Stochastic Gradient Boosting
##
## 78 samples
##  6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   n.trees  interaction.depth  shrinkage  RMSE       Rsquared   MAE
##    50      1                  0.01       0.6998425  0.3968528  0.5364712
##    50      1                  0.05       0.5794676  0.4725994  0.4458514
##    50      1                  0.10       0.5592615  0.4961727  0.4392305
##    50      3                  0.01       0.6700146  0.4352472  0.5090332
##    50      3                  0.05       0.5618518  0.4918224  0.4403975
##    50      3                  0.10       0.5828131  0.4621228  0.4576260
##    50      5                  0.01       0.6627767  0.4570804  0.5035832
##    50      5                  0.05       0.5779180  0.4623968  0.4493023
##    50      5                  0.10       0.5965265  0.4471140  0.4637714
##   100      1                  0.01       0.6432145  0.4372252  0.4912022
##   100      1                  0.05       0.5714151  0.4742294  0.4338028
```

```
##   100      1                   0.10       0.5856658  0.4568291  0.4514771
##   100      3                   0.01       0.6078501  0.4663083  0.4636006
##   100      3                   0.05       0.5828243  0.4648766  0.4520633
##   100      3                   0.10       0.5930642  0.4526387  0.4592789
##   100      5                   0.01       0.6022241  0.4686843  0.4601027
##   100      5                   0.05       0.5971189  0.4449550  0.4648022
##   100      5                   0.10       0.6311682  0.4121146  0.4933414
##   150      1                   0.01       0.6093732  0.4556827  0.4672260
##   150      1                   0.05       0.5704613  0.4784181  0.4348255
##   150      1                   0.10       0.5782997  0.4731647  0.4492974
##   150      3                   0.01       0.5808088  0.4758462  0.4475183
##   150      3                   0.05       0.5864977  0.4644896  0.4524814
##   150      3                   0.10       0.6026976  0.4463760  0.4700511
##   150      5                   0.01       0.5804916  0.4703207  0.4452269
##   150      5                   0.05       0.6061019  0.4381124  0.4685676
##   150      5                   0.10       0.6538225  0.3867369  0.5105376
##   300      1                   0.01       0.5710629  0.4821877  0.4383912
##   300      1                   0.05       0.5732918  0.4785964  0.4368527
##   300      1                   0.10       0.6042441  0.4442631  0.4693577
##   300      3                   0.01       0.5695869  0.4776849  0.4408936
##   300      3                   0.05       0.6215450  0.4232277  0.4821949
##   300      3                   0.10       0.6623977  0.3731690  0.5174547
##   300      5                   0.01       0.5706592  0.4770533  0.4401290
##   300      5                   0.05       0.6402085  0.4041171  0.4996875
##   300      5                   0.10       0.7053711  0.3381829  0.5444581
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 5
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were n.trees = 50, interaction.depth =
##  1, shrinkage = 0.1 and n.minobsinnode = 5.
```

Esses ajustes foram realizados com o objetivo de otimizar o desempenho do modelo, garantindo que ele fosse capaz de predizer os dados com alta precisão sem se tornar excessivamente complexo e suscetível ao *overfitting*. A combinação desses parâmetros buscou garantir que o modelo tivesse um bom desempenho ao fazer predições, afim de minimizar os erros enquanto mantinha a capacidade de generalizar bem para novos dados. O valor final de RMSE refletiu essa otimização, indicando o melhor ajuste possível do modelo aos dados de treinamento.

```r
# Estatísticas de precisão do modelo - MODO 1
y_pred <- predict(modelogb, newdata = dados)
y_orig <- dados$CP
ngb <- length(y_orig)

#R²
1-(sum((y_orig-y_pred)^2)/
  sum((y_orig-mean(y_orig))^2))
```

```
## [1] 0.6530459
```

```r
#MAPE
sum(abs(y_orig-y_pred)/y_orig)/ngb*100
```

```
## [1] 3.720681
```

```
#MAE
sum(abs(y_orig-y_pred))/ngb
```

```
## [1] 0.3566239
```

```
#RMSE
sqrt((sum((y_orig-y_pred)^2))/ngb)
```

```
## [1] 0.4640981
```

```
# Estatísticas de precisão do modelo - MODO 2
tes <- data.frame(obs = dados$CP, pred=y_pred)
caret::defaultSummary(tes)
```

```
##      RMSE  Rsquared       MAE
## 0.4640981 0.6670346 0.3566239
```

```
#Gráfico de predição Gradient Boosting
library(ggplot2)
ggplot(dados, aes(x = CP, y = y_pred)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2,
              linetype = "dashed") +
  labs(title = "Predicao do Modelo Gradient Boosting",
       x = "Valores Reais",
       y = "Valores Preditos")+
  theme_bw()
```

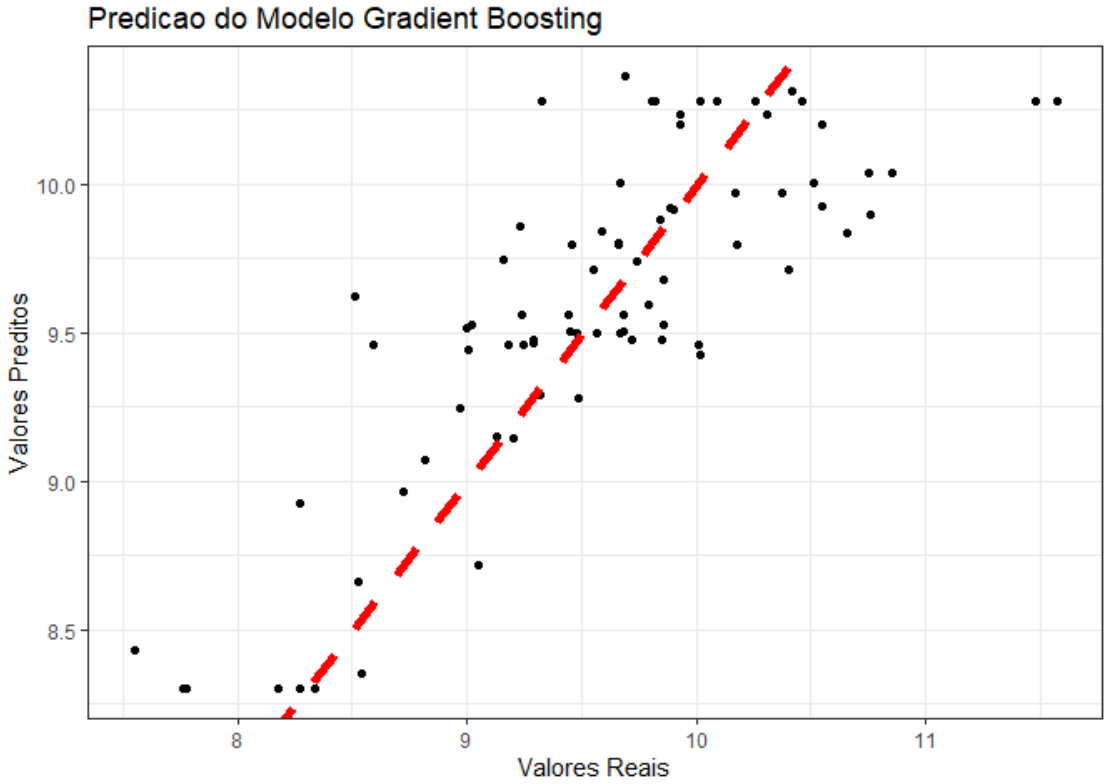

Figura 7. Relação entre os valores originais e os valores preditos por meio do modelo de Gradient Boosting.

4.2.6 Máquinas de vetores de suporte (SVM)

4.2.6.1 Teoria

O algoritmo de Máquinas de Vetores de Suporte (SVM) é conhecido por sua aplicação em classificação, mas também pode ser usado para regressão. No contexto de regressão, o objetivo do SVR é encontrar uma linha ou hiperplano (em espaços de mais dimensões) que melhor represente a relação entre as variáveis preditoras e a variável resposta (LORENA; CARVALHO, 2007). Contudo, ao contrário de métodos tradicionais como regressão linear, o SVR utiliza margens de tolerância (um conceito chamado de *epsilon-insensitive margin* para determinar quais pontos estão próximos o suficiente do modelo. O SVM pode ser utilizado tanto para dados lineares, quanto para dados não lineares.

O SVR permite definir um intervalo (ε) ao redor da linha predita, onde erros dentro dessa margem são ignorados (NOBLE, 2006). Apenas os pontos fora da faixa de tolerância são penalizados no ajuste do modelo. O SVR não tenta minimizar diretamente o erro de cada ponto, mas foca em minimizar o desvio dos pontos fora da margem. Esses pontos são chamados de *vetores de suporte*. O SVR usa um parâmetro *C*, que controla o *overfitting*, ou seja, busca um equilíbrio entre ajustar um modelo simples que generaliza bem para novos dados e ajustar um modelo complexo que pode ser mais preciso nos dados de treinamento, mas corre o risco de *overfitting*. Se o conjunto de variáveis

não apresenta relação de linearidade, pode-se utilizar funções kernel (linear, polinomial, radial e sigmoide), que transformam os dados em uma nova dimensão, o que permite ajustar hiperplanos em cenários complexos (BRERETON; LLOYD, 2010).

Por exemplo, um pesquisador tem objetivo de predizer a CP por meio de variáveis meteorológicas. Inicialmente, o algoritmo SVM de regressão ajusta uma linha preditiva que ignora pequenos desvios dentro de uma faixa aceitável (ε). O objetivo será ajustar apenas os pontos que estão fora dessa faixa, de modo que o modelo seja robusto a pequenas variações nos dados. Caso os dados não sejam lineares, pode-se aplicar um kernel ao modelo, que transforma os dados para uma dimensão de forma que exista uma relação linear.

4.2.6.2 Aplicação em R

Para esse exemplo, o objetivo será predizer a variável CP em função das variáveis RSFM, RSFF, RFMC, RFFC, SFM e SFFC, ou seja, busca-se predizer a CP por meio de variáveis meteorológicas. O pacote *caret* será utilizado para criar os modelos de SVM. Serão abordados três diferentes *kernels* para predição da CP (linear, polinomial e radial). Inicialmente o método *svmLinear* foi empregado, com validação cruzada *leave-one-out* (LOOCV). Utilizou-se um teste do parâmetro *C* nas faixas de 0,01 a 100. Verificou-se que o parâmetro *C* igual a 100 foi o mais adequado para reduzir o RMSE. O R^2 utilizando os dados originais foi de 42%, enquanto o R^2 da validação cruzada foi de 34,26%. Isso indica uma baixa capacidade de predizer novos dados.

```
library(caret)
set.seed(1)
modelosvm <- caret::train(CP~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
                  data = dados,
                  method = "svmLinear", #kernel linear
                  tuneGrid = expand.grid(C = c(0.01, 0.1, 1, 10, 100)),
                  trControl = trainControl(method = "LOOCV"))
modelosvm
```

```
## Support Vector Machines with Linear Kernel
##
## 78 samples
##  6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   C      RMSE       Rsquared   MAE
##   1e-02  0.6523869  0.3651933  0.5038620
##   1e-01  0.6529905  0.3285475  0.5161344
##   1e+00  0.6582048  0.3245028  0.5109957
##   1e+01  0.6485515  0.3388301  0.4831638
```

```
##   1e+02  0.6462822  0.3426027  0.4762579
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was C = 100.

#Estatísticas de precisão
y_predLinear <- predict(modelosvm, newdata = dados)
y_origLinear <- dados$CP
n_svm <- length(y_origLinear)
svmlinear <- data.frame(obs = dados$CP, pred=y_predLinear)
caret::defaultSummary(svmlinear)

##      RMSE  Rsquared       MAE
## 0.6034209 0.4200786 0.4297572
```

```
#Gráfico de predição SVM Linear
library(ggplot2)
ggplot(dados, aes(x = CP, y = y_predLinear)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2,
              linetype = "dashed") +
  labs(title = "Predicao do Modelo SVM Linear",
       x = "Valores Reais",
       y = "Valores Preditos")+
  theme_bw()
```

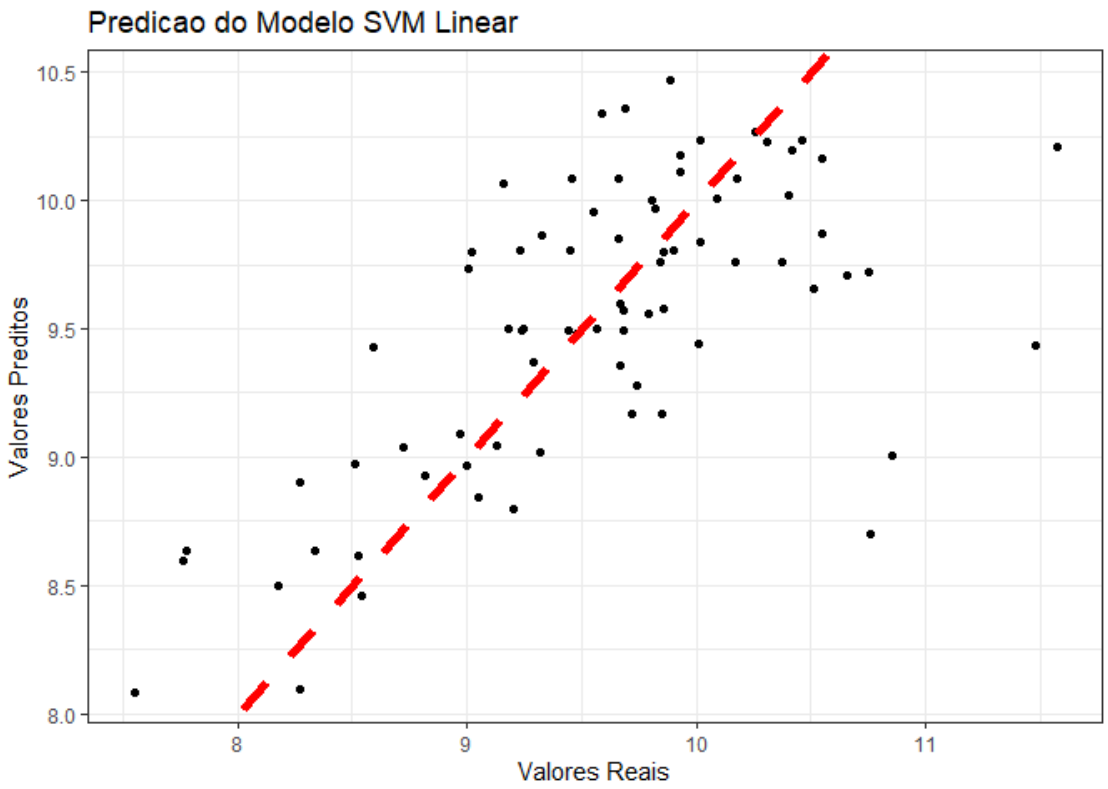

Figura 8. Relação entre os valores originais e os valores preditos por meio do modelo SVM com kernel linear.

Foi ajustado o modelo com o método *svmPoly*, com validação cruzada *leave-one-out*. Parâmetros testados incluíram *C* nas faixas de 0,01 a 100 e o grau do polinômio entre 2, 3 e 4. Um alto grau de polinômio pode promover um sobreajuste do modelo aos dados, causando *overfitting*. Verificou-se que o parâmetro *C* igual a 1 e o grau de polinômio igual a 2 promoveram os menores valores do RMSE. O R^2 utilizando os dados originais foi de 64,74%, enquanto o R^2 da validação cruzada foi de 49,36%. Isso indica que o modelo utilizando o kernel polinomial apresentou uma melhor capacidade preditiva de novos dados, ou seja, menor variância.

```
set.seed(1)
modelosvm1 <- caret::train(CP~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
                  data = dados,
                  method = "svmPoly",#kernel polinomial
                  tuneGrid = expand.grid(
                    degree = c(2, 3, 4),  # Grau do polinômio
                    scale = c(0.001, 0.01, 0.1),  # Escala do kernel
                    C = c(0.01, 0.1, 1, 10, 100)),#Parâmetro de regulariz.
                  trControl = trainControl(method = "LOOCV"))
modelosvm1
```

```
## Support Vector Machines with Polynomial Kernel
##
## 78 samples
##  6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   degree  scale  C      RMSE       Rsquared     MAE
##   2       0.001  1e-02  0.7995171  0.576888524  0.6150331
##   2       0.001  1e-01  0.7905685  0.003222128  0.6085105
##   2       0.001  1e+00  0.7188328  0.340939740  0.5502501
##   2       0.001  1e+01  0.6326458  0.371810745  0.4939084
##   2       0.001  1e+02  0.6430137  0.349250342  0.5040657
##   2       0.010  1e-02  0.7904524  0.002355709  0.6083990
##   2       0.010  1e-01  0.7179359  0.344900389  0.5498387
##   2       0.010  1e+00  0.6292319  0.378864979  0.4936600
##   2       0.010  1e+01  0.5806022  0.461559213  0.4516567
##   2       0.010  1e+02  0.5279510  0.553823233  0.4122302
##   2       0.100  1e-02  0.7125408  0.356525990  0.5483787
##   2       0.100  1e-01  0.5566332  0.530588440  0.4316011
##   2       0.100  1e+00  0.5135990  0.579390199  0.3978696
##   2       0.100  1e+01  0.5406833  0.533031316  0.4247977
##   2       0.100  1e+02  0.5496356  0.520538857  0.4252212
##   3       0.001  1e-02  0.7991168  0.541138950  0.6148230
##   3       0.001  1e-01  0.7846014  0.043039729  0.6036627
##   3       0.001  1e+00  0.7070026  0.320574231  0.5453033
##   3       0.001  1e+01  0.6370271  0.356498993  0.4985576
```

```
##   3       0.001  1e+02  0.6231426  0.383031042  0.4897262
##   3       0.010  1e-02  0.7842115  0.048118182  0.6032938
##   3       0.010  1e-01  0.7050538  0.324285794  0.5443769
##   3       0.010  1e+00  0.6137356  0.405750055  0.4817864
##   3       0.010  1e+01  0.5434301  0.528314605  0.4249219
##   3       0.010  1e+02  0.5659866  0.492528732  0.4382525
##   3       0.100  1e-02  0.6607248  0.444705660  0.5122751
##   3       0.100  1e-01  0.5553977  0.509310192  0.4231683
##   3       0.100  1e+00  0.6157314  0.423918773  0.4624769
##   3       0.100  1e+01  0.7905297  0.273956045  0.5118856
##   3       0.100  1e+02  0.9303052  0.168088616  0.6298161
##   4       0.001  1e-02  0.7987163  0.501125426  0.6146119
##   4       0.001  1e-01  0.7770371  0.146483475  0.5968109
##   4       0.001  1e+00  0.6921715  0.326940752  0.5365730
##   4       0.001  1e+01  0.6282718  0.374479230  0.4918622
##   4       0.001  1e+02  0.5990193  0.428078654  0.4645266
##   4       0.010  1e-02  0.7757193  0.167557616  0.5952108
##   4       0.010  1e-01  0.6845998  0.353501100  0.5311490
##   4       0.010  1e+00  0.5856806  0.462075849  0.4558245
##   4       0.010  1e+01  0.5418982  0.530455564  0.4178135
##   4       0.010  1e+02  0.6046802  0.434572708  0.4593245
##   4       0.100  1e-02  0.5981801  0.492306691  0.4664163
##   4       0.100  1e-01  0.7508781  0.254385886  0.4831061
##   4       0.100  1e+00  1.1060620  0.086811581  0.5880151
##   4       0.100  1e+01  1.0932339  0.051764215  0.6837409
##   4       0.100  1e+02  1.5780989  0.035294327  0.8291114
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were degree = 2, scale = 0.1 and C = 1.
```

```
#Estatísticas de precisão
y_predPoly <- predict(modelosvm1, newdata = dados)
y_origPoly <- dados$CP
n_svm <- length(y_origPoly)
svmPoly <- data.frame(obs = dados$CP, pred=y_predPoly)
caret::defaultSummary(svmPoly)
```

```
##      RMSE  Rsquared       MAE
## 0.4728262 0.6474707 0.3513191
```

```
#Gráfico de predição SVM Polinomial
ggplot(dados, aes(x = CP, y = y_predPoly)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2,
              linetype = "dashed") +
  labs(title = "Predicao do Modelo SVM Polinomial",
       x = "Valores Reais",
       y = "Valores Preditos")+
  theme_bw()
```

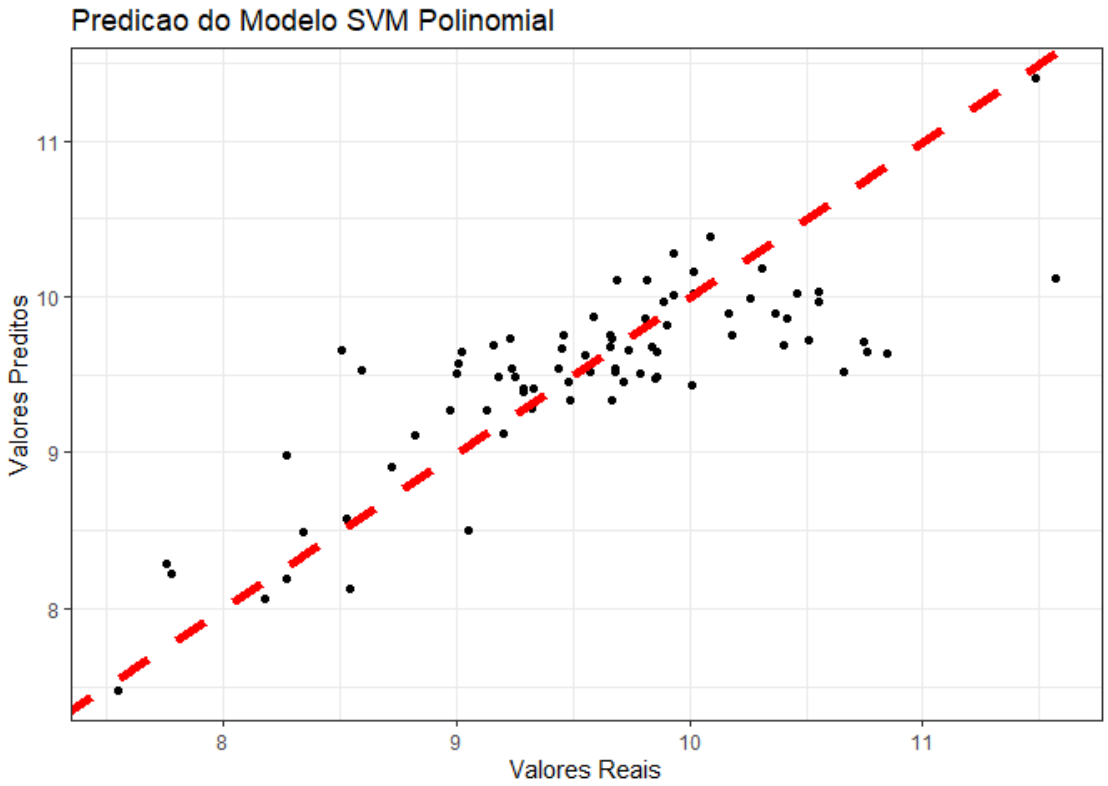

Figura 9. Relação entre os valores originais e os valores preditos por meio do modelo SVM com kernel polinomial.

Foi ajustado o modelo com o método *svmRadial*, com validação cruzada *leave-one-out*. Parâmetros testados incluíram *C* nas faixas de 0,01 a 100 e o sigma entre 0,001 a 1. Verificou-se que o parâmetro *C* igual a 1 e o sigma igual a 0,1 promoveram os menores valores do RMSE. O R^2 utilizando os dados originais foi de 64,75%, enquanto o R^2 da validação cruzada foi de 37,57%. Isso indica que o modelo utilizando o kernel radial apresentou uma menor capacidade preditiva de novos dados, ou seja, menor variância, em comparação com o *kernel* polinomial.

```
set.seed(1)
modelosvm2 <- caret::train(CP~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
                  data = dados,
                  method = "svmRadial",#kernel radial
                  tuneGrid = expand.grid(C = c(0.01, 0.1, 1, 10, 100),
                                         sigma = c(0.001, 0.01, 0.1, 1)),
                  trControl = trainControl(method = "LOOCV"))
modelosvm2

## Support Vector Machines with Radial Basis Function Kernel
##
## 78 samples
##  6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
```

```
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   sigma  C      RMSE       Rsquared     MAE
##   0.001  1e-02  0.7995335  0.578552734  0.6150398
##   0.001  1e-01  0.7907682  0.004858956  0.6086373
##   0.001  1e+00  0.7197408  0.339883078  0.5506062
##   0.001  1e+01  0.6315022  0.375289537  0.4930038
##   0.001  1e+02  0.6353410  0.361835485  0.4992553
##   0.010  1e-02  0.7923194  0.033751762  0.6096363
##   0.010  1e-01  0.7246802  0.348171345  0.5526712
##   0.010  1e+00  0.6352152  0.375725510  0.5004981
##   0.010  1e+01  0.5642424  0.494518978  0.4375464
##   0.010  1e+02  0.5628677  0.494792746  0.4360043
##   0.100  1e-02  0.7643493  0.349717518  0.5794633
##   0.100  1e-01  0.6322315  0.493353048  0.4711600
##   0.100  1e+00  0.5505418  0.522073857  0.3994555
##   0.100  1e+01  0.5882341  0.453498981  0.4332889
##   0.100  1e+02  0.7117754  0.304513233  0.5403901
##   1.000  1e-02  0.7906367  0.002780753  0.6053672
##   1.000  1e-01  0.7128168  0.381463782  0.5398187
##   1.000  1e+00  0.6153560  0.391590611  0.4584255
##   1.000  1e+01  0.7556687  0.241700371  0.5762912
##   1.000  1e+02  0.8983315  0.148894734  0.6985022
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were sigma = 0.1 and C = 1.
```

```
#Estatísticas de precisão
y_predRad <- predict(modelosvm2, newdata = dados)
y_origRad <- dados$CP
n_svm <- length(y_origRad)
svmRad <- data.frame(obs = dados$CP, pred=y_predRad)
caret::defaultSummary(svmRad)
```

```
##      RMSE  Rsquared       MAE
## 0.4832748 0.6475834 0.3403544
```

```
#Gráfico de predição SVM Radial
ggplot(dados, aes(x = CP, y = y_predRad)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2,
              linetype = "dashed") +
  labs(title = "Predicao do Modelo SVM Radial",
       x = "Valores Reais",
       y = "Valores Preditos")+
  theme_bw()
```

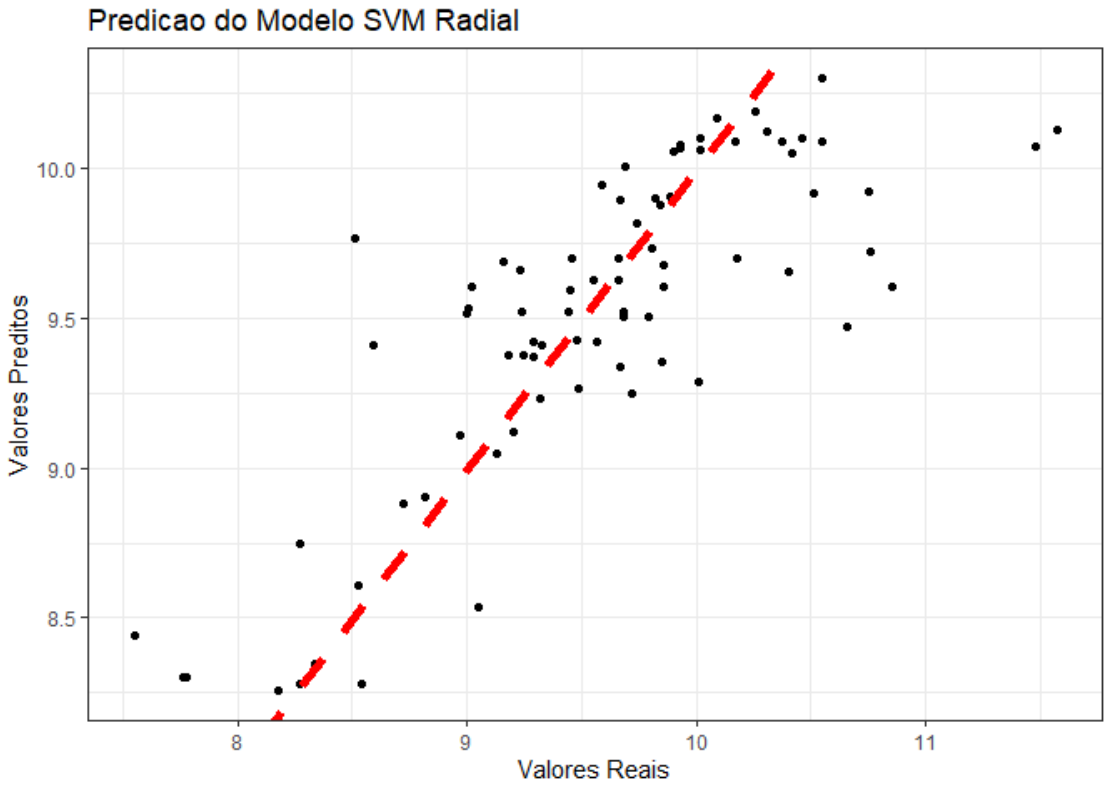

Figura 10. Relação entre os valores originais e os valores preditos por meio do modelo SVM com kernel radial.

4.2.7 Redes neurais

4.2.7.1 Teoria

Redes neurais são modelos computacionais inspirados no cérebro humano, compostos de camadas de neurônios artificiais. São eficazes para modelagem de relações complexas e não lineares, com a capacidade de aprender padrões detalhados (RAUBER, 2005). Útil para predições precisas em estudos onde as relações entre as variáveis são complexas, como resposta de cultivares a múltiplas variáveis ambientais.

Uma rede neural é um modelo computacional inspirado no funcionamento do cérebro humano, formado por um grande número de unidades interligadas chamadas de neurônios (RAUBER, 2005). Cada neurônio processa informações e aplica uma função matemática, conhecida como função de ativação, que transforma a entrada recebida em uma saída específica (WU; FENG, 2018). As conexões entre os neurônios possuem pesos, que determinam a intensidade com que um neurônio influencia outro. Esses pesos são ajustados durante o processo de aprendizado da rede, representando a memória da rede neural. O aprendizado ocorre quando a rede recebe dados de entrada e ajusta os pesos para melhorar a precisão das suas respostas (WANG, 2003).

A saída da rede neural depende de três fatores principais: a forma como os neurônios estão conectados, os valores dos pesos atribuídos às conexões e as funções de ativação utilizadas. Por meio

dessas características, a rede neural pode modelar e aproximar relações complexas entre variáveis, sendo capaz de resolver problemas como classificação, regressão e reconhecimento de padrões. Apesar de sua estrutura matemática, uma rede neural é, essencialmente, uma aproximação de algoritmos ou funções encontradas na natureza, podendo também ser usada para simular estratégias lógicas de tomada de decisão. Esse modelo se destaca pela flexibilidade e capacidade de lidar com problemas de grande complexidade, tornando-se uma ferramenta poderosa em diversas áreas do conhecimento.

4.2.7.2 Aplicação em R

O script abaixo demonstra como implementar um modelo de Redes Neurais utilizando o pacote *neuralnet* no R. Nesse exemplo, a variável dependente (CP) é predita com base em várias variáveis explicativas relacionadas (RSFM, RSFF, RFMC, RFFC, SFMC e SFFC). A estrutura da rede neural foi configurada com duas camadas ocultas, contendo 4 e 2 neurônios, respectivamente. Inicialmente, as variáveis foram padronizadas para garantir que todas tenham a mesma escala. A função *scale()* normaliza os dados para média zero e desvio padrão 1. A configuração *hidden = c(4, 2)* define duas camadas ocultas com 4 e 2 neurônios, respectivamente. O parâmetro *threshold = 0.1* define o critério de parada baseado no erro.

Durante o treinamento de uma rede neural, o algoritmo ajusta os pesos das conexões entre os neurônios para minimizar a diferença entre os valores preditos e os valores reais. O *threshold* define o valor mínimo que o erro da rede deve atingir para que o treinamento seja considerado concluído. Assim que o erro total ficar menor ou igual ao valor do *threshold*, o treinamento é interrompido. Se for informado um valor muito baixo, o modelo continuará treinando até atingir um erro pequeno, o que pode resultar em *overfitting*. No entanto, com um valor muito alto o treinamento é interrompido antes que o modelo tenha identificado os padrões dos dados, resultando em *underfitting*.

```
dados1 <- data.frame(scale(dados[,6:34]))#padronozação dos dados

set.seed(12)
modelorn <- neuralnet::neuralnet(CP~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
                  data=dados1,
                  hidden=c(4,2),
                  threshold = 0.1,
                  linear.output = TRUE)
```

A figura 11 apresenta a estrutura da rede neural artificial utilizada para predizer a variável dependente CP, de acordo com as especificações realizadas no modelo. Pode-se observar que a rede neural é composta por três tipos de camadas: a camada de entrada, representada pelas variáveis

preditoras, duas camadas ocultas com 4 e 2 nós, respectivamente, responsáveis pelo processamento intermediário e a camada de saída, que contém um único nó e fornece o valor predito para a variável CP. As conexões entre os neurônios indicam como os sinais das variáveis de entrada são processados e propagados até a camada de saída.

Cada conexão possui um peso associado, representado pelos valores em preto, que determina a influência de uma variável ou neurônio sobre o próximo neurônio. Pesos positivos indicam uma influência de aumento no sinal transmitido, enquanto pesos negativos indicam uma influência de redução. Além disso, os valores em azul correspondem aos *biases* dos nós, que são somados às entradas antes de serem processadas pela função de ativação. Os *biases* permitem que a rede ajuste os dados de forma mais flexível, deslocando as funções de ativação e aumentando a capacidade do modelo de aprender padrões complexos.

Durante o treinamento da rede, os pesos e os *biases* foram ajustados para minimizar o erro entre os valores preditos e os valores reais. Nesse modelo, o erro final obtido foi de 9,452672, com o treinamento convergindo após 734 iterações. Os pesos mais altos nas conexões indicam que determinadas variáveis têm maior influência para o modelo, enquanto as camadas ocultas e os *biases* ajudam a capturar relações não lineares e padrões complexos nos dados. Essa visualização permite compreender como a rede neural processa as informações e como as variáveis preditoras contribuem para a predição da variável resposta.

```
plot(modelorn, rep = "best")
```

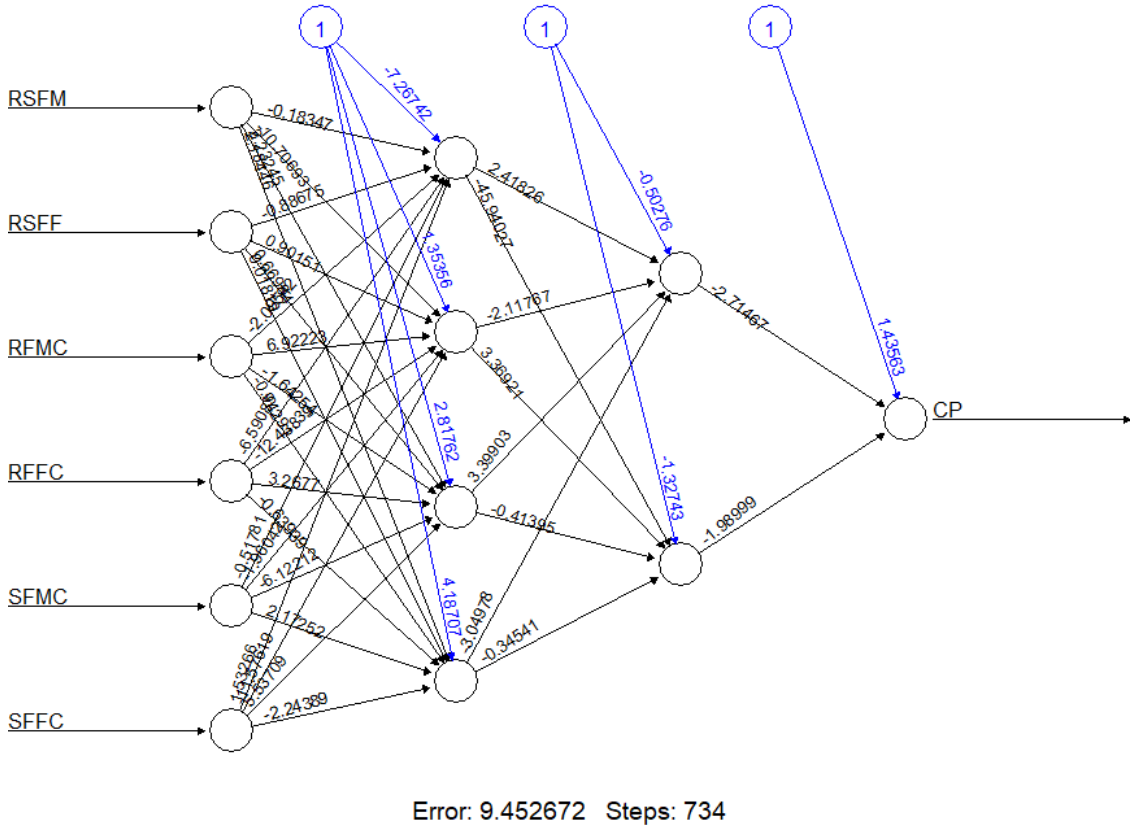

Figura 11. Estrutura da rede neural com a camada de entrada, representada pelas variáveis preditoras, duas camadas ocultas com 4 e 2 neurônios, respectivamente, e a camada de saída que fornece o valor predito para a variável CP. As conexões entre os neurônios indicam como os sinais das variáveis de entrada são processados e propagados até a camada de saída.

Para calcular as estatísticas de precisão é necessário obter novamente os dados na escala original. Isso foi realizado, por meio da multiplicação dos valores preditos pelo desvio padrão da escala original e somado a média da escala original. O modelo ajustado pela rede neural apresentou um R^2 de 0,7404, ou seja, explicou 74,04% da variabilidade total da CP por meio das variáveis preditores. Isso indica um ajuste adequado para predição da CP, por meio de variáveis meteorológicas.

```
y_predrnPadron <- predict(modelorn, dados1)
y_predrnSemPadr <- y_predrnPadron*sd(dados$CP) + median(dados$CP)
y_orgrn <- dados$CP
ngb <- length(y_orgrn)

#R²
1-(sum((y_orgrn-y_predrnSemPadr)^2)/
     sum((y_orgrn-mean(y_orgrn))^2))

## [1] 0.7404499
```

```r
#MAPE
sum(abs(y_orgrn-y_predrnSemPadr)/y_orgrn)/ngb*100
```

```
## [1] 3.248481
```

```r
#MAE
sum(abs(y_orgrn-y_predrnSemPadr))/ngb
```

```
## [1] 0.3101854
```

```r
#RMSE
sqrt((sum((y_orgrn-y_predrnSemPadr)^2))/ngb)
```

```
## [1] 0.4014065
```

```r
#Gráfico de predição SVM Radial
ggplot2::ggplot(dados1, ggplot2::aes(x = dados$CP, y = y_predrnSemPadr)) +
  ggplot2::geom_point() +
  ggplot2::geom_abline(intercept = 0,
               slope = 1,
               color = "red",
               linewidth = 2,
               linetype = "dashed") +
  ggplot2::labs(title = "Predicao do Modelo Rede Neural",
       x = "Valores Reais",
       y = "Valores Preditos")+
  ggplot2::theme_bw()
```

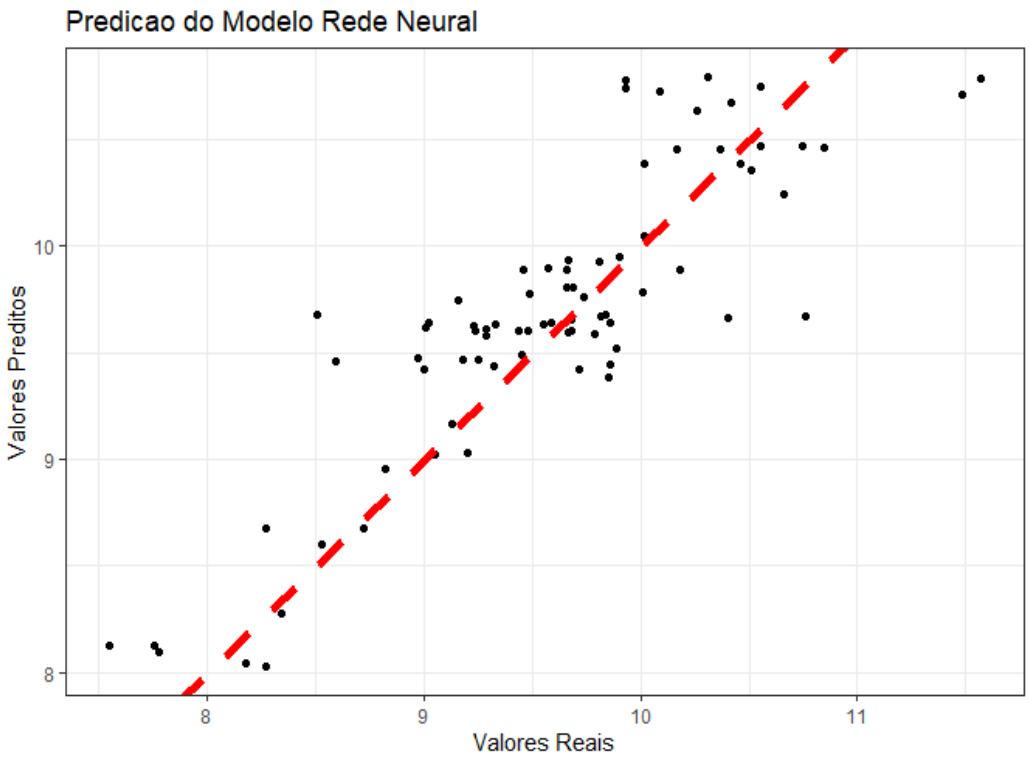

Figura 12. Relação entre os valores originais e os valores preditos por meio do modelo de Redes Neurais.

4.3 Referências

BATISTELLA, D. *et al*. Comparative analysis of orbital sensors in soybean yield estimation by the random forest algorithm. **Ciência e Agrotecnologia**, v. 47, p. e002423, 2023.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5-32, 2001.

BREIMAN, L. **Classification and regression trees**. Routledge, 2017, 368p.

BRERETON, R. G.; LLOYD, G. R. Support vector machines for classification and regression. **Analyst**, v. 135, n. 2, p. 230-267, 2010.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. São Paulo, 9. ed., n. 1, 2017.

CHEROBINI, L. Relações lineares entre caracteres de cultivares de alfafa. **Sigmae**, v. 13, n. 4, p. 41-51, 2024.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of Statistics**, p. 1189-1232, 2001.

HAIR, J. F. et al. **Análise multivariada de dados**. Bookman editora, 2009,688p.

KHAN, S. N.; LI, D.; MAIMAITIJIANG, M. A geographically weighted random forest approach to predict corn yield in the US corn belt. **Remote Sensing**, v. 14, n. 12, p. 2843, 2022.

KONRAD, M. *et al*. Predição da massa fresca e massa seca da parte aérea da planta de teosinto em função de caracteres morfológicos. **Sigmae**, v. 12, n. 3, p. 10-17, 2023.

LORENA, A. C.; CARVALHO, A. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43-67, 2007.

LORO, M. V. *et al*. Relações lineares entre caracteres do pendão e da espiga em bases genéticas de milho. **Journal of Environmental Analysis and Progress**, v. 9, n. 2, p. 065-078, 2024a.

LORO, M. V. *et al*. Relações lineares entre variáveis meteorológicas e caracteres fenológicos, morfológicos e produtivos em bases genéticas de milho. **Revista Vivências**, v. 20, n. 41, p. 95-111, 2024b.

MEUS, L. D. *et al*. Irrigated rice yield plateaus are caused by management factors in Argentina. **Agronomy for Sustainable Development**, v. 44, n. 6, p. 56, 2024.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in Neurorobotics**, v. 7, p. 21, 2013.

NICK, T. G.; CAMPBELL, K. M. Logistic regression. **Topics in biostatistics**, p. 273-301, 2007.

NOBLE, W. S. What is a support vector machine?. **Nature biotechnology**, v. 24, n. 12, p. 1565-1567, 2006.

RAUBER, T. W. Redes neurais artificiais. **Universidade Federal do Espírito Santo**, v. 29, 2005.

REIS, M. B. *et al*. Árvore de regressão para previsão da produtividade de matéria fresca da parte aérea de teosinto em função de variáveis meteorológicas. **Sigmae**, v. 12, n. 3, p. 24-31, 2023.

SCARTON, V. D. B. *et al*. Influence of meteorological variables and geographic factors in the selection of soybean lines. **Revista de Agricultura Neotropical**, v. 10, n. 3, p. e7246-e7246, 2023.

SMIDT, E. R. *et al*. Identifying field attributes that predict soybean yield using random forest analysis. **Agronomy Journal**, v. 108, n. 2, p. 637-646, 2016.

WANG, S. C. Artificial neural network. **Interdisciplinary computing in java programming**. V. 743, p. 81-100, 2003.

WU, Y.; FENG, J. Development and application of artificial neural network. **Wireless Personal Communications**, v. 102, p. 1645-1656, 2018.

5. LINK VIDEOAULA

https://drive.google.com/file/d/1qGEBum6yknq6wgJoEN02-Tplvxe5QVxN/view?usp=sharing

6. EXERCÍCIO COM GABARITO

6.1 Exercício proposto

Utilizando o banco de dados ***dados***, que contém as variáveis MG, RSFM, RSFF, RFMC, RFFC, SFMC e SFFC, crie modelos preditivos para estimar a variável MG com base nas variáveis explicativas RSFM, RSFF, RFMC, RFFC, SFMC e SFFC. Para isso, utilize os algoritmos Árvore de Regressão, *Random Forest* e *Gradient Boosting*. Avalie o desempenho de cada modelo. Compare os resultados obtidos por cada abordagem, considerando métricas como a raiz do erro quadrático médio (RMSE) e o coeficiente de determinação (R^2). Apresente o script!

6.2 Resolução do exercício

O estudo avaliou o desempenho de diferentes modelos preditivos para estimar a variável MG com base em variáveis meteorológicas. Inicialmente, utilizou-se o algoritmo de árvore de regressão, ajustado com valores variados do parâmetro de complexidade (*cp*). O modelo foi treinado com validação cruzada *Leave-One-Out*, sendo o valor ótimo de *cp* identificado como 0,05. As estatísticas de precisão do modelo indicaram um R^2 de 0,55 e RMSE de 19,46. Em seguida, foi treinado o modelo de *Random Forest*, também utilizando validação LOOCV. O modelo selecionou automaticamente o número ótimo de variáveis por nó (*mtry = 2*). Esse modelo apresentou melhor desempenho preditivo em comparação ao de árvore de regressão, com um R^2 de 0,83 e RMSE de 12,11.

Aplicou-se o modelo de *Gradient Boosting*, utilizando uma grade de hiperparâmetros para explorar combinações de número de árvores, profundidade de interação, taxa de aprendizado e

número mínimo de observações por nó terminal. O modelo selecionado otimizou os parâmetros para 300 árvores, profundidade de interação igual a 1, taxa de aprendizado de 0,01 e número mínimo de observações por nó igual a 5. Este modelo apresentou um desempenho intermediário, com R^2 de 0,58, e RMSE de 22,40. O modelo de *Random Forest* apresentou maior capacidade preditiva dos dados. Já o modelo de árvore de regressão apresentou maior interpretabilidade. No entanto, o *Gradient Boosting* exibiu o maior R^2 da validação cruzada, indicando a maior capacidade preditiva de novos dados. Logo, observa-se que o *Gradient Bosting* faz um equilíbrio entre viés e variância.

```
# URL do arquivo no GitHub (link para o arquivo .xlsx raw)
url <- "https://github.com/muriloloro/Modelos-de-Predicao/raw/main/dados.xlsx"
library(httr)
library(readxl)
# Definir o caminho temporário para salvar o arquivo
temp_file <- tempfile(fileext = ".xlsx")
GET(url, write_disk(temp_file, overwrite = TRUE))

dados <- read_excel(temp_file)

#----------------------------ARVORE DE REGRESSÃO

#Ajustando parametro cp
library(caret)
set.seed(123)
caret::train(MG~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
               data = dados,
               method = "rpart",
               tuneGrid = data.frame(cp = c(0, 0.01, 0.05, 0.1, 0.15, 0.2)),
               trControl = trainControl(method = "LOOCV"))

## CART
##
## 78 samples
##  6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   cp    RMSE      Rsquared   MAE
##   0.00  22.88898  0.4002642  18.00453
##   0.01  23.04066  0.3938595  18.15184
##   0.05  22.72931  0.4021351  18.34550
##   0.10  25.97180  0.2379318  21.25141
##   0.15  24.93364  0.2838509  20.04832
##   0.20  24.93364  0.2838509  20.04832
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.05.

#----------------------MODELO DE ÁRVORE DE REGRESSÃO
modeloar <- rpart::rpart(MG~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
```

```
              data = dados,
                     method = "anova",
                     cp = 0.05)

#contribuição das variáveis
barplot(modeloar$variable.importance)
```

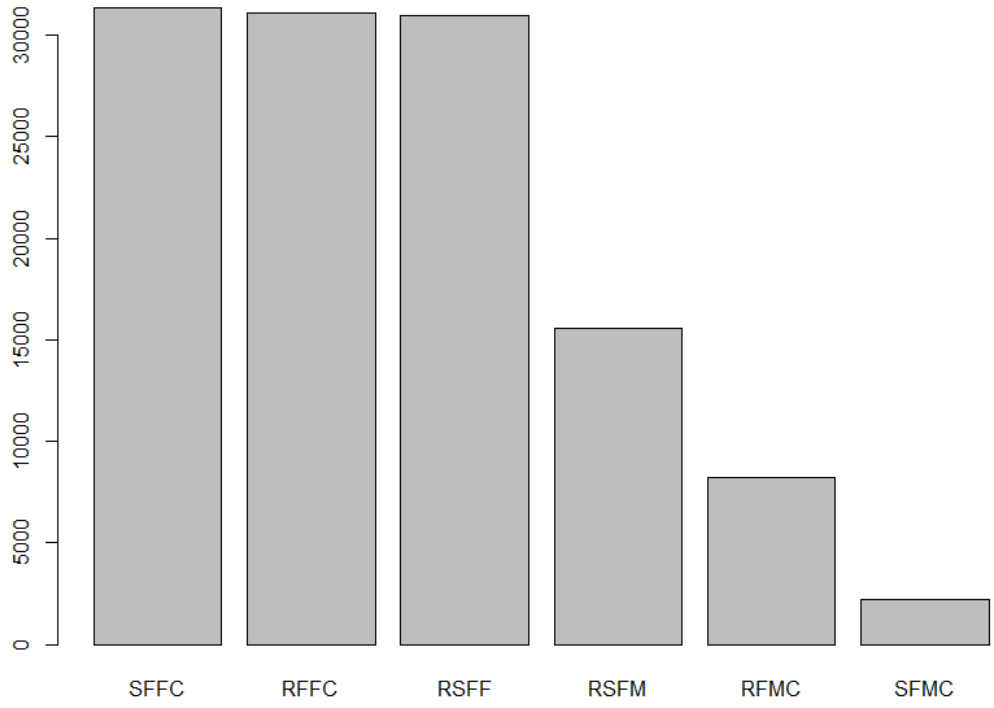

Figura 1. Contribuição das variáveis meteorológicas para predição CP por meio da árvore de decisão baseada em regressão.

```
# plotar a arvore

rpart.plot::rpart.plot(modeloar,
            type = 0,
            extra = 101,
            box.palette = c("yellow","green"),
            branch.lty=2,
            shadow.col = "black",
            nn=TRUE,
            cex = 1.2)
```

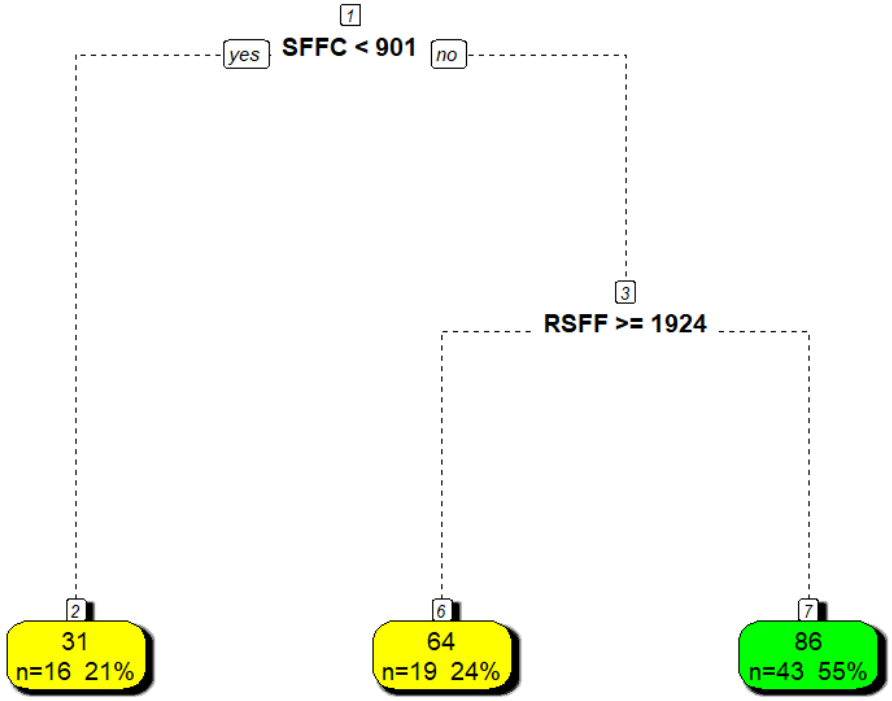

Figura 2. Árvore de decisão baseada em regressão para predição da CP em função de variáveis meteorológicas.

```
# Estatísticas de precisão do modelo - MODO 1
y_predito <- predict(modeloar, newdata = dados)
y_original <- dados$MG
n <- length(y_original)

#R²
1-(sum((y_original-y_predito)^2)/
  sum((y_original-mean(y_original))^2))

## [1] 0.5537455

#MAPE
sum(abs(y_original-y_predito)/y_original)/n*100

## [1] 128.704

#MAE
sum(abs(y_original-y_predito))/n

## [1] 15.66597

#RMSE
sqrt((sum((y_original-y_predito)^2))/n)

## [1] 19.45807

#------------------------------RANDOM FOREST
```

```r
set.seed(1)
modelorf <- caret::train(MG~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
                  data = dados,
                  method = "rf",
                  ntree = 100,
                  trControl = trainControl(method = "LOOCV"))

modelorf
```

```
## Random Forest
##
## 78 samples
##  6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##   2     22.76778  0.3951925  18.68565
##   4     23.21233  0.3767964  18.95591
##   6     23.21612  0.3759598  18.99790
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
```

```r
# Estatísticas de precisão do modelo - MODO 1
y_predit <- predict(modelorf, newdata = dados)
y_origina <- dados$MG
nrf <- length(y_origina)

#R²
1-(sum((y_origina-y_predit)^2)/
  sum((y_origina-mean(y_origina))^2))
```

```
## [1] 0.8271835
```

```r
#MAPE
sum(abs(y_origina-y_predit)/y_origina)/nrf*100
```

```
## [1] 89.94141
```

```r
#MAE
sum(abs(y_origina-y_predit))/nrf
```

```
## [1] 9.548622
```

```r
#RMSE
sqrt((sum((y_origina-y_predit)^2))/nrf)
```

```
## [1] 12.1088
```

```r
#---------------------------------------GRADIENT BOOSTING

#Treinamento do modelo
set.seed(1)
library(caret)
```

```r
modelogb <- caret::train(MG~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC,
                  data = dados,
                  method = "gbm",#define gradient boosting
                  verbose = FALSE,
                  trControl = trainControl(method = "LOOCV"),
                  tuneGrid = expand.grid(
  n.trees = c(50, 100, 150, 300),        # Número de árvores
  interaction.depth = c(1, 3, 5),    # Profundidade das árvores
  shrinkage = c(0.01, 0.05, 0.1),    # Taxa de aprendizado (learning rate)
  n.minobsinnode = c(5)               # Número mínimo de observações por nó ter
minal
))# Passando a grade de parâmetros

modelogb
```

```
## Stochastic Gradient Boosting
##
## 78 samples
##  6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 77, 77, 77, 77, 77, 77, ...
## Resampling results across tuning parameters:
##
##   n.trees  interaction.depth  shrinkage  RMSE      Rsquared   MAE
##    50      1                  0.01       25.11480  0.4192361  19.99676
##    50      1                  0.05       21.95671  0.4344097  17.67359
##    50      1                  0.10       21.50736  0.4564913  17.28342
##    50      3                  0.01       24.58289  0.4361448  19.75108
##    50      3                  0.05       22.05410  0.4272585  18.03856
##    50      3                  0.10       21.95465  0.4395226  17.62706
##    50      5                  0.01       24.75368  0.4084350  19.79946
##    50      5                  0.05       22.18525  0.4209909  17.59618
##    50      5                  0.10       23.31168  0.3840748  18.87609
##   100      1                  0.01       23.08516  0.4349510  18.54113
##   100      1                  0.05       21.64737  0.4480727  17.48399
##   100      1                  0.10       21.41159  0.4647289  17.39855
##   100      3                  0.01       22.72865  0.4354255  18.36070
##   100      3                  0.05       22.28267  0.4243216  17.98415
##   100      3                  0.10       22.95903  0.4031976  18.43292
##   100      5                  0.01       22.87835  0.4188999  18.31717
##   100      5                  0.05       22.34061  0.4218415  17.67095
##   100      5                  0.10       24.26636  0.3603846  20.07270
##   150      1                  0.01       22.17535  0.4438050  17.80917
##   150      1                  0.05       21.81309  0.4426386  17.54809
##   150      1                  0.10       21.71672  0.4534744  17.46570
##   150      3                  0.01       22.05112  0.4378284  17.73474
##   150      3                  0.05       22.68794  0.4094206  18.27502
##   150      3                  0.10       23.61963  0.3824627  19.08232
##   150      5                  0.01       22.34974  0.4175012  17.96550
##   150      5                  0.05       23.03477  0.3984559  18.40071
##   150      5                  0.10       25.08120  0.3375138  20.35990
##   300      1                  0.01       21.40028  0.4623407  17.20735
##   300      1                  0.05       22.12858  0.4334753  17.66289
##   300      1                  0.10       22.91474  0.4039613  18.44382
```

```
##   300      3                   0.01       21.92397  0.4351498  17.48975
##   300      3                   0.05       24.17705  0.3599113  19.51289
##   300      3                   0.10       25.39548  0.3261045  20.52796
##   300      5                   0.01       22.27152  0.4200642  17.93798
##   300      5                   0.05       23.90056  0.3736164  19.22800
##   300      5                   0.10       26.51865  0.3005596  21.62820
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 5
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were n.trees = 300, interaction.depth =
##  1, shrinkage = 0.01 and n.minobsinnode = 5.
```

```r
# Estatísticas de precisão do modelo - MODO 1
y_pred <- predict(modelogb, newdata = dados)
y_orig <- dados$MG
ngb <- length(y_orig)

#R²
1-(sum((y_orig-y_pred)^2)/
  sum((y_orig-mean(y_orig))^2))
```

```
## [1] 0.5847364
```

```r
#MAPE
sum(abs(y_orig-y_pred)/y_orig)/ngb*100
```

```
## [1] 144.3091
```

```r
#MAE
sum(abs(y_orig-y_pred))/ngb
```

```
## [1] 15.0823
```

```r
#RMSE
sqrt((sum((y_orig-y_pred)^2))/ngb)
```

```
## [1] 18.77026
```

7. ANÁLISE DE COMPONENTES PRINCIPAIS

7.1 Considerações sobre escalas de medida das variáveis

A Análise de Componentes Principais (PCA) é uma técnica utilizada para reduzir a dimensionalidade dos dados e identificar os principais padrões de variação das variáveis. No entanto, em um conjunto de dados com variáveis de diferentes escalas e unidades, como altura de planta (AP, em cm), radiação solar global da semeadura ao florescimento masculino (RSFM, em MJ m^{-2}) e produtividade (PROD, em Mg ha^{-1}), estratégias pré análises precisam ser adotadas. Por exemplo, em um conjunto de dados a AP média foi de 175,30 cm, a RSFM média de 1857,59 MJ m^{-2} e PROD média de 3,06 Mg ha^{-1}. Logo, pode-se observar que a escala de medida da RSFM é expressivamente superior as escalas de medida da AP e PROD. Se essas variáveis são analisadas diretamente na PCA, aquelas com maior variância, como a RSFM, podem dominar os componentes principais. Isso ocorre porque a PCA é sensível à escala das variáveis, o que pode reduzir as contribuições de variáveis com menor variância, como a AP e PROD.

Para evitar vieses na análise e interpretação dos componentes principais e garantir que as variáveis com maior variância não dominem os componentes é necessário a padronização das variáveis antes da análise. A padronização, transforma cada variável para ter média zero e desvio-padrão igual a 1. Esse ajuste garante que todas as variáveis contribuam igualmente para a formação dos componentes principais, independentemente de suas unidades de medida originais.

No exemplo abaixo, foi realizada a ACP em dois cenários: com padronização dos dados e sem padronização dos dados. Os dados para os exemplos podem ser obtidos na plataforma GitHub pelo código abaixo:

```
#URL do arquivo no GitHub (link para o arquivo .xlsx raw)
url <- "https://github.com/muriloloro/Modelos-de-Predicao/raw/main/dados.xlsx"
library(httr)
library(readxl)

# Definir o caminho para salvar o arquivo
temp_file <- tempfile(fileext = ".xlsx")
GET(url, write_disk(temp_file, overwrite = TRUE))
dados <- read_excel(temp_file)
```

7.1.1 Cenário com padronização dos valores das variáveis

Na análise de ACP com padronização, os dados são transformados de forma que cada variável tenha média 0 e desvio padrão 1. Isso reduz a influência das diferentes escalas das variáveis. Por exemplo, uma variável com valores de alta magnitude como a RFMC, que tem um desvio padrão de 70,68 MJ m^{-2} pode dominar a análise, impedindo a expressão da variabilidade de outras variáveis

com valores menores, como a AP com desvio padrão de 10,44 cm. A padronização assegura que todas as variáveis tenham a mesma importância na análise, sem que nenhuma delas seja mais influente por causa de sua escala. As componentes principais refletiram as direções de maior variabilidade nos dados, levando em consideração a contribuição proporcional de cada variável (Figura 2).

```
#--------------------COM PADRONIZAÇÃO DAS VARIÁVEIS
library(factoextra)

# 1 - Selecionando variáveis
res.pca1 <- dados |>
  dplyr::select(AP, DE, CE, ME, MG, PROD, RSFM, RFMC)

# 2 - Padronizando variaveis
res.pca <- prcomp(res.pca1, scale. = TRUE)
res.pca

## Standard deviations (1, .., p=8):
## [1] 2.0271449 1.1131158 0.9975911 0.8888044 0.7139629 0.4992737 0.3004045
## [8] 0.1312844
##
## Rotation (n x k) = (8 x 8):
##              PC1         PC2         PC3         PC4         PC5         PC6
## AP    0.01242088 -0.63454203  0.42720859 -0.63111606 -0.02369205  0.12509906
## DE    0.44384621  0.09176692 -0.25987978 -0.21186379 -0.08147804  0.28771602
## CE    0.15345702 -0.60493103  0.14325204  0.73102363  0.01052723  0.12795389
## ME    0.45286866 -0.15053247 -0.27596240 -0.01003020 -0.25519538  0.05017632
## MG    0.46457355 -0.05635724 -0.25863314 -0.06689777 -0.13262348  0.10625562
## PROD  0.42439006 -0.06913327  0.06044948 -0.05422152  0.43054175 -0.78210885
## RSFM -0.33463723 -0.31063246 -0.43792923 -0.06487933 -0.57236539 -0.47450992
## RFMC  0.25746768  0.30965400  0.62526129  0.10337059 -0.63008506 -0.18583777
##              PC7           PC8
## AP   -0.01072947 -0.0078270250
## DE    0.76857705 -0.0450083679
## CE    0.18228278 -0.0766984970
## ME   -0.34363605  0.7143290705
## MG   -0.46134487 -0.6840398626
## PROD  0.10704429 -0.0007866778
## RSFM  0.17486525 -0.1104283057
## RFMC  0.05464453 -0.0407276428

# 3- Extrair os autovalores
eig.val <- get_eigenvalue(res.pca)

# 4 - Escolha do número de componentes principais
fviz_eig(res.pca,
               addlabels = TRUE,
               barfill = "orange",
               xlab = "Componentes Principais",
               ylab = "Variância explicada (%)",
               title = "",
               barcolor = "black")+
  theme_bw()+
  theme_classic()+
```

```
  theme(
    panel.grid = element_blank(),
    axis.text = element_text(family = "serif", colour = "black", size = 12),
    axis.title.x = element_text(size = 14, family = "serif"),
    axis.title.y = element_text(size = 14, family = "serif"),
    legend.text = element_text(size = 14, family = "serif"))
```

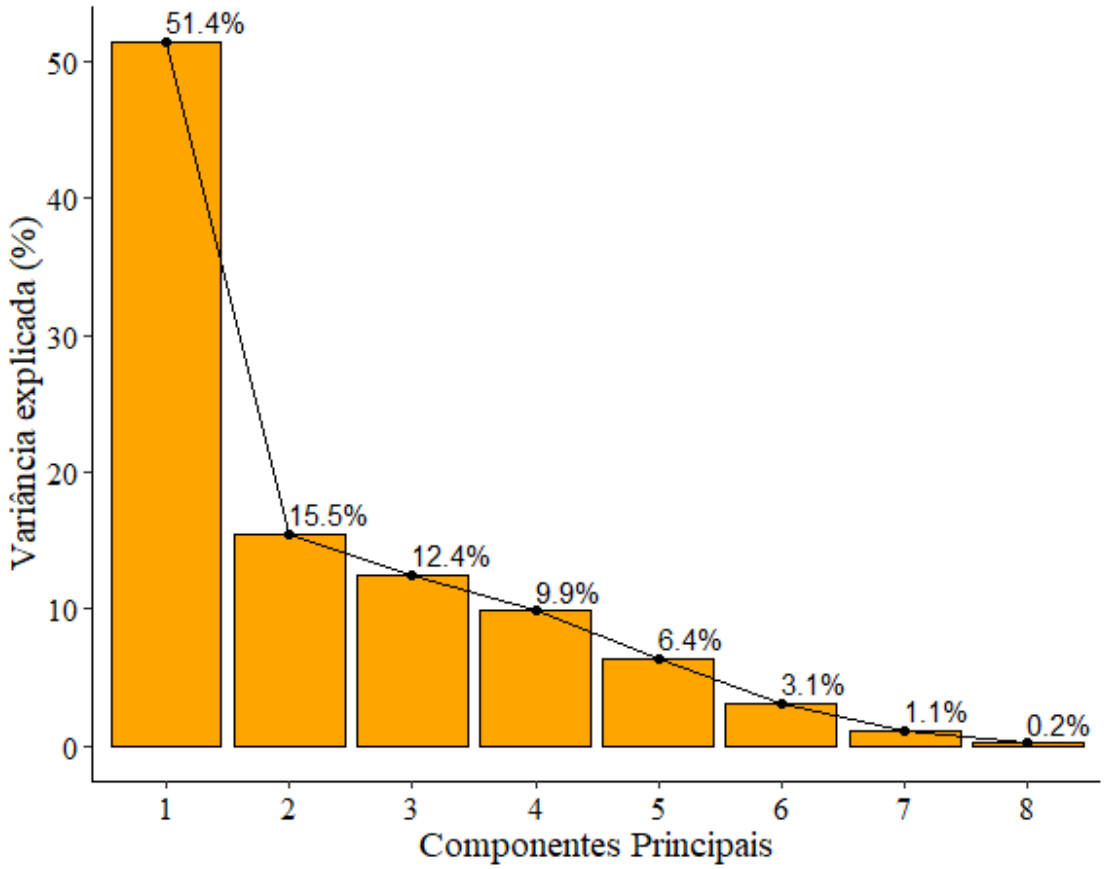

Figura 1. *Screeplot* para seleção do número de componentes principais baseado na porcentagem de variância explicada.

```
# Componentes principais
fviz_pca_biplot(res.pca,
                        # indivíduos
                        geom.ind = "point",
                        fill.ind = factor(dados$BaseGen, levels = c("SIMPLES", "
TRIPLO", "DUPLO", "VARIEDADE")),
                        col.ind = "black",
                        pointshape = 21, pointsize = 4,
                        palette = c("green", "purple", "yellow", "red"),  # Ajus
te a paleta conforme a ordem dos níveis
                        addEllipses = FALSE,
                        repel = TRUE,
                        title = "",
                        # variáveis
                        col.var = "black",       # Cor das siglas das variáveis
                        labelsize = 5,          # Tamanho da fonte das siglas da
```

```
s variáveis
                        font.var = 2,           # Negrito nas siglas das variáve
is
                        legend.title = list(fill = "")) +
  theme_bw() +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

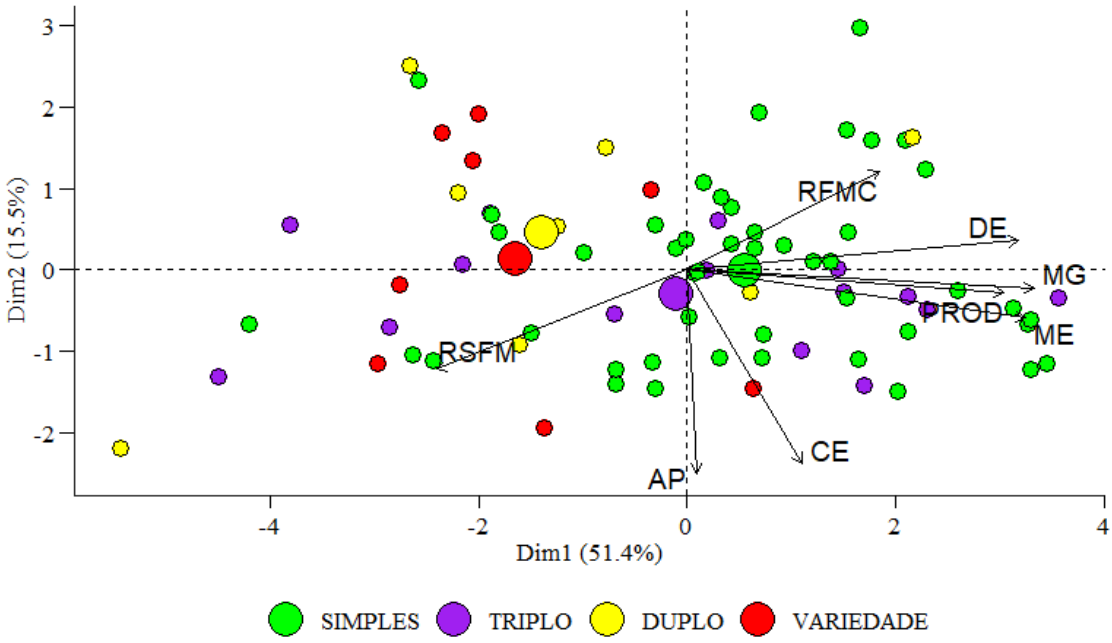

Figura 2. Análise de componentes principais com padronização dos dados.

7.1.2 Cenário sem padronização dos valores das variáveis

Sem a padronização dos dados, as variáveis com maiores escalas de variação tendem a dominar a análise, o que pode enviesar os resultados. No exemplo sem padronização, o desvio padrão das variáveis varia muito, o que significa que variáveis com maior dispersão como RSFM e RFMC tiveram maior peso nos componentes principais. Por exemplo, RSFM e RFMC contribuem

fortemente para o PC1 e PC2. Assim, as variáveis com menor dispersão como AP, DE e CE apresentaram uma contribuição menor para as componentes principais. Como resultado, a análise foi influenciada pelas variáveis RSFM e RFMC (Figura 4), enquanto variáveis com menor variação, tiveram uma contribuição subestimada.

```
# 1 - Selecionando variáveis
res.pca1 <- dados |>
  dplyr::select(AP, DE, CE, ME, MG, PROD, RSFM, RFMC)

# 2 - Padronizando variaveis
res.pca2 <- prcomp(res.pca1, scale. = FALSE)
res.pca2

## Standard deviations (1, .., p=8):
## [1] 118.4825037  54.8301491  36.3110701  10.4296288   4.3369404   1.1131785
## [7]   0.8976543   0.1801535
##
## Rotation (n x k) = (8 x 8):
##                PC1           PC2         PC3           PC4           PC5
## AP    0.004936641 -0.0079570030  0.01530067  0.997165077 -0.069412538
## DE   -0.002296661  0.0002917627  0.00946030 -0.002428577 -0.006463477
## CE   -0.001233516 -0.0002112223  0.01041056  0.023685961  0.126003918
## ME   -0.112083329 -0.0369374332  0.71495002  0.034934282  0.681357163
## MG   -0.132427814  0.0006818612  0.67638776 -0.058492707 -0.716956230
## PROD -0.008313527  0.0029001471  0.02133062  0.017104088  0.005021591
## RSFM  0.887599970 -0.4364992427  0.14387112 -0.011841213 -0.027377205
## RFMC -0.426574980 -0.8989058447 -0.09879364 -0.004507857 -0.014648855
##               PC6           PC7           PC8
## AP    0.01809291 -0.0139378808  0.0005434364
## DE    0.08453517  0.0164235468  0.9962136023
## CE   -0.97148538 -0.1795628283  0.0861707006
## ME    0.09659162 -0.0027158705 -0.0106827385
## MG   -0.08250784 -0.0262699658 -0.0040884769
## PROD -0.18050266  0.9831416091 -0.0010395243
## RSFM -0.00422166  0.0052434319  0.0008731660
## RFMC -0.00246849  0.0008883347  0.0003068027

# 3- Extrair os autovalores
eig.val <- get_eigenvalue(res.pca2)

# 4 - Escolha do número de componentes principais
fviz_eig(res.pca2,
               addlabels = TRUE,
               barfill = "orange",
               xlab = "Componentes Principais",
               ylab = "Variância explicada (%)",
               title = "",
               barcolor = "black")+
  theme_bw()+
  theme_classic()+
  theme(
    panel.grid = element_blank(),
    axis.text = element_text(family = "serif", colour = "black", size = 12),
```

```
    axis.title.x = element_text(size = 14, family = "serif"),
    axis.title.y = element_text(size = 14, family = "serif"),
    legend.text = element_text(size = 14, family = "serif"))
```

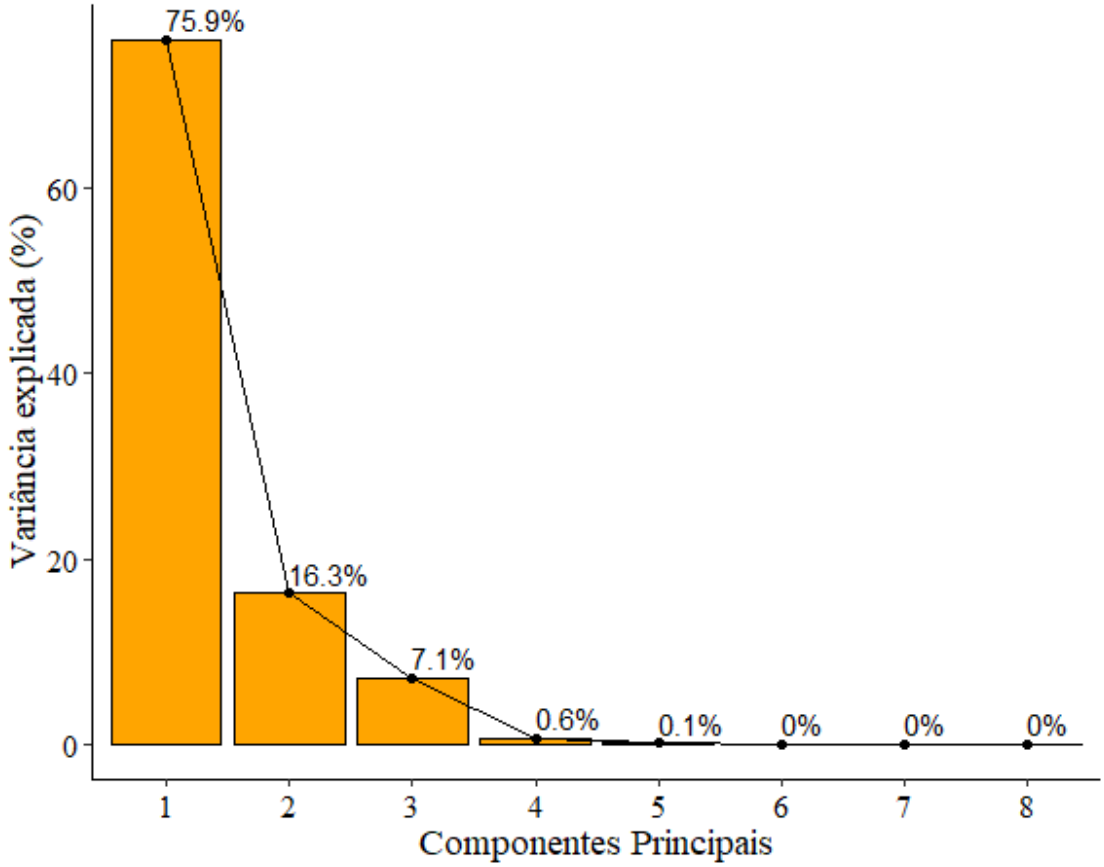

Figura 3. *Screeplot* para seleção do número de componentes principais baseado na porcentagem de variância explicada.

```
# Componentes principais
fviz_pca_biplot(res.pca2,
                        # indivíduos
                        geom.ind = "point",
                        fill.ind = factor(dados$BaseGen, levels = c("SIMPLES", "
TRIPLO", "DUPLO", "VARIEDADE")),
                        col.ind = "black",
                        pointshape = 21, pointsize = 4,
                        palette = c("green", "purple", "yellow", "red"),  # Ajus
te a paleta conforme a ordem dos níveis
                        addEllipses = FALSE,
                        repel = TRUE,
                        title = "",
                        # variáveis
                        col.var = "black",       # Cor das siglas das variáveis
                        labelsize = 5,          # Tamanho da fonte das siglas da
s variáveis
                        font.var = 2,           # Negrito nas siglas das variáve
is
```

```
                      legend.title = list(fill = "")) +
  theme_bw() +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

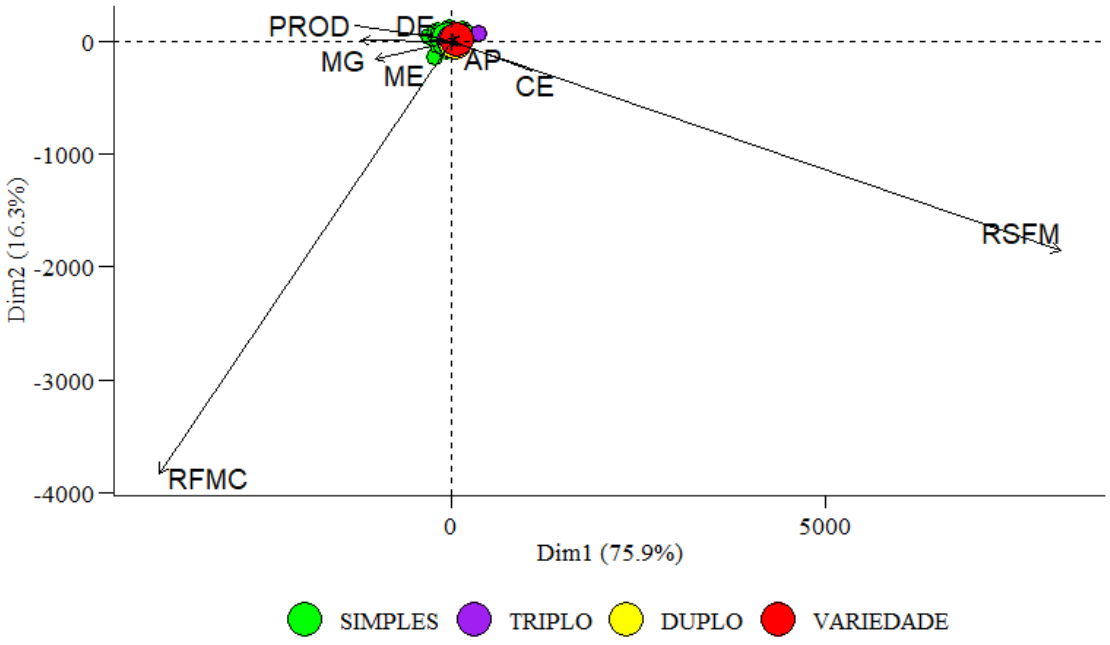

Figura 4. Análise de componentes principais sem a padronização dos dados

7.2 Considerações sobre a natureza das variáveis

No caso de variáveis qualitativas, como classe de qualidade nutricional (alta ou baixa), estas não podem ser diretamente incluídas na PCA, que foi projetada para variáveis quantitativas. Para incluir variáveis qualitativas, é necessário transformá-las em variáveis numéricas, neste caso como uma variável binária. Assim, a estratégia de converter variáveis categóricas em uma sequência de

variáveis binárias com valores 0 e 1 é uma maneira de fazer a ACP em um conjunto de dados com variáveis categóricas.

A PCA busca maximizar a variabilidade de todas as variáveis sobre os componentes principais, para que as contribuições das variáveis sejam equilibradas (JOLLIFFE; CADIMA, 2016; SCMITH, 2002). No entanto, quando a análise inclui variáveis categóricas, essas variáveis precisam ser transformadas em um formato numérico. A variabilidade total que uma variável categórica contribui no PCA depende do número de categorias e das frequências associadas a essas categorias. Por exemplo, variáveis com mais categorias geram mais informações na tabela de dados, resultando em uma soma de variabilidade maior. Por exemplo, considera-se uma variável que indica se os genótipos de milho têm alta ou baixa qualidade nutricional proteica dos grãos (coluna NUT do arquivo de dados utilizados nesse arquivo). Assim, essa variável será transformada em uma variável binária 0 (baixa) e 1 (alta). Agora, considera-se que ao invés de alta e baixa, as classes fossem: muito baixa, baixa, moderada, alta e muito alta, ou seja, a transformação dos dados categóricos para uma variável numérica seria 0, 1, 2, 3, 4 e 5, respectivamente. Assim, percebe-se que a variável com um maior número de classes possivelmente apresentaria uma maior variância em comparação com a variável com menor número de classes. Isso implicaria na contribuição de cada variável.

Além disso, se uma classe de uma variável categórica for muito frequente, sua contribuição à variabilidade também será elevada. Esse desequilíbrio faz com que variáveis categóricas com muitas categorias ou categorias muito frequentes dominem os componentes principais, distorcendo a análise. Isso compromete o objetivo do PCA, que é analisar todas as variáveis de forma equilibrada, para que suas contribuições aos componentes principais reflitam sua importância relativa no conjunto de dados. Quando variáveis categóricas com mais categorias ou categorias desequilibradas têm um peso maior, o PCA perde a capacidade de capturar os padrões gerais e passa a refletir um viés causado pela estrutura dos dados.

Uma solução eficaz para esse problema é o uso da Análise Fatorial de Dados Mistos (FAMD) (KASSAMBARA, 2017; PTAK; ERRICO; CHRISTENSEN, 2021). Esse método ajusta os cálculos da variabilidade para que tanto as variáveis contínuas quanto as categóricas sejam tratadas de maneira equilibrada. Para variáveis contínuas, a FAMD utiliza a variabilidade tradicional, como no PCA. Para variáveis categóricas, ajusta a contribuição das categorias, de forma que todas as variáveis originais, independentemente do número de categorias, tenham pesos similares (KASSAMBARA, 2017). Esse ajuste evita que variáveis categóricas com muitas categorias dominem a análise, proporcionando um equilíbrio entre variáveis contínuas e categóricas.

7.2.1 Cenário com variáveis mistas por ACP e FAMD, com frequência balanceada da variável categórica

Para este exemplo, foi considerada uma análise de componentes principais (PCA) e análise fatorial de dados mistos (FAMD) considerando as seguintes variáveis: AP, DE, CE ME e NUT (para PCA) e AP, DE, CE, ME e NUTB (para FAMD). Importante destacar que as variáveis NUT e NUTB indicam a qualidade nutricional proteica dos genótipos de milho. No entanto, NUT apresenta a classificação em dados binários (0 e 1) para utilização na ACP, enquanto NUTB apresenta a classificação categórica (ALTA e BAIXA) para uso na FAMD. Se a variável categórica apresenta uma frequência equilibrada de classes, ou seja, há um equilíbrio entre as classes 0 e 1 (ALTA e BAIXA) a ACP e FAMD apresentam resultados similares. Isso indica que a variável categórica pode ser utilizada por ACP ou FAMD. Pode-se observar que os dois componentes principais na ACP e FAMD apresentaram a mesma explicação (68,60%), indicando a similaridade dos resultados das duas análises, considerando uma variável categórica com frequência balanceada (Figuras 5 e 6).

```
#---------------------------------ANALISE DE COMPONENTES PRINCIPAIS
# 1 - Selecionando variáveis
res.pca5 <- dados |>
  dplyr::select(AP, DE, CE, ME, NUT)

# 2 - Padronizando variaveis
res.pca6 <- prcomp(res.pca5, scale. = TRUE)

# 3- Extrair os autovalores
eig.val <- get_eigenvalue(res.pca6)

# Componentes principais
fviz_pca_biplot(res.pca6,
                        # indivíduos
                        geom.ind = "point",
                        fill.ind = factor(dados$BaseGen, levels = c("SIMPLES", "
TRIPLO", "DUPLO", "VARIEDADE")),
                        col.ind = "black",
                        pointshape = 21, pointsize = 4,
                        palette = c("green", "purple", "yellow", "red"),  # Ajus
te a paleta conforme a ordem dos níveis
                        addEllipses = FALSE,
                        repel = TRUE,
                        title = "",
                        # variáveis
                        col.var = "black",       # Cor das siglas das variáveis
                        labelsize = 5,          # Tamanho da fonte das siglas da
s variáveis
                        font.var = 2,           # Negrito nas siglas das variáve
is
                        legend.title = list(fill = "")) +
  theme_bw() +
  theme_classic() +
```

```
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

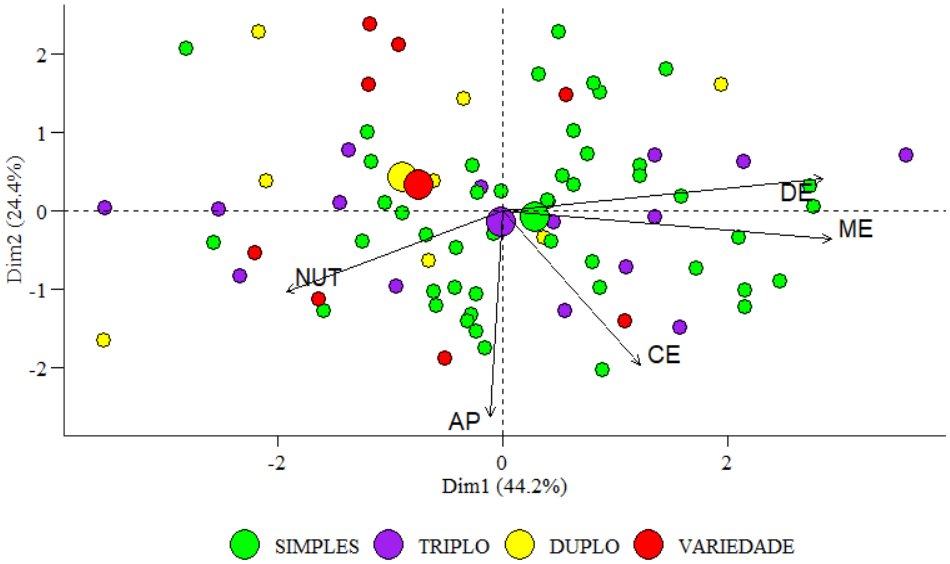

Figura 5. Análise de componentes principais com uma variável categórica transformada em binária com frequência balanceada.

```
#-----------------------ANÁLISE FATORIAL DE DADOS MISTOS

# Carregar pacotes necessários
library(FactoMineR)  # Para FAMD
library(factoextra) # Para visualização

res.AFM <- dados |>
  dplyr::select(AP, DE, CE, ME, NUTB)
# Realizar a FAMD
res.famd <- FAMD(res.AFM, graph = FALSE)

# Visualizar os indivíduos (gráfico FAMD)
fviz_famd_ind(res.famd, geom.ind = "point",
                        col.ind = "black",
              repel = TRUE)+
  theme_bw() +
```

```
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

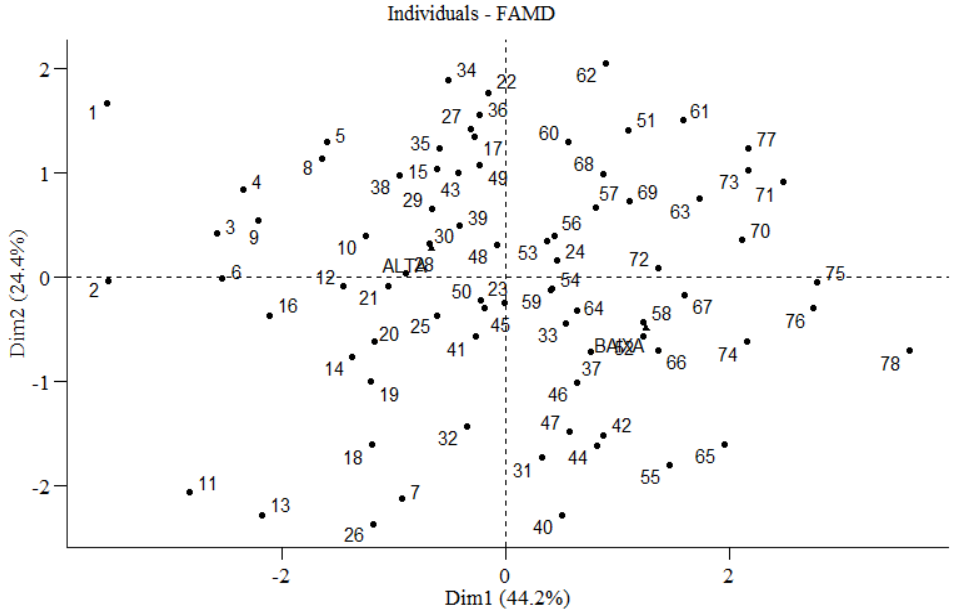

Figura 6. Análise fatorial de dados mistos com uma variável categórica com frequência balanceada.

7.2.2 Cenário com variáveis mistas por ACP e FAMD, com frequência desbalanceada da variável categórica

Para este exemplo, foi considerada uma análise de componentes principais (PCA) e análise fatorial de dados mistos (FAMD) considerando as seguintes variáveis: AP, DE, CE ME, NUT e CAT (para PCA) e AP, DE, CE, ME e NUTB e CATB (para FAMD). Importante destacar que as variáveis NUT e NUTB indicam a qualidade nutricional proteica dos genótipos de milho. As variáveis CAT e CATB indicam a categoria massa da espiga dos genótipos de milho.

No entanto, NUT apresenta a classificação em dados binários (0 e 1) para utilização na ACP, enquanto NUTB apresenta a classificação categórica (ALTA e BAIXA) para uso na FAMD. Essa variável apresenta uma frequência balanceada das classes. A variável CAT apresenta a classificação

em dados discretos (0, 1 e 2) para utilização na ACP, enquanto CATB apresenta a classificação categórica (P = pesada, M = moderada, L = leve) para uso na FAMD. Essa variável apresenta uma frequência desbalanceada das classes, ou seja, a frequência de classes M (moderada) é muito maior que as outras classes (P e L).

Se a variável categórica apresenta uma frequência desbalanceada de classes, ou seja, há um desequilíbrio entre as classes 0, 1 e 2 (P, M e L) a ACP e FAMD apresentam resultados distintos. Pode-se observar que os dois componentes principais na ACP e FAMD explicaram 68,30% e 61,90% da variabilidade total dos dados, respectivamente, indicando uma diferença dos resultados das duas e análises, considerando uma variável categórica com frequência desbalanceada (Figuras 7 e 8). Em situações com frequência desequilibrada das classes, a FAMD consegue ajustar as variáveis categóricas para que não ocorra influência da variância nos componentes. Enquanto a ACP não faz essa correção, e isso inflaciona os componentes. Isso pode ser observado pela superestimação da variância explicada pelos dois componentes da ACP em relação a FAMD. Portanto, embora seja possível utilizar variáveis categóricas na ACP, se existir um número de classes grande e um desequilíbrio entre a frequência das classes a FAMD é uma análise mais adequada.

```
#---------------------------------ANALISE DE COMPONENTES PRINCIPAIS
# 1 - Selecionando variáveis
res.pca3 <- dados |>
  dplyr::select(AP, DE, CE, ME, NUT, CAT)

# 2 - Padronizando variaveis
res.pca4 <- prcomp(res.pca3, scale. = TRUE)

# 3- Extrair os autovalores
eig.val <- get_eigenvalue(res.pca2)

# Componentes principais
fviz_pca_biplot(res.pca4,
                        # indivíduos
                        geom.ind = "point",
                        fill.ind = factor(dados$BaseGen, levels = c("SIMPLES", "
TRIPLO", "DUPLO", "VARIEDADE")),
                        col.ind = "black",
                        pointshape = 21, pointsize = 4,
                        palette = c("green", "purple", "yellow", "red"),  # Ajus
te a paleta conforme a ordem dos níveis
                        addEllipses = FALSE,
                        repel = TRUE,
                        title = "",
                        # variáveis
                        col.var = "black",       # Cor das siglas das variáveis
                        labelsize = 5,          # Tamanho da fonte das siglas da
s variáveis
                        font.var = 2,           # Negrito nas siglas das variáve
is
```

```
                        legend.title = list(fill = "")) +
  theme_bw() +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

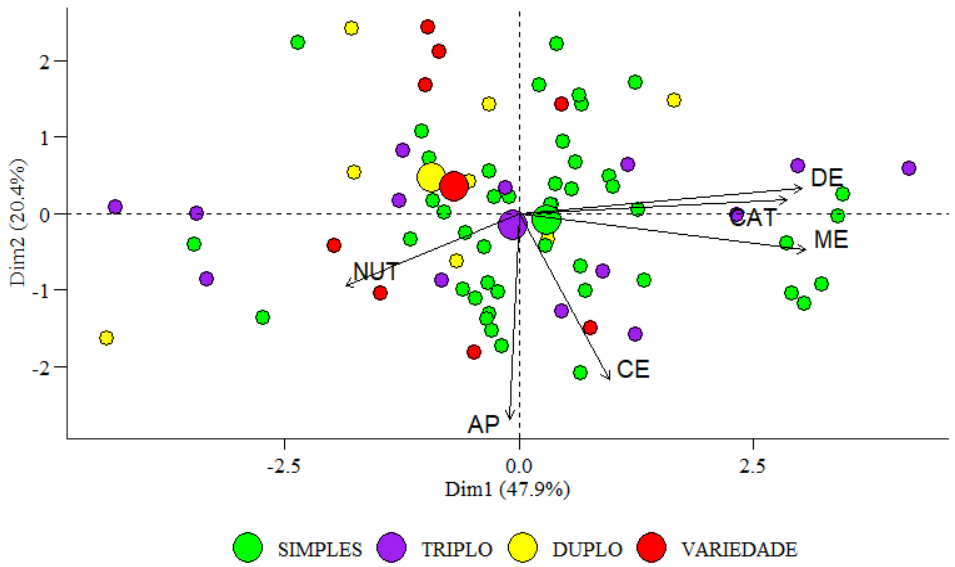

Figura 7. Análise de componentes principais com variáveis categóricas transformada em binária com frequência desbalanceada.

```
#-----------------------ANÁLISE FATORIAL DE DADOS MISTOS

# Carregar pacotes necessários
library(FactoMineR)  # Para FAMD
library(factoextra) # Para visualização

res.AFM <- dados |>
  dplyr::select(AP, DE, CE, ME, NUTB, CATB)
# Realizar a FAMD
res.famd <- FAMD(res.AFM, graph = FALSE)
```

```
# Visualizar os indivíduos (gráfico FAMD)
fviz_famd_ind(res.famd, geom.ind = "point",
                        col.ind = "black",
              repel = TRUE)+
  theme_bw() +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

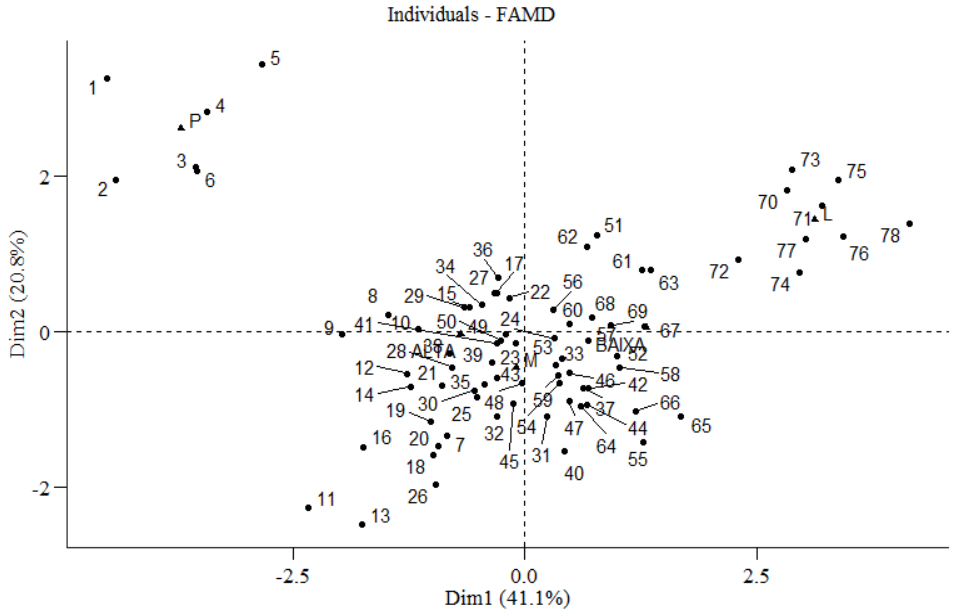

Figura 8. Análise fatorial de dados mistos com variáveis categóricas com frequência desbalanceada.

8. VARIAÇÕES DA ANÁLISE DE COMPONENTES PRINCIPAIS

8.1 Análise de Componentes Principais Robusta

A análise de componentes principais robusta é uma técnica para realizar a decomposição dos dados de forma mais eficiente à presença de *outliers* ou observações que não seguem uma distribuição normal multivariada. Essa abordagem foi apresentada no artigo *ROBPCA: A New Approach to Robust Principal Component Analysis* proposta por Hubert, Rousseeuw, Branden (2005) e busca superar as limitações da análise de componentes principais clássica, especialmente em cenários com valores extremos.

A PCA clássica transforma as variáveis originais em componentes principais, que são combinações lineares dessas variáveis. Esses componentes são ordenados pela variância explicada, sendo que o primeiro componente possui a maior variância possível, seguido dos demais em ordem decrescente. Além disso, os componentes principais são ortogonais entre si. Apesar de amplamente utilizada, a PCA clássica é sensível a *outliers*, uma vez que sua construção depende da matriz de covariância ou de correlação, que pode ser facilmente distorcida por valores extremos.

Por outro lado, a ROBPCA combina estimativas robustas para localizar o centro dos dados e calcular a matriz de covariância com base em métodos como o *Minimum Covariance Determinant (MCD)*. Essa técnica reduz inicialmente os dados para um subespaço de menor dimensão, onde os componentes principais são identificados de forma mais robusta. Além disso, a ROBPCA inclui um diagnóstico de outliers, permitindo separar as observações que desviam do padrão principal.

Ao comparar a PCA clássica com a ROBPCA, algumas diferenças são evidentes. Enquanto a PCA clássica é altamente sensível a outliers e pode gerar componentes distorcidos em dados contaminados, a ROBPCA preserva a estrutura dos dados subjacentes e identifica observações discrepantes. A PCA clássica baseia-se na matriz de covariância convencional, enquanto a ROBPCA utiliza estimativas robustas, como o método MCD. Ademais, a ROBPCA não depende da suposição de normalidade multivariada, sendo mais flexível em diferentes contextos. O estudo de Wang et al. (2022) mostrou que o ROBPCA supera o PCA padrão em termos de precisão e robustez.

A ROBPCA é particularmente útil em análises onde a presença de outliers pode comprometer os resultados ou em dados que não seguem uma distribuição normal. Ao aplicar a ROBPCA, obtêm-se componentes principais que refletem melhor a variabilidade subjacente dos dados, garantindo maior confiabilidade nas conclusões.

8.1.1 ACP Clássica e ACP Robusta com presença de *outliers*

Para avaliar as diferenças entre PCA clássica e PCA robusta foram simulados dados com 10% de valores *outliers*. No diagrama de dispersão é possível avaliar a presença de pontos discrepantes (Figura8).

```
#SIMULAÇÃO DOS DADOS COM OUTLIERS (10%)

set.seed(123)
n_outliers <- round(100 * 0.1)# Número de amostras normais e outliers
n_normais <- 100 - n_outliers

# Gerando dados normais para as cinco variáveis
dados_normais <- data.frame(
  AP = rnorm(n_normais, mean = 150, sd = 20),
  NG = rnorm(n_normais, mean = 300, sd = 50),
  AF = rnorm(n_normais, mean = 50, sd = 10),
  ME = rnorm(n_normais, mean = 500, sd = 80),
  MCG = rnorm(n_normais, mean = 20, sd = 5)
)

# Gerando dados de outliers para as cinco variáveis
dados_outliers <- data.frame(
  AP = rnorm(n_outliers, mean = 300, sd = 50), # Valores extremos
  NG = rnorm(n_outliers, mean = 100, sd = 30),     # Valores extremos
  AF = rnorm(n_outliers, mean = 10, sd = 5),     # Valores extremos
  ME = rnorm(n_outliers, mean = 1000, sd = 100),   # Valores extremos
  MCG = rnorm(n_outliers, mean = 50, sd = 10)    # Valores extremos
)

dadosOUT <- rbind(dados_normais, dados_outliers)# Combinando os dados normais e
os outliers
plot(dadosOUT)
```

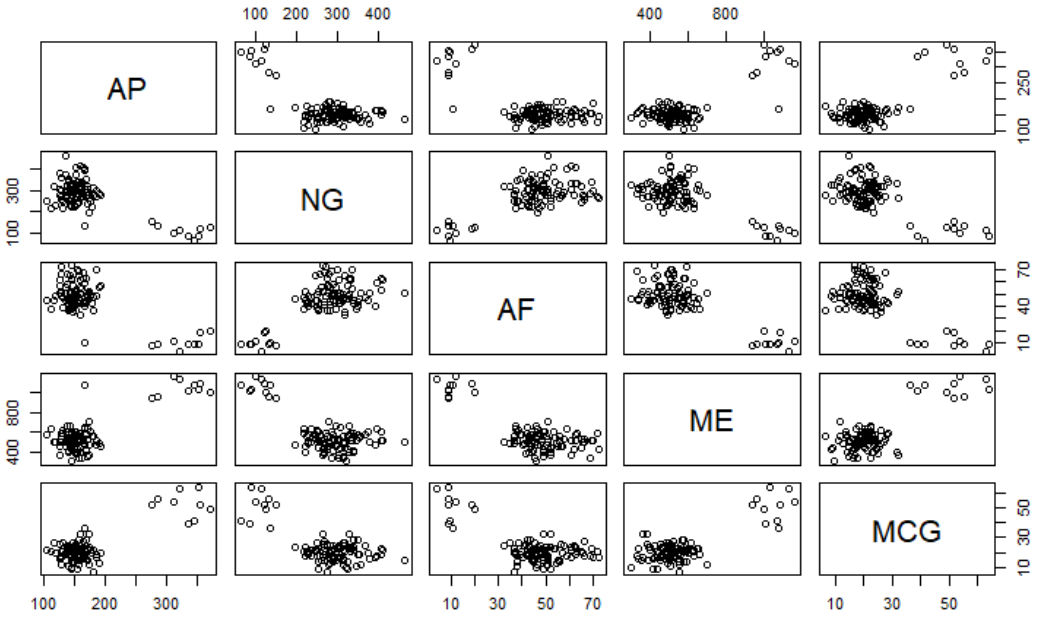

Figura 1. Diagrama de dispersão das variáveis para verificar a presença de outliers.

Os resultados obtidos entre a PCA clássica e a PCA robusta revelam diferenças significativas no tratamento de dados na presença de *outliers*. A PCA clássica demonstrou uma concentração da variância no primeiro componente principal (PC1), que explica 77,16% da variância total, enquanto o segundo componente explica apenas 7,71%. A variância cumulativa atinge 84,87% com os dois primeiros componentes e chega a 100% no quinto componente. Esse padrão reflete a sensibilidade da PCA clássica aos outliers, que distorcem a matriz de covariância ou correlação, levando a uma concentração excessiva de variância em um único componente.

A PCA robusta mostrou uma distribuição mais equilibrada da variância entre os componentes. O PC1 explicou 35,04% da variância total, enquanto o segundo componente contribui com 27,92%. Nos quatro primeiros componentes, a variância cumulativa alcança 92,45%, indicando uma representação mais uniforme e menos influenciada pelos *outliers*. Isso ocorre porque a PCA robusta minimiza os efeitos de dados extremos, o que promove uma análise mais estável e representativa da estrutura subjacente dos dados. Comparativamente, a PCA clássica é sensível a outliers, resultando em uma variância desproporcionalmente concentrada em poucos componentes. Em contraste, a PCA robusta distribui melhor a variância entre os componentes, evidenciando padrões mais confiáveis. Assim, a PCA robusta se destaca como uma abordagem superior em contextos onde há outliers, proporcionando uma análise mais precisa e realista dos dados.

```r
# PCA Clássica
PCA <- prcomp(dadosOUT, center = TRUE, scale. = TRUE)
summary(PCA)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5
## Standard deviation     1.9642 0.62092 0.59081 0.45848 0.44418
## Proportion of Variance 0.7716 0.07711 0.06981 0.04204 0.03946
## Cumulative Proportion  0.7716 0.84869 0.91850 0.96054 1.00000
```

```r
# ROBPCA
library(robustbase)
library(rrcov)
library(rospca)


PCARob <- PcaHubert(scale(dadosOUT), k = 4)# Usando método Hubert para PCA robu
sta
summary(PCARob)
```

```
##
## Call:
## PcaHubert(x = scale(dadosOUT), k = 4)
## Importance of components:
##                           PC1    PC2    PC3    PC4
## Standard deviation     0.7291 0.6509 0.4846 0.4610
## Proportion of Variance 0.3504 0.2792 0.1548 0.1401
## Cumulative Proportion  0.3504 0.6297 0.7844 0.9245
```

Por meio dos biplot é possível realizar uma comparação visual entre os resultados obtidos pela PCA clássica e pela PCA robusta em um cenário com 10% de outliers (Figuras 9 e 10). Na PCA clássica, observou-se que os outliers têm uma grande influência na definição do espaço dos componentes principais, o que resulta em uma dispersão maior dos pontos e na presença de observações afastadas do núcleo principal dos dados. Isso indica que os outliers distorcem a orientação e a escala dos eixos principais. A PCA Robusta demonstra uma maior resistência aos efeitos dos outliers. Os pontos principais estão mais concentrados em torno de um centro, e os componentes principais refletem a variabilidade inerente ao conjunto de dados, minimizando a influência das observações discrepantes. Essa característica permite que a PCA robusta capture de maneira mais confiável a estrutura subjacente dos dados, mesmo na presença de valores extremos.

```r
# Comparando resultados
plot(PCA$x, main = "PCA Clássica")
```

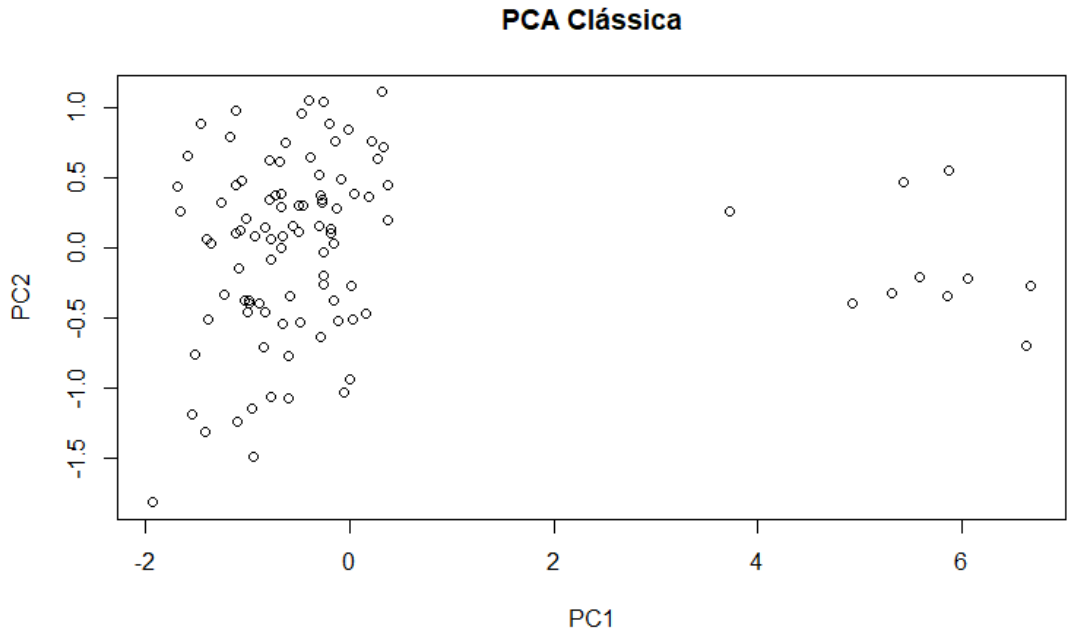

Figura 2. Análise de componentes principais clássica com a presença de outliers.

```
plot(PCARob$scores, main = "ROBPCA")
```

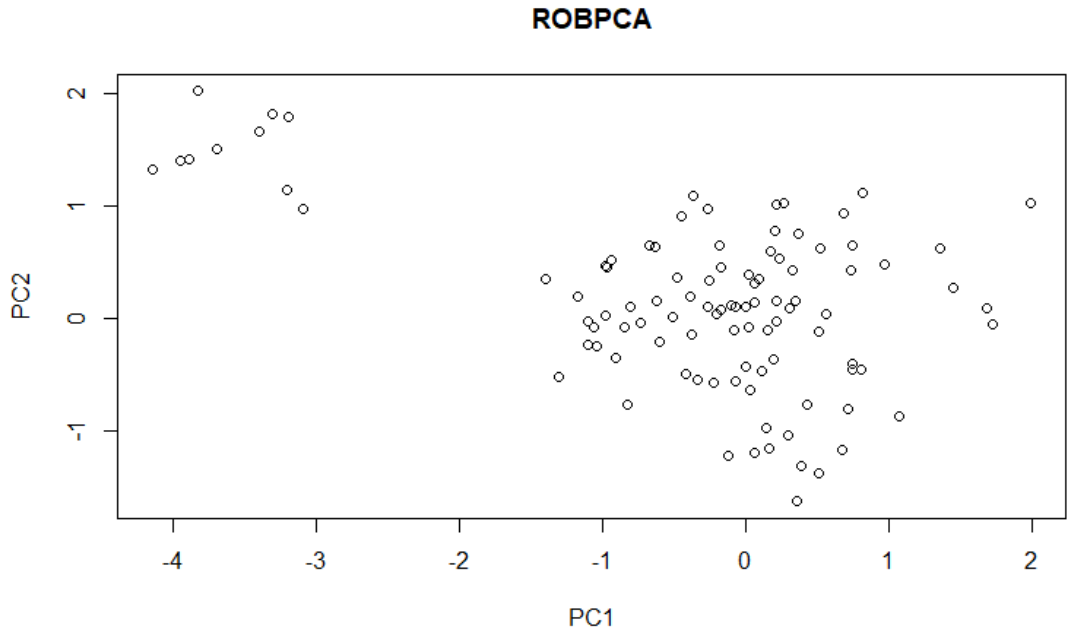

Figura 3. Biplot para identificação de outliers pela análise de componentes principais robusta.

A ACP clássica apresenta a capacidade em identificar os pontos outliers. Logo, pode-se observar no biplot que oito observações foram consideradas como outliers na análise (Figura 11).

```
diagPlot(PCARob, col = "green", id = 6)
```

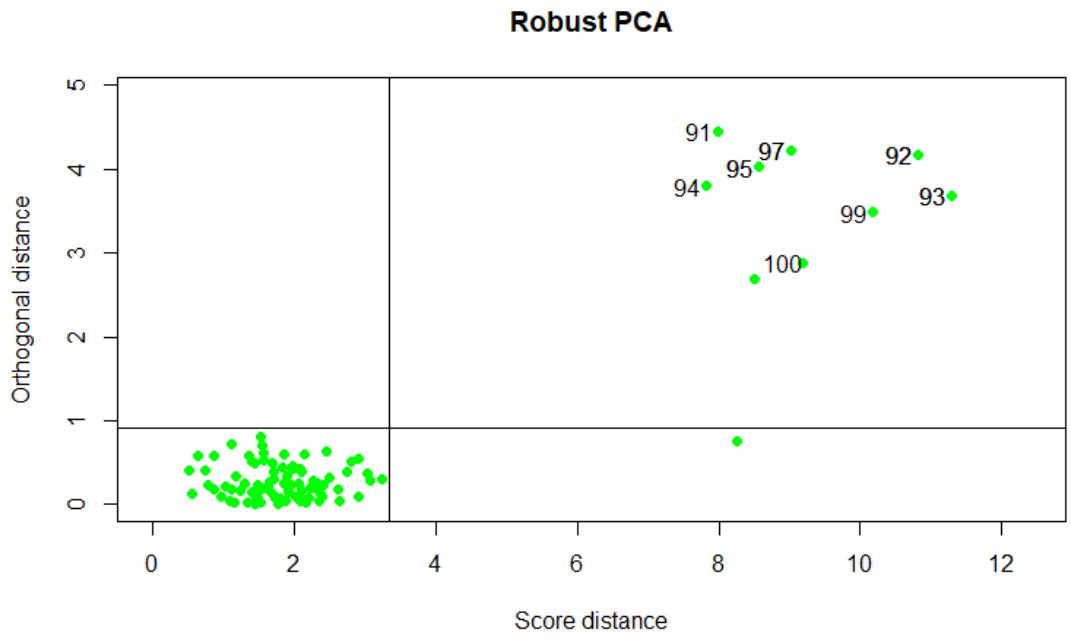

Figura 4. Biplot para identificação de outliers pela análise de componentes principais robusta

8.1.2 ACP Clássica e ACP Robusta sem presença de *outliers*

Para avaliar as diferenças entre PCA clássica e PCA robusta foram simulados dados sem a presença de *outliers*. No diagrama de dispersão é possível avaliar a ausência considerável de pontos discrepantes (Figura 12).

```
#SIMULAÇÃO DOS DADOS SEM OUTLIERS

set.seed(123)
# Gerando dados normais para as cinco variáveis
dadosNOR <- data.frame(
  AP = rnorm(100, mean = 150, sd = 10),
  NG = rnorm(100, mean = 300, sd = 50),
  AF = rnorm(100, mean = 50, sd = 5),
  ME = rnorm(100, mean = 500, sd = 10),
  MCG = rnorm(100, mean = 20, sd = 5)
)

plot(dadosNOR)
```

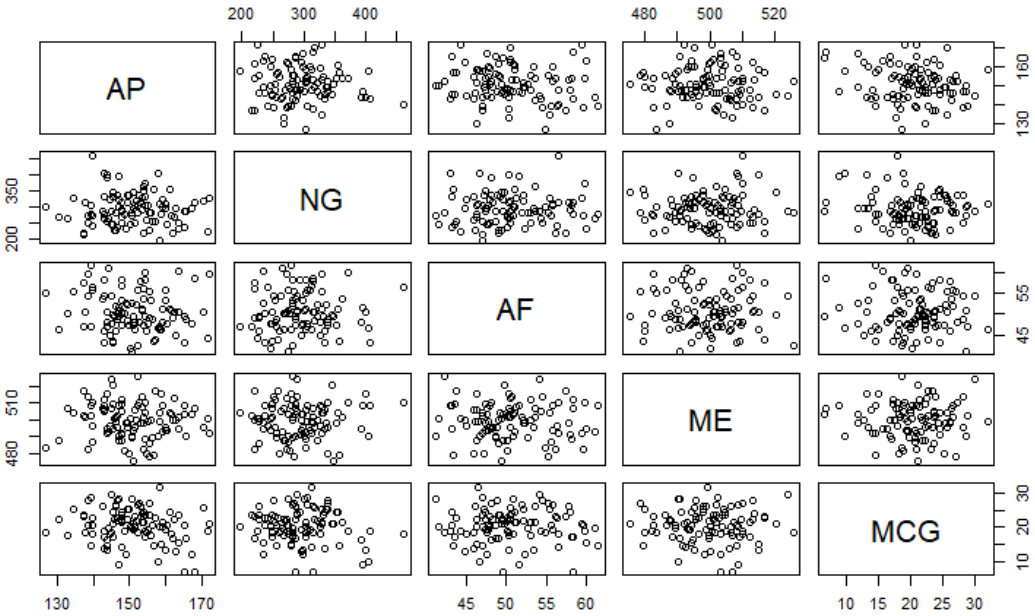

Figura 5. Diagrama de dispersão das variáveis para verificar a presença de outliers.

Os resultados obtidos entre a PCA clássica e a PCA robusta revelam similaridade dos resultados na ausência de *outliers*. A PCA clássica demonstrou uma concentração da variância no primeiro componente principal (PC1), que explica 24,54% da variância total, enquanto o segundo componente explicou 22,51%. A variância cumulativa atinge 47,05% com os dois primeiros componentes e chega a 85,75% no quarto componente. Isso reflete uma distribuição equilibrada da variância entre os componentes.

A PCA robusta também apresentou uma distribuição equilibrada da variância entre os componentes. O PC1 explicou 26,16% da variância total, enquanto o segundo componente contribuiu com 22,64%. Nos quatro primeiros componentes, a variância cumulativa foi de 85,14%, sendo muito similar a PCA clássica. Comparativamente, a PCA clássica e PCA robusta apresentam resultados similares quando não há presença expressiva de *outliers*.

```
# PCA Clássica
PCAnor <- prcomp(dadosNOR, center = TRUE, scale. = TRUE)
summary(PCAnor)

## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5
## Standard deviation     1.1077 1.0610 1.0191 0.9468 0.8440
## Proportion of Variance 0.2454 0.2251 0.2077 0.1793 0.1425
## Cumulative Proportion  0.2454 0.4705 0.6783 0.8575 1.0000

# ROBPCA
library(robustbase)
```

```
library(rrcov)
library(rospca)
PCARobNor <- PcaHubert(scale(dadosNOR), k = 4)# Usando método Hubert para PCA r
obusta
summary(PCARobNor)

##
## Call:
## PcaHubert(x = scale(dadosNOR), k = 4)
## Importance of components:
##                            PC1    PC2    PC3    PC4
## Standard deviation     1.1325 1.0536 0.9760 0.9106
## Proportion of Variance 0.2616 0.2264 0.1943 0.1691
## Cumulative Proportion  0.2616 0.4880 0.6823 0.8514
```

A similaridade entre as análises nesse cenário reflete-se nos gráficos biplot (Figuras 13 e 14). Na PCA clássica e PCA robusta os pontos estão dispersos em torno de um centro, ou seja, não há pontos discrepantes causando viés nas análises. Logo, os componentes principais refletem a variabilidade inerente ao conjunto de dados.

```
# Comparando resultados
plot(PCAnor$x, main = "PCA Clássica")
```

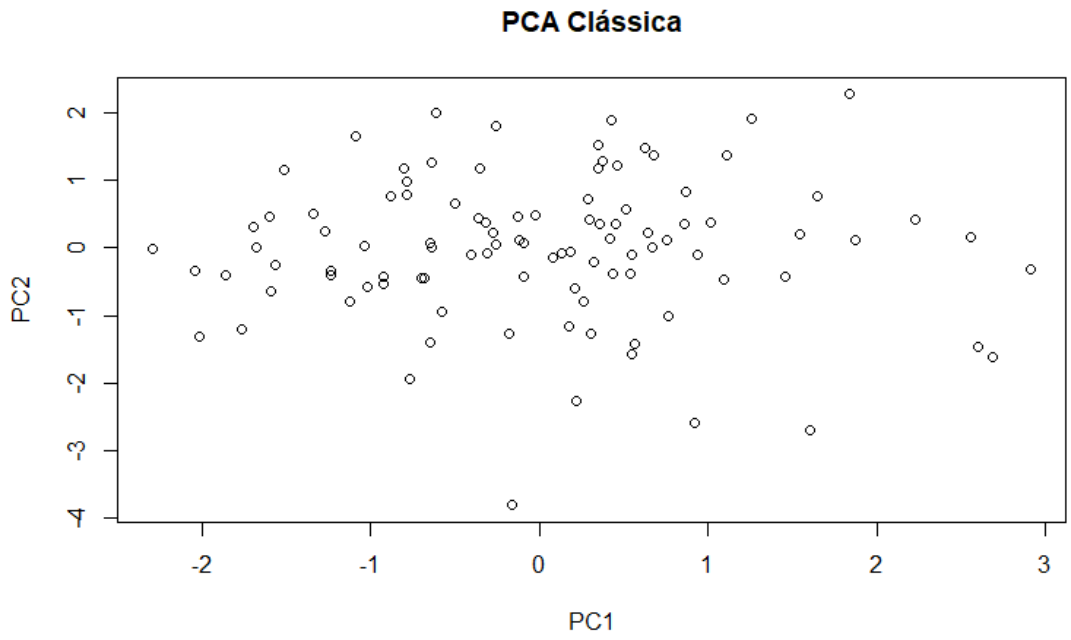

Figura 6. Análise de componentes principais clássica sem a presença de outliers.

```
plot(PCARobNor$scores, main = "ROBPCA")
```

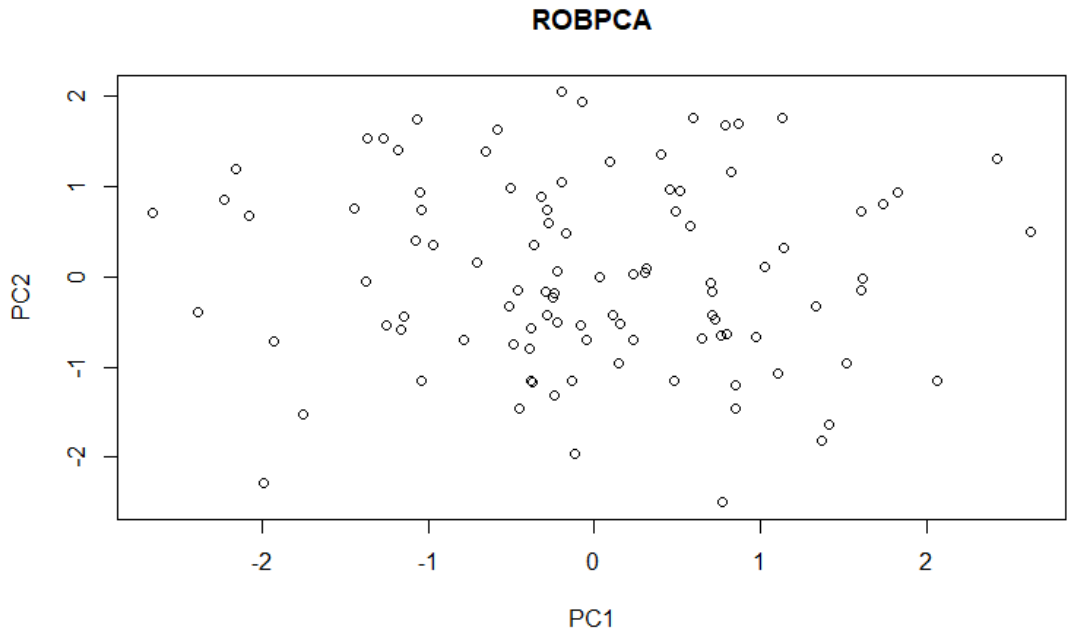

Figura 7. Análise de componentes principais robusta sem a presença de outliers.

8.2 Análise de Componentes Principais Não Linear (Kernel PCA)

A ACP não linear ou Kernel PCA, utiliza transformações de kernel para capturar relações não lineares entre variáveis. Em dados de milho, isso pode ser útil para modelar interações complexas, como a relação entre radiação solar, produtividade e teor de aminoácidos. Logo, o Kernel PCA introduz uma abordagem não linear ao problema, permitindo que os dados sejam transformados em um espaço de características de maior dimensão, onde as relações não lineares entre as variáveis podem ser capturadas de forma mais eficaz. Portanto, essa técnica que pode ser usada em dados que não se distribuem de maneira linear, proporcionando uma representação informativa dos dados originais (SCHOELKOPF; SMOLA; MUELLER, 1998). Pode-se fazer uma analogia à metodologia Máquinas de Vetores de Suporte que apresentam a possibilidade de identificar relações a partir de um kernel não linear como polinomial e radial.

O Kernel PCA, em vez de calcular os componentes principais diretamente nos dados originais, aplica uma transformação não-linear aos dados antes de realizar a análise de componentes principais. Isso possibilita a análise de conjuntos de dados mais complexos, que não poderiam ser separados de maneira linear. Para realizar essa transformação, diferentes tipos de kernels podem ser utilizados, cada um com características e aplicações específicas. Entre os kernels mais comuns estão o kernel linear, o kernel polinomial, o kernel radial e o kernel sigmoid. Cada tipo de kernel determina de maneira distinta a modelagem dos dados e o cálculo dos componentes principais, permitindo flexibilidade na adaptação do método a diferentes tipos de estruturas nos dados.

A seguir foram simulados dados com relações não lineares entre as variáveis para comparar os resultados entre uma PCA clássica e uma Kernel PCA. Na matriz de correlação é possível observar ausência de relação linear entre as variáveis. O gráfico de dispersão pode-se observar que a relação entre as variáveis, principalmente entre VAL e AP é não linear. A partir disso, verificou-se como a PCA clássica e a Kernel PCA avaliam essas relações.

```
library(kernlab)

# Configurações para simulação
set.seed(123)
AP <- rnorm(500, mean = 180, sd = 5)  # Altura média
LYS <- rnorm(500, mean = 4.5, sd = 0.5)
VAL <- 10 - 0.09 * (AP - 180)^2 + rnorm(500, sd = 0.05)  # Curva quadrática em
torno de AP=180

data <- data.frame(LYS, VAL, AP)
print(cor(data))

##              LYS          VAL           AP
## LYS  1.00000000  0.03674807 -0.05193691
## VAL  0.03674807  1.00000000 -0.11103788
## AP  -0.05193691 -0.11103788  1.00000000

plot(data)
```

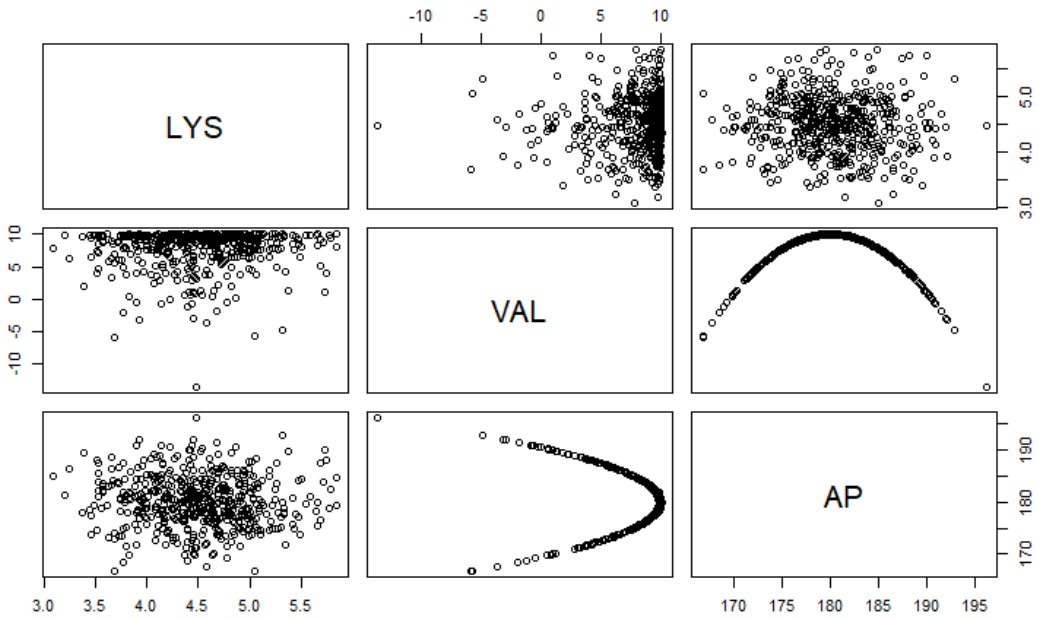

Figura 8. Diagrama de dispersão das variáveis com relações não lineares.

Os gráficos biplot refletiram as diferenças entre a análise clássica de componentes principais e o Kernel PCA com kernel radial (RBF) aplicado a dados com relações não lineares (Figuras 9 e 10).

No biplot do PCA clássico, observou-se que os dados foram projetados em um espaço onde os componentes principais capturam principalmente relações lineares (Figura 9). No entanto, as relações complexas entre as variáveis não foram adequadamente representadas, resultando em uma distribuição que se concentrou em um padrão linearizado. Essa limitação era esperada, uma vez que o PCA clássico pressupõe linearidade nas combinações das variáveis.

As relações não lineares entre as variáveis foram melhor captadas pelo Kernel PCA, utilizando um kernel radial (Figura 10). O kernel radial transforma o espaço original das variáveis em um espaço de maior dimensionalidade, o que permite que padrões mais complexos sejam representados. Isso resultou em uma projeção que demonstrou melhor classificação e distribuição dos dados em componentes principais neste caso com relações não lineares. Logo, o Kernel PCA demonstrou um desempenho superior em relação ao PCA clássico para os dados simulados, como era esperado, dada a introdução de relações não lineares entre as variáveis.

```
#---------------------PCA clássica
pca_model <- prcomp(data, center = TRUE, scale. = TRUE)

pca_var<- (pca_model$sdev)^2
pca_expl <- 100 * pca_var / sum(pca_var)
print(pca_expl) #Variância explicada, método manual

## [1] 37.97839 32.43017 29.59145

pca_comp <- as.data.frame(pca_model$x)

# Bliplot PCA clássico
ggplot(pca_comp, aes(x = PC1, y = PC2)) +
  geom_point(alpha = 0.6) +
  labs(title = "PCA Clássico", x = "PCA 1", y = "PCA 2") +
  theme_bw() +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

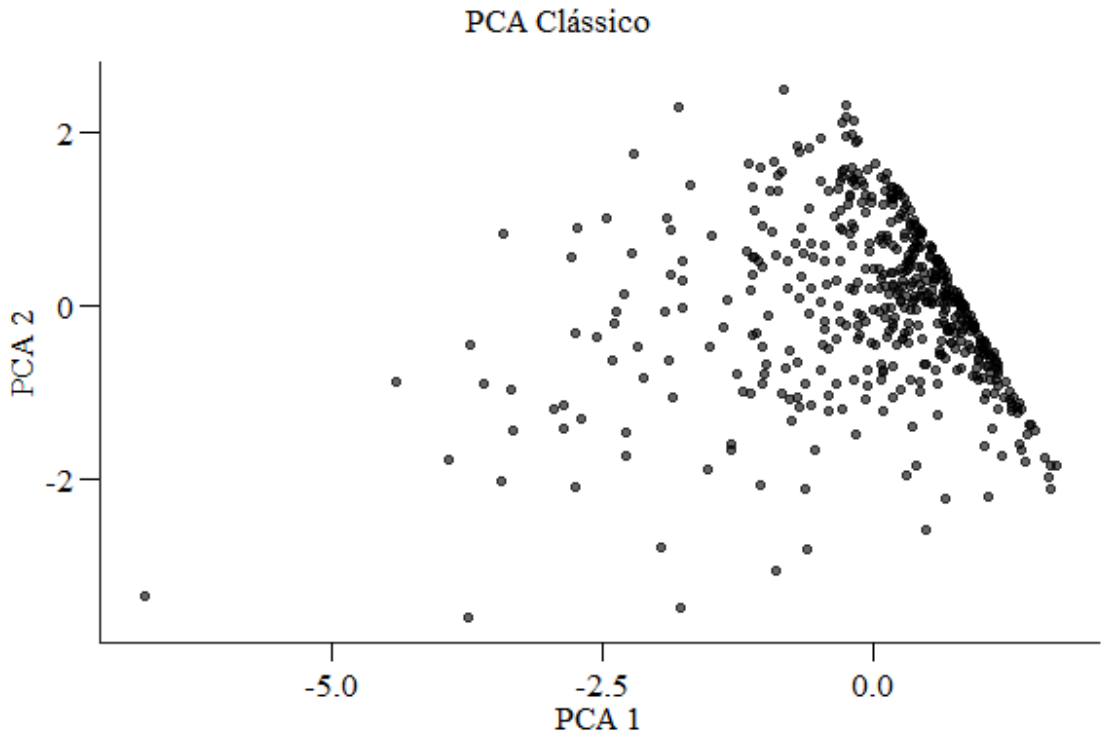

Figura 9. Biplot da análise de componentes principais clássica para dados não lineares

```
# Kernel PCA com Kernel RBF
data1 <- data.frame(scale(data)) #padronização
kpca_model <- kpca(~., data = data1,
                   kernel = "rbfdot",
                   kpar = list(sigma = 0.5),
                   features = 3) #numero PCA

# Componentes principais do Kernel PCA
kpca_comp <- as.data.frame(predict(kpca_model, data1))
# Calcular variância explicada no Kernel PCA
kpca_eig <- kpca_model@eig  # Autovalores no espaço kernel
kpca_exp <- 100 * kpca_eig / sum(kpca_eig)
print(kpca_exp)

##   Comp.1   Comp.2   Comp.3
## 36.93981 34.63121 28.42898

# Biplot Kernel PCA
ggplot(kpca_comp, aes(x = V1, y = V2)) +
  geom_point(alpha = 0.6, color = "blue") +
  labs(title = "Kernel PCA (RBF)", x = "PCA 1", y = "PCA 2") +
  theme_bw() +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
```

```
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

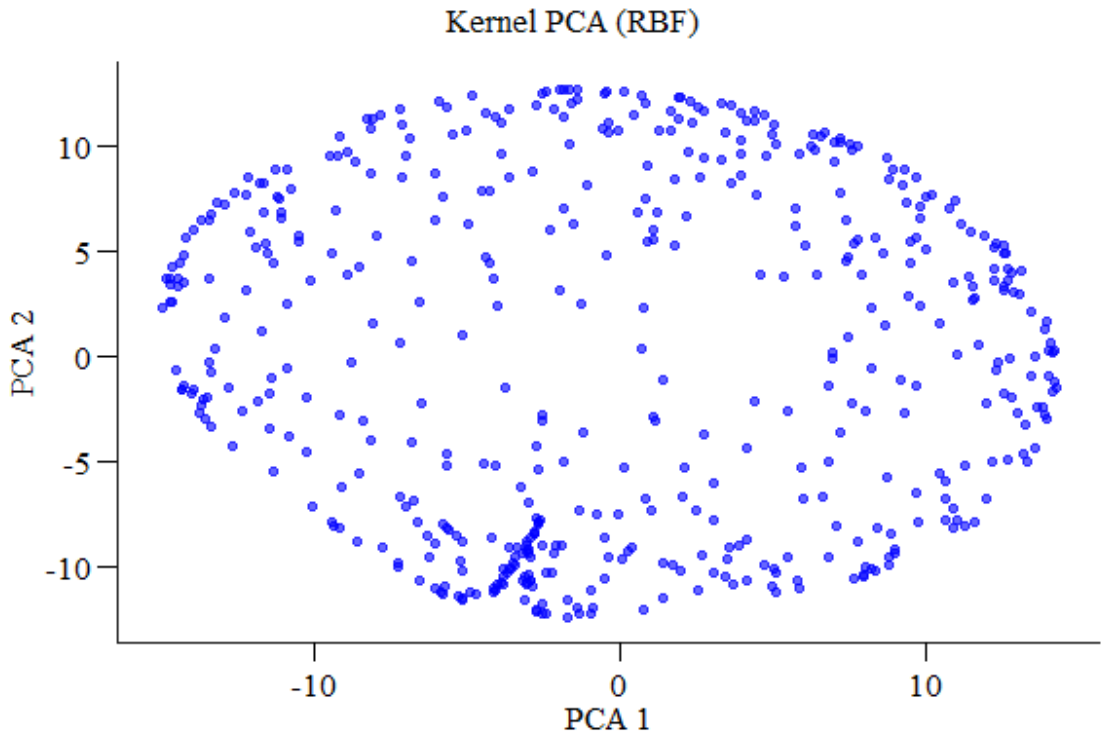

Figura 10. Biplot da análise de componentes principais não linear (PCA Kernel) para dados não lineares.

8.3 Análise de Componentes Principais Generalizada

A PCA Generalizada é uma extensão da PCA Clássica, com melhor desempenho em dados que não atendem às suposições de normalidade e linearidade. Baseia-se na teoria de modelos lineares generalizados, permitindo ajustar os componentes principais de acordo com a distribuição específica dos dados. Diferentemente da PCA Clássica, que maximiza a variância dos dados em um espaço linear, a PCA Generalizada utiliza funções apropriadas, para transformar os dados em um espaço onde as relações estruturais sejam preservadas (VIDAL; MA; SASTRY, 2005). Essa abordagem torna a PCA Generalizada adequada para modelar dados com características não lineares, assimétricas ou heterocedásticas, como aqueles provenientes de distribuições Poisson ou binomial, por exemplo.

A aplicação da PCA Generalizada é eficiente em situações em que os dados representam contagens, proporções ou outras variáveis com restrições intrínsecas que inviabilizam o uso de métodos clássicos. Por exemplo, em dados de contagem com distribuição Poisson, a PCA

Generalizada permite capturar as variações estruturais considerar a relação intrínseca entre média e variância, comum nesse tipo de distribuição. Da mesma forma, pode ser utilizada em dados ambientais, genômicos ou em qualquer contexto em que as distribuições dos dados sejam não normais.

A análise comparativa entre a PCA Clássica e a PCA Generalizada em dados com distribuição Poisson revelou diferenças na capacidade de ambas as técnicas em representar a estrutura dos dados. A PCA Clássica, fundamentada na decomposição em valores singulares (SVD) da matriz de correlação, assume que os dados seguem uma distribuição aproximadamente normal e possuem uma estrutura linear. Essa suposição, no entanto, é inadequada para dados de contagem, como os que seguem a distribuição Poisson, caracterizados por assimetria.

Os resultados obtidos demonstraram que a PCA Clássica apresentou limitações na representação dos dados (Figura 11). Os escores das observações foram deslocados para a direita, um efeito frequentemente observado em distribuições assimétricas. Além disso, a separação entre grupos ou padrões nos dados foi prejudicada, evidenciando uma baixa capacidade da PCA Clássica de capturar a estrutura subjacente em distribuições não normais. Esse comportamento decorre do fato de que, em dados Poisson, as altas variâncias associadas aos valores maiores dominam os componentes principais, comprimindo os valores menores e reduzindo a discriminação entre os dados.

Na PCA Generalizada, que expande o método clássico ao incorporar a teoria de modelos lineares generalizados, apresentou resultados superiores. Esse método ajusta os componentes principais com base em distribuições específicas, como Poisson, para modelar os dados. Com isso, a PCA Generalizada conseguiu melhorar representação dos escores, distribuindo-os de forma mais uniforme ao longo dos eixos principais e capturando melhor os padrões, refletindo uma maior capacidade de representar a estrutura real dos dados (Figura 12).

Os gráficos comparativos reforçaram essa diferença. Enquanto o gráfico da PCA Clássica apresentou dados aglomerados em uma região específica e com baixa discriminação, o gráfico da PCA Generalizada mostrou maior dispersão, com os dados bem distribuídos e os padrões mais evidentes. Essa comparação evidencia a limitação da PCA Clássica em contextos de dados não normais, como aqueles com distribuição Poisson, enquanto a PCA Generalizada se mostrou uma alternativa flexível e adequada para esse tipo de dado.

```
#Simulação de dados - Distribuicao de Poisson
mu<-rep(c(.5,3),each=10)
mu<-matrix(exp(rnorm(100*20)),nrow=100)
mu[,1:10]<-mu[,1:10]*exp(rnorm(100))

Y<-matrix(rpois(prod(dim(mu)),mu),nrow=nrow(mu))
```

```r
#PCA Classica
pca_model <- prcomp(Y, center = TRUE, scale. = TRUE)

pca_var<- (pca_model$sdev)^2
pca_expl <- 100 * pca_var / sum(pca_var)
print(pca_expl) #Variância explicada, método manual
```

```
##  [1] 18.7836352  9.1435487  7.8813271  6.8645347  6.3802153  5.9921960
##  [7]  5.4436772  5.2794800  4.7564470  4.6972777  4.2891154  4.1737211
## [13]  3.7717842  3.1367427  2.5834787  1.8862602  1.6468658  1.4675915
## [19]  1.0303948  0.7917066
```

```r
pca_comp <- as.data.frame(pca_model$x)

# Bliplot PCA clássico
ggplot(pca_comp, aes(x = PC1, y = PC2)) +
  geom_point(alpha = 0.6, size = 3) +
  labs(title = "PCA Clássico", x = "PCA 1", y = "PCA 2") +
  theme_bw() +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

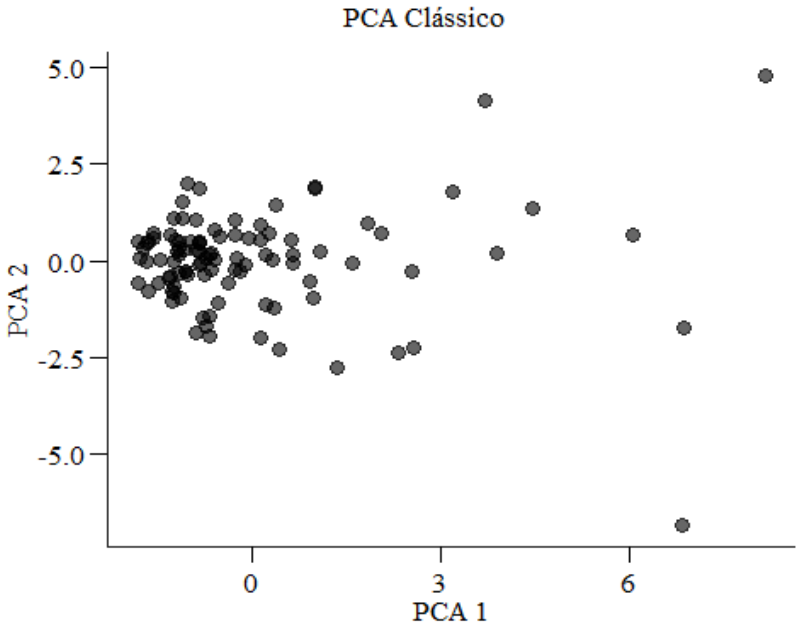

Figura 11. Biplot da análise de componentes principais clássica para dados com distribuição Poisson.

```r
#PCA Generalizada
library(glmpca)
res<-glmpca(Y, 2, fam = "poi")
fat <- res$loadings

# Supondo que você tenha um data.frame com `factors` e `clust`
factors_df <- data.frame(PC1 = fat[,1], PC2 = fat[,2])

ggplot(factors_df, aes(x = PC1, y = PC2)) +
  geom_point(alpha = 0.6, size = 3) +
  labs(title = "PCA Generalizada", x = "Fator 1", y = "Fator 2") +
  theme_bw() +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family = "serif", size = 13),
        axis.text.y = element_text(size = 13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color = "
black"),
        axis.ticks = element_line(colour = "black"),
        axis.ticks.length = unit(0.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust = 2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

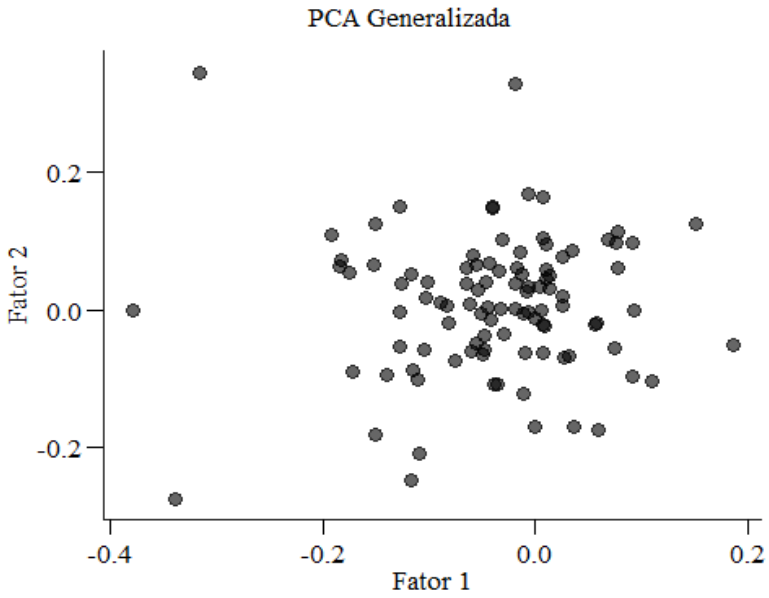

Figura 12. Biplot da análise de componentes principais generalizada para dados com distribuição Poisson.

8.4 Referências

HUBERT, M.; ROUSSEEUW, P. J.; BRANDEN, K. V. ROBPCA: a new approach to robust principal component analysis. **Technometrics**, v. 47, n. 1, p. 64-79, 2005.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences**, v. 374, n. 2065, p. 20150202, 2016.

KASSAMBARA, A. **Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra**. Sthda, 2017.

PTAK, S. H.; ERRICO, M.; CHRISTENSEN, K. V. Extracting activity patterns: Exploratory data analysis on a fucoidan extract data set with mixed variables. **Algal Research**, v. 54, p. 102220, 2021.

SMITH, L. I. **A tutorial on principal components analysis**. 2002.

VIDAL, R.; MA, Y.; SASTRY, S. Generalized principal component analysis (GPCA). **IEEE transactions on pattern analysis and machine intelligence**, v. 27, n. 12, p. 1945-1959, 2005.

WANG, S. *et al*. Robust principal component analysis via joint reconstruction and projection. **IEEE Transactions on Neural Networks and Learning Systems**, 2022.

9. RELAÇÕES NÃO LINEARES E PRESENÇA DE OUTLIERS

A correlação de Pearson é uma medida linear da relação entre duas variáveis, o que significa que ela só captura relações que podem ser representadas por uma linha reta. Em cenários onde as variáveis apresentam uma relação não linear ou são influenciadas por outliers, a correlação de Pearson pode fornecer uma visão incompleta ou até mesmo distorcida da verdadeira relação entre os dados. Abaixo, descrevo dois cenários em que a correlação de Pearson não seria o método ideal para analisar as relações entre caracteres agronômicos, nutricionais e variáveis meteorológicas.

9.1 Relações não lineares

A correlação de Pearson assume que a relação entre as variáveis é linear. Isso significa que, se a relação entre os caracteres agronômicos e as variáveis meteorológicas apresentar um padrão não linear, como uma relação quadrática, a correlação de Pearson não será capaz de identificar essa complexidade. Por exemplo, a relação entre a massa de grãos da espiga de milho e a temperatura do ar pode ser melhor representada por uma curva, onde existe um ponto ótimo de temperatura, além do qual a produtividade diminui (Figura 1). Nesse caso, uma correlação de Pearson pode sugerir uma relação fraca ou inexistente, mesmo que a verdadeira relação seja forte, mas não linear.

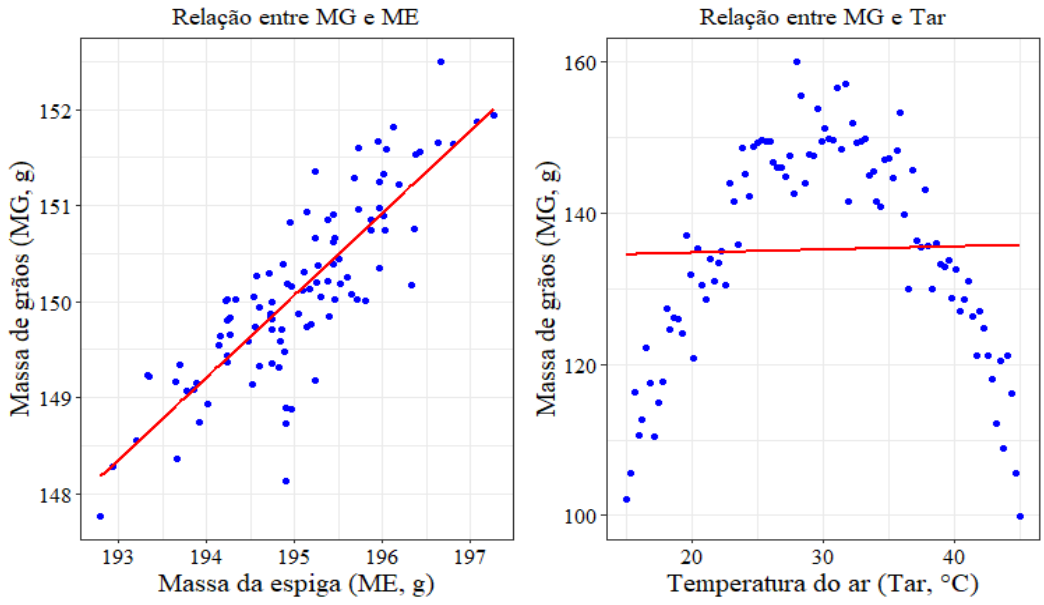

Figura 1. Exemplo de relação linear entre MG e ME e não linear entre MG e Tar.

9.2 Presença de *Outliers*

Outliers são valores extremos que podem distorcer a correlação de Pearson, pois essa medida é sensível a extremos. Se houver dados atípicos, como valores meteorológicos altos ou baixos (por exemplo, uma onda de calor extrema) ou valores nutricionais altos (como concentrações muito altas de um aminoácido nos grãos), a correlação de Pearson pode superestimar ou subestimar a força da

relação entre as variáveis. Em situações como essa, a correlação de Pearson pode não refletir com precisão a relação subjacente entre as variáveis.

Imagine uma análise da relação entre o teor de amido e duração do estádio reprodutivo. Em condições normais, espera-se que uma maior duração do estádio reprodutivo esteja associada a um maior teor de amido, pois esse período oferece mais tempo para o transporte e a conversão de fotoassimilados em carboidratos armazenados nos grãos. Para compreender a influência de uma outlier, será considerado o exemplo a seguir. Em um experimento, progênies de milho foram semeadas em uma área experimental homogênea, e monitorou-se desde a emergência até a colheita. A duração do estádio reprodutivo foi registrada para cada planta, e, ao final do ciclo, avaliou-se o teor de amido acumulado nos grãos. Contudo, algumas plantas apresentaram uma emergência cinco dias mais tarde em comparação à maioria. Esse atraso provocou diferenças significativas no ambiente de desenvolvimento dessas plantas. Devido à emergência tardia, estas plantas ficaram sob o sombreamento das plantas mais desenvolvidas, o que reduziu a incidência de radiação solar. Como consequência, a taxa de fotossíntese dessas plantas foi limitada, comprometendo o transporte e o acúmulo de carboidratos nos grãos.

Na análise dos dados, foi possível observar que essas plantas, embora apresentassem uma duração de estádio reprodutivo semelhante à das demais, exibiram valores de teor de amido significativamente menores, configurando-se como outliers na relação estudada (Figura 2). A correlação de Pearson é sensível a outliers e, nesse caso, o valor de correlação é distorcido pela presença de pontos extremos que puxam para baixo.

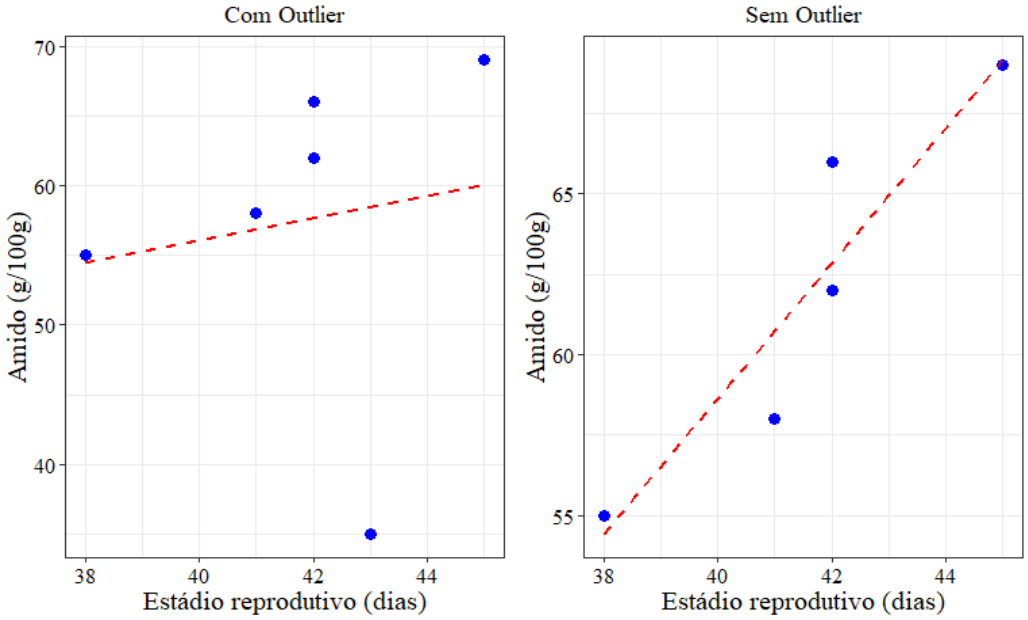

Figura 2. Exemplo de relação linear entre amido e duração do estádio reprodutivo com ausência e presença de outlier.

10. AVALIAÇÃO DE PADRÕES NÃO LINEARES E MONOTÔNICOS

10.1 Correlação de Spearman

A correlação de Spearman é uma medida baseada na transformação dos dados em rankings. Em vez de usar os valores absolutos das variáveis, é calcula a correlação linear entre os *rankings* dessas variáveis. Logo, é possível capturar relações monotônicas, sejam crescentes ou decrescentes, independentemente da linearidade. Por exemplo, uma relação exponencial ou logarítmica será adequadamente descrita pela correlação de Spearman, desde que o padrão seja monotônico. No entanto, sua eficácia é limitada em capturar relações não monotônicas, como padrões em forma de "U" ou "S".

Primeiro, é necessário classificar os valores das duas variáveis, X e Y, de forma crescente. Para cada valor, será atribuído um número de classificação (*ranking*). Em seguida, calcula-se a diferença entre as classificações de cada par de observações nas duas variáveis. Para cada par, subtrai-se o ranking de uma variável do ranking da outra. Depois, essas diferenças são elevadas ao quadrado. O próximo passo é somar todos esses quadrados das diferenças e, a partir dessa soma, calcular a correlação de Spearman.

10.2 Correlação de Kendall

A correlação de Kendall, assim como a de Spearman, avalia relações monotônicas, mas sua abordagem é diferente. Baseia-se na contagem de pares concordantes e discordantes entre as variáveis, refletindo a probabilidade de que o aumento em uma variável esteja associado ao aumento (ou diminuição) na outra. A correlação de Kendall é mais robusta para dados com muitos empates e pode ser interpretada de maneira mais intuitiva em contextos de probabilidade. Entretanto, assim como a correlação de Spearman, é mais apropriada para padrões monotônicos e menos eficiente em capturar relações não monotônicas complexas. Além disso, o coeficiente de Kendall é menos sensível a outliers, o que o torna uma escolha preferida em muitas análises estatísticas.

Portanto, o coeficiente de Kendall é uma medida estatística utilizada para avaliar a associação entre duas variáveis, levando em consideração a ordem relativa dos valores nos pares de observações. Essa medida é particularmente útil para identificar relações monotônicas, ou seja, quando uma variável tende a aumentar ou diminuir em conjunto com a outra, sem necessariamente seguir uma relação linear. Para calcular o coeficiente de Kendall, é necessário analisar todos os pares possíveis de observações presentes no conjunto de dados. Para cada par de observações, verifica-se se é concordante ou discordante. Um par é considerado concordante quando ambas as variáveis mantêm a mesma ordem: se uma variável aumenta, a outra também aumenta, ou se uma diminui, a outra

também diminui. Por outro lado, um par é considerado discordante quando a ordem das variáveis é oposta: quando uma variável aumenta, a outra diminui, ou vice-versa.

O coeficiente de Kendall é calculado a partir da diferença entre o número de pares concordantes e discordantes, dividida pelo número total de pares possíveis. Quando há muitos pares concordantes em relação aos discordantes, o coeficiente de Kendall será positivo, indicando uma relação positiva entre as variáveis. Se houver mais pares discordantes, o coeficiente será negativo, indicando uma relação inversa. Caso os pares concordantes e discordantes se equilibrem, o coeficiente se aproxima de zero, o que sugere que não há uma relação monotônica entre as variáveis.

10.3 Coeficiente de Máxima Informação (MIC)

O Coeficiente de Máxima Informação (MIC) é uma abordagem mais flexível, projetada para identificar tanto relações lineares quanto não lineares de qualquer tipo. Baseia-se no conceito de maximização da informação mútua entre duas variáveis, procurando padrões em que o conhecimento de uma variável reduz a incerteza sobre a outra. O MIC é capaz de identificar padrões não monotônicos, como relações em forma de "U" ou "S", tornando-se uma ferramenta eficaz para explorar dados em busca de relações complexas. O coeficiente MIC varia de 0 a 1 e não informa a direção nem o tipo de relacionamento.

O MIC é baseado na ideia de maximizar a informação mútua entre as variáveis, ou seja, ele busca a maior dependência possível entre as variáveis observadas, independentemente da forma dessa dependência. O processo de cálculo do MIC envolve dividir o espaço de dados em uma grade, de modo a identificar a melhor maneira de segmentar os dados para revelar a relação mais forte. O valor do MIC varia de 0 a 1, onde 0 indica que não há relação entre as variáveis, e 1 indica uma relação perfeita de máxima dependência. Diferentemente de métodos como a correlação de Kendall ou de Spearman, que se concentram em detectar relações monotônicas, o MIC é capaz de identificar relações mais complexas, incluindo aquelas que não são monotônicas.

11. CORRELAÇÃO CANÔNICA, SPEARMAN, KENDALL E MIC

A correlação canônica é uma técnica estatística avançada usada para explorar e quantificar as relações entre dois conjuntos multivariados de variáveis. Ao contrário das correlações simples, que avaliam a relação entre pares de variáveis, a correlação canônica permite analisar o relacionamento entre dois grupos de variáveis simultaneamente. Essa abordagem busca encontrar combinações lineares de variáveis em cada conjunto, chamadas de variáveis canônicas, que maximizam a correlação entre os dois conjuntos. Por exemplo, ao trabalhar com dois grupos de variáveis, como variáveis nutricionais proteicas (lisina, metionina e triptofano) e agronômicas (massa de grãos, diâmetro da espiga, comprimento da espiga e comprimento de grão) a correlação canônica pode

ajudar a entender como diferentes combinações dessas variáveis se relacionam, fornecendo uma visão mais global das interações multivariadas. A principal vantagem da correlação canônica é a identificação de padrões complexos entre múltiplas variáveis.

No entanto, ao comparar a correlação canônica com os métodos de correlação não linear como Spearman, Kendall e Coeficiente de Máxima Informação (MIC), há diferenças no tipo de relação que cada método consegue identificar e nas situações em que são mais apropriados. Tanto a correlação de Spearman quanto a correlação de Kendall são utilizadas para identificar relações monotônicas entre variáveis. Isso indica que, mesmo que a relação entre as variáveis não seja linear, ambas as abordagens conseguem identificar uma tendência consistente de aumento ou diminuição entre as variáveis. A diferença entre os métodos está na maneira com que os coeficientes são estimados e no tipo de classificação que utilizam, mas ambas são adequadas para verificar as relações em dados não lineares ou quando há presença de outliers. No entanto, ambos os métodos, são univariados, ou seja, avaliam a relação entre duas variáveis de cada vez, enquanto a correlação canônica avalia múltiplas variáveis.

Por outro lado, o MIC é um método que busca identificar relações não lineares complexas, inclusive aquelas que não são monotônicas. É um método adequado quando há padrões complexos entre as variáveis, como relações curvas que não podem ser capturados por correlações tradicionais. Ao contrário de Spearman e Kendall, o MIC é capaz de detectar relações tanto com relações monotônicas quanto com não monotônicas.

Enquanto a correlação canônica é uma técnica multivariada que explora a relação entre dois conjuntos de variáveis, os métodos de correlação como Spearman, Kendall e MIC são univariados que avaliam as relações entre pares de variáveis. Logo, a correlação canônica é mais adequada quando o objetivo é avaliar padrões lineares em conjuntos de dados com múltiplas variáveis. Portanto, se as relações entre as variáveis forem monotônicas, então a correlação de Spearman ou Kendall pode ser uma boa escolha. Para relações não lineares mais complexas, o MIC seria mais apropriado. Já a correlação canônica é ideal para casos em que se deseja entender a interação entre múltiplas variáveis simultaneamente.

12. AVALIAÇÃO DOS MÉTODOS EM CENÁRIOS LINEARES, MONOTÔNICA NÃO LINEAR E COMPLEXA

Para analisar diferentes tipos de correlação, como Pearson, Spearman, Kendall e MIC, foram simulados dados representativos de várias relações entre variáveis, e calculados os coeficientes de correlação para cada cenário.

12.1 Cenário com relações lineares

Neste cenário, os dados simulam uma relação linear entre massa dos grãos (MG) e massa da espiga (ME). A correlação de Pearson (0,667) indicou uma relação linear moderada entre as variáveis. Este valor é esperado, uma vez que o método de Pearson é considera dependências lineares. A correlação de Spearman (0,597) também apresentou um valor positivo, embora mais baixo, indicando uma relação monotônica crescente. A correlação de Kendall (0,426), reflete uma dependência moderada, porém mais fraca que Pearson e Spearman.

O MIC para este cenário foi calculado em 0,376, indicando uma dependência moderada entre as variáveis. O valor de MIC mais baixo sugere que a relação entre MG e ME pode ser bem modelada por uma dependência linear, mas com algumas relações não lineares, que não são identificadas completamente pelas correlações tradicionais. O MIC, ao considerar tanto relações lineares quanto não lineares, revela que uma parte da variação nas variáveis pode não ser completamente explicada por uma simples relação linear.

Portanto, Pearson, Spearman e Kendall identificaram a natureza da relação entre as variáveis, mas Pearson foi o mais eficaz, devido à linearidade da relação. O MIC demonstrou sua utilidade ao avaliar relações não lineares, mas seu valor relativamente baixo sugere que a relação entre as variáveis pode ser suficientemente explicada por uma dependência linear. Logo, em cenários em que há dependência linear, a correlação de Pearson mostrou-se suficiente para avaliar o grau de associação entre as variáveis.

```r
cor(MG_linear, ME_linear, method = "pearson")
```

```
## [1] 0.6671146
```

```r
cor(MG_linear, ME_linear, method = "spearman")
```

```
## [1] 0.5971077
```

```r
cor(MG_linear, ME_linear, method = "kendall")
```

```
## [1] 0.4262626
```

```r
MIC1 <- minerva::mine(MG_linear, ME_linear)
MIC1$MIC
```

```
## [1] 0.3760096
```

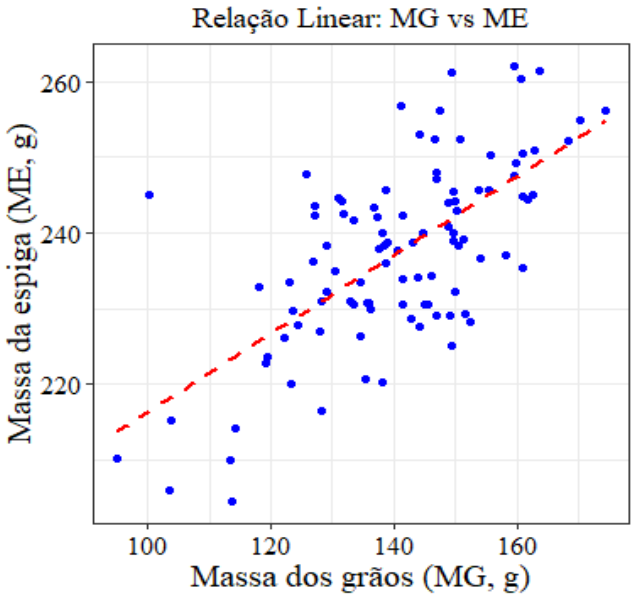

Figura 3. Exemplo de relação linear entre massa de grãos (MG, g) e massa da espiga (ME, g).

12.2 Cenário com relações monotônicas não lineares

Neste cenário, a relação entre MG e AMI segue uma curva monotônica crescente, mas com uma relação não linear. O coeficiente de correlação de Pearson (0,843) foi relativamente alto, indicando que há uma boa relação linear, embora a forma da relação seja na realidade não linear. A análise de correlação de Spearman (0,864) e Kendall (0,678), identificaram uma relação monotônica de forma mais robusta, com a correlação de Spearman sendo o mais sensível a essa forma de dependência. Logo, a correlação de Spearman é mais adequada para detectar relações monotônicas não lineares, como observado neste cenário. O valor de MIC = 1 para esse cenário reflete uma dependência forte e não linear entre MG e AMI, o que confirma que o MIC é extremamente eficaz em identificar relações não lineares complexas. O valor máximo de MIC sugere que a dependência entre essas variáveis não pode ser totalmente explicada por uma relação linear.

Portanto, a correlação de Spearman foi a mais eficaz para este cenário, devido à sua capacidade de capturar relações monotônicas não lineares. Já a correlação de Pearson, embora positiva, falhou em capturar completamente a natureza não linear da relação. O MIC, com um valor de 1, foi o método mais eficiente para identificar a dependência não linear neste cenário, destacando sua superioridade sobre os métodos tradicionais para esse tipo de relação.

```
cor(MG_monotonic, AMI_monotonic, method = "pearson")
## [1] 0.8426995
cor(MG_monotonic, AMI_monotonic, method = "spearman")
## [1] 0.8644944
```

```r
cor(MG_monotonic, AMI_monotonic, method = "kendall")

## [1] 0.6779798

MIC2 <- minerva::mine(MG_monotonic, AMI_monotonic)
MIC2$MIC

## [1] 1
```

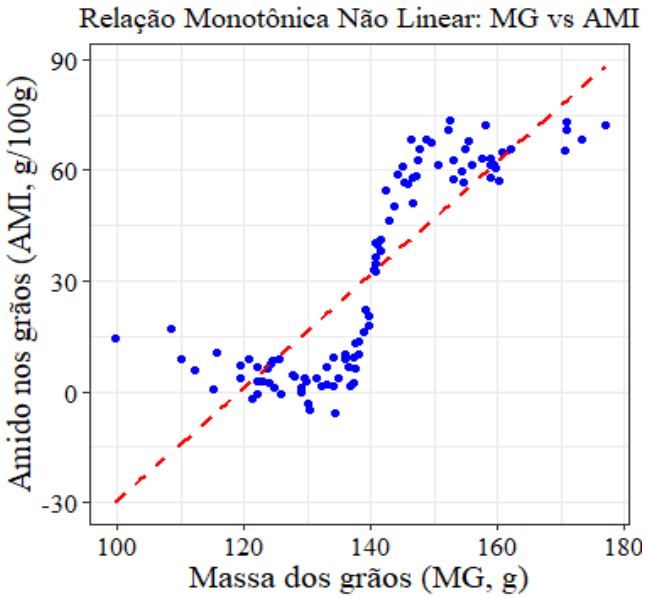

Figura 4. Exemplo de relação não linear monotônica entre massa de grãos (MG, g) e o teor de amido (AMI, g/100g).

12.3 Cenário com relações complexas não monotônica

Neste cenário, a relação entre MG e RS foi complexa e não monotônica, o que torna a captura dessa dependência mais difícil para os métodos de correlação tradicionais. A correlação de Pearson (0,152), Spearman (0,134), e Kendall (0,089) indicam uma dependência fraca, praticamente inexistente. Os métodos de correlação clássicos não conseguiram identificar as variações não lineares ou a estrutura mais complexa da relação, o que é esperado dada a relação não monotônica entre as variáveis.

No entanto, o MIC (0,264) indicou uma dependência fraca, mas ainda assim existe uma relação mais substancial que a capturada pelas correlações. O MIC, sendo mais sensível a relações complexas e não lineares, detecta uma pequena dependência entre as variáveis, algo que não é evidente nas correlações tradicionais. Portanto, os métodos de correlação clássica falharam em capturar qualquer dependência significativa, o que reforça sua limitação em cenários onde as relações são não lineares e complexas. O MIC foi o único método a identificar uma dependência, mesmo que fraca.

```r
cor(MG_complex, RS_complex, method = "pearson")
```

```
## [1] 0.1524039
```

```r
cor(MG_complex, RS_complex, method = "spearman")
```

```
## [1] 0.1338374
```

```r
cor(MG_complex, RS_complex, method = "kendall")
```

```
## [1] 0.08929293
```

```r
MIC3 <- minerva::mine(MG_complex, RS_complex)
MIC3$MIC
```

```
## [1] 0.2644765
```

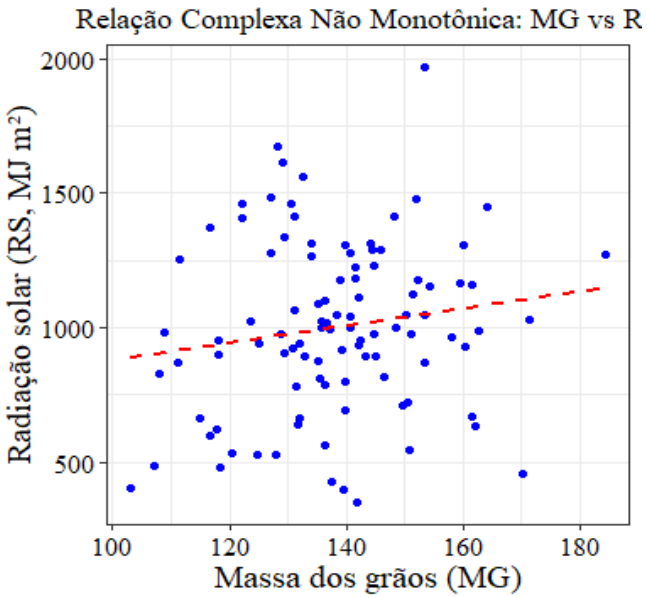

Figura 5. Exemplo de relação não linear complexa entre massa de grãos (MG, g) e radiação solar global (RS, MJ m²).

12.4 Cenário com relações complexas tipo U

Neste cenário, a relação entre MG e Tar segue uma curva em forma de U, o que caracteriza uma relação não linear altamente complexa. As correlações de Pearson (próximo de zero), Spearman (−0,0017), e Kendall (−0,0008) mostram valores muito baixos, indicando que esses métodos não identificaram qualquer tipo de dependência devido à forma não linear da relação. Esses métodos são inadequados para detectar padrões em U, que não seguem uma tendência linear ou monotônica.

Enquanto, o MIC (1,00) revelou uma dependência extremamente forte entre as variáveis. O MIC foi eficaz em capturar a complexidade de relações como a observada neste cenário. O valor de MIC máximo indica que, apesar de não existir uma relação linear ou monotônica, há uma relação robusta identificada pelo MIC. Logo, os métodos tradicionais de correlação não foram eficazes para

identificar a relação tipo U, destacando suas limitações quando as relações são altamente não lineares. O MIC, ao identificar a dependência não linear complexa, demonstrou vantagem em cenários com relações em formas não convencionais, como U.

```
cor(MG, Tar, method = "pearson")

## [1] -1.665263e-16

cor(MG, Tar, method = "spearman")

## [1] -0.00172837

cor(MG, Tar, method = "kendall")

## [1] -0.0008112005

MIC4 <- minerva::mine(MG, Tar)
MIC4$MIC

## [1] 1
```

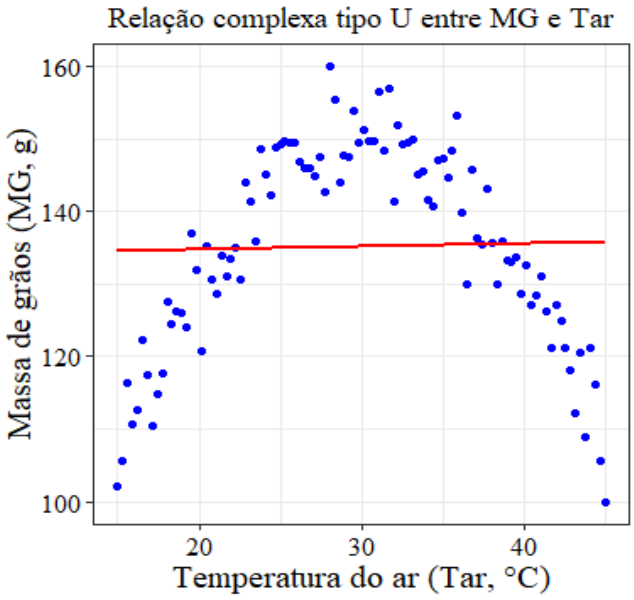

Figura 6. Exemplo de relação não linear complexa tipo U entre massa de grãos (MG, g) e temperatura do ar (Tar, ºC).

A análise dos diferentes cenários mostrou como cada metodologia identifica as relações entre as variáveis, dependendo da natureza dessas relações. Os métodos tradicionais de correlação (Pearson, Spearman, Kendall) são eficazes para identificar relações lineares ou monotônicas. No entanto, apresentam limitações quando a relação entre as variáveis não é linear ou monotônica. O Pearson é altamente eficaz para relações lineares, mas falha quando a relação é monotônica não linear ou complexa. O Spearman e o Kendall funcionam melhor em relações monotônicas, mas não conseguem lidar com padrões complexos e não lineares como U.

O MIC se destaca ao capturar relações não lineares complexas e padrões difíceis de identificar com os métodos tradicionais. Mesmo em cenários em que as correlações tradicionais falharam (como nas relações tipo U ou complexas não monotônicas), o MIC conseguiu identificar uma dependência, destacando sua flexibilidade e capacidade de detectar padrões mais complexos. Em resumo, enquanto os métodos tradicionais são úteis para detectar dependências lineares ou monotônicas, o MIC oferece uma vantagem significativa na identificação de relações não lineares complexas.

QUESTÕES PROFESSOR DR. MARCOS TOEBE

Questão 1. Correlação genética, correlação fenotípica e correlação de ambiente. Como são calculadas e qual a utilidade prática de suas estimativas para o melhoramento e seleção de plantas? Qual destas correlações seria a mais adequada para ser utilizada na análise de trilha, nas correlações canônicas, regressão *stepwise* e demais análises do seu trabalho? Explique e justifique! **Resposta: 13. CORRELAÇÃO GENÉTICA, FENOTÍPICA E AMBIENTAL.**

Questão 2. Como a ausência de repetições pode afetar a avaliação da qualidade experimental do seu trabalho e qual estratégia poderia ser utilizada para minimizar o impacto da heterogeneidade experimental nos resultados do seu estudo? Qual seu entendimento sobre delineamentos experimentais complementares e qual delineamento alternativo poderia ter sido empregado para estimar os efeitos do erro experimental? **Resposta: 14. QUALIDADE EXPERIMENTAL.**

13. CORRELAÇÃO GENÉTICA, FENOTÍPICA E AMBIENTAL

13.1 Estimativas dos coeficientes de correlação

Os programas de melhoramento têm como objetivo principal desenvolver cultivares superiores em relação a um conjunto de características desejáveis. Nesse contexto, compreender a natureza e a magnitude das correlações entre os caracteres de interesse torna-se essencial. As inter-relações entre os caracteres são geralmente avaliadas por meio das correlações genotípicas, fenotípicas e ambientais. A correlação fenotípica é obtida diretamente das medições fenotípicas e reflete tanto causas genéticas quanto ambientais. No entanto, apenas a correlação genotípica, que representa a porção genética da correlação fenotípica, é utilizada como base para direcionar programas de melhoramento, uma vez que é a única de caráter herdável.

Para compreender como são calculados os coeficientes de correlação fenotípica, genotípica e ambiental será utilizado um exemplo com dados da tese, sendo considerados 29 genótipos de milho avaliados em três datas de semeadura. Para isso, as datas de semeadura foram consideradas como os blocos. Logo, para a análise considerou-se um delineamento experimental blocos casualizados com três repetições. Neste exemplo, consideram-se as variáveis massa de grãos da espiga (MG, g) e altura de planta (AP, cm). Assim, busca-se estimar os coeficientes de correlação fenotípico, genotípico e ambiental entre as variáveis MG (x) e AP (y).

Para iniciar as estimativas dos coeficientes, procedeu-se com a análise de variância para cada variável. Assim, obteve-se o quadrado médio de tratamento para MG (QMG_x) e AP (QMG_y) e quadrado médio do resíduo para MG (QMR_x) e AP (QMR_y). Em seguida, os valores originais de MG e AP foram somados, ou seja, para cada genótipo em cada bloco, somaram-se os valores de MG e AP criando uma nova variável chamada X+Y. Para essa variável X+Y realizou-se a análise de variância

para se obter o quadrado médio de tratamento de XY (QMG_{x+y}) e o quadrado médio do resíduo de XY (QMR_{x+y}) (Tabela 1).

Tabela 1. Análise de variância das variáveis massa de grãos da espiga (MG, g), altura de planta (AP, cm) e soma X+Y avaliados em 29 genótipos de milho.

FV	GL	QM			E(QM)
		MG (x)	AP (y)	MG+AP (XY)	
Blocos	2	12037,30	2130,23	23650,10	
Genótipos	28	1122,90	673,82	2650,20	$\sigma^2 + r\sigma_g^2$
Resíduo	86	863,70	204,36	1403,10	σ^2

FV: fonte de variação; GL: graus de liberdade; QM: quadrado médio; E(QM): esperança do quadrado médio.

Pode-se observar que a esperança do quadrado médio de genótipos para cada variável é dada pela por: $\sigma^2 + r\sigma_g^2$, ou seja, há efeitos do resíduo somado com a multiplicação do número de repetições pela variância genética. Em seguida, foram calculados os produtos médios associados a tratamentos (PMG_{xy}) e resíduo. O produto médio associado a tratamento é obtido pela expressão: $PMG_{xy} = (QMG_{x+y} - QMG_x - QMG_y)/2$, enquanto o produto médio associado ao resíduo é obtido pela expressão: $PMR_{xy} = (QMR_{x+y} - QMR_x - QMR_y)/2$. Assim, calcularam-se:

$PMG_{xy} = (2650,20 - 1122,90 - 673,82)/2 = 426,74$
$PMR_{xy} = (1403,10 - 863,70 - 204,36)/2 = 167,52$

Os valores dos produtos médios de tratamentos (genótipos) e resíduo estão apresentados na Tabela 2. Observa-se que a esperança do produto médio considera as duas variáveis, MG (x) e AP (y).

Tabela 2. Análise dos produtos médios e suas respectivas esperanças matemáticas.

FV	GL	PM	E(PM)
Bloco	2	-	
Genótipos	28	426,74	$\sigma_{xy}^2 + r\sigma_{gxy}^2$
Resíduo	86	167,52	σ_{xy}^2

FV: fonte de variação; GL: graus de liberdade; PM: produto médio; E(PM): esperança do produto médio.

A partir das estimativas dos quadrados médios e produtos médios é possível calcular os coeficientes de correlação fenotípica, genotípica e ambiental. Para estimativa do coeficiente de correlação, é realizada uma razão entre a covariância das duas variáveis com a raiz do produto das

variâncias das variáveis. Logo, para calcular o coeficiente de correlação fenotípica o PMG_{xy} corresponde a covariância entre MG e AP, enquanto o QMG_x e QMG_y correspondem a variância de MG e AP, respectivamente. O PMR_{xy} corresponde a covariância entre MG e AP, enquanto o QMR_x e QMR_y correspondem a variância de MG e AP, respectivamente, para calcular o coeficiente de correlação ambiental. Para as estimativas do coeficiente de correlação genotípica utiliza-se a σ^2_{gxy} que corresponde a covariância entre MG e AP e a σ^2_{gx} e a σ^2_{gy} que correspondem a variância da MG e a variância de AP, respectivamente. Logo, as estimativas são realizadas pelas equações a seguir:

i) Correlação fenotípica

$$r_f = \frac{PMG_{xy}}{\sqrt{QMG_x.\ QMG_y}}$$

ii) Correlação ambiental

$$r_a = \frac{PMR_{xy}}{\sqrt{QMR_x.\ QMR_y}}$$

iii) Correlação genotípica

$$r_g = \frac{\sigma^2_{gxy}}{\sqrt{\sigma^2_{gx}.\ \sigma^2_{gy}}}$$

Com essas expressões, serão calculados os coeficientes de correlação serão calculados para os dados do exemplo:

i) Correlação fenotípica

$$r_f = \frac{426{,}74}{\sqrt{1122{,}90\ .\ 673{,}82}} = 0{,}49$$

ii) Correlação ambiental

$$r_a = \frac{167{,}52}{\sqrt{863{,}70\ .\ 204{,}36}} = 0{,}40$$

iii) Correlação genotípica

Para estimar a correlação genotípica, é necessário obter as estimativas dos componentes de variância: σ^2_{gx}, σ^2_{gy} e σ^2_{gxy}. Esses componentes são obtidos observando a esperança dos quadrados

médios e dos produtos médios. Por exemplo, a E(QM) de genótipos é $\sigma^2 + r\sigma_g^2$, enquanto a E(QM) do resíduo é σ^2. Deste modo, pode-se isolar σ_g^2 da seguinte forma:

$$\sigma_g^2 = \frac{QMG - QMR}{r} = \frac{\sigma^2 + r\sigma_g^2 - \sigma^2}{r}$$

Observa-se que por meio da expressão é possível isolar a variância genotípica (σ_g^2) de cada variável. Portanto, as estimativas de σ_{gx}^2 e σ_{gy}^2 foram realizadas da seguinte forma:

$$\sigma_{gx}^2 = \frac{1122{,}90 - 863{,}70}{3} = \frac{352{,}80}{3} = 86{,}40$$

$$\sigma_{gy}^2 = \frac{873{,}82 - 204{,}36}{3} = \frac{669{,}46}{3} = 223{,}15$$

A estimativa de σ_{gxy}^2 pode ser realizada por meio da observação da E(PM). Por exemplo, a E(PM) de genótipos é $\sigma_{xy}^2 + r\sigma_{gxy}^2$, enquanto a E(PM) do resíduo é σ_{xy}^2. Deste modo, pode-se isolar σ_{gxy}^2 da seguinte forma:

$$\sigma_{gxy}^2 = \frac{PMG - PMR}{r} = \frac{\sigma_{xy}^2 + r\sigma_{gxy}^2 - \sigma_{xy}^2}{r}$$

Logo, obtêm-se σ_{gxy}^2 da seguinte forma:

$$\sigma_{gxy}^2 = \frac{426{,}74 - 167{,}52}{3} = \frac{259{,}22}{3} = 86{,}41$$

A partir dos componentes de variância σ_{gx}^2 (86,40) σ_{gy}^2 (223,15) e σ_{gxy}^2 (86,41) pode-se calcular o coeficiente de correlação genotípica:

iii) Correlação genotípica

$$r_g = \frac{86{,}41}{\sqrt{86{,}50\ .\ 223{,}15}} = 0{,}62$$

13.2 Utilidade dos coeficientes de correlação no melhoramento de plantas

A correlação genética entre caracteres procura explicar por meio de mecanismos genéticos, a variação conjunta de duas variáveis. Essas correlações podem ser explicadas por dois fenômenos: pleiotropia e ligação gênica. A pleiotropia consiste em um mesmo gene determinar uma ou mais características (RAMALHO, 2012). Por exemplo, um gene que determina o ângulo da folha de milho pode, simultaneamente, determinar o número de ramificações do pendão, como descrito por Bertolini et al. (2023). Já a ligação gênica refere-se à proximidade física de dois ou mais genes no mesmo cromossomo, o que reduz a probabilidade de que sejam separados durante a recombinação meiótica (RAMALHO et al., 2012). Essa proximidade pode resultar em uma associação não independente entre os caracteres controlados por esses genes. Por exemplo, genes que determinam o número de

fileiras de grãos e a massa de grãos na espiga podem estar fisicamente ligados, levando a uma correlação genética entre esses caracteres.

Logo, pode-se observar que a correlação genotípica entre a MG e AP foi superior a correlação fenotípica. Essa diferença está relacionada às origens das variações que contribuem para as correlações e ao fato de que o coeficiente genotípico reflete apenas a variação genética comum entre dois caracteres, excluindo os efeitos do ambiente. A correlação genotípica mede a relação entre dois caracteres considerando apenas a variação genética, ou seja, a proporção da variação fenotípica que é herdável. Como considera apenas os efeitos genéticos, a correlação genotípica tende a ser mais alta quando os caracteres possuem bases genéticas fortemente associadas (pleiotropia ou ligação gênica).

A correlação fenotípica é uma combinação das correlações genéticas e ambientais. Por ser influenciada por efeitos ambientais que podem atenuar ou confundir a relação entre os caracteres, a correlação fenotípica tende a ser menor do que a genotípica, dependendo da magnitude dos efeitos ambientais. Enquanto a correlação ambiental, indica a relação entre dois caracteres atribuída exclusivamente às influências ambientais. Em muitos casos, os efeitos ambientais são aleatórios ou não sistemáticos, o que geralmente resulta em correlações ambientais menores. A correlação ambiental de 0,40 entre a MG e a AP em milho indica que os fatores ambientais tendem a influenciar esses dois caracteres de forma similar. Por exemplo, condições ambientais favoráveis, podem promover o aumento simultâneo da MG e AP.

Se essa correlação fosse próxima de zero, significaria que os fatores ambientais teriam pouco impacto conjunto sobre MG e AP. Por exemplo, a quantidade de radiação solar global poderia influenciar mais a MG do que a AP. Por outro lado, uma correlação ambiental negativa indicaria que os fatores ambientais exercem efeitos opostos sobre os dois caracteres, como no caso de uma alta densidade de plantio que aumenta a altura das plantas devido à competição por luz, mas reduz a produção de grãos por planta.

As correlações fenotípica, genotípica e ambiental possuem diferentes utilidades práticas no melhoramento e seleção de plantas, cada uma contribuindo de forma específica para o entendimento das relações entre caracteres e suas implicações na seleção de genótipos superiores. No entanto, a correlação genotípica promove maior eficiência na seleção de plantas. A correlação genotípica considera apenas a parte da variação que é geneticamente controlada, sendo derivada dos efeitos genéticos estimados para os caracteres. Por representar a relação herdável entre os caracteres, é a mais relevante para orientar programas de melhoramento, pois indica quais associações podem ser exploradas para ganhos genéticos simultâneos.

A produtividade de grãos tem sido a principal característica utilizada para identificar genótipos adaptados as condições de estresse abiótico (ERTIRO et al., 2017; NDLOVU et al., 2022). No entanto, em razão da menor contribuição genética na expressão desta característica, a seleção baseada

em caracteres secundários tem se tornado eficiente (LIMA et al., 2022). Principalmente para adaptação a ambientes com estresse causado por variáveis meteorológicas, o intervalo entre florescimento masculino e feminino tem sido o principal caractere secundário utilizado (LIMA et al., 2022; ZAIDI et al., 2022). De acordo com Parajuli et al. (2018), uma das melhores abordagens nos programas de melhoramento é utilizar características secundárias para selecionar os genótipos com melhor desempenho sob estresse. No entanto, uma característica secundária precisa preencher o critério de ser geneticamente variável e hereditária, rápida de medir e associada à produção sob estresse.

Assim, imagina-se que um pesquisador está avaliando a produtividade de grãos de 300 progênies de meios-irmãos. Após avaliar as 300 progênies, realizou a estimativa de herdabilidade para a produtividade de grãos, que foi de 0,20, ou seja, uma baixa herdabilidade. Isso indica que ao selecionar as progênies que exibiram o maior desempenho produtivo, ou seja, baseado no fenótipo, apenas 20% da expressão da produtividade de grãos foi determinada por efeitos genéticos, enquanto o ambiente contribuiu com 80%. Assim, a elevada produtividade das progênies poderia ter sido em função do ambiente (maior fertilidade do solo onde estavam as parcelas) e não devido a maior potencial genético. Para minimizar esse problema, pode-se identificar variáveis que apresentar uma maior herdabilidade e que tenham uma alta correlação genética com a produtividade de grãos para realizar uma seleção indireta. Por exemplo, o intervalo entre o florescimento masculino e feminino apresenta uma alta herdabilidade e tem uma elevada correlação com a produtividade de grãos (ABU et al., 2021; LIMA et al., 2022). Portanto, podem ser identificadas as progênies que exibem um menor intervalo entre o florescimento masculino e feminino e assim, estaria selecionando indiretamente as progênies de maior produtividade de grãos.

Portanto, no contexto de análises como análise de trilha, correlações canônicas e regressão *stepwise*, a correlação genotípica é a mais adequada. Isso ocorre porque essas análises visam compreender relações causais ou identificar caracteres que possam ser usados como preditores confiáveis em programas de melhoramento, e apenas a correlação genotípica reflete a porção herdável e, portanto, diretamente útil para ganhos genéticos. Por exemplo, na análise de trilha, utilizar a correlação genotípica assegura que os efeitos diretos e indiretos identificados são baseados em variações genéticas, evitando a interferência de fatores ambientais.

14. QUALIDADE EXPERIMENTAL

A ausência de repetições em um experimento, como o realizado com 78 genótipos de milho avaliados em 10 datas de semeadura dificulta a separação precisa da variação atribuída aos efeitos genotípicos da variação causada por erros experimentais ou pela heterogeneidade ambiental. Em

experimentos sem repetições, não há uma estimativa direta da variância experimental, o que dificulta a obtenção de inferências, como a comparação confiável entre os genótipos. Nesta situação, a variabilidade ambiental ou microambiental pode ser confundida com variação genotípica, o que dificulta a avaliação do desempenho dos genótipos. Em cada data de semeadura, as condições, como temperatura do ar, radiação solar global, soma térmica e precipitação, determinam o desempenho dos genótipos avaliados. Logo, um experimento sem repetições dentro de cada data dificulta a avaliação do desempenho genotípico dos genótipos.

Uma alternativa para avaliar o valor genotípico dos genótipos em experimentos sem repetições é o uso de modelos mistos. Nesses modelos, os efeitos genotípicos podem ser considerados como fixos e os efeitos ambientais, como datas de semeadura, podem ser modelados como aleatórios. Esses modelos podem incorporar covariáveis ambientais, como a temperatura do ar, radiação solar global, soma térmica, para explicar parte da heterogeneidade. Incluir covariáveis que descrevam as condições ambientais melhora a explicação da variação e aumenta a precisão das estimativas genotípicas. Essas covariáveis contínuas determinam diretamente o desempenho dos genótipos.

Para fins de exemplo, aplicou-se o modelo linear misto para verificar o valor genotípico de diferentes genótipos, incluindo covariáveis ambientais como a radiação solar global no estádio reprodutivo (RFMC) e a soma térmica no estádio reprodutivo (SFMC) sobre a massa de grãos (MG). O modelo incluiu os genótipos e as variáveis RFMC e SFMC como efeitos fixos, enquanto as datas de semeadura (DataC) foram consideradas como um efeito aleatório. Em relação aos efeitos fixos, o intercepto foi significativo, com uma estimativa de 67,99. Entre os genótipos, destacaram-se estimativas positivas como o genótipo B2620 (39,30) e negativas como IPR164 (-23,53). As variáveis RFMC e SFMC foram significativas. RFMC teve um efeito positivo (0,1571) e SFMC apresentou um efeito negativo (-0,2311), com ambos os efeitos indicando contribuição significativa na predição da MG.

```
#------------1. ACESSAR O BANCO DE DADOS

#URL do arquivo no GitHub (link para o arquivo .xlsx raw)
url <- "https://github.com/muriloloro/Modelos-de-Predicao/raw/main/dados-MLM.xl
sx"
library(httr)
library(readxl)
# Definir o caminho temporário para salvar o arquivo
temp_file <- tempfile(fileext = ".xlsx")
GET(url, write_disk(temp_file, overwrite = TRUE))

dados <- read_excel(temp_file)


#------------2. APLICAÇÃO DO MODELO MISTO
```

```r
library(lme4)

dados$Genotipo <- as.factor(dados$Genotipo)

# Modelo misto
modelo <- lmer(MG ~ Genotipo + RFMC + SFMC + (1|DataC), data = dados)
summary(modelo)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: MG ~ Genotipo + RFMC + SFMC + (1 | DataC)
##    Data: dados
##
## REML criterion at convergence: 6568.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1869 -0.5946 -0.0238  0.5954  3.1832
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  DataC    (Intercept) 511.0    22.61
##  Residual             537.3    23.18
## Number of obs: 773, groups:  DataC, 10
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)           71.49141   15.56495   4.593
## Genotipo20A78          1.40698   10.36745   0.136
## Genotipo30A95         -2.83853   10.37024  -0.274
## Genotipo36680          2.00737   10.36797   0.194
## Genotipo36770         29.09659   10.37094   2.806
## Genotipo36790         10.24113   10.36907   0.988
## Genotipo36799          0.78715   10.38549   0.076
## GenotipoAG1051        -9.96683   10.37092  -0.961
## GenotipoAG8690         4.16144   10.37764   0.401
## GenotipoAG8780         1.43635   10.36612   0.139
## GenotipoAG9021        -3.57348   10.39825  -0.344
## GenotipoAG9025        -5.18368   10.39115  -0.499
## GenotipoALVARÉ        -9.13921   10.36651  -0.882
## GenotipoAS1555        -5.17289   10.39301  -0.498
## GenotipoAS1633         3.29252   10.37164   0.317
## GenotipoAS1677        -3.08863   10.37764  -0.298
## GenotipoAS1730         7.54898   10.37928   0.727
## GenotipoB2401         24.47943   10.36726   2.361
## GenotipoB2418        -12.77869   10.42790  -1.225
## GenotipoB2620         33.78063   10.36650   3.259
## GenotipoB2688         10.68513   10.37016   1.030
## GenotipoB2801         13.59239   10.37010   1.311
## GenotipoBANDEIRANTE  -25.55566   10.40601  -2.456
## GenotipoBM270         19.11481   10.37630   1.842
## GenotipoBM3063        12.86916   10.36997   1.241
## GenotipoBM3066        15.63277   10.37560   1.507
## GenotipoBM3069        16.28269   10.37554   1.569
## GenotipoBM915          3.14772   10.37124   0.304
## GenotipoBR106         -6.17739   10.36823  -0.596
## GenotipoBRS PLANALTO  -5.33070   10.39542  -0.513
```

```
## GenotipoCODIGO       -15.45584   10.40599  -1.485
## GenotipoDKB177         5.45451   10.36734   0.526
## GenotipoDKB230        -4.06336   10.39740  -0.391
## GenotipoDKB235        -1.16050   10.37180  -0.112
## GenotipoDKB240        -4.00115   10.38771  -0.385
## GenotipoDKB255         3.45313   10.36890   0.333
## GenotipoDKB290        -1.97660   10.41298  -0.190
## GenotipoFEROZ          7.90408   10.37019   0.762
## GenotipoFS400         18.74327   10.37308   1.807
## GenotipoFS533         19.55526   10.37669   1.885
## GenotipoFS670         16.30069   10.37723   1.571
## GenotipoIPR164       -24.85083   10.37329  -2.396
## GenotipoK0167          6.61928   10.37501   0.638
## GenotipoK3100          5.74513   10.37491   0.554
## GenotipoK8774          5.85193   10.37255   0.564
## GenotipoK9300         -9.62582   10.40702  -0.925
## GenotipoK9606         14.97092   10.37421   1.443
## GenotipoK9660         12.78458   10.37069   1.233
## GenotipoLAVRADOR     -17.70368   10.37403  -1.707
## GenotipoLG3040        -3.39416   10.37860  -0.327
## GenotipoM274           1.62924   10.66793   0.153
## GenotipoMAXIMUS       -4.91950   10.37320  -0.474
## GenotipoMG300         -4.85671   10.38378  -0.468
## GenotipoMG580          9.40694   10.37865   0.906
## GenotipoMG593         21.85779   10.36900   2.108
## GenotipoMG618         10.49214   10.36981   1.012
## GenotipoMG652          7.91269   10.37378   0.763
## GenotipoMG699         19.26352   10.37136   1.857
## GenotipoNK467         -0.03321   10.36812  -0.003
## GenotipoNK520         -0.95678   10.37718  -0.092
## GenotipoNS45         -10.82352   10.40978  -1.040
## GenotipoNS75          18.50230   10.37418   1.783
## GenotipoNS80           4.82298   10.36989   0.465
## GenotipoNTX303        -8.27684   10.37242  -0.798
## GenotipoNTX454         8.31725   10.36768   0.802
## GenotipoNTX468        -8.16110   10.65639  -0.766
## GenotipoP3016         17.89213   10.37282   1.725
## GenotipoP3565         11.97829   10.36856   1.155
## GenotipoPIRATININGA  -12.47012   10.37101  -1.202
## GenotipoPR27D28       10.07314   10.66327   0.945
## GenotipoPR27D29       12.37232   10.66153   1.160
## GenotipoPR27D30       -9.80901   10.65780  -0.920
## GenotipoROBUSTO      -27.63220   10.38350  -2.661
## GenotipoSH5050        -2.35382   10.66791  -0.221
## GenotipoSHS5560        4.86483   10.37126   0.469
## GenotipoSHS7939       10.90456   10.37200   1.051
## GenotipoSHSUPER       16.91075   10.38012   1.629
## GenotipoSOBERANO       1.77858   10.66150   0.167
## RFMC                   0.16844    0.03556   4.737
## SFMC                  -0.24924    0.05358  -4.652

##
## Correlation matrix not shown by default, as p = 80 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```r
anova(modelo)
```

```
## Analysis of Variance Table
##          npar Sum Sq Mean Sq F value
## Genotipo   77 109397  1420.7  2.6444
## RFMC        1    507   506.8  0.9434
## SFMC        1  11626 11625.6 21.6383
```

```r
#3--------------EXTRAÇÃO DE INFORMAÇÕES NECESSÁRIAS

# Extrair os coeficientes fixos do modelo
blue <- fixef(modelo)
# Filtrar apenas os efeitos dos genótipos
efeitos_genotipicos <- blue[grep("Genotipo", names(blue))]
```

```r
#4---------------------CONSTRUÇÃO DO GRÁFICO COM OS EFEITOS GENOTÍPICOS

# Preparar os dados para o gráfico
efeitos_genotipicos_df <- data.frame(
  Genótipo = names(efeitos_genotipicos),
  Efeito = efeitos_genotipicos
)
library(ggplot2)
# Plotar os efeitos genotípicos
ggplot(efeitos_genotipicos_df, aes(x = reorder(Genótipo, Efeito), y = Efeito))
+
  geom_bar(stat = "identity", fill = "blue") +
  theme_minimal() +
  labs(title = "Efeito Genotípico sobre MG",
       x = "Genótipo",
       y = "Efeito Genotípico") +
  coord_flip() +
  theme_bw() +
  theme_classic() +
  theme(legend.position = "bottom",
        legend.text = element_text(size = 13),
        panel.background = element_blank(),
        axis.line = element_line(linewidth = 0.5, color = "#222222"),
        text = element_text(family="serif", size = 13),
        axis.text.y = element_text(size=13, color = "black"),
        axis.text.x = element_text(size = 13, angle = 0, vjust = 0.4, color="bl
ack"),
        axis.ticks = element_line(colour = 'black'),
        axis.ticks.length = unit(.25, "cm"),
        axis.ticks.x = element_line(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        plot.title = element_text(hjust = 0.45, vjust=2.12,
                                  colour = "black", size = 13, family = "serif"
))
```

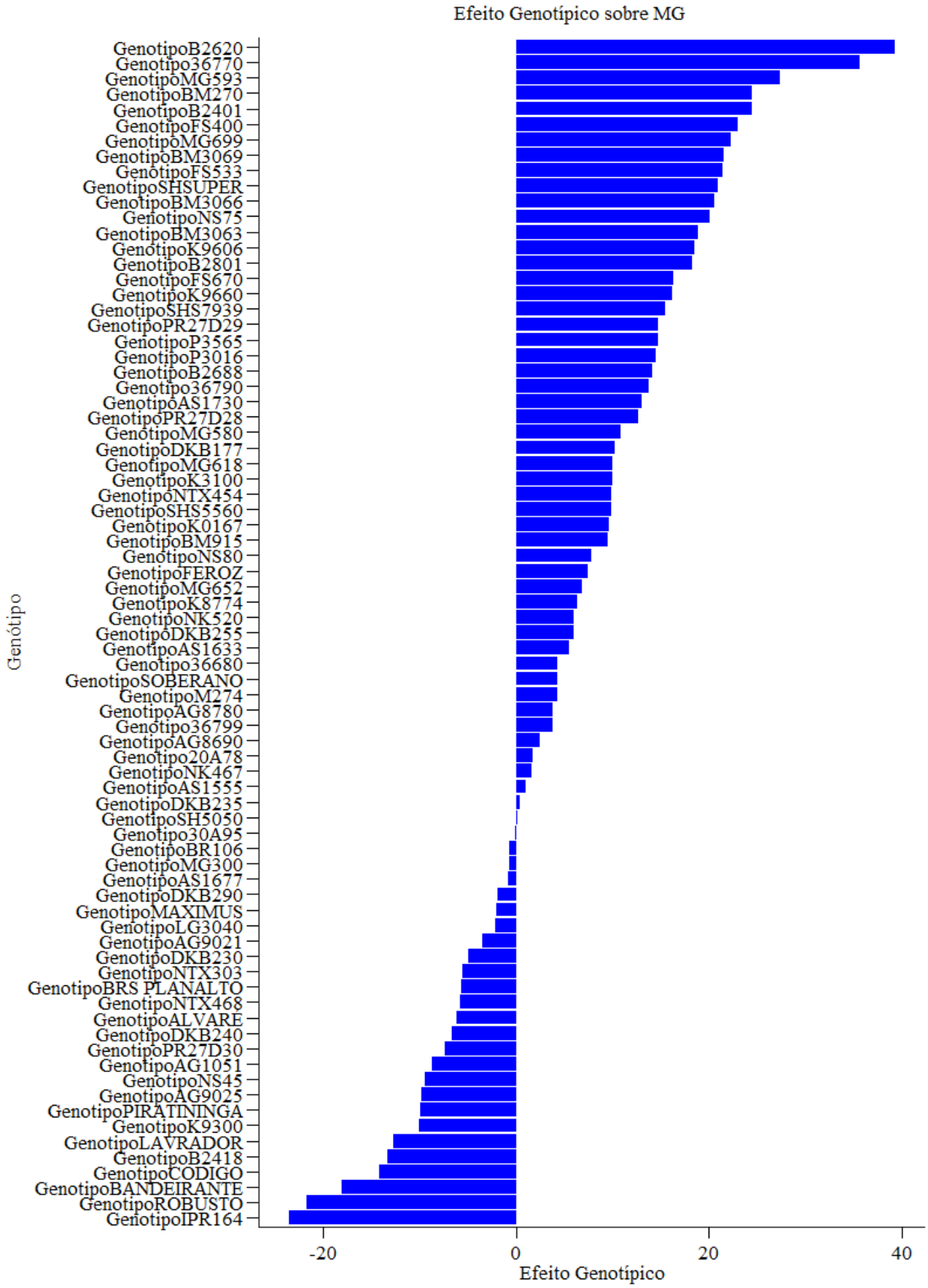

Figura 1. Valores genotípicos dos genótipos de milho para a massa de grãos da espiga (MG, g).

No planejamento do experimento, o adequado seria a utilização de um delineamento experimental para poder avaliar a precisão dos experimentos e, assim, poder verificar de forma precisa os efeitos de genótipos, datas de semeadura e interação genótipos x datas de semeadura. Os delineamentos básicos a serem utilizados em estudos agronômicos são o delineamento inteiramente casualizado, blocos casualizados e quadrado latino. Em função da existência de heterogeneidade entre os blocos (datas de semeadura) o delineamento de blocos ao acaso seria o mais adequado. No entanto, considerando o grande número de genótipos avaliados, seria difícil de garantir uma homogeneidade dentro dos blocos. Logo, os blocos incompletos são alternativas para essas situações.

Quando o número de tratamento é grande há necessidade de ter blocos maiores, que em muitas situações não é possível obter uma homogeneidade dentro dos blocos (PIMENTEL GOMES, 1982). Por exemplo, em experimentos com milho, muitas vezes tem-se centenas de híbridos ou progênies a serem avaliadas. Neste caso, poderia se utilizar de blocos ao acaso, mas os blocos seriam grandes que só em casos muitos favoráveis poderiam ser considerados uniformes. Uma solução consiste em usar delineamentos em blocos incompletos, dos quais são importantes os blocos incompletos balanceados e os reticulados quadrados, ou seja, latice quadrado.

Nos delineamentos em blocos incompletos balanceados cada tratamento aparece no mesmo bloco com cada um dos outros tratamentos e sempre o mesmo número de vezes. Por exemplo, em um experimento avaliando seis cultivares de milho (1 a 6), em 10 blocos de três parcelas, o par 2 e 4 (híbridos) aparecem nos blocos 2, 3 e 4 e 2, 4 e 6. Neste caso, tem-se que cada tratamento aparece no mesmo bloco com um dos tratamentos e sempre em duas vezes. Assim, todos os pares têm a mesma chance de serem comparados diretamente. Esse tipo de delineamento é usado para situações onde não consegue colocar todos os tratamentos em um bloco, mas ainda assim precisa de comparações justas. O balanceamento garante que todos os tratamentos sejam avaliados em relação aos outros de forma uniforme.

Há casos em que os blocos podem ser reunidos de modo a formar repetições. Assim, nos delineamentos em blocos incompletos balanceados do tipo II, ocorre uma modificação em relação ao delineamento básico: os blocos incompletos podem ser organizados de modo a formar repetições completas. Isso significa que, embora cada bloco isoladamente contenha apenas uma parte dos tratamentos, é possível combinar vários blocos para incluir todos os tratamentos pelo menos uma vez, formando uma repetição completa do experimento.

Os delineamentos em latice quadrado são utilizados quando o número de tratamentos é elevado. Os latices quadrado permitem dispor um número de $v = k^2$ de tratamentos em b blocos e de k parcelas. O número de tratamentos deve ser um quadrado perfeito como exemplo, 25, 36, 64, 81,

100. Nos latices quadrados os tratamentos de um bloco em uma repetição se distribuem por todos os blocos de qualquer das outras repetições.

Assim, os blocos incompletos balanceados e os látices quadrados são delineamentos experimentais amplamente utilizados em situações em que o número de tratamentos é grande e a casualização completa se torna inviável ou menos eficiente. Ambos os métodos buscam minimizar os efeitos da heterogeneidade ambiental, proporcionando comparações mais precisas entre os tratamentos. Esses delineamentos são amplamente aplicados em experimentos agrícolas e genéticos, especialmente na avaliação de variedades e genótipos, onde o controle da variabilidade ambiental é essencial para garantir resultados confiáveis e estatisticamente robustos. Portanto, no estudo poderiam ter sido utilizados os blocos incompletos ou latice quadrados para avaliação dos genótipos de milho. Isso permitiria o controle da heterogeneidade dos blocos, promovendo uma melhor precisão experimental.

14.1 Referências

ABU, P. *et al*. Genetic diversity and inter-trait relationship of tropical extra-early maturing quality protein maize inbred lines under low soil nitrogen stress. **Plos One**, v. 16, n. 6, p. e0252506, 2021.

BERTOLINI, E. *et al*. Regulatory variation controlling architectural pleiotropy in maize. **bioRxiv**, p. 08-19, 2023.

ERTIRO, B. T. *et al*. Combining ability and testcross performance of drought-tolerant maize inbred lines under stress and non-stress environments in Kenya. **Plant Breeding**, v. 136, n. 2, p. 197-2005, 2017.

LIMA, D. C. *et al*. Utility of anthesis–silking interval information to predict grain yield under water and nitrogen limited conditions. **Crop Science**, v. 63, n. 1, p. 151-163, 2023.

NDLOVU, N. *et al*. Genome-wide association studies of grain yield and quality traits under optimum and low-nitrogen stress in tropical maize (*Zea mays* L.). **Theoretical and Applied Genetics**, p. 1-20, 2022.

PARAJULI, S. *et al*. Quantification of secondary traits for drought and low nitrogen stress tolerance in inbreds and hybrids of maize (*Zea mays* L.). **Journal of Plant Breeding and Genetics**, v.2, p.106–118, 2018.

PIMENTEL GOMES, F. **Estatística Experimental**. 10ª Edição, 1982.

RAMALHO, M. A. P. *et al*. **Genética na Agropecuária**. 5ª Edição Revisada - 2012.

ZAIDI, P. H. *et al*. Genomic regions associated with salinity stress tolerance in tropical maize (*Zea Mays* L.). **Frontiers in Plant Science**, v. 13, 2022.

QUESTÕES PROFESSOR DR. MAICON NARDINO

Questão 01. Explique do ponto de vista genético, como as bases genéticas de milho podem impactar a magnitude das correlações fenotípicas, genéticas e ambientais. Pode ser desenvolvida para exemplificar a questão análises com dados reais ou simulados. Resposta: **15. BASES GENÉTICAS E AS CORELAÇÕES FENOTÍPICAS, GENÉTICAS E AMBIENTAIS**

Questão 02. Os elementos climáticos podem influenciar a performance dos híbridos de milho durante o ciclo de desenvolvimento, dependendo da intensidade. Desenvolva uma estratégia para um programa de melhoramento visando obtenção de híbridos com elevado valor nutricional e indique quais fontes de variabilidade seriam utilizadas. Resposta: **16. MELHORAMENTO DE MILHO PARA QUALIDADE NUTRICIONAL**

15 BASES GENÉTICAS E AS CORELAÇÕES FENOTÍPICAS, GENÉTICAS E AMBIENTAIS

As bases genéticas de milho são distintas em termos de variabilidade genética (LIMA; BORÉM, 2018). A média da produtividade de grãos dos híbridos simples geralmente é maior do que a produtividade dos híbridos triplo e duplo, como relatado por Emygdio et al. (2007). Já as variedades caracterizam-se por exibir maior estabilidade fenotípica diante das variações do ambiente devido à sua ampla base genética, conforme observado por Carpentieri-Pípolo et al. (2010).

Geneticamente, as bases genéticas de milho apresentam um grau específico de homogeneidade ou heterogeneidade genética, que pode determinar a forma como os caracteres se relacionam. Os híbridos simples, obtidos do cruzamento entre duas linhagens puras, apresentam alta uniformidade genética entre as plantas e aproveitam ao máximo a heterose. Essa homogeneidade tende a aumentar as correlações genéticas entre os caracteres, pois os genes compartilhados entre os indivíduos do híbrido influenciam de maneira consistente os caracteres estudados. Em ambientes controlados, também é comum observar altas correlações fenotípicas, já que a influência ambiental é minimizada.

Os híbridos triplos, desenvolvidos pelo cruzamento de um híbrido simples com uma terceira linhagem pura, combinam um equilíbrio entre variabilidade genética e vigor híbrido. Isso resulta em correlações genéticas e fenotípicas que, embora ainda elevadas, tendem a ser menores do que as observadas em híbridos simples. Os híbridos duplos, gerados pelo cruzamento entre dois híbridos simples, apresentam maior heterogeneidade genética devido à combinação de quatro linhagens parentais. Essa maior variabilidade genética em teoria pode reduzir as correlações genéticas, pois diferentes combinações alélicas podem impactar os caracteres de maneiras distintas. Como resultado,

as correlações fenotípicas também tendem a ser menores, especialmente em condições ambientais mais variáveis.

Já as variedades de polinização aberta possuem a maior heterogeneidade genética entre os tipos de materiais. Essa diversidade genética pode promover correlações genéticas mais baixas, uma vez que a ampla variabilidade pode reduzir a consistência na expressão dos caracteres. As correlações fenotípicas, por sua vez, tornam-se mais influenciadas pelo ambiente, podendo variar amplamente dependendo das condições em que as variedades são avaliadas. Portanto, do ponto de vista genético, as correlações fenotípicas refletem uma combinação dos efeitos genéticos e ambientais sobre os caracteres. Em materiais mais homogêneos, como os híbridos simples, as correlações genéticas tendem a ter um peso maior, enquanto nos materiais heterogêneos, como as VPAS, a influência ambiental é mais acentuada.

16. MELHORAMENTO DE MILHO PARA QUALIDADE NUTRICIONAL

16.1 Objetivo aumentar os teores de amilose e amilopectina

O avanço das pesquisas sobre a composição nutricional dos grãos foi determinante para evidenciar que os caracteres produtivos devem ser associados com qualidade mercadológica dos produtos resultantes. No caso do milho, majoritariamente destinado à produção escalonada de proteína animal, as pesquisas em melhoramento têm voltado esforços na melhoria de sua qualidade nutricional (LI et al., 2018; ZHONG et al., 2022), de modo a possibilitar sua maior eficiência de conversão no crescimento e terminação dos animais. Devido ao seu elevado teor em amido (70 a 75%), associado a baixos valores de fatores antinutricionais intrínsecos, o milho é largamente usado como ingrediente energético nas rações (BUTTS-WILMSMEYER et al., 2019; RODRÍGUEZ et al., 2020). Outros nutrientes igualmente importantes, como proteínas (~10%) (JAHANGIRLOU et al., 2022), aminoácidos, fibras, lipídios, vitaminas e minerais, também constituem os grãos deste cereal.

Ao longo dos anos foram desenvolvidas pesquisas associando atributos nutricionais deste cereal com o desempenho animal (LIU et al., 2014; MA et al., 2020; WANG et al., 2019). Contudo, a grande maioria dos estudos não abordam adequadamente as variantes do principal constituinte deste cereal - o amido - e seus efeitos sobre o metabolismo e desempenho dos animais. O amido é formado por amilose e amilopectina (BUTTS-WILMSMEYER et al., 2019). A amilose é um polissacarídeo pelo qual as plantas, preferencialmente, armazenam energia, em razão de sua cadeia de ligação linear que ocupa menor espaço, facilitando o armazenamento de amido nas plantas (TAIZ et al., 2017). Já a amilopectina apresenta uma cadeia ramificada sendo uma molécula que ocupa um maior espaço.

A natureza do amido (relação entre amilose e amilopectina) também altera a digestão do amido pelos animais. A amilopectina torna o amido mais solúvel o que promove uma digestão mais rápida

do amido, quando comparado a maior concentração de amilose (LIANG et al., 2023). Cadeias de amilopectina mais longas foram positivamente correlacionadas com a taxa de digestão (ZHONG et al., 2020). Os maiores valores de amilopectina promovem ganhos na conversão alimentar de animais monogástricos, quando inserida nas dietas. A maior concentração de amilopectina na dieta de suínos promoveu o maior peso médio diário em estádio de terminação (WANG et al., 2019). Resultados similares foram observados em aves, na qual a digestibilidade do amido aumentou quando a relação entre amilose e amilopectina diminui (MA et al., 2020). Em peixes, a menor relação entre amilose e amilopectina potencializou o ganho de peso (LIU et al., 2014). Esses estudos indicam que a menor relação entre a amilose e amilopectina promove os maiores ganhos de peso diário em animais monogástricos.

Compreender o processo digestivo dos amidos permite explorar a capacidade de desenvolver genótipos com altos teores de amilose ou amilopectina. Por exemplo, com base nos estudos citados, o desenvolvimento de genótipos com maior potencial de amilopectina nos grãos é benéfico para indústrias de rações para animais monogástricos (ZHONG et al., 2020). Enquanto genótipos com maior expressão de amilose podem ser preferíveis no segmento de alimentação humana e desenvolvimento de subprodutos como filmes de amido (ZHAI et al., 2021).

16.1.1 Síntese de amilose e amilopectina em grãos de milho

A cultura do milho apresenta uma variedade de genes mutantes que codificam proteínas e enzimas responsáveis pela produção de amidos especiais, tais como o amido ceroso (com ausência de amilose), o amido com alta proporção de amilose, e os mutantes açucarados (BROWN et al., 2015; FUJITA et al., 2007; TALUKDER et al., 2022a; YANG et al., 2013). Identificar os genes e alelos que determinam o desempenho produtivo, a concentração de amido e qualidade do amido é essencial para promover avanços na agricultura e indústria.

O milho ceroso (sem amilose) é uma variação do amido do milho normal e foi primeiramente encontrado na China, em 1908. Essa característica é controlada geneticamente por um único gene recessivo, o gene *waxy* (*wx*) (LI et al., 2018). Já a presença do gene mutante *amylose extender* (*ae*), na forma homozigota, aumenta significativamente o conteúdo de amilose no endosperma (>50%) (ZHONG et al., 2022).

Nas últimas décadas, houve avanços significativos na compreensão da genética e biossíntese do amido, incluindo a identificação das principais enzimas envolvidas. O processo da síntese de amido no endosperma do milho, inicia-se com a conversão da sacarose, originada na fotossíntese, em frutose e UDP-glicose por meio da enzima sacarose sintase (LI et al., 2018). Posteriormente, a frutose e UDP-glicose são transformadas em ADP-glicose pela ação da enzima AGPase. Tanto a amilose

quanto a amilopectina utilizam o ADP-glicose como doador de glicosil ativado para a síntese, porém, são produzidas por diferentes enzimas (DONG et al., 2019).

A amilose é sintetizada pela enzima amido sintase ligada a grânulos (GBSSI), enquanto a amilopectina requer a coordenação de um complexo de enzimas, incluindo as amido solúvel sintases, a enzima ramificadora de amido e a enzima desramificadora de amido, conforme indicado por Li et al. (2018). O gene recessivo *amylose extender* (*ae*) sintetiza uma enzima ramificadora de amido não funcional, o que promove o acúmulo de até 50% de amilose devido à menor produção de amilopectina (HAN et al., 2022). Já a presença do gene recessivo mutante (*waxy*) reduz a produção da enzima amido sintase ligada a grânulos, que produz amilose, resultando em teores de amilopectina no endosperma superior a 97% (DONG et al., 2019). Ambos os mutantes têm sido usados no melhoramento de milho para desenvolver linhagens e híbridos de milho com alto ou baixo teor de amilose para alterar as propriedades e a utilidade do amido.

O fenótipo de alta amilose no milho é controlado pelo gene *amylose extender* (*ae*) na forma recessiva. A presença de *ae* pode ser facilmente identificada na maioria das fontes de germoplasma, pela expressão de uma característica de endosperma vítreo e manchado, que são conferidas pelos alelos mutantes *ae* (LIMA; BORÉM, 2018). Esse gene é predominantemente expresso no endosperma e embriões durante o desenvolvimento do grão (HAN et al., 2022). A compreensão das relações entre a biossíntese e a estrutura molecular do amido é essencial para que melhoristas possam desenvolver estratégias previsíveis de melhoramento de culturas de amido, atendendo às importantes demandas da agricultura relacionadas ao rendimento, desempenho e funcionalidade do amido.

16.1.2 Melhoramento para teor de amilose

Do conjunto de genes responsáveis pela síntese de amilose no endosperma dos grãos, a enzima sintase ligada aos grânulos é umas das mais importantes, uma vez que é o único gene envolvido na produção de amilose (BROWN et al., 2015). Essa enzima é expressa por um único gene dominante, responsável pela síntese da amilose. Inicialmente o gene GBSSI expresso na forma dominante e recessivo sintetizavam amilose. No entanto, uma mutação do gene recessivo (SPRAGUE, 1939) foi originalmente identificada na China, sendo essa mutação ocorrida antes de 1760 (ZENG, 1987).

Logo, as plantas sem uma cópia funcional do GBSSI produzem grãos com amido formado majoritariamente por amilopectina (>97%) (BROWN et al., 2015). A amilose é um caráter monogênico, ou seja, determinada por um único gene. Por isso, um dos métodos mais rápidos e eficientes para a obtenção de linhagens de milho ceroso (sem amilose) é o de retrocruzamento com sucessivas autofecundações (TALUKDER et al., 2022a; YANG et al., 2013). Pesquisas realizadas por Talukder et al. (2022b) e Yang et al. (2013) utilizam o retrocruzamento para inserir o gene *waxy* em populações recorrentes de milho. Além disso, realizaram cruzamentos entre linhagens elites com

linhagens *waxy* realizando sucessivas autofecundações das progênies para a expressão do gene recessivo e, assim, desenvolvimento de linhagens cerosas. Convencionalmente, a avaliação do fenótipo *waxy* pode ser realizada no momento da emissão do pendão com uso do teste de iodeto de potássio nos grãos de pólen (LIMA; BORÉM, 2018; TALUKDER et al., 2022a). Com base nesse teste, é possível identificar, pela coloração dos grãos de pólen, se o indivíduo tem o gene *waxy*.

Estudos utilizam uma solução com iodeto de potássio para identificar a presença do gene *waxy* em progênies de milho (TALUKDER et al., 2022a; YANG et al.; 2013). Em estudo realizado com Quinoa (BROWN et al., 2015), os autores relataram que todos os genótipos testados coraram roxo-azulado, indicando a presença de amilose, ou seja, dada a natureza dominante do gene GBSSI esses acessos possuíam pelo menos uma cópia funcional de GBSSI. Os autores identificaram que um dos genótipos tinha um alelo do gene mutante recessivo, apesar disso, o fenótipo era não ceroso, pois o gene dominante era funcional, consistente com a natureza dominante dos genes GBSSI.

O desenvolvimento e seleção de genótipos de milho ceroso consiste inicialmente na obtenção de uma linhagem mutante que tenha o gene recessivo *waxy*. Essa linhagem pode ser obtida de outros programas de melhoramento genético, bancos ativos de germoplasma ou realizando a mutação por meio de engenharia genética, conforme Qi et al. (2020). Após a obtenção da linhagem mutante o próximo passo é inserir o gene *waxy* na linhagem elite, de alto desempenho produtivo. Assim, realiza-se o cruzamento entre as duas linhagens, obtendo-se os indivíduos F_1.

Ao realizar cruzamentos de uma linhagem homozigota *waxy* (gene recessivo) com uma progênie homozigota sem o gene *waxy* (dominante), todas as progênies oriundas do cruzamento possuem o fenótipo não *waxy*, uma vez que os genótipos resultantes do cruzamento são híbridos F_1 heterozigotos (RAMALHO et al., 2012). Esses indivíduos F_1 apresentam o alelo do amido ceroso, no entanto não irá se expressar em razão do caráter ceroso ser homozigoto recessivo (*wxwx*).

As plantas da geração F_1, oriundas do cruzamento de dois homozigotos são todas heterozigotas, o que impede a expressão do gene recessivo. Isso indica, que realizando esse modelo de cruzamento não é possível desenvolver híbridos F_1 com o fenótipo ceroso. Logo, a partir do cruzamento de uma linhagem elite com uma linhagem *waxy* é possível utilizar duas estratégias de melhoramento a partir da obtenção dos indivíduos F_1: realizar autofecundações dos indivíduos F_1 para obter linhagens com o gene *waxy* ou proceder com retrocruzamentos para recuperar a constituição genética da linhagem elite com o gene *waxy*.

Autofecundar ou realizar um novo ciclo de cruzamento entre as plantas F_1s permite gerar sementes que tenham segregação para o amido ceroso (YANG et al., 2013). A autofecundação é a forma mais comum utilizada para a obtenção de linhagens. A planta é selecionada e antes do aparecimento do estigma a espiga é protegida. Após a emissão dos estigmas a espiga é autofecundada, ou seja, o pólen da própria planta poliniza o estigma da espiga. A autofecundação leva a homozigose

das linhagens, mas essencialmente só conduz ao melhoramento se algum processo de seleção for aplicado. Se uma amostra aleatória de linhagens endogâmicas for obtida de uma população heterogênea, os híbridos simples entre elas representam as combinações genotípicas que ocorrem na população de onde foram extraídas. Portanto, para aumentar a probabilidade de obtenção de híbridos superiores é necessário aumentar a probabilidade de genótipos superiores na população. Para isso é existem vários métodos que podem ser utilizados para a obtenção de linhagens.

Nesta situação, em razão do principal objetivo ser o desenvolvimento de híbridos de maior valor nutricional (*waxy*) e o gene ser recessivo, será utilizado o método padrão para conduzir as plantas a homozigose. As progênies S_1 oriundas da autofecundação das progênies F_1 que manifestar o gene *waxy* podem ser selecionadas. Uma vez identificadas, o gene *waxy* já estará fixo, assim, o aumento da homozigose por meio da autofecundação busca reduzir a carga genética das progênies selecionadas. A segregação esperada é de três para um, sendo três partes com fenótipos sem *waxy*, portanto, não desejados, e apenas uma parte, 25%, com fenótipo *waxy* (RAMALHO et al., 2012). Assim, essas sementes são semeadas e, convencionalmente, pode-se identificar os genótipos com o gene *waxy* a partir do florescimento masculino com o teste de iodeto de potássio. No entanto, essa estratégia resulta em linhagens homozigotas com o gene *waxy*, mas com elevada carga genética. Geralmente a linhagem doadora do gene *waxy* apresenta uma carga genética associada, isso reduz significativamente o potencial produtivo das linhagens devido a manifestação de genes deletérios. Assim, a cada autofecundação é necessário a seleção de plantas com gene *waxy* e com redução da carga genética.

Inicialmente autofecunda-se centenas de planta da população F_1. Cerca de 20 a 30 sementes de cada espiga de autofecundação (S1) serão semeadas em fileiras. Realiza-se a seleção das entre e dentro de progênies, autofecundando 2 a 5 plantas das progênies selecionadas. Escolher de 1 a 3 espigas das progênies autofecundadas. Semear as sementes S2, autofecundar as plantas e selecionar, repetindo esse processo até S6 ou S8. Com a obtenção das linhagens *waxy* com remoção dos genes deletérios, o objetivo será testar a capacidade específica de combinação entre as linhagens para desenvolver híbridos simples com o gene *waxy*. Para que o híbrido simples tenha o gene *waxy* é necessário que o cruzamento seja realizado com duas linhagens que tenham o gene recessivo *waxy*. Uma vez que o cruzamento entre linhagem com gene dominante (*WAXY*) e uma linhagem com o gene recessivo (*waxy*) é realizado, o híbrido F1 não manifestará o gene recessivo. Portanto, por meio da introdução do gene *waxy* à população do programa de melhoramento genético, é possível desenvolver linhagens com remoção de genes deletérios para o desenvolvimento de híbridos F_1 de alta qualidade nutricional.

A estratégia por meio do retrocruzamento consiste em inserir o gene de interesse, como o caráter *waxy*, em um material genético específico (linhagem elite), por meio do cruzamento com a

fonte do gene desejado, seguidos por cruzamentos sucessivos com o material genético superior parental (linhagem elite) (BORÉM; MIRANDA; FRITSCHE-NETO, 2021). O objetivo do método é recuperar o genótipo do genitor recorrente, exceto para uma ou poucas características que o melhorista procura transferir a partir do genitor doador (LIMA; BORÉM, 2018). Assim, são realizados sucessivos retrocruzamentos dos indivíduos F_1, oriundos do cruzamento entre uma linhagem elite (genitor recorrente) com uma linhagem mutante *waxy*, com o genitor recorrente (linhagem elite).

Após são realizadas sucessivas autofecundações para identificar as progênies *waxy* e fixar os demais caracteres de interesse. Não há um número específico de retrocruzamentos a serem realizados. No entanto, alguns estudos utilizaram dois retrocruzamentos (TALUKDER et al., 2022; QI et al., 2020; YANG et al., 2013), obtendo uma recuperação do material genético da linhagem recorrente superior a 85% (QI et al., 2020). Após esse processo de retrocruzamentos e autofecundações tem-se a recuperação da linhagem elite, mas com o incremento do gene *waxy*, ou seja, uma linhagem de alto desempenho produtivo associado com a produção de amido ceroso.

O cruzamento entre uma linhagem elite e uma linhagem mutante *waxy* tem por objetivo introduzir um gene monogênico na linhagem elite ou desenvolver novas linhagens com o gene de interesse. No entanto, como o gene da característica cerosa é recessivo, não é possível desenvolver híbridos F_1 a partir do cruzamento de uma linhagem elite e uma linhagem mutante *waxy*, visto que o híbrido F_1 é heterozigotos dominante.

O desenvolvimento de híbridos F_1 *waxy* só é possível realizando o cruzamento entre duas linhagens *waxy*, conforme realizado por Talukder et al. (2022). Uma linhagem com gene recessivo *wx* ao cruzar com outra linhagem com gene recessivo *wx*, resulta em um híbrido F_1 com gene homozigoto *wxwx*. Portanto, em um programa de melhoramento genético para obtenção de genótipos superiores para amilopectina as estratégias são baseadas em cruzamentos entre linhagens, retrocruzamento e autofecundações. Para cada caso especial descrito acima, utiliza-se determinadas técnicas associadas umas às outras.

16.2 Objetivo aumentar a qualidade nutricional proteica

As propriedades familiares produzem e utilizam os grãos de milho diretamente na alimentação de suínos e aves, sem depender de rações comerciais para otimizar os recursos disponíveis. Entretanto, os grãos de milho apresentam baixos teores de aminoácidos essenciais como metionina (0,16 g/100 g), lisina (0,28 g/100 g) e triptofano (0,06 g/100 g) (SIMÕES et al., 2023; VIDAL et al., 2024). Animais monogástricos não conseguem sintetizar esses aminoácidos (essenciais), sendo a única via de ingestão externa através de alimentos. Logo, o melhoramento genético surge como uma

das principais estratégias para desenvolver genótipos de milho de alto desempenho agronômico e melhor qualidade nutricional dos grãos.

Pesquisas realizadas na região de Santa Maria, RS, revelam a existência da variabilidade genética entre híbridos e variedades de milho quanto à qualidade nutricional dos grãos (SIMÕES et al., 2023; VIDAL et al., 2024). Esses resultados indicam a viabilidade em selecionar genótipos superiores. O melhoramento genético de milho é essencial para enfrentar os desafios dos produtores, que utilizam os grãos na alimentação animal. A seleção de genótipos que combinem alto desempenho produtivo e qualidade nutricional promove ganhos econômicos, segurança na produção e maior sustentabilidade das propriedades. Variedades acessíveis, como as de polinização aberta e híbridos, podem atender à demanda local, ao oferecer alimentos de alta qualidade para os animais, aumentar a produtividade agrícola e pecuária. Esses avanços fortalecem a permanência dos agricultores no campo e a sucessão familiar, além de trazer inovação ao setor agrícola. Assim, o objetivo será desenvolver, avaliar e selecionar híbridos intervarietais, populações melhoradas e progênies parcialmente endogâmicas de milho com alto desempenho agronômico e melhor qualidade nutricional dos grãos.

Para isso, serão conduzidas três estratégias distintas para o desenvolvimento de materiais genéticos de milho, com foco no desempenho agronômico e qualidade nutricional dos grãos. Cada uma dessas estratégias apresentará diferentes abordagens de melhoramento genético, visando a obtenção de híbridos intervarietais, a seleção de progênies de meios-irmãos para obtenção de população melhorada e o desenvolvimento de linhagens endogâmicas para obtenção de híbridos. A estratégia geral envolve a utilização de diferentes técnicas de melhoramento genético, utilizando cruzamentos entre genitores de variedades de polinização aberta, bem como métodos de seleção recorrente e autofecundação. Essas estratégias permitirão o desenvolvimento de novos materiais genéticos, tanto para o uso direto em programas de melhoramento quanto para o desenvolvimento de novos híbridos simples e híbridos intervarietais.

A população inicial será formada por variedades de polinização aberta de milho avaliadas na região de Santa Maria (LORO et al., 2024a, LORO et al., 2024b; LUZ et al., 2014). As variedades serão semeadas na área do Departamento de Fitotecnia da Universidade Federal de Santa Maria, RS, Brasil, na safra de 2025/2026. As variedades serão semeadas em fileiras individuais espaçadas em 0,80 metros e 0,20 m entre plantas. Será utilizada a adubação com 250 kg ha^{-1} de adubo químico da fórmula (NPK) 05-20-20. A adubação nitrogenada com ureia (N - 46%) será fracionada, sendo a primeira aplicação de 100 kg ha^{-1} no estádio V4 e a segunda de 75 kg ha^{-1} no estádio V6 da cultura. Os demais manejos culturais, como controle de plantas daninhas, pragas e doenças, serão realizados de acordo com as indicações técnicas para a cultura de milho. A partir desta população, será utilizada duas estratégias de desenvolvimento de materiais genéticos de milho: desenvolvimento de híbridos intervarietais, seleção recorrente entre famílias de meios-irmãos para extração de linhagens.

16.2.1 Obtenção e avaliação dos híbridos intervarietais

Essa estratégia será realizada em duas etapas, sendo a primeira a realização dos cruzamentos entre as variedades em um dialelo completo sem recíproco, para obtenção dos híbridos intervarietais e a segunda etapa será a avaliação híbridos intervarietais. Os cruzamentos entre as variedades serão realizados manualmente da seguinte forma: nas fileiras femininas e, antes da emissão dos estigmas, a espiga será protegida com uma pequena embalagem plástica. Nas plantas masculinas, o pendão será protegido 24 horas antes de realizar a polinização das espigas, com uma embalagem de papel encerado. Os cruzamentos serão realizados manualmente com a coleta e mistura do pólen de no mínimo 30 plantas de cada variedade para a polinização de 30 espigas da outra variedade receptora de pólen, visando a representatividade de cada genitor para obtenção dos híbridos intervarietais.

Os híbridos intervarietais, juntamente com os genitores e dois genótipos comerciais (testemunhas) serão avaliados em duas datas de semeadura (outubro e novembro), na região de Santa Maria (exemplo de um programa de melhoramento na UFSM). Em cada ensaio será utilizado o delineamento experimental de blocos casualizados com quatro repetições, com parcelas constituídas por 3 fileiras de 5 metros, espaçadas em 0,80 m. Nesses híbridos, serão avaliados os seguintes caracteres fenológicos, morfológicos e produtivos. Uma amostra de 100 g de grãos de cada genótipo será retirada para avaliar os caracteres nutricionais proteicos em g/100 g de matéria seca: proteína bruta (CP, g/100 g), lisina (LYS, g/100 g), metionina (MET, g/100 g) e triptofano (TRP, g/100 g).

16.2.2 Seleção recorrente entre e dentro de meios irmãos

Na população inicial, formada pelas variedades, serão selecionadas 200 plantas, com base no fenótipo. Inicialmente, as plantas serão selecionadas de acordo com a altura de planta, posição relativa da espiga e qualidade de espigas. As espigas das plantas selecionadas serão colhidas individualmente e as sementes armazenadas em câmara fria. Na próxima safra, a metade das sementes de cada espiga (família de meios-irmãos) serão semeadas em experimentos com repetições, semeadas em fileiras individuais de 5 m, em látice quadrado, para avaliar o desempenho agronômico e a qualidade nutricional dos grãos para seleção entre progênies de meios-irmãos.

Na próxima safra, as sementes remanescentes correspondentes às progênies de meios-irmãos selecionadas no ensaio com repetições, serão recombinadas utilizando o método irlandês para completar o primeiro ciclo de seleção recorrente (C_1). No campo de recombinação, será realizada a seleção dentro de progênies de meios-irmãos para obter as progênies para iniciar o segundo ciclo de seleção recorrente. A seleção no bloco de recombinação será baseada na altura de planta, posição relativa da espiga e qualidade de espigas. Os ciclos de seleção recorrente serão realizados

continuamente, com a avaliação de cada população melhorada, a fim de aumentar a proporção de alelos favoráveis na população.

A partir da população melhorada, podem ser extraídas linhagens. Essas linhagens podem ser obtidas pelo método padrão ou da cova única, dependendo da mão de obra e área disponível. Essas linhagens poderão ser avaliadas quanto a capacidade geral de combinação com cruzamentos *top crosses* e por fim quanto a capacidade especifica de combinação para o desenvolvimento de híbridos simples com alto desempenho agronômico e qualidade nutricional dos grãos.

16.3 Referências

BORÉM, A.; MIRANDA, G. V.; FRITSCHE-NETO, R. **Melhoramento de plantas**. 8. ed. 2021, 453p.

BROWN, D. C. *et al*. Characterization of the Granule-Bound Starch Synthase I gene in Chenopodium. **The Plant Genome**, v. 8, n. 1, p. plantgenome2014.09.0051, 2015.

BUTTS-WILMSMEYER, C. J. *et al*. Weather during key growth stages explains grain quality and yield of maize. **Agronomy**, v. 9, n. 1, p. 16, 2019.

DONG, L. *et al*. Supersweet and waxy: meeting the diverse demands for specialty maize by genome editing. **Plant Biotechnology Journal**, v. 17, n. 10, p. 1853, 2019.

CARPENTIERI-PÍPOLO, C. *et al*. Avaliação de cultivares de milho crioulo em sistema de baixo nível tecnológico. Acta Scientiarum. Agronomy, v. 32, n. 2, p. 229-233, 2010

EMYGDIO, B. M; IGNACZAK, J. C; CARGNELUTTI FILHO, A. Potencial de rendimento de grãos de híbridos comerciais simples, triplos e duplos de milho. Revista Brasileira de Milho e Sorgo, v. 6, n. 1, p. 95-103, 2007.

FUJITA, N. *et al*. Characterization of SSIIIa-deficient mutants of rice: The function of SSIIIa and pleiotropic effects by SSIIIa deficiency in the rice endosperm. **Plant Physiol**. v. 144, p. 2009-2023, 2007.

LUZ, G. R. *et al*. Susceptibility of corn hybrids to corn stunt complex transmitted by the corn leafhopper, *Dalbulus maidis* (DeLong & Wolcott, 1923) in Brazil. **Observatório De La Economía Latinoamericana**, v. 22, n. 7, p. e5584-e5584, 2024.

HAN, J. *et al*. Using the dominant mutation gene Ae1-5180 (amylose extender) to develop high-amylose maize. **Molecular Breeding**, v. 42, n. 10, p. 57, 2022.

JAHANGIRLOU, M. R. *et al*. Phenotypic predictors of dent maize grain quality based on different genetics and management practices. **Journal of Cereal Science**, v. 103, p. 103388, 2022.

LI, C. *et al*. The genetic architecture of amylose biosynthesis in maize kernel. **Plant Biotechnology Journal**, v. 16, n. 2, p. 688-695, 2018.

LIANG, W. *et al*. The relationship between starch structure and digestibility by time-course digestion of amylopectin-only and amylose-only barley starches. **Food Hydrocolloids**, v. 139, p. 108491, 2023.

LIMA, R.; BORÉM, A. **Melhoramento de milho**. 1ª ed., Viçosa: Editora UFV, 2018. 396p.

LIU, X. H. *et al*. Effects of dietary amylose/amylopectin ratio on growth performance, feed utilization, digestive enzymes, and postprandial metabolic responses in juvenile obscure puffer Takifugu obscurus. **Fish physiology and biochemistry**, v. 40, p. 1423-1436, 2014.

LORO, M. V. *et al*. Relações lineares entre variáveis meteorológicas e caracteres fenológicos, morfológicos e produtivos em bases genéticas de milho. **Revista Vivências**, v. 41, p. 95-111, 2023a.

LORO, M. V. *et al*. Relações lineares entre caracteres do pendão e da espiga em bases genéticas de milho. **Journal of Environmental Analysis and Progress**, v. 9, n. 2, p. 065-078, 2024.

MA, J. *et al*. Effects of dietary amylose/amylopectin ratio and amylase on growth performance, energy and starch digestibility, and digestive enzymes in broilers. **Journal of Animal Physiology and Animal Nutrition**, v. 104, n. 3, p. 928-935, 2020.

QI, X. *et al*. Conversion of a normal maize hybrid into a *waxy* version using in vivo CRISPR/Cas9 targeted mutation activity. **The Crop Journal**, v. 8, n. 3, p. 440-448, 2020.

RAMALHO, M. A. P. *et al*. **Genética na Agropecuária**. 5. ed., 2012, 566p

RODRIGUEZ, D. A. *et al*. Digestibility of amino acids, fiber, and energy by growing pigs, and concentrations of digestible and metabolizable energy in yellow dent corn, hard red winter wheat, and sorghum may be influenced by extrusion. **Animal Feed Science and Technology**, v. 268, n. 114602, p. 1-11, 2020.

SPRAGUE, G. F. An estimation of the number of the top crossed plants required for adequate representation of a corn variety. **American Society of Agronomy**, v. 31, p. 11-16, 1939.

TAIZ, L. *et al*. **Fisiologia e Desenvolvimento Vegetal**. 6ª ed., Porto Alegre: Artmed, 2017. 888p.

TALUKDER, Z. A. *et al*. Combining higher accumulation of amylopectin, lysine and tryptophan in maize hybrids through genomics-assisted stacking of waxy1 and opaque2 genes. **Scientific Reports**, v. 12, n. 1, p. 706, 2022a.

WANG, H. *et al*. Effects of dietary amylose and amylopectin ratio on growth performance, meat quality, postmortem glycolysis and muscle fibre type transformation of finishing pigs. **Archives of Animal Nutrition**, v. 73, n. 3, p. 194-207, 2019.

YANG, L. *et al*. Marker-assisted selection for pyramiding the waxy and opaque-16 genes in maize using cross and backcross schemes. **Molecular Breeding**, v. 31, p. 767-775, 2013.

ZENG, M. Q. The relationship of waxy maize in China. **Crop Breed Resource**, v. 6, n. 3, p. 1-8, 1987.

ZHAI, X. *et al*. Cationized high amylose maize starch films reinforced with borax cross-linked nanocellulose. **International Journal of Biological Macromolecules**, v. 193, p. 1421-1429, 2021.

ZHONG, Y. *et al*. Amylose content and specific fine structures affect lamellar structure and digestibility of maize starches. **Food Hydrocolloids**, v. 108, n. 105994, p. 1-9, 2020.

QUESTÕES PROFESSOR DR. IVAN RICARDO CARVALHO

Questão 01. Demonstre um fluxograma de tomada de decisão e o Script em R para a construção de um data.frame de dados com co-variáveis ambientais, comprovando a qualidade dos dados e seu poder informativo. Utilize a análise de fatores como proposta para reducionalidade das covariáveis ambientais, proponha uma rotina estatística em R para isso. Empregue a regressão fatorial, utilizando as covariáveis ambientais (variáveis puras: Cenário 1) e utilizando as cargas fatoriais (variáveis reduzidas: Cenário 2), proponha a rotina estatística em R. **Resposta: 17. REGRESSÃO FATORIAL**

17. REGRESSÃO FATORIAL

A regressão fatorial é técnica estatística que consiste em modelar a relação entre uma variável dependente e fatores latentes. Esses fatores são variáveis que não podem ser medidas diretamente, mas são inferidas a partir de variáveis observadas. O objetivo é identificar como esses fatores latentes determinam a variável de interesse, simplificando a análise ao agrupar variáveis correlacionadas.

Por exemplo, ao estudar o teor de valina nos grãos de milho, podem ser considerados dois fatores latentes: condições meteorológicas no estádio vegetativo e no estádio reprodutivo. Esses fatores não são diretamente mensuráveis, mas podem ser representados por variáveis observáveis, como radiação solar global e precipitação acumulada no estádio vegetativo (para as condições meteorológicas no estádio vegetativo) e radiação solar global e soma térmica no estádio reprodutivo (para as condições meteorológicas no estádio reprodutivo). A regressão fatorial permite modelar o teor de valina como uma função desses fatores latentes, que, por sua vez, são expressos como combinações das variáveis observáveis.

No modelo, o teor de valina é representado como uma soma dos fatores latentes, onde os pesos indicam a contribuição de cada fator. Os fatores são formados pelas variáveis observadas, também ponderadas por coeficientes que indicam sua relevância. Após o ajuste do modelo, é possível interpretar os resultados para identificar a influência de cada fator latente sobre o teor de valina, bem como a contribuição das variáveis observadas para a formação desses fatores. Essa estratégia é importante em situações onde há alta correlação entre as variáveis independentes, pois os fatores latentes ajudam a reduzir a multicolinearidade e simplificam a interpretação.

17.1 Aplicação da regressão fatorial

Para compreender a aplicação da regressão fatorial será utilizado um exemplo do banco de dados "*dados-RFA*" que pode ser obtido pela plataforma GitHub, conforme script abaixo.

```r
#URL do arquivo no GitHub (link para o arquivo .xlsx raw)
url <- "https://github.com/muriloloro/Modelos-de-Predicao/raw/main/dados-RFA.xl
sx"
library(httr)
library(readxl)
# Definir o caminho temporário para salvar o arquivo
temp_file <- tempfile(fileext = ".xlsx")
GET(url, write_disk(temp_file, overwrite = TRUE))
dados <- read_excel(temp_file, sheet = "Covar")
```

Nesse banco de dados tem-se as seguintes variáveis meteorológicas de 78 genótipos de milho avaliados em 10 datas de semeadura: radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), soma térmica do florescimento feminino à colheita (SFFC, °C dia), precipitação pluviométrica acumulada da semeadura ao florescimento masculino (PSFM, mm), precipitação pluviométrica acumulada do florescimento masculino à colheita (PFMC, mm) e eficiência no uso da água (EUA, kg mm^{-1}).

A partir da matriz de dados, foram calculados os coeficientes de correlação linear de Pearson entre os pares de caracteres, com a significância avaliada pelo teste t de *Student* a 5% de probabilidade. A adequação da matriz de correlação para a análise fatorial foi verificada pelo teste de Kaiser-Meyer-Olkin (KMO), que indica se os fatores identificados na análise fatorial podem descrever adequadamente as variações nos dados originais. Também foi aplicado o teste de esfericidade de Bartlett, o qual testa se a matriz de correlação é uma matriz identidade, ou seja, se as variáveis não apresentam correlação significativa na população (hipótese nula).

A análise fatorial foi aplicada, sendo que as cargas dos fatores foram estimadas pelo método de componentes principal. O número de fatores a serem retidos foi determinado pelo critério de Kaiser-Guttman (GUTTMAN; 1954; KAISER, 1960), ou seja, são retidos os fatores com autovalor maior que 1,0. Simultaneamente a esse critério, verificou-se a porcentagem de explicação acumulada dos fatores, tendo por objetivo reter fatores de modo que a explicação fosse superior a 60% (HAIR et al., 2009).

A escolha do tipo de rotação de fatores a ser realizada considerou a possibilidade de verificar as relações entre os fatores. A utilização de rotação ortogonal promove a independência entre os fatores, algo que em ciências agrárias é difícil de ocorrer. Já a rotação oblíqua mantém as relações entre os fatores, o que possibilita verificar o sentido das relações entre os fatores retidos. De acordo com Hair et al. (2009), em geral, as duas formas de rotação produzem resultados similares. Logo, a rotação *oblimin* (obliqua) foi utilizada, uma vez que é esperado uma relação de dependência entre os

dados agronômicos, pois fatores como as condições meteorológicas frequentemente estão correlacionadas.

Os escores dos fatores rotacionados, para cada genótipo, foram estimados pelo método de regressão. Em seguida, aplicou-se a análise de regressão linear múltipla com validação cruzada *leave-one-out* considerando as covariáveis ambientais em dois cenários: Cenário 1: aplicou-se a regressão linear múltipla para predição da valina nos grãos por meio das variáveis originais (RSFM, RSFF, RFMC, RFFC, SFMC, SFFC e EUA) e Cenário 2: aplicou-se a regressão linear múltipla para predição da valina por meio dos escores dos três primeiros fatores que representam as variáveis originais.

Portanto, pode-se representar esses passos para aplicação da regressão fatorial por meio do fluxograma de tomada de decisão abaixo.

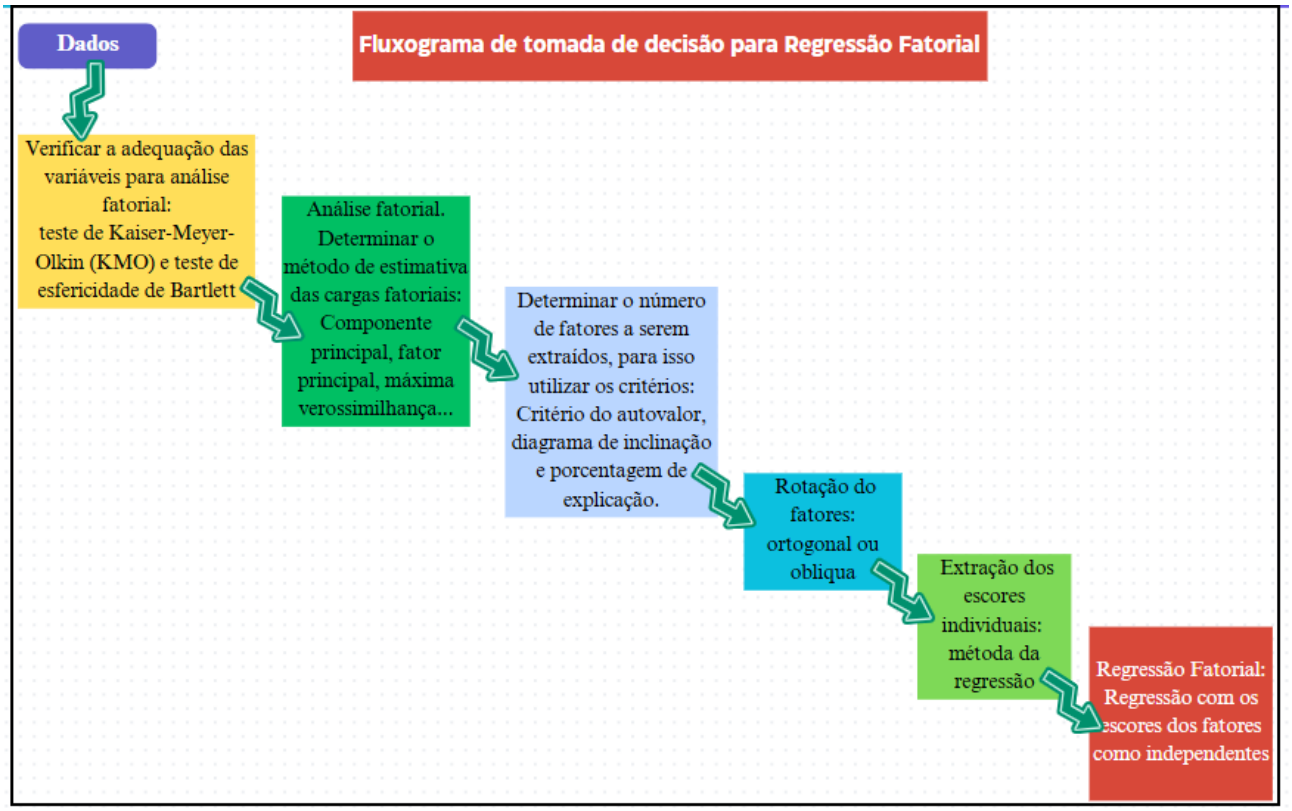

Figura 1. Fluxograma de tomada de decisão para aplicação da regressão fatorial.

17.2 Resultado da aplicação

Para a análise fatorial exploratória (AF), quanto maior o tamanho da amostra, mais adequada se torna para a realização de análises precisas. Segundo Hair et al. (2009), a amostra deve conter mais de 50 observações por caractere. Além disso, a relação entre o número de observações e a quantidade de caracteres deve ser superior a cinco para um, o que ajuda a fortalecer a confiabilidade das inferências feitas a partir dos dados. Neste estudo, a relação entre o número de observações e o número de caracteres foi de (773 observações/9 caracteres) 85,89 para 1, revelando a adequação do tamanho da amostra para a análise fatorial. O teste de Kaiser-Meyer-Olkin (KMO) apresentou um

valor de 0,63, indicando que os fatores identificados na análise fatorial conseguem descrever de forma satisfatória as variações dos dados originais. Segundo Hair et al. (2009), valores acima de 0,50 já são considerados aceitáveis, e um valor de 0,63 sugere que a estrutura de fatores é adequada para representar as relações entre os caracteres. O teste de esfericidade de Bartlett foi significativo ($p < 0,05$, qui-quadrado = 22.348,86), rejeitando a hipótese nula que a matriz de correlação é uma matriz identidade, ou seja, as variáveis não são correlacionadas na população. Logo, existe relação suficiente entre os caracteres para aplicação da análise fatorial (Figura 2).

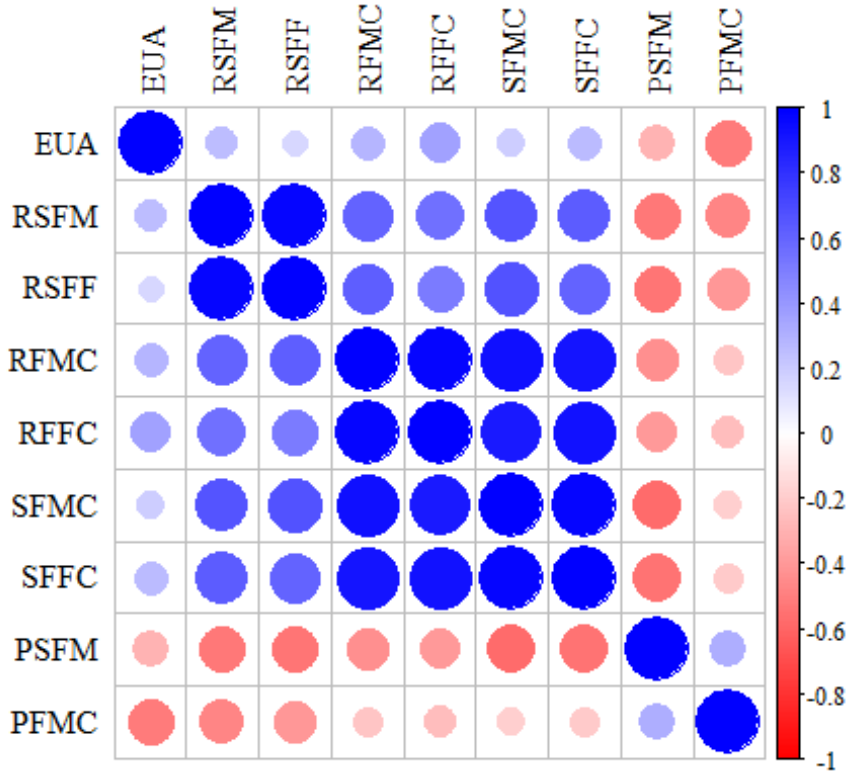

Figura 2. Matriz de coeficientes de correlação entre os caracteres fenológicos, morfológicos, produtivos, nutricionais e meteorológicos de 78 genótipos de milho avaliados em 10 datas de semeadura nas safras 2021/2022 e 2022/2023, Santa Maria, RS, Brasil. Variáveis: radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), soma térmica do florescimento feminino à colheita (SFFC, °C dia), precipitação pluviométrica acumulada da semeadura ao florescimento masculino (PSFM, mm), precipitação pluviométrica acumulada do florescimento masculino à colheita (PFMC, mm) e eficiência no uso da água (EUA, kg mm^{-1}).

Os três primeiros fatores apresentaram autovalores acima de 1,00 e, juntos, explicaram 87% da variabilidade total dos dados (Tabela 1 e Figura 3). Logo, de acordo com Hair et al. (2009) fatores com autovalores acima de 1,00 podem ser utilizados para a análise fatorial. Portanto, foram utilizados os três primeiros fatores para representar a variabilidade dos dados.

Tabela 1. Cargas fatoriais associadas aos fatores extraídos pelo método de componentes principais, com rotação *oblimin*, a partir de variáveis meteorológicas avaliadas em 78 genótipos de milho semeados em 10 datas de semeadura nas safras 2021/2022 e 2022/2023.

Variáveis	FA1	FA2	FA3	h²
EUA	0,21	-0,16	**0,91**	0,88
RSFM	0,14	**0,87**	0,04	0,93
RSFF	0,14	**0,91**	-0,07	0,94
RFMC	**0,95**	0,05	0,02	0,95
RFFC	**0,99**	-0,09	0,12	0,96
SFMC	**0,89**	0,20	-0,10	0,97
SFFC	**0,93**	0,09	-0,01	0,96
PSFM	-0,21	**-0,52**	-0,16	0,49
PFMC	0,22	-0,48	**-0,73**	0,79
Var (%)	44,00	27,00	16,00	-
Var Ac (%)	44,00	71,00	87,00	-

h²: comunalidade; FA1: fator 1, FA2: fator 2; FA3: fator 3; Var: variância explicada; Var Ac: variância explicada acumulada. Variáveis: radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), soma térmica do florescimento feminino à colheita (SFFC, °C dia), precipitação pluviométrica acumulada da semeadura ao florescimento masculino (PSFM, mm), precipitação pluviométrica acumulada do florescimento masculino à colheita (PFMC, mm) e eficiência no uso da água (EUA, kg mm^{-1}).

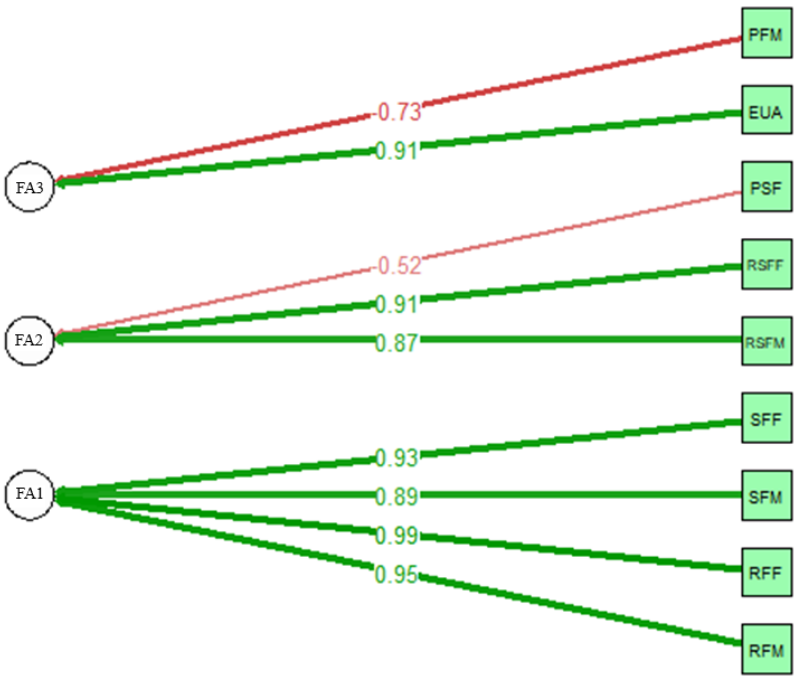

Figura 3. Cargas fatoriais das variáveis meteorológicas em cada fator. FA1: fator 1, FA2: fator 2; FA3: fator 3; Var: variância explicada; Var Ac: variância explicada acumulada. Variáveis: radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), soma térmica do florescimento feminino à colheita (SFFC, °C dia), precipitação pluviométrica acumulada da semeadura ao florescimento masculino (PSFM, mm), precipitação pluviométrica acumulada do florescimento masculino à colheita (PFMC, mm) e eficiência no uso da água (EUA, kg mm^{-1}).

As cargas fatoriais dos caracteres indicam sua contribuição para cada fator (Tabela 1 e Figura 3). O fator FA1 foi relacionado à radiação solar global e a soma térmica acumuladas no estádio reprodutivo dos genótipos, pois apresenta altas cargas positivas em variáveis como RFMC, RFFC, SFMC e SFFC. Assim, FA1 pode ser interpretado como uma dimensão ligada ao acúmulo de radiação e soma térmica no estádio reprodutivo. O FA2 captou a influência da radiação solar global e precipitação pluviométrica no estádio vegetativo, uma vez que as variáveis RSFM, RSFF e PSFM apresentaram maiores cargas fatoriais para esse fator. Já o FA3 associou-se à EUA e a PFMC.

Cada fator agrupou variáveis que possuem um sentido comum, facilitando a interpretação: FA1 foi o fator mais relevante para as condições meteorológicas no estádio reprodutivo do milho, FA2 para condições meteorológicas no estádio vegetativo e FA3 para eficiência no uso da água e precipitação pluviométrico no estádio reprodutivo. Esses fatores podem ser usados para predizer variáveis dependentes, o que promove um modelo mais parcimonioso, devido a redução do número de variáveis. Os valores de comunalidade foram altos para a maioria dos caracteres, indicando que o modelo de análise fatorial explicou satisfatoriamente a variabilidade total dos caracteres.

A comparação entre os dois cenários, um baseado em variáveis originais e outro em fatores, permitiu avaliar o desempenho de diferentes abordagens para a predição da VAL nos grãos de milho (Tabela 2). No primeiro cenário, utilizou-se as variáveis originais, resultando em um R^2 ajustado de 0,51, o que indica que o modelo é capaz de explicar 51% da variabilidade da variável dependente. No entanto, o R^2 obtido na validação cruzada *leave-one-out* foi de apenas 0,007, o que indica uma baixa capacidade de generalização do modelo. Além disso, a raiz do erro quadrático médio (RMSE) na validação foi relativamente alto, com um valor de 0,2, reforçando que o modelo apresenta problemas de *overfitting*, ajustando-se bem aos dados de treino, mas sendo ineficaz na predição de novos dados.

Tabela 2. Modelos de predição da valina nos grãos de milho por meio das variáveis originais (cenário 1) e cargas fatoriais (cenário 2).

Cenários	Modelo	R^2_{ajus}	R^2_{val}	$RMSE_{val}$
Cenário 1	y = 0,3783 - 0,02325RSFM + 0,0234RSFF -0,0234RFMC + 0,0234 RFFC + 0,000098SFMC - 0,00015SFFC -3,47EUA - 0,00009PSFM - 0,00012PFMC	0,51	0,007	0,20
Cenário 2	y = 0,46 - 0,02FA1 - 0,008FA2 + 0,04FA3	0,49	0,48	0,04

R^2_{ajus}: coeficiente de determinação; R^2_{val}: coeficiente de determinação da validação cruzada; $RMSE_{val}$: raiz do erro quadrático médio da validação cruzada; Variáveis: radiação solar global da semeadura ao florescimento masculino (RSFM, MJ m^{-2}); radiação solar global da semeadura ao florescimento feminino (RSFF, MJ m^{-2}); radiação solar global do florescimento masculino à colheita (RFMC, MJ m^{-2}), radiação solar global do florescimento feminino à colheita (RFFC, MJ m^{-2}), soma térmica do florescimento masculino à colheita (SFMC, °C dia), soma térmica do florescimento feminino à colheita (SFFC, °C dia), precipitação pluviométrica acumulada da semeadura ao florescimento masculino (PSFM, mm), precipitação pluviométrica acumulada do florescimento masculino à colheita (PFMC, mm) e eficiência no uso da água (EUA, kg mm^{-1}). Cenário 1: aplicou-se a regressão linear múltipla para predição da valina nos grãos por meio das variáveis originais (RSFM, RSFF, RFMC, RFFC, SFMC, SFFC e EUA) e Cenário 2: aplicou-se a regressão linear múltipla para predição da valina por meio dos escores dos três primeiros fatores que representam as variáveis originais.

No segundo cenário, o modelo foi ajustado com base em fatores obtidos por análise fatorial, o que reduziu a dimensionalidade e simplificou a estrutura do modelo. Nesse caso, o R^2 ajustado foi de 0,49, inferior ao do modelo com variáveis originais. Contudo, o desempenho na validação cruzada foi melhor, com um R^2 de 0,48 e um *RMSE* de 0,04. Esses resultados indicam que o modelo baseado em fatores possui uma maior capacidade de generalização e uma melhor precisão preditiva, além de ser menos suscetível ao *overfitting*.

O modelo ajustado para a predição do teor de VAL nos grãos de milho utilizou três fatores obtidos por meio da análise fatorial. Os coeficientes indicaram a contribuição de cada fator para o teor de valina. Os fatores FA1 e FA2 apresentaram coeficientes negativos (-0.02 e 0,008, respectivamente), sugerindo que valores elevados deste fator estavam associados a menores teores de valina. Enquanto o fator FA3, com coeficiente positivo (+0.04), teve maior influência positiva na predição do teor de VAL. O modelo se mostrou parcimonioso ao reduzir a dimensionalidade dos dados de nove variáveis originais para três fatores principais, sem comprometer sua capacidade explicativa. Essa simplificação reduziu os riscos de multicolinearidade e tornou o modelo mais eficiente e interpretável. A análise indicou que o aumento no teor de VAL pode ser obtido ao otimizar a variável EUA em ambientes com menor PFMC, associadas a FA3.

Portanto, o uso de fatores permitiu consolidar informações das variáveis originais em um número reduzido de fatores, eliminando redundâncias e o impacto da multicolinearidade. Essa simplificação resultou em um modelo mais robusto, que mantém um desempenho consistente entre o treinamento e a validação cruzada. Apesar da ligeira redução no R^2 ajustado, o ganho em precisão e generalização faz com que o segundo cenário seja mais adequado para a predição da VAL. Logo, o modelo baseado em fatores (regressão fatorial) apresentou melhor desempenho em relação ao modelo com variáveis originais, sendo a escolha mais indicada para aplicações práticas devido à sua maior precisão.

17.3 Referências

HAIR, J. F. *et al*. **Análise multivariada de dados**. Bookman editora, 2009,688p.

GUTTMAN, L. Some necessary conditions for common factor analysis. **Psychometrika**, v.19, p.149-162, 1954.

KAISER, H.F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, v.23, p.187-200. 1958.

17.4 Rotina em R

```
#URL do arquivo no GitHub (link para o arquivo .xlsx raw)
url <- "https://github.com/muriloloro/Modelos-de-Predicao/raw/main/dados-RFA.xlsx"
library(httr)
library(readxl)
# Definir o caminho temporário para salvar o arquivo
temp_file <- tempfile(fileext = ".xlsx")
GET(url, write_disk(temp_file, overwrite = TRUE))
dados <- read_excel(temp_file, sheet = "Covar")
```

###1. Matriz de correlação de Pearson

```
library(RColorBrewer)
library(corrplot)
library(ggcorrplot)
library(psych)

matcor <- cor(dados)
print(matcor, digits = 2)

#Figura de correlação
custom_colors <- colorRampPalette(c("red", "white", "blue"))(200)
par(family = "serif")
cor <- corrplot(matcor,
         method = "circle",
         col = custom_colors,
         tl.col = "black")
```

###2. Verificar a adequação da matriz de correlação para analise fatorial

```
#Podem ser utilizados os seguintes testes:
#  * Teste de esfericidade de Bartlett
#  * Kaiser-Meyer-Olkin (KMO)

#Ho: A matriz de correlação da população é uma matriz identidade, ou seja as
#variáveis não são correlacionadas na população.
```

```
#H1: A matriz de correlação da população não é uma matriz identidade, ou seja
#as variáveis são correlacionadas na população.

psych::cortest.bartlett(dados)
psych::KMO(dados)
```

###3. Selecionar o número de fatores

```
#----------PASSO 4 - VERIFICAR A VARIâNCIA EXPLICADA POR CADA FATOR
fit<-princomp(dados,cor=TRUE)
fit

summary(fit)

a <- screeplot(fit)

b <- plot(fit,type="lines")
```

###4. Rotação dos fatores

```
#Rotação obliqua (oblimin) - fatores são correlacionados
PCAoblimin <- principal(dados, nfactors=3,
                        n.obs=773,rotate="oblimin", scores=TRUE)
PCAoblimin
PCAoblimin$values #acessar autovalores
PCAoblimin$loadings #contribuição de cada variável (cargas fatoriais)
biplot(PCAoblimin)  #grafico biplot
```

###5. Obtenção das cargas fatoriais de cada genótipo

```
library(psych)
# verificando cada genótipo
escores <- factor.scores(dados,PCAoblimin,
              Phi = TRUE,
              method = c("Thurstone"),
              rho=NULL)

esc <- data.frame(escores$scores)
dados$TC1 <- esc$TC1
dados$TC2 <- esc$TC2
dados$TC3 <- esc$TC3

library(semPlot)
# Supondo que `PCAoblimin` é o modelo de análise fatorial ajustado
# Extrair a matriz de cargas do modelo
loadings_matrix <- PCAoblimin$loadings

# Criar uma nova matriz de cargas, zerando valores abaixo de 0.51
loadings_matrix[abs(loadings_matrix) < 0.51] <- 0

# Reatribuir a matriz de cargas ao objeto PCAoblimin
PCAoblimin$loadings <- loadings_matrix

# Agora, plotar com semPaths usando a nova matriz de cargas
c <- semPaths(PCAoblimin, "par",edge.label.cex = 1.3,
         sizeMan = 6, sizeLat = 6, shapeInt = 5,
         rotation = 2, layout = "tree",
```

```
        color = list(
  lat = rgb(253, 253, 253, maxColorValue = 255),
  man = rgb(155, 253, 175, maxColorValue = 255)),
  mar = c(1, 1, 1, 2))
```

###6. Regressão com dados originais e regressão fatorial

```
#URL do arquivo no GitHub (link para o arquivo .xlsx raw)
url <- "https://github.com/muriloloro/Modelos-de-Predicao/raw/main/dados-RFA.xl
sx"
library(httr)
library(readxl)
# Definir o caminho temporário para salvar o arquivo
temp_file <- tempfile(fileext = ".xlsx")
GET(url, write_disk(temp_file, overwrite = TRUE))
dados1 <- read_excel(temp_file, sheet = "Val")

attach(dados1)
dados$VAL <- dados1$VAL

#Cenário 1
library(caret)
modeloC1 <- caret::train(VAL~RSFM+RSFF+RFMC+RFFC+SFMC+SFFC+EUA+PSFM+PFMC,
           data = dados,
           method = "lm",
           trControl = trainControl(method = "LOOCV"))

summary(modeloC1)

#Cenário 2
modeloC2 <- caret::train(VAL~TC1 + TC2 + TC3,
           data = dados,
           method = "lm",
           trControl = trainControl(method = "LOOCV"))

summary(modeloC2)
```