Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Análise de agrupamento via método do k-médias para seleção genética em ovinos

Francisco Canindé Assis de Oliveira

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2024

Francisco Canindé Assis de Oliveira
Estatístico

Análise de agrupamento via método do k-médias para seleção genética em ovinos

versão revisada de acordo com a Resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **CRISTIAN MARCELO VILLEGAS LOBOS**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2024

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Oliveira, Francisco Canindé Assis de

Análise de agrupamento via método do k-médias para seleção genética em ovinos / Francisco Canindé Assis de Oliveira. – – versão revisada de acordo com a Resolução CoPGr 6018 de 2011. – – Piracicaba, 2024 .

73 p.

Dissertação (Mestrado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Análise de componentes prinicipais 2. Método do k-means 3. Ovinocultura 4. Análise de agrupamentos . I. Título.

AGRADECIMENTOS

Primeiramente, expresso minha gratidão a **Deus**, pois desde o dia do meu nascimento até o momento presente de minha vida, devo tudo à Sua graça.

À minha mãe, **Dona Telma**, uma mulher batalhadora que me criou praticamente sozinha, contando com a ajuda dos meus irmãos e irmã. Ela sempre esteve ao meu lado, nos bons e maus momentos.

À minha irmã, **Francisca Oliveira**, por ser mais que uma irmã, sendo uma grande amiga. Ela sempre me ouviu, motivou e incentivou.

Ao meu irmão, **Francisco de Assis**, pelo exemplo de força, garra e, acima de tudo, humildade. Ao meu irmão **Horlando Oliveira**, agradeço pela sabedoria transmitida.

Aos meus amigos, **Lucas Rafael** e **Gleison Américo**, por acompanharem toda a minha trajetória. O universo nos fez amigos, mas a vida nos tornou irmãos.

Aos amigos **Rene Airton**, **José Veraldo**, **Kevin Santos**, **Caio Mateus** e **Valmir Correia**, pelas conversas, companhia e parceria durante nossa vivência na residência da graduação na UFRN.

Às amizades feitas durante o período de mestrado, como **Carlos Chimarro**, **João Thiago**, **Maike Lovato**, **Silas Alves**, **Alana Uchoa**, **Thiago Moraes**, **Santiago Delgado**, **Christian Boinx**, **Laudeline Dantas**, **Edmundo Caetano**, **Rodrigo Domiciano**, **João Xavier**, pelas conversas e trocas de conhecimentos.

Ao meu orientador, **Cristian Villegas**, que aceitou a responsabilidade de orientar este nordestino. Agradeço pela oportunidade e pelas conversas, tanto no contexto acadêmico quanto no de amigo.

Ao professor **Helder Louvandini**, por aceitar a parceria neste trabalho.

Aos professores **Marcus Nunes**, **Marcelo Andrade**, **André Pinho** e **Germán Mauricio** por aceitarem fazer parte da comissão avaliadora deste trabalho.

A todos os professores do programa de pós-graduação em Estatística e Experimentação Agronômica, pelo conhecimento e ensinamentos compartilhados.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro para a realização desta pesquisa.

Não sou hábil com palavras, mas expresso de todo coração um sincero **Obrigado**.

EPÍGRAFE

"Cada dia é um presente, não um direito adquirido."
Nickelback

Se não puder voar, corra,
Se não puder corre, ande,
Se não puder andar, rasteje,
Mas faça o que fizer,
Continue seguindo em frente.
Martin Luther King

SUMÁRIO

RESUMO

Análise de agrupamento via método do k-médias para seleção genética em ovinos

A ovinocultura brasileira vem mostrando ao longo do tempo um aumento no potencial produtivo e de consumo. Entretanto, enfrenta barreiras relacionadas a criação dos pequenos ruminantes, barreiras relacionadas as infecções causadas por nematóides gastrointestinais. Buscando quebrar essas barreiras, os criadores costumam fazer uso de tratamentos a base de anti-helminticos, no entanto os nematóides ao longo do tempo criaram o que chamam de resistência anti-helmintica. Na busca por métodos que não envolvam esses tratamentos, alguns estudos surgiram objetivando a seleção genética via análise de agrupamentos. Nesse trabalho, são utilizados três variáveis, baseadas no número médio de ovos por grama de fezes (OPG), ganho de peso diário que foi calculado de duas formas e a porcentagem média para o teste do hematócrito (HT), as quais foram coletadas durante o período de 2020 a 2022 em ovinos da raça Santa Inês. Os dados passaram por uma análise de componentes principais e uma análise de agrupamentos pelo algoritmo do k-médias para a obtenção dos grupos com as características de resiliência, resistências e sensíveis a infecção pelos parasitas. A análise foi feita via software R e a determinação dos grupos foi feita através de uma análise detalhada dos dados.

Palavras-chave: Similaridade, Dissimilaridade, Componentes principais, Ovinocultura, Nematóides gastrointestinais, Parasitas helmintos.

ABSTRACT

Cluster analysis using the k-means method for genetic selection in sheep

Brazilian sheep farming has shown an increase in production and consumption potential over time. However, it faces barriers related to the breeding of small ruminants, barriers related to infections caused by gastrointestinal nematodes. Seeking to break down these barriers, breeders usually use treatments based on anthelmintics, however, nematodes over time have created what they call anthelmintic resistance. In the search for methods that do not involve these treatments, some studies have emerged aiming at genetic selection via cluster analysis. In this work, three variables are used, based on the average number of eggs per gram of feces (OPG), daily weight gain which was calculated in two ways and the average percentage for the hematocrit test (HT), which were collected during the period from 2020 to 2022 in Santa Inês sheep. The data underwent a principal component analysis and a cluster analysis using the k-means algorithm to obtain groups with the characteristics of resilience, resistance and sensitivity to infection by parasites. The analysis was done via *R* software and the groups were determined through a detailed analysis of the data.

Keywords: Similarity, Dissimilarity, Principal components, Sheep farming, Gastrointestinal nematodes, Helminth parasites

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

OPG	Ovos por grama de fezes
HT	Teste do hematócrito
GPI	Ganho de peso diário I
GPII	Ganho de peso diário II
ACP	Análise de componentes principais

1 INTRODUÇÃO

O aumento de casos de resistência dos parasitas gastrointestinais em ovinos aos diferentes princípios ativos tem gerado uma demanda por métodos alternativos de controle não-químico, como a seleção de animais que são geneticamente resistentes aos parasitos. A contagem de ovos por grama de fezes (OPG), valores do hematócrito (HT), contagem de eosinófilos sanguíneos e a classificação pelo método FAMACHA[1] são alguns indicadores bastantes utilizados em análises que auxiliam na identificação dos animais resistentes a infecção por esses parasitores (SOTOMAIOR et al., 2007).

Estudos relatam sobre a facilidade dos ovinos em carregar diversas variedades de espécies de parasitas helmintos. Por exemplo, no trabalho de Sargison (2011) destaca que as ovelhas são hospedeiras de diversos gêneros e espécies de parasitas helmintos. Para a maioria, desenvolveu-se um equilíbrio entre eles e os seus hospedeiros ovinos, em que o hospedeiro fornece o ambiente e nutrientes exigidos pela população parasitária, enquanto o parasita não compromete o hospedeiro a um ponto que ameace a sobrevivência das futuras gerações. O equilíbrio entre os hospedeiros ovinos e parasitas helmintos evoluiu ao longo de milhões de anos, mas foi pertubado em tempos relativamente recentes pela domesticação e práticas agricolas que favorecem os parasitas, pela seleção inadvertida de hospedeiros mais suscetíveis ou pela criação de ambientes que permitem o estabelecimento de populações maiores de estágios de vida livre dos parasitas. Esta pertubação no equilíbrio evolutivo parasita e hospedeiro afeta diferentes espécies de parasitas em diferentes extensões permitindo que algumas espécies de nematóides sejam potencialmente limitantes na produção dos ovinos e produtos derivados.

Helmintos parasitas como *Haemonchus contortus*, *Bunostomum trigonocephalum* e *Fasciola hepatica* são exemplos de parasitas limitantes da produção de ovinos, devido a efeitos diretos do seu comportamento de alimentação sanguínea. Entretando, as principais espécies de nematóides que limitam a produção em lugares com climas temperados[2] são *Teladorsagia circuncincta*, *Haemonchus contortus*, *Trichostrongylus vitrinus* e *Nematodirus battus*. Esses parasitas limitam a produtividade de animais suscetivéis devido às suas incubêmcias alimentares diretas que removem nutrientes da ingesta e devido aos efeitos indiretos na resposta imune de seus hospedeiros, danificando o revestimento absortivo do trato gastrointestinal, ou em alguns casos alimentando-se de sangue, segundo Sargison (2011).

A resistência por parte dos parasitas a anti-helmintos levou alguns pesquisadores a estudarem e utilizarem métodos alternativos para a seleção de animais geneticamente

[1] Um recurso para o controle da verminose em ovinos, (CHAGAS; CARVALHO; MOLENTO, 2007).

[2] Climas temperados são marcados por chuvas recorrentes de um volume anual variável, podendo ser continentais e oceânicos. No Brasil, correspondem ao clima subtropical. Para mais informações sobre os climas temprerados consultar Guitarrara (2023), disponível em (https://brasilescola.uol.com.br/geografia/clima-temperado.htm), acesso em 24 de outubro de 2023.

resistentes a infecção causada por esses parasitas. No estudo de Oliveira (2021), tem-se a abordagem de análise de agrupamentos por meio de indicadores que contribuem para a identificação desses animais. Considerando a seleção genética, este trabalho busca utlizar a metodologia da análise de agrupamento pelo método do k-médias para a obtenção dos animais geneticamente resistentes aos nematóides e separá-los em grupos.

O propósito deste estudo consiste em utilizar a técnica de análise de agrupamento para segmentar um conjunto de dados referentes a ovinos da raça Santa Inês. O objetivo é identificar, por meio das variáveis disponíveis, animais que apresentem características de resistência a infecção. A abordagem adotada envolve a aplicação do método de k-médias para a separação dos grupos, sendo que a seleção dos indivíduos em cada grupo é baseada nos centróides gerados para variáveis em análise e os indíviduos pertecentes a cada grupo.

O processo de seleção é respaldado por referências literárias, as quais desempenham o papel de indicadores fundamentais na distinção dos grupos. Estes grupos, por sua vez, são classificados conforme sua predisposição à resistência, resiliência ou suscetibilidade. Esse enfoque metodológico visa fornecer uma análise mais precisa e direcionada, contribuindo para a identificação e seleção de animais com características desejadas de resistência a infecção gerada pelos nematóides gastrointestinais.

Este trabalho está estruturado da seguinte forma: a Introdução, que pode ser encontrada no Capítulo 1; a Revisão Bibliográfica, detalhada no Capítulo 2, onde apresentaremos uma análise crítica da literatura existente; os Materiais e Métodos, descritos no Capítulo 3, que abrange a explicação dos procedimentos e dos materiais empregados; os Resultados, expostos no Capítulo 4, onde serão apresentados os dados obtidos; e, por fim, as Considerações Finais, que podem ser encontradas no Capítulo 5, onde discutiremos as conclusões.

2 REVISÃO BIBLIOGRÁFICA

2.1 Ovinocultura no Brasil

O Serviço Nacional de Apredizagem Rural SENAR (2021) informa que a ovinocultura de corte brasileira está com o seu crescimento impulsionado devido ao potencial elevado do mercado consumidor. Entretanto, o Brasil não tem capacidade de produção suficiente para atender a demanda, dessa forma aumentando a importação dos animais e produtos derivados, como a carne, o leite e a lã.

Considerando seu enorme potencial na produção pecuária, a realidade de produção insuficiente pode ser contornada, objetivando a criação de animais a serem abatidos em idade precoce, com carcaça de alta qualidade e custos compensadores. Para uma produção de animais mais pesados e em menor tempo, indica-se a adoção de cuidados na escolha das raças, os cruzamentos e o sistema de criação adequados a realidade e ao clima da propriedade associadas com a utilização de técnicas reprodutivas e conhecimentos de nutrição e prevenção de doenças, (SENAR, 2021).

A ovino cultura é uma importante área de captação de recursos para a economia brasileira segundo Osório et al. (2020). Encontrando-se em expansão, pois há um bom tempo existe uma maior demanda do mercado consumidor por produtos de melhor qualidade, porém em decorrência da situação produtiva, representada por uma pequena parcela na produção de carne, a baixa produtividade, o manejo inadequado e pouca tecnologia são fatores desafiadores para produção eficiente dos ovinos de acordo com Mendes et al. (2020). Em seu trabalho, Fernandes (2021) relata que as maiores dificuldades que a prática da ovinocultura enfrenta é a infecção por parasitores gastrintestinais como por exemplo o nematóide *Haemonchus Contortus* ilustrado na **Figura** 1.

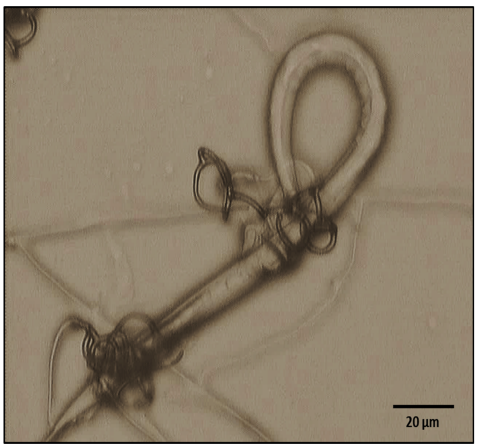


Figura 1. Imagem ilustrativa do nematóide *Haemonchus Contortus*.

Fonte: Pérez (2022). Disponível em: https://www.mdpi.com/2076-0817/11/10/1068

2.1.1 Nematoides gastrintestinais em ovinos

A maior parte do rebanho de corte no Brasil, principalmente as que econtram-se em sistemas de criação extensivo e semi-intensivo. Isto é, utilizam a pastagem como principal fonte nutricional, esses sistemas de criação apresentam um desempenho produtivo baixo. Os aniamis são mantidos em instalações que não permitem um manejo sanitário adequado, dificultando assim a prevenção e controle de doenças. Dessa forma, apresentam um elevado nível de contaminação por agentes infecciosos e parasitários. Entre as causas que interferem no desempenho produtivo dos animais em países tropicais e sub-tropicais, os parasitas gastrintestinais estão entre as principais informa Fernandes (2021).

Sargison (2011) descreve que os efeitos fisiopatológicos causados pela infecção dos parasitas gastrintestinais são a utilização ineficiente dos alimentos, induzindo a um estado de deficiência relativa de proteínas, desequilíbrios de fluidos e eletrólitos ou macroelementos e anemia, levando a sinais clínicos como a redução do apetite, baixo ganho em peso, diarreia e morte. Ainda condizente com o autor, os helmintos parasitas são indiscutivelmente as causas mais importantes da produtividade abaixo do ideal em ovinos, embora muitas vezes ocorram simultaneamente com outros problemas.

Devido à baixa produtividade causada pela infecção que afeta a digestão e a absorção dos nutrientes, é provável que a criação torne-se inviável economicamente. Isto está relacionado a saúde do rebanho a qual depende de um controle parasitário efetivo para que se mantenha os animais saudáveis. No Brasil, a ovinocultura tem como base a produção à pasto que favorece a contaminação dos animais por endoparasitas, pois estes se mantem endêmicos. O controle das pastagens torna-se difícil devido a resistência dos parasitas no ambiente, podendo afetar todas as etapas de produção. Além disso, os animais jovens e as fêmeas em periodo de gestação são mais suscetivéis, sofrendo maiores danos, e até mesmo a morte (FERNANDES, 2021).

A saúde de pequenos ruminantes ovinos e caprinos, há muito tempo sofre grande ameaça causada pelo parasitimo por nematoides gastrintestinais. Em ambas as espécies hospedeiras, os parasitas são responsáveis pelos maiores custos econômicos devidos às perdas diretas (produção reduzida, redução da qualidade de produtos derivados e a mortalidade dos animais) e as perdas indiretas (custos associados ao tratamento e controle, como os diagnósticos laboratoriais, medicamentos e mão de obra para a administração dos fármacos e o manejo dos animais), (HOSTE; SOTIRAKI; TORRES-ACOSTA, 2011).

As perdas econômicas enfatizam a necessidade de melhorias nas medidas de controle e tratamento, a rotina do uso de anti-helmínticos químicos tem sido de certa forma um obstáculo em programas de controle dos nematoides. Esses medicamentos são aplicados de formas curativas, buscando salvar os animais da morte, preservar a produtividade, quebrar o ciclo de vida dos parasitas antes que os animais atinjam níveis críticos. No entanto, o rápido desenvolvimento da resistência anti-helmíntica nas populações de ne-

matoides, é um problema importante. O controle é feito pela combinação de mais de uma abordagem, ou seja, não apenas pelo o uso de tratamentos anti-helmínticos, mas também pela redução da contaminação das pastagens e pela melhoria da imunidade do hospedeiro, de acordo com Hoste, Sotiraki e Torres-Acosta (2011).

2.1.2 Métodos para o controle dos nematoides gastrintestinais

Na literatura, muitos trabalhos exploraram os mecanismos de resistência à algumas classes de anti-helmínticos aumentando o conhecimento relevante. No trabalho de Arsenopoulos et al. (2021), os autores descrevem alguns estudos referentes à resistência anti-helmíntica para as principais classes de anti-helmínticos (*benzimidazóis*, *imidazotiazóis*, *lactonas macrocíclicas*), analisandos os desafios relativos a infecção pelo parasita *Haemonchus*.

A resistência anti-helmíntica leva a procura de outros métodos para o controle dos nematoides que não envolvam o uso de medicamentos e tratamentos clínicos, como a adesão de diferentes tipos de cultura e manejo na criação dos animais, conforme Arsenopoulos et al. (2021).

O controle dos parasitas gastrintestinais dependia estritamente do uso de anti-helmínticos, levando ao desenvolvimento da resistência parasitária, causando sérios prejuízos à criação de ovinos na Austrália. Por esse motivo, surgiu a proposta do "Controle Sustentável de Parasitas", objetivando a seleção genética de ovinos resistentes à verminose, com base na baixa contagem de ovos por grama de fezes, associadada a outros métodos integrados de controle, como a suplementação alimentar, manejo de pastagens e a redução do uso de compostos químicos em épocas estratégicas (VIEIRA et al., 2009).

O método de seleção baseado em uma baixa contagem de ovos por grama de fezes mostrou vantagens, tanto na redução da carga parasitária dos animais quanto na redução da contaminação da pastagem. A resistência dos hospedeiros impede o estabelecimento da infecção. Ainda que, a redução de OPG ano a ano seja pequena, apresenta reflexo permanente na melhoria da resistência dos animais (VIEIRA et al., 2009).

As infecções por nematoides gastrintestinais provocam prejuízos aos produtores de caprinos e a utilização de anti-helmintos é método de controle preferido pelos criadores. O tratamento via anti-helmintos é feito pela vermifugação[1] do rebanho com um acompanhamento mensal, a aplicação é feita quando o valor médio para o número de ovos por grama de fezes for superior a 500 (HASSUM, 2009). Entretanto, o uso sem orientação tem levado ao desenvolvimento de multirresistência por parte dos nematoides, isso estimulou Oliveira (2021) estudar uma opção para a solução desse problema, a seleção de animais resistentes aos parasitas analisando a respostas de caprinos a infecção por verminosa em rebanho experimental.

[1]Vermifugação é um método que faz uso de produtos químicos objetivando destruir ou expulsar vermes instestinais, Dicionário Português. Disponível em: https://www.dicio.com.br/vermifuga

A seleção foi feita pela combinação da lógica de programação Fuzzy e informações de OPG, escore da condição corporal (ECC), e escore FAMACHA [2]. Os resultados obtidos pelos recursos da lógica Fuzzy mostrou-se eficiente, embora comparados com análise de agrupamento multivariado apresentou o menor percentual de acerto global (OLIVEIRA, 2021).

2.2 Análise de componentes principais

Existem alguns métodos que são utilizadas previamente em algumas análises, como por exemplo a análise componentes principais, que antecedem uma análise de regressão multivariada e análise agrupamentos (JOHNSON; WICHERN, 2007).

Segundo Koritiaki et al. (2019), o principal objetivo da análise de componentes principais (ACP) é descrever a estrutura de variância e da covariância de uma nuvem de n pontos no espaço de dimensão p, $\mathbb{R}^p$, extraindo dessa nuvem de n pontos um novo conjunto de variáveis de mesma dimensão, ortogonais e não-correlacionadas.

Em seu trabalho, Bortoluzzi (2018) utilizou a metodologia da análise de componentes principais (ACP) para avaliar a variabilidade e as relações entre os valores genéticos de uma análise heptacarater para parâmetros da curva de lactação de ovinos da raça Lacaune. Em seu estudo foi possível obter três componentes principais responsáveis por explicar 94,59% da variabilidade total dos dados.

As análises de componentes principais são um meio que pode levar a ter facilidade, em outras palavras, é um método que pode ajudar (ou servir como entrada) a alguma outra análise, frequentemente servem como etapas intermediárias em investigações muito mais amplas. Por exemplo, podem servir como entrada para uma análise regressão múltipla ou análise de cluster (agrupamentos). Além disso, os componentes principais (escaladas) são uma "fatoração" da matriz de covariância para um modelo de análise fatorial detalhado no Capítulo 9 de Johnson e Wichern (2007).

Análise de componentes principais tem como objetivo reduzir e interpretar a estrutura de variância e covariância de um conjunto de variáveis através de combinações lineares dessas variáveis. Entretanto, são necessários p componentes para reduzir a variabilidade total do sistema, muitas vezes grande parte dessa variabilidade pode ser explicada por um subconjunto pequeno de k componentes principais. Assim, há quase tanta informação nos k componentes quanto nas p variáveis iniciais. Dessa forma, os k componentes principais podem então substituir as p variáveis, e o conjunto de dados original que possui n mensurações em p variáveis é reduzido a um conjunto de dados que consiste em n mensurações em k componentes principais.

[2] FAMACHA é um método seletivo que tem como objetivo vermifugar animais que apresentam sinais de anemia, facilmente visualizada na mucosa ocular dos ovinos segundo Chagas, Carvalho e Molento (2007).

2.2.1 Componentes Principais da População

Componentes principais são combinações lineares específicas das p variáveis aleatórias $X_1, X_2, \cdots, X_p$, em sua contextualização algébrica. Geometricamente, essas combinações representam uma seleção de um novo sistema de coordenadas obtido pela rotação do sistema original com $X_1, X_2, \cdots, X_p$ como eixo das coordenadas. Esses novos eixos representam as direções com máxima variabilidade e fornecem uma descrição mais simples e parcimoniosa da estrutura da covariância.

As componentes principais dependem unicamente da matriz de covariâncias $\mathbf{\Sigma}$ (ou da matriz de correlação $\boldsymbol{\rho}$) das variáveis. O seu desenvolvimento não requer uma suposição da distribuição normal multivariada. Porém, os componentes principais derivados para populações com distribuição normal multivariada tem interpretação útil em termos de elipsóides de densidade constante. Além disso, podem ser feitas inferências a partir dos componentes da amostra quando a população tem distribuição normal multivariada.

Seja o vetor aleatório $\boldsymbol{X'} = [X_1, X_2, \cdots, X_p]$ de dimensão $p \times 1$ com matriz de covariâncias $\mathbf{\Sigma}$ e autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ e seja $\boldsymbol{a}_i' = [a_{i1}, a_{i2}, \cdots, a_{ip}]$ um ventor de constantes de dimensão $1 \times p$. Assim, considere as seguintes combinações lineares

$$\begin{aligned} Y_1 &= \boldsymbol{a}_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= \boldsymbol{a}_2'\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ Y_p &= \boldsymbol{a}_p'\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p, \end{aligned} \tag{1}$$

usando as definições de variância e covariâncias de vetores aleatórios dadas por $Var(\boldsymbol{aY}) = \boldsymbol{a'}Var(\boldsymbol{Y})\boldsymbol{a}$ e $Cov(\boldsymbol{aY}, \boldsymbol{aY'}) = \boldsymbol{a'}Cov(\boldsymbol{Y}, \boldsymbol{Y'})\boldsymbol{a}$, respectivamente. Obtemos,

$$\begin{aligned} Var(\boldsymbol{Y}_i) &= \boldsymbol{a}_i'\mathbf{\Sigma}\boldsymbol{a_i} \qquad\qquad i = 1, 2, \cdots, p \\ Cov(\boldsymbol{Y}_i, \boldsymbol{Y}_k) &= \boldsymbol{a}_i'\mathbf{\Sigma}\boldsymbol{a_k} \qquad\qquad i, k = 1, 2, \cdots, p \quad \text{sendo} \quad i \neq k, \end{aligned} \tag{2}$$

os componentes principais são aquelas combinações lineares não correlacionadas $Y_1, Y_2, \cdots, Y_p$ cuja variância dada na equação (2) seja a maior possível.

A primeira componente principal é a combinação linear com máxima variância, ou seja, maximiza $Var(Y_1) = \boldsymbol{a}_1'\mathbf{\Sigma}\boldsymbol{a}_1$, embora $Var(Y_1)$ pode ser aumentada multiplicando qualquer $\boldsymbol{a}_1'$ por alguma constante. Logo, para eliminar essa indeterminação, é considerado conveniente restringir a atenção aos vetores de coeficientes de comprimento unitário. Portanto, define-se

> Primeira componente principal como a combinação linear $\boldsymbol{a}_1'\boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_1'\boldsymbol{X})$ sujeita a $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$;
>
> Segunda componente principal como a combinação linear $\boldsymbol{a}_2'\boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_2'\boldsymbol{X})$ sujeita a $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$ e $Cov(\boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{a}_2'\boldsymbol{X}) = 0$

no i-ésimo passo temos,

> i-ésima componente principal como a combinação linear $\boldsymbol{a}_i'\boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_i'\boldsymbol{X})$ sujeita a $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$ e $Cov(\boldsymbol{a}_i'\boldsymbol{X}, \boldsymbol{a}_k'\boldsymbol{X}) = 0$, $i \neq k$.

Seja $\boldsymbol{\Sigma}$ a matriz de covariâncias associada ao vetor aleatório $\boldsymbol{X}' = [X_1, X_2, \cdots, X_p]$. Considere que $\boldsymbol{\Sigma}$ tenha os seguintes pares de autovalores e autovetores $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \cdots, (\lambda_p, \boldsymbol{e}_p)$, em que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ e $\boldsymbol{e}_i' = [e_{i1}, e_{i2}, \cdots, e_{ip}]$ para $i = 1, 2, \cdots, p$. Logo a i-ésima componente principal é dada por

$$\boldsymbol{Y}_i = \boldsymbol{e}_i'\boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \qquad i = 1, 2, ..., p, \tag{3}$$

dessa forma, temos

$$\begin{aligned} Var(\boldsymbol{Y}_i) &= \boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i \qquad i = 1, 2, \cdots, p \\ Cov(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}) &= \boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_k = 0 \qquad i \neq k, \end{aligned} \tag{4}$$

se algum λ_i for igual a escolha do vetor de coeficientes correspondente $\boldsymbol{e}_i$ e, portanto, $\boldsymbol{Y}_i$ não é unico. E desse resultado, as componentes principais não são correlacionadas e suas variâncias são iguais aos autovalores de $\boldsymbol{\Sigma}$.

Seja $\boldsymbol{X}' = [X_1, X_2, \cdots, X_p]$ com matriz de covariâncias $\boldsymbol{\Sigma}$, com os pares de autovalores e autovetores $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \cdots, (\lambda_p, \boldsymbol{e}_p)$ em que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Seja $Y_1 = \boldsymbol{e}_1'\boldsymbol{X}, Y_2 = \boldsymbol{e}_2'\boldsymbol{X}, \cdots, Y_p = \boldsymbol{e}_p'\boldsymbol{X}$ as componentes principais e $\sigma_{11}, \sigma_{22}, \cdots, \sigma_{pp}$ as variâncias relacionadas a cada componente principal. Então,

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i), \tag{5}$$

isto indica que a variância total da população é σ, ou seja, a soma das variâncias relacionadas a cada componente é a soma de cada autovalor associado às suas respectivas componentes

$$\sigma = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_p, \tag{6}$$

consequentemente, a propoção π_k da variância total devido (explicada pelas) k-ésimas componentes principais é

$$\pi_k = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \cdots, p, \tag{7}$$

se a maior parte (por exemplo de 80% a 90%, nesse caso valores arbitrários) da variabilidade total da população, para um p grande, pode ser atribuída até as três primeiras componentes principais então essas componentes podem "substituir" as p variáveis originais

sem muita perda de informação, para mais informações sobre o número de componentes principais consultar o livro de Johnson e Wichern (2007).

Cada componente do vetor de coeficientes $\boldsymbol{e}_i' = [e_{i1}, \cdots, e_{ik}, \cdots, e_{ip}]$ também merece inspeção. A magnitude de e_{ik} mede a importância da k-ésima variável para i-ésima componente principal, independentemente das outras variáveis. Em particular, e_{ik} é proporcional ao coeficiente de correlção entre Y_i e X_k.

Sejam $Y_1 = \boldsymbol{e}_1'\boldsymbol{X}, Y_2 = \boldsymbol{e}_2'\boldsymbol{X}, \cdots, Y_p = \boldsymbol{e}_p'\boldsymbol{X}$ as componentes principais obtidas através da matriz de covariâncias $\boldsymbol{\Sigma}$, então

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i,k = 1,2,\cdots,p, \tag{8}$$

são os coeficientes de correlação entre os componentes Y_i e as variáveis X_k. Assim, $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), \cdots, (\lambda_p, \boldsymbol{e}_p)$ são os pares de autovalores e autovetores associados a $\boldsymbol{\Sigma}$, para mais detalhes consultar Johnson e Wichern (2007).

2.2.2 Componentes Principais Obtidas a Partir de Variáveis Padronizadas

As componentes principais também podem ser obtidas através da padronização das variáveis, na forma

$$\begin{aligned} Z_1 &= \tfrac{(X_1-\mu_1)}{\sqrt{\sigma_{11}}} \\ Z_2 &= \tfrac{(X_2-\mu_2)}{\sqrt{\sigma_{22}}} \\ &\vdots \\ Z_p &= \tfrac{(X_p-\mu_p)}{\sqrt{\sigma_{pp}}}, \end{aligned} \tag{9}$$

em sua forma matricial $\boldsymbol{Z} = (\boldsymbol{V}^{1/2})^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$, em que $\boldsymbol{V}^{1/2} = diag(\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \cdots, \sqrt{\sigma_{pp}})$ é a matriz diagonal de desvios padrão. Obviamente, $E(\boldsymbol{Z}) = \boldsymbol{0}$ e $Cov(\boldsymbol{Z}) = (\boldsymbol{V}^{1/2})^{-1}\boldsymbol{\Sigma}$ $(\boldsymbol{V}^{1/2})^{-1} = \boldsymbol{\rho}$. As componentes principais de $\boldsymbol{Z}$ podem ser obtidas através dos autovetores da matriz de correlação $\boldsymbol{\rho}$ de $\boldsymbol{X}$. Todos os resultados mostrados anteriormente podem ser aplicados, com alguma simplificação, uma vez que a variância de cada Z_i é única. Continua-se usando as notações Y_i e $(\lambda_i, \boldsymbol{e}_i)$ para referir-se a i-ésima componente principal e os pares de autovalores e autovetores, respecitvamente, de $\boldsymbol{\rho}$ ou $\boldsymbol{\Sigma}$. No entanto, os $(\lambda_i, \boldsymbol{e}_i)$ derivados de $\boldsymbol{\Sigma}$ não são, no geral, iguais aos derivados de $\boldsymbol{\rho}$, para mais detalhes consultar Johnson e Wichern (2007).

2.2.3 Componentes Principais da Amostra

Suponha que os dados $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p$ representam n ensaios independentes de alguma população p-dimensional com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. Produzindo o vetor amostral de médias $\bar{\boldsymbol{x}}$ e as matrizes de covariância amostral $\boldsymbol{S}$ e correlação amostral $\boldsymbol{R}$. Objetiva-se construir uma combinação linear não-correlacionada

da caracteristica medida que explique grande parte da variação na amostra. Todas as combinações não-correlacionadas com as maiores variâncias são chamadas de componentes principais da amostra.

Lembrando que os n valores de qualquer combinação linear

$$\boldsymbol{a}_1'\boldsymbol{x} = \sum_{i=1}^{p} a_{1i}x_{ji}. \qquad j = 1, 2, \cdots, n,$$

tem média e variância amostral $\boldsymbol{a}_1'\bar{\boldsymbol{x}}$ e $\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1$. Além disso, os pares de valores $(\boldsymbol{a}_1'\boldsymbol{x}_j, \boldsymbol{a}_2'\boldsymbol{x}_j)$, para duas combinações lineares tem covariância amostral $\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_2$.

As componentes principais da amostra são definidas como aquelas combinações lineares que apresentam variância amostral máxima. Tal como acontece com as quantidades populacionais, restringe-se os vetores de coeficientes $\boldsymbol{a}_i$ para satisfazer $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$. Especificamente,

> Primeira componente principal da amostra é a combinação linear $\boldsymbol{a}_1'\boldsymbol{x}_j$ que maximiza a variância amostral de $\boldsymbol{a}_1'\boldsymbol{x}_j$ sujeita a $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$.
>
> Segunda componente principal da amostra é a combinação linear $\boldsymbol{a}_2'\boldsymbol{x}_j$ que maximiza a variância amostral de $\boldsymbol{a}_2'\boldsymbol{x}_j$ sujeita a $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$ e covariância amostral igual a zero para o par $(\boldsymbol{a}_1'\boldsymbol{x}_j, \boldsymbol{a}_2'\boldsymbol{x}_j)$.

Dessa forma no i-ésimo passo, tem-se

> A i-ésima componente principal da amostra é a combinação linear $\boldsymbol{a}_i'\boldsymbol{x}_j$ que maximiza a variância amostral de $\boldsymbol{a}_i'\boldsymbol{x}_j$ sujeita a $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$ e covariância amostral igual a zero para todos os pares $(\boldsymbol{a}_i'\boldsymbol{x}_j, \boldsymbol{a}_{i'}'\boldsymbol{x}_j)$, $i' < i$.

A primeira componente principal maximiza $\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1$ ou, equivalentemente

$$\frac{\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1}{\boldsymbol{a}_1'\boldsymbol{a}_1}, \tag{10}$$

o máximo é o maior autovalor alcançado $\hat{\lambda}_1$ para a escolha do autovetor $\boldsymbol{a}_1 = \hat{\boldsymbol{e}}_1$ de $\boldsymbol{S}$. Escolhas sucessivas de $\boldsymbol{a}_i$ maximizam a equação (10) sujeita a $\boldsymbol{a}_i'\boldsymbol{S}\hat{\boldsymbol{e}}_k = \boldsymbol{a}_i'\hat{\lambda}_k\hat{\boldsymbol{e}}_k = 0$, ou seja, $\boldsymbol{a}_i$ é perpendicular a $\hat{\boldsymbol{e}}_k$. Seja $\boldsymbol{S} = s_{ik}$ a matriz de covariância amostral de ordem $p \times p$ com os pares de autovalores e autovetores $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), (\hat{\lambda}_2, \hat{\boldsymbol{e}}_2), \cdots, (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$, a i-ésima componente principal da amostra é dada por

$$\hat{\boldsymbol{y}}_i = \hat{\boldsymbol{e}}_i'\boldsymbol{x} = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p, \qquad i = 1, 2, \cdots, p,$$

em que $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ e $\boldsymbol{x}$ é qualquer observação sobre as variáveis $X_1, X_2, \cdots, X_p$. Também,

$$\begin{aligned} S(\hat{y}_k) &= \hat{\lambda}_k, \qquad k = 1, 2, \cdots, p \\ r(\hat{y}_i, \hat{y}_k) &= 0, \qquad i \neq k, \end{aligned} \tag{11}$$

$S(\hat{y}_k)$ e $r(\hat{y}_i, \hat{y}_k)$ são a variância e covariância amostral, respectivamente. Além disso, a variância amostral total é dada por

$$S_t = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p \tag{12}$$

e

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_k}}{\sqrt{s_{kk}}}, \qquad i, k = 1, 2, \cdots, p.$$

As observações $\boldsymbol{x}_j$ são frequentemente "centradas" subtraindo $\bar{\boldsymbol{x}}$. Isso não afeta a matriz de covariância amostral $\boldsymbol{S}$ e fornece a i-ésima componente principal

$$\hat{\boldsymbol{y}}_i = \hat{\boldsymbol{e}}_i'(\boldsymbol{x} - \bar{\boldsymbol{x}}), \qquad i = 1, 2, \cdots, p, \tag{13}$$

para qualquer vetor de observações $\boldsymbol{x}$. Considerando os valores da i-ésima componente

$$\hat{\boldsymbol{y}}_{ji} = \hat{\boldsymbol{e}}_i'(\boldsymbol{x}_j - \bar{\boldsymbol{x}}), \qquad j = 1, 2, \cdots, n,$$

gerada substituindo cada observação $\boldsymbol{x}_j$ por uma observação abritária $\boldsymbol{x}$, então

$$\bar{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{\boldsymbol{e}}_i' \left(\sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}}) \right) = \frac{1}{n}\hat{\boldsymbol{e}}_i'\boldsymbol{0} = 0,$$

ou seja, a média amostral de cada componente principal é zero. As variâncias amostrais ainda são fornecidas pelos $\hat{\lambda}_i$'s, assim como na equação (13), para mais detalhes consultar Johnson e Wichern (2007).

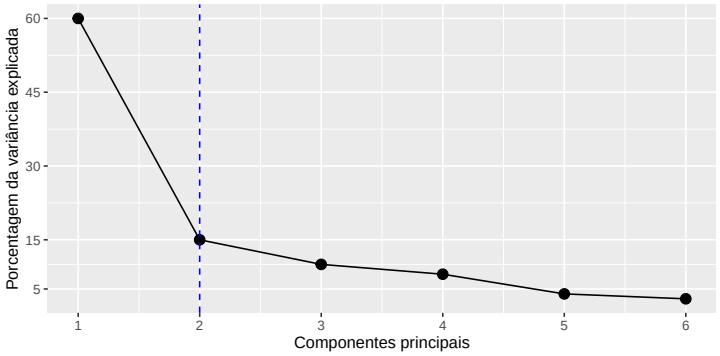
2.2.4 Número de Componentes Principais

Uma questão sempre abordada é quando temos que escolher o número de componentes principais. Embora não exista uma resposta definitiva para essa questão, alguns itens a serem considerados são a quantidade da variância total da amostra explicada, os tamanhos reais dos autovalores (a variância das componentes da amostra) e as interpretações das componentes do assunto. Além disso, uma componente associada a um autovalor próximo de zero pode indicar uma dependência linear inesperada nos dados.

Entretanto, existem algumas ferramentas úteis para determinar um número apropriado de componentes, uma delas seria o *Scree plot*. Ordenando os autovalores do maior para o menor, um gráfico dos $\hat{\lambda}_i$ versus i, um gráfico da magnitude de um autovalor versus o seu índice. Para determinar o número apropriado de componentes, buscamos um

cotovelo (uma curva) no *Scree Plot* como mostra a **Figura** 2. O número de componentes é considerado o ponto em que os autovalores restantes são relativamente pequenos e todos aproximadamente do mesmo tamanho.

Figura 2. Ilustração do gráfico *Scree plot*.



2.2.5 Interpretação das Componentes Principais da Amostra

Diversas interpretações podem ser feitas sobre as componentes principais da amostra. Primeiramente, suponha que a distribuição adjacente de $\boldsymbol{X}$ seja aproximadamente $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Então, as componentes principais da amostra $\hat{y}_i = \hat{\boldsymbol{e}}_i'(\boldsymbol{x} - \bar{\boldsymbol{x}})$ são a realização das componentes principais da população $Y_i = \boldsymbol{e}_i'(\boldsymbol{X} - \boldsymbol{\mu})$ que possuem uma distribuição $N_p(\boldsymbol{0}, \boldsymbol{\Lambda})$, em que $\boldsymbol{\Lambda}$ é uma matriz diagonal dos autovalores $\lambda_1, \lambda_2, \cdots, \lambda_p$ e $(\lambda_i, \boldsymbol{e}_i)$ são os pares de autovalores e autovetores de $\boldsymbol{\Sigma}$.

Além disso, podemos estimar $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ a parti de $\bar{\boldsymbol{x}}$ e $\boldsymbol{S}$ usando os valores amostrais $\boldsymbol{x}_j$. Se $\boldsymbol{S}$ for positiva definida, o contorno que consiste em todos os vetores $p \times 1$ de $\boldsymbol{x}$ satisfazendo

$$(\boldsymbol{x} - \bar{\boldsymbol{x}})' \boldsymbol{S}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}) = c^2, \tag{14}$$

estimando o contorno da densidade constante $(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c^2$ da densidade normal subjacente. Os contornos aproximados os quais podem ser desenhados no gráfico de dispersão indicam a distribuição normal que gerou os dados. A suposição da normalidade é util para procedimentos de inferência, mas não há necessidades para o desenvolvimento das propriedades das componentes principais da amostra.

Mesmo quando a suposição da distribuição normal é suspeita e o gráfico de dispersão se distancie de um padrão elíptico, ainda podemos extrair os autovalores de $\boldsymbol{S}$ e obter as componentes principais da amostra. Geometricamente, os dados podem ser plotados como n pontos em p-espaços. Podendo então serem expressos nas novas coordenadas, que coincidem com os eixos do contorno da equação (14). Definindo um hiperelipsóide centrado em $\bar{\boldsymbol{x}}$ cujos eixos são dados pelos autovetores de $\boldsymbol{S}^{-1}$, ou equivalentemente, de $\boldsymbol{S}$. Os comprimentos dos eixos dos hiperelipsóide são proporcionais a $\sqrt{\lambda_i}$, $i = 1, 2, \cdots, p$, em que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ são os autovalores de $\boldsymbol{S}$.

Como $\hat{\boldsymbol{e}}_i$ tem comprimento 1, o valor absoluto da i-ésima componente principal, $|\hat{y}_i| = |\hat{\boldsymbol{e}}_i'(\boldsymbol{x} - \bar{\boldsymbol{x}})|$, fornece o comprimento da projeção do vetor $(\boldsymbol{x} - \bar{\boldsymbol{x}})$ no vetor unitário $\hat{\boldsymbol{e}}_i$. Dessa forma, as componentes principais da amostra $\hat{y}_i = \hat{\boldsymbol{e}}_i'(\boldsymbol{x} - \bar{\boldsymbol{x}})$, $i = 1, 2, \cdots, p$, ficam ao longo dos eixos do hiperelipsóide, e seus valores absolutos são os comprimentos das projeções de $\boldsymbol{x} - \bar{\boldsymbol{x}}$ nas direções dos eixos de $\hat{\boldsymbol{e}}_i$. Consequentemente, as componentes principais da amostra podem ser vistas como o resultado da translação da origem do sistema de coordenadas original de $\bar{\boldsymbol{x}}$ e, em seguida, da rotação dos eixos de coordenadas até que passem pela dispersão nas direções de variância máxima.

2.2.6 Componentes Principais Padronizadas da Amostra

As componentes principais da amostra, no geral, não são invariantes em relação à mudança da escala. As variáveis medidas em escalas de direção ou em uma escala comum com faixas muito difentes são frequentemente padronizadas. A padronização é realizada através de

$$\boldsymbol{z}_j = \boldsymbol{D}^{-1/2}(\boldsymbol{x}_j - \bar{\boldsymbol{x}}) = \begin{bmatrix} \frac{x_{j1} - \bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{j2} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \qquad j = 1, 2, \cdots, n, \tag{15}$$

em que $\boldsymbol{D}^{-1/2} = \left[\frac{1}{\sqrt{s_{11}}}, \frac{1}{\sqrt{s_{22}}}, \cdots, \frac{1}{\sqrt{s_{pp}}}\right]$, assim a matriz $n \times p$ de observações padronizadas é dada por

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1' \\ \boldsymbol{z}_2' \\ \vdots \\ \boldsymbol{z}_n' \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} = \begin{bmatrix} \frac{x_{11} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{p1} - \bar{x}_p}{\sqrt{s_{pp}}} \\ \frac{x_{21} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{22} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{2p} - \bar{x}_p}{\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1} - \bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{n2} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{np} - \bar{x}_p}{\sqrt{s_{pp}}}, \end{bmatrix} \tag{16}$$

produzindo o vetor de médias amostrais

$$\bar{\boldsymbol{z}} = \frac{1}{n}(\mathbf{1}'\boldsymbol{Z})' = \frac{1}{n}\boldsymbol{Z}'\mathbf{1} = \frac{1}{n}\begin{bmatrix} \sum_{j=1}^{n} \frac{x_{j1} - \bar{x}_1}{\sqrt{s_{11}}} \\ \sum_{j=1}^{n} \frac{x_{j2} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \sum_{j=1}^{n} \frac{x_{jp} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = \mathbf{0} \tag{17}$$

e matriz de covariância amostrais

$$
\begin{aligned}
\boldsymbol{S}_z &= \frac{1}{n-1}\left(\boldsymbol{Z}-\frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{Z}\right)'\left(\boldsymbol{Z}-\frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{Z}\right)\\
&= \frac{1}{n-1}(\boldsymbol{Z}-\bar{\boldsymbol{z}})'(\boldsymbol{Z}-\bar{\boldsymbol{z}})\\
&= \frac{1}{n-1}\boldsymbol{Z}'\boldsymbol{Z}\\
&= \frac{1}{n-1}\begin{bmatrix}
\frac{(n-1)s_{11}}{s_{11}} & \frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}}\\
\frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \frac{(n-1)s_{22}}{s_{22}} & \cdots & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}}\\
\vdots & \vdots & \ddots & \vdots\\
\frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \frac{(n-1)s_{pp}}{s_{pp}}
\end{bmatrix}\\
&= \boldsymbol{R},
\end{aligned}
\tag{18}
$$

as componentes principais da amostra são dadas na forma da equação (12), com a matriz $\boldsymbol{R}$ no lugar de $\boldsymbol{S}$. Como as observações já estão "centradas" pela construção, não há necessidade de escrever as componentes na forma $\hat{y}_i = \hat{\boldsymbol{e}}_i'(\boldsymbol{x}-\bar{\boldsymbol{x}})$.

Se $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n$ são as observações padronizadas com matriz de covariâncias $\boldsymbol{R}$, as i-ésimas componentes principais da amostra são

$$
\hat{y}_i = \hat{\boldsymbol{e}}_i'\boldsymbol{z} = \hat{e}_{i1}z_1 + \hat{e}_{i2}z_2 + \cdots + \hat{e}_{ip}z_p, \qquad i = 1, 2, \cdots, p, \tag{19}
$$

em que $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ são os i-ésimos pares de autovalores e autovetores de $\boldsymbol{R}$ com $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Assim,

$$
\begin{aligned}
&V_a(\hat{y}_i) = \hat{\lambda}_i \qquad i = 1, 2, \cdots, p\\
&cov_a(\hat{y}_i, \hat{y}_{i'}) = 0,
\end{aligned}
\tag{20}
$$

$V_a(\hat{y}_i)$ e $cov_a(\hat{y}_i, \hat{y}_{i'})$ representam a variância e covariância amostral. Além disso, a variância total amostral é dada por

$$
V_{at} = tr(\boldsymbol{R}) = p = \sum_{i=1}^{p}\hat{\lambda}_i \tag{21}
$$

e, também

$$
\boldsymbol{r}_{\hat{y}_i, z_k} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i} \qquad i, k = 1, 2, \cdots, p. \tag{22}
$$

Logo, a proporção π_i da variância total da amostra explicada pelas i-ésimas componentes principais da amostra é

$$
\pi_i = \frac{\hat{\lambda}_i}{p}, \qquad i = 1, 2, \cdots, p. \tag{23}
$$

Uma regra prática sugere escolher apenas aquelas componentes cujas variâncias $\hat{\lambda}_i$ são maiores que unidade ou, equivalentemente, apenas aquelas componentes que, individualmente, explicam pelo menos uma proporção $1/p$ da variância total. Esta regra não tem muito apoio teórico e não deve ser aplicada cegamente. No entanto, o *Scree Plot* já é util para selecionar o número apropriado de componentes. Algumas notações foram levemente alteradas e para mais detalhes consultar o Capítulo 8 de Johnson e Wichern (2007).

2.3 Análise de agrupamentos

As técnicas multivariadas consistem em uma poderosa e valiosa ferramenta para programas de melhoramento genético. Auxiliando na identificação de grupos de animais produtivamente semelhantes, mesmo que sejam geneticamente divergentes, segundo (OLIVEIRA, 2021).

A análise de agrupamentos é um método estatístico de análise multivariada, que possui aplicabilidade em várias áreas na análise de dados voltados ao melhoramento genético animal. Essa análise visa a obtenção de grupos de indivíduos reunidos de acordo com a proximidade genética, com relação a um determinado conjunto de variáveis. Esse método eleva a homogeneidade dos indíviduos de um mesmo grupo, e a heterogeneidade dos indivíduos de grupos distantes, condizente com Oliveira (2016).

No estudo de Oliveira (2016), com base nos marcadores FAMACHA, contagem de ovos por grama de fezes, volume globular (VG), proteína plasmática total (PPT), escore de condição corporal e o peso corporal (PC) utilizados em uma análise de agrupamentos para obtenção dos grupos resistência/resiliência a infecção. A análise foi feita com o uso das técnicas de agupamentos hierárquicos para averiguar a possibilidade de divisão, distância euclidiana para a definição do número de agrupamentos a serem utilizados na análise de agrupamento não-hierárquica, o método não-hierárquico utilizado foi o método do k-médias.

Em seu trabalho, Sotomaior et al. (2007) envolveram a análise de agrupamentos baseado na similaridade ou na distância, utilizando as ligações completas e a distância euclidiana. Porém, condizente com o autor, a análise de agrupamentos por si só foi utilizada para o agrupamento, a definição dos grupos foi feita pela a análise bruta dos dados. Menegatto (2023) estudou a seleção de ovinos da raça Santa Inês buscando as caracteristicas de resistência e resiliência a infeção causada pelo nematóide *Haemonchus contortus*, realizando análise de agrupamentos para caracterização desses grupos. O método utilizado foi o algoritmo de Ward (1963) e a utilização da distância euclidiana como medida de similaridade entre os animais.

2.3.1 Métodos para Distância e Agrupamentos

Procedimentos exploratórios que são muitas vezes bastante úteis na compreensão da natureza complexa das relações multivariadas. Uma das técnicas explicativas muito importante é a busca de estrutura de agrupamentos "naturais" nos dados. Os agrupamentos podem fornecer um meio informal para avaliar a dimensionalidade, identificar valores discrepantes e sugerir hipóteses interessantes sobre as relações das variáveis.

Análise de agrupamentos consiste em classificar um número conhecidos de grupos com o objetivo operacional de atribuir uma nova observação a um desses grupos. A análise de agrupamentos (análise de *clusters*) é uma técnica simples, pois nenhuma suposição é feita em relação ao número de grupos ou à estrutura dos grupos. O agrupamento é feito com base em semelhanças (similaridades) ou distâncias (dissimilaridades). As entradas necessárias são medidas de similaridade ou dados apartir dos quais a similaridade pode ser calculada.

Nesta seção, serão abordados os métods para distância e agrupamentos, para os métodos de distância teremos as *Medidas de Similaridade* para itens e variáveis, já os *Métodos de Agrupamentos* serão apresentados os *Métodos de Agrupamentos Hierárquicos e Não-Hierárquicos*, com foco maior no *Método do k-médias*, método de agrupamento adotado neste trabalho.

2.4 Medidas de similaridade

A maioria dos esforços para produzir uma estrutura de grupo amostral a partir de um conjunto de dados complexo requer uma medida de "proximidade" ou "similaridade". Embora que, muitas vezes há muita subjetividade envolvida na escolha de uma medida de similaridade. Considerações importantes devem ser feitas como a natureza das variáveis se são (discretas ou continuas), sua escala de medição (nominal, ordinal, intervalar, proporção) e conhecimento do assunto.

2.4.1 Distâncias e coeficientes de similaridade para pares de itens

Nesta seção será abordado um pouco sobre *distância euclidiana*, *métrica de Minkowski*, *medida de Canberra*, *coeficiente de Czekanowski*, para relembrar consultar o Capitulo 1, Seção 1.5 de Johnson e Wichern (2007).

A distância euclidiana entre observações p-dimensionais pode ser definida da seguinte forma: sejam $\boldsymbol{x}' = [x_1, x_2, \cdots, x_p]$ e $\boldsymbol{y}' = [y_1, y_2, \cdots, y_p]$ conjuntos de observações de dimensão p, a distância euclidiana entre os dois é dada por

$$\begin{aligned} d(\boldsymbol{x}, \boldsymbol{y}) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \\ &= \sqrt{(\boldsymbol{x} - \boldsymbol{y})'(\boldsymbol{x} - \boldsymbol{y})}, \end{aligned} \tag{24}$$

estatisticamente a distância entre esses dois conjuntos de observações é dada por

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{y})}, \tag{25}$$

em que $\boldsymbol{S}$ contem as variâncias e covariâncias amostrais. Contudo, sem o conhecimento prévio dos grupos distintos, estas quantidades amostrais não podem ser calculadas.

Uma outra medida de distância é a métrica de Minkowski, dada por

$$d(\boldsymbol{x},\boldsymbol{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}, \qquad m \neq 0, \tag{26}$$

para $m = 1$, $d(\boldsymbol{x},\boldsymbol{y})$ mede a distância "city-block" (quarteirão) entre dois pontos em p dimensões. Para $m = 2$, $d(\boldsymbol{x},\boldsymbol{y})$ torna-se a distância euclidiana. No geral, ao aumentarmos o valor de m alteramos o peso dado a diferenças grandes e pequenas.

Ambas medidas a seguir são definidas apenas para variáveis não negativas. Sendo elas, a medida de Canberra e o coeficiente de Czekanowski, apresentadas por

$$\begin{array}{rl} Medida\ de\ Canberra: & d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)} \\ Coeficiente\ de\ Czekanowski: & d(\boldsymbol{x},\boldsymbol{y}) = 1 - \frac{2\sum_{i=1}^{p} min(x_i, y_i)}{\sum_{i=1}^{p}(x_i + y_i)}. \end{array} \tag{27}$$

Sempre que possível, é aconselhável usar distâncias que satisfaçam as propriedades de distâncias encontradas no Capítulo 1 de Johnson e Wichern (2007). Embora a maioria dos algoritmos de agrupamentos aceitam diversas medidas de distâncias há algumas que subjetivamente não são aceitas, por exemplo, a desigualdade triângular.

Quando os itens não puderem ser representados por medidas p-dimensionais significativas, os pares de itens são frequentemente comparados com base na presença ou ausência de certas características. Itens semelhantes tendem a ter mais características em comum do que itens diferentes.

A presença ou ausência de uma característica pode ser descrita matematicamente introduzindo uma variável binária, em que assume-se o valor 1 para referie-se a existência da presença da característica e 0 para refere-se a ausência da característica. Por exemplo, para $p = 5$ as "pontuações" sobre as características podem ser feitas da forma apresentadas na **Tabela** 1, na qual encontram-se duas compatibiliade $1-1$, uma compatibilidade $0-0$ e duas descompatibilidade, sendo uma $0-1$ e a outra $1-0$.

Tabela 1. Tabela de pontuações para $p = 5$

	Variáveis				
	1	2	3	4	5
Item i	1	0	0	1	1
Item k	1	1	0	1	0

Sejam x_{ij} a pontuação (0 ou 1) da i-ésima variável binária no i-ésimo item, x_{kj} a pontuação (0 ou 1) da j-ésima variável binária no k-ésimo item, $j = 1, 2, \cdots, p$. Consequentemente,

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & se \quad x_{ij} = x_{kj} = 1 \quad ou \quad x_{ij} = x_{kj} = 0 \\ 1 & se \quad x_{ij} \neq x_{kj}, \end{cases} \tag{28}$$

e a distância euclidiana, $\sum_{j=1}^{p}(x_{ij} - x_{kj})^2$, fornece uma contagem de incompatibilidades. Uma grande distância corresponde a muitas incompatibilidade, isto é, itens diferentes. No exemplo anterior, o escore da distância entre os itens i e k seria

$$\sum_{j=1}^{5}(x_{ij} - x_{kj})^2 = (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 = 2. \tag{29}$$

A distância dada na equação (28) sofre com uma ponderação igual para os pares $1-1$ e $0-0$. Em alguns casos, um jogo $1-1$ é uma indicação mais forte de semelhança do que um jogo $0-0$. Por exemplo, ao agrupar pessoas do departamento de estatística da ESALQ, a evidência de que duas pessoas sejam formadas em estatística é mais forte do que pessoas formadas em agronomia, ou seja, ser formado em estatística é uma evidência mais forte do que não ser formado em estatística, ou seja, há evidências mais fortes na presença do que na ausência. Dessa forma, pode ser razoável desconsiderar alguns pares $0-0$ ou até mesmo desconsiderá-los por completo. Para permitir o tratamento diferenciado dos pares $1-1$ ou $0-0$, existem vários esquemas que podem definir o coeficiente de similaridade.

Organizando as frequências de compatibilidades e incompatibilidades para os itens i e k na forma de tabela de contingência:

Tabela 2. Tabela da frequência de compatibilidades e incompatibilidades para itens.

i/k	0	1	Totais
0	a	b	a+b
1	c	d	c+d
Totais	a+c	b+d	p=a+b+c+d

Na **Tabela** 2, p refere-se ao número de dimensões, a representa o número de pares $0-0$, b representa o números de pares $0-1$, e assim sucessivamente. Pegando o exemplo anterior, temos que $d = 2$, $a = b = c = 1$ e $p = 5$.

A próxima tabela apresenta alguns dos coeficientes de similaridade mais comuns definidos em termos das frequências da tabela anterior:

Para visualizar a tabela completa consultar o Capítulo 12 página 675 de (JOHNSON; WICHERN, 2007). Os coeficientes 1, 2 e 3 da tabela estão relacionados monoticamente. A monotonicidade é importante, por que alguns procedimentos de agrupamento não são afetados se a definição de similaridade for alterada de uma maneira que deixe inalterada a ordem relativa das semelhanças.

Tabela 3. Tabela de coeficientes de similaridade para agrupamento de itens.

Coeficientes	Razões
$\frac{a+d}{p}$	Quando os pesos dos pares 0-0 e 1-1 forem iguais
$\frac{2(a+d)}{2(a+d)+b+c}$	Quando os pesos dos pares 0-0 e 1-1 forem o dobro dos outros
$\frac{a+d}{a+d+2(b+c)}$	Quando o peso dos pares incomparaveis forem o dobro
$\frac{a}{p}$	Nenhuma correspondência de pares 0-0 no numerador

2.4.2 Associação e medidas de similaridade para pares de variáveis

Medidas de similaridade para variáveis geralmente assumem a forma de correlações amostrais e são substituidas pelos seus valores absolutos. Quando as variáveis são binárias, os dados podem novamente serem organizados em forma de uma tabela de contigência. Desta vez, são as variáveis que delineiam as categorias. Para cada par de variáveis, existem n itens categorizados na **Tabela** 4. A tabela é dada da seguinte forma:

Tabela 4. Tabela de coeficientes de similaridade para agrupamento de variáveis.

Variável i/ Variável k	0	1	Totais
0	a	b	a+b
1	c	d	c+d
Totais	a+c	b+d	n=a+b+c+d

em que n é o número de itens existente nas variáveis, e por exemplo, b representa o número de itens em que a variável i é igual 0 e a variável k é igual a 1. A fórmula usual de correlação aplicada às variáveis binárias na tabela de contingência é dada por

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \tag{30}$$

podendo ser considerado como uma medida de similaridade entre duas variáveis. O coeficiente r está relacionado com estatística qui-quadrado ($r^2 = \chi^2/n$) para testar a independência de duas variáveis categóricas, para n fixo, uma grande similaridade (ou correlação) é constante com a presença de dependência.

2.5 Métodos de Agrupamentos Hierárquicos

Os algoritmos hierárquicos criam uma hierarquia de relacionamentos entre os elementos. São muitos populares, apesar de não terem nenhuma justificativa teórica baseada em estatística ou teoria da informação, comenta Linden (2009). Segundo Murtagh e Contreras (2012), os algoritmos hierárquicos são uma sequência de etapas irreversíveis usadas para construir uma estrutura de dados desejada.

As técnicas de agrupamentos hierárquicos procedem de uma série de fusões ou divisões sucessivas. Esses métodos começam os agrupamentos com itens indiviuais, dessa

forma tem a existncia incialmente iguais para grupos e itens. Os itens que mais se assemelham são agrupados nesses grupos iniciais de acordo com suas similaridades. Eventualmente, à medida que a similaridade diminui, todos os subgrupos são fundidos em um único agrupamento.

Os métodos hierárquicos que trabalham com uma série de divisões funcionam de forma oposta. Os objetos são dividos em dois grupos inicias de modo que os objetos em grupo se "distancie" dos objetos presentes no outro. Então, os dois grupos iniciais são divididos novamente em subgrupos diferentes, o processo continua até que haja tantos subgrupos quanto objetos, isto é, até que cada objeto forme um grupo. Os resultados dos métodos aglomerativos e divisivos podem ser exibidos em um dendrograma, ilustrando as fusões ou divisões que foram feitas em níveis sucessivos.

2.5.1 Métodos de ligação

Métodos de ligação são adequados para agrupar itens e também variáveis, mas não são tão adequados para alguns métodos hierárquicos aglomerativos. O algoritmo para métodos de agrupamentos hierárquicos funcionam da seguinte forma:

(i)- Começa criando N grupos, cada um contendo uma única entidade e uma matriz simétrica $N \times N$ de distâncias (ou similaridades) $\boldsymbol{D} = \{d_{ik}\}$;

(ii)- Procura na matriz de distâncias o par de grupos mais próximo (mais semelhante), por exemplo (U) e (V);

(iii)- Combina os dois grupos que foram selecionados, nomeia esse novo grupo (UV) e atualiza a matriz de distâncias, removendo as linhas e colunas referentes aos grupos (U) e (V) utilizados na formação do novo grupo, e em seguida, adiciona uma linha e coluna fornecendo as distâncias entre esse novo grupo e os grupos remanescentes;

(iv)- Repete os itens (ii) e (iii) $N - 1$ vezes até que todos os objetos estejam em um único grupo. Registra as identidades e os níveis (distâncias ou similaridades) dos grupos que foram combinados.

2.5.2 Ligação simples

As entradas para os algoritmos de ligação simples podem ser as distâncias ou as semelhanças entre pares de objetos. Os grupos são formados a partir dos objetos individuais pela junção de vizinhos mais próximos, em que o termo vizinho mais próximo denota a menor distância ou maior semelhança.

Primeiramente, devemos encontrar a menor distância em $\boldsymbol{D} = \{d_{ik}\}$ e juntar com o objeto correspondente, por exemplo, U e V, obtendo o grupo (UV). Para o passo iii do algoritmo geral, as distâncias entre (UV) e o um outro grupo W são calculadas por

$$d_{(UV)W} = min\{d_{UW}, d_{VW}\}, \tag{31}$$

em que as quantidades d_{UW} e d_{VW} são as distâncias entre os vizinhos mais próximos dos grupos U e W e dos grupos V e W, respectivamente.

2.5.3 Ligação completa

O método da ligação completa funciona de forma semelhante ao de ligação simples, porém com uma importante exceção: em cada estágio, a distância (semelhança) entre os grupos é determinada pela distância (semelhança) entre os dois elementos, um de cada grupo, que sejam mais distante. Dessa forma, a ligação completa garante que todos os itens de um grupo estejam dentro da distância máxima (ou menor semelhança) uns dos outros.

O algoritmo geral começa novamente encontrando a entrada mínima em $\boldsymbol{D} = \{d_{ik}\}$ e juntando com os objetos correspondentes como exemplo, U e V, para criar o grupo (UV). Para o passo iii do algoritmo geral, as distâncias entre (UV) e qualquer outro grupo W são dadas por

$$d_{(UV)W} = max\{d_{UW}, d_{UW}\}, \tag{32}$$

em que d_{UW} e d_{VW} representam as distâncias entre os membros mais distantes dos grupos U e W e dos grupos V e W, respectivamente.

2.5.4 Ligação média

O método da ligação média trata a distância entre dois grupos como a distância média entre todos os pares de itens de forma que um membro de um par pertença a cada um dos grupos.

O algoritmo de ligação média funciona da mesma maneira do algoritmo geral. Começa buscando na martiz de distâncias $\boldsymbol{D} = \{d_{ik}\}$ os objetos mais próximos (mais semelhantes), por exemplo, U e V formando o grupo (UV). Para o passo iii do algoritmo geral, as distâncias entre (UV) e qualquer outro grupo W são determinadas por

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}, \tag{33}$$

em que d_{ik} é a distância entre o objeto i no grupo (UV) e o objeto k no grupo W, e $N_{(UV)}$ e N_W são o número de itens grupo (UV) e W, respectivamente. Para mais informações sobre os métodos de agrupamentos hierárquicos consultar o Capítulo 12 do livro de Johnson e Wichern (2007).

2.6 Métodos de agrupamento não hierárquico

Técnicas de agrupamentos não hierárquicas foram projetadas para agrupar itens, em vez de variáveis, em um conjunto de K grupos. Segundo Giordani et al. (2020), os métodos de agrupamento não hierárquicos exigem que o usuário especifique antecipadamente o número de clusters. O número de clusters K pode ser especificado ou determinado como parte dos procedimentos do agrupamento. Nesse caso, uma matriz de distâncias (semelhanças) não precisa ser determinada e os dados não precisam ser armazenados durante a execução, esses métodos podem ser aplicados a conjunto de dados muito maiores do que os métodos hierárquicos.

Os métodos não hierárquicos partem de uma partição inicial de itens em grupos de um conjunto inicial de pontos, que vem dos núcleos dos grupos. Boas escolhas para as configurações inicias devem ser livres, podendo-se iniciar selecionando aleatoriamente os pontos iniciais entre os itens ou particionar aleatoriamente os itens em grupos iniciais.

2.6.1 Método do k-médias

O método do k-médias que foi incialmente proposto por MacQueen et al. (1967), é um algorítmo que particiona as observações em K grupos, cada partição é denomida de cluster e o número de cluster deve ser considerado a priori. Trata-se de um método de agrupamento não-hierárquico, métodos de agrupamentos não-hierárquicos podem ser aplicados a conjuntos de dados de elevada dimensionalidade ou cardinalidade sem afetar grandemente a sua eficiência computacional (OLIVEIRA, 2021).

Utilizado por ARAUJO (2017) com o objetivo de determinar a característica de resistência à verminose, a partir das combinações das características FAMACHA, OPG, HCT, plaquetas (PLT), hemoglobina copuscular médio (MCHC), dentre outras. Os animais do grupo em que os indivíduos apresentaram menores valores para o OPG e FAMACHA e maiores valores para as características sanguíneas e moformétricas foram classificados como resistêntes. Os animais com maiores números para o OPG e FAMACHA e menores valores para as características sanguíneas e morfométricas foram classificados como suscetíveis e os animais que apresentaram valores entre os dois grupos (resistente e suscetível) foram classificados como de resistência intermediária.

Objetivando identificar padrões na produção de leite nos municípios do estado de São Paulo, considerando diferentes tipos HESPANHOL, PELOZO e PEREIRA (2016) utilizaram o algoritmo do k-médias para o reconhecimento desses padrões. Assim, foi possível obter, robustamente, os padrões de produção de leite do estado de São Paulo.

O método do k-médias cria para cada grupo centróides (médias) relacionados as variáveis, dessa forma o algoritmo atribui cada observação ao grupo que possui o valor do centróide mais próximo do valor da observação. Resumidamente, o processo é composto por três etapas:

(i) - Divide os itens em k grupos iniciais;

(ii) - Atribui um item ao grupo cujo o centróide é mais próximo. A distância é geralmente calculada usando a distância euclidiana com observações padronizadas ou não padronizadas (embora situações onde utilizar observações não padronizadas sejam raras). Calcula novamente o centróide para o grupo que recebe o novo item e para o grupo que perdeu o item;

(iii) - Repete a etapa (ii) até que todos os itens estejam alocados em algum grupo.

Em vez de começar com uma partição inicial de todos os itens em k grupos preliminares na etapa (i), informamos os k centróides iniciais (pontos de sementes) e então prosseguimos para a etapa (ii). A atribuição final dos itens aos grupos dependerá, até certo ponto, da partição inicial ou da seleção inicial dos pontos iniciais. A sugestão é que a maioria das mudanças importantes nas atribuições ocorra na primeira etapa da realocação.

2.6.2 Método do k-médias e o critério da soma de quadrados para agrupamentos

O método do k-médias é baseado historicamente no critério da soma de quadrados para agrupamentos, sendo esse critério divido em duas versões, uma versão discreta e outra contínua.

Critério da soma de quadrados discreto: Sejam n pontos de dados $x_1, x_2, \cdots, x_n$ em $\mathbb{R}^p$ e k-partições, $\mathscr{C} = (C_1, C_2, \cdots, C_k)$, do conjunto $\mathscr{O} = \{1, 2, \cdots, n\}$ de objetos subjacentes de classes não vazias $C_i \subset \mathscr{O}$, o critério da soma de quadrado discreto (também conhecido como: critério de variância, inércia ou critério de traço) é dado por

$$g_n(\mathscr{C}) = \sum_{i=1}^{k} \sum_{l \in C_i} ||x_l - \bar{x}_{C_i}||^2, \tag{34}$$

em que $\bar{x}_{C_i}$ denota o centróide dos pontos de dados x_l pertencentes a classe C_i, ou seja, com $l \in C_i$. Procuramos uma partição k de $\mathscr{O}$ com o menor valor do critério $g_n(\mathscr{C})$. O problema da otimização de um parâmetro está relacionado, e até equivalente, ao problema da otimização com dois parâmetros

$$g_n(\mathscr{C}, \mathscr{Z}) = \sum_{i=1}^{k} \sum_{k \in C_i} ||x_l - z_i||^2, \tag{35}$$

em que a minimização também é válida para todo o sistema $\mathscr{Z} = (z_1, \cdots, z_k)$ de k pontos $z_1, \cdots, z_k$ de $\mathbb{R}^p$.

Teorema 1:

(i) Para qualquer partição k fixa $\mathscr{C}$, o critério $g_n(\mathscr{C}, \mathscr{Z})$ é particulamente minimizado em relação ao valor Z pelo sistema de centróides de classe $\mathscr{Z}^* = (\bar{x}_{C_1}, \cdots, \bar{x}_{C_k}) = \mathscr{Z}(\mathscr{C})$:

$$g_n(\mathscr{C}, \mathscr{Z}) \geq g_n(\mathscr{C}, \mathscr{Z}^*) = \sum_{i=1}^{k} \sum_{k \in C_i} ||x_k - \bar{x}_{C_i}||^2 = g_n(\mathscr{C}), \quad \text{para todo} \mathscr{Z}. \tag{36}$$

(ii) Para qualquer sistema protótipo fixo $\mathscr{Z}$, o critério $g_n(\mathscr{C}, \mathscr{Z})$ é particulamente minimizado em relação ao valor C por qualquer partição de distância mínima $\mathscr{C}^* = (C_1^*, \cdots, C_k^*) = \mathscr{C}(\mathscr{Z})$ induzida por $\mathscr{Z}$, ou seja, com as classes dadas por $C_i^* = \{l \in \mathscr{O} | d(x_l, z_i) = min_{j=1,\cdots,k} d(x_l, z_j)\}$, $i = 1, \cdots, n$, em que $d(x, z) = ||x - z||^2$ é a distância euclidiana quadrada:

$$g_n(\mathscr{C}, \mathscr{Z}) \geq g_n(\mathscr{C}^*, \mathscr{Z}) = \sum_{l=1}^{n} min_{j=1,\cdots,k}\{||x_l - z_j||^2\} \quad \text{para todo} \mathscr{C}. \tag{37}$$

O algoritmo do k-médias tenta aproximar uma partição k ótima iterando as etapas de minimização parcial (i) e (ii) do teorema 1. Prosseguindo da seguinte forma:

- Começa com um sistema protótipo arbritário $\mathscr{Z}^{(0)} = (z_1^{(0)}, \cdots, z_k^{(0)})$, sendo a primeira iteração $t = 0$.

- $t \to t + 1$

  (i) Minimiza o critério $g_n(\mathscr{C}, \mathscr{Z}^{(t)})$ com relação a partição k de $\mathscr{C}$, ou seja, determina a partição de distância mínima $\mathscr{C}^{(t+1)} = \mathscr{C}(\mathscr{Z}^{(t)})$.

  (ii) Minimiza o critério $g_n(\mathscr{C}^{(t+1)}, \mathscr{Z})$ em relação ao sistema protótipo $\mathscr{Z}$, ou seja, calcula o sistema de classes dos centróides $\mathscr{Z}^{(t+1)} = \mathscr{Z}(\mathscr{C}^{(t+1)})$.

- Para quando as etapas (i) e (ii) estiverem na estacionariedade.

Por construção, este algoritmo produz uma sequência $\mathscr{Z}^{(0)}, \mathscr{C}^{(1)}, \mathscr{Z}^{(1)}, \mathscr{C}^{(2)}, \cdots$ de protótipos e partições com valores decrescente dos critérios que convergem para um valor mínimo, tipicamente local.

Critério da soma de quadrados contínuo: Considere $x_1, \cdots, x_n$ sendo n realizações de um vetor aleatório $\boldsymbol{X}$ com distribuição P em $\mathbb{R}^p$, assim de forma análoga ao critério de soma de quadrados discreto, podemos procurar uma k partição $\mathscr{B} = (B_1, \cdots, B_k)$ em $\mathbb{R}^p$ com valor mínimo dado por

$$g(\mathscr{B}) = \sum_{i=1}^{k} \int_{B_i} ||x - \mathbb{E}[X|X \in B_i]||^2 dP(x). \tag{38}$$

Da mesma forma, podemos relacionar a equação anterior com o problema de otimização para dois parâmetros

$$g(\mathscr{B}, \mathscr{Z}) = \sum_{i=1}^{k} \int_{B_i} ||x - z_i||^2 dP(x), \tag{39}$$

e de forma análoga temos o seguinte teorema.

Teorema 2:

(i)- Para qualquer k partição fixa de $\mathscr{B}$ de $\mathbb{R}^p$, o critério $g(\mathscr{B}, \mathscr{Z})$ é particulamente minimizado em relação a $\mathscr{Z}$ pelo sistema protótipo $\mathscr{Z}^* = (Z_1^*, \cdots, Z_k^*) = \mathscr{Z}(\mathscr{B})$ dados pelas esperanças condicionais $z_i^* = \mathbb{E}[X|X \in B_i]$

$$g(\mathscr{B}, \mathscr{Z}) \geq g(\mathscr{B}, \mathscr{Z}^*) = \sum_{i=1}^{k} \int_{B_i} ||x - E[X|X \in B_i]||^2 dP(x) = g(\mathscr{B}) \textit{ para todo } \mathscr{Z}. \tag{40}$$

(ii)- Para qualquer protótipo fixo $\mathscr{Z}$, o critério $g(\mathscr{B}, \mathscr{Z})$é particulamente minimizado em relação a $\mathscr{B}$ por qualquer partição de distância mínima $\mathscr{B}^* = (B_1^*, \cdots, B_k^*) = \mathscr{B}(\mathscr{Z})$ geradas por $\mathscr{Z}$, ou seja, fornecendo as classes $B_i^* = \{x \in \mathbb{R}^p | d(x, z_i) = min_{j=1,\cdots,k}\{d(x, z_j)\}, i = 1, \cdots, n\}$, dadas por

$$g(\mathscr{B}, \mathscr{Z}) \geq g(\mathscr{B}^*, \mathscr{Z}) = \int_X min_{j=1,\cdots,k} ||x - z_j||^2 dP(x) \text{para todo} \mathscr{B}. \tag{41}$$

Esse teorema formula, e justifica, uma versão contínua do algoritmo do k-médias.

2.6.3 Método do k-médias generalizado

Os criterios da soma de quadrados de dois parâmetros para agrupamentos foram generalizados de diversas formas para atender a tipos de dados especiais ou algumas propriedades de agrupamentos. No caso discreto, temos

$$g_n(\mathscr{C}, \mathscr{Z}) = \sum_{i=1}^{k} \sum_{l \in C_i} d(l, z_i), \tag{42}$$

em que $d(l, z_i)$ mede a dissimilaridade entre um objeto l e uma protótipo de classe z. Há muita flexibilidade nessa abordagem, uma vez que

- não se tem muitas restrições quanto ao tipo de dados subjacentes (dados quantitativos, categóricos, etc)
- muitas maneiras de especificar uma família $\mathscr{P}$ de protótipos de classe z apropriados ou admissíveis para apresentar aspectos específicos dos grupos (pontos, hiperespaço em $\mathbb{R}^p$, subconjuntos de $\mathscr{O}$, relações de oredem)

- uma ampla gama de possibilidades para escolher a medida de dissimilaridade d, e ainda, adicionar ou introduzir pesos w_l para os objetos $l \in \mathscr{O}$.

Em todos os casos, o algoritmo do k-médias generalizado pode ser aplicado para atingir uma configuração ótima (local ou globalmente):

- Inicia com um sistema de protótipos arbitrários $\mathscr{Z}^{(0)} = z_1^{(0)}, \cdots, z_k^{(0)}$, com $t = 0$;
- Para $t = t + 1$

  (i)- Minimiza o critério $g_n(\mathscr{C}, \mathscr{Z}^{(t)})$ em relação uma partição k em $\mathscr{C}$ de $\mathscr{P}$. Produzindo uma partição de distância mínima $\mathscr{C}^{(t+1)} = \mathscr{C}(\mathscr{Z})^{(t)}$ com k classes $\mathscr{C}_i^{(t+1)} = \{l \in \mathscr{O} | d(l, z_i^{(t)}) = min_{j=1,\cdots,k} d(l, z_j^{(t)})\}$.

  (ii)- Minimiza o critério $g_n(\mathscr{C}^{(t+1)}, \mathscr{Z})$ em relação ao sistema de protótipos $\mathscr{Z}$. Consequentemente, isso equivale a determinar, para cada classe $C_i = C_i^{(t+1)}$, uma configuração $z_i^{(t+1)}$ mais típica no sentido de:

$$\mathscr{Q}(C_i, z) = \sum_{l \in C_i} d(l, z). \tag{43}$$

- Para as iterações quando todos os itens forem alocados nos grupos.

Para mais detalhes consultar Johnson e Wichern (2007), Diday (2007) e Rencher e Christensen (2002).

2.7 Classificação dos grupos resistentes, resilientes e suscetíveis

A ideia do k-médias é resumidamente separar um conjunto de dados em grupos, denominados de *clusters*. Entretanto, a classificação desses *clusters* é feita por uma análise exploratória detalhada dos dados.

Na literatura existem estudos que já envolveram alguns das variáveis aqui utilizados, como nos trabalhos de VIEIRA et al. (2009) e Oliveira (2021). Neste caso, temos a contagem de ovos por grama de fezes, porcentagem do teste do hematócrito e o ganho de peso diário do animal, o ganho de peso foi obtido de duas formas, considerando a variável peso e data da coleta presente no banco de dados. A primeira forma foi considerando o ganho de peso do indivíduo ao longo do tempo, a segunda considerando a primeira e a última mensuração.

O grupo dos animais resistentes deverá ter, com relação ao ganho de peso diário (gp), um ganho razoável de peso que indique que o animal não esteja de certa forma afetado pela a infecção por nematoides. Parente et al. (2009) estudaram diferentes tipos de dietas, visando o desempenho produtivo de ovinos em confinamento em um experimento inteiramente causalizado. O ganho médio diário para as diferentes dietas ultrapassou

150amml/dia, baseado nesse estudo essa marca irá caracterizar para a variável ganho de peso animais do grupo resistente. Indivíduos que não conseguiram ganhar mais do que 20g diárias será considerado do grupo suscetivél, entre esses valores os animais resilientes.

No trabalho de Oliveira (2021) o qual teve como base o estudo de Coutinho (2012) que classificou os animais resistentes aqueles que tiveram uma média menor ou igual a 763 OPG, e os animais suscetivéis os animais que tiveram a média do número de ovos por grama de fezes superior a 3.631 OPG. Alguns trabalhos fazem a classificação de forma um pouco diferente, Chagas, Carvalho e Molento (2007) classificaram os níveis da infeccção da seguinte forma, se a média de OPG for menor que 500 a infecção é considerada leve, entre 500 a 1.500 é a infecção é considerada moderada, de 1.501 à 3.000 caracterizada como pesada e se for maior que 3.000 a infecção é dada como fatal.

No artigo de Chagas, Carvalho e Molento (2007), ele relaciona o método FAMACHA com a coloração da conjuntiva ocular e o hematócrito. Nesse caso, o interesse específico é se basear na porcentagem do teste do hematócrito, logo o animal que teve porcentagem do hematócrito igual ou superior a 23% não seria tratado, sendo inferior a 23% o animal deveria ser tratado. Assim, a classificação dos grupos resilientes, resistentes e sucetivéis será feita a partir da combinação desses critérios.

3 MATERIAL E MÉTODOS

3.1 Descrição dos dados

O conjunto de dados diz respeito a ovinos da raça Santa Inês, provenientes do Laboratório de Nutrição Animal do Centro de Energia Nuclear na Agricultura (CENA/USP). Esses dados não foram publicados anteriormente a este trabalho, organizados em séries temporais para cada animal, as medições de peso, número de ovos por grama de fezes e os valores do hematócrito foram coletadas mensalmente ao longo do período de fevereiro de 2017 a dezembro de 2022.

Ao considerar as variáveis relacionadas à data da coleta e ao peso dos animais, adicionou-se uma nova variável aos dados: o ganho de peso. Esse ganho de peso foi calculado de duas maneiras distintas, resultando em duas novas variáveis denominadas ganho de peso I (GPI) e ganho de peso II (GPII).

O cálculo do ganho de peso I foi realizado considerando a evolução do animal ao longo do estudo, levando em conta se houve aumento ou diminuição de peso de um mês para o outro. Já o ganho de peso II foi determinado pela diferença entre as medidas da última e da primeira mensuração. Após a obtenção dessas variáveis, foram calculadas as médias para cada indivíduo e para cada classe de animais. Posteriormente, os resultados relativos às variáveis GPI e GPII foram normalizados pela quantidade de dias em que cada animal esteve envolvido no estudo, proporcionando assim o ganho de peso diário para cada animal.

Inicialmente, as classes dos animais foram designadas como matrizes (fêmeas selecionadas para reprodução), reprodutores (machos selecionados para reprodução), machos adultos, fêmeas adultas, animais adultos não destinados à reprodução, borregos e borregas, além de filhotes fêmeas e machos. No entanto, notou-se uma escassez de animais nas categorias de machos adultos, fêmeas adultas, borregos e borregas.

Diante disso, houve uma reorganização das classes, consolidando os animais das categorias reprodutores, machos adultos e borregos na classe de machos adultos, enquanto as fêmeas adultas, matrizes e borregas foram agrupadas exclusivamente como matrizes. Portanto, as classes finais ficaram definidas como matrizes, machos adultos e filhotes machos e fêmeas.

3.2 Aspectos computacionais

Toda a análise, incluindo a análise descritiva, a análise de componentes principais e a análise de agrupamentos, foi conduzida utilizando o software (R Core Team, 2023). No entanto, para cada uma dessas análises, foram empregados pacotes distintos. Alguns desses pacotes já estão integrados à base do R, enquanto outros são componentes adicionais que podem ser instalados no software.

3.2.1 Descrição dos pacotes

Quanto ao método de k-médias, o algoritmo está incorporado à base do *R*. Por padrão, o comando segue o algoritmo desenvolvido por Hartigan e Wong (1979). No entanto, é necessário realizar algumas especificações, como a padronização das variáveis e qual o algoritmo será utilizado para que o comando aplique o método proposto por MacQueen et al. (1967).

O pacote *readODS* desenvolvido por Schutten et al. (2023) foi empregado para a leitura dos dados no formato *ods*. Em relação à criação de gráficos, foram utilizados os pacotes *lawstat*, *ggplot2* e *factoextra*, elaborados por Gastwirth et al. (2023), Wickham et al. (2019) e Kassambara e Mundt (2020), respectivamente. Para funcionalidades adicionais, o pacote *MASS* de Venables e Ripley (2002) foi empregado em algumas funções, o pacote *dplyr* para a manipulação dos dados desenvolvido por Wickham et al. (2023). Além disso, os pacotes *stats* e *factominer*, desenvolvidos por R Core Team (2023) e Lê, Josse e Husson (2008), respectivamente, foram úteis no desenvolvimento da análise de componentes principais.

4 RESULTADOS E DISCUSSÕES

4.1 Análise descritiva

Em trabalhos que englobam alguma análise estatística, a primeira coisa que fazemos é uma prospecção dos dados, aqui não foi diferente. O banco de dados que foi utilizado neste trabalho passou por uma organização e lhe foi acrescentadas algumas variáveis. As variáveis acrescentadadas foram relacionadas ao ganho de peso dos animais considerando o que aconteceu ao longo do tempo e considerando somente a primeira e última mensuração. Na **Tabela** 5 encontram-se as médias das variáveis para cada classe e o número de animais presente em cada uma.

Tabela 5. Médias para as variáveis por classe dos animais.

Classes	Média do peso (kg)	Ganho de peso diário I (g)	Ganho de peso diário II (g)	Média do HT (%)	Média do OPG	Número de animais
Machos adultos	46,9	2,670	7,820	30,9	662	20
Matrizes	43,0	-0,368	1,240	32,9	931	45
Filhotes machos	22,7	1,460	8,160	33,9	445	17
Filhotes fêmeas	23,6	1,700	10,800	37,4	634	24

A classe das matrizes tive o menor ganho de peso em ambas as formas que foram obtidas. A média do OPG para essa classe foi superior a 500, junto com a classe dos filhotes e adultos machos, e todas as classes apresentaram em média valores dentro da faixa normal para a espécie no exame do hematócrito. As próximas tabelas apresentam um resumo detalhado para cada classe com os 5 animais que tiveram em média o maior número de ovos por grama de fezes.

A **Tabela** 6 apresenta os 5 animais da classe dos machos adultos que tiveram as maiores médias para OPG, as médias foram superior a 500 OPG, apresentando uma porcentagem do teste do hematócrito acima de 23% e um ganho de peso diário considerado baixo. Baseando-se no estudo de Chagas, Carvalho e Molento (2007) esses animais seriam caracterisados com uma infecção moderada e não seriam levados ao tratamento.

As 5 matrizes com as maiores médias para o OPG apresentadas na **Tabela** 7 seriam consideradas com um estágio grave da infecção, considerando a porcentagem do HT elas não seriam tradadas. Entretando, com a combinação das variáveis esses animas deveriam passar por tratamento, pois perderam peso e a média do OPG ultrapassou 1500, sendo que uma dessas matrizes chegou a ter em média 6000 OPG.

Os filhotes machos apresentados na **Tabela** 8 mostraram as menores médias para a variável OPG com base nos 5 animais apresentados, mesmo assim ainda seriam caracterizado com uma infecção moderada. Embora, o ganho de peso desses animais não foi tão baixo lenvando em consideração a média do OPG.

Tabela 6. Médias para as variáveis: os 5 animais com os maiores valores para o OPG, classe dos machos adultos.

Animal	Média do Peso (Kg)	Ganho de peso diário I (g)	Ganho de peso diário II (g)	Média da % do HT	Média do OPG	Tempo no estudo em meses
2026	29,2	7,296	21,890	30	2250	8
1417	76,34	0,0326	0,7182	37,57	1340	29
2036	19,33	12,935	12,935	26,5	1190	8
IZ-1564	74,99	0,5885	13,535	36,28	1097	29
2027	28,17	14,965	14,965	23,5	833	6

Tabela 7. Médias para as variáveis: os 5 animais com os maiores valores para o OPG, classe das matrizes.

Animal	Média do Peso (Kg)	Ganho de peso diário I (g)	Ganho de peso diário II (g)	Média da % do HT	Média do OPG	Tempo no estudo em meses
1812	39	-47,826	-47,826	32	6075	4
1735	40,77	-50,543	-50,543	34	2375	4
1723	46,77	-32,916	-32,916	36	1916	5
2103	21	0,7225	2,890	25	1900	7
1416	47,62	3,513	17,567	31	1822	10

Tabela 8. Médias para as variáveis: os 5 animais com os maiores valores para o OPG, classe dos filhotes machos.

Animal	Média do Peso (Kg)	Ganho de peso diário I (g)	Ganho de peso diário II (g)	Média da % do HT	Média do OPG	Tempo no estudo em meses
2012	10,55	3,296	6,593	28	1280	4
2004	28,57	1,638	14,750	32	950	10
2002	23,43	2,698	24,285	33	750	10
2026	24,17	2,876	25,892	33	621	10
2010	23,45	7,509	22,527	36	600	4

Tabela 9. Médias para as variáveis: os 5 animais com os maiores valores para o OPG, classe dos filhotes fêmeas.

Animal	Média do Peso (Kg)	Ganho de peso diário I (g)	Ganho de peso diário II (g)	Média da % do HT	Média do OPG	Tempo no estudo em meses
2009	22,2	11,363	56,818	38	1566	6
2003	31,36	6,428	57,857	35	1340	10
2011	24,32	6,593	39,560	35	1120	7
2032	17,24	11,525	46,103	40	1075	6
2008	11,95	14,010	28,021	28	1000	4

Os cinco filhotes fêmeas com as maiores médias do OPG conforme a **Tabela** 9, podem ser caracterizadas com uma infecção moderada. Entretanto, considerando o ganho

de peso baseado na primeira e última mensuração foi razoavelmente bom, dessa forma visivelmente antes de entrar no método de agrupamento poderiamos dizer que elas seriam consideradas animais com a característica de resiliencia.

Na **Figura** 3 observa-se a distribuição da variável ganho de peso considerando o que aconteceu com os animais ao decorrer do tempo para o GPI nas diferentes classes. A variável ganho de peso diário I está bem distribuida entre os valores $0g$ e $20g$ gramas por dia, para as classes dos machos adultos e dos filhotes machos e fêmeas. Embora que dois animais da classe dos machos adultos tiveram um ganho de diário acima de $50g$. Para as matrizes a variável encontra-se distribuida entre os valores $-5g$ e aproximadamente $0g$ com alguns pontos ultrapassando a marca de $-50g$ gramas por dia, três desses animais que mostraram uma perda peso estão exibidos na tabela 7, note que esse animais manifestaram uma média superior a 1500 OPG.

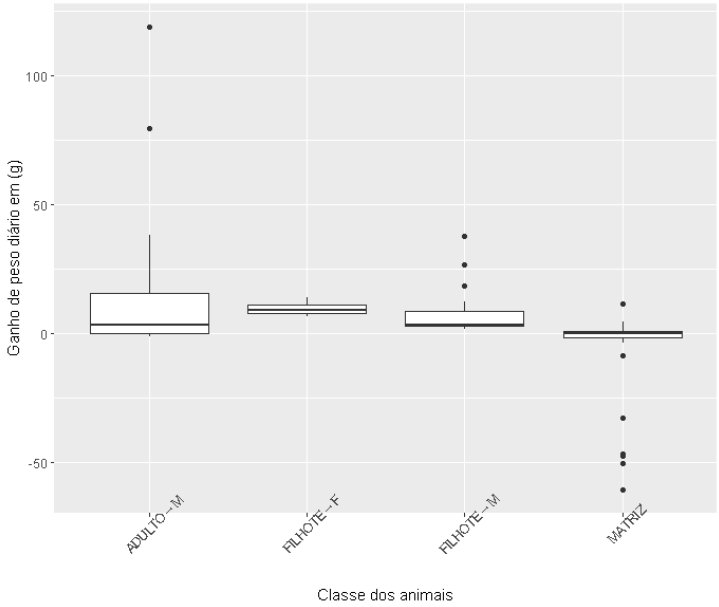


Figura 3. Distribuição do ganho de peso diário I nas diferentes classes

A **Figura** 4 mostra a relação da variável ganho de peso diário I com o peso médio nos animais, no gráfico fica mais visível o comportamento da variável ganho de peso diário I em torno do valor 0. Além disso, com relação a classe das matrizes esta variável encontra-se em torno de zero e valores negativos, como já mencionado as fêmeas em periodo de gestação são mais suscetíveis à infecção, o que poderia explicar os valores negativos para o ganho de peso de alguns animais presentes nessa classe, embora pode ter ocorrido das matrizes em questão terem passo por trabalho de parto ocasionando em uma perda de peso razoavelmente grande.

A **Figura** 5 apresenta a relação entre a variável ganho de peso diário I com o número médio de ovos por grama de fezes. No gráfico, é vísivel o mesmo comportamento que acontece na figura anterior. Entretanto, observando a classe das matrizes para valores de 0 a 1000 OPG, o ganho de peso diário está entre $-10g$ e $10g$, e para os animais que

tiveram uma perda de peso superior a $25g$ o OPG supera 1000 ovos por grama. Note que, há um ponto especifico na parte inferior direita no quadro das matrizes esse ponto é referente ao animal 1812. Durante os quatro meses que passou no estudo, este animal perdeu, em média, cerca de $47,82g$ de peso por dia. Durante esse tempo, foi observado que o animal tinha, em média, 6075 ovos por grama de fezes. No entanto, apesar desse alto número de OPG, é importante verificar se o animal teve filhotes durante esse período. Isso poderia ser outra razão para a grande perda de peso observada.

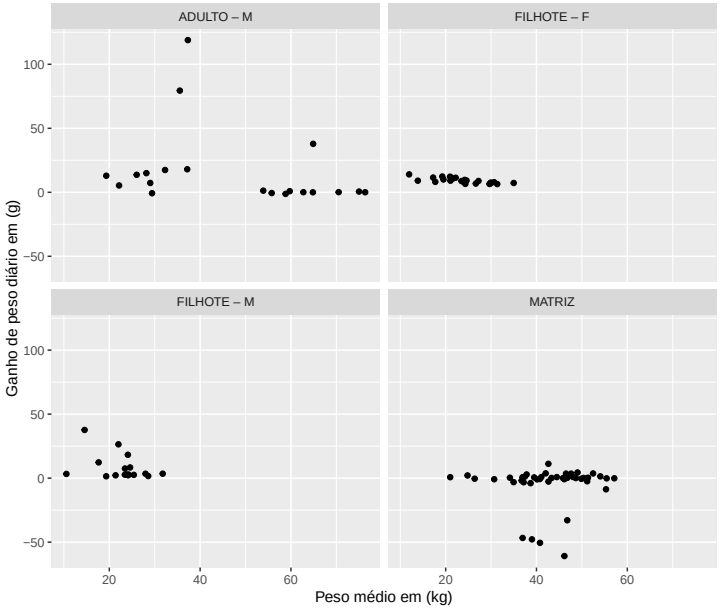


Figura 4. Dispersão entre o ganho de peso diário I e o peso médio dos animais.

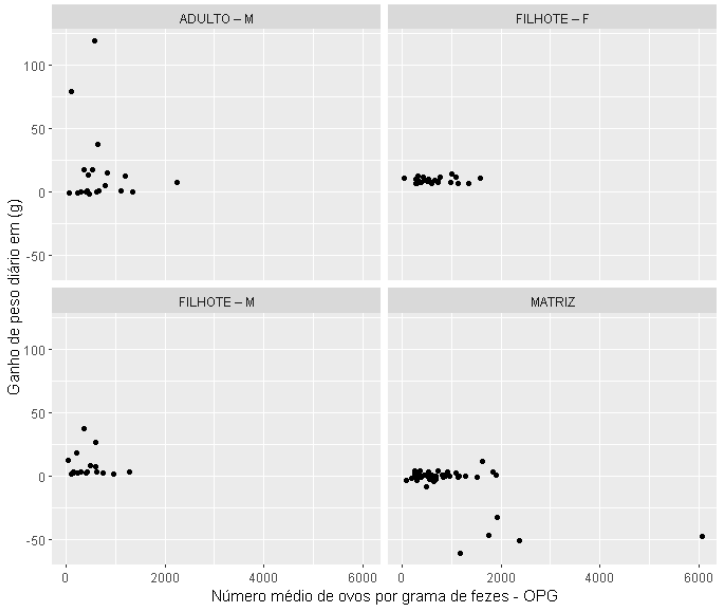


Figura 5. Dispersão entre o ganho de peso diário I e número médio de ovos por grama de fezes.

Em relação com a média do teste do hematócrito (%), todas as classes apresen-

taram valores dentro da faixa normal para a espécie, pois quase todos os animais tiveram uma média superior a 25%. Os animais da classe das matrizes que perderam peso durante o estudo também mostraram bons valores para o hematócrito, ou seja, com uma média superior a 25% basta observar os valores negativos no eixo y no quadro das matrizes apresentados no gráfico da **Figura** 6.

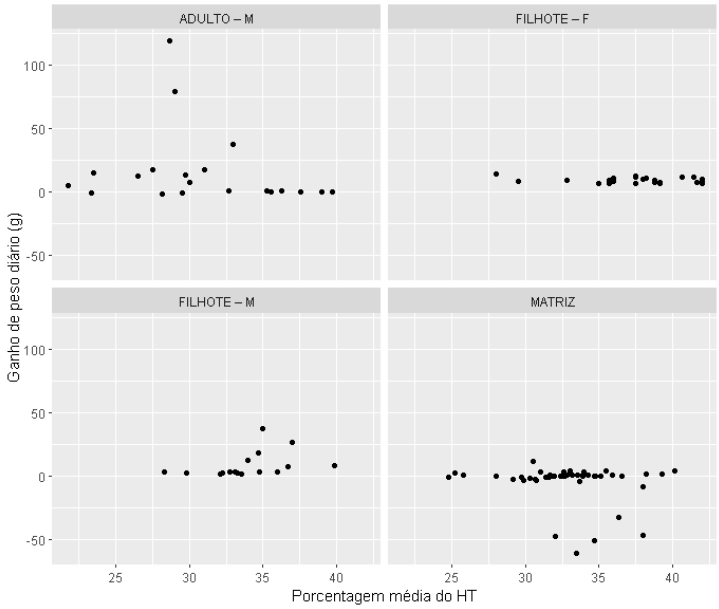


Figura 6. Dispersão entre o ganho de peso diário I e a porcentagem médio do teste do hematócrito.

As próximas **Figuras** 7, 8, 9, 10 serão em relação ao ganho de peso diário II, ou seja, o ganho de peso calculado considerando somente a primeira e ultima mensuração. A **Figura** 7 mostra a distribuição do GPII para cada classe. No gráfico, conseguimos ver que com o cálculo do ganho de pesso dessa forma, a classe dos filhotes fêmeas teve um melhor desempenho, as matrizes mudaram um pouco os valores, mas o cenário para essa classe não chegou a mudar, ainda é notável que para alguns indivíduos há uma perda de peso.

Na **Figura** 8 podemos ver a relação da variável ganho de peso diário II em função do peso médio. Ao separarmos pelas classes no gráfico, vemos que muitos animais da classe das matrizes tiveram uma perda de peso considerável. A classe dos filhotes fêmeas mostraram um bom desempenho, os animais machos adultos e filhotes machos não mostraram uma situação a gravente comparados com as matrizes.

Relacionando o ganho de peso diário II com o número de ovos por grama de fezes, para as classes dos machos adultos, filhotes machos e fêmeas poucos animais ultrapassaram 1000 OPG. Por outro lado, as matrizes muitos animais apresentaram uma média para o número de ovos por grama de fezes superior a 1000 OPG, como mostra a **Figura** 9.

Em função da porcentagem média do teste do hematócrito, como mostra a **Figura** 10. O que pode-se comentar, é que semelhante ao cenário com o ganho de peso diário I

as médias para o teste do hematócrito obtidas mostraram-se dentro dos valores normais para a espécie.

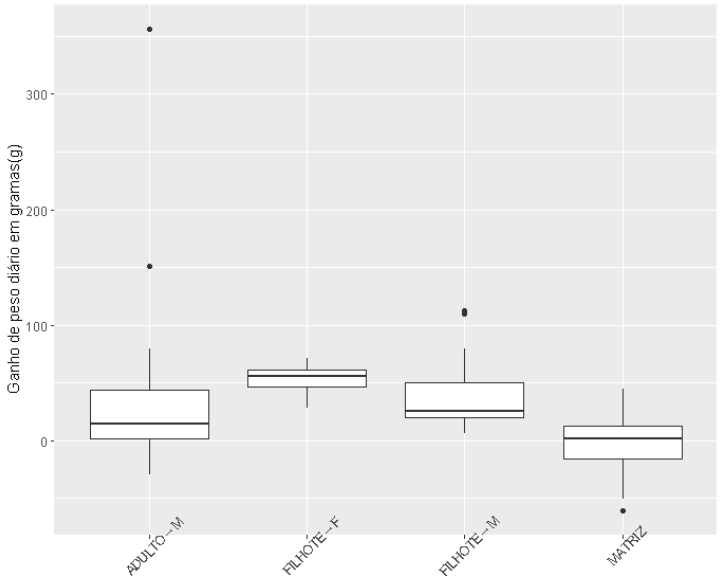


Figura 7. Distribuição do ganho de peso diário II nas diferentes classes de animais

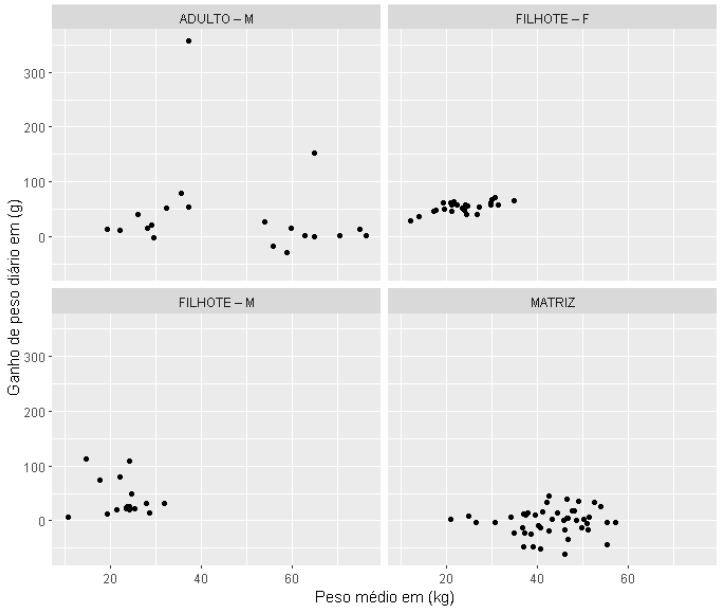


Figura 8. Dispersão entre o ganho de peso diário II e o peso médio dos animais.

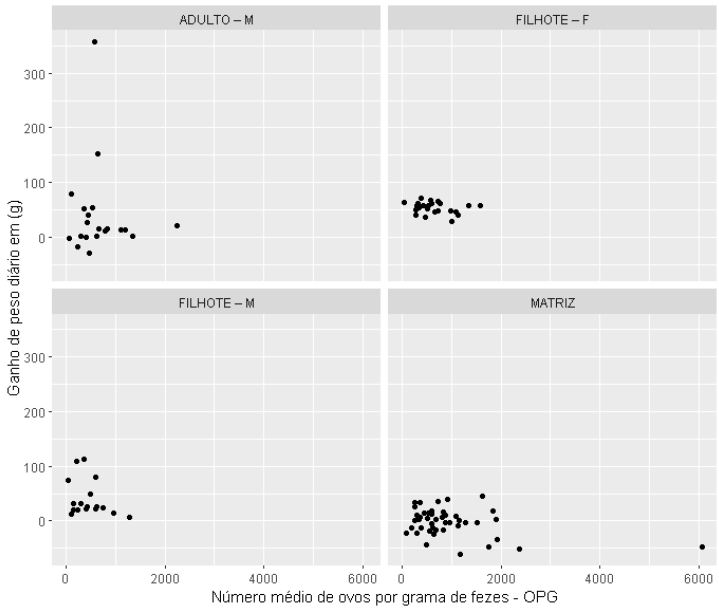


Figura 9. Dispersão entre o ganho de peso diário II e o número médio de ovos por grama de fezes.

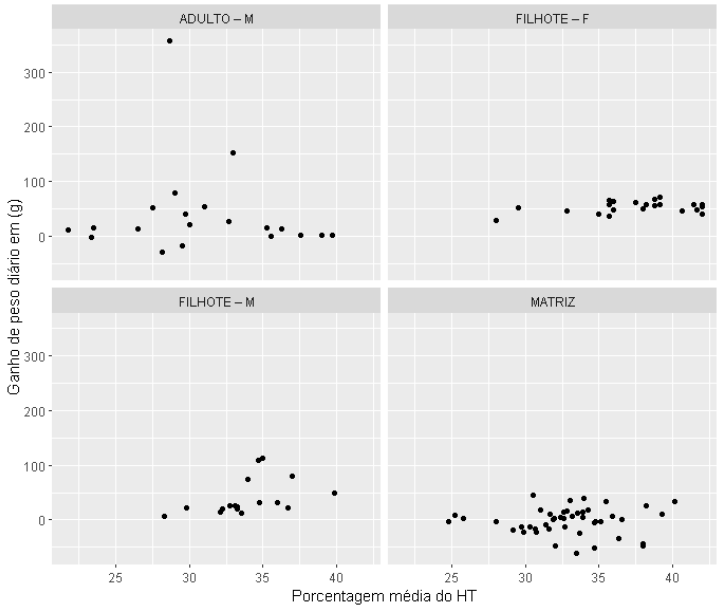


Figura 10. Dispersão entre o ganho de peso diário II e a porcentagem média do teste do hematócrito.

4.2 Análise de componentes principais

A análise de componentes principais nesse trabalho busca mostrar que as variáveis são não correlacionadas e que poderiamos usar os indicadores no método de agrupamento k-médias, embora exista correlação entre as variáveis e as componentes principais. Com a ACP poderíamos utilizar três componentes principais, como os dados foram separados em dois subconjuntos de dados, sendo um com a variável ganho de peso diário I (GPI) e o outro com a variável ganho de peso diário II (GPII), assim a análise será dividida em cenário I e cenário II.

A **Tabela** 10 retorna que as duas primeiras componentes principais explicam 81,9% da variabilidade total dos dados, na **Tabela** 11 as duas primeiras componentes explicam 75% da variabilidade total. Entretanto, mesmo que nos dois cenários as duas primeiras componentes pricipais expliquem mais de 70% da variabilidade, não há necessidade de utilizar as componentes principais como os indicadores.

Tabela 10. Análise de componentes principais para as variáveis do cenário I.

	PC I	PC II	PC III
Variância	1,40	1,05	0,54
% da variância	46,75	35,15	18,10
% acumulada	46,75	81,90	100,00

Tabela 11. Análise de componentes principais para as variáveis do cenário II.

	PC I	PC II	PC III
Variância	1,36	0,89	0,74
% da variância	45,36	29,79	24,85
% acumulada	45,36	75,15	100,00

As **Tabelas** 12 e 13 são referentes à correlação das variáveis com as componentes principais, veja que em ambas as tabelas as duas primeiras variáveis estão bem correlacionadas com a primeira componente e a segunda componente com a terceira variável. Dessa forma, temos que as variáveis GPI e a média do OPG foram importantes para que a primeira componente explicasse aproximadamente 46% da variabilidade total dos dados e a média do HT importante para a segunda componente principal para ACP no cenário I conforme mostra a **Tabela** 12. Já a ACP feita para o cenário II na **Tabela** 13, mostra que o GPII e a média do peso também foram importantes para que a primeira componente explicasse 45%, embora com uma correlação menor comparado com o cenário anterior, para a segunda componente a variável média do HT também mostrou-se importante.

Tabela 12. Correlação das variáveis com as componentes principais.

	PC I	PC II	PC III
Ganho de peso diário I	0,81	-0,31	0,48
Média do OPG	-0,84	-0,17	0,49
Média da % do HT	0,11	0,96	0,24

Tabela 13. Correlação das variáveis com as componentes principais.

	PC I	PC II	PC III
Ganho de peso diário II	0,72	-0,32	0,60
Média do OPG	-0,72	0,31	0,61
Média da % do HT	0,55	0,83	0,006

As **Figuras** 11 e 12 mostram a contribuição de cada variável nas componentes e também indícios de não correlação entre as variáveis para os cenários I e II, respectivamente. Observa-se que as setas formam quase uma angulação de 90° indicando uma correlação de quase 0 entre as variáveis para o primeiro cenário. No segundo cenário não temos o mesmo comportamento, pois a média do OPG e o GPII não chegaram a formar uma ângulo de quase 90°, veja que as setas estam em sentidos opostos indicando, nesse caso uma correlação negativa.

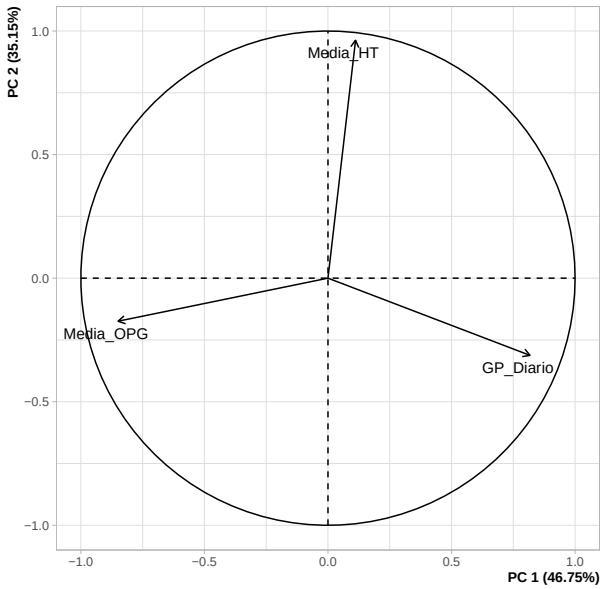


Figura 11. *Biplot* para as variáveis do cenário I.

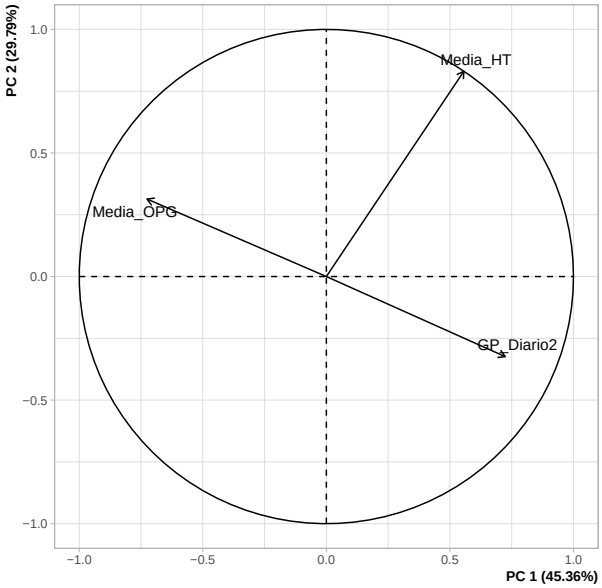


Figura 12. *Biplot* para as variáveis do cenário II.

As **Tabelas** 14 e 15 mostram as correlações pelo coeficiente de correlação de *Pearson* para cada cenário. Podemos perceber que as correlações entre as variáveis em ambos os cenários são bem baixas. Apenas as variáveis OPG e ganho de peso I e II são pouco mais correlacionadas, com valores de $-0,39$ e $-0,25$, respectivamente, mesmo assim podemos considerar correlações baixas.

Tabela 14. Tabela de correlações entre as variáveis cenário I.

Variáveis	GPI	Média do OPG	Média da % do HT
GPI	-	-0,39	-0,08
Média do OPG	-0,39	-	-0,13
Média da % do HT	-0,08	-0,13	-

Tabela 15. Tabela de correlações entre as variáveis cenário II.

Variáveis	GPII	Média do OPG	Média da % do HT
GPII	-	-0,25	0,13
Média do OPG	-0,25	-	-0,13
Média da % do HT	0,13	-0,13	-

Segundo Lira e Neto (2006) o coeficiente de correlação de *Pearson* pode ser interpretado como um indicador que descreve a interdepêndencia entre duas variáveis e sua significância é verificada atráves de um teste de hipoteses em que a estatística de teste é dada por $t = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}$, em que $t \sim t_{n-2}$, ou seja, t segue uma distribuição t com $n-2$ graus de liberdade.

Para verificar se existe correlação, as **Tabelas** 16 e 17 apresentam os teste de correlação de *Pearson*, com as hipóteses:

$$H_0 : \text{não existe correlação entre as variáveis}$$
$$H_1 : \text{existe correlação entre as variáveis}$$

com base em um nível de significância de $\alpha = 0,05$ rejeita a hipótese de que existe correlação entre as variáves do GPI e a média do OPG, mas não rejeita a hipótese de que existe correlação entre as variáveis GPI e a média do HT, e entre a média do HT com a média do OPG, para o primeiro cenário. Já para o segundo cenário, o resultado se repete na **Tabela** 17. Embora, o teste de correlação tenha dado positivo para algumas variáveis, vimos que essas correlações são fracas. Nesse caso, podendo-se prosseguir para a análise de agrupamentos.

Tabela 16. Teste de correlação cenário I.

Variáveis	p-valor
GPI e Média do OPG	0,00
GPI e Média da % do HT	0,35
Média do HT e Média do OPG	0,15

Tabela 17. Teste de correlação cenário II.

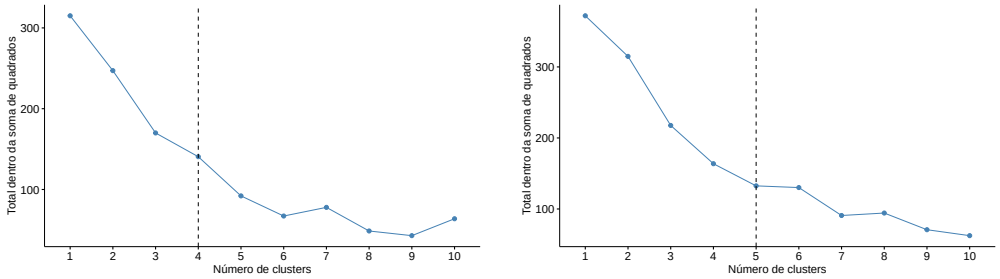
Variáveis	p-valor
GPII e Média do OPG	0,00
GPII e Média da % do HT	0,15
Média do HT e Média do OPG	0,15

No próximo tópico será apresentada a análise de agrupamentos, o número de clusters pré-determinadospara o método do k médias foi $k = 3$, mas no próprio método do k-médias tem-se uma função que determina a quantidade de grupos que a análise poderia ter. Entretanto, nesse estudo buscou-se por apenas três grupos os animais suscetíveis, resilientes e resistentes a infecção.

4.3 Análise de agrupamentos

Em premissa, o número de clusters que poderia ser usado na análise de agrupamentos no primeiro cenário seria de 4 clusters, e para o segundo cenário seria de 5 clusters. Veja na **Figura** 13, a linha tracejada no eixo x indica onde se teria algo parecido como uma dobra (cotovelo), onde a linha partindo do eixo y muda um pouco o sentido, dessa forma indicando que o número de clusters dito “ótimo” seria o respectivo valor.

Entretanto, para os fins desse estudo não faz sentido utilizar os valores acima de 3 para determinar os grupos, por isso o valor utilizado para o método do k-médias foi de $k = 3$.



(a) Cenário I. (b) Cenário II.

Figura 13. Número ótimo de clusters para os cenários I e II.

As **Tabelas** 18 e 19 apresentam os valores dos centróides calculados pelo método do k-médias para as variáveis e a quantidade de animais presente em cada grupo. O resultado para o primeiro cenário exibido na **Tabela** 18, mostra que o grupo 1 apresentou o melhor ganho de peso com o número médio do OPG superior a 800, o grupo 2 mostra o menor número médio do OPG, mas com o ganho de peso inferior ao do grupo 1. Já o grupo 3 com somente 5 animais apresenta o pior desempenho, tanto para o ganho de peso quanto para o OPG, em todos os grupos os animais apresentaram uma boa porcentagem para o teste do hematócrito.

Tabela 18. Valores dos centróides cenário I

Clusters	Ganho de peso diário I	Média da % do HT	Média do OPG	Número de indivíduos
1 - resilientes	10,42	28,87	820	34
2 - resistentes	5,13	36,05	548	67
3 - suscetíveis	-47,77	34,9	2658	5

Tabela 19. Valores dos centróides cenário II

Clusters	Ganho de peso diário II	Média da % do HT	Média do OPG	Número de indivíduos
1 - suscetíveis	-3,24	30,39	1712	26
2 - resilientes	8,91	34,39	493	71
3 - resistentes	76,03	37,45	528	28

O resultado para o segundo cenário exibido na **Tabela** 19, apresenta que o grupo 3 mostrou o melhor desempenho para o ganho de peso. Os grupos 2 e 3 mostraram um número médio do OPG bem próximo nos valores de 493 e 528 OPG, respectivamente. O

grupo 1 apresentou o pior desempenho tanto para o ganho de peso quanto para o OPG e todos grupos também apresentaram uma boa média para teste do hematócrito.

A **Figura** 14 mostra o gráfico dos grupos para o cenário I, note que a nuvem que representa o grupo três ficou no canto superior esquerdo. Se voltarmos na **Figura** 11 percebemos que a variável número médio do OPG e a média da porcentagem teste do HT influenciaram para que os indivíduos desses grupos fossem alocados nessa região. Na **Tabela** 18, temos que o centróide para o OPG desse grupo é de $\bar{x} = 2658$ e o centróide para o HT é de $\bar{x} = 34,9$. Embora o a porcentagem do teste do hematócrito para esses animais estejam em valores bons o número médio de ovos por grama de fezes e o valor negativo para o ganho de peso diário levariam esses animais a serem caracterisados com uma infecção pesada considerando os valores apresentados no capítulo anterior.

A **Figura** 15 mostra o gráfico dos grupos para o cenário II, o grupo três foi o que teve o maior valor para o centróide para o ganho de peso com o centróide para o número médio do OPG $\bar{x} = 528$. Note que, os centróides para o HT para os três grupos ficaram bem próximo, por isso as três nuvens de pontos no gráfico coincidiram em alturas próximas. Entretanto, as variáveis ganho de peso II e OPG puxam as nuvens de pontos para direita e esquerda, respectivamente. Dessa forma, o grupo 3 ficou a esquerda por causa do centróide do OPG.

Figura 14. Representação dos grupos usando as componentes principais para o cenário I.

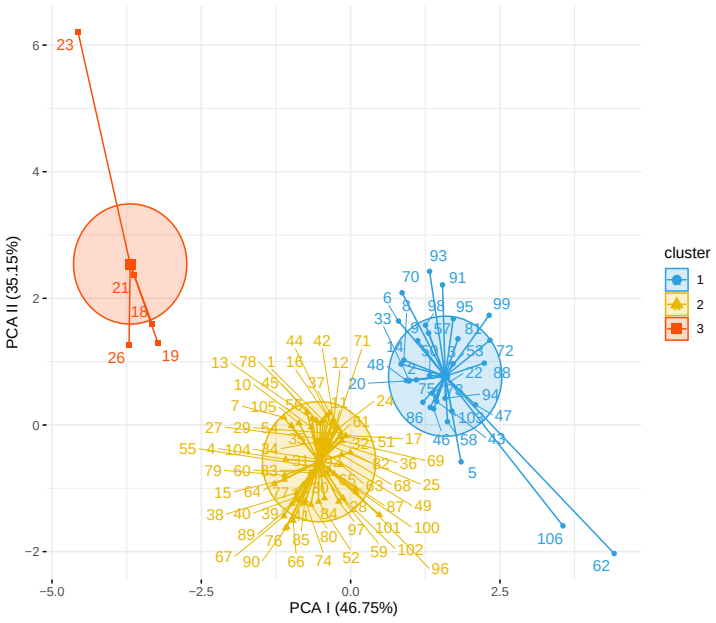
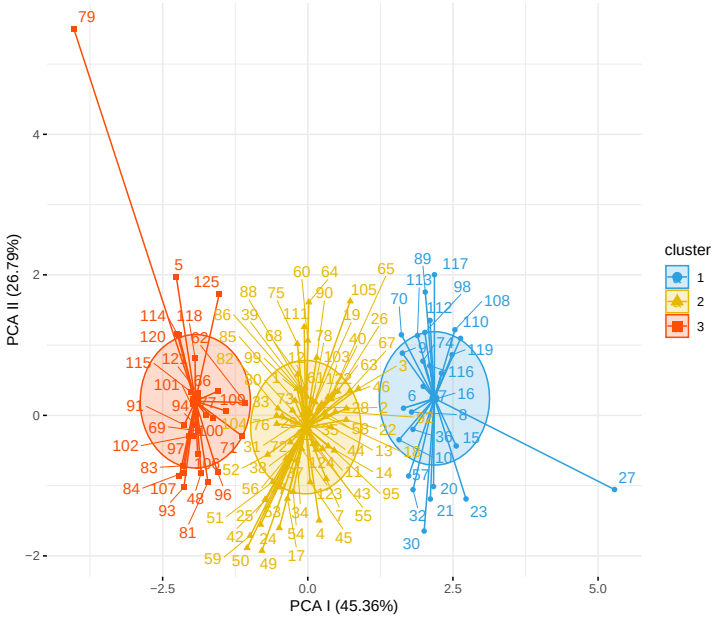


Figura 15. Representação dos grupos usando as componentes principais para o cenário II.



Os próximos passos, tem como finalidade caracterizar quais os grupos em ambos os cenários seriam referidos aos animais resilientes, suscetíveis e resistentes. Para isso, de cada grupo foi selecionado uma amostra aleatória de 5 animais para uma breve comparação e determinação das caraterísticas dos grupos.

4.4 Descrição dos grupos

4.4.1 Determinação dos grupos para o cenário I

No primeiro cenário, o grupo 1 obteve o melhor desempenho em termos de ganho de peso diário. No entanto, o número de ovos por grama de fezes excedeu o valor médio de 1000 para alguns animais, considerando-se a porcentagem do HT estão dentro dos parâmetros normais. Por sua vez, o grupo 2 registrou o menor valor médio de OPG, mas o ganho de peso diário não pode ser considerado extremamente baixo. É importante ressaltar que os animais resilientes apresentam características de infecção, embora não manifestem sinais clínicos. Já o grupo 3, composto apenas por 5 animais, apresentou os maiores valores médios de OPG e os menores valores de ganho de peso diário.

Tabela 20. Amostra de 5 animais presentes no grupo 1 do cenário I.

ID do animal	Classe	Ganho de peso diário I	Média da % do HT	Média do OPG
1427	MATRIZ	11,2	30,5	1629
2012	FILHOTE - M	3,30	28,3	1280
2015	ADULTO - M	119,00	28,7	580
2036	ADULTO - M	12,90	26,5	1190
S/N	ADULTO - M	79,40	29,0	100

Tabela 21. Amostra de 5 animais presentes no grupo 2 do cenário I.

ID do animal	Classe	Ganho de peso diário I	Média da % do HT	Média do OPG
1912	MATRIZ	3,66	40,1	350
1917	MATRIZ	1,38	38,2	247
2010	FILHOTE - M	7,51	36,7	600
2029	FILHOTE - F	7,90	39,1	367
2106	FILHOTE - F	10,60	36,0	25

Observando os valores apresentados nas **Tabelas** 20, 21 e 22 e comparando os valores que caracterizam níveis da infecção no trabalho de Chagas, Carvalho e Molento (2007), a classificação dos grupos poderia ser feita da seguinte forma para o primeiro cenário, os animais do grupo 1 com a característica de resiliência, pois tiveram um alto número médio do OPG e bom ganho de peso diário. Os animais do grupo 2 com a característica de resistência, apresentaram uma baixo valor médio para o OPG, um ganho de peso razoável e os animais do grupo 3 com a característica de suscetíveis, apresentando grandes valores para o número médio do OPG e uma perda de peso.

Tabela 22. Amostra de 5 animais presentes no grupo 3 do cenário I.

ID do animal	Classe	Ganho de peso diário I	Média da % do HT	Média do OPG
1722	MATRIZ	-60,9	33,5	1175
1723	MATRIZ	-32,9	36,3	1917
1735	MATRIZ	-50,5	34,7	2375
1812	MATRIZ	-47,8	32,0	6075
1820	MATRIZ	-46,7	38,0	1750

4.4.2 Determinação dos grupos para o cenário II

No segundo cenário, os animais do grupo 1 apresentaram o menor desempenho no ganho de peso diário e um alto valor para o número médio do OPG. Os animais do grupo 2, apresentaram um valor para o número médio do OPG bem próximo dos animais do grupo 3, porém apresentaram ganho de peso diário bem menor comparado com os animais do grupo 3. Os animais do grupo 3, apresentaram o melhor desempenho do ganho de peso diário. Como o valor do centróide para o número médio do OPG entre os grupos 2 e 3 foram bem próximos, podemos considerar o valor do centróide para o teste do hematócrito que foi maior para o grupo 3.

Considerando os valores apresentados nas **Tabelas** 23, 24 e 25 e os valores apresentadados no trabalho de Chagas, Carvalho e Molento (2007), a classificação dos grupos poderia ser feita da seguinte forma os animais do grupo 1 seriam caracterizados como suscetíveis a infecção considerando o ganho de peso diário e o número médio para o OPG. Os animais pertecentes ao grupo 2 com a característica de resiliência e os animais do

grupo 3 com a característica de resistência. Embora, o número médio do OPG não seja tão discrepante entre os grupos 2 e 3, a porcentagem média do teste do hematócrito e o ganho de peso diário do grupo 3 foram melhores que o grupo 2, isso leva a considerar o grupo 3 contendo os animais com as característica de resistência a infecção.

Tabela 23. Amostra de 5 animais presentes no grupo 1 do cenário II.

ID do animal	Classe	Ganho de peso diário II	Média da % do HT	Média do OPG
1427	MATRIZ	17,60	31,0	1822
1812	MATRIZ	-47,80	32,0	6075
1820	MATRIZ	-46,70	38,0	1750
2036	ADULTO - M	12,90	26,5	1190
2103	MATRIZ	2,89	25,8	1900

Tabela 24. Amostra de 5 animais presentes no grupo 2 do cenário II.

ID do animal	Classe	Ganho de peso diário II	Média da % do HT	Média do OPG
1612	MATRIZ	35,20	33,0	715
1902	MATRIZ	6,11	35,9	350
1903	ADULTO - M	27,10	32,7	430
1904	MATRIZ	10,80	31,7	866
1910	MATRIZ	15,80	32,8	821

Tabela 25. Amostra de 5 animais presentes no grupo 3 do cenário II.

ID do animal	Classe	Ganho de peso diário II	Média da % do HT	Média do OPG
2023	FILHOTE - F	58,4	42,0	533
2030	FILHOTE - F	40,9	42,0	262
2101	FILHOTE - M	50,0	39,8	488
2103	FILHOTE - F	48,9	36,0	725
2112	FILHOTE - F	113,0	35,0	350

Com a ideia de gerar duas variáveis diferentes sobre o ganho de peso do animal, buscou-se verificar se, ao observar o que aconteceu com os animais ao longo do tempo, isto resultaria em mais informações sobre a perda de peso. Entretanto, ao observar exatamente se os animais perderam peso como feito no segundo cenário pegando a diferença entre a última e primeira mensuração, resultou em um maior número para o grupo dos animais com a característica de suscetíveis a infecção.

Na coluna referente ao número de indivíduos em cada grupo exibidos na **Tabela** 18, podemos ver que apenas 5 animais foram alocados no grupo 3, sendo elas as matrizes com os piores desempenhos referentes ao ganho de peso e o número de ovos por grama de fezes. Dessa forma, o cenário II mostrou-se melhor para que o método de agrupamentos do k-médias reunisse de forma mais eficaz os animais em seus respectivos grupos. Note

que, os animais presentes no grupo 3 do primeiro cenário o qual caracterizamos como os animais suscetíveis, também foram alocados no grupo 1 do segundo cenário grupo referente aos animais com a característica de suscetíveis. Isto é perceptível com base nas duas matrizes de ID's 1812 e 1820. O próximo tópico apresenta as considerações finais.

5 CONSIDERAÇÕES FINAIS

Nesse estudo foram considerados três indicadores sendo eles o ganho de peso diário I, onde foi considerado todo o processo do animal ao longo do tempo que em que foi avaliado. O ganho de peso diário II, considerando somente a primeira e ultima mensuração de cada animal, dessa forma tem-se exatamente se o animal perdeu peso ao final do estudo. O número médio de ovos por grama de fezes e a média da porcentagem do teste do hematócrito. Além disso, o cenário II mostrou-se mais razoável para a descrição dos grupos.

Embora esses sejam ótimos indicadores mesmo que fracamente correlacionados, ambos contribuíram para a determinação dos grupos resilientes, resistentes e suscetíveis à infecção pelos nematóides gastrointestinais. Entretanto, é indicado considerar mais indicadores os quais não estejam correlacionados, mas fazendo o uso da análise de componentes principais esses indicadores poderiam ser reduzidos a duas ou três componentes.

Utilizar fortes indicadores para a identificação da infecção é de fato a melhor opção. Sobre o método de análise de agrupamento, o algoritmo k-médias é uma ótima ferramenta para agrupar indíviduos considerando um conjunto de variáveis. Ainda que a determinação dos grupos deva ser feita pelo pesquisador, pois mesmo com os indíviuos agrupados a ferramenta utilizada não destingue os grupos por isso a determinação foi feita baseando-se em estudos já realizados sobre seleção genética de ovinos e estudos que avaliaram e estudaram o grau da infecção em ovinos por nematóides gastrointestinais.

Por fim, o método do k-médias conseguiu distribuir bem os indivíduos em cada grupo. Assim, propiciou a identificação dos animais com as características desejáveis que contribuem para a não disseminação da infecção causada por nematóides gastrointestinais. Além disso, ajuda a ter um controle melhor dos animais nas pastagens, pois os animais caracterizados como suscetíveis podem ser direcionados ao tramento ou eliminados do plantel.

REFERÊNCIAS

ARAUJO, J. I. M. **Estudo gentico da resistê à verminoses gastrintestinais em ovinos tropicais**. Dissertação (Mestrado) — Universidade Federal do Piaui, 2017.

ARSENOPOULOS, K. V.; FTHENAKIS, G. C.; KATSAROU, E. I.; PAPADOPOULOS, E. Haemonchosis: A challenging parasitic infection of sheep and goats. **Animals**, MDPI, v. 11, n. 2, p. 363, 2021.

BORTOLUZZI, C. **Análise de componentes principais dos valores genéticos dos parâmetros da curva de lactação de ovinos leiteiros**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2018.

CHAGAS, A. C. d. S.; CARVALHO, C. O. d.; MOLENTO, M. B. Método famacha: um recurso para o controle da verminose em ovinos. São Carlos, SP: Embrapa Pecuária Sudeste, 2007., 2007.

COUTINHO, R. M. A. **Marcadores fenotípicos para caracterização de caprinos com diferentes níveis de resistência às endoparasitoses gastrintestinais**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2012.

DIDAY, E. **Selected contributions in data analysis and classification**. [S.l.]: Springer Science & Business Media, 2007.

FERNANDES, M. A. **Dinâmica do metabolismo do fósforo em cordeiros submetidos a infecção mista de Haemonchus contortus e Trichostrongylus colubriformis utilizando o 32P**. Tese (Doutorado) — Universidade de São Paulo, 2021.

GASTWIRTH, J. L.; GEL, Y. R.; HUI, W. L. W.; LYUBCHICH, V.; MIAO, W.; NOGUCHI, K. **lawstat: Tools for Biostatistics, Public Policy, and Law**. [S.l.], 2023. R package version 3.6. Disponível em: <https://CRAN.R-project.org/package=lawstat>.

GIORDANI, P.; FERRARO, M. B.; MARTELLA, F.; GIORDANI, P.; FERRARO, M. B.; MARTELLA, F. Non-hierarchical clustering. **An Introduction to Clustering with R**, Springer, p. 75–109, 2020.

GUITARRARA, P. **Clima temperado**. 2023. Disponível em: <https://brasilescola.uol.com.br/geografia/clima-temperado.htm>.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. **Journal of the Royal Statistical Society. Series c (applied statistics)**, JSTOR, v. 28, n. 1, p. 100–108, 1979.

HASSUM, I. C. Dicas gerais para controle da verminose na produção de pequenos ruminantes. Bagé: Embrapa Pecuária Sul, 2009., 2009.

HESPANHOL, R. M.; PELOZO, P. D. F.; PEREIRA, D. R. Identificação de padrões na cadeia produtiva de leite do estado de São Paulo utilizando o k-means. II Simposio Internacional de Agronegocio e Desenvolvimento - SIAD, 2016.

HOSTE, H.; SOTIRAKI, S.; TORRES-ACOSTA, J. F. de J. Control of endoparasitic nematode infections in goats. **Veterinary Clinics: Food Animal Practice**, Elsevier, v. 27, n. 1, p. 163–173, 2011.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis, Pearson Education Inc.** [S.l.]: Upper Saddle River New Jersey, 2007.

KASSAMBARA, A.; MUNDT, F. **factoextra: Extract and Visualize the Results of Multivariate Data Analyses**. [S.l.], 2020. R package version 1.0.7. Disponível em: <https://CRAN.R-project.org/package=factoextra>.

KORITIAKI, N. A.; RIBEIRO, E. L. de A.; MUNIZ, C. A. S. D.; MARESTONE, B. S.; JUNIOR, F. F. Análise de componentes principais para características de crescimento pré-desmame em ovinos da raça santa inês. **Semina: Ciências Agrárias**, v. 40, n. 6Supl2, p. 3269–3278, 2019.

LÊ, S.; JOSSE, J.; HUSSON, F. FactoMineR: A package for multivariate analysis. **Journal of Statistical Software**, v. 25, n. 1, p. 1–18, 2008.

LINDEN, R. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, n, v. 4, n. 4, p. 18–36, 2009.

LIRA, S. A.; NETO, A. C. Coeficientes de correlação para variáveis ordinais e dicotômicas derivados do coeficiente linear de pearson. **Ciência & Engenharia**, v. 15, n. 1/2, p. 45–53, 2006.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [S.l.], 1967. v. 5.1, n. 14, p. 281–297.

MENDES, J. P.; TSUZUKI, T. T.; FERREIRA, M. B.; GARCIA, W. R.; VALENTIM, J. K.; PIETRAMALE, R. T. R. Haemonchus contortus e medidas estratégicas de controle para ovinos. **Ensaios e Ciência C Biológicas Agrárias e da Saúde**, v. 24, n. 2, p. 105–110, 2020.

MENEGATTO, L. S. **Evidências e componentes de seleção em ovinos para resistência e tolerância a verminoses**. Tese (Doutorado) — Universidade de São Paulo, 2023.

MURTAGH, F.; CONTRERAS, P. Algorithms for hierarchical clustering: an overview. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 2, n. 1, p. 86–97, 2012.

OLIVEIRA, E. J. **Critérios de seleção para características de importância econômica em ovinos da raça Santa Inês**. Tese (Doutorado) — Universidade de São Paulo, 2016.

OLIVEIRA, W. P. d. S. **Lógica fuzzy para discriminar a resposta de caprinos a verminose: resistência, resiliência e sensibilidade.** Tese (Doutorado) — Universidade Federal do Piaui, 2021.

OSÓRIO, T. M.; MENEZES, L. de M.; ROSA, K. B. da; ESCOBAR, R. F.; LENCINA, R. M.; MAYDANA, G. de M.; SOUZA, V. Q. de. Resistência anti-helmíntica em nematódeos gastrointestinais na ovinocultura: uma revisão. **Brazilian Journal of Development**, v. 6, n. 11, p. 89194–89205, 2020.

PARENTE, H.; MACHADO, T.; CARVALHO, F.; GARCIA, R.; ROGÉRIO, M.; BARROS, N.; ZANINE, A. Desempenho produtivo de ovinos em confinamento alimentados com diferentes dietas. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, SciELO Brasil, v. 61, p. 460–466, 2009.

PéREZ, J. A. O. A. G. **Arthrobotrys musiformis (Orbiliales) Kills Haemonchus contortus Infective Larvae (Trichostronylidae) through Its Predatory Activity and Its Fungal Culture Filtrates**. 2022. Disponível em: <https://www.mdpi.com/2076-0817/11/10/1068>.

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

RENCHER, A. C.; CHRISTENSEN, W. Méthods of multivariate analysis. a john wiley & sons. **Inc. Publication**, v. 727, p. 2218–0230, 2002.

SARGISON, N. D. Pharmaceutical control of endoparasitic helminth infections in sheep. **Veterinary Clinics: Food Animal Practice**, Elsevier, v. 27, n. 1, p. 139–156, 2011.

SCHUTTEN, G.-J.; CHAN, C. hong; LEEPER, T. J.; STEUER, D. **readODS: Read and Write ODS Files**. [S.l.], 2023. R package version 1.8.0. Disponível em: <https://CRAN.R-project.org/package=readODS>.

SENAR, S. N. D. A. R. **Ovinocultura: Criação e manejo de ovinos de corte.** 2021. Disponível em: <https://www.cnabrasil.org.br/assets/arquivos/265_Ovino_corte.pdf>.

SOTOMAIOR, C. S.; CARLI, L. M. D.; TANGLEICA, L.; KAIBER, B. K.; SOUZA, F. P. de. Identificação de ovinos e caprinos resistentes e susceptíveis aos helmintos gastrintestinais. **Revista Acadêmica Ciência Animal**, v. 5, n. 4, p. 397–412, 2007.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <https://www.stats.ox.ac.uk/pub/MASS4/>.

VIEIRA, L. d. S.; LÔBO, R. N. B.; CAVALCANTE, A. C. R.; NEVES, M. R. M. das; NAVARRO, A. d. C.; BENVENUTI, C. L.; ZAROS, L. G. Panorama mundial dos métodos de controle de endoparasitoses. In: SIMPÓSIO INTERNACIONAL SOBRE CAPRINOS E OVINOS DE CORTE, 4.; FEIRA ..., 2009.

WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, Taylor & Francis, v. 58, n. 301, p. 236–244, 1963.

WICKHAM, H.; AVERICK, M.; BRYAN, J.; CHANG, W.; MCGOWAN, L. D.; FRANçOIS, R.; GROLEMUND, G.; HAYES, A.; HENRY, L.; HESTER, J.; KUHN, M.; PEDERSEN, T. L.; MILLER, E.; BACHE, S. M.; MüLLER, K.; OOMS, J.; ROBINSON, D.; SEIDEL, D. P.; SPINU, V.; TAKAHASHI, K.; VAUGHAN, D.; WILKE, C.; WOO, K.; YUTANI, H. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.

WICKHAM, H.; FRANçOIS, R.; HENRY, L.; MüLLER, K.; VAUGHAN, D. **dplyr: A Grammar of Data Manipulation**. [S.l.], 2023. R package version 1.1.1. Disponível em: <https://CRAN.R-project.org/package=dplyr>.

APÊNDICES

Apêndice I

No **Apêndice I** são apresentados a lista de pacotes utilizados no desenvolvimento da análise, assim como a lista de códigos implementados no software *R*.

```
# Lista de pacotes:

library(readODS) # para a leitura dos dados
library(lawstat) # para auxílio na criação de alguns gráficos
library(MASS) # para auxílio em algumas funções
library(tidyverse) # para auxílio na criação de alguns gráficos
library(stats) # para o auxílio na análise de ACP
library(factoextra) # para auxílio na criação de alguns gráficos
library(FactoMineR) # para o auxílio na análise de ACP

# Leitura dos dados:

dados <- read_ods("inventario.ods", col_names = T)


# Atribuindo o formato de data a variável referente a data da coleta:

dados$Data <- as.Date(dados$Data, format = "%d/%m/%y")


# Criação das tabelas:

tabela_geral_media <- dados %>%
    group_by(Animal, Categoria) %>%
    summarise(Media_Peso = mean(Peso, na.rm = T),
    Media_HT = mean(HT, na.rm = T),
    Media_OPG = mean(OPG, na.rm = T),
    Tempo_no_estudo = n())

# Machos Adultos:

media_MA <- tabela_geral_media %>%
                filter(Categoria == "ADULTO - M")
```

```
# Filhotes fêmeas:

media_FF <- tabela_geral_media %>%
                filter(Categoria == "FILHOTE - F")

# Filhotes machos:

media_FM <- tabela_geral_media %>%
                filter(Categoria == "FILHOTE - M")

# Matrizes:

media_M <- tabela_geral_media %>%

  filter(Categoria == "MATRIZ")

# Obtenção das variáveis referentes ao ganho de peso

tabela_geral_gp <- dados %>%
                        arrange(Data) %>%
                        group_by(Animal, Categoria) %>%
                        summarise(Data = Data, Peso = Peso, HT = HT, OPG = OPG) %>%
                        mutate(GP = Peso-lag(Peso, default = NA),
                                Dias = Data-lag(Data, default = NA))

tabela_media_gp <- tabela_geral_gp %>%

group_by(Animal, Categoria) %>%
    summarise(Media_Peso = mean(Peso, na.rm = T),
                Media_GP = mean(GP, na.rm = T),
                Media_GP2 = sum(GP, na.rm = T),
                Media_HT = mean(HT, na.rm = T),
                Media_OPG = mean(OPG, na.rm = T),
                Total_dias = sum(Dias, na.rm = T),
                n = n())

tabela_media_gp <- na.omit(tabela_media_gp)
tabela_media_gp$Media_GP <- tabela_media_gp$Media_GP*1000
```

```
tabela_media_gp$GP_Diario <- tabela_media_gp$Media_GP/tabela_media_gp$Total_dias
tabela_media_gp$Media_GP2 <- tabela_media_gp$Media_GP2*1000
tabela_media_gp$GP_Diario2 <- tabela_media_gp$Media_GP2/tabela_media_gp$Total_dias

# Criação dos gráficos:
# Box-plot

ggplot(tabela_media_gp, aes(Categoria, GP_Diario)) +
    geom_boxplot()+
    labs(y = "Ganho de peso diário em (g)", x = NULL)+
    theme_grey()+
    theme(axis.text.x = element_text(angle = 45))

ggplot(tabela_media_gp, aes(Categoria, GP_Diario2)) +
    geom_boxplot()+
    labs(y = "Ganho de peso diário em gramas(g)", x = NULL)+
    theme(axis.text.x = element_text(angle = 45))

# Gráficos de pontos

ggplot(tabela_media_gp, aes(Media_Peso, GP_Diario))+
    geom_point()+
    facet_wrap(~Categoria)+
    theme(legend.position = "none")+
    labs(y = "Ganho de peso diário em (g)",
         x = "Peso médio em (kg)", title = NULL)

ggplot(tabela_media_gp, aes(Media_OPG, GP_Diario))+
    geom_point()+
    facet_wrap(~Categoria)+
    theme(legend.position = "none")+
    labs(y = "Ganho de peso diário em (g)",
         x = "Número médio de ovos por grama de fezes - OPG",
         title = NULL)

ggplot(tabela_media_gp, aes(Media_HT, GP_Diario))+
    geom_point()+
    facet_wrap(~Categoria)+
    theme(legend.position = "none")+
```

```
    labs(y = "Ganho de peso diário (g)",
    x = "Porcentagem média do HT",
    title = NULL)

ggplot(tabela_media_gp, aes(Media_Peso, GP_Diario2))+
    geom_point()+
    facet_wrap(~Categoria)+
    theme(legend.position = "none")+
    labs(y = "Ganho de peso diário em (g)",
         x = "Peso médio em (kg)",
         title = NULL)

ggplot(tabela_media_gp, aes(Media_OPG, GP_Diario2))+
    geom_point()+
    facet_wrap(~Categoria)+
    theme(legend.position = "none")+
    labs(y ="Ganho de peso diário em (g)",
         x = "Número médio de ovos por grama de fezes - OPG",
         title = NULL)

ggplot(tabela_media_gp, aes(Media_HT, GP_Diario2))+
    geom_point()+
    facet_wrap(~Categoria)+
    theme(legend.position = "none")+
    labs(y = "Ganho de peso diário em (g)",
         x = "Porcentagem média do HT",
         title = NULL)

# Análise de componentes principais e alguns gráficos:

dadospca1 <- tabela_media_gp %>%
                select(GP_Diario, Media_OPG, Media_HT)

scale1 <- scale(dadospca1[,-1])

pca_cor <- PCA(scale1, scale.unit = T) # aqui as variáveis foram padronizadas com
                                       # o comando scale, mas dentro do comando PCA
                                       # pode-se indicar "scale = T" que o
                                       # comando automaticamente faz a padronização.
```

```
summary(pca_cor)

gr <- plot(pca_cor,choix="var", title = "Gráfico das variáveis")

gr +labs(x = "PC 1 (46.75%)", y = "PC 2 (35.15%)")

dadospca2 <- tabela_media_gp %>%
                select(GP_Diario2, Media_OPG, Media_HT)

scale2 <- scale(dadospca2[,-1])

pca_cor2 <- PCA(scale2, scale.unit = T)

summary(pca_cor2)

gr2 <- plot(pca_cor2,choix="var", title = "Gráfico das variáveis")

gr2 + labs(x = "PC 1 (45.36%)", y = "PC 2 (29.79%)")



# Tabelas referente as correlações e o teste de correlação de Pearson

matr1 <- with(tabela_media_gp,
            cbind(GP_Diario, Media_OPG, Media_HT))

cor(matr1)

cor.test(matr1[,1],matr1[,2] , method = "pearson")

cor.test(matr1[,1],matr1[,3] , method = "pearson")

cor.test(matr1[,2],matr1[,3] , method = "pearson")

matr2 <- with(tabela_media_gp,
            cbind(GP_Diario2, Media_OPG, Media_HT))

cor(matr2)
```

```
cor.test(matr2[,1],matr2[,2] , method = "pearson")

cor.test(matr2[,1],matr2[,3] , method = "pearson")

cor.test(matr2[,2],matr2[,3] , method = "pearson")

# Análise de agrupamento

df <- scale(pca_borregos[,-c(1,2)])

df2 <- scale(cluster1[,-c(1,2)]) # Obs.: O banco de dados cluster1 foi criado
                                 # para separar as variáveis GPII da GPI

# Gráficos para o número "ótimo" de grupos

fviz_nbclust(df, kmeans, method = "wss")+
    geom_vline(xintercept = 4, linetype = 2)+
    labs(x = "Número de clusters",
         y = "Total dentro da soma de quadrados",
         title = " ")

fviz_nbclust(df2, kmeans, method = "wss")+
    geom_vline(xintercept = 5, linetype = 2)+
    labs(x = "Número de clusters",
         y = "Total dentro da soma de quadrados",
         title = " ")

# Método do k-médias

kms1 <- kmeans(df, 3, algorithm = "MacQueen")

aggregate(pca_borregos[,-c(1,2)],
        by=list(cluster=kms1$cluster), mean) # Obtenção dos centróides

pca_borregos3 <- cbind(pca_borregos,
                    cluster=kms1$cluster) # Adicionando os grupos ao banco de dados

kms4 <- kmeans(df2, 3, algorithm = "MacQueen")
```

```
aggregate(cluster1[,-c(1,2)],
        by=list(cluster=kms4$cluster), mean) # Obtenção dos centróides

cluster4 <- cbind(cluster1,
        cluster=kms4$cluster) # Adicionando os grupos ao banco de dados

# Gráfico dos grupos

fviz_cluster(kms1, data=pca_borregos3[,-c(1,2)],
             palette = c("#2E9FDF", "#E7B800","#FC4E07"),
             main = NULL,
             ellipse.type="euclid",
             star.plot=TRUE,
             repel=TRUE,
             xlab = "PCA I (46.75%)",
             ylab = "PCA II (35.15%)",
             ggtheme=theme_minimal())

fviz_cluster(kms4, data=cluster4[,-c(1,2)],
             palette = c("#2E9FDF", "#E7B800","#FC4E07"),
             main = NULL,
             ellipse.type="euclid",
             star.plot=TRUE,
             repel=TRUE,
             xlab = "PCA I (45.36%)",
             ylab = "PCA II (26.79%)",
             ggtheme=theme_minimal())
```