UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Antonio César Holanda Mangueira

Métodos Multivariados no estudo das relações entre variáveis Socioeconômicas do Estado de Mato Grosso

Campina Grande - PB

Agosto 2017

Antonio César Holanda Mangueira

Métodos Multivariados no estudo das relações entre variáveis Socioeconômicas do Estado de Mato Grosso

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Ricardo Alves de Olinda

Campina Grande - PB

Agosto 2017

M277m Mangueira, Antonio César Holanda.
Métodos multivariados no estudo das relações entre variáveis socioeconômicas do Estado de Mato Grosso [manuscrito] / Antonio César Holanda Mangueira. - 2017.
46 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2017.
"Orientação: Prof. Dr. Ricardo Alves de Olinda, Departamento de Estatística".

1. Análise multivariada. 2. Biplot. 3. Desenvolvimento sustentável. 4. MANOVA. I. Título.

21. ed. CDD 519.53

Antonio César Holanda Mangueira

Métodos Multivariados no estudo das relações entre algumas variáveis Socioeconômicas do Estado de Mato Grosso

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 09 de Agosto de 2017.

BANCA EXAMINADORA

Prof. Dr. Ricardo Alves de Olinda
Universidade Estadual da Paraíba

Prof. Dr. João Gil de Luna
Universidade Estadual da Paraíba

Prof. Dr. Macio Augusto de Albuquerque
Universidade Estadual da Paraíba

Agradecimentos

Em primeiro lugar a Deus por iluminar o meu caminho durante esta caminhada.

De modo especial a minha família e amigos, que de forma carinhosa me deram força e coragem, apoiando nos momentos difíceis, e que mesmo não sabendo, iluminaram de maneira especial os meus pensamentos, e levando a buscar de mais conhecimentos.

Ao meu orientador Doutor Ricardo Alves Olinda pela sua dedicação, profissionalismo, humildade e sua disponibilidade de ajudar. E aos demais professores do departamento de estatística que, direta ou indiretamente, me ajudaram a chegar onde cheguei. Ao longo desses anos, tive a oportunidade de aprender com os melhores e mais competentes professores que eu poderia ter.

Aos meus colegas que também contribuíram para meu aprendizado e por estarem unidos nos momentos difíceis ao longo do curso, enriquecendo minha vida acadêmica.

"A política serve a um momento no presente,
mas uma equação é eterna."
(Albert Einstein)

"Estatística prova que se você se preocupa com a sua vida,
sobra menos tempo para falar da minha."
(Ana Carolina)

"Alguns usam a estatística como os bêbados usam postes,
mais para apoio do que para iluminação."
(Andrew Lang)

Resumo

A grande quantidade de dados e como obter informações tem sido um desafio para especialistas das mais variadas áreas do conhecimento. Dentre os tais problemas existentes está o de relacionar e buscar informações para a tomada de decisão. Sendo assim, neste trabalho busca-se estudar a influência que a produção de soja exerce no Índice de Desenvolvimento Sustentável para Municípios (IDSM) do Estado de Mato Grosso, fazendo uso de algumas técnicas da análise multivariada. A disseminação do uso das técnicas da análise multivariada pode melhorar a qualidade das pesquisas, proporcionando uma economia relativa de tempo e de custo, facilitando a interpretação das estruturas dos dados e diminuindo a perda de informação. Foram estudados 137 municípios e seis variáveis. Calculou-se a matriz de covariâncias e a matriz de correlações, que por sua vez apresentaram correlações significativas. Foram realizados testes univariados de Shapiro-Wilk que apresentaram normalidade univariada em três variáveis simultaneamente. Em seguida foram aplicados os testes de Mardia, Henze-Zirkle, Royston que indicaram haver fortes evidências para não se rejeitar a hipótese de normalidade multivariada. Realizou-se uma análise de componentes principais visando obter variáveis latentes em função das já existentes com variância conhecida (autovalores). Por fim, foi aplicado a técnica de biplot com objetivo de representar graficamente a matriz de dados com os respectivos grupos que compõem as mesorregiões bem como suas variáveis. As análises foram realizadas com auxilio do software R.

Palavras-chave: Desenvolvimento Sustentável. MANOVA. Componentes Principais.

Abstract

The vast amount of data and how to obtain information have been a challenge to the experts from various areas of knowledge. Among those problems there is the difficult of connecting and collecting information to the decision-making. Therefore, this research aims at studying the influence of the production of soybean on the Index of Sustainable Development for Municipalities (IDSM) in the state of Mato Grosso, by making use of some techniques of multivariate analysis. The dissemination of the use of the multivariate analysis techniques can improve the quality of the research, for it promotes the saving of time and cost, besides, it facilitates the interpretation of the data structures by decreasing the information loss. 137 municipality and six variables were studied. It was calculated the covariance matrix and the correlation matrix, which in turn presented meaningful correlations. Shapiro-Wilk univariate tests were performed and presented univariate normality in three variables simultaneously. Next, the Mardia, Henze-Zirkle and Royston tests were done and indicated that there were strong evidences to not reject the hypothesis of multivariate normality. An analysis of main components was performed in order to obtain latent variables in function of those already existing with known variance (eigenvalues). Finally, the biplot technique was applied in order to graphically represent the data matrix with the respective groups that belong to the mesoregions as well as their variables. The analyses were performed using the software R.

Keywords: Sustainable Development. MANOVA. Main Components.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O tema sustentabilidade está em evidência em virtude da necessidade da busca por novas formas de organização do processo produtivo que priorizem a manutenção da capacidade de suporte dos sistemas ambientais e, desta forma, o bem-estar das gerações presentes e futuras. O tema em questão tem proporcionado o surgimento de metodologias que buscam mensurar e avaliar a sustentabilidade e subsidiar as tomadas de decisão de organizações públicas e privadas. Partindo desses pressupostos, adquire cada vez mais importância a realização de estudos que explorem as relações entre o surgimento, desenvolvimento e consolidação de atividades e setores econômicos em dado escopo geográfico e a partir disso, mostrar as implantações destas mudanças na geração do desenvolvimento sustentável (MACEDO et al., 2016).

Para Sachs (2007) a busca pela sustentabilidade envolve seis dimensões como ponto de partida, a saber: social, demográfica, econômica, político-institucional, ambiental e cultural. Essa definição evidência a necessidade de se observar a questão por uma ótica multidimensional, mediante análise integrada e sistêmica, não tendo como ponto de observância de fatores isolados, mas sim, os fatores em seu conjunto. De acordo com Topa (2009) quando se toma uma decisão, muitos fatores costumam estar envolvidos. Porém, nem todos têm a mesma importância para essa decisão.

Quando se analisa o mundo que nos cerca, identifica-se que todos os acontecimentos, sejam eles culturais ou naturais, envolvem um grande número de variáveis. As diversas ciências têm a pretensão de conhecer a realidade, e de interpretar os acontecimentos e os fenômenos, baseadas no conhecimento das variáveis intervenientes, consideradas importantes nesses eventos. Estabelecer relações, encontrar, ou propor, leis explicativas, é papel próprio da ciência. Para isso, é necessário controlar, manipular e medir as variáveis que são consideradas relevantes ao entendimento do fenômeno analisado. Muitas são as dificuldades em traduzir as informações obtidas em conhecimento, principalmente quando se trata da avaliação estatística das informações (MAGNUSSON, 2003).

Segundo Rencher (2003) a análise multivariada consiste de uma coleção de métodos que pode ser usado quando diversas medições são feitas em cada indivíduo ou objeto em uma ou mais amostras. Refere-se às medições como variáveis e aos indivíduos ou objetos como unidades (unidades de pesquisa, unidades amostrais ou unidades experimentais) ou observações. Na prática, conjuntos de dados multivariados são comuns, embora não sejam sempre analisados como tal. Mas o uso exclusivo de procedimentos univariados com tais dados não é desculpável, haja vista a disponibilidade de técnicas multivariadas e recursos computacionais de baixo custo para realizá-los

Para Vicini e Souza (2005) os métodos estatísticos, para analisar variáveis, estão dispostos em dois grupos: um que trata da estatística, que considera as variáveis de maneira isolada a estatística univariada, e outro que considera as variáveis de forma conjunta a estatística multivariada. Ao realizar um estudo estatístico quer seja univariado ou multivariado, sempre existirá a perda de informação, pois no momento que se esta reduzindo um conjunto de dados para ser representado pela sua média, no caso univariado se perde informação. O mesmo ocorre quando se aplica uma técnica multivariada, pois ao reduzir a dimensionalidade de um problema também se perde informação.

A análise multivariada compreende, dentre outras, as técnicas da Análise de Componentes Principais(ACP) a Análise de Variância Multivariada (MANOVA) (PONTES, 2005). A MANOVA é utilizada para comparar vetores de médias, os dados normalmente são provenientes de delineamentos estatísticos. A formulação de um teste estatístico para comparar vetores de médias, depende da partição do total da variância em: variância devido ao efeito de tratamentos e variância devido ao erro. Esta partição da variância total é denominada de MANOVA (JOHNSON; WICHERN, 2002).

A ACP é uma técnica estatística multivariada com o objetivo de rotacionar um sistema de coordenadas com alta dimensão de maneira que alcance a máxima variabilidade dos dados em baixa dimensão. A rotação realizada é linear e após, pode-se definir uma projeção com os componentes de máxima variância do espaço original. Desta forma, a interpretação dos dados é mais simples. Isso se justifica por estes componentes serem combinações lineares das variáveis originais e representam suas projeções nas direções de máxima variabilidade dos dados (MONTGOMERY, 2009).

Objetivo Geral:

Realizar um estudo dos principais aspectos teóricos e práticos acerca das técnicas da MANOVA e ACP aplicando aos dados que definem o Índice de Desenvolvimento Sustentável para Municípios(IDSM) do Estado de Mato Grosso.

Objetivos Específicos:

A fim de alcançar o objetivo geral deste trabalho, estabeleceram-se os seguintes objetivos específicos:

1. Aplicar a MANOVA e verificar a comparação entre as médias das diferentes variáveis simultaneamente no Índice Desenvolvimento Sustentável para Municípios do Estado de Mato Grosso;

2. Aplicar a ACP aos dados do Índice Desenvolvimento Sustentável para Municípios do Estado de Mato Grosso com objetivo de redimensionar as variáveis originais;

3. Analisar e interpretar os gráficos biplot aplicado as variáveis do IDSM.

2 Fundamentação Teórica

2.1 Técnicas da análise Multivariada

De acordo com Hair et al. (2009) a análise multivariada é uma ferramenta estatística que processa informações, de modo a simplificar a estrutura dos dados e a sintetizar informações, quando o número de variáveis envolvidas é muito grande, facilitando o entendimento do relacionamento existente entre as variáveis do processo. A análise multivariada auxilia na formulação de questões relativamente complexas de forma específica e precisa, possibilitando a condução de pesquisas teoricamente significativas

Segundo Steiner (1995) a necessidade de entender a relação entre diversas variáveis aleatórias faz da análise multivariada uma metodologia com grande potencial de uso. Para Lourenço e Matias (2000) as técnicas estatísticas multivariadas são mais complexas do que aquelas da estatística univariada. Além disso, apesar de uma razoável complexidade teórica, fundamentada na matemática, as técnicas multivariadas, por permitirem o tratamento de diversas variáveis ao mesmo tempo, podem oferecer ao pesquisador um material bastante robusto para a análise dos dados da pesquisa.

2.1.1 Normalidade Multivariada

Para Gouvêa, Prearo e Romeiro (2012) as técnicas de análise multivariada que se utilizam de variáveis métricas e testes estatísticos, a normalidade multivariada é a condição mais fundamental de aplicação. Entretanto, no caso da análise de regressão, há a premissa de normalidade univariada, considerando-se apenas a variável referente aos resíduos.

A distribuição normal multivariada assumirá a forma de sinos tridimensionais simétricos nas seguintes condições: quando o eixo x apresenta valores de uma determinada variável, o eixo y apresenta contagem para cada valor da variável em x, e o eixo z apresenta valores de qualquer outra variável em consideração.

Entretanto, Johnson e Wichern (2002) alertam que, para dados reais, a presença de variáveis com distribuição normal multivariada exata dificilmente ocorre. Nesses casos, a densidade normal é frequentemente uma aproximação útil à verdadeira distribuição da população.

2.1.2 Testes de Normalidade Multivariada

Uma das fases mais delicadas no planejamento de experimentos é a fase inicial, quando são feitas as suposições a serem válidas para a análise dos dados, ou seja, determina-

se um modelo ao qual supõe-se que se ajustem aos dados. Os métodos usuais nesta fase são os gráficos box-plot, esquemas de ramos e folhas e testes para detectar a possível distribuição dos dados, ou seja, a adequação, ou não, dos dados à uma determinada distribuição teórica. Na maior parte da metodologia utilizada nos trabalhos estatístico, essa pressuposição refere-se à normalidade dos dados, homogeneidade de variâncias e não existência de dados discrepantes.

A validade dos procedimentos a serem utilizados, em geral está associada á possibilidade de assumir uma determinada distribuição teórica, geralmente a distribuição normal. Quando isso não ocorre, transformações de dados podem ser utilizadas em alguns casos particulares. Andrews, Gnanadesikan e Warner (1971) apresentam uma extensão do método de Box e Cox para a obtenção de transformações de dados multivariados mas tais transformações podem não ser adequadas devido à impossibilidade de se obterem conclusões confiáveis quando se realiza a transformações inversa.

Segundo Mardia (1970), os testes relacionados à análise de variância multivariada são, em geral, robustos mesmo quando a normalidade não ocorre, o mesmo não acontece com os teste de igualdade das matrizes de covariâncias. A rejeição do ajuste ao modelo teórico aos dados pode ainda levar ao uso de procedimentos que não são baseados em qualquer modelo específico, ou seja, aos métodos não-paramétricos (ou de distribuição livre), em que não se assume uma forma específica de distribuição dos dados e sim formas gerais para tal distribuição, como a simétrica, por exemplo.

2.1.3 Teste de Mardia

O teste de Mardia (1970), baseia-se nas medidas de assimetria $\hat{y}_{1,p}$ e curtose $\hat{y}_{2,p}$ utilizando a distância de Mahalanobis. A assimetria é dada por $\hat{y}_{1,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i=1}^{n} m_{ij}^3$ e a curtose por $\hat{y}_{2,p} = \frac{1}{n} \sum_{i=1}^{n} m_{ij}^2$, onde $m_{ij} = (x_i - \bar{x})' \, \mathbf{\Sigma}^{-1} \, (x_j - \bar{x})$ é o quadrado da distância de Mahalanobis e p o número de variáveis. O teste estatístico para a assimetria $(\frac{\nu}{6})$ $\hat{y}_{1,p}$ tem uma distribuição aproximada da qui-quadrado (χ^2) com um número de graus de liberdade dados por

$$\frac{p(p+1)(p+2)}{6}.$$

Para amostra de dimensão inferior a 20, Mardia introduziu a seguinte correção no teste da assimetria para controlar o erro tipo I. Assim para amostras pequenas a estatística da assimetria é dada por $(\frac{nk}{6})\hat{y}_{1,p}$, onde $k = \frac{(p+1)(n+1)(n+3)}{(n(n+1)(p+1)-6)}$ esta estatística mantém a distribuição de (χ^2) com $\frac{p(p+1)(p+2)}{6}$ graus de liberdade. Uma distribuição multivariada normal deverá ter

$$\hat{y}_{1,p} = 0$$

$$\hat{y}_{2,p} = p(p+2).$$

Segundo Sharma (1995) há poucos métodos disponíveis para testar a normalidade multivariada. O índice de Mardia parece ser o teste de normalidade multivariada mais disponível para os usuários de software estatísticos. Baseado nas funções de assimetria e curtose, o índice de Mardia está disponível no pacote estatístico LISREL e no pacote estatístico EQS.

2.1.4 Teste de Henze-Zirkler

O teste de Henze Zirkler (1990) é baseado numa distância funcional não negativa que mede a distância entre duas funções de distribuição. Quando a distribuição dos dados é normal multivariada a distribuição da estatística (HZ) do teste tem distribuição log-normal aproximada.
A estatística é dada pela seguinte fórmula:

$$HZ = \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}\epsilon^{\frac{\beta^2}{2}Dij} - 2(1+\beta^2)^{-\frac{p}{2}}\sum_{i=1}^{n}\epsilon^{-\frac{\beta^2}{2(1+\beta^2)}D_i} + n(1+2\beta^2)^{-\frac{p}{2}},$$

onde: p é o número de variáveis;

$$\beta = \frac{1}{\sqrt{2}}(\frac{n(2p+1)}{4})^{\frac{1}{p+4}}; D_{ij} = (x_i - x_j)^{'}\mathbf{\Sigma}^{-1}(x_i - x_j); D_i = (x_i - \bar{x})^{'}\mathbf{\Sigma}^{-1}(x_i - \bar{x}) = m_{ij},$$

D_i e D_{ij} representam, respectivamente, as distâncias de Mahalanobis entre a observação de ordem i e o centróide e entre observação de ordem i e a de ordem j. Se os dados forem multivariados normais a distribuição HZ tem média μ e variância σ^2 dadas pelas seguintes fórmulas:

$$\mu = 1 - \frac{-\frac{p}{a2}(1 + p\beta^{\frac{2}{a}} + (p(p+2)\beta^4))}{2a^2};$$

$$\sigma^2 = 2(1+4\beta^2)^{\frac{-p}{2}} + \frac{2a^{-p}(1+2p\beta^4)}{a^2} + \frac{3p(p+2)\beta^8}{4a^4} - 4w\beta^{-\frac{p}{2}}(1 + \frac{3p\beta^4}{2w\beta} + \frac{p(p+2)\beta^8}{2w\beta^2});$$

em que:

$$a = 1 + 2\beta^2 \;\; e \;\; w\beta = (1+\beta^2)(1+3\beta^2).$$

A média e a variância log-normalizadas das estatística HZ são dadas por:

$$ln(\mu) = ln(\sqrt{\frac{\mu^4}{\sigma^2 + \mu^2}})$$

$$ln(\sigma^2) = ln(\frac{\sigma^2 + \mu^2}{\sigma^2})$$

Com os parâmetros μ e σ da distribuição log-normalizada podemos testar a significância da normalidade multivariada com o teste de Wald sendo z dado pela equação seguinte:

$$z = \frac{ln(HZ) - ln(\mu)}{ln(\sigma)}, \tag{2.1}$$

em que ln é o logaritmo na base 10.

2.1.5 Teste de Royston

Royston (1992) propôs uma extensão do teste Shapiro-Wilk/Shapiro-Francia (SW/SF) para testar a normalidade multivariada. Seja w_j o teste SW/SF para a variável de ordem j com $j = 1, 2, 3, ..., p$ e z_j os valores obtidos pela transformação para normal proposta:
se $4 \leq n \leq 11$ então $x = n$ e $w_j = -ln[\gamma - ln(1 - w_j)]$, para $12 \leq n \leq 2000$ então $x = ln(n)$ e $w_j = ln(1 - w_j)$.

Após usar as equações os valores transformados de cada variável aleatória são dados pela seguinte equação $Z_j = \frac{w_j - \mu}{\sigma}$, onde γ, μ, e σ resultam das aproximações polinomiais dadas pelas equações:

$$\begin{aligned} \gamma &= a_{0\gamma} + a_{1\gamma}x + a_{2\gamma}x^2 + \ldots + a_{d\gamma}x^d \\ \mu &= a_{0\mu} + a_{1\mu}x + a_{2\mu}x^2 + \ldots + a_{d\mu}x^d \\ ln(\sigma) &= a_{0\sigma} + a_{1\sigma}x + a_{2\sigma}x^2 + \ldots + a_{d\sigma}x^d \end{aligned}$$

A estatística (H) de Royston (1992) para o teste de normalidade multivariada é dada por:

$$H = \frac{e\sum_{j=1}^{p}\psi_j}{p} \sim \chi^2_\nu \tag{2.2}$$

onde e são graus de liberdade dados por $e = \frac{p}{1+(p-1)\bar{c}}$ e $\psi_j = \{\phi^{-1}[\frac{\phi(-z_j)}{2}]\}^2$ com $j = 1, 2, ..., p$ para prosseguir com o cálculo de H determinamos o termo $\bar{c}$ da seguinte forma: sendo R a matriz das correlações e a correlação entre variáveis de ordem i e j

$$\bar{c} = \sum_i \sum_j \frac{c_{ij}}{p(p-1)}, \{c_{ij}\}_{i \neq j},$$

em que

$$c_{ij} = \begin{cases} g(r_{ij}, n) & \text{para } i \neq j; \\ 1 & \text{para } i = j. \end{cases}$$

impondo a g(.) as seguintes restrições g(0, n) = 0 e g(1, n) = 1 a função g(.) é definida da seguinte forma:

$$g(r, n) = \sigma^{\lambda}[1 - \frac{\mu}{\nu}(1 - \sigma)^{\mu}],$$

os valores dos parâmetros $\mu = 0{,}715$, $\lambda = 5$ foram estimados por estudos de simulação e ν é uma função cúbica da seguinte forma $\nu(n) = 0,21364 + 0,015124x^2 - 0,0018034x^2$, onde $x = ln(n)$.

Para Johnson e Wichern (2002), as pesquisas referentes à normalidade podem se concentrar apenas em variáveis isoladas ou grupos bivariados (Distribuições marginais e scatterplots), pois é difícil construir um bom teste para normalidade conjunta em mais do que duas dimensões. No caso multivariado, os testes de normalidade univariada têm como principal objetivo verificar a normalidade de distribuições marginais. Dentre eles, tem-se o exame do histograma e das causas da distribuição e a verificação de normalidade através de gráficos, como por exemplo, o Q-Qplot (Quantis vs Quantis plot). Entretanto, as verificações gráficas têm utilidade apenas nos casos em que o ajuste de uma determinada distribuição teórica a um conjunto de dados é graficamente óbvio, ou ainda quando existem dados muito discrepantes em relação à distribuição proposta do ajuste, a subjetividade do método pode levar a conclusões diferentes, dependendo do pesquisador.

Shapiro e Wilk (1965) propõem que o método gráfico seja complementado por testes objetivos. Um teste bastante citado na literatura baseando na regressão das observações ordenadas contra os valores das estatísticas de ordem da distribuição padronizada assumida, comparações entre os diversos testes para normalidade são feitas em (SHAPIRO; WILK; CHEN, 1968).

2.1.6 Distribuição Normal Multivariada

Para Prado (2016) uma generalização da densidade normal com diversas dimensões desempenha um papel fundamental na análise multivariada. Embora os dados reais não sejam exatamente normal multivariados, a densidade normal é frequentemente uma aproximação usual para a verdadeira distribuição da população. Uma vantagem da distribuição normal multivariada decorre do fato de que ela é matematicamente tratável e bons resultados pode ser obtidos. Isto frequentemente não é o caso de outras distribuições geradoras de dados. É claro que, a atração matemática por si só é de pouca utilidade para o experimentador.

Segundo Rao (1952), tentativas iniciais de generalização das análises univariadas de variância para o caso de variáveis múltiplas foi dado por Wishart (1928), o qual estudou a distribuição amostral simultânea de variâncias e covariâncias em amostras de uma população normal multivariada. Posteriormente, Hotelling e Frankel (1931) verificam a distribuição T, que é uma extensão natural da distribuição de t-Student para uma

população normal multivariada. Wilks (1932), seguindo o método da razão de verossimilhança (Neyman e Pearson, 1908 e 1931; Pearson e Neyman 1930), obteve generalizações apropriadas na análise de variância aplicáveis a diversas variáveis. A estatística proposta por estes autores tem sido útil em uma variedade de problemas.

Segundo Demétrio (1985), a análise de variância multivariada (MANOVA), além de fornecer resultados com base na análise conjunta de todos as variáveis utilizadas, levando-se em consideração um nível de significância conhecido, permite estimar a melhor combinação de variáveis que leva a um valor de F máximo.

Quando um conjunto de variáveis apresenta respostas diferentes e não existe correlação entre elas, a análise de variância univariada ANOVA é o procedimento correto; entretanto, quando as variáveis são mutuamente correlacionadas, o que geralmente acontece, deve-se pressupor multi-normalidade dos dados e então realiza-se a análise multivariada da variância MANOVA(Multivariate Analysis Of Variance), que permite o melhor aproveitamento das informações conjunta das variáveis. A MANOVA é uma importante ferramenta que nos permite estimar e comparar médias de duas ou mais populações normais multivariadas independentes (JÚNIOR et al., 2005). A formulação de um teste estatístico para comparar vetores de médias, depende da partição do total da variância em: variância devido ao efeito de tratamentos e variância devido ao erro . Um ponto relevante da análise multivariada é o aproveitamento da informação conjunta das variáveis envolvidas (JOHNSON; WICHERN, 2002).

A MANOVA é usada para investigar se os vetores de médias dos grupos (ou tratamentos) são iguais, ou seja, faz a comparação entre as médias das diferentes variáveis simultaneamente. A MANOVA é uma extensão da análise de variância (ANOVA) e ambas utilizam dois passos sequências:

a) Testa-se a hipótese global de igualdade de médias entre os grupos;

b) Se o resultado do passo anterior for significativo, utilizam-se testes adicionais no sentido de explicar as diferenças entre os grupos (comparações múltiplas)

A MANOVA, no entanto, tem vantagens sobre a realização de sucessivas ANOVAS para diferentes variáveis, pois na primeira considera-se o nível de significância conjunto dos testes e aproveita-se as informações conjuntas das variáveis envolvidas.

De acordo com Hair et al. (2006), para os procedimentos de teste multivariado de MANOVA serem válidos, quatro suposições devem ser atendidas, que podem ser entendidas como generalizações dos pressupostos da ANOVA:

i) Modelo aditivo para efeitos de tratamentos, blocos (se houver) e erro;

ii) Independência dos erros;

iii) Igualdade da matriz de covariância $\mathbf{\Sigma}$ para todas as amostras;

iv) O conjunto de p variáveis dependentes deve seguir uma distribuição normal multivariada, ou seja, os erros devem ter distribuição normal multivariada, com matriz de $\boldsymbol{\Sigma}(\epsilon_i \sim N_p(\boldsymbol{\phi}; \boldsymbol{\Sigma}))$, sendo $\boldsymbol{\phi}$ o vetor nulo essa condição tem relevância diminuída quando as amostras são de grande dimensão.

A MANOVA pode ser utilizada para qualquer delineamento experimental, sem apresentar dificuldades adicionais. Mas é preciso trabalhar com todas as variáveis simultaneamente, obtendo para elas matrizes de somas de quadrados e somas de produtos cruzados (SQPC).

Considere, por exemplo, um delineamento inteiramente casualizado (DIC) com g grupos (tratamentos) onde avaliou-se p variáveis de n observações, tal que $n = n_1+n_2, ..., n_g$, sendo n_k o número de indivíduos do grupo k, $k = 1, 2, ..., g$. assim, o modelo linear de análise de variância multivariada para este caso é dado matricialmente por:

$$\boldsymbol{Y}_{(n\times p)} = \boldsymbol{X}_{(n\times(g+1))}\boldsymbol{\beta}_{((g+1)\times p)} + \boldsymbol{E}_{(n\times p)}, \tag{2.3}$$

sendo:

$$\boldsymbol{Y} = \begin{bmatrix} y_{111} & y_{121} & \cdots & y_{1p1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{11n_1} & y_{12n_1} & \cdots & y_{1pn_1} \\ y_{211} & y_{221} & \cdots & y_{1p1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{21n_2} & y_{22n_2} & \cdots & y_{2pn_2} \\ & \vdots & & \\ y_{g11} & y_{g21} & \cdots & y_{gp1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{g1n_g} & y_{g2n_g} & \cdots & y_{gpn_g} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 1 & \ldots & 0 \\ & & \vdots & & \\ 1 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} \dots & \beta_{0p} \\ \beta_{11} & \beta_{12} \dots & \beta_{1p} \\ \beta_{21} & \beta_{22} \dots & \beta_{2p} \\ & \vdots & \\ \beta_{g1} & \beta_{g2} \dots & \beta_{gp} \end{bmatrix}$$

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{111} & \epsilon_{121} & \dots & \epsilon_{1p1} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{11n_1} & \epsilon_{12n_1} & \dots & \epsilon_{1pn_1} \\ \epsilon_{211} & \epsilon_{221} & \dots & \epsilon_{1p1} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{21n_2} & \epsilon_{22n_2} & \dots & \epsilon_{2pn_2} \\ & \vdots & & \\ \epsilon_{g11} & \epsilon_{g21} & \dots & \epsilon_{gp1} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{g1n_g} & \epsilon_{g2n_g} & \dots & \epsilon_{gpn_g} \end{bmatrix}$$

onde os índices kjm são: k = grupo $(1, ..., g)$, j = variáveis $(1, ..., p)$ e m = repetição $(1, ..., n_k)$. O modelo linear de análise de variância multivariada também pode ser escrito como:

$$\boldsymbol{y}_{km} = \boldsymbol{\mu} + \boldsymbol{\tau}_k + \epsilon_{km}, \tag{2.4}$$

$k = 1, 2, ..., g$ e $m = 1, 2, ..., n_k$, em que: $\boldsymbol{y}_{km}$ é o vetor $(p \times 1)$ de observações do k-ésimo grupo, do m-ésimo indivíduo; $\boldsymbol{\mu}$ é o vetor de constantes $(p \times 1)$ comuns a todos os grupo; $\boldsymbol{\tau}_k$ representa o vetor $(p \times 1)$ de efeitos do k-ésimo grupo; ϵ_{km} é o vetor de erros associado a y_{km} e tem distribuição $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ para todo k e m.

Segundo Sartorio (2008) para estimar os parâmetros de um modelo com um único fator (one way) pode-se estabelecer a restrição $\sum_g^{k=1} n_k \boldsymbol{\tau}_k = 0$. Observe que cada componente do modelo é um vetor de p componentes e os erros associados ao vetor $\boldsymbol{y}_{km}$ são correlacionados.

Antes de calcular as estatísticas de teste para diferenças de médias dos grupos, o pesquisador deve primeiramente determinar se as medidas dependentes são significativamente correlacionadas. Se forem, a hipótese a ser testada é:

$$\begin{aligned} H_0 &: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g \\ H_1 &: \boldsymbol{\mu}_l \neq \boldsymbol{\mu}_k, l \neq k, l, k = 1, 2, \ldots, g. \end{aligned}$$

H_0 equivale a testar que os vetores de médias dos g grupos são todos iguais, e H_1, que ao menos um par de grupos é diferente em pelo menos uma variável.

Existem vários critérios para testes de hipóteses multivariados, contudo, segundo (JOHNSON; WICHERN, 2002; REIS, 1997), os quatro mais utilizados são:

2.1.7 Lambda de Wilks

Segundo Cecatto e Belfiore (2015) lambda de Wilks, que varia de 0 a 1, propicia a avaliação da existência de diferenças de médias entre os grupos para cada variável. Valores elevados desta estatística indicam ausência de diferenças entre os grupos, e sua expressão é dada por:

$$\Lambda = |\boldsymbol{W}|/|\boldsymbol{T}| \tag{2.5}$$

em que;
$|\boldsymbol{W}|$ = determinante da matriz das somas de quadrados e produtos cruzados dentro da amostra;
$|\boldsymbol{T}|$ = determinante da matriz das somas totais de quadrados e produtos cruzados.
Se Λ for pequeno, a variação dentro das amostras é baixa em comparação com a variação total. Isso quer dizer que as amostras não vêm de populações com o mesmo vetor de médias.

2.1.8 Maior raiz de Roy

Para Song (2013) a base para este teste é que se a combinação linear das variáveis de X_1 à X_p que maximiza a razão entre a soma dos quadrados entre amostras e a soma dos quadrados dentro das amostras é encontrada, então essa razão máxima é igual a ao autovalor λ_1. Esta estatística é definida por:

$$\lambda_1(1 - \lambda_1) \tag{2.6}$$

Portanto, o autovalor máximo λ_1 pode ser uma boa estatística para testar se a variação entre amostras é significantemente grande, e que há, portanto, evidência de que as amostras sendo consideradas não vêm de populações com o mesmo vetor médio. O valor λ_1 é comparado com um valor tabelado da tabela F. Rejeitamos a igualdade para valores grandes de λ_1.

2.1.9 Traço de Pillai

Bray e Maxwell (1985) concluíram que quando tamanhos amostrais são iguais o traço de Pillai é o mais robusto a violação das hipóteses, contudo, quando os tamanhos amostrais são diferentes, essa estatística é afetada pela violação da hipótese de igualdade

das matrizes de covariâncias. A estatística é dada por:

$$V = \sum_{i=1}^{p} \lambda/(1+\lambda_i), \tag{2.7}$$

Onde os λ_i são autovalores obtidos, $i = 1, 2, \ldots, p$. Temos novamente que valores grandes de V fornecem evidências de que as amostras consideradas vêm de populações com vetores médias diferentes.

2.1.10 Traço de Lawley-Hotelling

O traço de Lawley-Hotelling, (Lawley (1938); Hotelling (1951)) é dado por:

$$U = \sum_{i=1}^{p} \lambda_i \tag{2.8}$$

Essa estatística é apenas a soma dos autovalores da matriz $\boldsymbol{W}^{-1}\boldsymbol{B}$, onde grandes valores fornecem evidência contra a hipótese nula de igualdade.
Observação:
$\boldsymbol{W}$ = matriz das somas de quadrados e produtos cruzados dentro da amostra;
$\boldsymbol{B} = \boldsymbol{T} - \boldsymbol{W}$;
$\boldsymbol{T}$ = matriz das somas totais de quadrados e produtos cruzados.

A Tabela 1 fornece as estatísticas de comparação com os valores tabelados da tabela F-Snedecor, dos quatro testes analisados.

Tabela 1 – Estatísticas multivariadas para testes de hipóteses sobre vetores de médias a suas equivalências aproximadas a estatística com o teste F, com ν_W e ν_B graus de liberdade do resíduo e de efeito dos grupos sobre as variáveis resposta.

Critério	Estatística	Aproximações F	Graus de Liberdade
Wilks	$\Lambda = \begin{cases} \frac{\lvert\mathbf{W}\rvert}{\lvert\mathbf{W}+\mathbf{B}\rvert} \\ \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \end{cases}$	$F = \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \frac{rt-2f}{p\nu_B}$	$\begin{cases} \nu_1 = p\nu_H \\ \nu_2 = rt - 2f \end{cases}$
Pillai	$V = \begin{cases} tr\left[\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1}\right] \\ \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} \end{cases}$	$F = \begin{cases} \left(\frac{V}{s-V}\right)\left(\frac{2N+s+1}{2m+s+1}\right) \\ \left(\frac{V}{s-V}\right)\left[\frac{s(\nu_W-\nu_B+s)}{p\nu_B}\right] \end{cases}$	$\begin{cases} \nu_1 = s(2m+s+1) \\ \nu_2 = s(2N+s+1) \\ e \\ \nu_1 = s(2m+s+1) \\ \nu_2 = s(2N+s+1) \end{cases}$
Hotelling Lawley	$U = \begin{cases} tr(W^{-1}B) \\ \sum_{i=1}^{s} \lambda_i \end{cases}$	$F = \begin{cases} \frac{U}{c} \\ \frac{2(sN+1)U}{s^2(2m+s+1)} \end{cases}$	$\begin{cases} \nu_1 = a \\ \nu_2 = b \\ e \\ \nu_1 = s(2m+s+1) \\ \nu_2 = 2(sN+1) \end{cases}$
Roy	$\theta = \frac{\lambda_1}{1+\lambda_1}$	$F = \frac{(\nu_W-d+\nu_B)\lambda_1}{d}$	$\begin{cases} \nu_1 = d \\ \nu_2 = \nu_W - d + \nu_B \end{cases}$

$r = \nu_W - (p - \nu_B + 1)/2$ $\quad f = (p\nu_B - 2)/4$ $\quad t = \begin{cases} \sqrt{\frac{p^2\nu_B^2-4}{p^2+\nu_B^2-5}} & se\ p^2+\nu_B^2-5>0 \\ 1 & caso\ contrário \end{cases}$

$s = min(p, \nu_B)$ $\quad N = (\nu_E - p - 1)/2$ $\quad m = (|p-\nu_\beta|-1)/2$

$a = p\nu_B$ $\quad b = 4 + (a+2)/(\Upsilon - 1)$ $\quad \Upsilon = \frac{(\nu_W+\nu_B-p-1)}{(\nu_W-p-3)(\nu_W-p)}$

$c = a(b-2)/[b(\nu_B - p - 1)]$ $\quad d = \max(p, \nu_B)$

Os quatro testes têm níveis de significância similares, geralmente, o que nos dá possibilidade de utilizar qualquer um deles quando a distribuição das p variáveis é aproximadamente normal multivariada com a mesma matriz de covariância dentro das amostras para todas as g populações das quais as amostras das m observações foram extraídas, além da independência entre os grupos. Tais testes são considerados robustos (isto é, podem ser aplicados mesmo se as suposições não se verificarem na totalidade dos grupos ou variáveis)

se os tamanhos amostrais forem aproximadamente iguais para as m amostras. No entanto, se houver alguma questão sobre essa suposição, estudos sugerem que a estatística de Pillai possa ser mais eficiente. Altas correlações entre as variáveis sugerem maior confiança no teste de Pillai; baixas correlações sugerem escolher o teste de Roy.

Apesar disso, os quatro testes costumam fornecer conclusões similares e nenhum deles pode ser considerado "o melhor", em geral. Cada teste capta diferentes características das diferenças entre as médias.

2.2 Análise dos Componentes Principais

A técnica de Analise Componentes Principais (ACP) foi originalmente descrita por Karl Pearson, em 1901, em um artigo onde deu ênfase à sua utilização no ajustamento de um subespaço a uma nuvem de pontos. Posteriormente, a técnica foi consolidada por Hotelling (1933), para o propósito particular de analisar estruturas de correlações (MORRISON, 1976; MARDIA; KENT; BIBBY, 1979; MANLY, 1986; CRUZ, 1990). Entretanto, o uso da análise só foi difundida após desenvolvimento dos computadores e atualmente, devido a grande disponibilidade de recursos de computadores sofisticados e de software aplicados, a técnica tornou-se amplamente disponível e utilizada nas várias áreas da ciência.

A técnica de ACP procura explicar a estrutura de variâncias-covariâncias através de poucas combinações lineares das variáveis originais, com os objetivos de reduzir os dados, colocá-los numa forma mais adequada para análise, evidenciar as tendências e facilitar sua interpretação. Segundo Liberato (1995), a utilização da análise de componentes principais tem por finalidade proporcionar simplificação estrutural dos dados, de modo que a diversidade, influenciada a princípio por um conjunto p-dimensional (p = números de caracter considerados no estudo), possa ser representada geometricamente em duas ou três dimensão de fácil interpretação. Ou ainda, a análise por componentes principais, segundo Cruz et al. (1994), consiste em transformar um conjunto original de variáveis em outro conjunto, de dimensões equivalentes, mas com propriedades importantes de grande interesse em certos estudos.

Segundo Montgomery (2009) os componentes principais de um conjunto de variáveis por exemplo, $x_1, x_2, ..., x_p$ são um conjunto especial de combinações lineares representadas por

$$\begin{aligned} z_1 &= \boldsymbol{l}_1' x = l_{11}x_1 + l_{12}x_2 + \ldots + l_{1p}x_p \\ z_2 &= \boldsymbol{l}_2' x = l_{21}x_1 + l_{22}x_2 + \ldots + l_{2p}x_p \\ &\vdots \\ z_p &= \boldsymbol{l}_p' x = l_{p1}x_1 + l_{p2}x_2 + \ldots + l_{pp}x_p, \end{aligned}$$

em que l_{ij} são constantes a serem determinadas e representam a rotação que será realizada. Geometricamente, as variáveis dos componentes principais (escores), isto é, $z_1, z_2, ..., z_p$ são os novos eixos que representam as diretrizes de variabilidade máxima. É importante dizer que, para criar esse novo sistema é necessário levar em consideração algumas restrições. A primeira restrição é que a soma das constantes c_{ij} ao quadrado devem ser iguais a *um*. Por exemplo, para o primeiro componente (z_1) a única restrição é

$$l_{11}^2 + l_{12}^2 + ... + l_{1p}^2 = 1. \tag{2.9}$$

A partir do segundo componente z_2, além da restrição das constantes $l_{2j}, j = 1, 2, ..., p$, esses novos componentes devem obedecer ao critério de que sejam ortogonais às anteriores, ou seja, as constantes do vetor l_{1j} com as constantes do vetor l_{1j} devem ter o produto escalar igual à zero, isto é, $\sum_{j=1}^{p} l_{1j} l_{2j} = 0$.

Manly (2008) estabelece os seguintes passos para realizar análise de componentes principais:

1º) - Comece codificando as variáveis $x_1, x_2, ..., x_p$ para terem médias zero e variâncias unitárias.

2º) - Calcule a matriz de covariâncias $\boldsymbol{S}_{p\times p}$. Essa é uma matriz de correlação se o primeiro passo já foi executado.

3º) - Encontre as estimativas dos autovalores $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p$ e os correspondentes autovetores $\hat{\epsilon}_1, \hat{\epsilon}_2, ..., \hat{\epsilon}_p$, os coeficientes da j-ésima componente principal são então os elementos de ϵ_j, enquanto que λ_j é a sua variância.

4º) - Descarte quaisquer componentes que explicam apenas uma pequena proporção da variação dos dados.

A análise é realizada com o intuito de resumir o padrão de correlação entre as variáveis e muitas vezes é possível chegar a conjuntos de variáveis que sejam não correlacionadas uns com os outros, levando assim a um agrupamento delas.

A ACP depende somente da matriz de covariâncias ($\boldsymbol{\Sigma}$) ou da matriz de correlação ($\boldsymbol{\rho}$) de $X_1, ..., X_p$. Não requer qualquer suposição sobre a forma da distribuição multivariada dessas variáveis. Se a normalidade existe, a análise é engrandecida, senão ela ainda vale a pena (TABACHNICK; FIDELL; OSTERLIND, 2001).

2.3 Biplot

Em muitas áreas do conhecimento os pesquisadores utilizam os primeiros e mais importantes componentes de uma análise para agrupar observações e itens de acordo com a representação em duas ou no máximo três dimensões. Um gráfico bidimensional pode ser obtido representando as informações das observações para duas variáveis em um diagrama,

já para um gráfico biplot, a ideia é de acrescentar informações sobre várias variáveis em um único gráfico obtido pelos componentes principais.

Biplot é uma técnica multivariada proposta por Gabriel (1971) com o objetivo de representar graficamente uma matriz de dados, de tal forma que esta representação permita visualizar em um plano as relações e inter-relações entre as linhas e colunas desta matriz. Fatorando a matriz de dados original pela Decomposição em Valores Singulares (DVS) como a soma de produtos de matrizes que contém os marcadores de linhas e colunas que constituem os elementos para a representação gráfica, consegue-se uma avaliação visual da estrutura da matriz de dados (GOWER; HAND, 1996).

Independente da forma de fatoração, as aplicações do método Biplot dividem-se em representação da matriz de dados, com finalidade descritiva e diagnóstico de modelos. A técnica é bastante útil, pois o gráfico utilizado para representar simultaneamente as linhas e colunas de uma matriz de dados pode indicar a existência de agrupamentos entre as observações, assim como mostrar as variâncias e correlações entre as variáveis (CÁRDENAS; GALINDO; VICENTE-VILLARDÓN, 2007).

3 Materiais e Métodos

3.1 Materiais

Os dados utilizados são referentes ao banco de dados sobre as variáveis Social, Demográfico, Econômico Pollins, Ambiental e Cultural que compõem o Índice Desenvolvimento Sustentável para Municípios do estado de Mato Grosso com 141 cidades e cinco Mesorregião. Sera feito testes de normalidade multivariada e aplicação de técnicas da Analise de Variância Multivariada - MANOVA e Analise de Componentes Principais - ACP. O tratamento dos dados no decorrer do trabalho bem como as análises estatísticas foram realizadas com auxilio do software (R Core Team, 2016).

Conforme Hair et al. (2009) a análise multivariada auxilia na formulação específica e precisa de questões relativamente complexas, possibilitando a condução de pesquisas teoricamente significativas.

A análise estatística multivariada envolve um conjunto de métodos estatísticos e matemáticos, destinada a descrever e interpretar os dados que provém da observação de várias variáveis estudadas conjuntamente e algumas estruturas de correlação (JOHNSON; WICHERN, 2002).

Conforme Giri (2004), algebricamente os componentes principais são combinações lineares das variáveis originais.

3.2 Métodos

3.2.1 MANOVA

Para proceder a MANOVA, calcula-se:

i) A matriz $\boldsymbol{T}(p\times p)$ de soma dos quadrados e produtos cruzados total ($SQPCT_{Total}$):

$$\boldsymbol{T} = \sum_{k=1}^{g}\sum_{m=1}^{n_k}(y_{km}-\bar{y})(y_{km}-\bar{y})'; \tag{3.1}$$

ii) A matriz $\boldsymbol{H}(p\times p)$ de soma dos quadrados e produtos cruzados entre os grupos ou matriz de soma dos quadrados e produtos cruzados da hipótese ($SQPC_{Hip}$):

$$\boldsymbol{H} = \sum_{k=1}^{g} n_k(\bar{y}_k-\bar{y})(\bar{y}_k-\bar{y})'; \tag{3.2}$$

iii) A matriz $\boldsymbol{W}(p \times p)$ de soma dos quadrados e produtos cruzados dentro dos grupos ou matriz de soma de quadrados e produtos cruzados do resíduos ($SQPC_{Ris}$):

$$\boldsymbol{W} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (y_{km} - \bar{y}_k)(y_{km} - \bar{y}_k)^{'} = \sum_{k=1}^{g} (n_k - 1)\boldsymbol{S}_k \tag{3.3}$$

em que $\boldsymbol{S}_k$ é a matriz de variâncias e covariâncias amostrais do k-ésimo grupo; $\bar{y}$ é o vetor $(p \times 1)$ de constantes amostrais comuns a todos os grupos; e $\bar{y}_k$ é o vetor $(p \times 1)$ de médias amostrais do k-ésimo grupo.

Vale lembrar que, as matrizes, $\boldsymbol{T}$, $\boldsymbol{H}$ e $\boldsymbol{W}$ são simétricas e que demonstra-se facilmente que $\boldsymbol{T} = \boldsymbol{H} + \boldsymbol{W}$. Logo, para se obter a mariz de SQPCRes basta calcular: $\boldsymbol{W} = \boldsymbol{T}$ - $\boldsymbol{H}$. Com estas matrizes estruturamos a MANOVA e apresentamos na Tabela 2.

Tabela 2 – Análise da Variância Multivariada - MANOVA para comparar vetores de médias das k populações

Fonte de Variação(FV)	**graus de liberdade(gl)**	**Matriz de SQPC**
Grupos	$g-1$	$\boldsymbol{H}$
Resíduo	$n-g$	$\boldsymbol{W}$
Total	$n-1$	$\boldsymbol{T} = \boldsymbol{H} + \boldsymbol{W}$

3.2.2 Obtenção dos Componentes Principais

Algebricamente, componentes principais são combinações lineares particulares das p variáveis aleatórias $X_1, X_2, ..., X_p$. Geometricamente, estas combinações lineares representam a seleção de um novo sistema de coordenadas obtidas pela rotação do sistema original com $X_1, X_2, ..., X_p$ como eixos. Os novos eixos representam as direções com variabilidade máxima e fornece uma descrição mais simples e mais parcimoniosa da estrutura de covariâncias (JOHNSON; WICHERN, 2002).

O seu desenvolvimento não necessita de normalidade. No entanto, a análise de componentes derivada de populações normais multivariadas têm suas interpretações usuais em termos de elipsóides de densidade constante (JOHNSON; WICHERN, 2002). Entretanto, embora a análise, formalmente não requeira a distribuição normal multivariada, ela é mais apropriada para variáveis quantitativas contínuas. Quando os dados são constituídos de contagem, razões, proporções ou percentagens, a transformação é recomendada para tornar sua distribuição mais apropriada, previamente à análise de componentes principais. Como exemplo,(STAUFFER; GARTON; STEINHORST, 1985) recomenda a transformação de arco seno da raiz quadrada para dados provenientes de percentagem e os dados de contagem a transformação de raiz quadrada (GOMES, 1984).

Seja o vetor aleatório $\boldsymbol{X}^{'} = [X_1, X_2 ... X_p]$ que tem a matriz de covariâncias ($\boldsymbol{S}$) com auto-valores ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$) e considerando as seguintes combinações lineares:

$$Y_1 \;=\; \boldsymbol{l}_1^{'}\boldsymbol{X} = \boldsymbol{l}_{11}X_1 + \boldsymbol{l}_{21}X_2 + \ldots + \boldsymbol{l}_{p1}X_p$$

$$\begin{aligned} Y_2 &= \boldsymbol{l}_2'\boldsymbol{X} = \boldsymbol{l}_{12}X_1 + \boldsymbol{l}_{22}X_2 + \ldots + \boldsymbol{l}_{p2}X_p \\ &\vdots \\ Y_p &= \boldsymbol{l}_p'\boldsymbol{X} = \boldsymbol{l}_{1p}X_1 + \boldsymbol{l}_{2p}X_2 + \ldots + \boldsymbol{l}_{pp}X_p \end{aligned}$$

Sendo:

$$\begin{aligned} Var(Y_i) &= \mathbf{1}_i'\boldsymbol{S}\mathbf{1}_i \qquad\qquad i = 1, 2, ..., p \\ Cov(Y_i, Y_k) &= \mathbf{1}_i'\boldsymbol{S}\mathbf{1}_k \qquad\qquad i = 1, 2, ..., p \end{aligned}$$

Os componentes principais são combinações lineares não correlacionadas,cujas variâncias são tão grandes quanto possível assim:

i) O primeiro componente principal (Y_1) é a combinação linear com variância máxima, isto é, a combinação linear $\boldsymbol{l}_1'\boldsymbol{X}$ que maximiza $Var(\boldsymbol{l}_1'\boldsymbol{X})$ sujeito a $\boldsymbol{l}_1'\boldsymbol{l}_1 = 1$

ii) O segundo componente principal (Y_2) é a combinação linear $\boldsymbol{l}_2'\boldsymbol{X}$ que maximiza $Var(\boldsymbol{l}_2'\boldsymbol{X})$, sujeito a $\boldsymbol{l}_2'\boldsymbol{l}_2$ e com $Cov(\boldsymbol{l}_1'\boldsymbol{X}, \boldsymbol{l}_2'\boldsymbol{X}) = 0$;

iii) O i-ésimo componente principal (Y_i) é a combinação linear $\boldsymbol{l}_i'\boldsymbol{X}$ que maximiza $Var(\boldsymbol{l}_i'\boldsymbol{X})$, sujeito a $\boldsymbol{l}_i'\boldsymbol{l}_i = 1$ e, em todos os casos, a $Cov(\boldsymbol{l}_i'\boldsymbol{X}, \boldsymbol{l}_k'\boldsymbol{X}) = 0$.

Desta forma, verifica-se que entre todos os componentes principais, Y_1 apresenta a maior variância, Y_2 a segunda maior e, assim sucessivamente, e independente entre si.

Assim, segundo Cruz et al. (1994), o problema estatístico consiste fundamentalmente em estimar os coeficientes de ponderação dos caracteres em cada componente e a variância a eles associada.
Sendo Y_1 o primeiro componente principal, sua variância é dada por:

$$Var(Y_1) = \boldsymbol{l}_1'\boldsymbol{S}\boldsymbol{l}_1.$$

O que se deseja é obter estimativas para o vetor $\boldsymbol{l}_1$ de tal forma que a variância de Y_1 seja a maior de todas. Para atingir este objetivo impõe-se a restrição $\boldsymbol{l}_1'\boldsymbol{l} = 1$, a qual é introduzida na expressão $Var(Y_1) = \boldsymbol{l}_1^1\boldsymbol{S}\mathbf{1}_1$ pelo multiplicador λ_1 de Lagrante. Assim:

$$W_1 = \boldsymbol{l}_1'\boldsymbol{S}\boldsymbol{l}_1 + \lambda_1(1 - \boldsymbol{l}_1'\boldsymbol{l}_1). \tag{3.4}$$

A solução que maximiza $Var(Y_1)$ é obtida pela derivação de W_1 em relação a $\mathbf{1}_1$ que é dada por:

$$|\boldsymbol{S} - \lambda_1\boldsymbol{I}| \quad a = 0.$$

A solução deste sistema deve ser tal que $\mathbf{1} \neq \mathbf{0}$, assim é necessário que o determinante de $(\boldsymbol{S} - \lambda_1\boldsymbol{I})$ seja nulo, para que o sistema se torne indeterminado e a solução possa ser escolhida entre aquelas que satisfaçam a condição $\boldsymbol{l}_1'\boldsymbol{l}_1 = 1$.

Sendo λ_1 o valor que satisfaz $|S - \lambda_1 \boldsymbol{I}| = 0$, então, por definição, λ_1 é a raiz característica (ou autovalor) de $\boldsymbol{S}$ e $\boldsymbol{l}_1$, o vetor característico (autovetor) associado. Sendo o vetor $\boldsymbol{l}_1^{'}$ o escolhido para maximizar $Var(Y_1)$, tem-se que λ_1 é o maior valor entre o conjunto de autovalores de $\boldsymbol{S}$.

A variância do segundo componente principal é dada por $Var(Y_2) = \boldsymbol{l}_2^{'}\boldsymbol{S}\boldsymbol{l}_2$. Para obtenção das estimativas do vetor $\boldsymbol{l}_2^{'}$, deve-se considerar as restrições $\boldsymbol{l}_2^{'}\boldsymbol{l}_2 = 1$ e $\boldsymbol{l}_2^{'}\boldsymbol{l}_1 = \boldsymbol{l}_1^{'}\boldsymbol{l}_2 = 0$, as quais são incorporadas na função de maximização por meio dos multiplicadores λ_2 e $\boldsymbol{\theta}$ de Lagrande. assim, é estabelecido que:

$$W_2 = \boldsymbol{l}_2^{'}\boldsymbol{S}\boldsymbol{l}_2 + \lambda_2(1 - \boldsymbol{l}_2^{'}\boldsymbol{l}_2) + \boldsymbol{\theta}\boldsymbol{l}_2^{'}\boldsymbol{l}_1. \tag{3.5}$$

A solução que maximizar $Var(Y_2)$, obtida pela derivação de W_2 em relação ao $\boldsymbol{l}_2$, é dada por:

$$(\boldsymbol{S} - \lambda_2 \boldsymbol{I})\boldsymbol{l}_2 = \boldsymbol{0},$$

em que λ_2 é a segunda maior raiz característica de $\boldsymbol{S}$ e $\boldsymbol{l}_2$ o seu autovetor associado.

As restrições consideradas neste segundo componente principal atendem aos seguintes propósitos:

i) A primeira restrição é necessária para garantir a unicidade de $\boldsymbol{l}_2$;

ii) A segunda restrição garante que $\boldsymbol{l}_1$ e $\boldsymbol{l}_2$ sejam ortogonais.

Os demais componentes principais são estimados de maneira análoga ao descrito para os dois primeiros.

3.2.3 Construindo biplot

A construção de um biplot bidimensional origina-se dos componentes principais amostrais, de $\boldsymbol{X}_{n\times p}$. Procura-se por uma aproximação $\boldsymbol{Y}_{n\times p}$ de posto 2 da matriz original $\boldsymbol{X}$, então obteremos um exato biplot bidimensional de $\boldsymbol{Y}$.

A melhor aproximação de posto 2 de $\boldsymbol{X}$ é obtida decompondo-a em valor singular (DVS), que consiste em escrevê-lo como produto de três matrizes:

$$\boldsymbol{X}_{n\times p} = \boldsymbol{U}_{n\times p}\boldsymbol{\Lambda}_{p\times p}\boldsymbol{V}_{p\times p}^{'} \tag{3.6}$$

em que $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$, que são valores singulares que obedecem $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$,
$\boldsymbol{U}$: matriz ortogonal de autovetores de $\boldsymbol{X}_{n\times p}\boldsymbol{X}_{p\times n}^{'}$;
$\boldsymbol{V}$: matriz ortogonal de autovetores de $\boldsymbol{X}_{p\times n}^{'}\boldsymbol{X}_{n\times p}$.
(JOHNSON; WICHERN, 2002)

Assim, pela decomposição em valores singulares

$$\boldsymbol{X} = \lambda_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{'} + \lambda_2 \boldsymbol{u}_2 \boldsymbol{v}_2^{'} + \ldots + \lambda_p \boldsymbol{u}_p \boldsymbol{v}_p^{'}$$

ou seja,

$$\boldsymbol{X} = \sum_{i=1}^{p} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^{'}$$

em que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, $\boldsymbol{u}_i$ é o vetor de autovetores associados ao i-ésimo autovalor da matriz $\boldsymbol{X}\boldsymbol{X}^{'}$ com dimensão $(n \times 1)$ e $\boldsymbol{v}_i^{'}$ é o vetor linha de autovetores associados ao i-ésimo autovalor da matriz $\boldsymbol{X}^{'}\boldsymbol{X}$ com dimensão $(1 \times p)$. Se considerarmos os 2 primeiros componentes, obteremos uma aproximação através de uma matriz de classificação igual a 2. Ou seja,

$$\boldsymbol{Y}_{n \times p} \approx \boldsymbol{Y}_{(2)} = \lambda_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{'} + \lambda_2 \boldsymbol{u}_2 \boldsymbol{v}_2^{'}$$

ou

$$\boldsymbol{Y}_{n \times p} \approx \boldsymbol{Y}_{(2)} = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1^{'} \\ \boldsymbol{v}_2^{'} \end{pmatrix}$$

Para obter o biplot da matriz $\boldsymbol{Y}$, é preciso escrevê-lo como o produto de duas matrizes $\boldsymbol{G}\boldsymbol{H}^{'}$, onde $\boldsymbol{G}$ é uma matriz $(n \times 2)$ e $\boldsymbol{H}$ é uma matriz $(2 \times p)$. Isso pode ser feito de várias formas, dependendo da fatoração da matriz $\boldsymbol{Y}$ em produtos de matrizes, tal que:

$$\boldsymbol{Y}_{n \times p} \approx (\boldsymbol{U}\boldsymbol{\Lambda}^{\alpha})(\boldsymbol{\Lambda}^{1-\alpha}\boldsymbol{V}) = \boldsymbol{G}\boldsymbol{H}^{'} \tag{3.7}$$

em que α variando geralmente em 0, $\frac{1}{2}$ e 1 determina a ênfase que será dada as linhas ou colunas de $\boldsymbol{Y}$. (GABRIEL, 1971) propôs 3 formas básicas de construção das matrizes $\boldsymbol{G}$ e $\boldsymbol{H}$.

i) Biplot SQRT, em que nenhuma ênfase é dada a linha ou coluna de $\boldsymbol{Y}$, $\alpha = \frac{1}{2}$

$$\boldsymbol{Y}_{n \times p} = \begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{21}\sqrt{\lambda_2} \\ \vdots & \vdots \\ u_{n1}\sqrt{\lambda_1} & u_{n2}\sqrt{\lambda_2} \end{pmatrix} \begin{pmatrix} v_{11}\sqrt{\lambda_1} & \ldots & v_{1p}\sqrt{\lambda_1} \\ v_{21}\sqrt{\lambda_2} & \ldots & v_{2p}\sqrt{\lambda_2} \end{pmatrix}$$

ii) Biplot JK, em que a ênfase é nas linhas $\boldsymbol{Y}$, $\alpha = 1$.

$$\boldsymbol{Y}_{n \times p} = \begin{pmatrix} u_{11}\lambda_1 & u_{21}\lambda_2 \\ \vdots & \vdots \\ u_{n1}\lambda_1 & u_{n2}\lambda_2 \end{pmatrix} \begin{pmatrix} v_{11} & \ldots & v_{1p} \\ v_{21} & \ldots & v_{2p} \end{pmatrix}$$

iii) Biplot GH, em que a ênfase é nas colunas de $\boldsymbol{Y}$; $\alpha = 0$.

$$\boldsymbol{Y}_{n\times p} = \begin{pmatrix} u_{11} & u_{21} \\ \vdots & \vdots \\ u_{n1} & u_{n2} \end{pmatrix} \begin{pmatrix} v_{11}\lambda_1 & \dots & v_{1p}\lambda_1 \\ v_{1n}\lambda_2 & \dots & v_{2p}\lambda_2 \end{pmatrix}$$

Considere a i-ésima linha de $\boldsymbol{G}$, $\boldsymbol{g}_j^{'}$ e a j-ésima coluna de $\boldsymbol{H}^{'}, \boldsymbol{h}_j$

$$\boldsymbol{Y}_{n\times p} = \begin{pmatrix} g_1^{'} \\ \vdots \\ g_n^{'} \end{pmatrix} \begin{pmatrix} \boldsymbol{h}_1 & \dots & \boldsymbol{h}_p \end{pmatrix}$$

O biplot consiste em plotar os vetores $\boldsymbol{g}_i^{'}$, $i = 1, \dots, n$ e $\boldsymbol{h}_j$, $j = 1, \dots, p$ o espaço tridimensional. Eles são vetores que partem da origem até o ponto de coordenadas dadas pelos elementos de $\boldsymbol{g}_i$ e $\boldsymbol{h}_j$.

Cada elemento y_{ij} de $\boldsymbol{Y}$ é representado como produto interno de $\boldsymbol{g}_i^{'}\boldsymbol{h}_j$. O comprimento da projeção de $\boldsymbol{g}_i^{'}$ sobre $\boldsymbol{h}_j$ é dado por:

$$L_{\frac{pg_i}{h_j}} = \frac{|g_i^{'}h_j|}{L_{h_j}} \Rightarrow |g_i^{'}h_j| \begin{cases} L_{h_j}L_{\frac{pg_i}{h_j}} \\ ou \\ L_{g_j}L_{\frac{pg_j}{h_j}} \end{cases} = L_{g_j}L_{h_j}cos(\theta) \tag{3.8}$$

O produto interno $\boldsymbol{g}_i^{'}\boldsymbol{h}_j$ pode ser visto como o produto do comprimento de um dos vetores vezes o comprimento da projeção do outro no primeiro (DIAS, 2008).

4 Aplicações

4.1 Análise descritivas e exploratória dos dados

Pode-se observar na Figura 1 a correlação entre as variáveis que compõem o Índice de Desenvolvimento Sustentável para Municípios do Estado de Mato Grosso. Foi utilizado a correlação de Spearmam devido a falta de normalidade entre algumas variáveis, fato este que será apresentado posteriormente com aplicação dos testes de normalidade univariada.

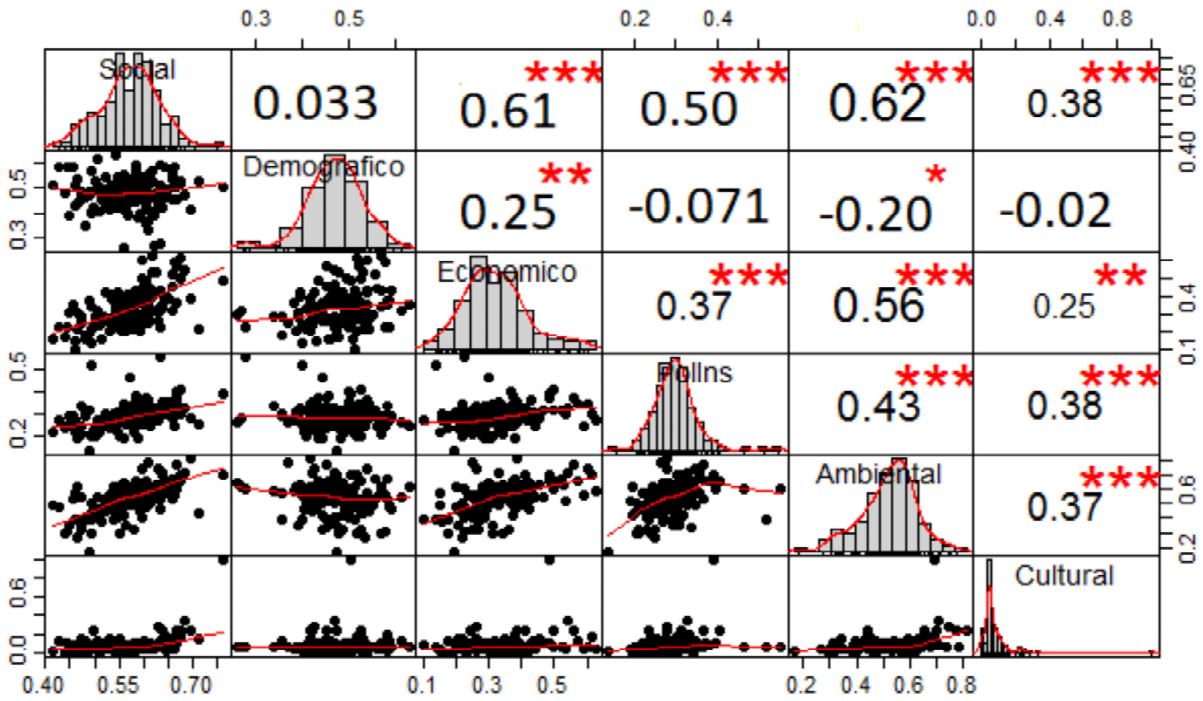


Figura 1 – Visualização de uma Matriz de Correlação. Em cima do valor (absoluto) da correlação mais o resultado do cor.test como estrelas (***) significativo a 1%, (**) significativo a 5% e (*) significativo a 10%. Na parte inferior, os diagramas de dispersão bivariados, com uma curva ajustada.

Uma vez que as variáveis apresentam correlações significativas e não significativas foi iniciado a análise univariada das seis variáveis em estudo (Social, Demográfico, Econômico, Politico-Institucional, Ambiental, Cultural). na Tabela 3 apresenta-se as principais estatística descritivas obtidas pela medida de tendência central (média, mediana), e dispersão (desvio padrão, assimetria e curtose) para as variáveis que compõem IDSM. Pode-se observar valores de curtose maiores que zero para as variáveis Politico-Institucional, Ambiental e Cultural. De fato, observando-se o histograma de ambas na Figura 2, confirma-se

tal assimetria positiva e negativa como também é apresentado na Tabela 3. É provável que esta assimetria contribua para a não normalidade dessas variáveis, este fato será confirmado após aplicação de testes de normalidade.

Tabela 3 – Estatística descritiva relacionado as variáveis que compõem o Índice desenvolvimento Sustentável para Municípios do Estado de Mato Grosso.

Variavel	**n**	**Média**	**Desvio**	**Mediana**	**Assimetria**	**Curtose**
Social	137	0,570	0,062	0,574	-0,107	-0,045
Demográfico	137	0,692	0,041	0,692	-0,053	-0,232
Econômico	137	0,571	0,090	0,574	0,069	0,103
polins	137	0,535	0,048	0,535	0,265	1,852
Ambiental	137	0,726	0,080	0,738	-0,844	1,280
Cultural	137	0,260	0,118	0,243	1,897	10,534

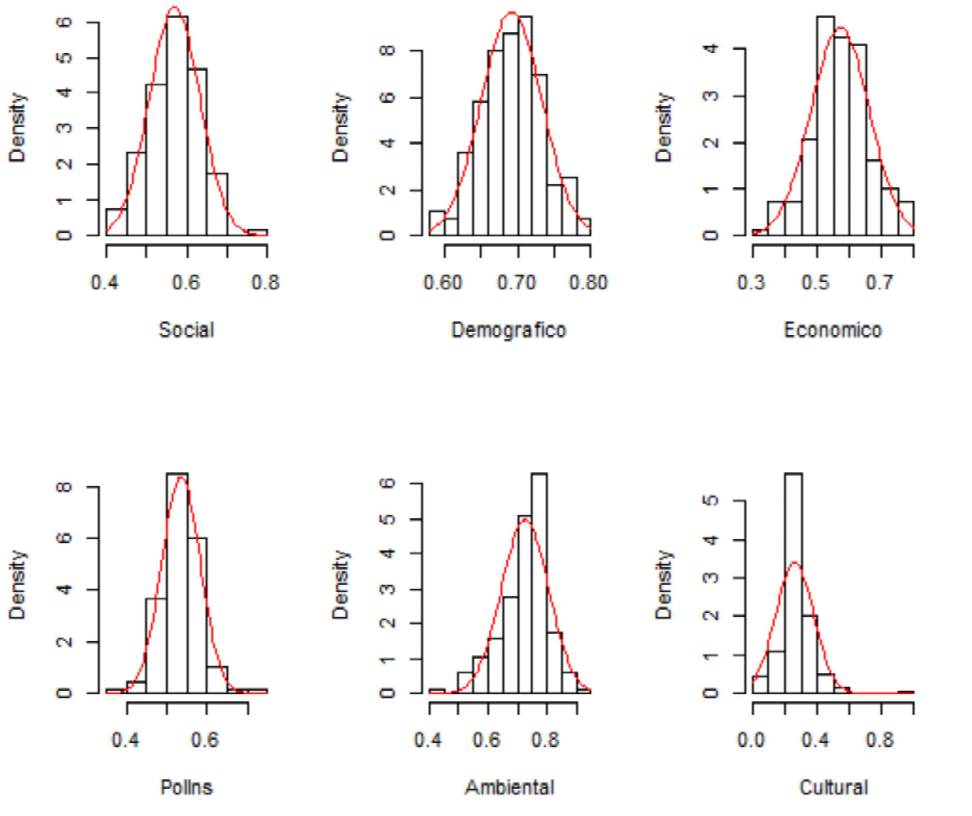


Figura 2 – Histograma das variáveis que compõem o Índice desenvolvimento Sustentável para Municípios do Estado de Mato Grosso.

Apresenta-se na Tabela 4 o Teste de normalidade univariado de Shapiro-Wilk com seus respectivos nível descritivos p. Pode-se observar por meio do teste de Shapiro-Wilk que 50% das variáveis segue um comportamento normal univariado. As variáveis que não seguem normalidade Polins, Ambiental e Cultural serão analisadas utilizando-se métodos

não paramétricos comparando-se suas respectivas medianas. É importante ressalta que mesmo as variáveis sugerindo o comportamento normal univariado, não garante que as mesmas seguem um comportamento normal multivariada.

Tabela 4 – Teste de normalidade univariado de Shapiro-Wilk para as variáveis do Índice Desenvolvimento Sustentável para Municípios do Estado de Mato Grosso.

Variável	**Estatística**	**Valor de p**	**Normalidade**
Social	0,9913	0,5565	Sim
Demográfico	0,9968	0,9920	Sim
Econômico	0,9919	0,6186	Sim
polins	0,9755	0,0142	Não
Ambiental	0,9543	0,0002	Não
Cultural	0,8400	0,0000	Não

Conforme foi observado na Tabela 4, nem todas as variáveis apresentam normalidade univariada pelo o teste de Shapiro-Wilk. Sendo assim, neste momento será verificado a normalidade multivariada, isto é, será aplicado testes para verificar a normalidade multivariada levando-se em consideração todas as variáveis conjuntamente. Retirando-se a variável cultural que não apresenta normalidade univariada e aplicando-se os testes de normalidade multivariados de Mardia, Henze-Zirkle e Royston, para as demais variáveis pode-se observar na Tabela 5 que há fortes evidências ($p > 0,05$) para não se rejeitar a hipótese de normalidade multivariada, sugerindo assim, um comportamento normal multivariado para as variáveis em estudo (KORKMAZ; GOKSULUK; ZARARSIZ, 2014).

Tabela 5 – Teste de Normalidade Multivariada para as variáveis: Social, Demográfico, Econômico.

Teste de normalidade	**Estatística(Valor de p)**	**Normalidade Multivariada**
Teste de Mardia	14,9843 (0,9866)	Sim
Teste de Henze-Zirkle	0,5043 (0,9742)	Sim
Teste de Royston	0,5205 (0,9115)	Sim

4.2 MANOVA

Veremos agora a estatística T^2 de Hotelling, assim chamada em homenagem a Harold Hotelling, um pioneiro em análises multivariadas, o qual foi o primeiro a obter sua distribuição amostral. Observamos que H_0 será rejeitada se um ou mais componentes médios ou algumas combinações das médias diferir muito dos valores fixados na hipótese nula.

O vetor das estimativas das médias é dado por:

$$\bar{\boldsymbol{X}} = \begin{bmatrix} 0,570 \\ 0,692 \\ 0,571 \end{bmatrix}$$

A matriz de variâncias e covariâncias estimada foi:

$$\boldsymbol{S} = \begin{bmatrix} 0,003878 & 0,000039 & 0,003768 \\ 0,000039 & 0,001691 & 0,001003 \\ 0,003768 & 0,001003 & 0,008076 \end{bmatrix}$$

E por fim a matriz de correlação foi:

$$\boldsymbol{R} = \begin{bmatrix} 1,00000 & 0,01522 & 0,67339 \\ 0,01522 & 1,00000 & 0,27140 \\ 0,67339 & 0,27140 & 1,00000 \end{bmatrix}$$

Comparando o valor observado $T^2 = 17{,}929$ com valor crítico $\frac{(n-1)p}{(n-p)}F_{[p;(n-p);\alpha]} = \frac{(137-1)3}{(137-3)}F_{[3;(137-3);0,05]} = 8{,}1362$ observamos que $T^2 = 17{,}929 > 8{,}1362$, consequentemente rejeitamos H_0 ao nível $\alpha = 0{,}05$ de significância. O nível descritivo p é maior que 0,0001 altamente significativo.

Dando sequência, será aplicado a Análise de Variância Multivariado (MANOVA) para verificar o efeito da influência que a produção de soja exerce aos Índice de Desenvolvimento Sustentável para Municípios no Estado do Mato Grosso. A MANOVA é utilizada para investigar a comparação entre as médias das diferentes variâncias simultaneamente. A MANOVA é uma extensão da Análise de Variância (ANOVA) e ambas utilizam os seguintes passos:

i) Testa-se a hipótese global de igualdade de médias entre o efeito da influência;

ii) Se o resultado dos passo anterior for significativo, utilizam-se testes adicionais no sentido de explicar as diferenças entre o efeito da influência (comparações múltiplas)

A MANOVA, no entanto, tem vantagens sobre a ANOVA, pois considera-se o nível de significância conjunto dos testes aproveitando-se das informações conjuntas das variáveis envolvidas.

Tabela 6 – Análise da Variância Multivariada (MANOVA) para verificar o efeito do vetor de médias das variáveis do Índice Desenvolvimento Sustentável para Municípios.

Critério	**Estatística**	**Aproximação do teste F**	**Valor de p**
Lambda Wilks	0,63004	5,6411	$< 0,0001$
Traço de Pillai	0,40978	5,3789	$< 0,0001$
Traço de Hotelling Lawley	0,52513	5,8056	$< 0,0001$
Maior Raiz Roy	0,36793	12,51	$< 0,0001$

De acordo com os testes apresentados na Tabela 6, pode-se concluir com um nível de significância de 5% que não houve diferenças significativas (p > 0,05) entre os vetores de médias e que existe evidências que a produção de soja exerce influência sobre o Índice de Desenvolvimento Sustentável para Municípios do Estado do Mato Grosso.

Assim veremos agora as diferenças de médias para fazer comparações múltiplas de Bonferroni. E vamos ilustrar os intervalos de confiança para as diferenças dos pares de médias ($\mu_k - \mu_l$) dos tratamentos para os fatores (Centro-sul, Nordeste, Norte, Sudeste, Sudoeste). Então o vetor de médias para os grupos é dados por:

Tabela 7 – Médias dos tratamentos e fatores para comparação Mútiplas de Bonferroni.

Mesorregiões	**Social**	**Demográfico**	**Econômico**
Centro-sul	0,5482	0,6604	0,5134
Nordeste	0,5598	0,6996	0,5302
Norte	0,5789	0,7090	0,5965
Sudeste	0,5949	0,6440	0,5832
Sudoeste	0,5602	0,6886	0,5748

i) Intervalo de confiança simultâneos de Bonferrini.

$$\bar{x_1} = 0,570 \pm 0,0558 \;\; ou \;\; 0,5142 \leq \mu_1 \leq 0,6258$$

$$\bar{x_2} = 0,692 \pm 0,0821 \;\; ou \;\; 0,6099 \leq \mu_2 \leq 0,7741$$

$$\bar{x_3} = 0,571 \pm 0,1078 \;\; ou \;\; 0,4632 \leq \mu_3 \leq 0,6788$$

ii) Intervalo de confiança simultâneos de T^2.

$$\bar{x_1} = 0,570 \pm 0,1731 \;\; ou \;\; (0,3969; 0,7431)$$

$$\bar{x_2} = 0,692 \pm 0,1041 \;\; ou \;\; (0,5879; 0,7961)$$

$$\bar{x_3} = 0,571 \pm 0,1366 \;\; ou \;\; (0,4344; 0,7076)$$

Apresentado os intervalos com 95% de confiança de Bonferroni, juntamente com os correspondentes intervalos simultâneos de T^2. Para cada componentes médio, o intervalo de confiança de Bonferroni fica dentro do intervalo T^2. Consequentemente, a região formada pelos três intervalos de Bonferroni está contida na região formada pelos três intervalos de

T^2. Se estamos interessados apenas em componentes médios, os intervalos de confiança de Bonferroni fornecem estimativas mais precisas do que os intervalos T^2. Por outro lado, a região com 95% de confiança para μ fornece valores realísticos para os pares (μ_1, μ_2, μ_3) quando a correlação entre as características medidas é levada em conta (JOHNSON; WICHERN, 2002).

4.3 Análise de Componentes Principais

A Análise de Componentes Principais (CP) discriminou as Variáveis segundo os indivíduos analisados nesta amostra. Foram considerados os três primeiros autovalores obtidos da matriz de covariância dos dados originais, que determinaram as combinações lineares das variáveis originais, isto é, os componentes principais. Na Tabela 8 é possível observar as correlações das variáveis e dos componentes principais.

Tabela 8 – Autovalores gerados a partir da matriz de correlação R das variáveis padronizadas.

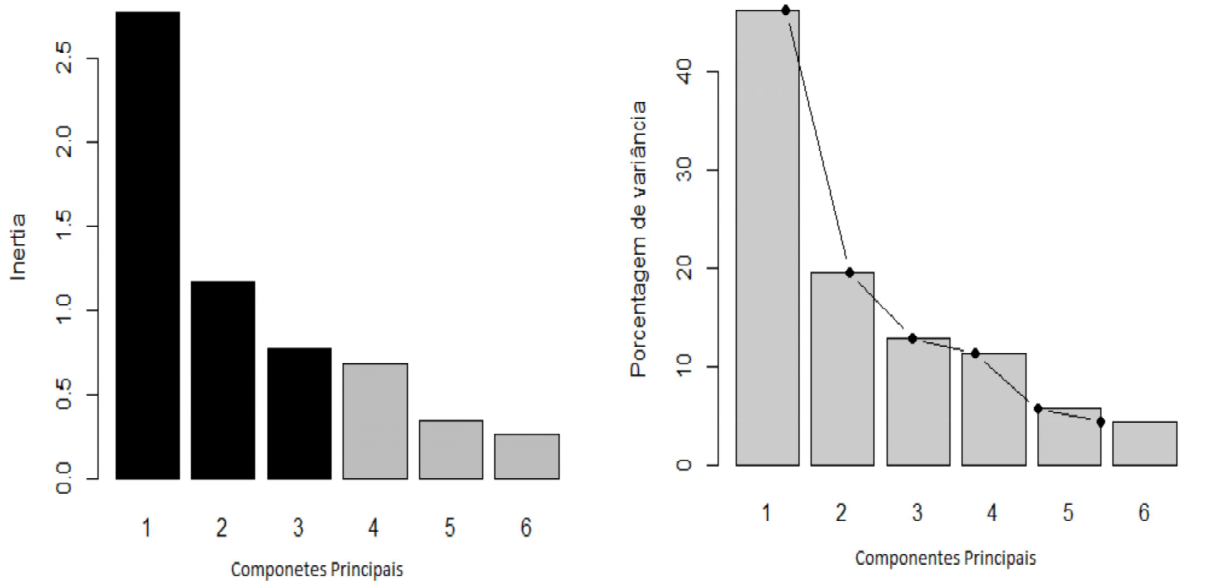
Componentes Principais (CP)	**Autovalores**	**Variância total (%)**	**Autovalores Acumulado**	**Variância Acumulada (%)**
1	2,772	46,21	2,772	46,21
2	1,174	19,57	3,946	65,78
3	0,774	12,89	4,720	78,67
4	0,676	11,26	5,396	89,93
5	0,341	5,68	5,737	95,61
6	0,263	4,39	6,000	100,00

Na análise de componentes principais a variância contida em cada componente principal gerado é expressa pelos autovalores da matriz padronizada, de tal forma que, o maior autovalor está associado ao primeiro componente principal, o segundo maior autovalor ao segundo, e assim por diante, até que o menor autovalor esteja associado ao último componente principal, colocando os primeiros como os mais importantes. Sendo assim, os primeiros componentes gerados pela análise de componentes principais explicam a maior parte da variância dos dados originais que explicaram mais de 70% da variância conforme sugerido por (KAISER, 1960), sendo assim, adotando-se o lambda maior que $1(\lambda > 1)$, tem-se que, os três primeiros componentes principais retêm 78,67% da variabilidade total, ou seja, das informações contidas nas variáveis, pode-se observar também que o primeiro componentes principal (CP1) 46,21% da variância total; o segundo (CP2), 19,57% ; e o terceiro (CP3), 12,89%.

Isto indica que as seis variáveis, Social Demográfica Econômica Politico- Institucional Ambiental e Cultural estudadas podem ser substituídas por estes três componentes, com perda miníma de informações. Diversos estudos utilizaram a análise de componentes principais (SANTI et al., 2012; HONGYU; SANDANIELO; JUNIOR, 2016).

Em estudos que utilizam a técnica dos componentes principais como meio de descartes de variáveis com a finalidade de redução de mão de obra, tempo e custo dependendo da análise e interpretação dos dados experimentais, a importância relativa das características pode ser avaliada pela magnitude do coeficiente de ponderação (CRUZ et al., 1994).

Bratchell (1989), utilizou a análise de componentes principais na modelagem de superfície de resposta multivariada e verificou que o método traz luz para solução. Também recomendou a utilização de métodos de rotação das componentes principais para melhorar a explicação dos modelos.



(a) Scree-plot dos 6 componentes principais

(b) Barplot porcentagem da Variância dos componentes

Figura 3 – Visualização gráfica da variabilidade dos componentes principais gerados a partir da matriz de correlação $\boldsymbol{R}$.

O Scree-plot Figura 3(a) indicaram que apenas os três primeiros componentes são suficientes para explicar a maior parte da variação total dos dados, 78,67% Tabela 8, ou seja, elas podem substituir as variáveis originais em análises subsequentes. A redução do número de variáveis já era esperada, pela forte correlação entre elas.

Tabela 9 – Autovetores da matriz de correlação dos componentes principais e das variáveis do Índice Desenvolvimento Sustentáveis para Municípios.

Variáveis	**Comp-1**	**Comp-2**	**Comp-3**
Social	$\mathbf{0,8694}$	-0,0426	-0,0702
Demográfico	-0,0203	$\mathbf{-0,9241}$	0,1247
Econômico	$\mathbf{0,7534}$	-0,4410	-0,2962
Polins	0,5904	0,3920	0,0986
Ambiental	$\mathbf{0,8128}$	0,1602	-0,2734
Cultural	0,5866	-0,0176	$\mathbf{0,7685}$

As variáveis Social, Ambiental e Econômico apresentaram contribuições similares para o componente principal 1 (Comp-1), isto foi verificado pelas variáveis que possam têm vetor de maior comprimento Tabela 9 e que foram mais próximas ao eixo Comp-1, mostrado na Figura 4.

As variáveis Demográfico e Cultural apresentam a maior contribuição para os componentes principais 2 e 3 (Comp-2 e Comp-3) respectivamente como apresentado na Figura 4.

ACP foi usada para reduzir as dimensões das variáveis originais sem perda de informações, por definição, a correlação entre os principais componentes é zero, isto é, a variação explicada no componente principal (Comp-1) é independente da variação explicada no componente principal 2 (Comp-2) e assim por diante. Isto implica que para qualquer componente principal não vai causar uma resposta correlacionada em termos de outros componentes principais, isto é, eles são ortogonais (SAVEGNAGO et al., 2011) (FRAGA et al., 2015).

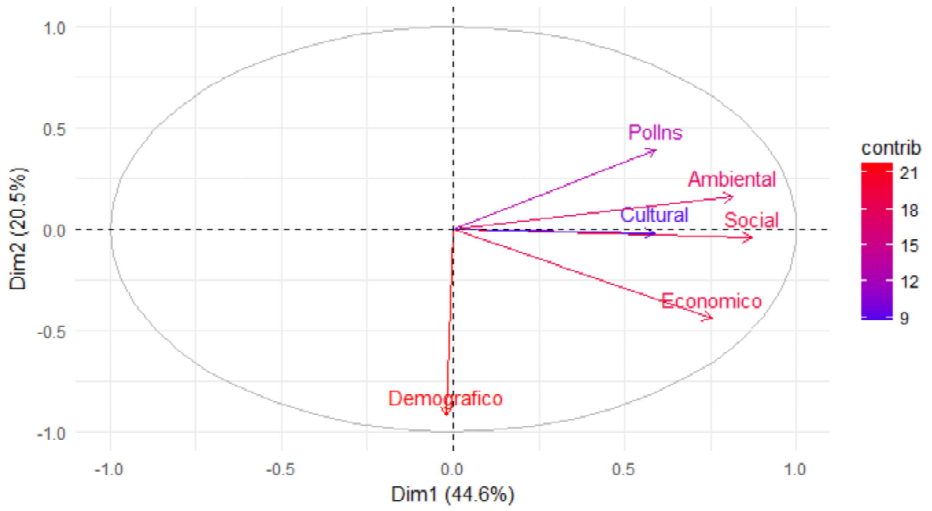


Figura 4 – Representação gráfica dos componentes principais e da contribuição relativa a cada coluna.

O Biplot é um método para representar de forma bidimensional dados multivariados. As variáveis de sustentabilidade na Figura 5 são identificadas pelas cores verde, rosa, azul, laranja e verde claro, mesorregiões do estado de Mato Grosso que são Centro sul, Nordeste, Norte, Sudeste e Sudoeste respectivamente, ao passo que os autovalores permitem representar as variáveis. Cada observação é representada pelo par de escores dos dois primeiros componentes principais. Os ângulos entre os vetores estão relacionados às correlações entre as variáveis, sendo que quando menor o ângulo, mais correlacionadas estão. As posições dos pontos (Cidades que pertence as mesorregiões) no gráfico indicam semelhanças e diferenças entre elas.

Na construção das elipses apresentadas no Biplot Figura 5 representando as mesorregiões do (Estado de Mato Grosso), é possível observar que as mesorregiões se encontram sobrepostas, corroborando assim, para o que foi apresentado nos testes da MANOVA, isto é, existe uma similaridade considerável entre os vetores de médias desses cincos grupos, indicando assim que não será possível classificá-las em populações distintas.

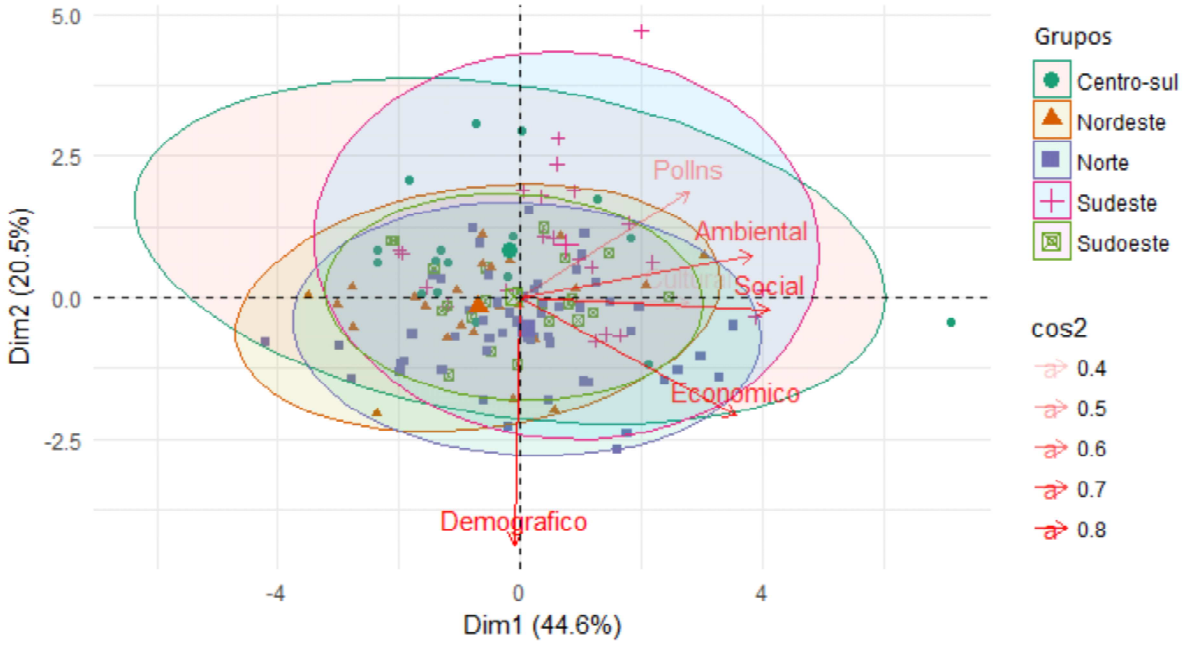


Figura 5 – Biplot de indivíduos e variáveis com suas respectivas elipses de confiança.

5 Conclusão

A utilização da MANOVA em substituição a ANOVA, perante a evidência de normalidade multivariada nas variáveis: (Social, Demográfica, Econômica, Política- Institucional, Ambiental e Cultural), para Índice Desenvolvimento Sustentável para Municípios do estado do Mato Grosso, apresenta-se como vantagem importante, visto que os testes de Lambda de Wilks, Traço de Pillai, Traço de Lawley Hotelling e Maior raiz de Roys foram satisfatório para concluir com um nível de significância de 5% que não houve diferenças significativas ($p > 0{,}05$) entre os vetores de médias e que existe evidências que a produção de soja exerce influência ao Índice de Desenvolvimento Sustentável para Municípios do Estado do Mato Grosso.

Utilizamos a ACP para redução do números das variáveis Social, Demográfico Econômica Politico-Institucional Ambiental e Cultural do Índice de Desenvolvimento para Municípios, sendo assim, os primeiros componentes gerados pela análise de componentes principais explicam a maior parte da variância dos dados originais retendo 78,67% da variabilidade total , ou seja, das informações contidas nas variáveis. Isso indica que as seis variáveis estudadas podem ser substituídas pelos três primeiros componentes, com perda miníma de informações.

Os gráficos biplot é um método de análise multivariada que por si só revela resultados satisfatórios para qualquer conjunto de dados multivariados. Este aliado a outro método de análise, neste caso a análise de componentes principais, tem um papel importantíssimo como complemento para a interpretação dos resultados gráficos.

Referências

ANDREWS, D.; GNANADESIKAN, R.; WARNER, J. Transformations of multivariate data. *Biometrics*, JSTOR, p. 825–840, 1971. Citado na página 13.

BRATCHELL, N. Multivariate response surface modelling by principal components analysis. *Journal of Chemometrics*, Wiley Online Library, v. 3, n. 4, p. 579–588, 1989. Citado na página 38.

BRAY, J. H.; MAXWELL, S. E. *Multivariate analysis of variance*. [S.l.]: Sage, 1985. Citado na página 20.

CÁRDENAS, O.; GALINDO, M.; VICENTE-VILLARDÓN, J. L. Los métodos biplot: evolución y aplicaciones. *Revista Venezolana de Análisis de Coyuntura*, Universidad Central de Venezuela, v. 13, n. 1, p. 279–303, 2007. Citado na página 25.

CECATTO, C.; BELFIORE, P. O uso de métodos de previsão de demanda nas indústrias alimentícias brasileiras. *Gest. Prod., São Carlos*, v. 22, n. 2, p. 404–418, 2015. Citado na página 20.

CRUZ, C. *Aplicação de algumas técnicas multivariadas no melhoramento de plantas*. [S.l.]: Esalq, 1990. Citado na página 23.

CRUZ, C. et al. Diversidade genética. *CRUZ, CD; REGAZZI, AJ Modelos biométricos aplicados ao melhoramento genético. Viçosa: UFV*, p. 287–313, 1994. Citado 3 vezes nas páginas 23, 28 e 38.

DEMÉTRIO, C. *Análise multidimensional para dados de cana-de-açúcar*. [S.l.]: ESALQ, 1985. Citado na página 17.

DIAS, C. Intranet do departamento de ciências exatas. *Arquivos de aulas. LCEESALQ-USP. Disponıvel em:< http://www. lce. esalq. usp. br/tadeu/intranet/aula12. pdf>. Acesso em*, v. 29, 2008. Citado na página 31.

FRAGA, A. B. et al. Multivariate analysis to evaluate genetic groups and production traits of crossbred holstein× zebu cows. *Tropical animal health and production*, Springer, v. 48, n. 3, p. 533–538, 2015. Citado na página 39.

GABRIEL, K. R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, JSTOR, p. 453–467, 1971. Citado 2 vezes nas páginas 25 e 30.

GIRI, N. C. *Mu1 tivaria t e Statistical Analysis*. [S.l.]: Marcel Dekker, Inc., New York, 2004. Citado na página 26.

GOMES, F. P. *A estatística moderna na pesquisa agropecuaria*. [S.l.], 1984. Citado na página 27.

GOUVÊA, M. A.; PREARO, L. C.; ROMEIRO, M. do C. Avaliação da adequação de aplicação de técnicas multivariadas em estudos do comportamento do consumidor em teses e dissertações de duas instituições de ensino superior. *Revista de Administração*, Elsevier, v. 47, n. 2, p. 338–355, 2012. Citado na página 12.

GOWER, J.; HAND, D. *Biplots Chapman & Hall*. [S.l.]: London, 1996. Citado na página 25.

HAIR, J. F. et al. *Análise multivariada de dados*. [S.l.]: Bookman Editora, 2006. Citado na página 17.

HAIR, J. F. et al. *Análise multivariada de dados*. [S.l.]: Bookman Editora, 2009. Citado 2 vezes nas páginas 12 e 26.

HENZE ZIRKLER, B. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 19, n. 10, p. 3595–3617, 1990. Citado na página 14.

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. de O. Análise de componentes principais: Resumo teórico, aplicação e interpretação. *E&S Engineering and Science*, v. 5, n. 1, p. 83–90, 2016. Citado na página 37.

HOTELLING. A generalized t test and measure of multivariate dispersion. In: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA. *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. [S.l.], 1951. Citado na página 21.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, Warwick & York, v. 24, n. 6, p. 417, 1933. Citado na página 23.

HOTELLING, H.; FRANKEL, L. R. The transformation of statistics to simplify their distribution. *The Annals of Mathematical Statistics*, JSTOR, v. 9, n. 2, p. 87–96, 1931. Citado na página 16.

JOHNSON, R. A.; WICHERN. *Applied multivariate statistical analysis*. [S.l.]: Prentice hall Upper Saddle River, NJ, 2002. v. 5. Citado 9 vezes nas páginas 11, 12, 16, 17, 20, 26, 27, 29 e 37.

JÚNIOR, H. et al. Análise multivariada de dados. *Análise multivariada de dados*, Bookman Porto Alegre, 2005. Citado na página 17.

KAISER, H. F. The application of electronic computers to factor analysis. *Educational and psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 141–151, 1960. Citado na página 37.

KORKMAZ, S.; GOKSULUK, D.; ZARARSIZ, G. Mvn: an r package for assessing multivariate normality. *The R Journal*, Citeseer, v. 6, n. 2, p. 151–162, 2014. Citado na página 34.

LAWLEY, D. A generalization of fisher's z test. *Biometrika*, JSTOR, v. 30, n. 1/2, p. 180–187, 1938. Citado na página 21.

LIBERATO, J. R. Aplicações de técnicas de análise multivariada em fitopatologia. Universidade Federal de Viçosa, 1995. Citado na página 23.

LOURENÇO, A.; MATIAS, R. P. Estatística multivariada. *Instituto Superior de*, 2000. Citado na página 12.

MACEDO, L. O. B. et al. Influências da produção de soja sobre a sustentabilidade dos municípios do estado de mato grosso–mt. *Revista ESPACIOS/ Vol. 37 (Nº 07) Año 2016*, 2016. Citado na página 10.

MAGNUSSON, W. E. *Estatística [sem] matemática: a ligação entre as questões ea análise*. [S.l.]: Planta, 2003. Citado na página 10.

MANLY, B. J. 1986. multivariate statistical methods, a primer. *Chapman and Hall. London. 160p*, 1986. Citado na página 23.

MANLY, B. *Métodos estatísticos multivariados. Tradução de Sara Ianda Correa Carmona*. [S.l.]: Porto Alegre: Bookman, 2008. Citado na página 24.

MARDIA, K.; KENT, J.; BIBBY, J. Multivariate analysis academic press inc. *London) LTD*, 1979. Citado na página 23.

MARDIA, K. V. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, JSTOR, p. 519–530, 1970. Citado na página 13.

MONTGOMERY, D. C. *Statistical quality control*. [S.l.]: Wiley New York, 2009. v. 7. Citado 2 vezes nas páginas 11 e 23.

MORRISON, D. The structure of multivariate observations: I. principal components. In: *Multivariate statistical methods*. [S.l.]: McGraw-Hill Book Company, New York, 1976. p. 266–301. Citado na página 23.

PONTES, A. *Análise de variância multivariada com a utilização de testes não-paramétricos e componentes principais baseados em matrizes de posto. 2005. 106 p.* Tese (Doutorado) — Tese (Doutorado em estatística e experimentação agronômica). Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2005. Citado na página 11.

PRADO, T. K. L. d. Regressão não linear multivariada no crescimento do coco variedade anã verde. Universidade Federal de Lavras, 2016. Citado na página 16.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. Disponível em: <https://www.R-project.org/>. Citado na página 26.

RAO, C. R. Advanced statistical methods in biometric research. Wiley, 1952. Citado na página 16.

REIS, E. Discriminating analysis (analise discriminante). *Estatistica Multivariada Aplicada. Edições Silabo, Lisbon, Portugal*, p. 201–244, 1997. Citado na página 20.

RENCHER, A. C. *Methods of multivariate analysis*. [S.l.]: John Wiley & Sons, 2003. v. 492. Citado na página 10.

ROYSTON, P. Approximating the shapiro-wilk w-test for non-normality. *Statistics and Computing*, Springer, v. 2, n. 3, p. 117–119, 1992. Citado na página 15.

SACHS, I. *Rumo à ecossocioeconomia: teoria e prática do desenvolvimento*. [S.l.]: Cortez, 2007. Citado na página 10.

SANTI, A. L. et al. Análise de componentes principais de atributos químicos e físicos do solo limitantes à produtividade de grãos. *Pesquisa Agropecuária Brasileira*, v. 47, n. 9, p. 1346–1357, 2012. Citado na página 37.

SARTORIO, S. D. *Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R.* Tese (Doutorado) — Escola Superior de Agricultura "Luiz de Queiroz, 2008. Citado na página 19.

SAVEGNAGO, R. et al. Estimates of genetic parameters, and cluster and principal components analyses of breeding values related to egg production traits in a white leghorn population. *Poultry science*, Oxford University Press, v. 90, n. 10, p. 2174–2188, 2011. Citado na página 39.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, Biometrika Trust, v. 52, n. 3-4, p. 591–611, 1965. Citado na página 16.

SHAPIRO, S. S.; WILK, M. B.; CHEN, H. J. A comparative study of various tests for normality. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 63, n. 324, p. 1343–1372, 1968. Citado na página 16.

SHARMA, S. *Applied multivariate techniques.* [S.l.]: John Wiley & Sons, Inc., 1995. Citado na página 14.

SONG, F. N. Técnicas de análise multivariada com aplicações a dados de natureza biológica. Universidade Estadual Paulista (UNESP), 2013. Citado na página 20.

STAUFFER, D. F.; GARTON, E. O.; STEINHORST, R. K. A comparison of principal components from real and random data. *Ecology*, Wiley Online Library, v. 66, n. 6, p. 1693–1698, 1985. Citado na página 27.

STEINER, M. T. A. *Uma metodologia para o reconhecimento de padrões multivariados com resposta dicotômica.* Tese (Doutorado) — Universidade Federal de Santa Catarina, Centro Tecnologico., 1995. Citado na página 12.

TABACHNICK, B. G.; FIDELL, L. S.; OSTERLIND, S. J. Using multivariate statistics. Allyn and Bacon Boston, 2001. Citado na página 24.

TOPA, M. A. Análise multivariada como ferramenta de gerenciamento de fornecedores visando um relacionamento com vantagem competitiva. *Universidade Federal do Paraná*, 2009. Citado na página 10.

VICINI, L.; SOUZA, A. M. Análise multivariada da teoria à prática. *Santa Maria: UFSM, CCNE*, 2005. Citado na página 11.

WILKS, S. S. Certain generalizations in the analysis of variance. *Biometrika*, Biometrika Trust, v. 24, n. 3-4, p. 471–494, 1932. Citado na página 17.

WISHART, J. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, JSTOR, p. 32–52, 1928. Citado na página 16.