ISSN 1679-0154
Dezembro/2023

BOLETIM DE PESQUISA E DESENVOLVIMENTO

258

Análise Não Paramétrica Aplicada em Componente Principal como uma Alternativa à Análise de Variância Multivariada

Estudo da Qualidade Biológica do Solo

Embrapa

ISSN 1679-0154
Dezembro/2023

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Milho e Sorgo
Ministério da Agricultura e Pecuária

BOLETIM DE PESQUISA E DESENVOLVIMENTO 258

Análise Não Paramétrica Aplicada em Componente Principal como uma Alternativa à Análise de Variância Multivariada

Estudo da Qualidade Biológica do Solo

Antônio Carlos de Oliveira
Ivanildo Evódio Marriel
Álvaro Vilela Resende
Enilda Alves Coelho

Embrapa Milho e Sorgo
Sete Lagoas, MG
2023

Embrapa Milho e Sorgo
Rod. MG 424, Km 45
Caixa Postal 151
CEP 35701-970 Sete Lagoas, MG
Fone: (31) 3027-1100
www.embrapa.br/fale-conosco/sac

Revisão de texto
Antonio Claudio da Silva Barros

Normalização bibliográfica
Rosângela Lacerda de Castro (CRB-6/2749)

Tratamento das ilustrações
Márcio Augusto Pereira do Nascimento

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Márcio Augusto Pereira do Nascimento

Foto da capa
Enilda Alves Coelho

1ª edição
Publicação digital (2023): PDF

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Milho e Sorgo

Análise não paramétrica aplicada em componente principal como uma alternativa à análise de variância multivariada: estudo da qualidade biológica do solo / Antônio Carlos de Oliveira... [et al.]. – Sete Lagoas : Embrapa Milho e Sorgo, 2023.

PDF (29 p.) : il. color. – (Boletim de Pesquisa e Desenvolvimento / Embrapa Milho e Sorgo, ISSN 1679-0154; 258).

1. Análise de variância. 2. Análise multivariada. 3. Método estatístico. 4. Biologia do solo. I. Oliveira, Antônio Carlos de. II. Marriel, Ivanildo Evódio. III. Resende, Álvaro Vilela de. IV. Coelho, Enilda Alves. V. Série.

CDD (21.ed.) 519.5

Rosângela Lacerda de Castro (CRB-6/2749)

Sumário

Análise Não Paramétrica Aplicada em Componente Principal como uma Alternativa à Análise de Variância Multivariada

Estudo da Qualidade Biológica do Solo

Antônio Carlos de Oliveira[1]

Ivanildo Evódio Marriel[2]

Álvaro Vilela Resende[3]

Enilda Alves Coelho[4]

Resumo - A aplicação da análise de variância multivariada (Manova) requer algumas pressuposições que muitas vezes não se verificam, como a distribuição normal multivariada e a igualdade das matrizes de covariância. Frequentemente, esses problemas se mantêm, mesmo após o uso de métodos de transformação das variáveis. Para superar esses problemas, propõe-se neste trabalho a aplicação do teste não paramétrico de Kruskal-Wallis sobre os escores da primeira componente principal (PC_1), obtida da análise de componentes principais (PCA), como uma alternativa à Manova, em dados sobre indicadores de qualidade biológica do solo. Essa abordagem, além de agregar o efeito das variáveis em um único "índice", de simples interpretação, não requer as pressuposições de distribuição normal nem homogeneidade de variância. Como uma alternativa para a análise de experimentos, cujos dados não atendem a esses requisitos, foram utilizados dados experimentais, sobre atividades de enzimas envolvidas na ciclagem de nitrogênio e de fósforo, em amostras de solo de Cerrado, no sentido restrito, sob diferentes manejos. Os

[1] Engenheiro-agrônomo, doutor em Agronomia (Estatística e Experimentação Agronômica), pesquisador da Embrapa Milho e Sorgo, Sete Lagoas, MG;
[2] Engenheiro-agrônomo, doutor em Agronomia (Solos e Nutrição de Plantas), pesquisador da Embrapa Milho e Sorgo, Sete Lagoas, MG;
[3] Engenheiro-agrônomo, doutor em Ciência do Solo, pesquisador da Embrapa Milho Sorgo, Sete Lagoas. MG;
[4] Cientista da computação, mestre em Ciência da computação, Analista da Embrapa Milho Sorgo, Sete Lagoas, MG.

resultados mostraram que a primeira componente principal (PC_1) resultou em um índice global adequado para discriminar o efeito dos manejos sobre a qualidade biológica do solo.

Termos para indexação: análise de variância multivariada, componentes principais, testes não paramétricos, qualidade biológica do solo.

Nonparametric Analysis Applied in Principal Component as an Alternative to Multivariate Analysis of Variance

Abstract - The application of multivariate analysis of variance (Manova) requires some assumptions that are often not verified, such as multivariate normal distribution and equality of covariance matrices. These problems often persist, even after using variable transformation methods. In order to overcome these problems, this work proposes the application of the non-parametric Kruskal-Wallis test on the scores of the first principal component (PC_1) , obtained from principal component analysis (PCA), as an alternative to Manova, in data on soil biological quality indicators. This approach, in addition to aggregating the effect of variables into a single “index”, which is simple to interpret, does not require the assumptions of normal distribution or homogeneity of variance. As an alternative to the analysis of experiments, whose data do not meet these requirements, experimental data were used on the activities of enzymes involved in nitrogen and phosphorus cycling, in Cerrado soil samples, in the strict sense, under different management. The results showed that the first principal component (PC_1) resulted in a global index suitable for discriminating the effect of management on the biological quality of the soil.

Index terms: multivariate analysis of variance, principal components, non-parametric tests soil health.

Introdução

A análise de variância univariada (Anova) tem sido frequentemente utilizada para testar a hipótese de igualdade entre médias de tratamentos (H_0) em experimentos de diferentes áreas do conhecimento. Os testes, em geral, são realizados de forma univariada, ou seja, a Anova avalia apenas uma variável dependente de cada vez. Essa limitação pode ser um grande problema em determinadas situações, pois pode impedir a detecção de efeitos que realmente existem. Como as variáveis de resposta são medidas em uma mesma unidade experimental, é possível a ocorrência de correlação entre elas, sendo então mais adequada a realização de um teste multivariado, visto que, nesse caso, o teste da hipótese considera, simultaneamente, todas as variáveis dependentes. Por levar em consideração a correlação entre as variáveis, os testes multivariados tendem a ser mais poderosos que os testes univariados, sendo mais elevada a probabilidade de rejeitar a hipótese nula (H_0) quando, de fato, ela é falsa (Johnson; Wichern, 2007; Reis, 2001). Neste contexto a análise de variância multivariada (Manova) substitui com vantagens a análise de variância univariada. No entanto, a aplicação adequada da Manova requer algumas pressuposições que muitas vezes não se verificam, como a distribuição normal multivariada e a igualdade das matrizes de covariância. Frequentemente, esses problemas se mantêm, mesmo após o uso de métodos de transformação das variáveis.

Visando resolver o problema da não distribuição normal multivariada, Abapihi et al. (2021) propuseram, ao invés da Manova, o uso da Anova aplicada sobre os escores da primeira componente principal (PC_1), obtida por meio da técnica multivariada "Análise de Componentes Principais" (PCA). Conforme salienta Mingoti (2007), esse procedimento, ou seja, a utilização dos escores das componentes principais, é prática comum para condução de análise estatística de dados. A PC_1 é utilizado por ser a componente com a maior contribuição para a variação total das variáveis originais, podendo ser considerada um "indicador agregado" das variáveis originais. Abapihi et al. (2021) argumentam que, sendo essas variáveis normalmente distribuídas, as componentes também o serão, o que, segundo eles, viabiliza a aplicação da Anova sobre os escores da PC_1. No entanto, o procedimento não garante a homogeneidade da variância dos resíduos, que constitui requisito de fundamental importância para a aplicação da Anova.

Para superar esses problemas, propõe-se neste trabalho o uso da Análise de Componentes Principais, conforme Abapihi et al. (2021), mas substituindo a Anova pela aplicação do teste não paramétrico de Kruskal-Wallis (1952) sobre os escores da primeira componente principal, para testar a hipótese de igualdade dos tratamentos. O teste de Wicoxon é realizado para as comparações múltiplas

Essa abordagem não paramétrica, apesar de não explicar toda a informação contida nos dados, não requer as pressuposições de distribuição normal nem homogeneidade de variância. Portanto, constitui uma alternativa para a análise de experimentos cujos dados não atendem a esses requisitos, como é o caso dos ensaios de avaliação da qualidade biológica do solo, avaliada por meio da atividade de enzimas envolvidas na ciclagem de nutrientes, em agroecossistemas, caso tratado neste trabalho.

Material e Métodos

O trabalho propõe a utilização do teste não paramétrico de Kruskal-Wallis (1952) ao invés da Anova, usada por Abapihi et al. (2021), conforme já mencionado. A análise de PCA é utilizada para a obtenção dos escores da PC_1, e os testes não paramétricos são usados para os testes de hipóteses sobre os tratamentos a serem avaliados. A seguir, um resumo dessas ferramentas.

Análise de componentes principais (PCA)

A análise de componentes principais (PCA) foi concebida por Karl Pearson em 1901 e consolidada por Hotellig em 1933, conforme citação de Duarte (1998). A PCA explica a estrutura de variância-covariância de um conjunto de variáveis por meio de algumas combinações lineares dessas variáveis. Seus objetivos gerais são a interpretação dessa estrutura e/ou a redução das variáveis a serem avaliadas. Embora *p* componentes sejam necessárias para reproduzir a variabilidade total do sistema, muitas vezes grande parte dessa variabilidade pode ser explicada por um pequeno número *k* de componentes principais. Nesses casos, as *k* componentes principais podem então substituir as *p* variáveis iniciais do conjunto original (Johnson; Wichern, 2007).

Algebricamente, as componentes principais são combinações lineares das *p* variáveis aleatórias $X_1, X_2, \ldots, X_p$ e dependem somente da matriz de covariância **Σ** (ou da matriz de correlação **R**) dessas variáveis, não requerendo a suposição de distribuição normal multivariada. Em geral, deseja-se reduzir o número de variáveis a ser avaliadas.

Segue abaixo o desenvolvimento algébrico, para a obtenção das componentes principais populacionais, apresentado por Johnson e Wichern (2007), no caso de se utilizar a matriz de covariâncias.

Suponha-se que o vetor aleatório **X**' = [X_1, X_2, . . ., X_p.] tenha matriz de covariâncias **Σ** com raízes características (autovalores) $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0$.

Considerem-se as combinações lineares:

$$Y_1 = \boldsymbol{a}_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$
$$Y_2 = \boldsymbol{a}_2'\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$
$$\vdots$$
$$Y_p = \boldsymbol{a}_p'\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{2pp}X_p$$

Logo,

$$Var(Y_i) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_i \qquad i = 1,2, \ldots, p$$
$$Cov\,(Y_i, Y_k) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_k' \quad i,k = 1,2 \ldots, p$$

As componentes principais são as combinações lineares $Y_1, Y_2, \ldots, Y_p$, que sejam não correlacionadas e apresentem variâncias máximas. Os autores apresentam as seguintes definições:

Primeira componente principal = combinação linear $\boldsymbol{a}_1'\boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_1'\boldsymbol{X})$ sujeita à restrição $\boldsymbol{a}_1'\boldsymbol{a}_1 = \boldsymbol{1}$.

Segunda componente principal = combinação linear $\boldsymbol{a}_2'\boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_2'\boldsymbol{X})$ sujeita à restrição $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$ e $Cov(\boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{a}_2'\boldsymbol{X}) = 0$.

i-ésima componente principal = combinação linear $\boldsymbol{a}_i'\boldsymbol{X}$ que maximiza $Var(\boldsymbol{a}_i'\boldsymbol{X})$ sujeita à restrição $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$ e $Cov(\boldsymbol{a}_i'\boldsymbol{X}, \boldsymbol{a}_k'\boldsymbol{X} = 0$ para *k* < *i*.

As restrições impostas nas definições anteriores são satisfeitas quando o vetor $\boldsymbol{a}_i = \boldsymbol{e}_i$, ou seja, o autovetor normalizado correspondente a λ_i,

conforme relata Mingoti (2007). Logo, a i-ésima componente principal fica definida como:

$$Y_i = PC_i = \boldsymbol{e'}_i \boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p \,,$$

satisfeitas as condições:

$\boldsymbol{e'}_i \boldsymbol{e_k}$=0 , $i \neq k$; $\boldsymbol{e'}_i \boldsymbol{e_i}$=1 , $i = 1,2,\cdots,p$; $\Sigma \boldsymbol{e_i} = \boldsymbol{\lambda_i e_i}$ $i = 1,2,\cdots p$

Johnson e Wichern (2007) apresentam os seguintes resultados e considerações:

a) As componentes principais têm variâncias iguais aos autovalores de **Σ,** e são não correlacionadas, ou seja:

$$Var(Y_i) = \lambda_i \quad i = 1,2,\cdots p$$

$$Cov\,(Y_i, Y_k) = 0 \quad i \neq k$$

b) A proporção da variância total da população, devida à *k*-ésima componente principal, é:

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

Se a maior parte da variação total (acima de 80%, por exemplo) puder ser atribuída ao primeiro, ou a dois ou três componentes, então esses componentes podem substituir as *p* variáveis originais sem muita perda de informação.

Os coeficientes de correlação entre as componentes PC_i e as variáveis X_k são obtidos pela expressão:

$$\rho_{Y_i X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i,k = 1,2,\cdots p$$; sendo σ_{kk} a variância da k-ésima variável.

A magnitude de e_{ik} mede a contribuição da k-ésima variável para o i-ésimo componente principal, independentemente das demais variáveis, sendo proporcional ao coeficiente de correlação entre Y_i *e* X_k . Tanto os

coeficientes e_{ik} como as correlações devem ser examinados para ajudar a interpretar as componentes principais.

Quando as variáveis originais são medidas em escalas diferentes, há a necessidade de se proceder a uma padronização antes da obtenção das componentes principais, a partir da matriz de covariâncias. Neste trabalho, as variáveis são padronizadas pelas suas médias e seus desvios padrões. As novas variáveis são representadas por Z_i, em que $Z_i = (X_i - \mu_i)/\sigma_i$, sendo μ_i e σ_i a média e o desvio padrão de X_i, respectivamente.

Após a redução de *p* para as *k* componentes principais consideradas de maior relevância, seus valores numéricos são determinados para cada observação, de modo a substituir os valores das variáveis originais na execução da análise subsequente. Esses novos valores são chamados de escores das componentes. A Tabela 1, extraída de Varella (2008), mostra a organização de um conjunto de dados composto por *m* tratamentos, *p* variáveis e *k* componentes principais.

Tabela1. Variáveis originais padronizadas e escores correspondentes para os *m* tratamentos e as *p* variáveis.

Tratamentos	Variáveis padronizadas				Escores das componentes			
	Z_1	Z_2	⋯	Z_p	PC_1	PC_2	⋯	PC_K
1	Z_{11}	Z_{12}	⋯	Z_{1p}	PC_{11}	PC_{12}	⋯	PC_{1k}
2	Z_{21}	Z_{22}	⋯	Z_{2p}	PC_{21}	PC_{22}	⋯	PC_{2k}
⋮	⋮	⋮		⋮	⋮	⋮		⋮
m	Z_{m1}	Z_{m2}	⋯	Z_{mp}	PC_{m1}	PC_{m2}	⋯	PC_{mk}

Assim, os escores da primeira componente principal, para os *m* tratamentos, são:

Tratamentos	Primeira componente principal
1	$PC_{11} = e_{11}Z_{11} + e_{12}Z_{12} + \cdots + e_{1p}Z_{1p}$
2	$PC_{21} = e_{11}Z_{21} + e_{12}Z_{22} + \cdots + e_{1p}Z_{2p}$
⋮	⋮
m	$PC_{m1} = e_{11}Z_{m1} + e_{12}Z_{m2} + \cdots + e_{1p}Z_{mp}$

Johnson e Wichart (2007) afirmam que “as análises de componentes principais são mais um meio para se chegar a um fim do que um fim em si mesmas, servindo, frequentemente, como etapas intermediárias em pesquisas mais abrangentes”. Neste trabalho, essa estratégia é utilizada. A primeira componente principal é usada como um índice, para substituir as variáveis originais. Para o teste da hipótese de igualdade entre os tratamentos, utiliza-se o teste não paramétrico de Kruskal-Wallis (1952) sobre os escores obtidos com base nessa componente.

Teste de Kruskal-Wallis

O teste de Kruskal-Wallis é uma ferramenta usada na análise de dados quando se deseja comparar a distribuição de vários grupos (tratamentos) independentes. Ele é utilizado quando os dados não atendem aos pressupostos da análise de variância (Anova) por causa da violação da suposição de normalidade ou homogeneidade de variâncias. Ao invés de testar a hipótese de igualdade de médias entre os tratamentos, o teste de Kruskal-Wallis testa a hipótese de que os tratamentos provêm da mesma distribuição, ou não. A estatística *H*, que é calculada, em geral, com base nos postos (ranks) dos valores originais, é o critério de referência para o teste (Campos, 1983). Neste trabalho, a estatística *H* é calculada com base nos postos dos escores da primeira componente principal.

$$H = \frac{12}{N(N+1)} \sum_{i-1}^{v} \frac{R_i^2}{n_i} - 3(N+1),$$

em que:

n_i é o número de observações do i-ésimo tratamento, *N* é o total de observações.

$N = \sum_{i-1}^{v} n_i$ e R_i é a soma dos postos atribuídos ao *i*-ésimo tratamento.

Quando *N* é grande, *H* tem distribuição quiquadrado (χ^2) com *v*-1 graus de liberdade. Se o valor de *H* calculado for significativo, rejeita-se a hipótese de igualdade, indicando a presença de pelo menos dois tratamentos estatisticamente diferentes entre si. Caso contrário, não há evidência suficiente para rejeitar a hipótese nula. Campos (1983) representa essas hipóteses da seguinte forma:

$$H_0: t_1 = t_2 \cdots t_v$$

H_1: *pelo menos dois tratamentos diferem entre si.*

Para as comparações múltiplas entre os tratamentos, utiliza-se aqui uma variação do *Wilcoxon rank sum-test*, com a correção de Bonferroni, conforme Daxue Consulting (2023). Essa correção é importante para controlar o erro do tipo I, que é a probabilidade de rejeitar erroneamente a hipótese nula, quando ela é verdadeira. Para realizar uma correção de Bonferroni, deve-se dividir o valor crítico *p* (α) pelo número de comparações que está sendo feito (Napierala, 2014; Groppe, 2023).

Para ilustrar a proposta de análise aqui apresentada, é considerado um experimento, conduzido na Embrapa Milho e Sorgo (Oliveira, 2021), cujo objetivo foi estimar o impacto de sistemas do manejo agrícola sobre a qualidade biológica/saúde do solo, em área de Cerrado no sentido restrito. Como variáveis, foram utilizados os indicadores enzimáticos envolvidos na dinâmica de fósforo (fosfatase ácida e alcalina– µmoles de nitrofenol g^{-1} solo), e na dinâmica de nitrogênio (arginase e urease–µmoles de amônio g^{-1} solo), uma vez que atividades microbianas têm sido usadas para avaliar a influência de práticas de manejo agrícola sobre a saude biológica do solo, que modula sua capacidade produtiva (Mendes; Cherubin, 2023). De modo geral, maior atividade enzimática indica maior abundância e diversidade da comunidade microbiológica do agrossistema, resultando na maior qualidade/saúde e capacidade reprodutiva do solo.

As atividades dessas enzimas foram quantificadas em amostras de solo, coletadas em áreas sob seis sistemas intensificados de cultivo de soja e milho (Simão, 2020), e em uma área sob vegetação de Cerrado, vizinha às demais, que constituíram os tratamentos, conforme a seguir:

a) Sistema 1: monocultura de soja, com médio investimento em adubação (Sm).

b) Sistema 2: milho monocultura, com médio investimento (Mm).

c) Sistema 3: sucessão anual de soja e milho, com médio investimento (SMm).

d) Sistema 4: sucessão de soja e milho, com inserção da braquiária (*Urochloa ruziziensis*), com médio investimento (SMBm).

e) Sistema 5: sucessão de soja e milho, com alto investimento (SMa).

f) Sistema 6: sucessão de soja e milho, com inserção de braquiária, com alto investimento (SMBa).

g) Tratamento referência: área sob vegetação de Cerrado, adjacente aos sistemas de cultivo (VCer).

Em cada área dos tratamentos, foram coletadas dez amostras aleatórias, constituindo um experimento em delineamento completamente casualizado. As análises estatísticas foram realizadas utilizando-se o software R. Os códigos são apresentados no Anexo A.

Resultados e Discussão

Os valores médios das atividades das enzimas fosfatase ácida, fosfatase alcalina, arginase e urease nas sete áreas, submetidas a seis sistemas de cultivos e uma área sob vegetação natural são apresentados na Tabela 2. Valores de maiores magnitudes, para as quatro enzimas, ocorreram na área sob vegetação de Cerrado, e os menores, nas áreas de monocultura.

Tabela 2. Valores médios das atividades das enzimas fosfatase ácida, fosfatase alcalina, arginase e urease nas áreas, submetidas a seis manejos de cultivo e em uma área sob vegetação natural.

Tratamento	Fosfatase ácida	Fosfatase alcalina	Arginase	Urease
1. Monocultura de soja, com médio investimento (Sm)	3.154,19	1.993,17	12,90	128,34
2. Monocultura de milho, com médio investimento (Mm)	3.267,31	2.103,50	17,06	127,69
3. Sucessão soja e milho, com médio investimento (SMm)	3.689,65	2.210,13	23,38	166,97
4. Sucessão soja e milho, com inclusão de braquiária, com médio investimento (SMBm)	3.514,51	2.226,60	23,25	153,25
5. Sucessão de soja e milho, com alto investimento (SMa)	3.702,59	2.226,96	15,61	161,08
6. Sucessão de soja e milho, com inserção de braquiária, com alto investimento (SMBa)	3.681,36	2.222,66	19,54	155,54
7. Área sob vegetação de Cerrado (VCer)	3.892,20	2.501,38	25,61	188,93

As estatísticas descritivas das enzimas são apresentadas na Tabela 3.

Tabela 3. Estatísticas descritivas das variáveis fosfatase ácida, fosfatase alcalina, arginase e urease.

Variável	Fosfatase ácida	Fosfatase alcalina	Arginase	Urease
Média	3.557,40	2.212,06	19,62	154,54
Mediana	3.609,86	2.196,00	19,31	154,05
Desvio padrão	417,64	252,58	4,73	24,34
Mínimo	2.504,38	1.670,00	11,99	119,06
Máximo	5.360,00	3.270,00	34,48	241,77

Os resultados da PCA aplicada à matriz de covariância, considerando as variáveis padronizadas, resultaram nas seguintes componentes principais:

$$\widehat{PC}_1 = 0{,}5243(F.ácida) + 0{,}5178(F.alcalina) + 0{,}4212(Arginase) + 0{,}5285(Urease)$$

$$\widehat{PC}_2 = -0{,}3697(F.ácida) - 0{,}3346(F.alcalina) + 0{,}8667(Arginase) + 0{,}0039(Urease)$$

$$\widehat{PC}_3 = 0{,}0137(F.ácida) - 0{,}6042(F.alcalina) - 0{,}2308(Arginase) + 0{,}7624(Urease)$$

$$\widehat{PC}_4 = 0{,}7668(F.ácida) - 0{,}5046(F.alcalina) + 0{,}1340(Arginase) - 0{,}3731(Urease)$$

As variâncias estimadas ($\hat{\lambda}_i$), as porcentagens da variância total explicadas pelas componentes e as porcentagens acumuladas estão apresentadas na Tabela 4. Destaca-se a primeira componente, que explica 76,44% da variação amostral total. Assim, essa componente pode substituir as *p* variáveis originais sem muita perda de informação. Regazzi (2000), citado por Varella (2023), menciona que são aceitáveis percentuais acima de 70% de explicação.

Tabela 4. Variância, porcentagem da variância total explicada, para cada componente principal, e porcentagem acumulada pelas componentes.

Componente principal	Variância	Porcentagem	Porcentagem acumulada
$\widehat{PC}_1$	3,0577	76,440	76,440
$\widehat{PC}_2$	0,5901	14,750	91,200
$\widehat{PC}_3$	0,2186	5,465	96,666
$\widehat{PC}_4$	0,1336	3,339	100,00

A primeira componente principal $\widehat{PC}_1$, é um índice global da qualidade biológica do solo. Observa-se que os coeficientes dessa componente são todos positivos. Os valores desses coeficientes, para as variáveis fosfatase ácida, fosfatase alcalina e urease, foram similares, com os valores 0,5243, 0,5178 e 0,5285, respectivamente, havendo pequena diferença em relação à arginase, cujo valor foi igual a 0,4212. Portanto, as quatro variáveis são igualmente importantes para a $\widehat{PC}_1$, e, quanto maiores as atividades dessas enzimas, maior será o valor numérico da componente. A $\widehat{PC}_1$ está significativamente correlacionada com essas variáveis (Tabela 5). As demais componentes principais são índices comparativos. Por exemplo, a segunda componente compara os indicadores enzimáticos, envolvidos na dinâmica de fósforo (fosfatase ácida e fosfatase alcalina) com os indicadores envolvidos na dinâmica de nitrogênio (arginase e urease).

Tabela 5. Coeficientes de correlação entre as componentes principais e as variáveis originais.

Variável	PC_1	PC_2	PC_3	PC_4
Fosfatase ácida	0,9169**	-0,2840*	0,0064 NS	0,2802*
Fosfatase alcalina	0,9056**	-0,2570*	-0,2825*	-0,1844 NS
Arginase	0,7366**	0,6658**	-0,1079 NS	0,0490 NS
Urease	0,9242**	0,0030 NS	0,3565**	-0,1364 NS

NS: não significativo. *Significativo a 5% de probabilidade. **Significativo a 1%.

Os escores da primeira componente principal, determinados para as dez amostras das sete áreas em estudo, são apresentados na Tabela 6. As somas dos postos (ranks) desses escores, para cada área, estão na Tabela 7.

Tabela 6. Escores da primeira componente principal determinados em dez amostras das áreas estudadas (tratamentos: Som, Mm, SMm, SMBm, SMa. SMBa e VCer).

Tratamento	Escore	Tratamento	Escore	Tratamento	Escore	Tratamento	Escore
Sm	-2,5067	SMm	0,1237	SMa	-0,2124	VCer	-2,4587
Sm	-2,2924	SMm	0,4518	SMa	-0,1239	VCer	-1,2581
Sm	-2,2877	SMm	0,4985	SMa	-0,1067	VCer	0,4606
Sm	-2,2726	SMm	0,5713	SMa	-0,0777	VCer	1,6832
Sm	-2,1454	SMm	0,7738	SMa	-0,0677	VCer	2,0154
Sm	-2,1021	SMm	0,8938	SMa	-0,0118	VCer	2,5109
Sm	-2,0699	SMm	0,9509	SMa	0,0798	VCer	2,8183
Sm	-1,8969	SMm	1,0372	SMa	0,0858	VCer	3,3132
Sm	-1,8522	SMm	1,0685	SMa	0,1336	VCer	6,2141
Sm	-1,794	SMm	1,2933	SMa	0,2775	VCer	7,6326
Mm	-1,5695	SMBm	-0,136	SMBa	0,0318		
Mm	-1,557	SMBm	0,0285	SMBa	0,0421		
Mm	-1,5019	SMBm	0,0874	SMBa	0,0443		
Mm	-1,4132	SMBm	0,1271	SMBa	0,1005		
Mm	-1,4079	SMBm	0,178	SMBa	0,1982		
Mm	-1,3803	SMBm	0,3671	SMBa	0,2274		
Mm	-1,3334	SMBm	0,4302	SMBa	0,2584		
Mm	-1,3063	SMBm	0,4658	SMBa	0,3035		
Mm	-1,2895	SMBm	0,5042	SMBa	0,3424		
Mm	-1,2202	SMBm	0,6577	SMBa	0,3696		

Tabela 7. Somas dos postos (ranks) dos escores da primeira componente principal, obtidas nas áreas submetidas aos seis diferentes manejos agrícolas e a uma área sob Cerrado.

Tratamento	Soma dos postos dos escores da primeira componente principal (R_i)
1. Monocultura de soja, com médio investimento (Sm)	64
2. Monocultura de milho, com médio investimento (Mm)	166
3. Sucessão soja e milho, com médio investimento (SMm)	562
4. Sucessão soja e milho, com inclusão de braquiária, com médio investimento (SMBm)	433
5. Sucessão de soja e milho, com alto investimento (SMa)	312
6. Sucessão de soja e milho, com inserção de braquiária, com alto investimento (SMBa)	404
7. Área sob vegetação de Cerrado (VCer)	544

A estatística *H*, determinada de acordo com o critério Kruskal-Wallis, apresentou um valor igual a 50,537, resultando na rejeição da hipótese de que os postos dos escores dos tratamentos têm a mesma distribuição, com um nível de significância menor do que 1% de probabilidade (p valor < 0,01).

As comparações par a par, pelo teste de Wilcoxon (*Wilcoxon rank sum-test*), ajustado pelo método de Bonferroni, são apresentadas Tabela 8:

Tabela 8. Valores de *p* (erro tipo I), obtidos pelo *Wilcoxon rank sum-test*, para as comparações entre os sete tratamentos avaliados.

	Sm	Mm	SMm	SMBm	SMa	SMBa
Mm	0,00023					
SMm	0,00023	0,00023				
SMBm	0,00023	0,00023	0,08161			
SMa	0,00023	0,00023	0,00091	0,30848		
SMBa	0,00023	0,00023	0,00682	1	0,24142	
VCer	0,02205	0,04387	1	0,9884	0,48784	0,48784

p<0,05 indica significância menor de 5% de probabilidade.

Observa-se, pela Tabela 8, que não houve diferenças significativas entre o tratamento referência (VCer: área sob vegetação de Cerrado) e os sistemas SMm, SMBm, SMa e SMBa, que têm em comum a sucessão de culturas. Por outro lado, o sistema Sm (monocultivo de soja) e o sistema Mm (monocultivo de milho) diferem significativamente entre si, com atividades enzimáticas inferiores aos demais sistemas (p< 0,01). Esses resultados indicam que sistemas envolvendo sucessão de culturas favorecem a diversidade funcional da comunidade microbiana do solo. Sabe-se que resíduos e raízes de plantas desempenham um papel importante na estruturação dessas comunidades microbianas, liberando uma vasta gama de compostos que diferem entre espécies de plantas (Pausch; Kuzyakov, 2018). Populações microbianas diversas participam na ciclagem biogeoquímica de nutrientes e, consequentemente, impactam significativamente a fertilidade do solo e a sustentabilidade dos sistemas produtivos (Sofo et al., 2022; Mendes; Cherubin, 2023).

Conclusões

O uso da análise proposta permite a comparação entre tratamentos mesmo quando os pressupostos sobre a distribuição dos dados e homogeneidade de variância, exigidos nos testes paramétricas, não forem atendidos, o que frequentemente ocorre em estudos sobre atributos biológicos do solo.

O índice global, representado pela primeira componente principal, apresenta sensibilidade para monitorar a qualidade biológica do solo, considerando enzimas envolvidas na ciclagem de nitrogênio e fósforo.

A incorporação de enzimas envolvidas na ciclagem de outros nutrientes, para a obtenção da primeira componente principal, pode ampliar a representatividade do índice supracitado quanto à qualidade biológica ou à saúde do solo, componentes-chave para a sustentabilidade agrícola.

Referências

ABAPIHI, B.; WIBAWA, G. N. A.; BHARUDDIN; MUKHSAR; AGUSRAWATI; LAOME. ANOVA on principal component as an alternative to MANOVA. **Journal of Physics: Conference Series**, v. 1899, 012103, 2021. DOI: http://doi.org/10.1088/1742-6596/1899/1/012103.

CAMPOS, H. **Estatística experimental não paramétrica**. 4. ed. Piracicaba: Escola Superior de Agricultura “Luiz de Queiroz”, 1983.

DAXUE CONSULTING. **Brief introduction to statistic**. Disponível em: https://bookdown.org/thomas_pernet/Tuto/non-parametric-tests.html#wilcoxon-whitney-wilcoxon-test. Acesso em: 19 maio 2023.

DUARTE, J. B. **Introdução à análise de componentes principais**. Piracicaba: [s.n.], 1998. Disponível em: https://files.cercomp.ufg.br/weby/up/396/o/ACP.pdf. Acesso em: 20 maio 2023.

GROPPE, D. **Bonferroni-Holm correction for multiple comparisons**: version 1.1.0.0. Disponível: https://www.mathworks.com/matlabcentral/fileexchange/28303-bonferroni-holm-correction-for-multiple-comparisons. Acesso em: 23 jun. 2023.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th ed. Englewood Cliffs: Prentice-Hall, 2007. 772 p.

KRUSKAL, W. H.; WALLIS, W. A. Use of ranks in one-criterion variance analysis. **Journal of the American Statistical Association**, v. 47, n. 260, p. 583-621, 1952. DOI: https://doi.org/10.2307/2280779.

MENDES, I. C.; CHERUBIN, M. R. (ed.). **Soil health and sustainable agriculture in Brazil**. Madison: Soil Science Society of America, 2023. 275p.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivada**: uma abordagem aplicada. Belo Horizonte: Universidade Federal de Minas Gerais, 2007. 297 p.

NAPIERALA, M. A. **What is the Bonferoni Correction**. 2014. Disponível em: https://www.semanticscholar.org/paper/What-Is-the-Bonferroni-Correction-Napierala/d981fdd547036e35d80fa771341c2d71e196dd82. Acesso em: 23 jun. 2023

OLIVEIRA, T. S. S. **Atributos microbiológicos do solo como indicadores de sustentabilidade de sistemas de produção de grãos no cerrado central mineiro**. 2021. 75 f. Dissertação (Mestrado em Bioengenharia) - Universidade Federal de São João del-Rei, Sete Lagoas.

PAUSCH, J.; KUZYAKOV, Y. Carbon input by roots into the soil: quantification of rhizodeposition from root to ecosystem scale. **Global Change Biology**, v. 24, n. 1, p. 1-12, 2018. DOI: https://doi.org/10.1111/gcb.13850.

REIS, E. **Estatística multivariada aplicada**. 2. ed. Lisboa: Sílabo, 2001. 343 p.

SIMÃO, E. P. **Desempenho produtivo e econômico de sistemas intensificados de cultivo de soja e milho na região central de Minas Gerais**. 2020. 71 f. Tese (Doutorado em Fitotecnia) - Universidade Federal de Viçosa, Viçosa, MG.

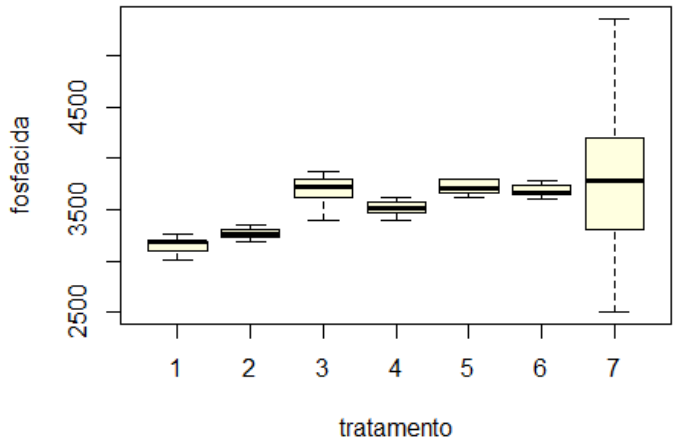
SOFO, A.; ZANELLA, A.; PONGE. J. F. Soil quality and fertility in sustainable agriculture, with a contribution to the biological classification of agricultural soils. **Soil Use and Management**, v. 38, n. 2, p. 1085-1112, 2022. DOI: https://doi.org/10.1111/sum.12702.

VARELLA, A. A. C. **Análise de componentes principais**. 2023. 12 p. Monografia (Pós-graduação em Agronomia) - Seropédica: Universidade Federal Rural do Rio de Janeiro, Seropédica. Disponível em: http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20as%20ciencias%20agrarias/Aulas/analise%20de%20componentes%20principais.pdf. Acesso em: 12 maio 2023.

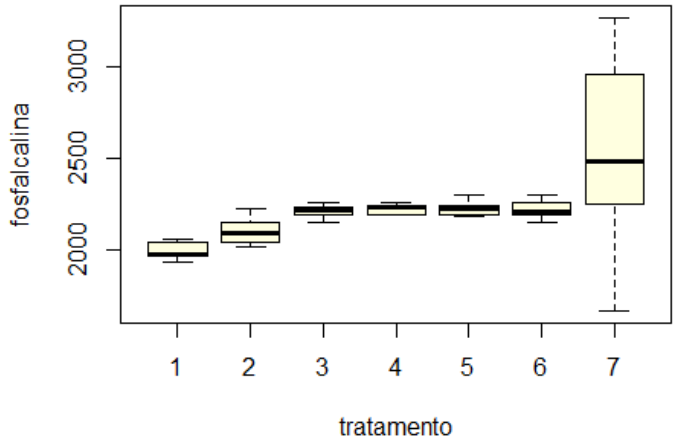
Anexo A

Análise de experimento sobre qualidade biológica de solo com base em quatro indicadores enzimáticos em sete ambientes. Sete Lagoas, MG, 2023.

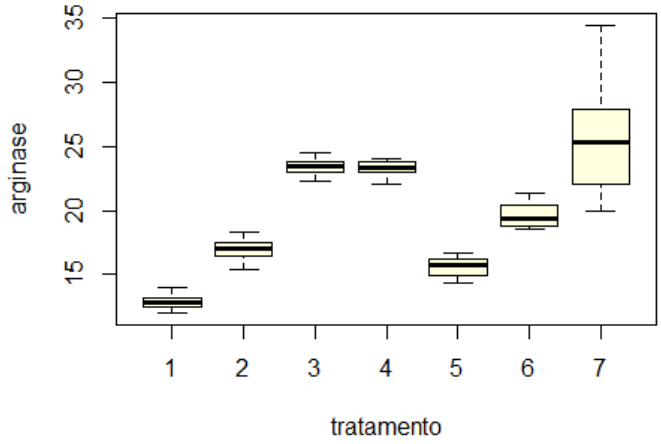
```
#Bibliotecas

##install.packages("GGally")
##install.packages("factoextra")
##install.packages("corrr")
##install.packages("ggcorrplot")
##install.packages("FactoMineR")
##install.packages("factoextra")
library("ggpubr")
library("tidyverse")
library("rstatix")
library("car")
library("readxl")
library("corrr")
library("ggcorrplot")
library("FactoMineR")
library("factoextra")

#Conjunto de dados
qsolo_data=read_excel('dados/dadosenzimas.xlsx',col_names=TRUE)

#Seleção das colunas
qsolo_data_trat <- qsolo_data %>%
  select(tratamento, fosfacida,fosfalcalina, arginase,urease,) %>%
  add_column(id = 1:nrow(qsolo_data), .before = 1)
head(qsolo_data_trat)

## # A tibble: 6 x 6
##      id tratamento fosfacida fosfalcalina arginase urease
##   <int>      <dbl>     <dbl>        <dbl>    <dbl>  <dbl>
## 1     1          1     3180.        1976.     12.4   123.
## 2     2          1     3232.        1933.     12.0   126.
## 3     3          1     3006.        2042.     14.0   138.
## 4     4          1     3093.        2020.     12.9   128.
## 5     5          1     3181         1978      13.9   137.
## 6     6          1     3180.        1976.     12.4   123.

##Boxplot das enzimas (variáveis originais) para cada  tratamento de
Produção
with(qsolo_data_trat, boxplot(qsolo_data_trat$fosfacida ~
qsolo_data_trat$tratamento,ann=TRUE, col='lightyellow',
xlab='tratamento', ylab='fosfacida'))
```



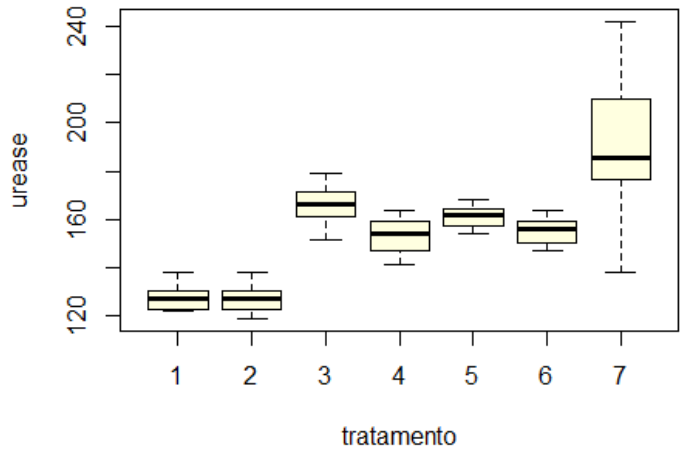

```
with(qsolo_data_trat, boxplot(qsolo_data_trat$fosfalcalina  ~
qsolo_data_trat$tratamento,ann=TRUE, col='lightyellow',
xlab='tratamento', ylab='fosfalcalina'))
```



```
with(qsolo_data_trat, boxplot(qsolo_data_trat$arginase  ~
qsolo_data_trat$tratamento,ann=TRUE, col='lightyellow',
xlab='tratamento', ylab='arginase'))
```




```
with(qsolo_data_trat, boxplot(qsolo_data_trat$urease  ~
qsolo_data_trat$tratamento,ann=TRUE, col='lightyellow',
xlab='tratamento', ylab='urease'))
```



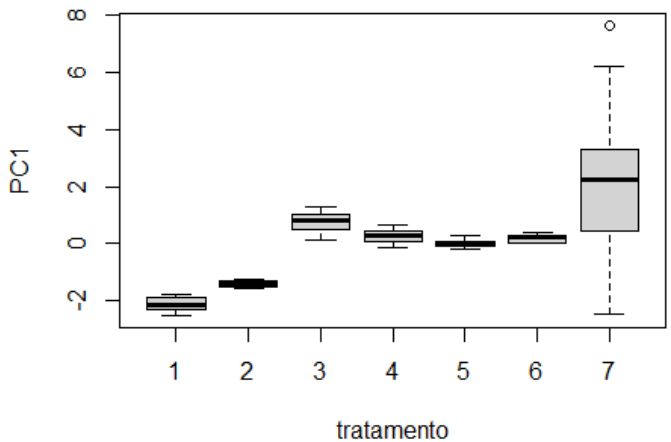

```
#Análise de Componentes Principais (PCA)

##Conjunto de dados
dadosPCA<-qsolo_data
dadosPCA<-qsolo_data[,-c(1)]
head(dadosPCA)

## # A tibble: 6 x 4
##   fosfacida fosfalcalina arginase urease
##       <dbl>        <dbl>    <dbl>  <dbl>
## 1     3180.        1976.     12.4   123.
## 2     3232.        1933.     12.0   126.
## 3     3006.        2042.     14.0   138.
## 4     3093.        2020.     12.9   128.
## 5     3181         1978      13.9   137.
## 6     3180.        1976.     12.4   123.

##Matriz de correlações
corr_matrix <- cor(dadosPCA)
corr_matrix

##             fosfacida fosfalcalina  arginase    urease
## fosfacida    1.0000000    0.8498439 0.4993136 0.8106881
## fosfalcalina 0.8498439    1.0000000 0.5173713 0.7606639
## arginase     0.4993136    0.5173713 1.0000000 0.6376849
## urease       0.8106881    0.7606639 0.6376849 1.0000000
```

```
##Aplicando a análise de componentes principais (PCA)
res.pca <- prcomp(dadosPCA, scale = TRUE)
res.pca

## Standard deviations (1, .., p=4):
## [1] 1.7486269 0.7681899 0.4675648 0.3654740
##
## Rotation (n x k) = (4 x 4):
##                    PC1          PC2         PC3        PC4
## fosfacida    0.5243567 -0.369773919  0.01373558  0.7668954
## fosfalcalina 0.5178789 -0.334644441 -0.60429021 -0.5046265
## arginase     0.4212574  0.866755437 -0.23089728  0.1340286
## urease       0.5285771  0.003919172  0.76245073 -0.3731753

##Importância dos componentes
summary(res.pca)

## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7486 0.7682 0.46756 0.36547
## Proportion of Variance 0.7644 0.1475 0.05465 0.03339
## Cumulative Proportion  0.7644 0.9120 0.96661 1.00000

##Obs: Os scores para cada tratamento foram  armazenados em res.pca$x


##Lendo os escores
head(res.pca$x)

##            PC1        PC2         PC3         PC4
## [1,] -2.292381 -0.6761067 -0.10143209  0.06573318
## [2,] -2.272602 -0.7442677  0.14652469  0.17792792
## [3,] -1.896933 -0.3178063  0.14766141 -0.57976591
## [4,] -2.145357 -0.5621839 -0.05987806 -0.25231466
## [5,] -1.852169 -0.4146412  0.27063652 -0.11358014
## [6,] -2.287683 -0.6827623 -0.08815911  0.05907400



#Incluindo a coluna dos tratamentos na matriz de escores
tratamento<- gl(7,10,labels=c(1:7))
datascore=cbind(tratamento,res.pca$x)
head(datascore)

##      tratamento       PC1        PC2         PC3         PC4
## [1,]          1 -2.292381 -0.6761067 -0.10143209  0.06573318
## [2,]          1 -2.272602 -0.7442677  0.14652469  0.17792792
## [3,]          1 -1.896933 -0.3178063  0.14766141 -0.57976591
## [4,]          1 -2.145357 -0.5621839 -0.05987806 -0.25231466
## [5,]          1 -1.852169 -0.4146412  0.27063652 -0.11358014
## [6,]          1 -2.287683 -0.6827623 -0.08815911  0.05907400
```

```
##Transformando matriz datascore em data-frame
dataframe_data=as.data.frame(datascore)

##Transformando tratamento em fator
dataframe_data$tratamento<-as.factor(dataframe_data$tratamento)
str(dataframe_data)
```

```
## 'data.frame':    70 obs. of  5 variables:
##  $ tratamento: Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1 1 1 1
1 1 ...
##  $ PC1       : num  -2.29 -2.27 -1.9 -2.15 -1.85 ...
##  $ PC2       : num  -0.676 -0.744 -0.318 -0.562 -0.415 ...
##  $ PC3       : num  -0.1014 0.1465 0.1477 -0.0599 0.2706 ...
##  $ PC4       : num  0.0657 0.1779 -0.5798 -0.2523 -0.1136 ...
```

```
#Boxplot para PC1 nos diferentes tratamentos
boxplot(PC1 ~ tratamento,
        data = dataframe_data
)
```




```
#Testes não paramétricos de Kruskal-Wallis e de Wilcoxon para o PC1, para
comparar os tratamentos

kruskal.test(PC1 ~ tratamento, data = dataframe_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
```

23

```
## data:  PC1 by tratamento
## Kruskal-Wallis chi-squared = 50.537, df = 6, p-value = 3.669e-09
```

```
pairwise.wilcox.test(dataframe_data$PC1,
             dataframe_data$tratamento,
             p.adjust.method="bonferroni")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum exact test
##
## data:  dataframe_data$PC1 and dataframe_data$tratamento
##
##   1       2       3       4       5       6
## 2 0.00023 -       -       -       -       -
## 3 0.00023 0.00023 -       -       -       -
## 4 0.00023 0.00023 0.08161 -       -       -
## 5 0.00023 0.00023 0.00091 0.30848 -       -
## 6 0.00023 0.00023 0.00682 1.00000 0.24142 -
## 7 0.02205 0.04387 1.00000 0.90840 0.48784 0.48784
##
## P value adjustment method: bonferroni
```

MINISTÉRIO DA
AGRICULTURA E
PECUÁRIA

GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO

CGPE 018403