BOLETIM DE PESQUISA E DESENVOLVIMENTO
58

ISSN 2764-2623
Dezembro/2023

Correlação de atributos do solo com a produção de café para auxiliar a análise de variabilidade espacial

Embrapa

ISSN 2764-2623
Dezembro/2023

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Agricultura Digital
Ministério da Agricultura e Pecuária

BOLETIM DE PESQUISA E DESENVOLVIMENTO 58

Correlação de atributos do solo com a produção de café para auxiliar a análise de variabilidade espacial

Adauto Luiz Mancini
Maria Fernanda Moura
Célia Regina Grego
Gustavo Costa Rodrigues
Cristina Aparecida Gonçalves Rodrigues

Embrapa Agricultura Digital
Campinas, SP
2023

Embrapa Agricultura Digital

Av. Dr. André Tosello, 209 - Cidade Universitária
Campinas, SP, Brasil
CEP. 13083-886
Fone: (19) 3211-5700
www.embrapa.br

Revisão de texto
Graziella Galinari

Normalização bibliográfica
Carla Cristiane Osawa

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Magda Cruciol

Imagem da capa
Magda Cruciol

1º edição
Publicação digital 2023: PDF

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Agricultura Digital

Correlação de atributos do solo com a produção de café para auxiliar a análise de variabilidade espacial / Adauto Luiz Mancini … [et al.]. – Campinas : Embrapa Agricultura Digital, 2023.
PDF (28 p.) : il. color. - (Boletim de pesquisa e desenvolvimento / Embrapa Agricultura Digital, ISSN 2764-2623 ; 58).

1. Análise multivariada. 2. Geoestatística. 3. Seleção de atributos. II. Moura, Maria Fernanda. III. Grego, Célia Regina. IV. Rodrigues, Gustavo Costa. V. Rodrigues, Cristina Aparecida Gonçalves. VI. Série.

CDD (21. ed.) 519.5

Carla Cristiane Osawa (CRB-8/10421)

Sumário

Correlação de atributos do solo com a produção de café para auxiliar a análise de variabilidade espacial

Adauto Luiz Mancini[1]
Maria Fernanda Moura[2]
Célia Regina Grego[3]
Gustavo Costa Rodrigues[4]
Cristina Aparecida Gonçalves Rodrigues[5]

Resumo - A identificação espacial dos atributos de solo mais fortemente correlacionados com a produção agrícola ajuda no mapeamento da produtividade do talhão, auxiliando o produtor rural na tomada de decisão quanto ao manejo e uso de insumos na lavoura. O uso de técnicas de análise de dados, como análise multivariada, mineração de dados e geoestatística permitem analisar adequadamente estes atributos. Este trabalho mostra o resultado após o emprego destas técnicas para identificar, a partir da análise de solo e da medição da condutividade elétrica de um talhão de café especial localizado em uma fazenda no sul de Minas Gerais, os atributos mais correlacionados com a produção amostrada em campo para o ano de 2022. Os mapas SOM (Self Organizing Map) e a análise de correlação simples mostraram-se ferramentas eficientes para selecionar os atributos, porém necessitando de análise manual/visual realizada pelos cientistas de dados (estes métodos não têm como objetivo direto a seleção de atributos). Os métodos CFS (Correlation-based Feature Selection) e os métodos baseados em ganho de informação (árvore de decisão e M5 Rules) do Weka, também apresentaram bom desempenho. Os macronutrientes pH em $CaCl_2$, H, H+Al e os micronutrientes S, Fe, B, Cu, Zn foram os mais importantes para determinação dos valores de produção no conjunto de dados analisado. A seleção desses atributos auxiliou na análise geoestatística e mapeamento direcionado apenas a essas variáveis de interesse.

[1] Bacharel em Ciências da Computação, mestre em Ciências da Computação, pesquisador da Embrapa Agricultura Digital, Campinas, SP

[2] Estatística, doutora em Ciências da Computação, pesquisadora da Embrapa Agricultura Digital, Campinas, SP.

[3] Engenheira-agronôma, doutora em Agronomia, pesquisadora da Embrapa Agricultura Digital, Campinas, SP.

[4] Agrônomo, mestre em Fisiologia Vegetal, pesquisador da Embrapa Agricultura Digital, Campinas, SP

[5] Zootecnista, doutora em Biologia Vegetal, pesquisadora da Embrapa Territorial, Campinas, SP.

Termos para indexação: Análise multivariada, geoestatística, seleção de atributos

Correlation of soil attributes with coffee production to assist the analysis of spatial variability

Abstract - The spatial identification of the soil attributes most strongly correlated with production helps in mapping the plot productivity, helping the producer in making decisions regarding the management and use of inputs in the field. The use of data analysis techniques, such as multivariate analysis, data mining and geostatistics, allows us to analyze these attributes. This work shows the effort made in using these techniques to identify, based on soil analysis of a specialty coffee plot located on a farm in the south of Minas Gerais, the attributes most correlated with the production sampled in the field for the year 2022. SOM (Self Organizing Map) maps and simple correlation analysis proved to be efficient tools for selecting attributes, but requiring manual/visual analysis performed by data scientists (these methods do not directly aim at attribute selection). The CFS (Correlation-based Feature Selection) method and methods based on information gain (decision tree and M5 Rules) from Weka also performed well. The macronutrients pH_$CaCL_2$, H, H+Al and the micronutrients S, Fe, B, Cu, Zn were the most important for determining production values in the analyzed data set. The selection of these attributes contributed to the geostatistical analysis and mapping directed only to these variables of interest.

Index Terms: Multivariate analysis, geostatistics, attribute selection

Introdução

A produção dos cafés do Brasil, que é mundialmente o maior produtor e exportador, só cresce com o passar dos anos. Em 2023, a previsão é um aumento de 7,9%, em relação à produção de café de 2022. Dentre os estados brasileiros, o maior produtor em relação à área total é Minas Gerais com 1,9 milhão de hectares neste ano de 2023 (Acompanhamento…, 2023).

É sabido que fatores ambientais, tais como o solo, clima e relevo, afetam o sistema de produção. Ferramentas de agricultura de precisão, segundo Silva e Alves (2013), são utilizadas para auxiliar na identificação da variabilidade espacial desses fatores dentro de um mesmo talhão, para otimizar a capacidade produtiva do café de qualidade. Dentro da agricultura de precisão, em conjunto com técnicas de análise de dados, como análise multivariada, mineração de dados e geoestatística, é possível identificar espacialmente quais atributos de solo estão correlacionados com a produção e, assim, ajudar no mapeamento do talhão e consequentemente auxiliar o produtor na tomada de decisão quanto ao manejo e uso de insumos na lavoura. Considerando o exposto, o objeto deste trabalho foi utilizar métodos estatísticos para selecionar atributos de solo que melhor se correlacionam com a produção de café para a espacialização dentro de um talhão de café especial numa fazenda do sul de Minas Gerais, no ano de 2022.

Material e métodos

Os dados obtidos foram coletados em trabalho de campo realizado no ano de 2022, em área de produção de café especial localizada no município de Paraguaçu (Figura 1), Minas Gerais, com variedade Catucaí Amarelo, totalizando aproximadamente três hectares.

Na fase de colheita, em junho de 2022, foi realizada a amostragem da produção do café em grade de 40 pontos georreferenciados. Em cada ponto de amostragem foram colhidas duas plantas de café pelo método do pano e, após separação das folhas e galhos, foi realizada a pesagem e obtida a produção de grãos por planta. Nos mesmos pontos da colheita, foram coletadas amostras de solo 0-20 cm de profundidade para determinação de macronutrientes (pH em $CaCl_2$ e em água, MO em mg/dm^3, P em mg/dm^3, K, Ca, Mg,

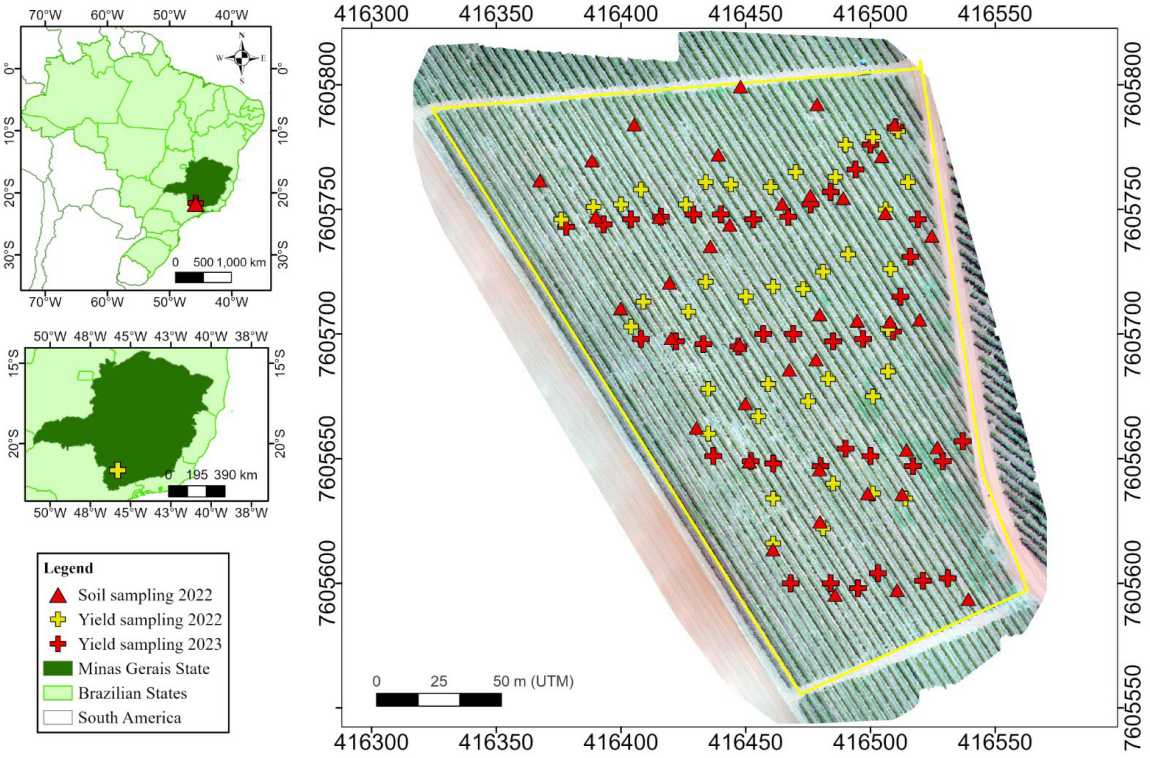

Figura 1. Área de experimentos da fazenda Santa Cruz.

H, H+Al, SB, CTC em mmolc/dm^3, V%, Ca/Mg, Mg/K) e dos micronutrientes do solo S, Na, Fe, Mn, Cu, Zn e B em mg/dm^3.

Também foram obtidos e analisados os dados de condutividade elétrica aparente do solo (CE) rasa 0-20 cm e profunda 0-40 cm a partir do sensor portátil de contato desenvolvido pela Embrapa Instrumentação descrito por Rabello et al. (2011).

Descrição dos dados

Os dados georreferenciados, todos na mesma grade de coordenadas de pontos, foram submetidos a vários métodos de análise estatística, multivariada e de correlação.

A grande quantidade de variáveis de solo, 22 (Tabela 1), dificultou o estabelecimento de uma correlação entre estas variáveis e a produção por ponto amostral. Ainda, na Tabela 1, encontram-se as descrições das variáveis observadas.

Tabela 1. Descrição das variáveis de solo observadas.

Variáveis de Macronutrientes do solo	Descrição dos Macronutrientes do solo	Variáveis de Micronutrientes do solo	Descrição dos Micronutrientes do solo
pH_$CaCl_2$	Concentração de hidrogênio na solução do solo	S	Teor de enxofre
pH_H2O	Concentração de hidrogênio na solução do solo	NA	Teor de sódio
MO	Teor de matéria orgânica	Fe	Teor de ferro
P	Teor de fósforo	Mn	Teor de manganês
K	Teor de potássio	Cu	Teor de cobre
Ca	Teor de cálcio	Zn	Teor de zinco
Mg	Teor de magnésio	B	Teor de boro
Al	Teor de alumínio	CE rasa	Condutividade elétrica na prof. rasa 0-20 cm
H	Teor de hidrogênio	CE profunda	Condutividade elétrica na prof. profunda 0-40 cm
H+Al	Teor de hidrogênio + alumínio	Produção	Produção de grãos de café verde em kg/planta
SB	Soma de Bases		
CTC	Capacidade de troca catiônica		
V%	Saturação por Bases		

Métodos para a seleção de variáveis

Em decorrência da grande quantidade de variáveis independentes (ou atributos, como serão referenciados a partir deste ponto) – ao todo vinte e duas (22) – e da pouca amostragem disponível – quarenta (40) valores observados –, optou-se por utilizar métodos de seleção de atributos não supervisionados mais robustos, tais como a análise de componentes principais e análise de agrupamentos, e também utilizar outros métodos supervisionados tais como a análise de correlação simples e outros com base em redução da entropia do sistema, a fim de chegar a um conjunto de variáveis que tivessem um certo poder de discriminação e representatividade em relação à produção observada. Uma breve descrição de cada um desses métodos é dada nos próximos itens.

Não supervisionados

Análise de Componentes Principais: é uma técnica de análise estatística multivariada para obter uma representação ortogonal das variáveis observadas, no caso, aplicada aos atributos observados; isto é, os vetores de atributos são representados no plano de seus autovalores e autovetores. Como o método é sensível à escala original das variáveis, ele é aplicado à matriz de correlação dos dados normalizados. A partir dos autovalores, pode-se escolher o número de componentes principais a utilizar a partir da porção explicada da variância para o conjunto de autovalores. O método prcomp do Software R executa o cálculo das componentes principais e das porções explicadas da variância, por exemplo:

```
# sendo RZ_Solo a matriz de correlação dos dados padronizados
PCA <- prcomp(RZ_Solo)  # calcula as componentes principais
summary(PCA)            # mostra os valores das componentes calculadas
# e para plotar o gráfico da proporção de variância explicada
screeplot(PCA, type = c("lines"), main = deparse(substitute(PCA)))
# ou o histograma
screeplot(PCA)
```

Agrupamento: escolheu-se a rede neural **SOM** (Self Organizing Map), também conhecido como rede de Kohonen. A rede SOM pertence à família dos algoritmos autocodificadores, que são algoritmos não supervisionados e determinam uma função codificadora que representa o dado de entrada em um outro espaço de dados (possivelmente com diferente número de dimensões) e uma função decodificadora que a partir de um dado codificado retorna um valor igual ou próximo do espaço original de entrada de dados. Na rede SOM, é gerado um mapa (grade) de vetores centróides. Os vetores de entrada mais parecidos com os valores do vetor centróide pertencem ao grupo representado pelo centróide. Centróides parecidos são posicionados próximos em uma grade bidimensional. Para o presente estudo foi utilizado o pacote Kohonen do software R e escolhidos mapas com grades 2 linhas por 2 colunas (4 grupos) e 3 linhas por 3 colunas (9 grupos) com vizinhança em formato hexagonal. Ainda, escolheu-se a versão supervisionada oferecida no pacote (classe xyf) para reforçar o agrupamento de vetores de entrada com produção semelhante em um mesmo centróide.

Supervisionados

Análise de Correlação: geralmente busca-se essa análise para determinar o grau de proximidade entre vetores em um plano, cuja relação entre eles é linear (de fato, supõe-se que seja linear). Utilizando o coeficiente de correlação de Pearson, tem-se para cada vetor de atributo e a variável de interesse (produção) o grau de proximidade entre eles, bem como pode-se calcular a proximidade entre os atributos. Como a correlação pode mostrar uma relação positiva (quando um cresce, o outro cresce), ou negativa (quando um cresce, o outro decresce), os valores apresentados vão de [-1,1] e quanto mais próximos dos limites do intervalo consideram-se mais correlacionados.

Entropia/Ganho de Informação: a entropia mede a quantidade de informação associada a uma variável aleatória e a entropia conjunta entre duas variáveis, sendo a interseção das conjuntas a quantidade de informação mútua. O ganho de informação corresponde à diferença entre a entropia de uma variável e a conjunta entre esta e alguma outra.

Para separar conjuntos menores de atributos, selecionaram-se nesta metodologia, os seguintes métodos do [Weka:

Árvore de Decisão com método de busca Best First e Greedy: no software Weka, com o algoritmo J48 baseado no C4.5 (SALZBERG, 1994). Busca pelos atributos que reduzem a entropia do sistema a fim de melhor explicar as classes de interesse; o foco neste trabalho, utilizando dois modelos de busca (Best First e Greedy), é separar um conjunto de atributos mais relevantes – isto é, os que melhor resumem a árvore de decisão gerada.

Algoritmo M5 Rules, com base em taxa de ganho de informação: adaptação do C4.5 para problemas de regressão, gera uma lista de decisões para problemas de regressão usando as estratégias de separar e conquistar, com base na taxa de ganho de informação (Holmes et al., 1999). Em cada iteração ele constrói uma árvore modelo usando M5 e transforma a "melhor" folha em uma regra. O foco, neste trabalho, é utilizar as melhores listas de atributos.

CFS – Correlação, com método de busca Best First e Greedy Stepwise, no Software Weka. Este seletor avalia o valor de um subconjunto de atributos considerando a capacidade preditiva individual de cada recurso juntamente com o grau de redundância entre eles. São preferidos subconjuntos de recursos que são altamente correlacionados com a classe, embora tenham baixa intercorrelação – entre os atributos selecionados (Hall, 1999).

Relief, do Software Weka: avalia o valor de um atributo amostrando repetidamente uma instância e considerando o valor do atributo fornecido para a instância mais próxima da mesma classe e de classe diferente. Pode operar em dados de classe discretos e contínuos (Kira; Rendell, 1992; Kononenko, 1994; Robnik-Sikonja; Kononenko, 1997), pois detecta condições de dependências entre atributos e fornece uma visão unificada sobre a estimativa de atributos em regressão e/ou classificação.

Escolha do conjunto de atributos

Para efeito de comparação entre os conjuntos de atributos selecionados podem-se utilizar classificadores e comparar o erro quadrático médio obtido em cada ajuste. O erro quadrático médio corresponde à média das distâncias (ao quadrado, para evitar a nulidade da soma) entre os valores estimados pelo classificador e os valores observados. É a comparação mais

indicada entre classificadores, pois se algum deles for viciado a medida de redução da variância pode não ser adequada. E, lembrando que o número de amostras é pequeno e que o comportamento da produção observada não é linear, procurou-se valorizar mais métodos não paramétricos, tal como a rede neural Multi-layer Perceptron, sem desconsiderar ajustes gaussianos e regressão linear por discretização. Dessa forma, foram escolhidos os métodos de classificação, para comparação, todos nas implementações disponíveis no software Weka:

MLP - Multi-layer perceptron: rede neural alimentada para a frente com neurônios conectados por diversas ligações com pesos associados e dividida em diversas camadas.

SMO com kernel RBF: o algoritmo Sequential Minimal Optimization (SMO) é derivado levando a ideia do método de decomposição ao extremo e otimizando um subconjunto mínimo de apenas dois pontos em cada iteração. Com Kernel RBF que é o mais semelhante à distribuição gaussiana.

Função Gaussiana: processos gaussianos para regressão sem ajuste de hiperparâmetros, partindo da normalização dos dados de entrada a fim de reduzir ruídos; valores ausentes são substituídos pela média/moda global.

Regressão por discretização: utilizando-se o estimador univariado (apenas produção) para os histogramas que dão origem às classes e utilizando J48 para a classificação.

Regressão linear simples: sem seleção de atributos, pois a seleção habitual para a implementação no Weka é o M5 Rules, que é um dos conjuntos alvo do teste.

Análise geoestatística

A partir das variáveis selecionadas pelos métodos descritos, que melhor discriminaram as que se correlacionaram com a produção do café, foi realizada a análise geoestatística para a detecção de variabilidade espacial e construção de mapas de valores interpolados utilizando software GEOEST desenvolvido por Vieira (2000).

Resultados e discussão

Inicialmente, tomando-se os dados de produção de 2022, procedeu-se a uma análise exploratória simples, na qual pode-se observar que: a distribuição dos dados de produção é não linear (observe o gráfico da Figura 2); e há uma grande diferença de escalas entre as variáveis observadas. Dessa forma, em vários dos métodos de seleção de atributos foram utilizados os valores normalizados de cada variável.

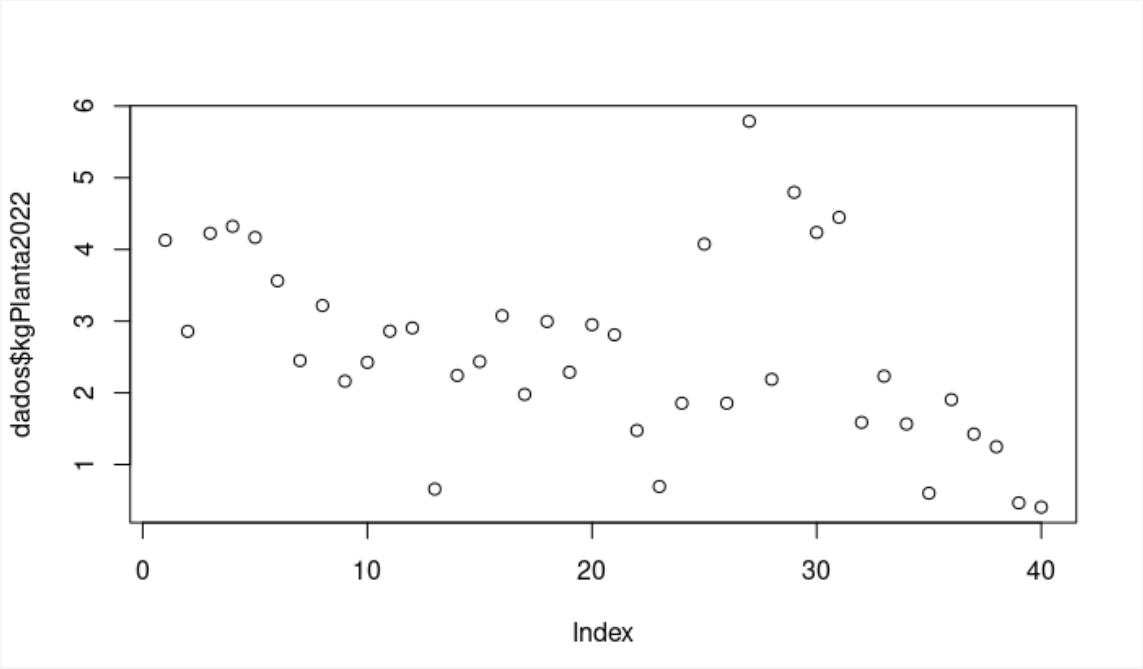

Figura 2. Distribuição da produção de café por ponto de observação.

A seguir, cada método de seleção de atributos foi aplicado obtendo-se a tabela de comparação dos erros quadráticos, escolhidos os melhores conjuntos de atributos e analisados os mapas de produção a partir de cada atributo dos melhores conjuntos selecionados.

Análise de componentes principais

O espaço de componentes principais é calculado a partir da normalização dos dados e de sua matriz de correlação. Assim, primeiramente os dados foram normalizados, calculou-se a matriz de correlação e aplicou-se um teste de significância das correlações. O teste de esfericidade busca verificar se a matriz de correlações não é próxima à matriz identidade, pois caso seja, as correlações não possuem significância. No software R, procedeu-se a:

```
z <- scale(dados)                          # normalizando os dados
RZ <- round(cor(z), 2)                     # obtendo a matriz de correlações
#install.packages("psych")                 # pacote com o teste de esfericidade
library(psych)
resTesteCorZ <- cortest.bartlett(RZ, n = nrow(dados)) # executando o teste
resTesteCorZ                               # obtendo o p-value do teste
```

Como o p-value obtido foi de 1.877763e-106, considera-se que as correlações obtidas não são desprezíveis. Assim, vamos analisar as componentes principais, utilizando o R:

```
PCA <- prcomp(RZ_Solo)
summary(PCA)
screeplot(PCA, type = c("lines"), main = deparse(substitute(PCA)))
screeplot(PCA)
```

Pelo gráfico da composição de variância resultante, Figura 3, observamos que dez componentes principais são suficientes para explicar praticamente a oralidade da variância. Assim, o novo conjunto de dados a ser utilizado nas predições (ou nos classificadores) foi obtido por:

```
novaMatriz <- dados[1:40,1:24]                    # apenas as variáveis independentes
novoX <- as.matrix(novaMatriz)%*%as.matrix(ComPR) # transformação para as PCAs
novos_dados <- cbind(novoX, dados[,25])           # incluindo a produção observada
```

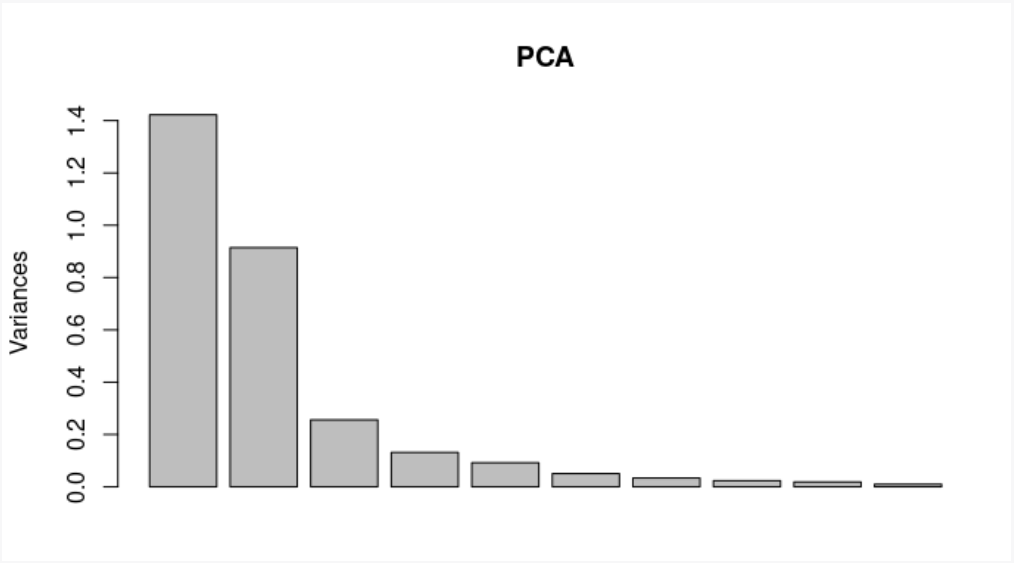

Figura 3. Percentual acumulado da proporção da variância explicada por componente.

Os novos dados foram gravados em arquivo csv para serem levados para o Software Weka, a fim de serem submetidos aos classificadores para escolha de atributos.

Análise de correlação

Esta análise foi realizada com os valores obtidos na análise de componentes principais, ainda no software R. Como os valores das correlações de cada variável com a produção observada foram muito pequenos, optou-se por uma análise visual da matriz de correlação, auxiliada pelos quartis dos valores das correlações. Considerou-se os seguintes quartis, sempre do valor do módulo da correlação:

- Para o terceiro quartil (Q3) valores acima de 0.21; e
- Para o segundo quartil (Q2) valores de 0.12 a menores que 0.21.

Assim, selecionaram-se quatro conjuntos de atributos, a partir de suas correlações com a variável de produção e a seguir entre os cruzamentos dos quartis – a seleção está ilustrada na Figura 4. Os conjuntos de atributos foram:

- CORR1 = {ph_$CaCl_2$, H, H+Al, V%, S, Fe}: atributos cuja correlação com a produção encontra-se no terceiro quartil (Q3);
- CORR2 = {ph_$CaCl_2$, H, H+Al, V%, S, Fe, P, Al, CTC, Mn}: atributos cuja correlação com a produção encontra-se no segundo quartil (Q2) e terceiro quartis (Q3);
- CORR3 = {ph_$CaCl_2$, H, H+Al, V%, S, Fe, P, Al, CTC, Mn, ph_H_2O, Ca, Mg, SB}: atributos cuja correlação com a produção encontra-se a partir do segundo quartil (Q2) e aqueles cujas intersecções com os do Q3 estão acima de 0.5; e
- CORR4 = {ph_$CaCl_2$, H, H+Al, V%, S, Fe, P, Al, CTC, Mn, MO, Cu, B}: atributos já selecionados e aqueles cujas intersecções com os do Q2 estão acima de 0.5.

Agrupamento por rede neural SOM

Inicialmente foi gerado um mapa de dimensões 2x2, ordenou-se os centróides em produção crescente e procurou-se por atributos que aparentemente aumentavam ou diminuíam com o crescimento da produção (Tabela 2).

	ce1	ce2	pH_CaCl2	pH_H2O	MO	P	K	Ca	Mg	Al	H	H_mais_Al	SB	CTC	V_perc	Ca_div_Mg	Mg_div_K	S	Na	Fe	Mn	Cu	Zn	B	kgPlanta2022	Quartis Abs(Corr)
ce1	1	0.83	0.4	0.46	0.19	0.14	0.09	0.5	0.17	-0.41	-0.3	-0.28	0.44	0.34	0.49	0.41	-0.18	-0.17	0.07	-0.21	0.36	0.02	0.08	0.24	0.02	
ce2	0.83	1	0.19	0.3	0.24	0.14	0.09	0.41	0.08	-0.4	-0.09	-0.07	0.36	0.33	0.31	0.38	-0.1	-0.01	0.02	-0.32	0.17	-0.01	0.1	0.18	0.03	
pH_CaCl2	0.4	0.19	1	0.68	0.14	-0.01	-0.05	0.44	0.27	-0.41	-0.9	-0.9	0.37	0.03	0.84	0.18	0.11	-0.34	0.01	0.33	0.12	-0.06	-0.2	0.05	-0.39	Q3 = (0.21,...]
pH_H2O	0.46	0.3	0.68	1	0.04	0.01	0.02	0.31	0.17	-0.31	-0.66	-0.64	0.26	0.02	0.6	0.2	-0.07	-0.47	-0.5	0.04	0.11	-0.14	-0.3	0.16	-0.1	
MO	0.19	0.24	0.14	0.04	1	0.75	0	0.76	0.68	-0.1	0.07	0.06	0.72	0.75	0.39	-0.08	0.54	0.4	0.3	-0.26	0.7	0.54	0.49	0.55	0.05	
P	0.14	0.14	-0.01	0.01	0.75	1	-0.03	0.62	0.56	0.12	0.16	0.15	0.59	0.65	0.23	-0.1	0.33	0.45	0.33	-0.08	0.55	0.44	0.28	0.49	0.17	Q2 = (0.12,0.21]
K	0.09	0.09	-0.05	0.02	0	-0.03	1	0	0.01	-0.05	0.03	0.04	0.35	0.37	0.16	-0.03	-0.48	0.24	0.15	-0.01	0.01	-0.08	-0.11	0.2	-0.04	
Ca	0.5	0.41	0.44	0.31	0.76	0.62	0	1	0.79	-0.36	-0.25	-0.26	0.93	0.84	0.75	0.09	0.47	0.16	0.26	-0.2	0.67	0.28	0.43	0.52	0.03	
Mg	0.17	0.08	0.27	0.17	0.68	0.56	0.01	0.79	1	-0.15	-0.19	-0.2	0.83	0.76	0.63	-0.5	0.59	0.12	0.21	-0.16	0.65	0.21	0.43	0.63	0.04	
Al	-0.41	-0.4	-0.41	-0.31	-0.1	0.12	-0.05	-0.36	-0.15	1	0.3	0.31	-0.32	-0.21	-0.42	-0.28	-0.05	0.27	0.03	0.02	-0.16	-0.04	-0.08	-0.02	0.17	Q2 = (0.12,0.21]
H	-0.3	-0.09	-0.9	-0.66	0.07	0.16	0.03	-0.25	-0.19	0.3	1	1	-0.21	0.16	-0.79	-0.06	-0.03	0.36	0.13	-0.37	-0.01	0.18	0.25	-0.03	0.34	Q3 = (0.21,...]
H_mais_Al	-0.28	-0.07	-0.9	-0.64	0.06	0.15	0.04	-0.26	-0.2	0.31	1	1	-0.22	0.16	-0.79	-0.05	-0.05	0.37	0.11	-0.4	-0.03	0.16	0.24	-0.03	0.36	Q3 = (0.21,...]
SB	0.44	0.36	0.37	0.26	0.72	0.59	0.35	0.93	0.83	-0.32	-0.21	-0.22	1	0.93	0.75	-0.06	0.32	0.23	0.31	-0.19	0.65	0.23	0.38	0.6	0.02	
CTC	0.34	0.33	0.03	0.02	0.75	0.65	0.37	0.84	0.76	-0.21	0.16	0.16	0.93	1	0.46	-0.08	0.31	0.37	0.36	-0.34	0.65	0.29	0.48	0.6	0.16	Q2 = (0.12,0.21]
V_perc	0.49	0.31	0.84	0.6	0.39	0.23	0.16	0.75	0.63	-0.42	-0.79	-0.79	0.75	0.46	1	0.01	0.25	-0.14	0.12	0.11	0.4	0.04	0.08	0.37	-0.21	Q3 = (0.21,...]
Ca_div_Mg	0.41	0.38	0.18	0.2	-0.08	-0.1	-0.03	0.09	-0.5	-0.28	-0.06	-0.05	-0.06	-0.08	0.01	1	-0.33	0.06	-0.01	-0.11	-0.1	0	-0.08	-0.29	-0.05	
Mg_div_K	-0.18	-0.1	0.11	-0.07	0.54	0.33	-0.48	0.47	0.59	-0.05	-0.03	-0.05	0.32	0.31	0.25	-0.33	1	0	0.11	-0.08	0.34	0.33	0.45	0.24	0.12	
S	-0.17	-0.01	-0.34	-0.47	0.4	0.45	0.24	0.16	0.12	0.27	0.36	0.37	0.23	0.37	-0.14	0.06	0	1	0.42	-0.2	0.12	0.14	0.25	0.17	0.24	Q3 = (0.21,...]
Na	0.07	0.02	0.01	-0.5	0.3	0.33	0.15	0.26	0.21	0.03	0.13	0.11	0.31	0.36	0.12	-0.01	0.11	0.42	1	0.08	0.23	0.11	0.2	0.02	-0.08	
Fe	-0.21	-0.32	0.33	0.04	-0.26	-0.08	-0.01	-0.2	-0.16	0.02	-0.37	-0.4	-0.19	-0.34	0.11	-0.11	-0.08	-0.2	0.08	1	-0.3	-0.17	-0.38	-0.31	-0.37	Q3 = (0.21,...]
Mn	0.36	0.17	0.12	0.11	0.7	0.55	0.01	0.67	0.65	-0.16	-0.01	-0.03	0.65	0.65	0.4	-0.1	0.34	0.12	0.23	-0.3	1	0.55	0.44	0.69	0.15	Q2 = (0.12,0.21]
Cu	0.02	-0.01	-0.06	-0.14	0.54	0.44	-0.08	0.28	0.21	-0.04	0.18	0.16	0.23	0.29	0.04	0	0.33	0.14	0.11	-0.17	0.55	1	0.33	0.48	0.09	
Zn	0.08	0.1	-0.2	-0.3	0.49	0.28	-0.11	0.43	0.43	-0.08	0.25	0.24	0.38	0.48	0.08	-0.08	0.45	0.25	0.2	-0.38	0.44	0.33	1	0.07	0.12	
B	0.24	0.18	0.05	0.16	0.55	0.49	0.2	0.52	0.63	-0.02	-0.03	-0.03	0.6	0.6	0.37	-0.29	0.24	0.17	0.02	-0.31	0.69	0.48	0.07	1	0.13	
kgPlanta2022	0.02	0.03	-0.39	-0.1	0.05	0.17	-0.04	0.03	0.04	0.17	0.34	0.36	0.02	0.16	-0.21	-0.05	0.12	0.24	-0.08	-0.37	0.15	0.09	0.12	0.13	1	
Corr Q3				Corr Q2			Intersecção Q3			Intersecção Q2																
	pH_CaCl2			P			pH_H2O			MO																
	H			Al			Ca			SB																
	H+Al			CTC			Mg			Cu																
	V_perc			Mn			SB			B																
	S																								
	Fe																								

Figura 4. Escolha de atributos pelos valores das correlações.

Tabela 2. Centróides do mapa SOM 2x2 com dados de análise de solo 2022 da Fazenda Santa Cruz.

vetor	pH_ $CaCl_2$	pH_H2O	MO	P	K	Ca	Mg	Al	H	SB	CTC
V3	5.917102	6.523816	31.19380	11.83898	3.359630	41.33844	11.46232	0.0600000	18.73877	56.37081	75.49761
V4	5.594760	6.437870	35.11316	23.89344	5.771973	47.61112	14.34769	0.1217703	22.94490	67.94964	91.32385
V1	5.981095	6.789427	48.40414	68.78037	3.642728	66.02620	17.38769	0.0600000	20.26946	87.40162	107.93043
V2	5.508644	6.416208	34.05379	25.85208	3.462765	46.48131	12.69911	0.1309103	23.34675	62.81337	86.99131
vetor	**V%**	**S**	**Na**	**Fe**	**Mn**	**Cu**	**Zn**	**B**	**ce1**	**ce2**	**KgPlanta**
V3	74.03795	12.12799	6.174242	72.85417	11.02192	2.280390	5.900547	0.4447584	4.694653	1.698897	0.9649833
V4	73.97491	15.22624	6.143110	44.22948	16.98452	2.962431	9.528926	0.7109124	4.536334	1.689220	2.4461098
V1	81.00314	16.42766	9.159311	64.73716	25.07215	8.642732	9.849113	0.8795227	5.115285	1.849471	2.4492041
V2	71.81360	15.73592	4.946090	44.20534	14.83857	2.930605	9.157540	0.5831042	4.872501	1.810130	4.4061531

Os atributos inicialmente escolhidos a partir da Tabela 2 foram P, H, SB, CTC, S, Na, Fe, Mn, Cu, Zn, B, ce1, ce2. O erro médio (valor absoluto) em Kg usando o mapeamento com os 21 atributos para a previsão de produção foi 0.4025481. Como segundo passo, ordenou-se as 40 amostras por ordem crescente de produção e para cada atributo foi contabilizada a porcentagem de ocorrências em que o atributo aumentava de valor na amostra seguinte. Por esta análise, foram selecionados os seguintes atributos:

“Al” 5.1% (inversamente proporcional)
“H” 53.8%
“CTC” 53.8%
“Fe” 56.4%

A partir destas duas seleções iniciais de atributos, testou-se variações de combinações destes atributos em mapas SOM de grade 2x2 e 3x3, com os erros médios mostrados na Tabela 3.

Tabela 3. Erros médios de mapas SOM por combinação de atributos.

Atributos	Erro médio (dimensão 2x2)	Erro médio (dimensão 3x3)
Al, H, CTC, Fe	0.4769916	0.3474321
H, CTC, Fe	0.4769916	0.3474321
H, CTC, Fe, Cu, Zn	0.4769916	0.313936
H, CTC, Cu, Zn	0.4242069	0.2806019
CTC, Cu, Zn	0.4067554	0.3068478
Fe, Cu, Zn	0.5909683	0.3706689

Considerando a grande importância dos atributos teor de cobre (Cu) e teor de zinco (Zn), gerando baixos níveis de erro quando são incluídos nas combinações de atributos, escolheu-se as combinações de atributos (CTC, Cu, Zn) e (Fe, Cu, Zn). Estas combinações permitem um mapeamento com baixo erro da produção com apenas três elementos da análise do solo. As Figura 5 e 6 mostram o mapa SOM 2x2 gerado para as combinações (CTC, Cu, Zn) e (Fe, Cu, Zn), respectivamente.

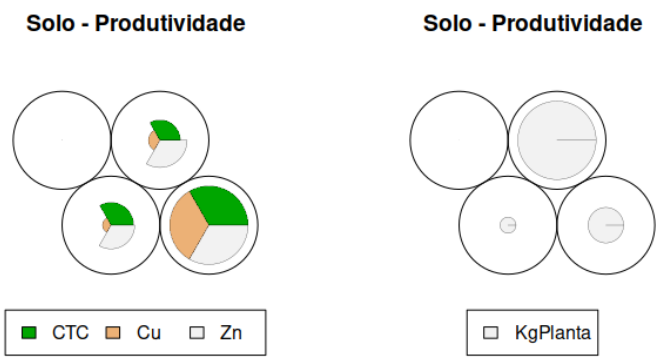

Figura 5. Mapa SOM para CTC, Cu e Zn.

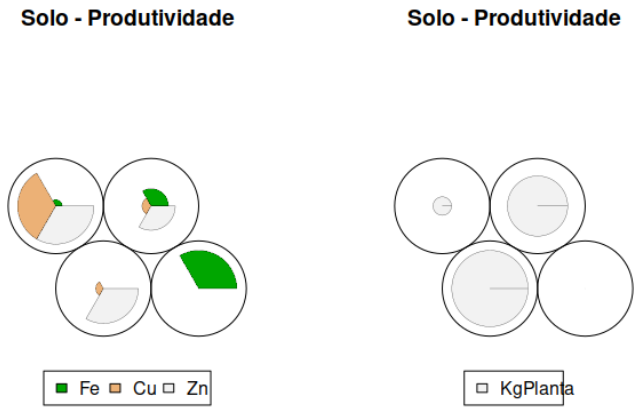

Figura 6. Mapa SOM para Fe, Cu e Zn.

Com o objetivo de se comparar os valores de produção gerados pelo mapa SOM com a produção real observada, foi gerado um mapa das amostras com produção real por tercis e dois mapas geográficos das amostras agrupadas pelo mapa SOM para as combinações CTC, Cu, Zn (Figura 5) e Fe, Cu, Zn (Figura 7).

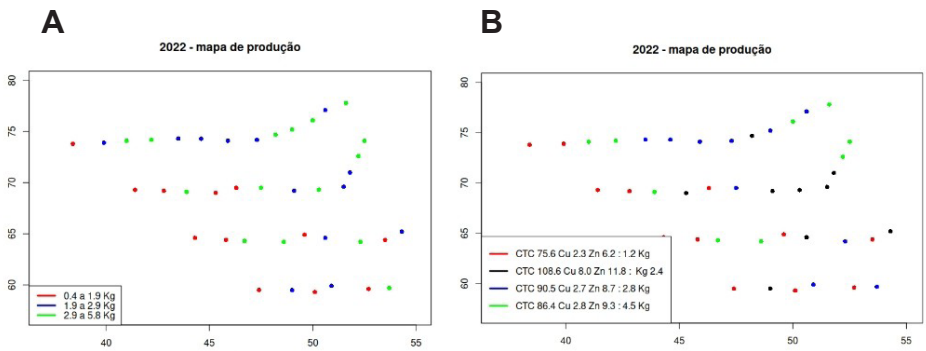

Figura 7. Mapa: (A) de produção observada; (B) SOM baseado em CTC, Cu, Z.

Comparando os pontos nos mapas da Figura 7, atributos CTC, Cu, Zn, há:

- 9 amostras com produção real entre 0.4 a 1.9 kg que aparecem na média 1.2 Kg no SOM
- 8 amostras com produção real entre 2.9 a 5.8 kg que aparecem na média 4.5 Kg no SOM
- 5 amostras com produção real entre 1.9 a 2.9 kg que aparecem na média 2.8 Kg no SOM
- 4 amostras com produção real entre 1.9 a 2.9 kg que aparecem na média 2.4 Kg no SOM
- 3 amostras com produção real entre 2.9 a 5.8 kg que aparecem na média 2.8 Kg no SOM
- 6 amostras com produção real entre 0.4 e 1.9 Kg que aparecem na média 2.8 Kg no SOM

Assim, há 35 pontos com boa ou razoável correlação entre a produção real e a produção mapeada pelos atributos CTC, Cu, Zn no mapa SOM. Os três atributos selecionados mostraram-se bastante relevantes para identificação da produção da fazenda para o ano de 2022.

Comparando os pontos nos mapas da Figura 8, atributos Fe, Cu, Zn, há:

- 5 amostras com produção real entre 0.4 a 1.9 kg que aparecem na média 1.2 Kg no SOM
- 6 amostras com produção real entre 2.9 a 5.8 kg que aparecem na média 4.7 Kg no SOM
- 3 amostras com produção real entre 1.9 a 2.9 kg que aparecem na média 3.5 Kg no SOM
- 9 amostras com produção real entre 1.9 a 2.9 kg que aparecem na média 2.1 Kg no SOM
- 4 amostras com produção real entre 2.9 a 5.8 kg que aparecem na média 3.5 Kg no SOM
- 6 amostras com produção real entre 0.4 e 1.9 Kg que aparecem na média 2.1 Kg no SOM

Assim, há 33 pontos com correlação boa ou razoável entre a produção real e a produção mapeada pelos atributos Fe, Cu, Zn no mapa SOM. Os três

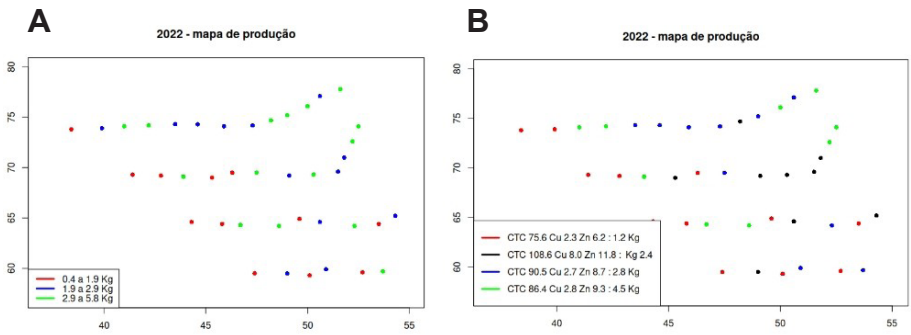

Figura 8. Mapa: (A) de produção observada; (B) SOM baseado em Fe, Cu, Zn.

atributos selecionados mostraram-se bastante relevantes para identificação da produção da fazenda para o ano de 2022.

A comparação visual entre a produção real e a produção prevista pelo mapa SOM 3x3 a partir dos atributos CTC, Cu, Zn e Fe, Cu, Zn é mostrada na Figura 9.

Desta forma, com o método SOM foram escolhidos quatro (4) conjuntos de atributos, que foram denominados:

- SOM1 = {CTC, Cu, Zn}
- SOM2 = {Fe, Cu, Zn}
- SOM3 = {CTC, Fe, Cu, Zn}
- SOM4 = {pH_$CaCl_2$, S, V_perc, Fe, Cu, Zn}

Seleção de atributos com os métodos do Weka

Essas seleções foram mais facilmente obtidas, apenas utilizando-se o conjunto de dados originais no software Weka, assim temos:

- Árvore de Decisão tanto utilizando o método de busca Best First ou Greedy, obteve o conjunto AD(BF/Greedy) = {ph_$CaCl_2$, Fe, Cu}
- O método M5 Rules obteve o conjunto M5Rules = {H, MO, P, Fe, Mn}
- O método CFS foi utilizado com dois algoritmos de busca e obteve dois diferentes conjuntos de atributos:

- CFS(BF) = {ph_$CaCl_2$, Al, H, V%, Fe, S, B}
- CFS(GSW) = {ph_$CaCl_2$, Fe, Al, H+Al, Mg/K, Na}
- O método Relief padrão obteve o conjunto Relief = {ph_H_2O, S, Fe, Al, V%, ph_$CaCl_2$, H}}

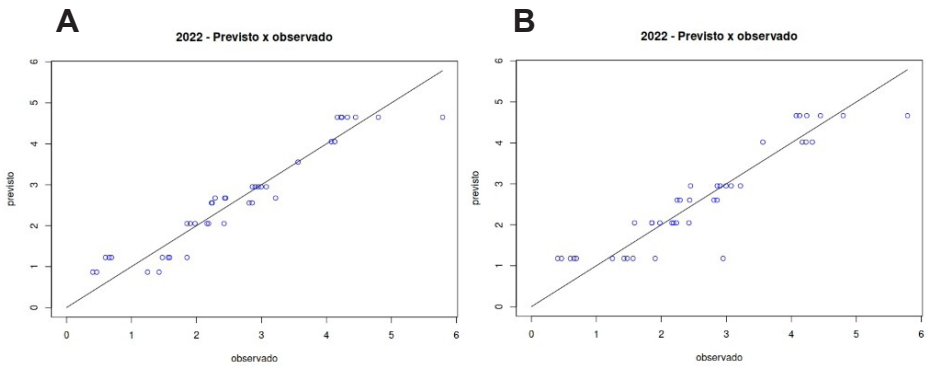

Figura 9. Comparação entre produção observada e prevista a partir de mapa SOM 3X3: (A) CTC, Cu, Zn; (B) Fe, Cu, Zn.

Seleção dos atributos

Na ferramenta Weka de aprendizado de máquina foram utilizados diferentes conjuntos de dados da seguinte forma:

- os novos dados, a partir da transformação com PCA, que trataremos como PCA;
- o conjunto completo de dados originais, com todos os atributos, que trataremos como Todos; e
- para cada conjunto de atributos selecionados, foram apresentados os dados apenas com os atributos selecionados para cada classificador utilizado, cujos nomes de cada qual seguem a mesma padronização dos nomes dos conjuntos de atributos.

Os resultados para cada conjunto de dados e classificador utilizado são apresentados na Tabela 4. E as médias dos erros quadráticos consideradas entre os classificadores, para cada conjunto de atributos, são apresentadas na Tabela 5.

Tabela 4. Valores dos Erros Quadráticos Médios obtidos.

	LinReg(NSel)		RegByDisc.J48		Gaussiana		SMO(RBF)		MLP	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
PCA	1.3092	1.5605	1.3633	1.6575	1.0814	1.2973	1.0915	1.3368	2.0238	2.6534
Todos	1.4437	1.6968	1.3332	1.5907	1.2629	1.5327	1.0723	1.3278	1.7488	2.4471
Corr1	1.0935	1.2956	1.1285	1.3765	1.0481	1.2708	1.0514	1.2988	1.2659	1.5291
Corr2	1.1966	1.4106	1.2695	1.5681	1.0843	1.3094	1.0287	1.29	1.9184	2.7148
Corr3	1.2307	1.4351	1.2061	1.4642	1.1157	1.3461	1.0256	1.29	1.8586	2.5052
Corr4	1.4855	1.7894	1.0681	1.3898	1.1432	1.355	1.0307	1.2959	1.627	2.3364
CFS(BF)	1.1923	1.3985	1.0038	1.2106	1.0648	1.2839	1.0334	1.2982	1.4251	1.8508
CFS (GreedSW)	1.1646	1.3943	1.2719	1.6058	1.0627	1.297	1.0447	1.3005	1.1498	1.4892
Relief	1.169	1.3567	1.1773	1.3676	1.0849	1.3064	1.0465	1.3031	1.5984	2.0335
SOM1	1.1945	1.4463	1.2286	1.4867	1.0998	1.3229	1.1029	1.359	1.3877	1.888
SOM2	1.1015	1.3139	1.0536	1.3728	1.0786	1.2903	1.1102	1.3533	1.1388	1.3772
SOM3	1.124	1.3369	1.2472	1.5064	1.0721	1.2869	1.1091	1.3585	1.2392	1.4948
SOM4	1.0922	1.3101	1.2571	1.4897	1.062	1.2802	1.0788	1.3199	1.7115	2.5918
AD(BF/Gred)	1.0897	1.294	1.1665	1.3943	1.0678	1.3199	1.0859	1.328	1.197	1.4677
M5 Rules	1.1276	1.3113	1.1333	1.3804	1.0421	1.2461	1.0821	1.332	1.4437	1.6968

Tabela 5. Classificação ascendente dos MAEs e RMSEs para cada conjunto de atributos.

Média RMSE	Média MAE	Conjunto de Atributos
1,34	1,1	SOM2
1,35	1,12	Corr1
1,36	1,12	AD(BF/Gred)
1,41	1,14	CFS(GreedSW)
1,41	1,14	CFS(BF)
1,4	1,16	SOM3
1,4	1,17	M5 Rules
1,5	1,2	SOM1
1,5	1,22	Relief
1,6	1,24	SOM4
1,6	1,27	Corr4
1,6	1,29	Corr3
1,7	1,3	Corr2
1,72	1,37	Todos
1,7	1,37	PCA

Nota-se que, com apenas três atributos (Fe, Cu, Zn) indicados no conjunto de atributos SOM2 observa-se um erro médio de 1.1 Kg, enquanto foram necessários seis atributos do CORR2 (ph_$ClCa_2$, H, H+Al, V%, S, Fe) para obter um erro médio de 1,12 Kg, embora também o AD (BF/Greedy) tenha utilizado apenas três atributos (ph_$CaCl_2$, Fe, Cu). Foi considerado que os três melhores métodos foram o SOM2, CORR2 e AD (BF/Greedy); seguidos dos CFS(BF)={ph_$CaCl_2$, Al, H, V%, Fe, S, B}, CFS(GSW)={ph _$CaCl_2$, Fe, Al, H+Al, Mg_div_K, Na}, SOM3={CTC, Fe, Cu, Zn) e M5={H, MO, P, Fe, Mn}.

Deve-se notar que os atributos mais frequentemente selecionados nesses métodos são: Fe, ph_$CaCl_2$, Cu, H, Zn, H+Al, V%, S e Al.

Mapas interpolados

A seleção das variáveis ph em $CaCl_2$, H, H+Al, V%, S, Fe, B, Cu, Zn pelo método de melhor resposta quanto à correlação com a produção, permitiu realizar a análise geoestatística direcionada apenas a essas variáveis de melhor correlação, onde a interpolação e produção de mapas de variabilidade espacial foi realizada produzindo os mapas conforme a Figura 10.

Conclusões

O uso de mapas SOM mostrou-se uma ferramenta eficiente para selecionar atributos para mapear a produção do café a partir da análise de solo, bem como a análise de correlação simples e aquelas implementadas pelo método CFS do Weka, e, também, os métodos baseados em ganho de informação – árvore de decisão e M5 Rules. Uma dificuldade encontrada no emprego dos agrupamentos SOM e da análise de correlação simples para a seleção de atributos foi a análise manual/visual realizada pelos cientistas de dados, uma vez que os atributos selecionados automaticamente por outros métodos implementados na ferramenta também obtiveram bons resultados, especialmente os que utilizam correlação e os que utilizam ganho de informação. A despeito disso, os conjuntos de atributos selecionados pelo SOM ainda obtiveram os menores erros, provavelmente porque os agrupamentos trabalham com combinações de faixas de valores e por isso tendem a ter maior tolerância a ruídos.

Os macronutrientes ph_$CaCl_2$, H, H+Al, V% e os micronutrientes S, Fe, B, Cu, Zn foram selecionados apresentando a melhor correlação com a produção, sendo utilizados na análise geoestatística e mapeamento da variabilidade espacial.

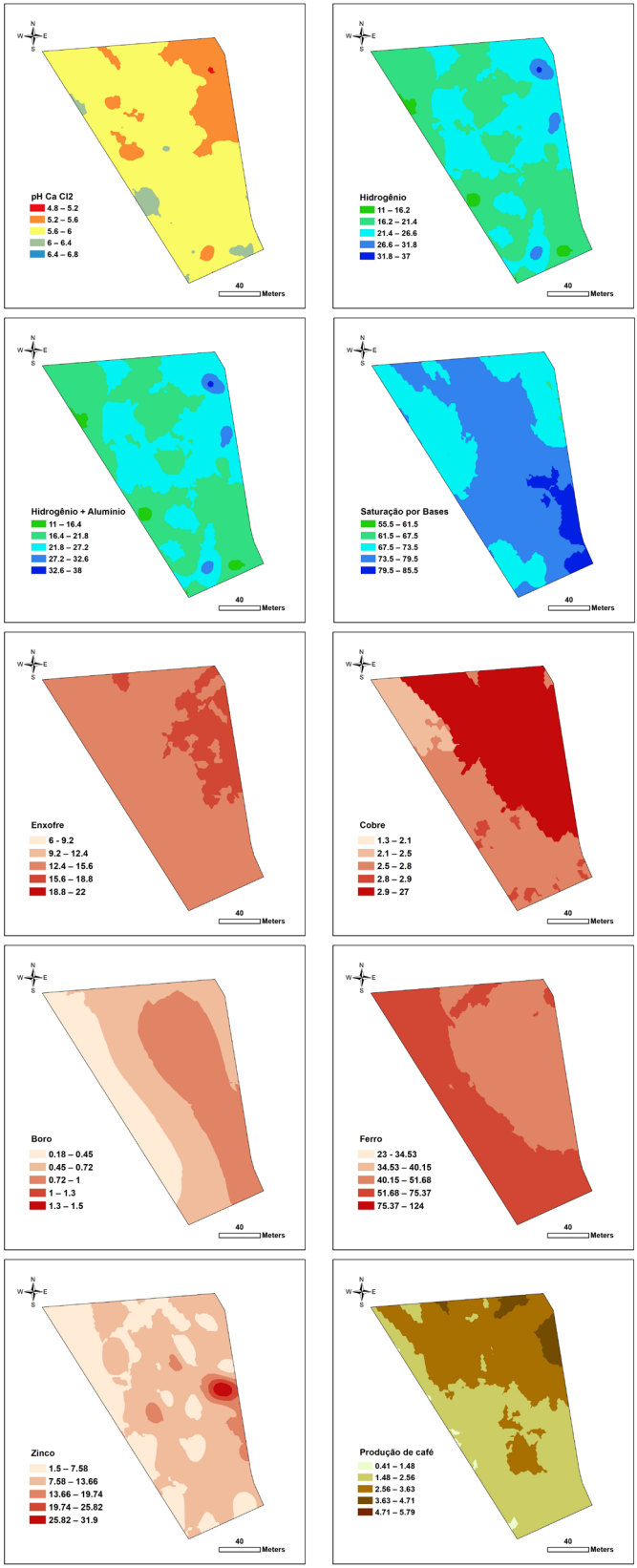

Figura 10. Mapas para as variáveis de solo (ph_$CaCl_2$, H, H+Al, V%, S, Fe, B, Cu, Zn) que apresentaram melhor correlação com a produção do café.

Referências

ACOMPANHAMENTO DA SAFRA BRASILEIRA [DE] CAFÉ: safra 2023: segundo levantamento, v. 10, n. 2, p. 1-44, maio 2023. Disponível em http://www.consorciopesquisacafe.com.br/images/stories/noticias/2021/2023/Maio/2_levantamento_safra_conab.pdf. Acesso em 30 nov. 2023.

HALL, M. A. **Correlation-based feature subset selection for machine learning**. 1999. 178 p. Thesis (Degree of Doctor of Philosophy) – Department of Computer Science, The University of Waikato, Hamilton, New Zealand.

HOLMES, G.; HALL, M.; FRANK, E. Generating rule sets from model trees. In: AUSTRALIAN JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 12., 1999, Sydney. **Advanced topics in artificial intelligence**: proceedings. Berlin: Springer, 1999. p. 1-12. Editor: Norman Foo. (Lecture notes in artificial intelligence, 1747).

KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: INTERNATIONAL WORKSHOP ON MACHINE LEARNING, 9., 1992, Aberdeen. **Machine learning**: proceedings. San Mateo: Morgan Kaufmann Publishers, 1992. p. 249-256. Editors: Derek Sleeman, Peter Edwards. ML 92. DOI: 10.1016/B978-1-55860-247-2.50037-1.

KONONENKO, I. Estimating attributes: analysis and extensions of RELIEF. In: EUROPEAN CONFERENCE ON MACHINE LEARNING, 1994, Catania. **Machine learning**: ECML-94: proceedings. Berlin: Springer, 1994. p. 171-182. Editors: Francesco Bergadano, Luc De Raedt.

RABELLO, L. M.; INAMASU, R. Y.; BERNARDI, A. C. de C.; NAIME, J. de M.; MOLIN, J. P. Mapeamento da condutividade elétrica do solo - sistema protótipo. In: INAMASU, R. Y.; NAIME, J. de M.; RESENDE, A. V. de; BASSOI, L. H.; BERNARDI, A. de C. (ed.). **Agricultura de precisão**: um novo olhar. São Carlos, SP: Embrapa Instrumentação, 2011. p. 41-45. Disponível em: https://www.alice.cnptia.embrapa.br/alice/handle/doc/1017337. Acesso em: 30 nov. 2023.

ROBNIK-SIKONJA, M.; KONONENKO, I. An adaptation of Relief for attribute estimation in regression. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 14., 1997, Nashville. **Proceedings** […]. San Francisco: Morgan Kaufmann Publishers, 1997. p. 296-304. Editor: Douglas H. Fisher.

SALZBERG, S. L. Book review: C4.5: programs for machine learning. Machine Learning, v. 16, p. 235-240, 1994. Resenha de: QUINLAN, J. R. **C4.5**: programs for machine learning. San Mateo: Morgan Kaufmann Publishers, 1993. DOI: 10.1007/BF00993309.

SILVA, F. M. da; ALVES, M. de C. **Cafeicultura de precisão**. Lavras: Editora Ufla, 2013. 227 p.

VIEIRA, S. R. Uso de geoestatística em estudos de variabilidade espacial de propriedades do solo. In: NOVAIS, R. F.; ALVAREZ V, V. H.; SCHAEFER, C. E. G. R. (ed.). **Tópicos em ciência do solo**. Viçosa: Sociedade Brasileira de Ciência do Solo, 2000. v. 1, p. 1-54.

MINISTÉRIO DA AGRICULTURA E PECUÁRIA

GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO

CGPE 018442