

DISCRIMINAÇÃO DE CELADONS DE LONGQUAN E JINGDEZHEN POR PLS-DA APLICADA À COMPOSIÇÃO QUÍMICA POR EDXRF

Juliano Ribeiro da Silva¹, Victor Hugo Cordeiro Vinhaes da Costa¹, Alexandre Xavier da Costa Carvalho¹, Gabriela de Sousa Ribeiro¹, Beatriz Kraucs Toledo de Oliveira¹, Igor Campos de Almeida Lima¹

¹Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil
(julianoribarqueo@gmail.com)

Resumo: A procedência de celadons de Longquan e Jingdezhen foi avaliada por PLS-DA aplicada a dados de EDXRF, com corpo cerâmico e esmalte analisados separadamente. Com três variáveis latentes, a sensibilidade foi 1,000 em ambas as frações; F_1 /MCC foram 0,857/0,810 no corpo cerâmico e 1,000/1,000 no esmalte. O teste de permutação indicou que a discriminação não se deveu ao acaso, sustentando a pertinência da abordagem nas condições avaliadas.

Palavras-chave: Arqueometria. Celadon. EDXRF. PLS-DA. Teste de permutação.

INTRODUÇÃO

A arqueometria integra abordagens da química, da física, da geologia, da estatística e da arqueologia para produzir inferências quantitativas sobre matérias-primas, tecnologia de produção, circulação e proveniência de vestígios materiais (Pollard, Heron e Gillard, 2008). Em cerâmicas arqueológicas, a composição química do corpo e do esmalte pode refletir diferenças nas argilas empregadas, nos fundentes, nas receitas de fabricação e nas condições de queima. A interpretação desses dados, contudo, depende da amostragem, da calibração instrumental, da heterogeneidade do material e de possíveis alterações pós-deposicionais (Tite, 2008; Pashkova et al., 2023).

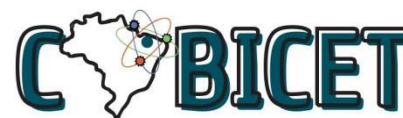
A espectrometria de fluorescência de raios X por dispersão de energia (EDXRF) é amplamente utilizada na caracterização elementar de materiais arqueológicos, pois requer preparo reduzido da amostra e, em muitos casos, permite análises não destrutivas ou minimamente invasivas (Shackley, 2011; Potts e West, 2008). Em cerâmicas, entretanto, sua aplicação exige atenção a efeitos de matriz, geometria de medição, rugosidade superficial, espessura da amostra e limites de detecção, fatores que podem afetar as concentrações estimadas e, consequentemente, a interpretação quimiométrica dos dados (Forster et al., 2011; Calparsoro et al., 2019; Pashkova et al., 2023).

No caso dos celadons chineses, Longquan e Jingdezhen constituem centros cerâmicos historicamente relevantes. He, Zhang e Zhang (2016) analisaram celadons desses centros por EDXRF e observaram diferenças composicionais mais

evidentes nos corpos cerâmicos, ao passo que os esmaltes apresentaram maior proximidade química e tecnológica. No mesmo estudo, os autores empregaram random forests para avaliar a importância discriminante dos elementos químicos. Trabalhos sobre cerâmicas chinesas e celadons também indicam que dados de EDXRF, associados a métodos multivariados, podem contribuir para a discussão de composição, período, centro produtor e tecnologia de fabricação (Wu et al., 2000; Li et al., 2011; Wu et al., 2013).

A literatura recente tem ampliado o uso de métodos quimiométricos e de aprendizado de máquina no estudo de cerâmicas arqueológicas. Essa tendência inclui aplicações de pXRF e EDXRF na classificação e caracterização composicional de cerâmicas (Forster et al., 2011; Calparsoro et al., 2019), o uso de fluorescência de raios X por reflexão total (TXRF) associada a PCA, k-means e SVM (Maltsev et al., 2023), bem como discussões sobre incerteza em análises elementares de materiais arqueológicos (Pashkova et al., 2023). Também se observam aplicações de métodos supervisionados para predição de proveniência ou classificação de amostras descritas por concentrações elementares (Anglisano et al., 2022; Ruschioni et al., 2023).

A análise discriminante por mínimos quadrados parciais (PLS-DA) é uma alternativa pertinente para esse tipo de problema, pois combina redução de dimensionalidade e modelagem supervisionada. O método é particularmente útil em situações nas quais os preditores químicos são correlacionados e o número de observações é limitado, como ocorre com frequência em estudos arqueométricos (Stähle e Wold, 1987; Barker e Rayens, 2003; Westerhuis et



al., 2008; Ruiz-Perez et al., 2020). Na arqueometria e na ciência do patrimônio, abordagens supervisionadas têm sido empregadas para atribuição de proveniência e classificação de amostras com base em concentrações elementares (Anglisano et al., 2022; Ruschioni et al., 2023), inclusive em estudos que combinam espectrometria de raios X e quimiometria (Maltsev et al., 2023).

O objetivo deste trabalho é avaliar a capacidade da PLS-DA, aplicada à composição química obtida por EDXRF, de discriminar a procedência de celadons entre Longquan e Jingdezhen. Para preservar a interpretação arqueométrica dos resultados, corpo cerâmico e esmalte foram analisados separadamente, uma vez que correspondem a frações materiais distintas quanto à composição, à função tecnológica e à relação com matérias-primas e processamento. A escolha da PLS-DA justifica-se por sua aplicação em problemas quimiométricos com preditores correlacionados e número limitado de amostras (Stähle e Wold, 1987; Barker e Rayens, 2003; Westerhuis et al., 2008; Ruiz-Perez et al., 2020). Além disso, o estudo original já havia explorado a discriminação dessas cerâmicas por meio de um método supervisionado, random forests, voltado à identificação de elementos químicos relevantes para a separação entre grupos.

MATERIAL E MÉTODOS

A base de dados, originalmente publicada por He, Zhang e Zhang (2015) como preprint e posteriormente revisada em 2016, está disponibilizada no UCI Machine Learning Repository sob o nome Chemical Composition of Ceramic Samples (UCI Machine Learning Repository, 2019), contém 88 registros de composição química obtidos por EDXRF. Os registros são identificados quanto à fração material analisada, corpo cerâmico ou esmalte, correspondendo a 44 medições de corpo cerâmico e 44 de esmalte, sem valores faltantes nas variáveis químicas consideradas. Neste trabalho, corpo cerâmico e esmalte não foram tratados como classes de discriminação, mas como frações materiais analisadas separadamente. O desfecho supervisionado foi a procedência, definida pelas classes Longquan e Jingdezhen. Foram mantidos os 17 preditores químicos listados na Tabela 1.

A variável de classe foi construída a partir do prefixo do identificador de cada amostra. Os registros iniciados por FLQ foram associados a Jingdezhen, enquanto aqueles iniciados por DY-BS, DY-NS, DY-Y e DY-M foram associados a Longquan. O grupo DY-QC foi excluído por não integrar o conjunto principal de 40 fragmentos descrito no estudo original (He, Zhang e Zhang, 2015). Após esse filtro, a base utilizada na modelagem continha 80 registros,

sendo 40 referentes a corpo cerâmico e 40 a esmalte. Em cada fração material, havia 27 amostras atribuídas a Longquan e 13 a Jingdezhen.

Tabela 1. Preditores químicos extraídos da base de dados *Chemical Composition of Ceramic Samples* (2019).

Variável	Nome do preditor
Na ₂ O	Óxido de sódio
MgO	Óxido de magnésio
Al ₂ O ₃	Óxido de alumínio
SiO ₂	Óxido de silício(IV)
K ₂ O	Óxido de potássio
CaO	Óxido de cálcio
TiO ₂	Óxido de titânio(IV)
Fe ₂ O ₃	Óxido de ferro(III)
MnO	Óxido de manganês(II)
CuO	Óxido de cobre(II)
ZnO	Óxido de zinco
PbO ₂	Óxido de chumbo(IV)
Rb ₂ O	Óxido de rubídio
SrO	Óxido de estrôncio
Y ₂ O ₃	Óxido de ítrio(III)
ZrO ₂	Óxido de zircônio(IV)
P ₂ O ₅	Óxido de fósforo(V)

Para cada fração material, Jingdezhen foi definida como classe positiva no cálculo das métricas binárias, por corresponder ao grupo minoritário no conjunto analisado. A divisão entre treino e teste foi realizada de forma estratificada, com 70% das amostras destinadas ao treinamento do modelo. Assim, em cada fração, o conjunto de treino ficou composto por 29 amostras, sendo 19 de Longquan e 10 de Jingdezhen, enquanto o conjunto de teste retido ficou composto por 11 amostras, sendo 8 de Longquan e 3 de Jingdezhen.

Antes do ajuste dos modelos, os preditores químicos foram padronizados por *escore-z*, conforme a (eq. 1)

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (1)$$

onde z_{ij} é o valor padronizado da observação i na variável j ; x_{ij} é o valor original da amostra i na variável j ; s_j é o desvio-padrão amostral da variável j , e $\bar{x}_j$ é a média da variável j . Na rotina computacional, essa padronização foi realizada por centralização e



escalonamento dentro do procedimento de ajuste, sendo estimada a partir dos dados de treino em cada reamostragem e aplicada aos dados correspondentes de validação ou teste.

Variáveis com variância nula ou quase nula foram avaliadas por meio da função `nearZeroVar()`, do pacote `caret` (Kuhn, 2008), e removidas quando aplicável. Essa etapa foi realizada antes do ajuste dos modelos, de modo a evitar a inclusão de preditores sem variabilidade informativa ou com baixa contribuição para a discriminação.

Como etapa exploratória, foi realizada uma análise de componentes principais (PCA) sobre a matriz de preditores químicos padronizados. A PCA cria combinações lineares ortogonais das variáveis originais, ordenadas segundo a variância explicada, permitindo investigar padrões de agrupamento, gradientes composicionais e variáveis associadas aos principais eixos de variação (Jolliffe e Cadima, 2016). Computacionalmente, a PCA pode ser obtida por decomposição em valores singulares (Wehrens, 2011). Seja $\mathbf{Z}$ a matriz padronizada, com n amostras e p variáveis. Sua decomposição pode ser escrita como (eq. 2):

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^T \quad (2)$$

em que $\mathbf{U}$ é uma matriz ortonormal de dimensão $n \times a$, formada pelos vetores singulares à esquerda; $\mathbf{D}$ é uma matriz diagonal $a \times a$, cujos elementos diagonais são os valores singulares; e $\mathbf{V}$ é uma matriz $p \times a$, formada pelos vetores singulares à direita. A decomposição também pode ser expressa como (eq. 3):

$$\mathbf{Z} = (\mathbf{U}\mathbf{D})\mathbf{V}^T = \mathbf{T}\mathbf{P}^T \quad (3)$$

em que $\mathbf{T} = \mathbf{U}\mathbf{D}$ é a matriz de escores, cujas linhas representam as coordenadas das amostras no espaço dos componentes principais, e $\mathbf{P} = \mathbf{V}$ é a matriz de cargas, cujas colunas indicam a contribuição das variáveis originais para cada componente. Assim, os escores descrevem a posição relativa das amostras no espaço reduzido, enquanto as cargas descrevem a orientação dos eixos em relação às variáveis originais (Brereton, 2003; Pomerantsev, 2014; Wehrens, 2011).

Os valores singulares contidos em $\mathbf{D}$ estão diretamente relacionados à variância explicada por cada componente principal. Para o i -ésimo componente, a variância associada pode ser escrita como indicado em Wehrens (2011) (eq. 4):

$$\lambda_i = \frac{d_i^2}{n-1} \quad (4)$$

onde d_i é o i -ésimo valor singular. A fração de variância explicada (Ferreira, 2015; Wehrens, 2011) pelo componente i é, então (eq. 5):

$$FV_i = \lambda_i / \sum_{j=1}^a \lambda_j \quad (5)$$

em que a representa o número de componentes principais não nulos considerados na decomposição, ou seja, o número de valores singulares presentes na matriz diagonal $\mathbf{D}$. Neste trabalho, a PCA foi utilizada apenas como ferramenta exploratória, por meio de gráficos de variância explicada e biplots. Nos biplots, escores das amostras e vetores das variáveis foram representados conjuntamente, seguindo a lógica proposta por Gabriel (1971). As elipses por classe foram incluídas apenas como auxílio visual para descrever a dispersão dos grupos, sem finalidade inferencial.

Diferentemente da PCA, que organiza o espaço latente segundo direções de maior variância do bloco $\mathbf{X}$, a regressão por mínimos quadrados parciais (PLS) procura componentes que sejam, simultaneamente, representativos dos preditores e relacionados à resposta. Assim, a extração das variáveis latentes não é conduzida apenas pela variabilidade interna de $\mathbf{X}$, mas pela associação entre combinações lineares de $\mathbf{X}$ e $\mathbf{Y}$ (Geladi e Kowalski, 1986; Wold, Sjöström e Eriksson, 2001). Para a componente h , essa ideia pode ser expressa pela maximização da covariância entre os escores latentes dos blocos $\mathbf{X}$ e $\mathbf{Y}$ (eq. 6):

$$\max_{\mathbf{w}_h, \mathbf{c}_h} [\text{cov}(\mathbf{X}\mathbf{w}_h, \mathbf{Y}\mathbf{c}_h)]^2 \quad (6)$$

onde $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h$ e $\mathbf{u}_h = \mathbf{Y}\mathbf{c}_h$ são os escores latentes da componente h , enquanto $\mathbf{w}_h$ e $\mathbf{c}_h$ são os vetores de pesos associados, respectivamente, aos blocos $\mathbf{X}$ e $\mathbf{Y}$.

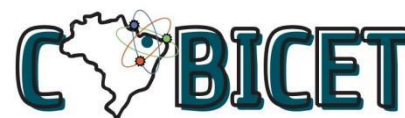
A análise discriminante por mínimos quadrados parciais (PLS-DA) pode ser entendida como a aplicação do PLS a uma resposta categórica codificada numericamente. Nesse caso, a matriz de preditores $\mathbf{X}$ é relacionada a uma matriz resposta $\mathbf{Y}$, construída a partir das classes das amostras (Ståhle e Wold, 1987; Barker e Rayens, 2003). Em uma formulação multiclasse, se há g classes, $\mathbf{Y}$ possui g colunas. Para cada amostra i , define-se $y_{ik} = 1$ quando a amostra pertence à classe k , e $y_{ik} = 0$ para as demais classes. Dessa forma, a classe de cada amostra passa a ser representada em uma estrutura numérica compatível com regressão multivariada.

Quando $\mathbf{Y}$ é representada por mais de uma coluna, a etapa de regressão da PLS-DA pode ser formulada como PLS2 (Brereton, 2003; Pomerantsev, 2014). Na notação adotada por Pomerantsev (2014), a decomposição dos blocos é escrita como (eq. 7-9):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (7)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \quad (8)$$

com



$$\mathbf{T} = \mathbf{XW}(\mathbf{P}^T \mathbf{W})^{-1} \quad (9)$$

As matrizes $\mathbf{T}$ e $\mathbf{U}$ contêm os escores dos blocos $\mathbf{X}$ e $\mathbf{Y}$, respectivamente; $\mathbf{P}$ e $\mathbf{Q}$ contêm as cargas (*loadings*) desses blocos; e $\mathbf{E}$ e $\mathbf{F}$ representam os resíduos das respectivas decomposições. A matriz $\mathbf{W}$, por sua vez, contém os pesos de projeção do bloco $\mathbf{X}$, estimados considerando a relação entre os preditores químicos e a codificação das classes em $\mathbf{Y}$ (Ferreira, 2015; Pomerantsev, 2014). Essa distinção é relevante porque os pesos de projeção participam da construção do espaço latente usado na discriminação das classes.

No conjunto de treino, a seleção do número de variáveis latentes foi realizada por validação cruzada 5-fold repetida 20 vezes. Foram avaliados modelos com 1 a 6 variáveis latentes. O critério primário de seleção foi o maior índice de Youden médio na validação cruzada repetida; em caso de empate, foram considerados, sucessivamente, sensibilidade, especificidade, coeficiente de correlação de Matthews, escore F_1 , exatidão e menor número de variáveis latentes. A padronização por escore-z foi incorporada ao procedimento de reamostragem, de modo que médias e desvios-padrão fossem estimados apenas a partir dos dados de treino correspondentes, reduzindo o risco de vazamento de informação (Kuhn, 2008; Westerhuis et al., 2008).

Os parâmetros de mérito foram calculados a partir das matrizes de confusão obtidas separadamente para corpo cerâmico e esmalte. Em ambas as frações materiais, Jingdezhen foi adotada como classe positiva e Longquan como classe negativa. Denotaram-se verdadeiros positivos por VP, verdadeiros negativos por VN, falsos positivos por FP e falsos negativos por FN. A disposição da matriz de confusão adotada é mostrada na tabela 2.

Tabela 2. Matriz de confusão.

Valor real	Valor predito	
	Positivo	Negativo
Positivo	VP	FN
Negativo	FP	VN

Fonte: Adaptado de Kuhn e Johnson, 2013.

A prevalência corresponde à fração de observações pertencentes à classe positiva no conjunto avaliado. Como Jingdezhen foi adotada como classe positiva, tem-se:

$$\text{Prev} = \frac{\text{VP} + \text{FN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}} \quad (10)$$

A sensibilidade descreve a proporção de amostras de Jingdezhen corretamente identificadas pelo modelo (Kuhn e Johnson, 2013):

$$\text{Sens} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (11)$$

A especificidade (Kuhn e Johnson, 2013) mede a proporção de amostras de Longquan corretamente identificadas:

$$\text{Esp} = \frac{\text{VN}}{\text{VN} + \text{FP}} \quad (12)$$

A precisão, ou valor preditivo positivo, descreve a confiabilidade das classificações atribuídas a Jingdezhen:

$$\text{Prec} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (13)$$

A exatidão (eq. 14) ou acurácia mede a proporção total de classificações corretas:

$$\text{Exat} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \quad (14)$$

O coeficiente kappa de Cohen corrige a concordância observada pela concordância esperada ao acaso, sendo definido a partir de p_o , a proporção observada de concordância, e p_e , a concordância esperada sob independência das margens (Cohen, 1960 *apud* Kuhn e Johnson, 2013):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (15)$$

O coeficiente de correlação de Matthews (Matthews, 1975; Powers, 2011) é uma medida global de associação entre a classe real e a classe predita, baseada nos quatro elementos da matriz de confusão. Por considerar simultaneamente verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, é adequado para avaliar o desempenho em problemas binários mesmo quando as classes não estão perfeitamente balanceadas.

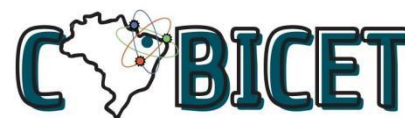
$$\text{MCC} = \frac{\text{VP} \times \text{VN} - \text{FP} \times \text{FN}}{\sqrt{(\text{VP} + \text{FP})(\text{VP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (16)$$

Os escores F foram usados para sintetizar precisão e sensibilidade. A forma geral F_β atribui peso beta à sensibilidade; F_1 corresponde a $\beta = 1$ e F_2 a $\beta = 2$, conforme a formulação de Van Rijsbergen (1979).

$$F_\beta = \frac{(1 + \beta^2) \times \text{Precisão} \times \text{Sensibilidade}}{\beta^2 \times \text{Precisão} + \text{Sensibilidade}} \quad (17)$$

O índice de Fowlkes-Mallows é a média geométrica entre precisão e sensibilidade (Fowlkes e Mallows, 1983):

$$\text{FMI} = \sqrt{\text{Precisão} \times \text{Sensibilidade}} \quad (18)$$



Assim, valores próximos de 1 indicam que o modelo apresenta, simultaneamente, elevada capacidade de identificar corretamente as amostras da classe de interesse e baixa proporção de falsas atribuições a essa classe. Por outro lado, valores reduzidos indicam perda de desempenho associada a falsos positivos, falsos negativos ou ambos.

O índice de Youden (Youden, 1950 *apud* Kuhn e Johnson, 2013) combina sensibilidade e especificidade, variando de 0, quando não há utilidade discriminante, a 1, no caso de separação perfeita.

$$J = \text{Sens} + \text{Esp} - 1 \quad (19)$$

O teste de permutação foi empregado para avaliar se o desempenho observado da PLS-DA poderia ser reproduzido quando a associação entre a composição química, representada pelos óxidos, e a proveniência das amostras fosse desfeita por aleatorização dos rótulos de classe. Em cada fração material, manteve-se fixa a matriz de preditores $\mathbf{X}$, enquanto os rótulos Longquan/Jingdezhen do conjunto de treino foram permutados aleatoriamente. Assim, a distribuição nula foi construída a partir de modelos ajustados em cenários nos quais não há relação sistemática entre os preditores químicos e a procedência atribuída às amostras.

Considerando duas classes com n_1 e n_2 amostras, o número exato de redistribuições possíveis dos rótulos, preservando os tamanhos das classes, é dado por:

$$N_{\text{perm,exato}} = \binom{n_1 + n_2}{n_1} = \frac{(n_1 + n_2)!}{n_1!n_2!} \quad (20)$$

Neste estudo, para cada fração material, o conjunto de treino continha 19 amostras de Longquan e 10 de Jingdezhen. Portanto, a enumeração completa exigiria

$$N_{\text{perm,exato}} = \binom{19+10}{10} = \frac{29!}{10!} = 20.030.010$$

redistribuições possíveis dos rótulos. Como a enumeração completa seria computacionalmente custosa, foi empregado um teste de permutação por Monte Carlo, com ($B=99.999$) permutações aleatórias dos rótulos do conjunto de treino. Em cada permutação, o modelo PLS-DA foi reajustado na configuração final selecionada, com três variáveis latentes, e avaliado no mesmo conjunto de teste retido. Para uma estatística de desempenho (T), adotou-se valor-p unilateral, pois o interesse era verificar se o desempenho observado era maior do que o esperado sob permutação aleatória dos rótulos. O valor-p de permutação foi calculado por:

$$\hat{p} = \text{valor} - p_{\text{perm}} = \frac{1 + \sum_{b=1}^B I(T_b \geq T_{\text{obs}})}{1 + B} \quad (21)$$

onde T_{obs} é a estatística observada no conjunto de teste retido; T_b é a estatística obtida na b -ésima permutação; $I(\cdot)$ é a função indicadora, que assume valor 1 quando $T_b \geq T_{\text{obs}}$ e 0 caso contrário, e B é o número de permutações válidas. A adição de uma unidade no numerador e no denominador evita valor-p igual a zero quando nenhuma estatística permutada é tão extrema quanto a observada, seguindo a formulação usual de valores-p empíricos em testes de Monte Carlo (Phipson e Smyth, 2010; Ojala e Garriga, 2010; North, Curtis e Sham, 2002). O menor valor-p possível nessa formulação é:

$$\text{valor} - p_{\text{perm,min}} = \frac{1}{1 + B} \quad (22)$$

Com $B = 99.999$, obtém-se $\text{valor} - p_{\text{perm,min}} p_{\text{min}} = 1/100.000 = 0,00001$. Como o teste foi executado por aproximação Monte Carlo, o valor-p obtido possui erro aleatório de simulação. Seja

$$R = \sum_{b=1}^B I(T_b \geq T_{\text{obs}}) \quad (23)$$

o número de permutações em que a estatística permutada foi pelo menos tão extrema quanto a estatística observada. Condicionalmente ao valor verdadeiro p da cauda nula que se deseja estimar, R pode ser tratado como uma contagem binomial, isto é, $R \sim \text{Bin}(B, p)$. Uma aproximação conveniente para o erro-padrão Monte Carlo do valor-p estimado é (North, Curtis e Sham, 2002):

$$\text{EP}_{\text{MC}}(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{B + 1}} \quad (24)$$

Essa expressão não altera o valor-p nem a decisão inferencial; apenas quantifica a precisão numérica decorrente do número finito de permutações aleatórias.

As hipóteses do teste foram formuladas de modo unilateral. A hipótese nula, H_0 , afirma que a associação entre $\mathbf{X}$ e os rótulos Longquan/Jingdezhen não produz desempenho superior ao esperado sob permutação dos rótulos. De modo equivalente, T_{obs} é compatível com a distribuição nula gerada por rótulos aleatórios. A hipótese alternativa, H_1 , afirma que a associação entre $\mathbf{X}$ e a procedência das amostras produz desempenho superior ao acaso, isto é, T_{obs} situa-se na cauda superior da distribuição nula.

RESULTADOS E DISCUSSÃO

A PCA indicou estrutura composicional associada à procedência nas duas frações materiais avaliadas. Para a fração corpo cerâmico, os dois primeiros componentes explicaram 49,2% da variância total, com 33,2% em PC1 e 16,0% em PC2. O biplot mostrou tendência de separação entre as amostras de Longquan e Jingdezhen principalmente ao longo de PC1, sugerindo diferenças composicionais no corpo cerâmico (figura 1).

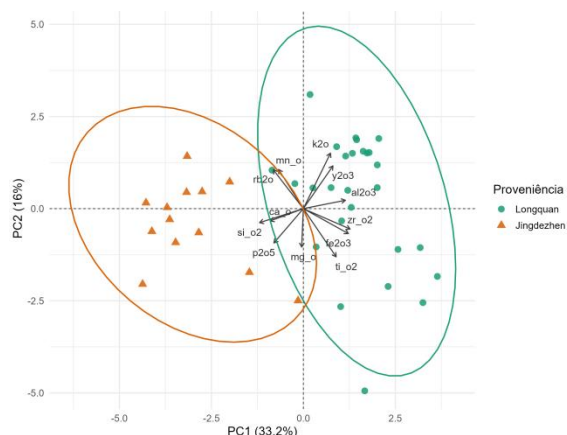


Figura 1. PCA da fração corpo cerâmico. Biplot PC1–PC2 das amostras segundo a procedência Longquan ou Jingdezhen; as elipses indicam apenas a dispersão visual dos grupos.

Para a fração esmalte, PC1 e PC2 explicaram, respectivamente, 30,7% e 17,0% da variância, totalizando 47,7%. Embora também se observe distinção entre as procedências, há maior proximidade entre os grupos no plano PC1–PC2 (figura 2).

Esses resultados devem ser interpretados como exploratórios, uma vez que a PCA não utiliza os rótulos de classe no ajuste dos componentes.

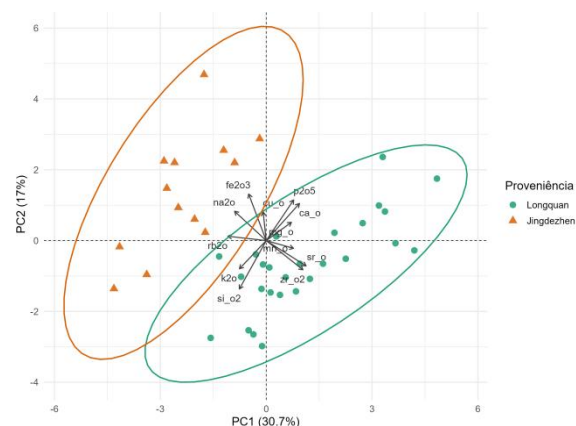


Figura 2. PCA da fração esmalte. Biplot PC1–PC2 das amostras segundo a procedência Longquan ou Jingdezhen; as elipses indicam apenas a dispersão visual dos grupos.

Na etapa supervisionada, a PLS-DA selecionou três variáveis latentes tanto para a fração corpo cerâmico quanto para a fração esmalte, considerando o maior índice médio de Youden na validação cruzada repetida. Para a fração corpo cerâmico, o modelo apresentou índice de Youden médio de 0,9825, exatidão média de 0,9933 e MCC médio de 0,9860. Para a fração esmalte, os valores médios foram ainda mais elevados: índice de Youden de 0,9975, exatidão de 0,9983 e MCC de 0,9971. Esses resultados indicam alto desempenho discriminante na etapa de reamostragem, para ambas as frações materiais.

No conjunto de teste retido, o modelo ajustado para a fração corpo cerâmico classificou corretamente as três amostras de Jingdezhen e sete das oito amostras de Longquan (tabela 3). O único erro correspondeu à classificação de uma amostra de Longquan como Jingdezhen. Considerando Jingdezhen como classe positiva, o modelo apresentou exatidão de 0,909, sensibilidade de 1,000, especificidade de 0,875, precisão de 0,750, escore F_1 de 0,857 e MCC de 0,810.

Tabela 3. Matriz de confusão no conjunto de teste retido para a fração corpo cerâmico.

Valor real	Valor predito	
	Longquan	Jingdezhen
Longquan	7	1
Jingdezhen	0	3

Embora os resultados no teste retido sejam elevados, sua interpretação deve considerar o tamanho reduzido desse conjunto, composto por apenas 11 amostras em cada fração material.

Para a fração esmalte, todas as amostras do teste foram corretamente classificadas (tabela 4), resultando em exatidão, sensibilidade, especificidade, escore F_1 , coeficiente de correlação de Matthews e coeficiente kappa iguais a 1,000.

Tabela 4. Matriz de confusão no conjunto de teste retido para a fração esmalte.

Valor real	Valor predito	
	Longquan	Jingdezhen
Longquan	8	0
Jingdezhen	0	3

O índice de Fowlkes-Mallows e o coeficiente kappa complementaram a avaliação dos modelos. Para a fração corpo cerâmico, o FMI foi 0,866, indicando bom equilíbrio entre a capacidade de reconhecer Jingdezhen e a confiabilidade das previsões atribuídas a essa procedência. Esse valor, entretanto, não foi máximo porque uma amostra de Longquan foi classificada como Jingdezhen, reduzindo a

precisão da classe positiva. O kappa foi 0,792, o que indica concordância elevada entre classes reais e previstas após correção pela concordância esperada ao acaso, mas ainda afetada pelo único falso positivo observado. Para a fração esmalte, FMI e kappa foram iguais a 1,000, coerentes com a classificação correta de todas as amostras do teste retido.

A regra bayesiana de decisão permitiu visualizar a separação entre as classes a partir dos valores previstos pelo PLS-DA. Para a fração corpo cerâmico, as densidades associadas a Longquan e Jingdezhen apresentaram separação clara, embora com maior proximidade em torno do limite de decisão (Figura 3).

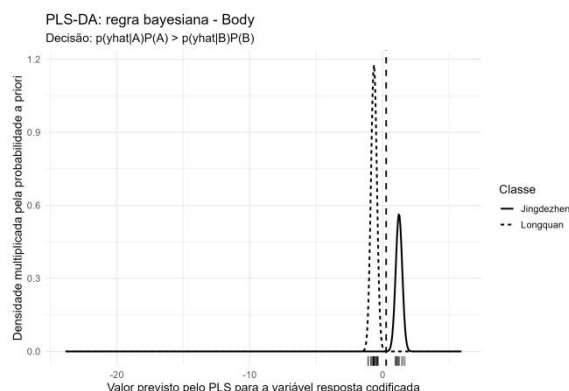


Figura 3. Regra bayesiana de decisão aplicada ao modelo PLS-DA para a fração corpo cerâmico. As curvas representam as densidades dos valores previstos pelo PLS multiplicadas pelas probabilidades *a priori* de cada classe. A linha vertical tracejada indica o limite de decisão bayesiano.

Essa configuração é compatível com a matriz de confusão do teste, na qual todas as amostras de Jingdezhen foram corretamente classificadas, mas uma amostra de Longquan foi atribuída a Jingdezhen. Na fração esmalte, as distribuições estimadas ficaram mais bem separadas em relação ao ponto de decisão (figura 4), o que é coerente com a classificação correta de todas as amostras no conjunto de teste retido.

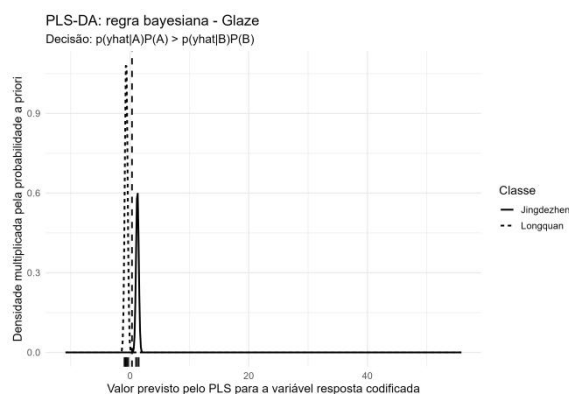


Figura 4. Regra bayesiana de decisão aplicada ao modelo PLS-DA para a fração esmalte. As curvas representam as densidades dos valores previstos pelo PLS multiplicadas pelas probabilidades *a priori* de cada classe. A linha vertical tracejada indica o limite de decisão bayesiano.

Esses resultados indicam que a regra bayesiana forneceu um critério de decisão consistente com o desempenho observado, ao mesmo tempo em que evidencia maior incerteza relativa no modelo ajustado para fração corpo cerâmico.

O teste de permutação foi empregado para verificar se o desempenho observado da PLS-DA poderia ser obtido apenas pela redistribuição aleatória dos rótulos de procedência. Nos histogramas, as barras representam os valores de escore F_1 obtidos após permutar os rótulos Longquan/Jingdezhen no conjunto de treino, enquanto a linha tracejada indica o escore F_1 observado no conjunto de teste retido.

Para a fração corpo cerâmico, o modelo apresentou escore $F_1 = 0,857$, índice de Youden = 0,875 e MCC = 0,810 no conjunto de teste retido. A Figura 5 mostra que o escore F_1 observado se localiza na cauda direita da distribuição permutada: embora algumas permutações tenham produzido desempenhos relativamente altos, a maior parte dos valores de F_1 permaneceu abaixo do valor observado. Os valores-p permutacionais associados ao escore F_1 , Youden e MCC foram, respectivamente, 0,04141, 0,02681 e 0,02711. Assim, para a fração corpo cerâmico, a evidência favorece a existência de informação discriminante na composição química, mas com menor margem em relação ao que foi observado para o esmalte.

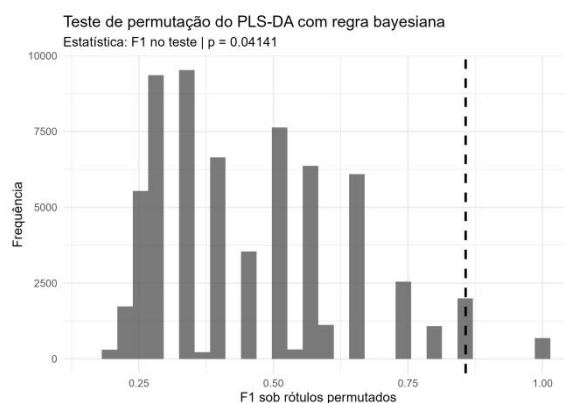


Figura 5. Teste de permutação da PLS-DA para a fração corpo cerâmico. Distribuição do F_1 sob permutação dos rótulos de procedência. A linha tracejada indica o valor observado no conjunto de teste retido.

Para a fração esmalte, o modelo apresentou escore $F_1 = 1,000$, índice de Youden = 1,000 e MCC = 1,000 no conjunto de teste retido. Na figura 6, o escore F_1

observado aparece no limite superior da escala, enquanto a distribuição obtida sob rótulos permutados concentra-se majoritariamente em valores inferiores.

Os valores-p permutacionais associados ao escore F_1 , índice de Youden e MCC foram, respectivamente, 0,006821, 0,00508 e 0,00516. Esse resultado indica que desempenhos equivalentes ao observado foram raros quando a relação entre composição química e procedência foi desfeita pela permutação dos rótulos. Portanto, a evidência de discriminação foi mais forte para a fração esmalte do que para a fração corpo cerâmico.

Resultados semelhantes foram observados para FMI e kappa. No corpo cerâmico, os valores-p permutacionais foram 0,02711 para FMI e 0,02681 para kappa, indicando que a combinação entre sensibilidade e precisão, bem como a concordância corrigida pelo acaso, dificilmente seriam obtidas apenas com rótulos de procedência embaralhados. No esmalte, os valores-p foram 0,00516 para FMI e 0,00508 para kappa, reforçando que o desempenho perfeito observado nessa fração foi raro sob permutação.

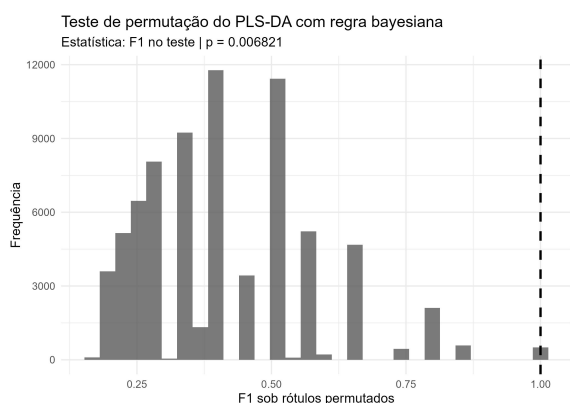


Figura 6. Teste de permutação da PLS-DA para fração esmalte. Distribuição do F_1 sob permutação dos rótulos de procedência. A linha tracejada indica o valor observado no conjunto de teste retido.

CONCLUSÃO

A PLS-DA aplicada à composição química obtida por EDXRF mostrou-se capaz de discriminar a procedência dos celadons de Longquan e Jingdezhen nas condições avaliadas. A análise separada de corpo cerâmico e esmalte preservou a interpretação arqueométrica das frações materiais e permitiu observar que ambas contêm informação química associada à procedência.

Na etapa exploratória, a PCA indicou tendência de separação entre as procedências, sobretudo ao longo do PC1, com maior nitidez visual para o corpo cerâmico. Na etapa supervisionada, os modelos PLS-DA selecionaram três variáveis latentes para ambas

as frações. No conjunto de teste retido, o modelo para a fração corpo cerâmico classificou corretamente 10 das 11 amostras, com apenas uma amostra de Longquan atribuída a Jingdezhen. Para a fração esmalte, todas as amostras do teste foram corretamente classificadas.

Os testes de permutação reforçaram que o desempenho observado não foi explicado apenas pela aleatorização dos rótulos de procedência. A evidência foi mais forte para a fração esmalte, cujos valores-p permutacionais foram menores para o escore F_1 , o índice de Youden e o MCC. Esse resultado indica que, embora o estudo original tenha destacado maior similaridade tecnológica entre esmaltes, a composição química dessa fração ainda preservou informação discriminante suficiente para separar Longquan e Jingdezhen no conjunto analisado.

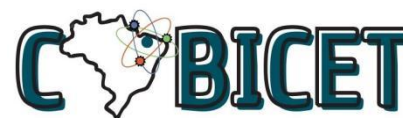
Esses achados devem ser interpretados com cautela, pois o número de amostras no teste retido foi reduzido. Ainda assim, os resultados indicam que a PLS-DA, associada à validação cruzada e ao teste de permutação, constitui uma abordagem quimiométrica pertinente para investigar procedência em dados arqueométricos de EDXRF. Estudos futuros podem ampliar a base amostral, incluir validação externa e comparar a PLS-DA com outros métodos supervisionados empregados em arqueometria.

AGRADECIMENTOS

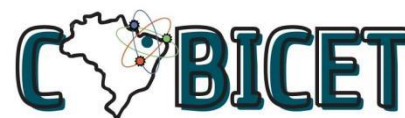
Os autores agradecem à Universidade do Estado do Rio de Janeiro pelo apoio institucional. B.K.T.O. agradece à Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) pelo apoio financeiro concedido por meio da Bolsa de Iniciação Científica (IC), processo nº E-26/200.744/2025, cuja contribuição foi fundamental para a realização e o desenvolvimento desta pesquisa.

REFERÊNCIAS

- ANGLISANO, A.; CASAS, L.; QUERALT, I.; DI FEBBO, R. Supervised Machine Learning Algorithms to Predict Provenance of Archaeological Pottery Fragments. *Sustainability*, v. 14, n. 18, 11214, 2022.
- BARKER, M.; RAYENS, W. Partial least squares for discrimination. *Journal of Chemometrics*, v. 17, n. 3, p. 166-173, 2003.
- BRERETON, R. G. *Chemometrics: Data analysis for the laboratory and chemical plant*. West Sussex: John Wiley & Sons, 2003.
- CALPARSORO, E.; MAGUREGUI, M.; MORILLAS, H.; ARANA, G.; IÑÁÑEZ, J. G. Non-destructive screening methodology based on ED-XRF for the classification of medieval and post-medieval archaeological ceramics. *Ceramics International*, v. 45, p. 10672-10683, 2019.



- CHEMICAL COMPOSITION OF CERAMIC SAMPLES [Dataset]. UCI Machine Learning Repository, 2019. DOI: 10.24432/C54P5X. Disponível em: <https://archive.ics.uci.edu/dataset/583/chemical+composition+of+ceramic+samples>.
- FERREIRA, M. M. C. Quimiometria: Conceitos, métodos e aplicações. Campinas: Editora da Unicamp, 2015.
- FORSTER, N.; GRAVE, P.; VICKERY, N.; KEALHOFER, L. Non-destructive analysis using PXRF: methodology and application to archaeological ceramics. *X-Ray Spectrometry*, v. 40, n. 5, p. 389-398, 2011.
- FOWLKES, E. B.; MALLOWS, C. L. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, v. 78, n. 383, p. 553-569, 1983.
- GABRIEL, K. R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, v. 58, n. 3, p. 453-467, 1971.
- GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, v. 185, p. 1-17, 1986.
- HE, Z.; ZHANG, M.; ZHANG, H. Data-driven research on chemical features of Jingdezhen and Longquan celadon by energy dispersive X-ray fluorescence. *arXiv:1511.07825*, 2015. Disponível em: <https://arxiv.org/abs/1511.07825>.
- HE, Z.; ZHANG, M.; ZHANG, H. Data-driven research on chemical features of Jingdezhen and Longquan celadon by energy dispersive X-ray fluorescence. *Ceramics International*, v. 42, n. 4, p. 5123-5129, 2016.
- JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, v. 374, 20150202, 2016.
- KUHN, M. Building predictive models in R using the caret package. *Journal of Statistical Software*, v. 28, n. 5, p. 1-26, 2008.
- KUHN, M.; JOHNSON, K. Applied Predictive Modeling. New York: Springer Science+Business Media, 2013.
- LI, L.; FENG, S.-L.; FENG, X.-Q.; XU, Q.; YAN, L.-T.; MA, B.; HUO, H. Study on the chemical composition features of Longquan celadon excavated from the Chuzhou site of Huai'an City in Jiangsu Province by EDXRF. *Chinese Physics C*, v. 35, n. 7, p. 689-694, 2011.
- MALTSEV, A. S.; UMAROVA, N. N.; PASHKOVA, G. V.; MUKHAMEDOVA, M. M.; SHERGIN, D. L.; PANCHUK, V. V.; KIRSANOV, D. O.; DEMONTEROVA, E. I. Combination of Total-Reflection X-Ray Fluorescence Method and Chemometric Techniques for Provenance Study of Archaeological Ceramics. *Molecules*, v. 28, n. 3, 1099, 2023.
- MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, v. 405, n. 2, p. 442-451, 1975.
- NORTH, B. V.; CURTIS, D.; SHAM, P. C. A note on the calculation of empirical P values from Monte Carlo procedures. *American Journal of Human Genetics*, v. 71, n. 2, p. 439-441, 2002.
- OJALA, M.; GARRIGA, G. C. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, v. 11, n. 3, p. 1833-1863, 2010.
- PASHKOVA, G. V.; STATKUS, M. A.; MUKHAMEDOVA, M. M.; FINKELSHTEIN, A. L.; ABDRAHITOVA, I. V.; BELOZEROVA, O. Y.; CHUBAROV, V. M.; AMOSOVA, A. A.; MALTSEV, A. S.; DEMONTEROVA, E. I.; SHERGIN, D. L. A Workflow for Uncertainty Assessment in Elemental Analysis of Archaeological Ceramics: A Case Study of Neolithic Coarse Pottery from Eastern Siberia. *Heritage*, v. 6, n. 5, p. 4434-4450, 2023.
- PHIPSON, B.; SMYTH, G. K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, v. 9, Article 39, 2010.
- POLLARD, A. M.; HERON, C.; GILLARD, R. D. Archaeological Chemistry. Cambridge: Royal Society of Chemistry, 2008.
- POMERANTSEV, A. L. Chemometrics in Excel. New Jersey: John Wiley & Sons, 2014.
- POTTS, P. J.; WEST, M. (ed.). Portable X-ray Fluorescence Spectrometry: Capabilities for In Situ Analysis. Cambridge: Royal Society of Chemistry, 2008.
- POWERS, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37-63, 2011.
- RUIZ-PEREZ, D.; GUAN, H.; MADHIVANAN, P.; MATHEE, K.; NARASIMHAN, G. So you think you can PLS-DA? *BMC Bioinformatics*, v. 21, 2, 2020.
- RUSCHIONI, G.; MALCHIODI, D.; ZANABONI, A. M.; BONIZZONI, L. Supervised learning



- algorithms as a tool for archaeology: Classification of ceramic samples described by chemical element concentrations. *Journal of Archaeological Science: Reports*, v. 49, 103995, 2023.
- SHACKLEY, M. S. (ed.). *X-Ray Fluorescence Spectrometry (XRF) in Geoarchaeology*. New York: Springer, 2011.
- STÅHLE, L.; WOLD, S. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *Journal of Chemometrics*, v. 1, n. 3, p. 185-196, 1987.
- TITE, M. S. Ceramic production, provenance and use: a review. *Archaeometry*, v. 50, n. 2, p. 216-231, 2008.
- VAN RIJSBERGEN, C. J. *Information Retrieval*. 2. ed. London: Butterworths, 1979.
- WEHRENS, R. *Chemometrics with R: Multivariate data analysis in the Natural Sciences and Life Sciences*. Heidelberg: Springer-Verlag, 2011.
- WESTERHUIS, J. A.; HOEFSLOOT, H. C. J.; SMILDE, A. K. Assessment of PLS-DA cross validation. *Metabolomics*, v. 4, p. 81-89, 2008.
- WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, v. 58, n. 2, p. 109-130, 2001.
- WU, J.; LEUNG, P. L.; LI, J. Z.; STOKES, M. J.; LI, M. T. EDXRF studies on blue and white Chinese Jingdezhen porcelain samples from the Yuan, Ming and Qing dynasties. *X-Ray Spectrometry*, v. 29, n. 3, p. 239-244, 2000.
- WU, J.; WU, Y. F.; WU, J. M.; ZHANG, M. L.; LI, Q. J.; WU, T. T.; XU, L. Research on the composition characteristics of imitated Longquan celadon in Jingdezhen and Longquan celadon. *Spectroscopy and Spectral Analysis*, v. 33, p. 2246-2250, 2013.