Breve discussão sobre a importância do pré-tratamento de dados na análise de componentes principais

Sergio DOVIDAUSKAS[1], Isaura Akemi OKADA[1]

[1]Núcleo de Ciências Químicas e Bromatológicas – Centro de Laboratório Regional – Instituto Adolfo Lutz de Ribeirão Preto VI

A análise de componentes principais (ACP) constitui-se em uma ferramenta básica e importante da análise multivariada de dados. É particularmente útil quando se dispõe de dados de muitas variáveis obtidas para um grande número de amostras. A partir de uma matriz de dados de n linhas (amostras) e m colunas (variáveis), procura-se por correlações significativas entre as variáveis de modo a substituir as variáveis originais por outras, as componentes principais – o objetivo é fazer que as correlações permitam que o conjunto de dados possa ser descrito com um menor número de variáveis, ou seja, duas ou três componentes principais. Dessa forma, será possível visualizar padrões e relações entre as amostras e entre as variáveis; em outras palavras, é uma análise exploratória de dados[1].

Uma vez construída a matriz de dados, pode ser necessário (e muito frequentemente o é) o pré-tratamento antes da análise multivariada propriamente dita. Esse pré-tratamento dos dados pode incluir: (i) uma transformação aplicada às linhas da matriz (amostras) como, por exemplo, utilizar técnicas de alisamento e de correção de linha base em espectros, e a normalização ou a mudança para logaritmo; (ii) um pré-processamento aplicado às colunas (variáveis), como a centralização dos dados na média, o escalamento pela variância, o autoescalamento e o escalamento pela amplitude, por exemplo[2]. É particularmente sobre esse pré-processamento aplicado às colunas que lida essa comunicação.

Assim, inicialmente é recomendável que se analise cada coluna (variável) no que diz respeito à distribuição dos dados, ou seja, se essa distribuição pode ser considerada normal ou não. A priori, se a distribuição não for normal, pré-processamentos paramétricos deveriam ser evitados, como a centralização pela média, o escalamento pela variância e o autoescalamento; esse último envolve subtrair a média dos valores de cada elemento da coluna, dividindo-se o resultado pelo respectivo desvio padrão – em outras palavras, para se obter o valor autoescalado (x_a) de uma determinada variável para uma dada amostra, subtrai-se do valor obtido para cada amostra (x) a média obtida para o conjunto de amostras (X_m), dividindo-se o resultado pelo desvio padrão de X_m (s_{Xm}), conforme indica a equação 1 (observe nessa equação a relação direta com a padronização em termos de escores z de uma distribuição normal[3]):

$$x_a = \frac{x - X_m}{s_{X_m}} \quad \textit{equação 1}$$

Não obstante, é importante avaliar o impacto que poderia ser observado nos resultados de uma ACP se pré-processamentos paramétricos fossem aplicados a dados com distribuições assimétricas (ou seja, não normais). Para ilustrar, tomemos como exemplo um estudo realizado em nosso laboratório[4] em que 88 municípios (linhas ou amostras) foram representados pelas respectivas séries em 12 variáveis (colunas). As variáveis consistiam nas medianas de medidas realizadas durante um ano nas águas de abastecimento público dos municípios, nos parâmetros pH, condutividade e concentrações de 10 íons (Li^+, Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, ClO_3^-, NO_3^-, PO_4^{3-} e SO_4^{2-}). Os 88 dados em cada coluna (x) foram centralizados pela mediana da coluna (X_{med}) e escalados pelo respectivo intervalo interquartil (ii), obtendo-se o valor centralizado e escalado x_{medii}, conforme indica a equação 2:

$$x_{medii} = \frac{x - X_{med}}{ii} \quad \textit{equação 2}$$

Esse pré-processamento foi escolhido devido às distribuições tipicamente assimétricas observadas nas variáveis consideradas. A **Figura 1A** traz o gráfico de escores da ACP realizada com o algoritmo NIPALS (*Non-linear Iterative Partial Least-Squares*) e usando-se 4 componentes principais; esse gráfico exibiu 4 grupos: um grupo constituído de apenas um município (Ibitinga); o segundo grupo (em azul) é composto por 4 municípios, enquanto o terceiro apresenta 8 municípios (cor magenta); o grupo mais numeroso (75 municípios, cor verde) situa-se próximo à origem de gráfico CP1/CP2 e foi denominado "grupo típico". O respectivo gráfico de pesos (**Figura 1B**) indica que o grupo em azul apresenta as concentrações de sódio e lítio como variáveis proeminentes, enquanto a concentração de sulfato é a variável mais importante para o grupo em magenta; o grupo em verde não apresenta variáveis proeminentes no modelo estudado (como indica a sua posição no gráfico de escores), enquanto Ibitinga apresentou variáveis físico-químicas incomuns (teores relativamente maiores de sulfato, cloreto, lítio e sódio, além de maiores valores de pH e condutividade), tendo sido estudado individualmente[4]. A formação desses mesmos grupos havia sido também observada quando se efetuou a análise hierárquica de agrupamentos pelo método Ward, utilizando-se os valores centralizados e escalados (x_{medii})[4].

Para comparação, consideremos agora representar cada um dos 88 municípios pela respectiva série de médias (e não medianas) nas mesmas 12 variáveis consideradas anteriormente; em adição, no pré-processamento utilizaremos o autoescalamento da equação 1 em lugar da centralização/escalamento da equação 2. A ACP nas mesmas condições (algoritmo NIPALS usando 4 componentes principais) resultará nos gráficos de escores e de pesos indicados nas **Figuras 1C** e **1D**, respectivamente. Observe-se inicialmente que, ao se passar de um modelo baseado em medianas para um modelo baseado em médias, a variância explicada para 2 componentes principais diminui de 75% (56% em CP1 19% em CP2) para 51% (30% em CP1 e 21% em CP2) – em outras palavras: a qualidade do modelo exibido nas **Figuras 1A** e **1B** (medianas) é melhor que a do modelo indicado nas **Figuras 1C** e **1D** (médias).

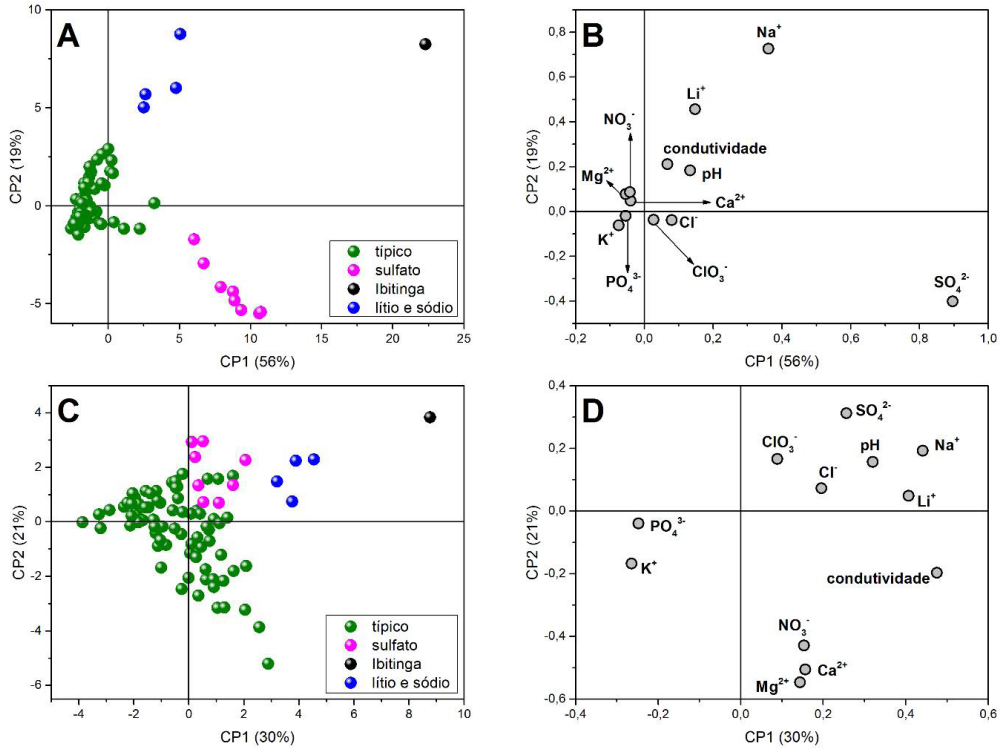

Fig 1. Análise de componentes principais de amostras de águas de abastecimento público de 88 municípios da região nordeste do Estado de São Paulo utilizando 12 variáveis: A/B, medianas; C/D, médias.

Essa queda na qualidade reflete na capacidade do modelo baseado em médias em produzir separações de grupos onde cada grupo inclua municípios com águas de abastecimento de perfis físico-químicos similares. Assim, observe-se que, se ainda é possível visualizar o grupo de apenas um município (Ibitinga), o "grupo do sulfato", claramente visualizado na **Figura 1A**, aparece na **Figura 1C** unido ao "grupo típico", ou seja, não há separação. Em adição, a distância entre o "grupo do cloreto" e o "grupo típico" é diminuída em relação ao modelo baseado em medianas. No que diz respeito à interpretação fornecida pelo gráfico de pesos (**Figura 1D**), observe-se que, se a interpretação para o município de Ibitinga não é muito prejudicada, no caso do grupo do cloreto não fica evidente que essa variável é a responsável pela formação do grupo em azul; situação pior é a do "grupo do sulfato": a análise simultânea dos gráficos das Figuras 1C e 1D não fornece indicações de que esse grupo exista.

Em conclusão: essa comunicação procurou demonstrar que, mesmo quando está se trabalhando com análises mais elaboradas, envolvendo muitas amostras e muitas variáveis, conhecimentos básicos sobre como lidar com os tipos de distribuição que se tem em mãos para as variáveis consideradas individualmente, pode ser a diferença entre a obtenção de um modelo interpretativo ou de um modelo que não fornece informação de qualidade.

AGRADECIMENTO

À Fundação de Amparo à Pesquisa do Estado de São Paulo pelo apoio financeiro (Processo FAPESP nº 2014/10034-2).

REFERÊNCIAS

1. Esbensen, K. H., *Multivariate Data Analysis - In Practice*. CAMO Process AS: Oslo, 2002; p 598.
2. Ferreira, M. M. C., *Quimiometria - Conceitos, Métodos e Aplicações*. Editora da Unicamp: Campinas (SP), 2015; p 495.
3. Moore, D. S.; McCabe, G. P., *Introdução à Prática da Estatística*. 3a ed.; LTC Editora: Rio de Janeiro, 2002; p 536.
4. Dovidauskas, S.; Okada, I. A.; Iha, M. H.; Cavallini, Á. G.; Okada, M. M.; Briganti, R. d. C., Parâmetros físico-químicos incomuns em água de abastecimento público de um município da região nordeste do Estado de São Paulo (Brasil). *Vigil. sanit. debate* **2017,** *5* (1), 106-115.