IV-238 - MONITORAMENTO DE QUALIDADE DE ÁGUA BASEADO EM CONTROLE ESTATÍSTICO MULTIVARIADO DE PROCESSO

Carolina Cristiane Pinto[(1)]
Engenheira Química pela Universidade Federal de São João del-Rei (UFSJ). Mestre em Saneamento, Meio Ambiente e Recursos Hídricos pela Escola de Engenharia da Universidade Federal de Minas Gerais (UFMG). Doutoranda em Engenharia Química na UFMG. Analista Ambiental do Instituto Mineiro de Gestão das Águas (Igam).
Fábio Palmer Caldeira Parreiras de Faria
Estudante do Curso de Graduação em Engenharia Química na Universidade Federal de Minas Gerais (UFMG).
Gustavo Matheus de Almeida
Engenheiro Químico pela Universidade Federal de Minas Gerais (UFMG). Mestre em Engenharia Química pela Escola de Engenharia da UFMG. Doutor em Engenharia Química pela Universidade de São Paulo (USP). Professor Associado do Departamento de Engenharia Química da UFMG (DEQ/UFMG).

Endereço[(1)]**:** Rua Jasmim, 24, Belvedere, Congonhas, MG, 36.415-000, Brasil, Telefone: (31) 9 8554-1679, e-mail: carol-cp@ufmg.br

RESUMO

Os Programas de Monitoramento da Qualidade das Águas (PMQAs) são de suma importância para a proteção do recurso hídrico, a manutenção da capacidade ambiental e o controle da poluição. Uma maneira de melhorar o processo de tomada de decisão sobre esses aspectos é incrementar as análises dos dados coletados. Neste trabalho, avaliou-se o conjunto de dados de uma estação de monitoramento localizada na calha do rio das Velhas (BV137), composto por doze parâmetros de qualidade da água, coletados ao longo de vinte e um anos (1997 a 2018). Realizou-se a análise dos dados com a técnica estatística multivariada denominada Análise por Componentes Principais (PCA), a partir de uma abordagem via *scores*. Essa técnica facilita a visualização e interpretação de problemas multivariados altamente correlacionados. A partir do modelo PCA obtido, característico de situações usuais, e utilizado como sistema de monitoramento, verificou-se a sua capacidade em identificar dados discrepantes, e em obter e associar regiões espaciais em seus gráficos de componentes principais aos fatores de pressão ambiental característicos da área em estudo, a saber, lançamentos de efluentes domésticos e industriais e atividades minerárias do Alto Velhas. Os parâmetros de qualidade da água com maior relevância foram coliformes termotolerantes, sólidos totais, turbidez e arsênio total. O uso do modelo PCA mostrou-se promissor para a identificação de padrões de comportamento e classificação de cenários de poluição, podendo ser utilizado como sistemas de suporte à decisão em atividades de monitoramento e gerenciamento de qualidade de água.

PALAVRAS-CHAVE: Qualidade de Água, Monitoramento, CEP Multivariado, Análise por Componentes Principais, Visualização de dados, Análise de Agrupamentos.

INTRODUÇÃO

A manutenção da qualidade das águas superficiais é crucial para a garantia da saúde pública e vida aquática em todo o mundo. Para se alcançar esse objetivo, é fundamental o controle de suas propriedades físicas, químicas e biológicas, a partir de Programas de Monitoramento da Qualidade das Águas (PMQAs) (Kannel *et al*., 2007; Ouyang, 2006). As análises dos dados obtidos com esses programas de monitoramento aumentam o entendimento sobre mudanças espaciais e/ou temporais na qualidade da água, com os objetivos de proteger o recurso hídrico, manter a capacidade ambiental e de controlar a poluição (Behmel *et al*., 2016; Chen *et al*., 2012). Em resumo, buscam-se com essas análises, processos de tomada de decisão mais racionais.

Com o avanço tecnológico das áreas de instrumentação, informática e bancos de dados, observou-se, ao longo dos últimos anos, a coleta e o acúmulo de grandes massas de dados. O desafio nesse novo cenário, seja no setor público ou privado, é transformar esses dados brutos em informação relevante.

Um modo de promover a geração de conhecimento é a atividade de monitoramento, com o uso de ferramentas estatísticas. A principal ferramenta estatística em monitoramento é a carta de controle, a partir de uma abordagem de CEP (Controle Estatístico de Processos) univariado. Os exemplos usuais são as cartas de controle de Shewhart, CUSUM e EWMA, com um número crescente de aplicações em problemas de monitoramento em engenharia ambiental (Samsudin *et al*., 2017; Sancho *et al*., 2016; Iglesias *et al*., 2015; Folladoret *al*., 2012; Corbett e Pan, 2002). Dada a correlação espacial entre variáveis ou parâmetros, a abordagem via CEP univariado, em geral, não é adequada. Portanto, é importante considerar o CEP multivariado, cuja técnica comumente utilizada é aquela denominada Análise por Componentes Principais (*Principal Components Analysis*; PCA). As aplicações de PCA em qualidade de água comumente são com os objetivos de agrupar estações de monitoramento; associar componentes principais a fontes de contaminação, como sedimentos, agricultura e esgoto; detectar fontes poluidoras; e identificar parâmetros com variações espaço-temporal (Olsen *et al*. 2012; Bhat *et al*., 2014; Zhang *et al*., 2011; Kowalkowski *et al*., 2006; Mendiguchía *et al*., 2004; Singh *et al*., 2004; Simeonov *et al*., 2003). PCA pode ser utilizado com diversos objetivos (Bro e Smilde, 2014). Porém, conforme relatado por Olsen *et al*. (2012) e Sergeant *et al*. (2016), ainda se encontram nos estudos modernos de aplicações de PCA em dados de qualidade de água erros analíticos e descuidos evitáveis.

A Análise por Componentes Principais, pertencente à área de estatística multivariada, é capaz de lidar com a alta dimensão de sistemas, ao considerar a correlação espacial entre suas variáveis ou parâmetros (Ge e Song, 2013). Portanto, PCA é útil em caso de processos multivariados altamente correlacionados, como aqueles em monitoramento de qualidade de água. A possibilidade de redução de dimensionalidade de um problema é útil por facilitar a sua visualização, e, consequentemente, a sua interpretação, e então, o seu entendimento.

O objetivo geral do presente trabalho é demonstrar o uso potencial de PCA como um sistema de suporte à decisão ao monitoramento de qualidade de água. Exemplifica-se a partir de duas aplicações. A primeira diz respeito à identificação de condições discrepantes em relação àquelas usuais, e a segunda, à classificação de eventos não usuais (condições críticas de qualidade de água), o que pode ser usado como um sinal de alerta. Um monitoramento mais eficiente sobre o estado da qualidade da água e das alterações em parâmetros-chave contribui para um processo de tomada de decisões mais racional. O estudo de caso refere-se a uma estação de monitoramento localizada na bacia hidrográfica do rio das Velhas, uma das principais sub-bacias da bacia do rio São Francisco.

Após uma descrição sucinta sobre PCA, apresenta-se a metodologia para as duas aplicações. Em seguida, apresentam-se e discutem-se os resultados, e por fim, as considerações finais. Utilizou-se, neste trabalho, a linguagem de programação Python (https://www.python.org/) de licença livre e código aberto, com os pacotes *pandas*, *numpy*, *seaborn*, *matplotlib* e *scipy*, além de sua biblioteca padrão.

ANÁLISE POR COMPONENTES PRINCIPAIS (PCA)

PCA é uma técnica de redução de dimensionalidade. A ideia é explicar a maior parte da estrutura de variância-covariância das variáveis originais a partir de um número menor de dimensões, denominadas componentes principais. Desse modo, dada a matriz de dados originais, $X_{[n,p]}$, em que n é o número de observações e p é o número de variáveis $(X_1, X_2, \cdots, X_p)$ (por exemplo, parâmetros de qualidade de água), busca-se um novo sistema de coordenadas onde k dimensões (componentes principais; ortogonais entre si) expliquem a máxima informação (variância) das p dimensões (variáveis) originais, com $k \ll p$. A Equação 1 resume a relação entre as variâncias das componentes (λ_i) e das variáveis originais (σ_i^2). Para tal, realiza-se a rotação ótima, segundo o critério de máxima variância, dos eixos do sistema original de coordenadas. As componentes principais (PC_i) são combinações lineares das variáveis originais (X_i), conforme a Equação 2, em que w_{ij} é o peso (*load*) associado à $j-$ésima variável na $i-$ésima componente principal. Pode-se considerar o peso como uma medida da influência da variável na componente principal. O valor de uma componente é denominado *score*. Em outras palavras, os *scores* são as coordenadas dos pontos originais no novo sistema (rotacionado) de coordenadas (Hair *et al*., 2009; Manly, 2008; Russell *et al*. 2000; Sharma, 1996).

$$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_i^2 \qquad (1)$$
$$i = 1,2,\cdots,p$$

$$PC_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{ip}X_p = \sum_{j=1}^{p} w_{ij}X_j \quad (2)$$

$$j = 1,2,\cdots,p$$

METODOLOGIA

Dividiu-se a metodologia em quatro etapas, conforme a Figura 1. Descrevem-se, a seguir, cada uma das etapas.

Figura 1: Etapas da metodologia.

ETAPA 1

Definição do Problema (área de estudo)

Definiu-se, como estudo de caso, a bacia hidrográfica do rio das Velhas, correspondente à Unidade de Planejamento e Gestão de Recursos Hídricos (UPGRH) SF5. É uma das principais sub-bacias da bacia do rio São Francisco, ao contribuir com expressivo volume de água e possuir o maior afluente em extensão da bacia do São Francisco, o rio das Velhas, com 806,84 km (PDRH Rio das Velhas, 2015). De acordo com Costa *et al*. (2017) e Trindade *et al*. (2017), essa sub-bacia é a mais poluída e a principal responsável pela deterioração da qualidade das águas superficiais do rio São Francisco. A bacia é dividida em três regiões fisiográficas (Baixo, Médio e Alto São Francisco) (Figura 2).

Banco de Dados

Utilizou-se um banco de dados históricos referente ao monitoramento da qualidade da água na sub-bacia do rio das Velhas, disponível pelo Instituto Mineiro de Gestão das Águas (Igam). Segundo Calazans (2015), uma das estações prioritárias para o monitoramento e controle do rio das Velhas é a BV137, no município de Lagoa Santa, no trecho médio do rio das Velhas. O conjunto de dados utilizado inclui os principais parâmetros físico-químicos e microbiológicos para a avaliação de uma rede de monitoramento de qualidade de água: matéria orgânica (*demanda bioquímica de oxigênio*), microrganismos (*coliformes termotolerantes*), íons dissolvidos (*condutividade elétrica* e *cloreto*), e nutrientes (*nitrato* e *fósforo*) da descarga de águas residuais não tratadas e da poluição difusa das atividades agrícolas. A descarga industrial e os processos erosivos podem alterar as características da água com partículas sólidas (*turbidez*, *sólidos totais* e *arsênio*, sendo este último um parâmetro característico da região, segundo Christofaro e Leão (2009)). Todas essas fontes de contaminantes podem modificar o *pH*, a *temperatura* e o *oxigênio dissolvido* na água. Em resumo, o banco de dados é composto por 12 parâmetros, coletados entre 1997 e 2018, num total de 148 amostras.

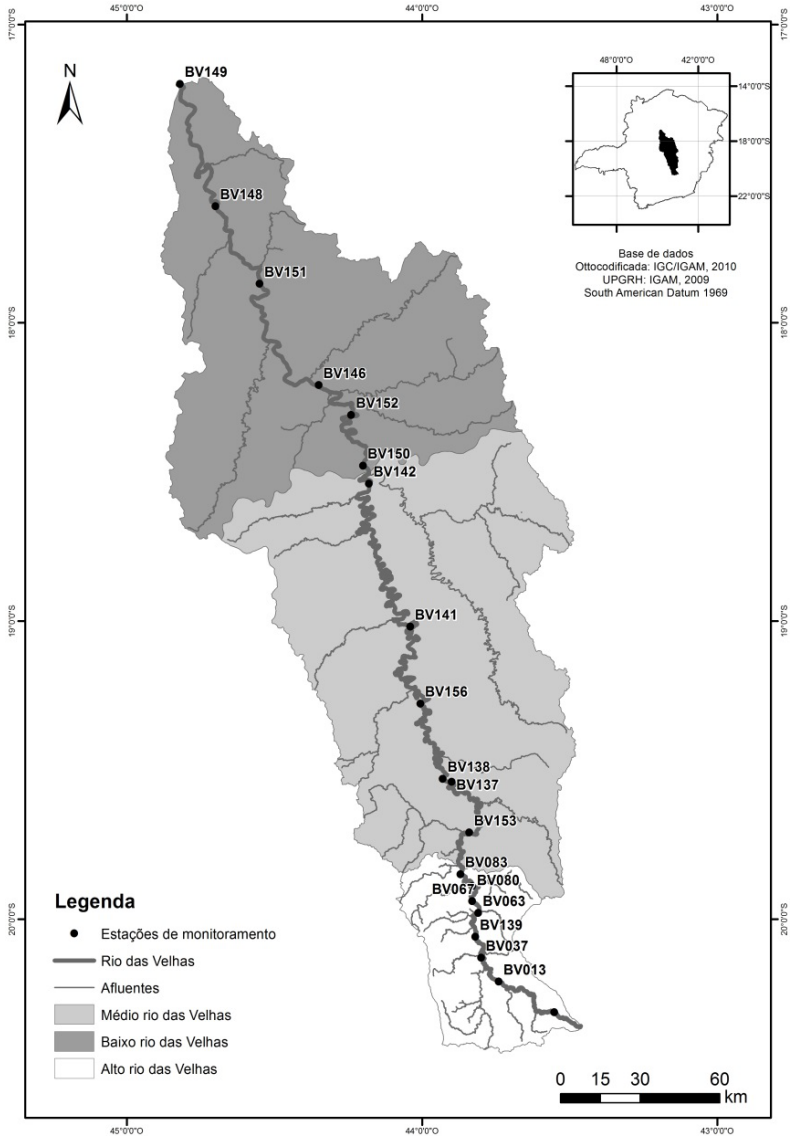

Figura 2: Localização geográfica da bacia do rio das Velhas.

ETAPA 2

Pré-Processamento de Dados

A etapa de pré-processamento é crucial em qualquer atividade de análise de dados. Os objetivos são: conhecer as principais características do conjunto de dados, e gerar um banco de dados de trabalho, a partir dos dados brutos. Inicialmente, realizou-se uma análise descritiva (cálculo de média, desvio-padrão, mínimo máximo, e percentis 25, 50 e 75) dos parâmetros. Em seguida, realizou-se uma inspeção visual dos dados, a partir de histogramas e gráficos de dispersão e de séries temporais, com o objetivo de identificar registros faltosos e dados discrepantes, por exemplo. Com essa análise exploratória dos dados, conheceu-se melhor o conjunto de dados e verificou-se a sua adequação para a presente proposta de estudo.

Num segundo momento, identificaram-se, de modo quantitativo, os dados discrepantes. Por usar $\bar{X}$, a média aritmética simples, o método ±3σ, comumente utilizado para a detecção de *outliers*, é sensível a dados extremos. Ressalta-se a presença significativa de *outliers* no conjunto de dados do presente estudo de caso. Com isso, adotou-se o filtro de Hampel, que utiliza a mediana (Lin *et al.*, 2006; Davies e Gather, 1981). Nesse caso, calcula-se a estatística MAD (*Median Absolute Deviation*; Equação 3(a)), a partir dos desvios absolutos das n observações (x_i; $i = 1,2,\cdots,n$) em relação à mediana ($\tilde{X}$), dado um parâmetro em particular. Os limites

inferior e superior (Equações 3(b) e 3(c), respectivamente) são mais estreitos em relação à abordagem tradicional.

$$MAD = 1{,}4868 \cdot Mediana\{|x_i - \tilde{X}|\} \quad (3a)$$

$$LS\ (Limite\ Superior) = \tilde{X} + 3 \cdot MAD \quad (3b)$$

$$LI\ (Limite\ Inferior) = \tilde{X} - 3 \cdot MAD \quad (3c)$$

Desse modo, em caso de pelo menos um valor discrepante para quaisquer parâmetros, classificou-se a respectiva amostra como observação discrepante. Utilizou-se esse procedimento como o ponto de partida para a construção dos dois modelos PCA, sendo um para cada aplicação proposta neste trabalho. O objetivo com a primeira aplicação é demonstrar a capacidade de PCA em detectar observações discrepantes. Após essa comprovação, o objetivo com a segunda aplicação é usar esse potencial para classificar eventos não usuais (críticos) em relação à qualidade de água, e associá-los a faixas, não usuais, de parâmetros-chave.

ETAPA 3

Aplicação 1: Reconhecimento de Observações Discrepantes

Inicialmente, padronizaram-se os parâmetros ($X \rightarrow X_p$) de modo a se ter média (μ) 0 e desvio-padrão (σ) 1 (Equação 4). Esse procedimento é crucial para evitar a prevalência daqueles parâmetros com maiores desvios-padrões sobre a definição dos eixos (as componentes principais) no novo sistema (rotacionado) de coordenadas (Sergeant *et al*., 2016), conforme descrição na seção Análise por Componentes Principais (PCA).

$$X_p = (X - \mu)/\sigma \quad (4)$$

Em seguida, identificou-se um modelo PCA com o conjunto completo de observações, discrepantes e não discrepantes. Analisam-se então os gráficos de *scores* com o objetivo de se verificar a sua capacidade em separar, de modo espacial, as observações discrepantes daquelas não discrepantes segundo classificação de Hampel.

ETAPA 4

Aplicação 2: Classificação de Eventos Não Usuais (Eventos Críticos)

Após a separação entre observações discrepantes e não discrepantes, identificou-se um modelo PCA apenas com as observações não discrepantes segundo o filtro de Hampel. Utilizou-se então esse modelo, característico de situações usuais, uma vez que foi obtido com observações não discrepantes, como o sistema de monitoramento de qualidade de água. Ressalta-se que a denominação “usual” não significa uma condição de qualidade de água satisfatória.

Em seguida, padronizaram-se as observações discrepantes com as médias e desvios-padrões calculados anteriormente com as observações não discrepantes, e, com o modelo PCA obtido com o conjunto de observações não discrepantes, calcularam-se os respectivos *scores* para as observações discrepantes padronizadas. Analisaram-se, então, os gráficos de *scores* com o objetivo de se associar regiões espaciais com eventos não usuais em relação à qualidade de água. Em seguida, buscou-se caracterizar essas regiões com as faixas de valores dos parâmetros de qualidade. Com a identificação daqueles parâmetros com maior influência sobre uma região em particular, associou-se essa região a um evento não usual, porém, factível, em relação à qualidade de água.

RESULTADOS E DISCUSSÕES

ETAPA 1

Definição do Problema (área de estudo), e Banco de Dados

Apresentou-se o estudo de caso e o seu conjunto de dados nas seções “Definição do Problema” e “Banco de Dados”.

ETAPA 2

Pré-Processamento de Dados

Inicialmente, excluíram-se duas amostras com dados faltantes. Com isso, o conjunto de dados de trabalho contém 146 registros. A Tabela 1 contém um sumário de estatísticas descritivas sobre os parâmetros. A partir dos valores médios, observam-se violações de coliformes termotolerantes, fósfoto total, oxigênio dissolvido e turbidez, conforme os respectivos limites legais da Deliberação Normativa Conjunta COPAM/CERH-MG 01/2008 para corpos d'água de classe 3, o que é um indicativo de degradação da qualidade das águas nessa região. Ressalta-se o nível de coliformes termotolerantes, com valor médio de 41.640,35 NMP/100 mL, aproximadamente dez vezes superior ao limite máximo permitido de 4.000 NMP/100 mL para classe 3.

Tabela 1: Estatística descritiva sobre os parâmetros de qualidade da água; estação de monitoramento BV137, entre 1997 e 2018.

Parâmetros	Abreviação	Unidade	Média	Desvio-padrão	Mínimo	Máximo	Percentil 25	Mediana	Percentil 75
Arsênio Total	As_T	mg/L	0,033	0,021	0,0012	0,138	0,020	0,029	0,040
Cloreto Total	Cl^-	mg/L	15,28	8,12	3,27	46,10	8,88	14,05	20,53
Coliformes Termotolerantes	Colif. Term.	NMP/100 mL	41640,35	58384,72	50,00	160000,00	3300,00	13000,00	49027,75
Condutividade Elétrica (*in loco*)	CE	µS/cm	254,87	97,35	95,60	524,00	177,00	238,00	311,25
Demanda Bioquímica de Oxigênio	DBO	mg/L	9,29	12,26	2,00	131,00	4,93	6,45	9,00
Fósforo Total	P_T	mg/L	0,41	0,27	0,01	1,61	0,23	0,34	0,57
Nitrato	$N\text{-}NO_3^-$	mg/L	0,89	0,95	0,01	6,66	0,24	0,67	1,12
Oxigênio Dissolvido	OD	mg/L	3,38	1,38	0,50	8,10	2,60	3,55	4,18
pH (*in loco*)	pH	–	7,12	0,37	6,00	8,00	6,90	7,20	7,38
Sólidos Totais	S_T	mg/L	394,71	459,93	130,00	3622,00	194,25	226,00	304,00
Temperatura da Água	$T_{água}$	°C	24,80	2,54	18,00	31,70	22,80	25,00	26,90
Turbidez	Turb.	NTU	174,10	382,22	4,18	3100,00	17,38	39,65	136,25

Em relação à inspeção visual dos dados, a Figura 3(a,b) mostra, respectivamente, o histograma para cloreto total e o gráfico temporal para sólidos totais. Em ambos os casos, observam-se dados discrepantes. Esse tipo de dado é comum em qualquer sistema, e o objetivo desse trabalho é utilizá-los para classificar eventos não usuais em relação à qualidade de água. Por fim, classificaram-se as observações entre discrepantes (74 registros) e não discrepantes (72 registros), segundo o filtro de Hampel. A Figura 3(c-d) mostra os gráficos temporais, com a identificação desses tipos de observações, para coliformes termotolerantes e nitrato, respectivamente. Empregou-se essa classificação para a obtenção de ambos os modelos PCA.

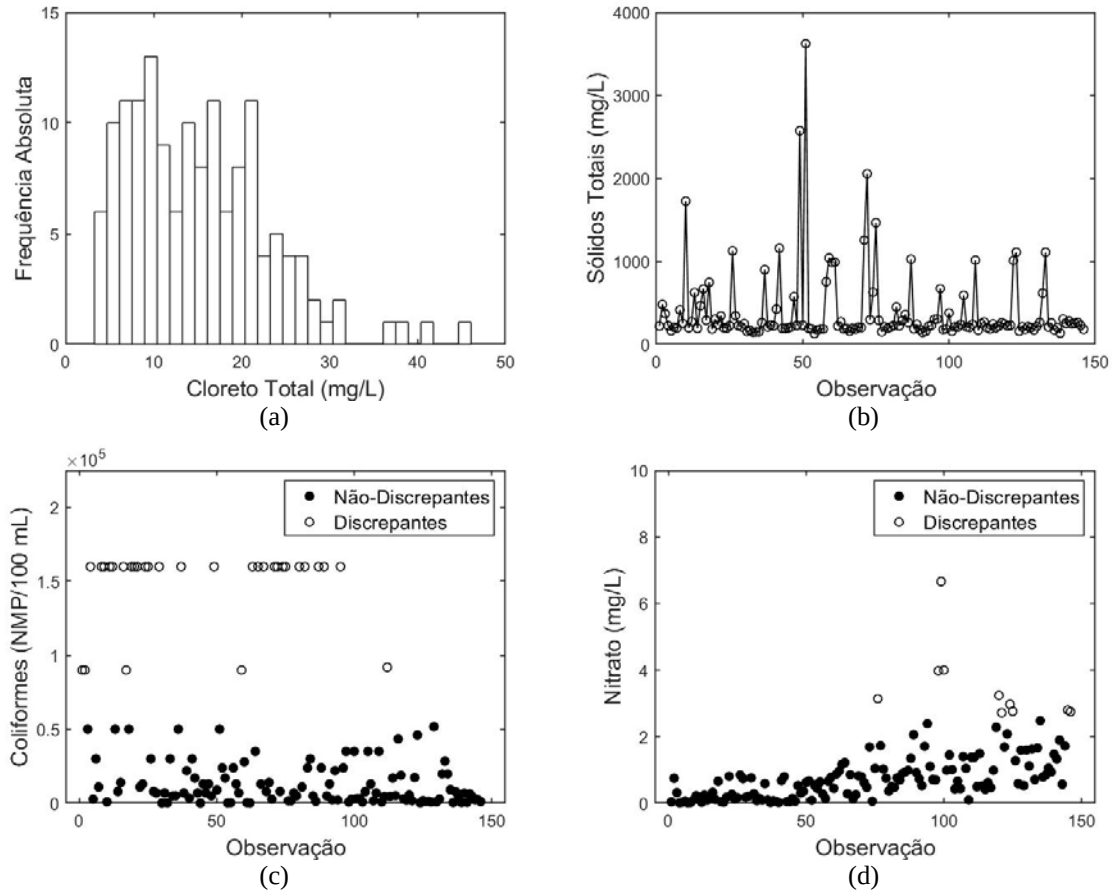

Figura 3: (a) Histograma, (b) gráfico temporal, e (c-d) gráficos com dados discrepantes e não discrepantes segundo o filtro de Hampel.

ETAPA 3

Aplicação 1: Identificação de Observações Discrepantes

Construiu-se o modelo PCA com o conjunto completo de observações, isto é, discrepantes e não discrepantes, segundo o filtro de Hampel. Em seguida, analisaram-se os gráficos de *scores* entre os pares de componentes principais, desde PC_1-PC_2, PC_1-PC_3, ⋯, até PC_{11}-PC_{12}. Nesse tipo de estudo, é recomendável analisar até as combinações entre componentes com variâncias relativamente baixas.

A Figura 4(a-d) mostra o resultado para PC_1-PC_2, PC_1-PC_3, PC_2-PC_3, e PC_1-PC_8. Em geral, observam-se os dados não discrepantes, segundo Hampel, concentrados numa nuvem de pontos (círculos sólidos), e os dados discrepantes, segundo Hampel, no entorno daqueles não discrepantes, concentrados ou dispersos (círculos vazados). Portanto, há uma correspondência entre a técnica PCA e a classificação segundo Hampel, utilizada como ponto de partida. A Figura 4(e-h) mostra histogramas para as componentes PC_1, PC_2, PC_3 e PC_8, respectivamente. Pode-se observar a diferença entre as distribuições de *scores* entre dados discrepantes e não discrepantes, principalmente para PC_1 e PC_8. As componentes, ainda que com percentual relativamente baixo de explicação da variância total dos dados originais, podem ser úteis para diferenciar eventos. As observações discrepantes resultam de valores discrepantes em pelo menos um dos parâmetros de qualidade, sendo um indicativo de situações indesejáveis. A apresentação desses resultados ilustra o uso potencial de PCA para a detecção de eventos não usuais (condições críticas) em atividades de monitoramento de qualidade de água. Esses eventos correspondem a dados discrepantes em relação a operações normais, cuja definição é a partir de dados não discrepantes.

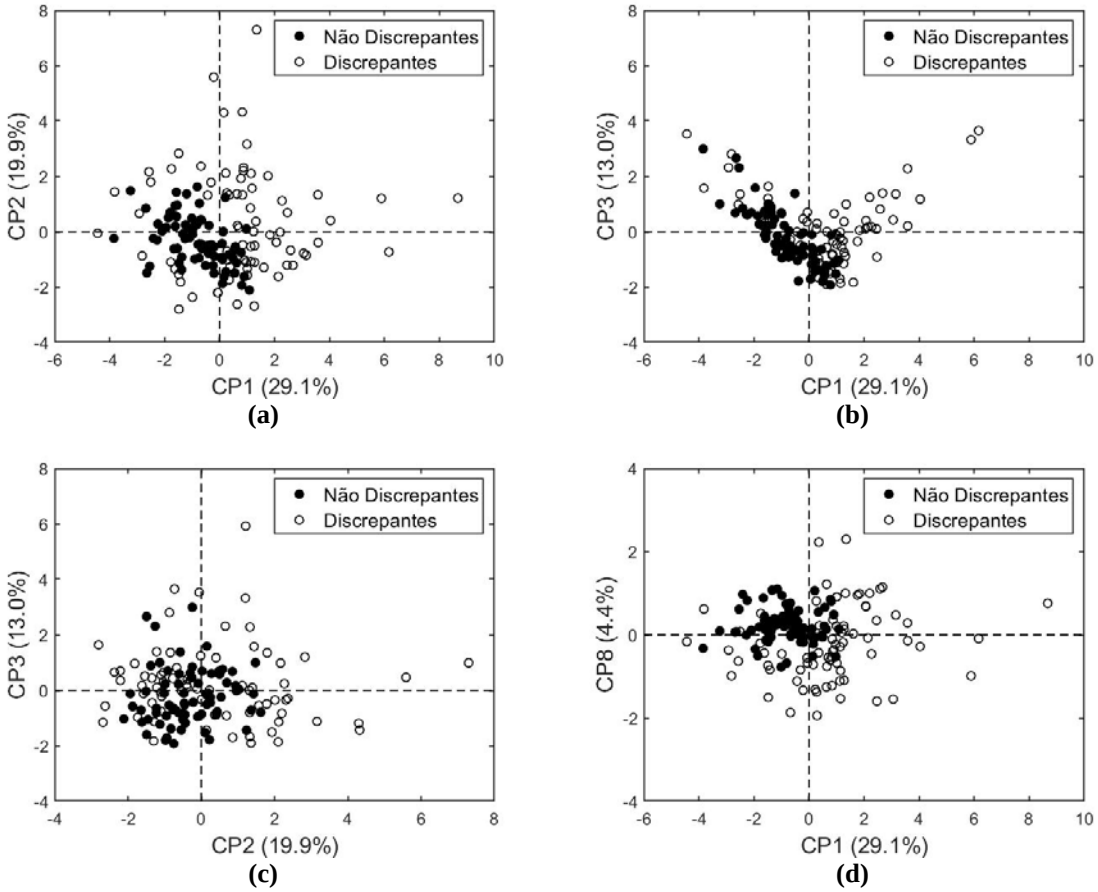

Figura 4(a-d): Gráficos de *scores*: (a) PC_1-PC_2, (b) PC_1-PC_3, (c) PC_2-PC_3, e (d) PC_1-PC_8, e histogramas: (e) PC_1, (f) PC_2, (g) PC_3, e (h) PC_4.

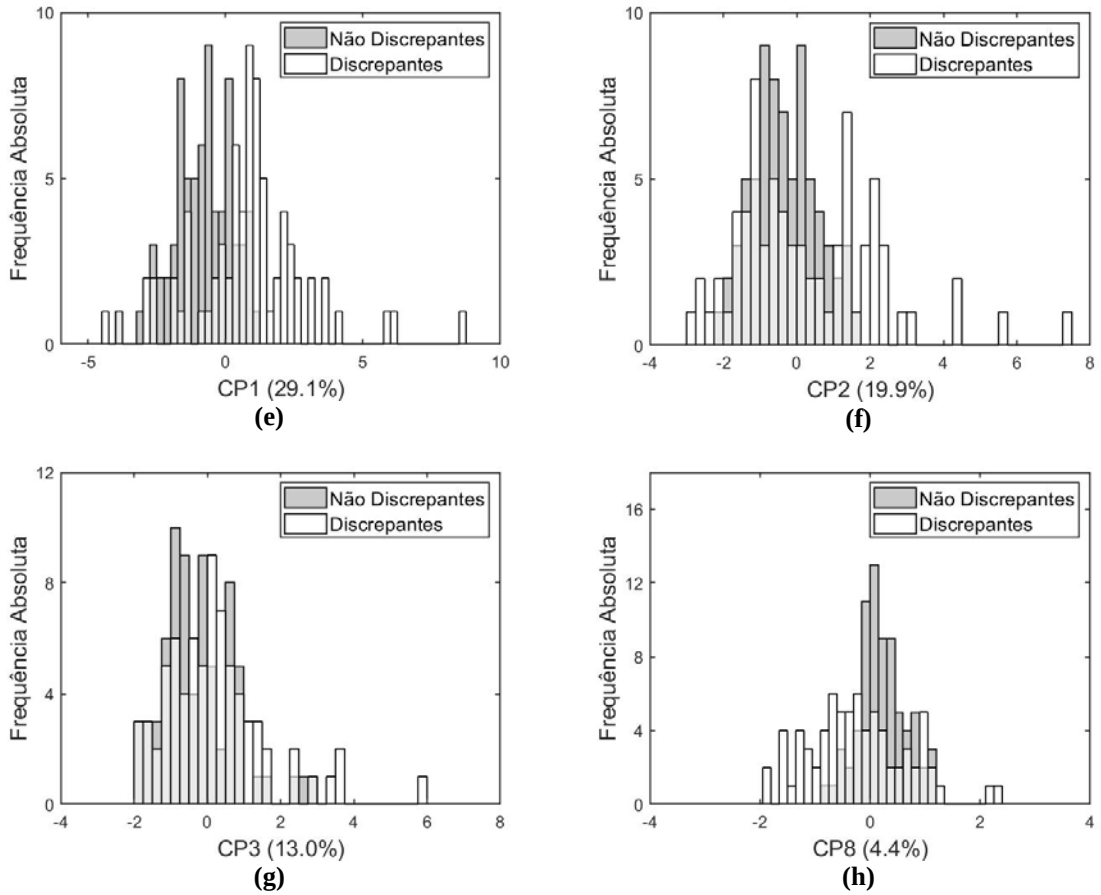

Figura 4(e-h): Gráficos de *scores*: (a) PC_1-PC_2, (b) PC_1-PC_3, (c) PC_2-PC_3, e (d) PC_1-PC_8, e histogramas: (e) PC_1, (f) PC_2, (g) PC_3, e (h) PC_4.

ETAPA 4

Aplicação 2: Classificação de Eventos Não Usuais

Construiu-se o modelo PCA apenas com o conjunto de observações não discrepantes, segundo o filtro de Hampel. Pela Tabela 2, o percentual de variância total dos dados originais explicada pelas cinco primeiras componentes é de aproximadamente 75% (74,6%), um valor significativo em problemas de engenharia.

Tabela 2: Modelo PCA obtido a partir de observações não discrepantes.

Variância (Componente Principal)	3,5	2,1	1,6	1,0	0,8	0,7	0,6	0,6	0,4	0,3	0,2	0,1
Variância Explicada (%)	29,0	17,5	13,1	8,5	6,6	5,9	5,1	4,9	3,6	2,8	2,1	1,0
Variância Explicada Acumulada (%)	29,0	46,5	59,5	68,0	74,6	80,5	85,7	90,5	94,1	96,9	99,0	100,0
Parâmetro / Componente Principal (CP)	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**	**CP8**	**CP9**	**CP10**	**CP11**	**CP12**
As_T	-0,292	0,119	0,289	-0,415	-0,231	-0,280	0,497	-0,313	0,309	-0,139	-0,221	0,069
Cl^-	-0,443	-0,074	0,119	0,025	0,314	0,300	-0,167	-0,260	0,102	0,237	-0,255	-0,607
Colif. Term.	0,287	-0,082	0,432	-0,148	-0,026	-0,111	-0,656	-0,311	0,084	-0,386	-0,079	0,021
CE	-0,471	-0,064	0,159	0,070	0,161	0,197	-0,199	-0,156	-0,028	0,232	0,069	0,745
DBO	-0,008	0,113	0,517	0,558	-0,408	-0,018	0,150	-0,118	-0,423	0,121	-0,046	-0,099
P_T	-0,214	-0,372	0,236	-0,312	0,310	-0,394	0,056	0,244	-0,557	-0,057	0,157	-0,102
$N\text{-}NO_3^-$	-0,207	0,444	-0,231	-0,262	-0,292	-0,189	-0,293	-0,264	-0,196	0,227	0,495	-0,163
OD	-0,283	0,378	-0,333	0,152	0,064	-0,227	-0,172	0,071	-0,297	-0,371	-0,566	0,082
pH	-0,363	0,030	0,204	0,164	-0,231	-0,231	-0,264	0,630	0,450	-0,018	0,111	-0,101
S_T	-0,061	0,465	0,288	-0,195	0,162	0,554	0,133	0,246	-0,121	-0,419	0,238	-0,032
$T_{ÁGUA}$	0,132	0,378	0,147	0,340	0,621	-0,421	0,122	-0,149	0,219	0,045	0,230	0,003
Turb.	0,305	0,344	0,239	-0,346	0,066	-0,023	-0,112	0,291	-0,070	0,581	-0,401	0,080

Em seguida, calcularam-se os *scores* para as observações discrepantes utilizando esse modelo PCA. Localizaram-se, então, três regiões (1, 2 e 3) entre o conjunto de dados não discrepantes, conforme a descrição a seguir. Em cada caso, são apresentados os gráficos de *scores* para os pares de combinações entre as três primeiras componentes, ou seja, PC_1-PC_2, PC_1-PC_3 e PC_2-PC_3. A região de referência, característica de valores usuais para os parâmetros de qualidade, é envolta por um limite de controle calculado com nível de confiança de 95%. Ela é relativa ao conjunto de dados não discrepantes, utilizado para a construção do modelo PCA (Tabela 2).

A região 1 (em vermelho) localiza-se no interior do limite de controle no plano PC_1-PC_2 (Figura 5(a)); porém, fora dele nos planos PC_1-PC_3 (Figura 5(b)) e PC_2-PC_3 (Figura 5(c)). Isso porque não há uma relação significativa com PC_2. Ela é característica, principalmente de valores relativamente altos de coliformes (Figura 5(d)) e de DBO e relativamente baixos de oxigênio dissolvido, de modo conjunto. Essa condição é um indicativo de um elevado grau de degradação desse trecho do rio das Velhas, sendo geralmente correspondente ao lançamento de esgoto doméstico. A Figura 5(d) compara as distribuições de coliformes entre a situação usual (dados não discrepantes) e a região 1. Observam-se mudanças em suas faixas. Pode-se observar também que a nuvem de dados não discrepantes localiza-se, sempre, dentro do limite de controle, conforme desejado.

Os demais pontos não discrepantes dentro do limite de controle correspondem a 28% do total de pontos, desconsiderando-se aqueles da região 1. Esses pontos, principalmente em PC_1-PC_3 (Figura 5(b)) e PC_2-PC_3 (Figura 5(c)) localizam-se no entorno da nuvem de pontos não discrepantes. Eles são característicos, principalmente, de valores mais altos de coliformes, DBO e turbidez, entre outros. Nesse trabalho, definiu-se o conjunto de registros usuais (dados não discrepantes) de modo automático, a partir do filtro de Hampel. A conferência de um especialista é importante para validar esses procedimentos. Como essa ação implicaria em

reduzir o conjunto de dados de observações não discrepantes, optou-se por manter o resultado da classificação segundo Hampel, o que não compromete a análise das regiões 1 a 3. No caso de redução do número de observações não discrepantes para a identificação do modelo PCA, a maior parte desses pontos se localizaria além do limite (recalculado) de controle.

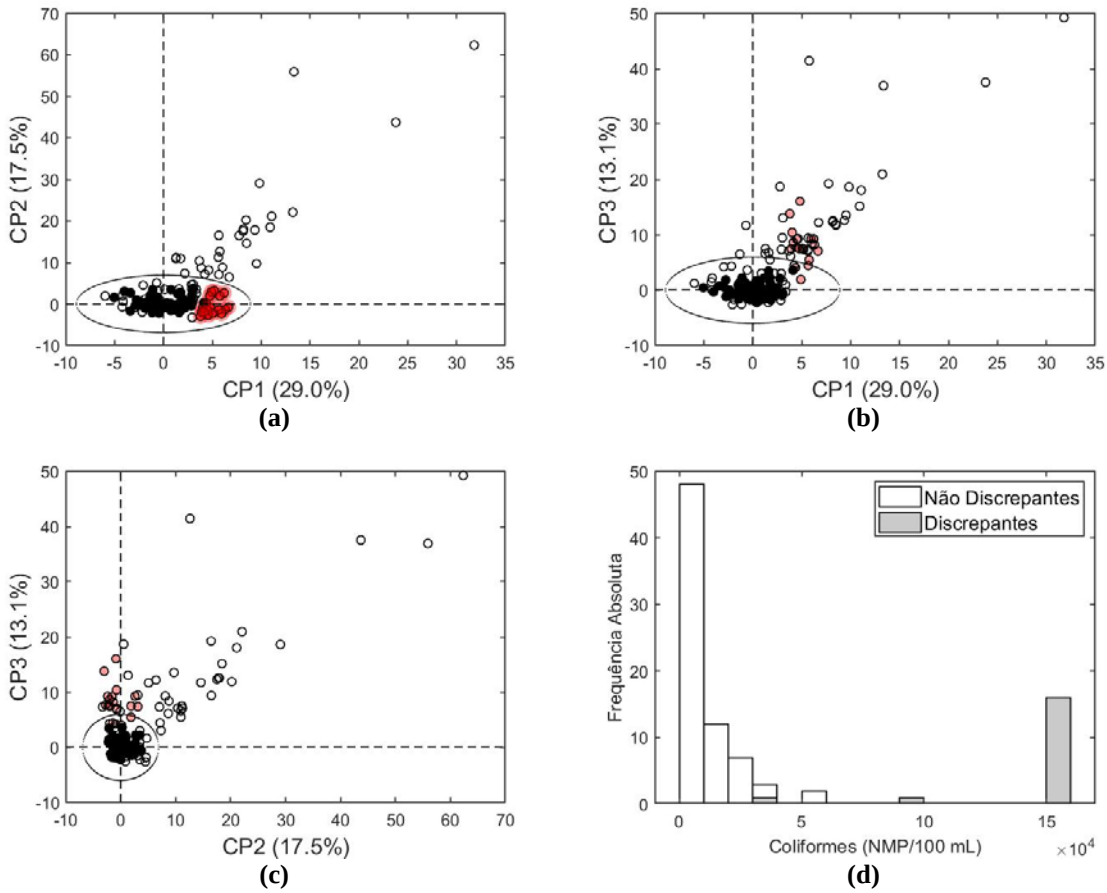

Figura 5: Gráficos de *scores*: (a) PC_1-PC_2, (b) PC_1-PC_3, e (c) PC_2-PC_3, e (d) histograma para DBO.

A Figura 6 mostra a região 2 (em vermelho). Ela é característica de concentrações elevadas, principalmente, de Sólidos Totais (Figura 6(d)) e Turbidez, em relação à situação usual. A facilidade de reconhecimento de um mesmo agrupamento varia entre as combinações de componentes principais. Para essa região, os planos candidatos em uma atividade de monitoramento seriam PC_1-PC_3 e PC_2-PC_3. A Figura 7 é um exemplo de uma condição extrema, contendo três observações consideravelmente discrepantes. Essa região 3 é característica de concentrações significativamente altas de arsênio total (Figura 7(d)), coliformes, sólidos totais e turbidez, em relação à situação usual. Essa condição é um indicativo de atividades minerárias (metalurgia de ouro). Dada a magnitude dessa discrepância, esse evento não usual é reconhecível em qualquer dos planos de componentes principais. Na prática, caso não fosse necessário utilizar a terceira componente principal, dever-se-ia optar pelo plano PC_1-PC_2, de modo a reduzir o número de gráficos de controle a acompanhar.

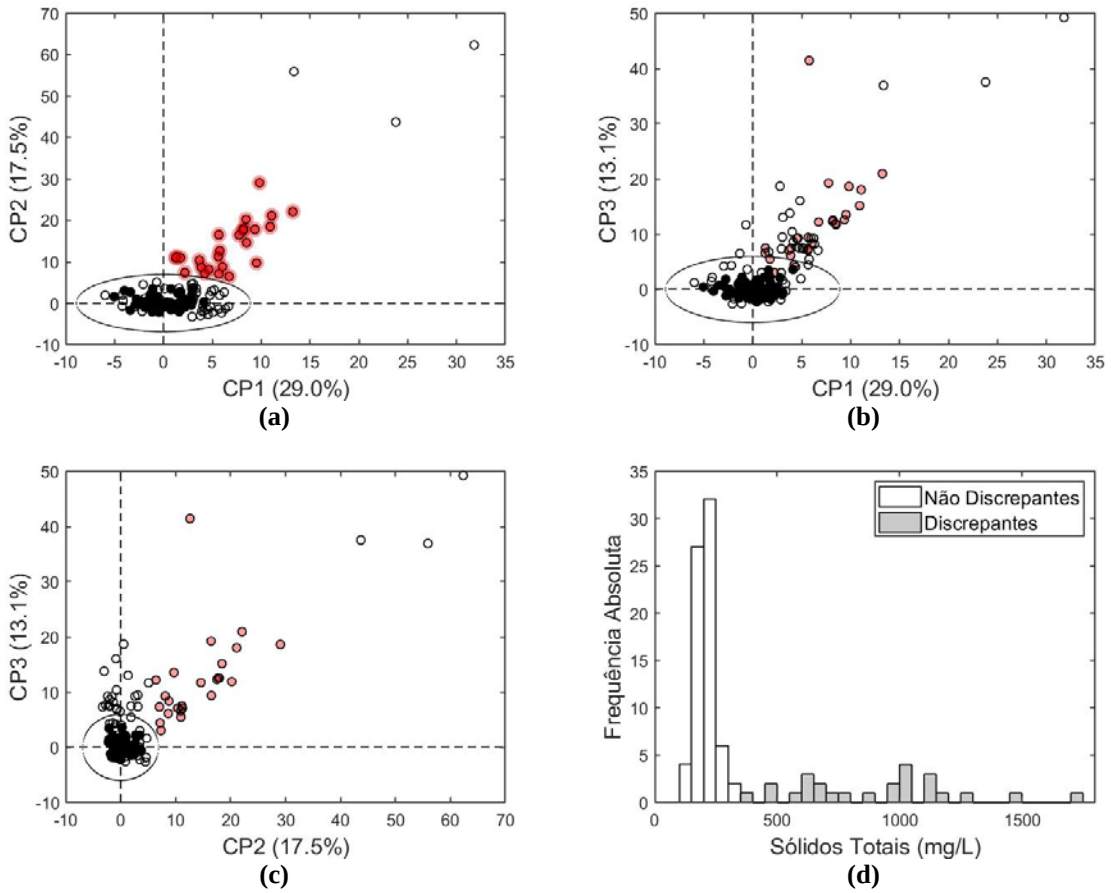

Figura 6: Gráficos de *scores*: (a) PC_1-PC_2, (b) PC_1-PC_3, e (c) PC_2-PC_3, e (d) histograma para sólidos totais.

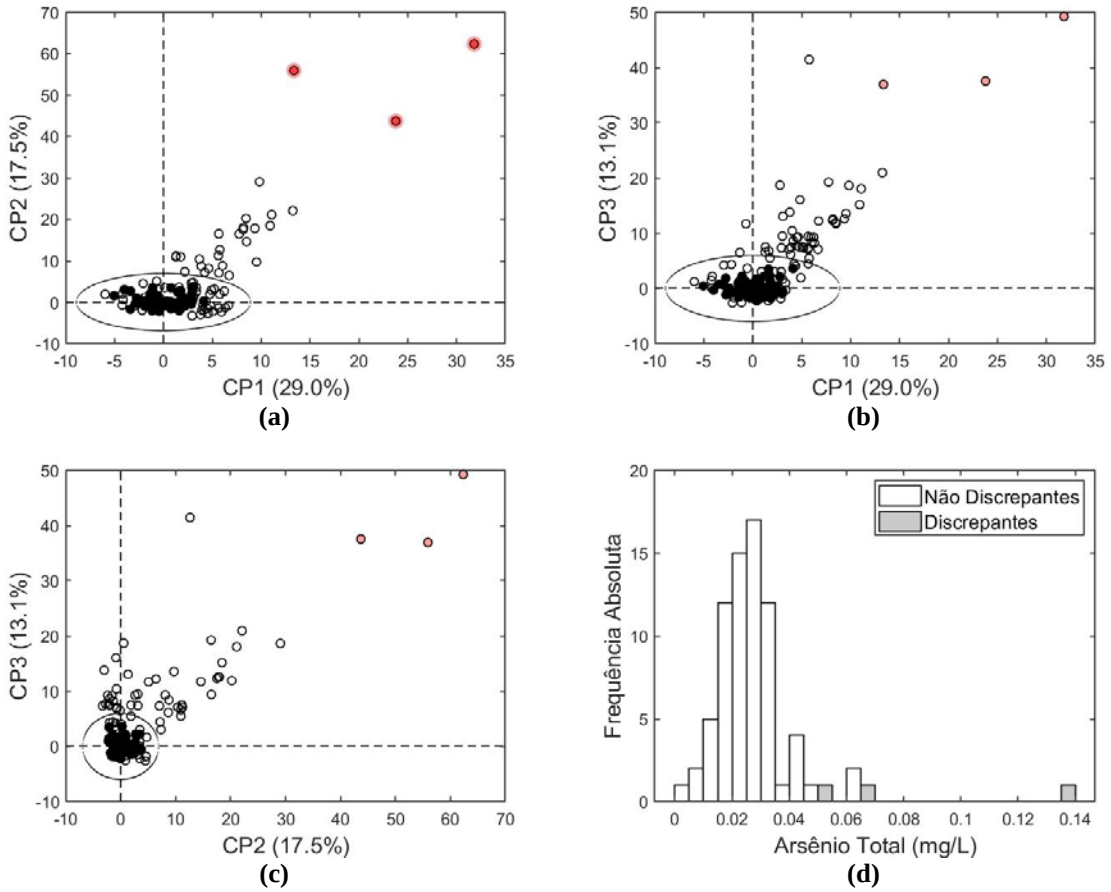

Figura 7: Gráficos de *scores*: (a) PC_1-PC_2, (b) PC_1-PC_3, e (c) PC_2-PC_3, e (d) histograma para arsênio total.

Os fatores de pressão ambiental neste ponto do rio das Velhas corroboram com as caracterizações das regiões (1 a 3) descritas pelo modelo PCA, o que é fundamental para a sua validação. A estação de monitoramento BV137 recebe influência dos lançamentos dos esgotos domésticos do município de Lagoa Santa (IGAM, 2010) e de efluentes industriais, principalmente de indústrias têxteis e de papel. As altas concentrações de arsênio, acima do valor permitido pela legislação, são consequência do beneficiamento de minério de ouro, que promove a disponibilização desse metal ao longo do corpo de água. De acordo com IGAM (2013), as fontes de arsênio na bacia do rio das Velhas concentram-se, principalmente, em seu alto curso, região de Nova Lima, onde se encontram fontes naturais; já no médio e baixo cursos são provenientes dos sedimentos, onde parte do arsênio vem sendo depositado ao longo dos anos.

A vantagem com esse tipo de análise é o monitoramento a partir de dois ou três gráficos, ao invés, por exemplo, de se ter um gráfico para cada parâmetro (nesse caso, doze), como comumente é realizado. Outra questão é quanto à consideração de correlação espacial, ao se ter uma análise conjunta dos parâmetros, o que é geralmente desconsiderado. Além dos valores, o tipo de correlação também é importante para o diagnóstico de uma situação. A visualização e a correlação de parâmetros através de análises individuais é uma tarefa não intuitiva ao ser humano, e, portanto, árdua.

CONCLUSÕES

Os bancos de dados gerados com os Programas de Monitoramento da Qualidade das Águas (PMQAs) constituem-se numa rica fonte de informação. A desconsideração de um número significativo de parâmetros altamente correlacionados entre si pode comprometer a análise de dados, o alcance de um processo de tomada de decisão mais efetivo. Nesse contexto, a técnica estatística multivariada, denominada Análise por Componentes Principais (PCA), é uma ferramenta útil a ser empregada em sistemas de suporte à decisão ao

monitoramento de qualidade de água. Neste estudo, explorou-se o uso de PCA, através de seus *scores*, para o monitoramento simultâneo de doze parâmetros de qualidade da água de uma estação da calha do rio das Velhas (BV137). Observou-se, ao final, um desempenho satisfatório, com a identificação de observações discrepantes e a associação de regiões espaciais nos gráficos de *scores* com eventos críticos em relação à qualidade da água. Essas informações podem servir como sinal de alerta sobre desvios em relação à condição usual de qualidade de água, e para diagnóstico, ao se estabelecer, *a priori*, associações entre essas regiões espaciais e eventos não usuais.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BEHMEL, S., DAMOUR, M., LUDWIG, R.Water quality monitoring strategies - A review and future perspectives. Science of The Total Environment, v. 571, p. 1312-1329, 2016.
2. BHAT, S.A., MERAJ, G., YASEEN, S., PANDIT, A.K. Statistical assessment of water quality parameters for pollution source identification in Suknag stream: An inflow stream of lake Wular (Ramsar Site), Kashmir Himalaya. Journal of Ecosystems, p. 1-18, 2014.
3. BRO, R., SMILDE, A. K. Principal component analysis. Anal. Methods, v. 6, p. 2812–2831, 2014.
4. CALAZANS, G.M. Avaliação e proposta de adequação da rede de monitoramento da qualidade das águas superficiais das sub-bacias do rio das Velhas e do rio Paraopeba, utilizando técnicas estatísticas multivariadas. Belo Horizonte, 2015. Dissertação de mestrado-Escola de Engenharia, Universidade Federal de Minas Gerais, 2015.
5. CHEN, Q., WU, W., BLANCKAERT, K., MA, J., HUANG, G. Optimization of water quality monitoring network in a large river by combining measurements, a numerical model and matter-element analyses. Journal of Environmental Management v. 110, p. 116-124, 2012.
6. CHRISTOFARO, C., LEÃO, M.M.D. Caracterização temporal do arsênio nos cursos d'água da bacia hidrográfica do Rio das Velhas, MG, Brasil, ao longo de uma década (1998-2007). Ambiente e Água - An Interdisciplinar Journal of Applied Science, Universidade de Taubaté, v. 4, n. 3, p. 54-66, 2009.
7. CORBETT, C.J., PAN, J.N. Evaluating environmental performance using statistical process control techniques. European Journal of Operational Research, v. 139, n. 1, p. 68-83, 2002.
8. COSTA, E.P., PINTO, C.C., SOARES, A.L.C., MELO, L.D.V., OLIVEIRA, S.C. Evaluation of Violations in Water Quality Standards in the Monitoring Network of São Francisco River Basin, the Third Largest in Brazil, Environmental Monitoring and Assessment, v. 189, p. 2-16, 2017.
9. DAVIES, L., GATHER, U. The identification of multiple outliers. Journal of the American Statistical Association, v. 88, p. 782–801, 1981.
10. FOLLADOR, F.A.C., VILAS BOAS, M.A., MALMANN, L., SCHOENHALS, M., VILLWOCK, R. Controle de Qualidade da Água Medido Através de Cartas de Controle de Shewhart, CUSUM e MMEP. Engenharia Ambiental - Espírito Santo do Pinhal, v. 9, n. 3, p. 183-197, jul /set. 2012.
11. GE, Z.; SONG, Z. Multivariate statistical process control: Process monitoring methods and applications. London: Springer Science & Business Media, 2013.
12. HAIR, J. F. JR., ANDERSON, R. E., TATHAM, R. L., BLACK, W. Análise Multivariada de dados. 6 ed. Porto Alegre: Bookman, 2009. 688 p.
13. LIN, B., RECKE, B., KNUDSEN, J.K.H., JØRGENSEN, S.B. A systematic approach for soft sensor development. Computers and Chemical Engineering, Denmark, n. 31(5-6), p. 419-425, 2006.
14. IGLESIAS, C., SANCHO, J., PIÑEIRO, J.I., MARTÍNEZ, J., PASTOR, J.J., TABOADA, J. Shewhart-type control charts and functional data analysis for water quality analysis based on a global indicator. Desalination and Water Treatment, p. 1–16, 2015.
15. KANNEL, P.R., LEE, S., KANEL, S.R., KHAN, S.P. Chemometric application in classification and assessment of monitoring locations of an urban river system. Analytica Chimica Acta, n. 582, p. 390-399, 2007.
16. KOWLKOWSKI, T., ZBYTNIEWSKI, R., SZPEJNA, J., BUSZEWSKI, B. Application of chemometrics in water classification. Water Research, v. 40, n. 4, p. 744-752, 2006.
17. MANLY, B.F.J. Multivariate Statistical Methods: a Primer. Second ed. Chapman and Hall/CRC, 2000.
18. MENDIGUCHÍA, C., MORENO, C., GALINDO-RIANO, M.D., GARCÍA-VARGAS, M. Using chemometric tools to assess antropogenic effects in river water a case study: Guadalquivir River (Spain). Analytica Chimica Acta, v. 515, p. 143-9, 2004.
19. OLSEN, R. L., CHAPPELL, R. W., LOFTIS, J. C. Water quality sample collection, data treatment and results presentation for principal components analysis – literature review and Illinois River watershed case study. Water Research, v. 46, p. 3110-3122, 2012.

20. OUYANG, Y. Evaluation of river water quality monitoring stations by principal component analysis. Water Research, v. 39, p. 2621-2635, 2005.
21. PLANO DIRETOR DE RECURSOS HÍDRICOS DA BACIA HIDROGRÁFICA DO RIO DAS VELHAS 2015 – PDRH do Rio das Velhas: Resumo Executivo. Comitê da Bacia Hidrográfica do Rio das Velhas. Belo Horizonte, 2015. 233 p.
22. RUSSELL, E.L., CHIANG, L.H., BRAATZ, R.D. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. Chemometrics and inteligente laboratory systems, v. 51, n. 1, p. 81-93, 2000.
23. SAMSUDIN, M.S., AZID, A., KHALIT, S.I., SAUDI, A.S.M., ZAUDI, M.A. River water quality assessment using APCS-MLR and statistical process control in Johor River Basin, Malaysia. International Journal of Advanced and Applied Sciences, v. 4, n. 8, p. 84-97, 2017.
24. SANCHO, J., IGLESIAS, C., PIÑEIRO, J., MARTÍNEZ, J., PASTOR, J. J., ARAÚJO, M., TABOADA, J. Study of Water Quality in a Spanish River Based on Statistical Process Control and Functional Data Analysis. Math Geosci, v. 48, p. 163-186, 2016.
25. SERGEANT, C. J., STARKEY, E. N., BARTZ, K. K., WILSON, M. H., MUETER, F. J. A practitioner's guide for exploring water quality patterns using principal components analysis and Procrustes. Environ Monit Assess, p. 188-249, 2016.
26. SHARMA, S. Applied multivariate techniques. John Wiley& Sons, Inc., New York, p. 509 1996.
27. SIMEONOV, V., STRATIS, J.A., SAMARA, C., ZACHARIADIS, G., VOUTSA, D., ANTHEMIDIS, A., SOFONIOU, M., KOUIMTZIS, T. Assessment of the surface water quality in northen Greece. Water Research, v. 37, n. 17, p. 4119-4124, 2003.
28. SINGH, K.P., MALIK, A., MOHAN, D., SINHA, S. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) – a case study. Water Research, v. 38, n. 18, p. 3980-3992, 2004.
29. TRINDADE, A.L.C., ALMEIDA, K.C.B., BARBOSA, P.E., OLIVEIRA, S.M.A.C. Tendências temporais e espaciais da qualidade das águas superficiais da sub-bacia do Rio das Velhas, estado de Minas Gerais. Engenharia Sanitária e Ambiental, v. 22, p. 13-24, 2017.
30. ZHANG, X., WANG, Q., LIU, Y., WU, J., YU, M. Application of multivariate statistical techniques in the assessment of water quality in the southwest new territories and Kowloon, Hong Kong. Environmental Monitoring and Assessment, v. 173, p. 17-27, 2011.