DOI: 10.1590/S1413-41522014019010000588

Artigo Técnico

Uso de métodos estatísticos robustos na análise ambiental

Use of robust statistical methods to analyze environmental data

Claudia Vilhena Schayer Sabino[1], Ludmila Vieira Lage[2], Katiane Cristina de Brito Almeida[3]

RESUMO

Dados ambientais frequentemente apresentam valores censurados, perdidos e/ou discrepantes (*outliers*). Além disto, as amostras devem ser consideradas dependentes por terem componentes espaciais e temporais. Outro fato frequente nestes dados é que dificilmente seguem uma distribuição Normal ou Log-normal. Devido a estas características e outras, técnicas estatísticas convencionais não devem ser utilizadas. O presente trabalho apresenta um estudo de caso do rio das Velhas, Minas Gerais, utilizando métodos estatísticos robustos após o tratamento adequado dos dados. A análise de componentes principais detectou as variáveis que mais contribuem para a degradação da qualidade das águas do rio das Velhas e a visualização espacial dos escores mostrou onde esta contaminação está mais evidente.

Palavras-chave: Estatísticas robustas; dados ambientais; análise de componentes principais; *Software* R; rio das Velhas.

ABSTRACT

Environmental data often presents censored, lost and/or outlier values. In addition, samples should be considered dependent for having spatial and temporal components. Another fact is that, frequently, these data won't follow a Normal or Log-normal distribution. Because of these and other characteristics, conventional statistical techniques should not be used. This article presents a case study of the Das Velhas river, Minas Gerais, using robust statistical methods after appropriate treatment of the data. The analysis of the main components found the variables that contribute the most for the degradation of water quality in the river, and the spatial visualization of the scores showed where this contamination is most evident.

Keywords: robust statistics; environmental data; principal components analysis; R software; Das Velhas river.

INTRODUÇÃO

A história da ocupação de Minas Gerais tem relação direta com o rio das Velhas e a sua degradação. A exploração começou com a descoberta e a extração de ouro e pedras preciosas e, posteriormente, do minério de ferro, seguido por um ciclo de industrialização e urbanização desordenada com a consolidação da Região Metropolitana de Belo Horizonte (RMBH).

Em 2004, o Governo do Estado de Minas Gerais implantou o Projeto estruturador de Revitalização da Bacia Hidrográfica do Rio das Velhas, que ficou conhecido como Meta 2010 (pescar, nadar e navegar no rio das Velhas em 2010). Desde sua implantação até o final de 2010, várias ações foram realizadas por diversos órgãos governamentais e da sociedade civil no sentido de revitalizar o rio. Destacam-se a implantação das Estações de Tratamento de Esgoto (ETE) do Ribeirão Onça e a ampliação da ETE Arrudas, ambas com a previsão de grande impacto na melhoria da qualidade da água do rio das Velhas. Os objetivos da Meta 2010 foram parcialmente e satisfatoriamente atingidos e, para dar continuidade aos trabalhos, a meta foi repactuada pelo Governo com o objetivo de consolidar a volta dos peixes e nadar no rio das Velhas na RMBH em 2014.

O Programa de monitoramento dos recursos hídricos no Estado de Minas Gerais é realizado pelo Instituto Mineiro de Gestão das Águas (IGAM). Em execução desde 1997, o monitoramento destaca-se por permitir a avaliação e o acompanhamento da condição de qualidade das águas nas principais bacias hidrográficas do Estado, possibilitando ao Sistema Estadual de Meio Ambiente e Recursos Hídricos de Minas Gerais e aos órgãos e entidades vinculados identificarem e implementarem estratégias de aperfeiçoamento de seus instrumentos gerenciais.

Trabalho realizado na Pontifícia Universidade Católica de Minas Gerais (PUC-MG) – Belo Horizonte (MG), Brasil.
[1]Doutora em Química pela Universidade Federal de Minas Gerais (UFMG); Professora do Programa de Pós-Graduação em Ensino de Ciências e Matemática da PUC-MG – Belo Horizonte (MG), Brasil.
[2]Estatística pela UFMG; Assistente de Pesquisa da PUC-MG – Belo Horizonte (MG), Brasil.
[3]Mestre em Engenharia Sanitária e Ambiental pela UFMG; Gerente de Monitoramento da Qualidade das Águas, Instituto Mineiro de Gestão das Águas (IGAM) – Belo Horizonte (MG), Brasil.
Endereço para correspondência: Claudia Vilhena Schayer Sabino – Avenida Dom José Gaspar, 500 – Coração Eucarístico – 30535-901 – Belo Horizonte (MG) - Brasil. – E-mail: sabinoc@pucminas.br
Recebido: 09/10/12 – **Aceito:** 17/05/13 – **Reg. ABES:** 588

Visando complementar a série histórica de dados do rio das Velhas, o IGAM adotou, a partir de 2008, uma freqüência de monitoramento mensal. Essas análises na calha principal do rio das Velhas servem de base para avaliar a efetividade das diversas ações realizadas na bacia, bem como proporcionar um estudo estatístico mais confiável.

A estatística, como parte da matemática aplicada, estuda os mais variados fenômenos das diversas áreas do conhecimento e representa um valioso instrumento de trabalho nos dias de hoje. Na área ambiental, destaca-se a análise multivariada. Os métodos multivariados são modelos estatísticos que consideram muitas variáveis ao mesmo tempo, o que demanda um exame detalhado e rigoroso dos dados, pois o tratamento inadequado pode ter efeitos "catastróficos" (SARTORI, 2008).

Dados provenientes de resultados da análise de amostras ambientais apresentam variáveis espaciais e temporais como, por exemplo, as coordenadas e a data de amostragem. Logo, pode-se dizer que há dependência entre amostras, pois pontos próximos têm mais chance de apresentar resultados semelhantes que pontos espacialmente distantes. Esses resultados também apresentam elevada variabilidade devido a variações sazonais e à influência de mudanças da vazão sobre as propriedades físico-químicas das variáveis (KUPPUSAMY & GIRIDHAR, 2006). Outro fato a ser considerado é que dados ambientais apresentam imprecisões relacionadas à amostragem, preparo e análise. Além disso, em muitos casos, são censurados pelo limite de detecção do método analítico.

Nos resultados de tais análises, é comum aparecerem valores discrepantes (*outliers*) que têm impacto expressivo na interpretação de análises estatísticas, causando distorções. Assim, métodos estatísticos que não dependem da distribuição dos dados ou da presença de *outliers*, como os métodos robustos, devem ser utilizados (REIMANN *et al.,* 2008).

O uso de técnicas robustas em dados ambientais já é uma metodologia consolidada internacionalmente há pelo menos uma década. No entanto, no Brasil, prevalece o uso de técnicas convencionais, não robustas.

O presente estudo teve como objetivo aplicar técnicas modernas para tratamento adequado dos dados disponíveis para o rio das Velhas, e, em seguida, utilizar métodos multivariados robustos. Foram utilizados os resultados das análises físico-químicas e biológicas do rio das Velhas entre os anos 1997 e 2010.

METODOLOGIA

Área de estudo

O rio das Velhas (Figura 1) é um dos principais afluentes do rio São Francisco e tem sua nascente dentro do Parque Municipal das Andorinhas, município de Ouro Preto, e deságua no rio São Francisco, na Barra do Guaicuí, município de Várzea da Palma, situado à margem direita do São Francisco, percorrendo aproximadamente 801 km e drenando uma bacia de 29.173 km^2.

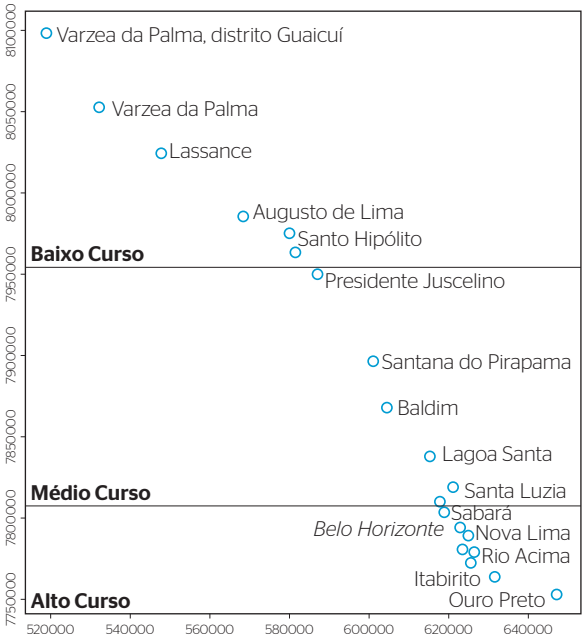


Figura 1 – Localização da área em estudo no rio das Velhas, indicando os munícipios em que há amostragem de dados.

O rio das Velhas é dividido em trechos, segundo os cursos alto, médio e baixo (GUIMARÃES, 1953). O Alto rio das Velhas compreende toda a região denominada Quadrilátero Ferrífero e apresenta o maior contingente populacional, com uma expressiva atividade econômica, concentrada principalmente na RMBH, onde estão presentes os maiores focos de poluição de toda a bacia. Os principais agentes poluidores são os esgotos industriais e domésticos e os efluentes gerados pelas atividades minerárias. Os trechos Médio e Baixo rio das Velhas possuem características diferenciadas em relação ao uso e ocupação do solo do alto trecho, apresentando uma menor concentração populacional, com o predomínio das atividades agrícolas e pecuárias. Essas atividades contribuem para processos de erosão na região, pois há um grande percentual de área mecanizada. As atividades agrícolas comprometem a qualidade da água devido à utilização de insumos (IGAM, 2010).

Características dos dados

Atualmente, o IGAM possui em sua rede de monitoramento qualitativo vinte e três (23) estações de amostragem localizadas ao longo do rio das Velhas. Para avaliar o grau de contaminação das águas, são analisadas 56 variáveis (Quadro 1). As coletas e respectivas análises físico-químicas e bacteriológicas das amostras de água são realizadas pela Fundação Centro Tecnológico de Minas Gerais (CETEC). A frequência das análises variou entre mensal e trimensal, ou menor para

algumas variáveis de ocorrência pouco comum ou que apresentaram pequena variabilidade anual. Ao todo, tem-se 1.378 coletas realizadas no rio das Velhas, entre os anos 1997 e 2010.

Pré-tratamento dos dados e análise estatística

Foram avaliadas as variáveis monitoradas em relação ao percentual de amostras cujos valores violaram os limites legais da Deliberação Normativa COPAM/CERH Nº01/08 considerando o enquadramento do corpo de água no local de cada estação, a fim de se verificar as principais interferências das atividades predominantes na bacia e se estas variáveis são convenientemente consideradas no cálculo estatístico.

Para os estudos estatísticos, foi utilizado o *software* R, que é gratuito e possui técnicas robustas para análise de dados. Neste trabalho, foram também utilizados os pacotes *StatDA*, *rrcov* e *robustbase,* que possuem as mesmas características.

Uma das primeiras etapas das análises estatísticas ambientais deve ser o estudo cuidadoso da distribuição das variáveis. No presente trabalho, realizou-se o teste Shapiro-Wilks (SW), considerando um nível de 5% significância. O teste SW baseia-se nos valores amostrais ordenados elevados ao quadrado.

Uma dificuldade usual de dados ambientais é que os resultados apresentam, para muitas amostras, valores censurados, ou seja, abaixo do limite mínimo de detecção do método analítico ou acima do limite máximo de detecção (LD). Gráficos, cálculos e mapas ficam distorcidos se usarem os valores <LD, sendo necessária a transformação destes dados. Os valores <LD foram substituídos pela sua metade do LD e os valores dos limites máximos de detecção foram mantidos. Outra dificuldade ao se analisar dados ambientais são os valores perdidos. Isto pode ocorrer por inúmeras razões, como volume de amostra insuficiente para todas as análises, erro na transcrição de resultados, dentre outros (MINGOTI, 2005). Ao contrário do valor <LD, que informa que a variável tem um valor baixo, o dado perdido não traz informação, não sendo possível substituí-lo.

De acordo com o sugerido por Reimman *et al.*, 2008, optou-se por excluir as variáveis (toda a coluna) que apresentaram 85% ou mais de dados censurados e/ou 50% ou mais de valores perdidos. Variáveis com muitos valores censurados ou perdidos apresentam geralmente valores de desvio-padrão da variância e desvio absoluto mediano (MAD) com valor zero, o que significa que tais variáveis não apresentam variabilidade suficiente para serem utilizadas em cálculos estatísticos. O MAD é calculado pela Equação 1:

$$MAD = 1.4256 * mediana_i * (|mediana - x_i|) \quad (1)$$

Antes de aplicar algum método multivariado, deve-se investigar a existência de *outliers*, que podem afetar os resultados finais da análise estatística. Em dados multidimensionais, uma observação é considerada *outlier*

Quadro 1 – Variáveis analisadas nas águas do rio das Velhas e sua unidade de medida.

Alcalinidade de bicarbonato ($mg.L^{-1}$)
Alcalinidade total ($mg.L^{-1}$)
Alumínio dissolvido ($mg.L^{-1}$)
Amônia não ionizável ($mg.L^{-1}$ NH_3)
Arsênio total ($mg.L^{-1}$)
Bário total ($mg.L^{-1}$)
Boro dissolvido ($mg.L^{-1}$)
Boro total ($mg.L^{-1}$)
Cádmio total ($mg.L^{-1}$)
Cálcio total ($mg.L^{-1}$)
Chumbo total ($mg.L^{-1}$)
Cianeto ($mg.L^{-1}$)
Cloreto total ($mg.L^{-1}$)
Clorofila a ($\mu g.L^{-1}$)
Cobre dissolvido ($mg.L^{-1}$)
Cobre total ($mg.L^{-1}$)
Coliformes termotolerantes ($NMP.100mL^{-1}$)
Coliformes totais ($NMP.100mL^{-1}$)
Condutividade elétrica ($\mu mho.cm^{-1}$)
Cor verdadeira ($mg\ Pt.L^{-1}$)
Cromo total ($mg.L^{-1}$)
Demanda Bioquímica Oxigênio ($mg.L^{-1}$)
Demanda Química de Oxigênio ($mg.L^{-1}$)
Densidade de cianobactérias ($cel.mL^{-1}$)
Dureza de cálcio ($mg.L^{-1}$)
Dureza de magnésio ($mg.L^{-1}$)
Dureza total ($mg.L^{-1}$)
Estreptococos fecais ($NMP.100mL^{-1}$)
Fenóis totais ($mg.L^{-1}$)
Feoftina ($\mu g.L^{-1}$)
Ferro dissolvido ($mg.L^{-1}$)
Fósforo total ($mg.L^{-1}$)
Magnésio total ($mg.L^{-1}$)
Manganês total ($mg.L^{-1}$)
Mercúrio total ($\mu g.L^{-1}$)
Níquel total ($mg.L^{-1}$)
Nitrato ($mg.L^{-1}$ N)
Nitrito ($mg.L^{-1}$ N)
Nitrogênio amoniacal total ($mg.L^{-1}$ N)
Nitrogênio orgânico ($mg.L^{-1}$ N)
Óleos e graxas ($mg.L^{-1}$)
Oxigênio dissolvido ($mg.L^{-1}$)
pH
Potássio dissolvido ($mg.L^{-1}$)
Selênio total ($mg.L^{-1}$)
Sódio dissolvido ($mg.L^{-1}$)
Sólidos dissolvidos totais ($mg.L^{-1}$)
Sólidos suspensão totais ($mg.L^{-1}$)
Sólidos totais ($mg.L^{-1}$)
Substâncias tensoativas ($mg.L^{-1}$)
Sulfato total ($mg.L^{-1}$)
Sulfeto ($mg.L^{-1}$)
Temperatura da água (ºC)
Temperatura do ar (ºC)
Turbidez (UNT)
Zinco total ($mg.L^{-1}$)

se apresentar valores extremos na distribuição multivariada e não apenas em uma ou outra variável. Com o intuito de detectar dados discrepantes, foi feita uma comparação multivariada entre a distância de Mahalanobis e a distância robusta (TODOROV & FILZMOSER, 2009). Após a identificação das observações atípicas, foram excluídas as que mostram verdadeira discrepância em comparação com o restante dos dados estudados.

Outro método aplicado foi a Análise de Correlação, através do qual é avaliado o grau de relacionamento entre duas variáveis quantitativas. O coeficiente de correlação pode ser utilizado para dados paramétricos e para dados não paramétricos. No entanto, esses testes são dependentes da presença de *outliers*, em maior (Pearson) ou menor extensão (Spearman) ou apresentam difícil visualização gráfica dos resultados (Kendall). Dessa maneira, optou-se por utilizar a correlação robusta, que é construída por meio do estimador da covariância robusta e é independente da presença de *outliers,* sendo ideais para cálculo de correlações de variáveis ambientais (REIMANN *et al.,* 2008).

Na análise por componentes principais (PCA), não foram utilizadas todas as variáveis correlacionadas, com o objetivo de aumentar o número dos graus de liberdade. Assim, apenas uma das variáveis correlacionadas foi escolhida para representar o conjunto.

Segundo Reimann *et al.* (2008), o cálculo da dimensão ideal de amostras para a PCA é dado pela Equação 2:

$$n \rangle p^2 + 3p + 1 \quad (2)$$

onde n é o número de observações (linha) e p é o número de variáveis (coluna) a serem analisadas.

A Análise de Componentes Principais é a técnica multivariada mais utilizada para explorar, interpretar e reduzir os dados, sem que haja perda de informação. Foi uma das primeiras desenvolvidas com métodos robustos. Neste caso, os autovalores, autovetores e matrizes de correlação e covariância são determinados por cálculos robustos não sujeitos à influência de *outliers*. As componentes principais (CP) obtidas constituem as novas variáveis respostas e são utilizadas nas análises subsequentes do estudo. A interpretação de cada CP é baseada nas variáveis que mais contribuem para a CP.

Para o cálculo de componentes principais robustas, foi feita a transformação log-centralizada (clr), o que visou equiparar a ordem de grandeza das diferentes variáveis, segundo a Equação 3:

$$clr\,(x_{ij}) = \log\left(\frac{x_{ij}}{G}\right) \quad (3)$$

onde x_{ij} é a amostra i da variável j, e G é a média geométrica de todas as variáveis.

O gráfico Scree-Plot, que representa a porcentagem de variância explicada por componente, foi utilizado para determinar quantas componentes deveriam ser utilizadas na análise multivariada.

Uma vez determinadas as componentes principais, os seus valores numéricos, denominados escores, foram calculados para cada elemento amostral. A distribuição espacial dos escores possibilitou localizar diferenças na contaminação entre as regiões do rio das Velhas.

O teste Mann-Whitney, aplicado aos escores de cada CP, visou determinar a eficácia das ações realizadas na bacia. Este teste, não paramétrico, compara tendências centrais de duas amostras independentes. O nível de significância adotado para o teste foi de 5%.

RESULTADOS E DISCUSSÃO

A primeira tarefa do estudo foi determinar as variáveis que apresentaram maiores números de desconformidade em relação à Deliberação Normativa COPAM/CERH Nº01/08 (Figura 2). De acordo com o IGAM (2010), os principais fatores de degradação ambiental que podem ser apontados como contribuintes dos resultados apresentados na Figura 2 são os lançamentos de esgotos domésticos nos corpos de água, as atividades minerárias, além de outras formas de uso ou presença natural nos solos da bacia de drenagem que podem afetar a qualidade da água. O objetivo desta etapa foi ao final verificar se tais variáveis são também estatisticamente significativas.

Em seguida, foi feita a análise das estatísticas descritivas e verificou-se que as variáveis chumbo total, cromo total, níquel total, óleos e graxas e substâncias tensoativas apresentaram MAD igual a zero e, por isso, também foram retiradas da tabela de resultados.

Após as exclusões, restaram as seguintes variáveis: arsênio total, cloreto total, coliformes termotolerantes, condutividade elétrica, demanda bioquímica de oxigênio, ferro dissolvido, fósforo total, manganês total, nitrato, nitrogênio amoniacal total, oxigênio dissolvido, pH, sólidos dissolvidos totais, sólidos em suspensão totais, sólidos totais, temperatura da água, temperatura do ar e turbidez.

As variáveis cloreto total, pH, sólidos dissolvidos totais, sólidos em suspensão totais, sólido totais, temperatura da água e turbidez mostraram-se altamente correlacionadas com outras variáveis (coeficiente de correlação robusta maior que 0,7) e foram excluídas.

Ao final do tratamento dos dados, restaram onze variáveis (Quadro 2), o que representa 19,6% daquelas do início do estudo. Essas foram utilizadas para a análise de componentes principais.

Após a eliminação de variáveis, foram excluídas coletas que possuíam dados perdidos e/ou *outliers*. Das 1.378 coletas iniciais, ficaram apenas 573 (42% do total de coletas iniciais), o que, de acordo com Reimann *et al.* (2008), representa ainda uma margem segura de dados para se realizar uma análise estável.

A avaliação do gráfico *Scree-plot* indicou que as quatro primeiras componentes são suficientes para explicar 77% da variação

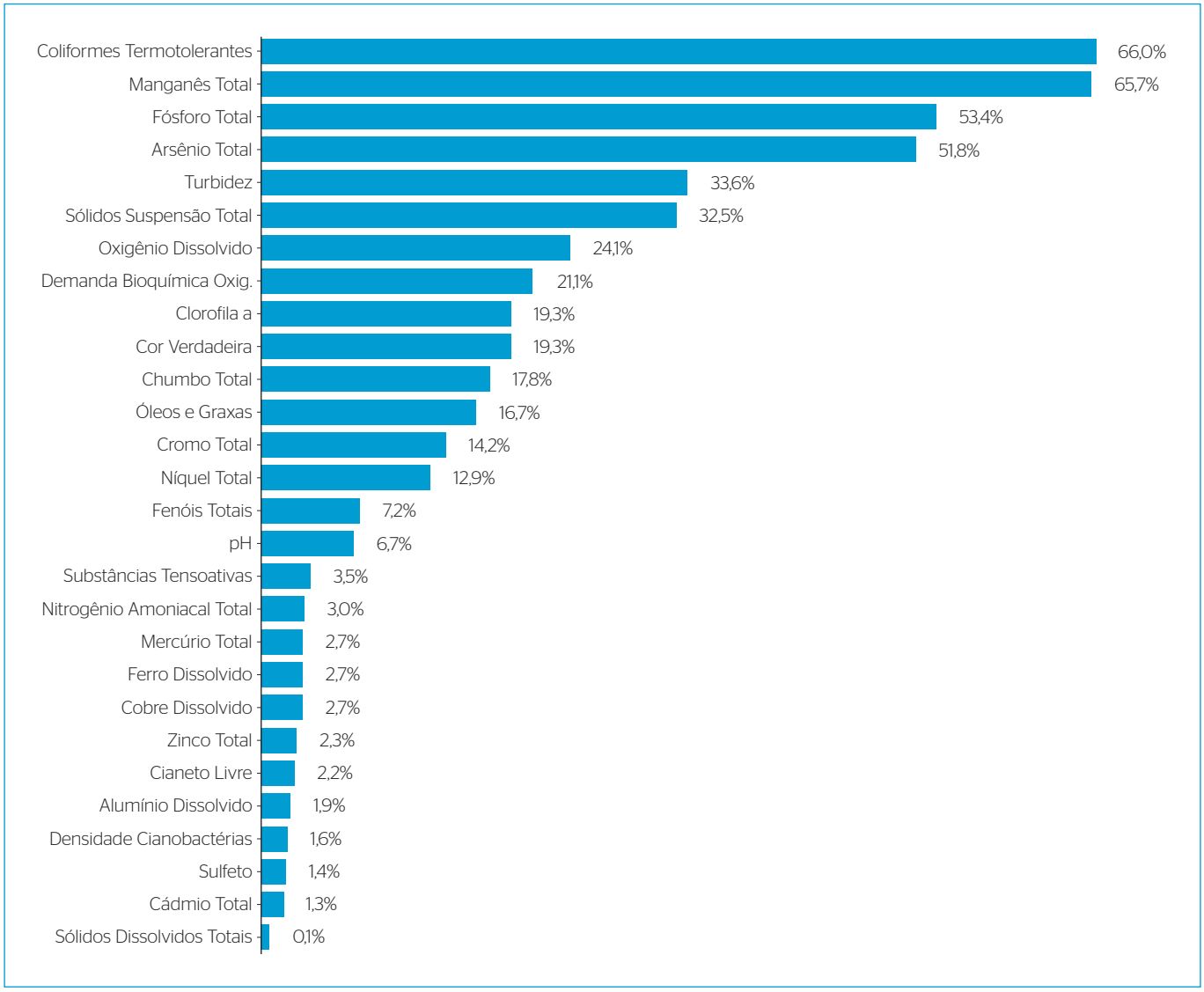


Figura 2 – Variáveis com percentual de não atendimento à legislação DN 01/08, 1997 a 2010.

total dos dados, ou seja, elas representam os aspectos principais da qualidade das águas superficiais do rio das Velhas. Ressalta-se que todas essas componentes apresentam autovalores superiores a 1, valor recomendado por Reid *et al.*, 2009, como critério para inclusão da CP na análise.

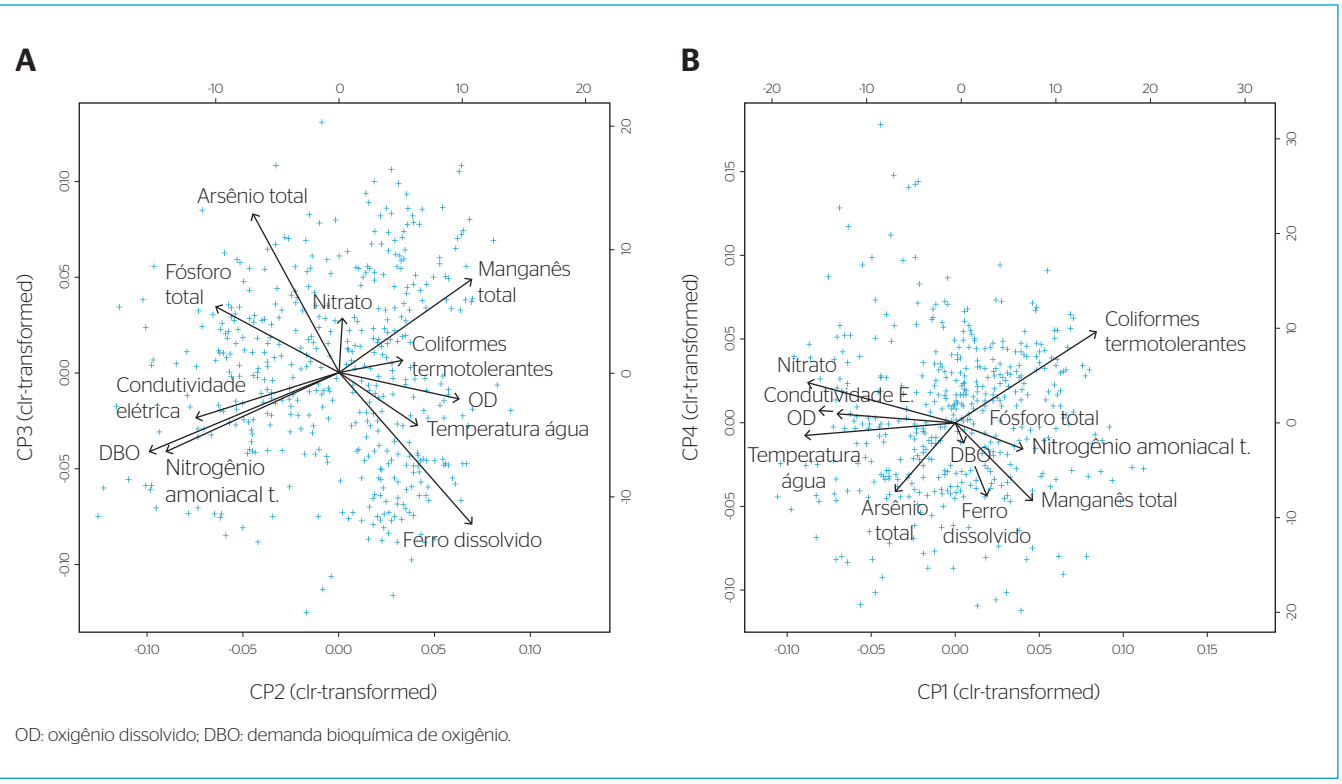
A Tabela 1 apresenta as cargas (pesos) das variáveis em cada componente principal. Na Figura 3, são apresentados gráficos *biplot* de escores (pontos identificando as coletas) e pesos (setas) das componentes principais. Quanto mais paralelo é o vetor de peso ao eixo da componente principal, maior é a importância da variável correspondente. Como exemplificação das interpretações das componentes principais, foi feito o mapeamento dos escores de cada CP no ano de 2010 (Figura 4). O ano de 2010 foi escolhido por retratar um quadro mais atual do rio das Velhas frente às ações de revitalizações implementadas no âmbito do programa de revitalização da bacia.

Quadro 2 – Variáveis utilizadas na análise de componentes principais.

Arsênio total
Coliformes termotolerantes
Condutividade elétrica
Demanda bioquímica de oxigênio
Ferro dissolvido
Fósforo total
Manganês total
Nitrato
Nitrogênio amoniacal total
Oxigênio dissolvido
Temperatura da água

Tabela 1 – Carga das variáveis em cada componente principal.

Variável	CP1	CP2	CP3	CP4
Arsênio total	-0.181	-0.209	0.557	-0.420
Coliformes termotolerantes	0.413	0.160	0.048	0.548
Condutividade elétrica	-0.371	-0.348	-0.156	0.064
Demanda bioquímica de oxigênio - DBO	0.024	-0.463	-0.276	-0.139
Ferro dissolvido	0.094	0.318	-0.524	-0.438
Fósforo total	0.311	-0.299	0.230	0.017
Manganês total	0.225	0.323	0.325	-0.462
Nitrogênio amoniacal total	0.200	-0.422	-0.276	-0.158
Nitrato	-0.432	0.009	0.191	0.239
Oxigênio dissolvido - OD	-0.403	0.293	-0.092	0.077
Temperatura da água	-0.445	0.190	-0.180	-0.077



OD: oxigênio dissolvido; DBO: demanda bioquímica de oxigênio.

Figura 3 – Gráfico *biplot* da CP2xCP3 (a) e CP1xCP4 (b) para os dados transformados pela log-centralização no rio das Velhas.

A CP1 explica 33% da variabilidade dos dados, estando a variável coliformes termotolerantes (carga positiva) associada aos lançamentos de esgotos na região do Alto rio das Velhas e na RMBH. As atividades agrícolas desenvolvidas no Médio e Baixo curso sugerem que o uso de fertilizantes é a principal fonte de nitrato (carga negativa) e fósforo (carga positiva) nesse trecho do rio das Velhas.

A CP2, que explica 24% da variabilidade, relaciona-se com a mineração e a presença natural de alguns elementos no solo (ferro e manganês – carga positiva), evidenciadas na qualidade das águas devido às ações antrópicas em todo o rio. Na região do médio e baixo rio, a CP2 indica a presença de matéria orgânica dissolvida (demanda bioquímica de oxigênio e nitrogênio amoniacal total – carga negativa).

Figura 4 – Distribuição Espacial dos Escores nas CPs 1, 2, 3 e 4 no ano de 2010 ao longo do rio das Velhas.

Tabela 2 – Teste Mann-Whitney comparando os escores das CPs antes e depois das ações realizadas na região do rio das Velhas.

	CP1	CP2	CP3	CP4
Estatística de teste U	38521,0	37942,0	37445,0	32868,0
Valor p	0,222	0,130	0,078	0,080

A CP3 explica 12% da variância e evidencia o impacto do arsênio (carga positiva) na mineração de ouro, a partir de Nova Lima, e do ferro (carga negativa) na mineração da região do quadrilátero ferrífero (baixo curso).

Por fim, a CP4, que explica 8% da variabilidade dos dados, sugere uma relação do arsênio, ferro e manganês (cargas negativas) com a presença natural desses elementos nos solos da região, e que é intensificada pelas atividades minerárias desenvolvidas em seu alto curso ao longo de três séculos. Já a variável coliformes termotolerantes (cargas positivas) pode estar relacionada com lançamentos de esgotos sanitários e às fontes não pontuais, como a presença de áreas de pastagens no baixo curso do rio das Velhas.

Com o intuito de verificar se houve melhora na qualidade das águas superficiais no rio das Velhas após as ações de revitalização realizadas na região, foi aplicado o teste não paramétrico Mann-Whitney aos escores de cada CP comparando os grupos "Antes das ações" (1997 a 2004) e "Depois das ações" (2005 a 2010). Com um nível de confiança de 95%, conclui-se que ainda não há diferença estatisticamente significativa entre os anos que antecederam e sucederam as ações (Tabela 2).

CONCLUSÕES

O uso da técnica multivariada robusta em dados ambientais mostrou-se realmente valiosa, uma vez que detectou com clareza as variáveis que mais contribuem para a degradação da qualidade das águas do rio das Velhas e onde esta contaminação está mais evidente.

O tratamento adequado dos resultados tem impacto expressivo na interpretação dos resultados de análises estatísticas, uma vez que esses possuem particularidades que precisam ser identificadas e compreendidas. No presente estudo, a escolha da transformação log-centralizada, a detecção multivariada dos *outliers* e a exclusão de variáveis altamente correlacionadas foi uma etapa extremamente relevante para que fossem empregadas as melhores ferramentas estatísticas. Muitas vezes, essa fase é ignorada por pesquisadores e pode deturpar toda a análise dos resultados.

Ficou evidenciado que os dados ambientais não consistem de amostras independentes, como é presumido nas técnicas estatísticas clássicas, uma vez que essas amostras estão ligadas adicionalmente por dependência espacial. Assim, o mapeamento desses dados demonstrou ser importante para a visualização e interpretação da qualidade das águas superficiais ao longo do rio das Velhas.

O teste estatístico, a um nível de significância de 5%, mostrou que ainda não há diferença significativa entre os anos que antecederam e sucederam as ações de revitalização na bacia. Apesar disso, nota-se que as ações realizadas na região podem ter representado uma melhora pontual, sobretudo nos trechos que foram beneficiados pela implantação da ETE do Ribeirão Onça e a ampliação da ETE Arrudas, mas essa melhora ainda não refletiu na qualidade do rio como um todo.

As variáveis que foram consideradas importantes na associação de cada componente estão entre aquelas que apresentaram a maior freqüência de violação dos limites ambientais. No entanto, algumas variáveis consideradas críticas na bacia, como chumbo e cromo total, não puderam entrar na análise em virtude do número insuficiente de dados, decorrentes da frequência de análise ou por não serem analisadas em todas as estações. Desse modo, recomenda-se verificar se a localização e a freqüência de análises praticada é adequada, considerado a importância dos dados de monitoramento no conhecimento e avaliação das condições da qualidade das águas.

REFERÊNCIAS

GUIMARÃES, A.P. (1953) *Paisagem física do Rio das Velhas*. Dissertação (Mestrado em Geologia). UFMG. Belo Horizonte, MG.

IGAM - Instituto Mineiro de Gestão das Águas. (2010) *Relatório Monitoramento da Qualidade das águas Superficiais da Bacia do Rio das Velhas 2009*. Belo Horizonte, MG.

KUPPUSAMY, M.R. & GIRIDHAR, V.V. (2006) Factor analysis of water quality characteristics including trace metal speciation in the coastal environmental system of Chennai Ennore. *Environmental International*, v.31, n. 2, p. 174-179.

MINGOTI, S.A. (2005) *Análise de dados através de estatística multivariada: uma abordagem aplicada*. Belo Horizonte. Ed. UFMG, 300 p.

REIMANN, C.; FILZMOSER, P.; GARRETT R.; DUTTER R. (2008) *Statistical data analysis explained. Applied environmental statistics with R*. 1 ed. Chichester. Ed. John Wiley & Sons, 362 p.

SARTORI, S.D. (2008) *Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R*. Dissertação (Mestrado em Agronomia). Escola Superior de Agricultura Luiz de Queiroz. Piracicaba, SP.

TODOROV, V. & FILZMOSER, P. (2009) An object oriented framework for robust multivariate analysis. *Journal of Statistical Software*, v. 32, n. 3, p. 1-47.