IV-070 - CONSEQÜÊNCIAS DA UTILIZAÇÃO DE MÉTODOS DE SUBSTITUIÇÃO DE VALORES CENSURADOS NOS RESULTADOS DAS ANÁLISES DE DADOS DE MONITORAMENTO AMBIENTAL

Sílvia Maria Alves Corrêa Oliveira[(1)]
Engenheira; mestre e doutora em Saneamento, Meio Ambiente e Recursos Hídricos pela Escola de Engenharia da UFMG, professora adjunta do Departamento de Engenharia Sanitária e Ambiental da UFMG.
Lenora Ludolf Gomes
Bióloga; mestre em Microbiologia pela UFMG; doutora em Saneamento, Meio Ambiente e Recursos Hídricos pela Escola de Engenharia da UFMG, professora adjunta do Departamento de Engenharia Civil e Ambiental da Universidade de Brasília.

Endereço[(1)]**:** Av. Antônio Carlos, 6627 - Bloco 1, 4º andar, sala 4525 - 30270-901 - Belo Horizonte - MG - Brasil - Tel: (31) 3409-3645 - Fax: (031) 3409-1879 - e-mail: silvia@desa.ufmg.br

RESUMO

Dados de monitoramento na área ambiental frequentemente apresentam valores censurados, ou seja, as concentrações de uma amostra podem estar muito perto de zero e, neste caso, o valor medido pode ficar abaixo do limite de detecção (LD) dos métodos analíticos. A presença de dados censurados interfere no cálculo de estatísticas descritivas, levando a estimadores enviesados de médias, variâncias e de outros parâmetros da população, além de dificultar a utilização de testes de diferenças entre grupos e desenvolvimento de modelos de regressão. A prática mais comumente adotada em estudos ambientais para tratamento de dados censurados é o da substituição destes por um valor correspondente a uma fração do limite de detecção. Para analisar as conseqüências deste tipo de substituição, este trabalho utilizou dados selecionados de uma série histórica de monitoramento trimestral da qualidade das águas do estado de Minas Gerais, efetuado pelo Instituto Mineiro de Gestão das Águas (IGAM) desde 1997. Os dados de nitrogênio amoniacal total correspondem ao período de setembro de 1997 a outubro de 2009 e foram selecionados de quatro estações de monitoramento localizadas na bacia do rio São Francisco, sub-bacia do rio das Velhas. As estações de monitoramento foram selecionadas em função do percentual de dados censurados apresentado, que foram de 24, 36, 64 e 80% dos valores medidos. O estudo concluiu que a prática de substituição dos dados censurados por qualquer valor entre zero e o limite de detecção é operacionalmente simples e pode ser adequada, em termos práticos, quando o percentual de dados censurados for baixo. Assim, para conjuntos de dados que apresentem um percentual elevado de observações abaixo do limite de detecção, a substituição dos dados censurados deve ser evitada. Para estes casos, existem alternativas que podem ser selecionadas e a escolha correta do método a ser utilizado depende tanto do grau de censura, que interfere diretamente nos resultados, quanto do tipo de aplicação (estatística descritiva, intervalos de confiança; testes de hipóteses, ajuste de distribuições de probabilidade, correlações, análises de regressão e tendências). Dependendo do método utilizado no seu tratamento, os resultados podem sofrer alterações consideráveis, tendo sua interpretação prejudicada.

PALAVRAS-CHAVE: Aumento de Capacidade, Melhoria da Qualidade, Água com Alcalinidade, Coagulante Adequado, Auxiliares de Floculação.

INTRODUÇÃO

Em algumas situações de monitoramento ambiental, as concentrações verdadeiras de uma amostra podem estar muito perto de zero e, neste caso, o valor medido pode ficar abaixo do limite de detecção (LD). Isto decorre das limitações inerentes aos métodos analíticos de mensuração e laboratórios de análise costumam se reportar a tais dados como “não detectado” ou “menor que”. Do ponto de vista estatístico, esses dados são denominados “censurados à esquerda” ou simplesmente censurados, uma vez que valores abaixo de limites de detecção não estão disponíveis para análise. Algumas vezes o dado é considerado “censurado à direita” ou “maior que”, principalmente em estudos médicos e industriais, onde o período de tempo é medido até que um evento ocorra, tal como a recorrência de uma doença ou a falha de um artigo manufaturado. Para essas observações o evento ainda não terá ocorrido quando terminar o tempo do experimento. Na área ambiental é possível encontrar tais

dados quando são efetuadas análises de dados biológicos (ex. coliformes termotolerantes, *Escherichia Coli*, etc.).

Os dados censurados interferem no cálculo de estatísticas descritivas, levando a estimadores enviesados de médias, variâncias e de outros parâmetros da população, além de dificultar a utilização de testes de diferenças entre grupos e desenvolvimento de modelos de regressão. Helsel (2005) enfatiza que, apesar dos problemas relatados, os dados censurados não devem ser eliminados da série estudada, pois nessas situações distorções ainda piores podem ser geradas. A prática mais comumente adotada em estudos ambientais para tratamento de dados censurados é o da substituição destes por um valor correspondente a uma fração do limite de detecção. Helsel (2006) comenta que a fração mais utilizada é um meio (1/2 LD), mas outros valores utilizados incluem zero, o próprio valor do limite de detecção ou ainda valores aleatórios. Christofaro (2009) comenta que a substituição por valores iguais a zero tende a produzir médias subestimadas em relação à média real, enquanto que a substituição por valores iguais ao limite de detecção tende a produzir médias superestimadas. Por distorcer também os valores do desvio-padrão das amostras, a substituição interfere em todos os testes paramétricos de hipóteses que utilizam essa estatística (Helsel, 2006). A falta de cuidado na utilização de dados censurados é evidência de um ceticismo generalizado acerca das informações contidas nessas observações. Na verdade, uma grande quantidade de informações estará disponível nos dados censurados, desde que métodos adequados para a sua extração sejam utilizados.

O objetivo deste trabalho é ilustrar, de uma maneira clara e simples, as conseqüências da substituição de medições não detectadas por um valor constante, seja abaixo ou igual ao limite de detecção, quando são calculadas estatísticas descritivas ou testes de hipóteses.

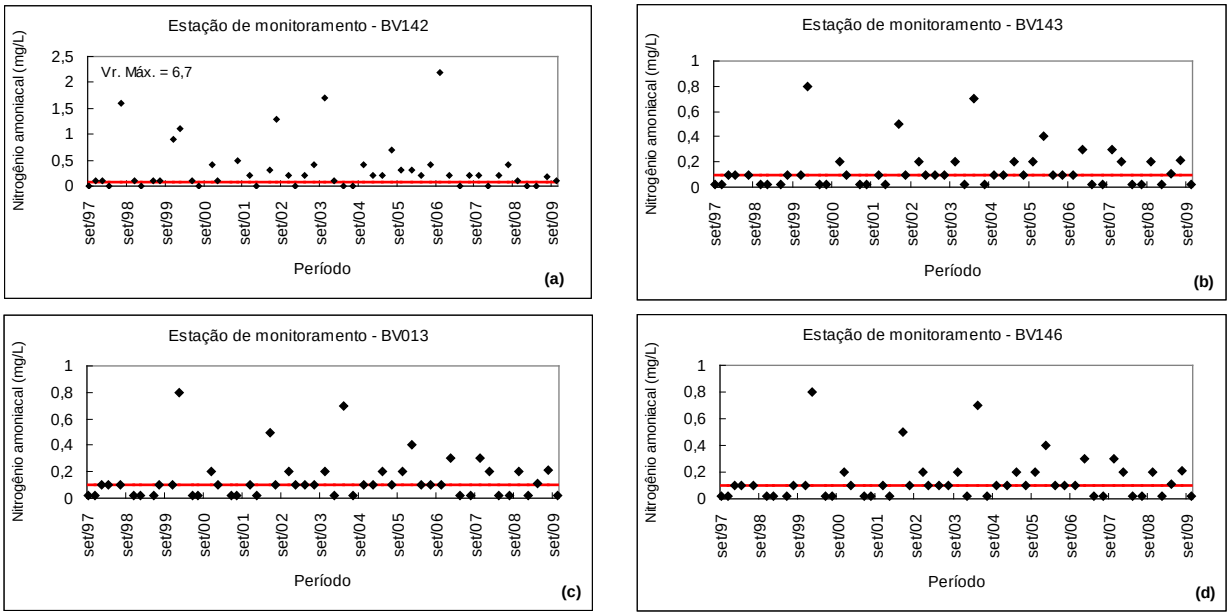
MATERIAIS E MÉTODOS

Os dados utilizados foram selecionados de uma série histórica de monitoramento trimestral da qualidade das águas do estado de Minas Gerais, efetuado pelo Instituto Mineiro de Gestão das Águas (IGAM) desde 1997. Os dados de nitrogênio amoniacal total correspondem ao período de setembro de 1997 a outubro de 2009 e foram selecionados de quatro estações de monitoramento localizadas na bacia do rio São Francisco, sub-bacia do rio das Velhas. As estações de monitoramento foram selecionadas em função do percentual de dados censurados apresentado, visando tornar mais didático o estudo efetuado. Os percentuais de dados censurados das estações BV142, BV143, BV013 e BV146 foram, respectivamente, 24, 36, 64 e 80% dos valores medidos.

RESULTADOS E DISCUSSÃO

A série histórica dos dados relativos às concentrações de nitrogênio amoniacal medidas nas quatro estações de monitoramento, no período de setembro de 2007 a outubro de 2009, considerando diferentes percentuais de censura são apresentados na Figura 1.

A linha vermelha representa o limite de detecção do método analítico para este parâmetro (0,1 mg/L), sendo possível observar número de dados censurados em cada estação de monitoramento. O impacto da substituição de um grande número de dados abaixo do limite de detecção. Ressalta-se que a estação BV142, com o menor percentual de dados censurados, apresenta concentrações de nitrogênio amoniacal expressivamente maiores que os observados nas outras estações. A opção por utilizar valores “fabricados” para cálculo de estatísticas descritivas pode gerar resultados distorcidos e não realísticos, principalmente quando o percentual de dados censurados for grande. Isto pode ser comprovado pela observação da Tabela 1, que apresenta os cálculos da média aritmética, desvio padrão e percentis 25, 50 (mediana) e 75% das concentrações de nitrogênio amoniacal das quatro estações de monitoramento. Os dados de cada estação foram mantidos inalterados, exceto pelas observações censuradas, que foram substituídas por zero, metade do limite de detecção (LD/2) e valores iguais ao limite de detecção (LD).



Nota: A linha vermelha representa o limite de detecção do método analítico (0,1 mg/L)

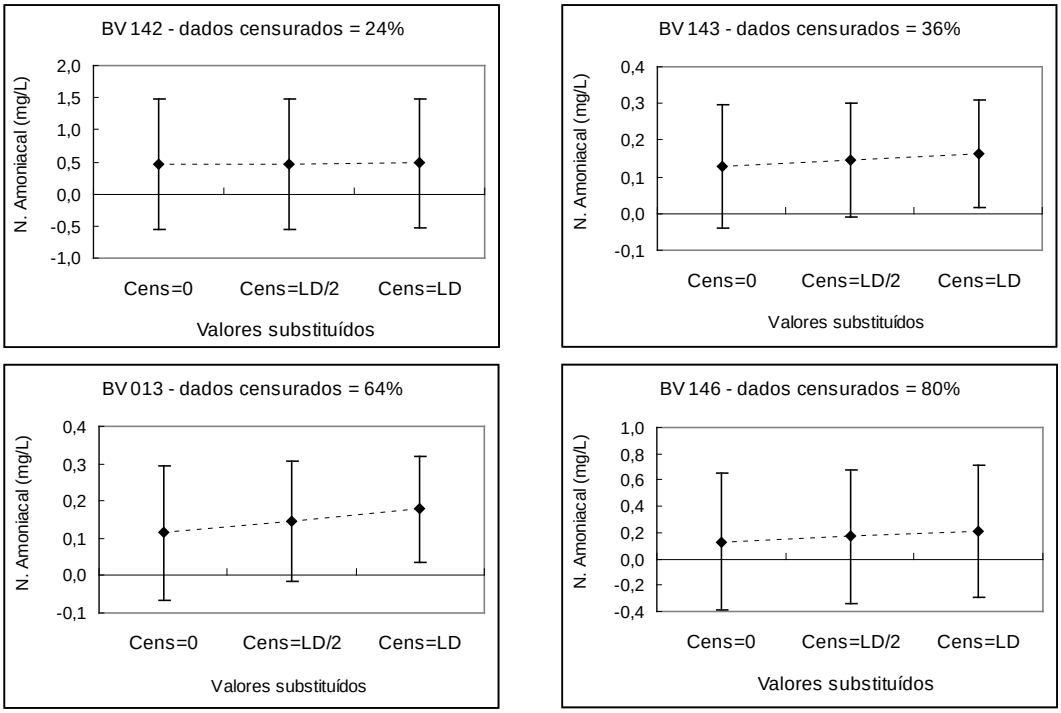
Figura 1: Dados trimestrais de Nitrogênio Amoniacal no período de Set/97 a Out/09 das estações de monitoramento com % de dados censurados igual a: (a) 24%; (b) 36%; (c) 64% e, (d) 80%.

Tabela 1: Estatística descritiva das concentrações de nitrogênio amoniacal das quatro estações de monitoramento, utilizando métodos de substituição.

% censura	Estação	Valor substituído	Média	Desv. Pad.	Perc. 25%	Mediana	Perc. 75%
24%	BV142	Zero	0,46	1,02	0,10	0,20	0,40
		1/2 LD	0,47	1,01	0,10	0,20	0,40
		LD	0,48	1,01	0,10	0,20	0,40
36%	BV143	Zero	0,13	0,17	0,00	0,10	0,20
		1/2 LD	0,15	0,16	0,05	0,10	0,20
		LD	0,16	0,15	0,10	0,10	0,20
64%	BV013	Zero	0,11	0,18	0,00	0,00	0,20
		1/2 LD	0,15	0,16	0,05	0,05	0,20
		LD	0,18	0,14	0,10	0,10	0,20
80%	BV146	Zero	0,13	0,52	0,00	0,00	0,00
		1/2 LD	0,17	0,51	0,05	0,05	0,05
		LD	0,21	0,50	0,10	0,10	0,10

Observa-se que, quanto maior o percentual de valores censurados presentes em uma série de observações, maior a diferença apresentada nos resultados estatísticos. Tomando-se como exemplo a média aritmética dos dados que apresentam 24% de censura, a opção por substituir os valores abaixo do limite de detecção tem pouco impacto no valor calculado. No entanto, para dados com elevado percentual de dados censurados, como é o caso da BV146 (80%), os resultados das médias aritméticas foram fortemente afetados, gerando valores significativamente diferentes entre si. Outra constatação é que as médias das concentrações de nitrogênio calculadas para os conjuntos de dados substituídos por zero foram sistematicamente menores que as outras.

Da mesma forma, estimativas de variabilidade como desvio padrão, utilizando diferentes valores de substituição, apresentam valores sempre distintos, sendo improvável que apresentem valores próximos daqueles que ocorreriam se não existissem dados censurados (Figura 2).



Nota: as barras verticais indicam os valores dos desvios padrão.

Figura 2: Gráficos das médias e desvios padrão das concentrações de nitrogênio amoniacal das quatro estações de monitoramento.

A Figura 3 mostra os gráficos Box-Whisker dos quatro conjuntos de dados após a substituição dos valores censurados. As estimativas dos percentis 25, 50 e 75% foram bastante diferentes para as três substituições, considerando os dados de todas as estações. As diferenças vão se tornando ainda maiores à medida que o percentual de valores censurados aumenta, como comprovado pelos testes estatísticos aplicados. Os testes não paramétricos de Kruskal-Wallis, seguido pelo teste de comparações múltiplas, indicaram diferenças significativas entre todos os conjuntos de dados da BV146, estação com 80% dos dados censurados. Os testes apontaram diferenças significativas também entre os valores substituídos por zero e LD das estações BV013 e BV143 e nenhuma diferença entre os três conjuntos de dados da BV142. É interessante observar que a prática de substituir os dados censurados por qualquer valor entre zero e o limite de detecção é operacionalmente simples e pode ser adequada, em termos práticos, quando o percentual de dados censurados for baixo. Gibbons e Coleman (2001) recomendam os métodos de substituição quando o conjunto de dados detectados for igual ou superior a 80%, ou seja, menos 20% de dados censurados.

No entanto, conforme apresentado na Tabela 1 e Figuras 1 e 2, mesmo para percentuais de censura um pouco acima deste valor recomendado, como o caso da BV142, com 24%, nenhum impacto significativo foi observado na estatística descritiva calculada, considerando as três opções de substituição. Outra observação

importante é que quando menos de 50% dos dados estão abaixo do limite de detecção, é possível o cálculo de alguns percentis, tais como a mediana e o percentil 25%. Do mesmo modo, quando menos de 25% dos dados são censurados, o intervalo interquartil (percentil 75% - percentil 25%) também é conhecido. No entanto, para o cálculo da média aritmética e do desvio padrão não existe processo similar disponível. Assim, para conjuntos de dados que apresentem um percentual elevado de observações abaixo do limite de detecção, a substituição dos dados censurados deve ser evitada. Para estes casos, existem outras alternativas que podem ser selecionadas e a escolha correta do método a ser utilizado depende tanto do grau de censura, que interfere diretamente nos resultados, quanto do tipo de aplicação (estatística descritiva, intervalos de confiança; testes de hipóteses, ajuste de distribuições de probabilidade, correlações, análises de regressão e tendências). Dependendo do método utilizado no seu tratamento, os resultados podem sofrer alterações consideráveis, tendo sua interpretação prejudicada.

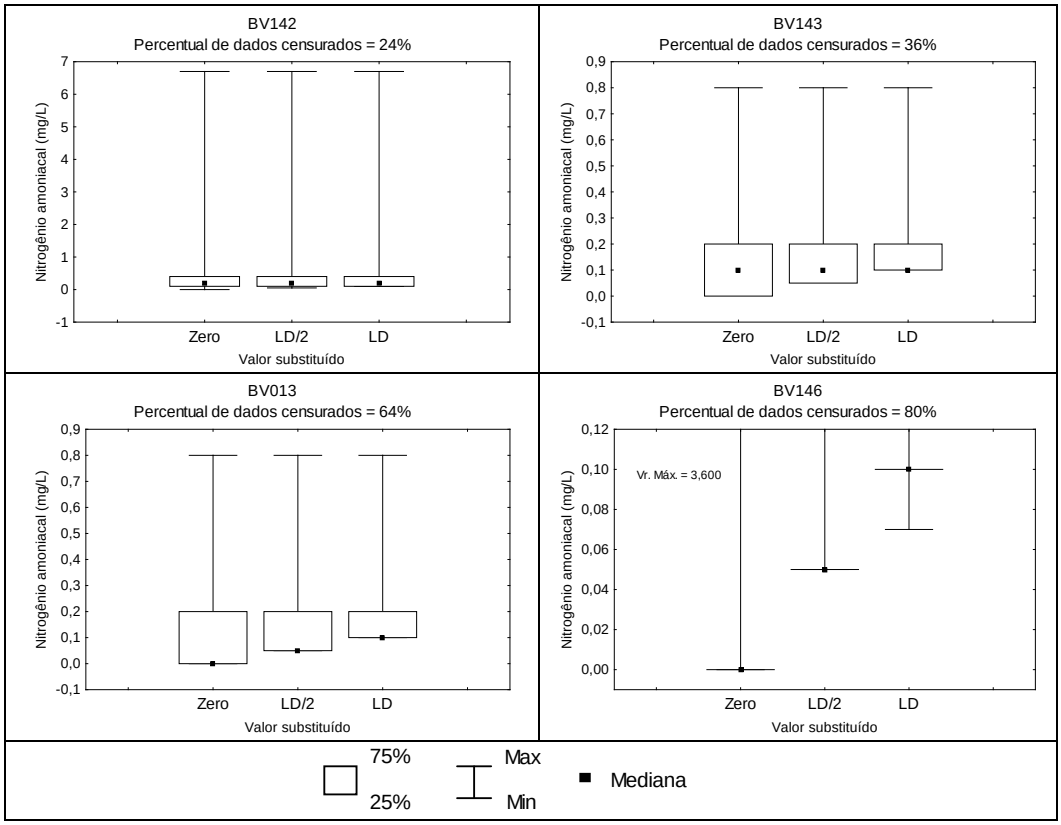


Figura 3: Gráficos Box-whisker das concentrações de nitrogênio amoniacal das quatro estações de monitoramento, considerando dados idênticos, exceto pelas observações censuradas.

Em algumas situações, todas as medições podem se encontrar abaixo do limite de detecção do método analítico, o que não inviabiliza a utilização de tais dados. Métodos baseados na distribuição de probabilidade binomial podem ser utilizados para extrair informações importantes destes dados. Dentre eles, destacam-se: determinação de intervalos de confiança, testes de hipóteses para comparação entre grupos considerando proporção, cálculo da probabilidade de violação de determinados padrões ambientais legais.

CONCLUSÕES

O método da substituição continua sendo o mais comumente utilizado para cálculo de estatística descritiva de dados ambientais censurados. No entanto, independentemente do valor escolhido, é extremamente improvável que todas as amostras reportadas como censuradas efetivamente apresentem o mesmo valor. Existe um ceticismo generalizado acerca das informações contidas nessas observações, quando na verdade, uma grande quantidade de informações estará disponível nos dados censurados, desde que métodos adequados para a sua extração sejam utilizados.

A prática de substituição dos dados censurados por qualquer valor entre zero e o limite de detecção é operacionalmente simples e pode ser adequada, em termos práticos, quando o percentual de dados censurados for baixo, por exemplo, menos de 20%. Assim, para conjuntos de dados que apresentem um percentual elevado de observações abaixo do limite de detecção, a substituição dos dados censurados deve ser evitada. Para estes casos, existem outras alternativas que podem ser selecionadas e a escolha correta do método a ser utilizado depende tanto do grau de censura, que interfere diretamente nos resultados, quanto do tipo de aplicação (estatística descritiva, intervalos de confiança; testes de hipóteses, ajuste de distribuições de probabilidade, correlações, análises de regressão e tendências). Dependendo do método utilizado no seu tratamento, os resultados podem sofrer alterações consideráveis, tendo sua interpretação prejudicada.

AGRADECIMENTOS

Os autores agradecem ao Instituto Mineiro de Gestão das Águas (IGAM) por terem disponibilizados os dados e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig) por ter viabilizado essa pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

1. CHRISTOFARO, C. *Avaliação probabilística de risco ecológico de metais nas águas superficiais da Bacia do rio das Velhas - MG.* 2009. 274 f. Tese (Doutorado em Saneamento, Meio Ambiente e Recursos Hídricos) - Escola de Engenharia, Universidade Federal de Minas Gerais, Belo Horizonte, 2009.
2. GIBBONS, R. D.; COLEMAN, D. E. *Statistical methods for detection and quantification of environmental contamination*. John Wiley & Sons, Inc., New York, 2001.
3. HELSEL, D. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. John Wiley, New York, 2005b.
4. HELSEL, D. Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*, v. 65, pp. 2434 –2439, 2006.