DOI: 10.1590/S1413-41522015020000121484

Artigo Técnico

Influência de dados censurados no cálculo da concentração média das variáveis de qualidade da água demanda química de oxigênio e fosfato

The influence of censored data on computing mean concentrations of water quality variables chemical oxygen demand and phosphate

Thaisa de Souza Contar[1], Cesar Augusto Medeiros Destro[1], Gilson Alberto Rosa Lima[2]

RESUMO

Um dos problemas comumente encontrado na análise estatística de dados provenientes do monitoramento da qualidade da água envolve os chamados dados censurados. Este estudo mostra como os valores calculados das concentração médias de amostras que contenham dados censurados são influenciadas pelo tamanho da amostra, percentual de dados censurados e método de cálculo. As concentrações médias foram calculadas com os métodos de substituição e *maximum likelihood estimator* (MLE) para amostras de demanda química de oxigênio e fosfato. O intervalo de confiança da média foi utilizado como métrica de comparação. Os resultados revelam que as concentrações médias calculadas com o método de substituição simples se posiciona fora do intervalo de confiança. Já o método MLE calcula valores aceitáveis, porém, a eficácia do método depende de um limite de elasticidade do nível de censura e do tamanho da amostra.

Palavras-chave: dados censurados; método MLE; DQO; fosfato; qualidade da água.

ABSTRACT

One of the most common problems in statistical analysis of data from water quality monitoring programs involves the so-called censored data. This study highlights how the sample mean concentration of chemical oxygen demand and phosphate variables are influenced by the estimation method, sample size and percentage of censored data. The confidence interval was used as a metric for comparison of calculated sample mean concentrations on both simple substitution and maximum likelihood estimator (MLE) methods. The results show that sample mean concentrations calculated by simple substitution method are positioned outside the confidence interval. Otherwise, the MLE method produces acceptable values for sample mean concentration, although the magnitude is constrained by sample size and percentage of censored data. The effectiveness of the method depends on an elastic limit of censoring level and sample size.

Keywords: censored data; MLE method; COD, phosphate; water quality.

INTRODUÇÃO

Um dos problemas comumente encontrados na análise estatística de dados provenientes de monitoramento da qualidade da água envolve os chamados dados ou valores censurados (HELSEL, 2004). Dados censurados ocorrem quando o valor de uma observação é apenas parcialmente conhecido, encontra-se abaixo ou acima do limite de detecção do instrumento ou técnica de medição. Para identificar um valor censurado é comum utilizar o indicador "< LD", o qual significa que o valor medido é menor do que o limite de detecção do método de medição, em outras palavras, o sinal produzido pela amostra é muito pequeno para ser distinguido do ruído de fundo do instrumento de medida (AHN, 1998). O valor do limite de detecção (LD) depende do método analítico utilizado e pode ser definido, segundo She (1997), como a concentração mínima de uma substância que pode ser medida e reportada, com 99% de confiança de que a concentração do analito é maior do que zero. É um valor de concentração inferior ao limite considerado confiável o suficiente para ser relatado como valor numérico observado. Devido à presença

[1]Engenheira Sanitarista pela Universidade Federal de Mato Grosso (UFMT). Mestre em Recursos Hídricos pela UFMT – Cuiabá (MT), Brasil.
[2]Bacharel em Física pela Universidade Federal de Santa Maria (UFSM). Mestre em Física pela Universidade Estadual de Campinas (UNICAMP). Doutor em Ciências pela Universidade de São Paulo (USP). Professor Adjunto do Departamento de Engenharia Sanitária e Ambiental, UFMT – Cuiabá (MT), Brasil.
Endereço para correspondência: Gilson Alberto Rosa Lima – Avenida Fernando Corrêa da Costa, 1050, Bloco F, Sala 240 – Boa Esperança – 78060-900 – Cuiabá (MT), Brasil – E-mail: garlima@ufmt.br
Recebido: 30/07/13 – **Aceito:** 10/08/14 – **Reg. ABES:** 121484

de valores censurados não é possível calcular com precisão a concentração média de uma variável que contenha valores censurados. Wendelberger e Campbell (1994) ressaltam que há dois métodos que permitem efetuar esse cálculo, substituindo os valores censurados, ou tratando-os estatisticamente.

É comum na literatura o uso do método de substituição simples, no qual os valores censurados são substituídos por zero, ½LD ou LD (HELSEL, 2004). Essa prática foi estimulada e recomendada pela Environmental Protection Agency (EPA) (EPA, 1998 *apud* HELSEL, 2004). Porém, nas duas ultimas décadas vários estudos mostraram que o método de substituição gera resultados imprecisos em comparação com outros métodos, especialmente quando o conjunto de dados possui poucas observações (n<50) ou um elevado percentual de valores censurados (pvc). Segundo Gleit (1985), Gilliom e Helsel (1986), Helsel e Cohn (1988), Newman e Dixon (1990), Helsel e Hirsch (2002), Singh e Nocerino (2002) e Helsel (2004), o método da substituição produz estimativa imprecisa e normalmente obscurece padrões e tendências que um determinado conjunto de dados possui. Gleit (1985) e Gilliom e Helsel (1986) abordaram o desempenho do método *maximum likelihood estimator* (MLE) e compararam detalhadamente a capacidade dos vários métodos de cálculo para milhares de conjuntos de dados simulados por computador. Gilliom e Helsel (1986) aplicaram esse método a vários conjuntos de dados de qualidade de água. Kuttatharmmakul *et al.* (2001) demonstraram que o método MLE produz estimativas confiáveis para amostras contendo pequeno pvc, menores do que 33%.

Como o método de substituição é o mais usado no Brasil, este estudo mostra a importância do tratamento estatístico adequado no cálculo da concentração média de variáveis que contenham valores censurados. Os dados utilizados neste estudo são provenientes do Programa de Monitoramento da Qualidade da Água do Estado de Mato Grosso, sob a coordenação da Secretaria de Estado de Meio Ambiente (SEMA) e estão disponíveis no site da Agencia Nacional de Água (ANA) no link http://hidroweb.ana.gov.br/.

METODOLOGIA

O estudo foi realizado em quatro etapas. A primeira etapa, chamada de pré-processamento dos dados, foi constituída por uma exaustiva análise dos valores observados de cada uma das variáveis de qualidade para eliminar todos os possíveis tipos de inconsistências provocadas por erros sistemáticos e não sistemáticos. Uma análise comparativa entre os valores registrados nos boletins de análise, elaborado pela gerência do laboratório da SEMA do Mato Grosso e os valores registrados no sistema *Hidroweb* foi realizada para identificar os valores censurados. Essa etapa foi de suma importância, já que o modelo de armazenamento de dados do sistema *Hidroweb* não distingue valores observados de valores censurados.

Na segunda etapa, o método gráfico de probabilidades (*probability plot*) (HELSEL, 2004) foi utilizado para determinar a melhor função de distribuição teórica para ajustar os valores observados e censurados de cada uma das variáveis de qualidade. O método assume como hipótese nula ou H_0 que os valores observados podem ser modelados estatisticamente por uma certa distribuição de probabilidades (no caso deste estudo foram consideradas as distribuições normal e lognormal). A distribuição "teórica" é representada no gráfico de probabilidades por uma reta e as observações são inseridas no gráfico individualmente e, se os valores observados ficarem distribuídos próximos dessa reta, assume-se que a hipótese nula é verdadeira (não se rejeita H_0). A hipótese é verificada através do teste de normalidade de Anderson-Darling. Esse procedimento é necessário por duas razões, primeiro porque normalmente se assume a priori que os dados podem ser modelados por uma distribuição normal e segundo porque o método MLE assume uma função de distribuição para estimar as medidas de tendência central e dispersão (AHN, 1998; HELSEL & HIRSCH, 2002; HELSEL, 2004).

Na terceira etapa foram calculados os intervalos de confiança da concentração média para $\alpha = 0,05$, média, mediana e desvio padrão para o maior conjunto de dados com n≥100 (Tabela 1) de cada uma das variáveis de qualidade. Os intervalos de confiança da concentração média das variáveis foram calculados para os métodos de substituição e MLE, utilizando tanto a distribuição normal quanto a distribuição lognormal.

Na quarta etapa foi realizada a análise da influência do método de cálculo, tamanho da amostra (n) e percentual de valores censurados no valor calculado da concentração média das variáveis de qualidade, tendo como métrica de comparação os intervalos de confiança calculados com o método MLE.

Os dados utilizados correspondem ao período de 1995 a 2008 e são provenientes da estação de amostragem Passagem da Conceição (código Hidroweb 66259200), a qual faz parte do programa de monitoramento da qualidade da água do Rio Cuiabá. Essa estação foi escolhida por conter o maior número de observações e por ser próxima ao ponto de captação de água para o abastecimento da capital do Estado de Mato Grosso. Cabe ressaltar que, embora as análises tenham sido realizadas para todas as variáveis que contêm valores censurados (fósforo, demanda química de oxigênio – DQO, demanda bioquímica de oxigênio – DBO, nitrato, nitrito, nitrogênio, n-amoniacal, nitrogênio Kjeldhak total – NKT, oxigênio dissolvido, alcalinidade, fosfato, coliformes fecais e coliformes totais), mostrar e discutir os resultados de todas as variáveis analisadas demanda mais páginas do que é permitido. Por isso, neste artigo são mostrados apenas os resultados obtidos para as variáveis DQO e fosfato.

O MÉTODO *MAXIMUM LIKELIHOOD ESTIMATOR*

O método MLE assume uma função de distribuição (normal, lognormal ou outra) para modelar os valores observados e censurados. O método constitui em resolver a função $L(a,b)$ (Equação 1), onde L é a função de máxima verossimilhança. Os parâmetros a e b são otimizados em uma rotina até que os valores computados sejam aqueles que maximizam L.

$$L = \prod_{i=1}^{N} f(x_i, \alpha, \beta) \quad (1)$$

Tabela 1 – Medidas de tendência central e dispersão para amostras das variáveis demanda química de oxigênio e fosfato utilizando os métodos de substituição simples e *maximum likelihood estimator*.

Método	n	Média (mg.L^{-1})	Desvio padrão (mg.L^{-1})	Mediana (mg.L^{-1})	IC95%
	Demanda química de oxigênio				
Substituição: zero	100	9,1040	6,8156	8,7000	7,7516-10,4564
Substituição: ½LD	100	9,9140	5,9555	9,0000	8,7323-11,0957
Substituição: LD	100	10,7240	5,6840	9,0000	9,5960-11,8520
MLE (normal)	100	10,0109	5,9127	10,0109	8,8324-11,1893
MLE (lognormal)	100	10,1280	5,5099	8,8967	9,0990-11,2734
	Fosfato				
Substituição: zero	104	0,1052	0,1245	0,0800	0,0810-0,1294
Substituição: ½LD	104	0,1128	0,1185	0,0800	0,0896-0,1357
Substituição: LD	104	0,1201	0,1134	0,0800	0,0981-0,1422
MLE (normal)	104	0,1131	0,1179	0,1131	0,0904-0,1358
MLE (lognormal)	104	0,1148	0,1214	0,0789	0,0943-0,1398

LD: limite de detecção; MLE: *maximum likelihood estimator*.

Seja um conjunto com n dados (observações e valores censurados) $x(x_1, x_2, \ldots x_n)$, conhecendo-se a distribuição de probabilidade que descreve x:$f(x,\alpha,b)$, o método MLE calcula os valores de α e b que maximizam a probabilidade dos valores observados $(x_1, x_2, \ldots x_n)$ serem descritos pela referida distribuição de probabilidade. Assim, a probabilidade de se medir cada um dos valores é representada matematicamente pela probabilidade de se medir x_1, que é $f(x_1,\alpha,b)dx$, a probabilidade de se medir x_2, que é $f(x_2,\alpha,b)dx$, e a probabilidade de se medir x_n, que é $f(x_n,\alpha,b)dx$. Como as medidas são consideradas independentes, a função L pode ser expressa pelas Equações 2 ou 3.

$$L = f(x_1, \alpha, \beta)dx \times f(x_2, \alpha, \beta)dx \ldots f(x_n, \alpha, \beta)dx \quad (2)$$

$$L = f(x_1, \alpha, \beta) \times f(x_2, \alpha, \beta)\mathrm{L}f(x_n, \alpha, \beta)dx^n \quad (3)$$

Como o termo dx^n é apenas uma constante de proporcionalidade, L pode ser escrito como um produtório, sendo α e b os parâmetros que se quer determinar.

Neste estudo α é a média e b é o desvio padrão (ou seja, os parâmetros que definem a distribuição normal teórica). Os valores de α e b que maximizam a função L são calculados através da Equação 4:

$$\left.\frac{\partial L}{\partial \alpha}\right|_{\alpha=\alpha^*} = 0 \quad \left.\frac{\partial L}{\partial \beta}\right|_{\beta=\beta^*} = 0, \quad (4)$$

Tanto L como ln (L) podem ser utilizados na Equação 1, ambos têm seu máximo na mesma posição. Como o logaritmo natural converte o produtório em um somatório, a Equação 4 pode ser reescrita como mostra a Equação 5:

$$\ln L = \sum_{i=1}^{N} \ln f(x_1, \alpha, \beta) \quad (5)$$

sendo a nova condição de maximização dada expressa pelas Equações 6 e 7.

$$\left.\frac{\partial \ln L}{\partial \alpha}\right|_{\alpha=\alpha^*} = \sum_{i=1}^{N} \frac{\partial L}{\partial \alpha} \left.\ln f(x_1,\alpha)\right|_{\alpha=\alpha^*} = 0 \quad (6)$$

$$\left.\frac{\partial \ln L}{\partial \beta}\right|_{\beta=\beta^*} = \sum_{i=1}^{N} \frac{\partial L}{\partial \beta} \left.\ln f(x_1,\alpha)\right|_{\beta=\beta^*} = 0 \quad (7)$$

A média e o desvio padrão da distribuição são calculados com base nos valores observados x_i e na proporção de valores abaixo do limite de detecção do método analítico de medida (medição). Sendo expressas pelas Equações 8 e 9 respectivamente.

$$\alpha = \mu = exp\left(\frac{\mu_{\ln X} + \sigma^2_{\ln X}}{2}\right) \quad (8)$$

$$\beta = \sigma^2 = \mu_x^2[\exp(\sigma^2_{\ln X})\text{-}1] \quad (9)$$

Em síntese, o método utiliza os valores observados, a proporção de dados censurados, os limites de detecção (se houver mais de um) e a distribuição teórica como dados de entrada. Os parâmetros α e b são computados de tal forma a maximizar a aderência dos valores observados e a proporção de dados censurados à distribuição teórica. Em outras palavras, a função L envolve os valores observados e a proporção de valores censurados. Mais detalhes sobre o método podem ser obtidos em Aitchison e Brown (1957), Gilbert (1987) e Helsel (2004). Esse método está implementado nos seguintes pacotes estatísticos: *NADA*, *MiniTab, SAS* e *R*.

RESULTADOS E DISCUSSÃO

O primeiro resultado se refere à escolha da função de distribuição que melhor ajusta os valores observados e censurados das concentrações de cada uma das variáveis de qualidade. Todos os testes foram realizados para (α = 0,05), ou seja, com nível de significância estatística de 95%. Nas seções a seguir são mostrados os resultados dos testes de aderência para escolha da função de distribuição teórica. Primeiro aplicando o método de substituição onde os valores censurados são substituídos por zero, ½LD e LD, e depois aplicando

o método MLE. Na seção seguinte são mostrados parâmetros estatísticos de medida de tendência central e dispersão, calculados pelos dois métodos. A próxima seção mostra o resultado da análise da influência do número de observações (n) e percentual de valores censurados no valor das concentrações médias calculadas com os métodos de substituição e MLE.

Teste de aderência – método de substituição

Embora o teste de aderência tenha sido realizado para a substituição dos dados censurados pelos valores zero, ½LD e LD, devido à limitação de espaço, a Figura 1 mostra apenas o resultado da substituição por ½LD, pois os resultados dos testes de aderência para a substituição por zero e LD são semelhantes. Na Tabela 1 são mostrados os valores dos parâmetros estatísticos média, mediana e desvio padrão para as três substituições.

A Figura 1 mostra o gráfico de probabilidade, o qual se fundamenta no teste de hipóteses, sendo que a hipótese H_0 estabelece que os valores observados e censurados podem ser modelados por uma distribuição teórica a escolher e a hipótese H_1 estabelece que tais valores não podem ser modelados pela distribuição escolhida. A Figura 1 deve ser interpretada da seguinte forma: a função de distribuição é representada no gráfico de probabilidade pela reta. Se os valores observados e censurados se distribuírem próximos à reta, significa que o valor p calculado é maior do que 0,05 (significância do teste, Tabela 1), logo, assume-se como verdadeira a hipótese H_0. Caso contrario, assume-se a hipótese H_1 como verdadeira. A proporção de valores censurados é calculada levando-se em conta o valor do LD. Logo, a Figura 1 mostra que os valores observados e censurados das concentrações medidas das variáveis DQO e fosfato não podem ser modeladas com a distribuição normal, pois a maioria das observações e dos valores censurados se encontram fora da reta de regressão e do intervalo de confiança do teste.

A Figura 2 mostra os testes de aderência para a distribuição lognormal. Embora a função de distribuição lognormal ajuste melhor aos valores observados, a existência dos valores censurados prejudica o ajuste da função de distribuição teórica com os valores observados, fazendo com que alguns pontos fiquem fora do intervalo de confiança do teste.

Como a análise foi realizada para um número de observações considerado satisfatório (Tabela 1) do ponto de vista da eficácia dos testes de aderência, o resultado mostra que, aplicando o método de substituição nesse conjunto de dados, tanto a hipótese de normalidade quanto de lognormalidade são rejeitadas Consequentemente, quando se utiliza o método de substituição é preferível usar o parâmetro estatístico mediana como medida de tendência central.

Teste de aderência – método *maximum likelihood estimator*

O mesmo procedimento de análise foi realizado para o método MLE. As Figuras 3 e 4 mostram o resultado do teste de aderência para a distribuição normal e lognormal, respectivamente.

A Figura 3 mostra que mesmo aplicando o método MLE com a função de distribuição normal, os valores observados das concentrações também não se ajustam à reta de regressão, indicando que a hipótese de normalidade deve ser rejeitada (Tabela 2). Ou seja, a função de distribuição normal não deve ser usada para modelar esse conjunto de dados. A Figura 4 mostra que ao aplicar o método MLE considerando a função de distribuição lognormal, observa-se uma significativa melhora no ajuste dos valores observados à reta de regressão.

Os resultados mostrados nas Figuras 1 a 4 estão de acordo com os resultados de Nascimento (2010), o qual mostrou que a maioria das variáveis de qualidade da água monitoradas deve ser modelada por uma função lognormal, especialmente aquelas que possuem valores censurados. A Tabela 1 mostra os valores estimados dos parâmetros estatísticos média, desvio padrão e mediana, com os métodos de substituição e MLE para as distribuições teóricas normal e lognormal.

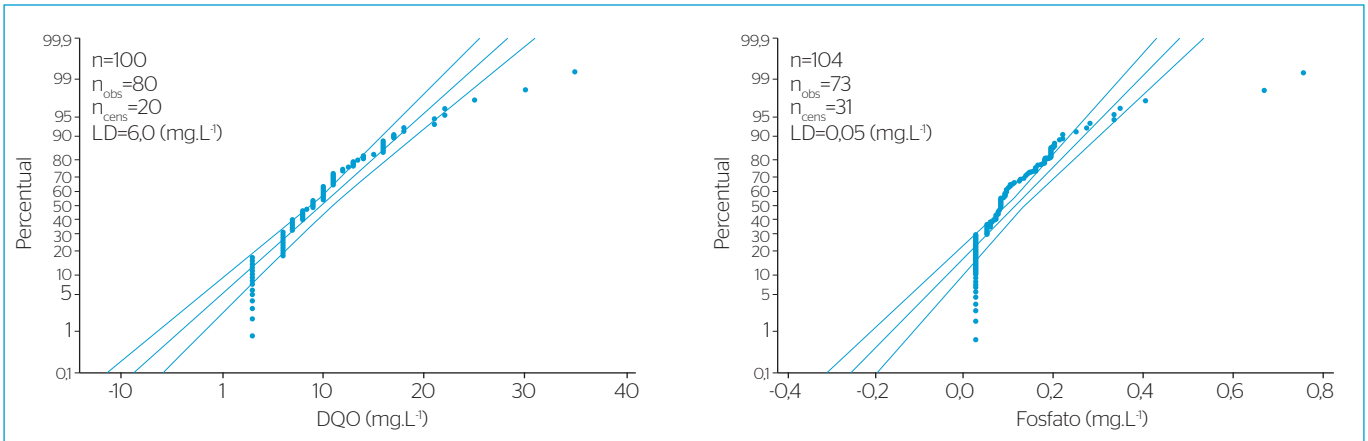
Levando em conta que o teste de aderência para o conjunto de dados (n≥100) indica a distribuição lognormal como aquela mais apropriada para descrever os valores observados das variáveis DQO e fosfato, e que o método de substituição não possui nenhum embasamento científico que o justifique, é possível concluir que o resultado mais plausível é obtido quando a concentração média é calculada com o método MLE utilizando-se a distribuição lognormal.

Tomando-se a distribuição lognormal e o método MLE como referência de melhor estimativa da concentração média de uma amostra que contenha dados censurados, o método de substituição por zero subestima o valor das concentrações médias, enquanto a substituição por LD superestima esse valor. A substituição por ½LD estima o valor da média próximo do valor estimado pelo método MLE, bem como dos intervalos de confiança.

Este resultado é consistente com o gráfico de probabilidade (Figura 3), no qual os valores observados estão dentro do intervalo de confiança do teste. Nota-se que, embora os valores da concentração média calculada pelo método de substituição para (n≥100) estejam dentro do intervalo de confiança da média, o teste de aderência mostra que a hipótese dos valores observados serem modelados pela distribuição normal é rejeitada. Logo, modelar esse conjunto de dados assumindo uma distribuição normal deve ser evitado. O método MLE (lognormal) ainda estimou os menores valores para o desvio padrão e para os intervalos de confiança da concentração média. A mesma análise foi realizada para a variável fosfato, sendo o resultado similar.

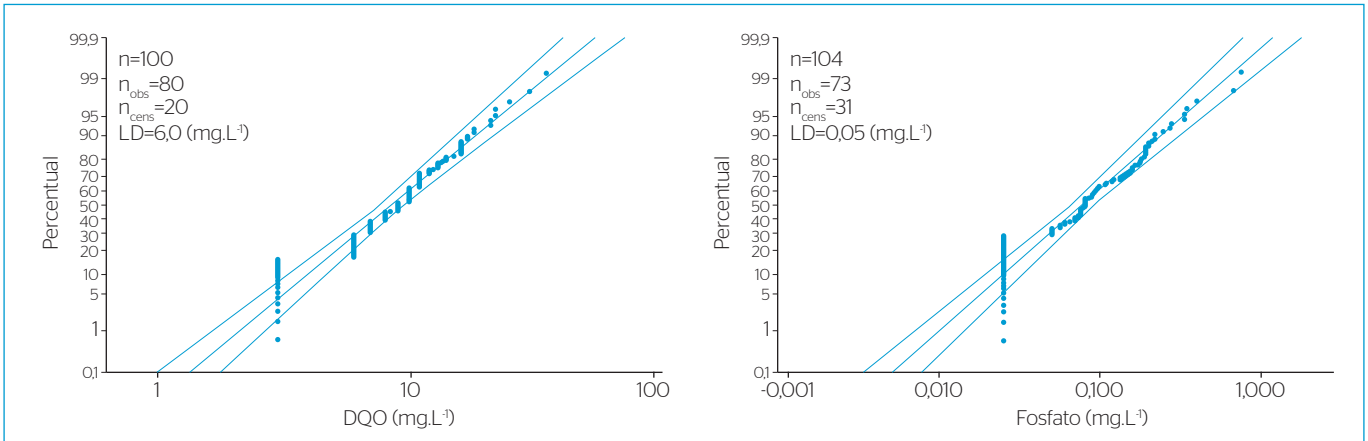
Influência dos valores censurados e tamanho da amostra no valor da concentração média das variáveis de qualidade

A Figura 5 mostra que o valor da concentração média de DQO é influenciada pelo método de cálculo, percentual de valores censurados e tamanho da amostra, porém, o percentual de valores censurados



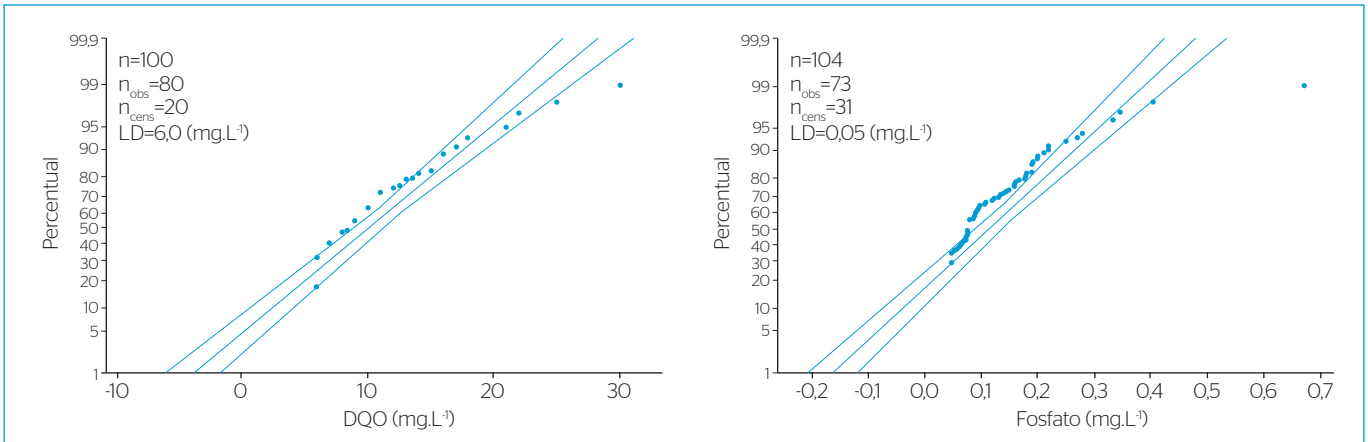
n: número total de dados da amostra; n_{obs}: número de observações; n_{cens}: número de valores censurados; LD: limite de detecção; DQO: demanda química de oxigênio.

Figura 1 – Ajuste dos valores observados da concentração média com a função de distribuição normal quando LD é substituindo por ½LD.



n: número total de dados da amostra; n_{obs}: número de observações; n_{cens}: número de valores censurados; LD: limite de detecção; DQO: demanda química de oxigênio.

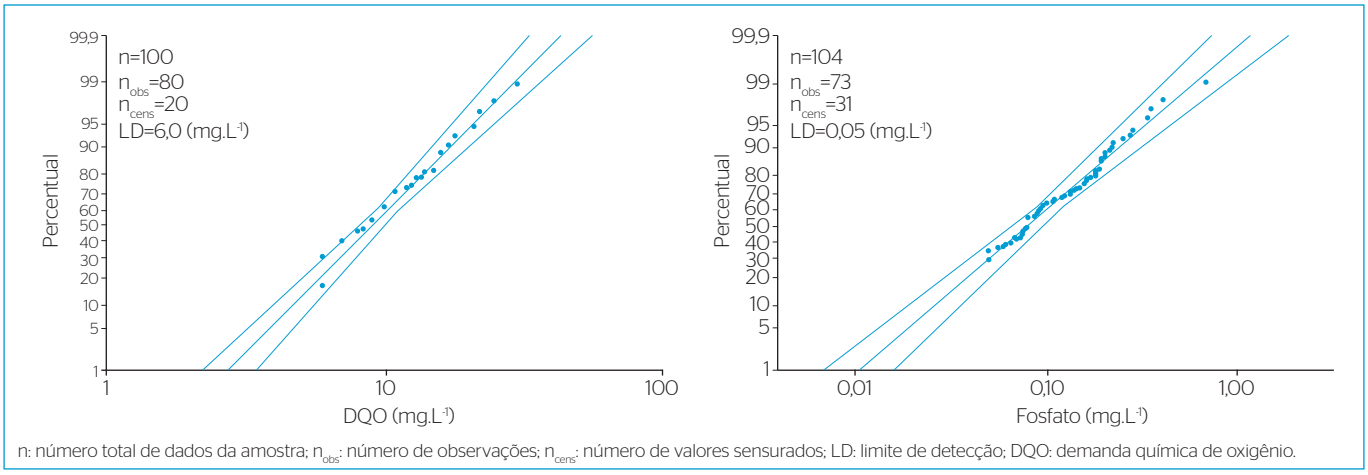
Figura 2 – Ajuste dos valores observados da concentração média com a função de distribuição lognormal quando LD é substituindo por ½LD.



n: número total de dados da amostra; n_{obs}: número de observações; n_{cens}: número de valores censurados; LD: limite de detecção; DQO: demanda química de oxigênio.

Figura 3 – Ajuste dos valores observados da concentração média com a função de distribuição normal utilizando o método *maximum likelihood estimator.*

exerce maior influencia. À medida que o percentual de valores censurados aumenta, o valor da concentração média diminui. A concentração média sofre maior variação quando o valor censurado é substituído por zero e menor quando substituído por LD. A substituição do valor censurado por ½LD mantém as concentrações médias dentro do intervalo de confiança para um percentual maior de valores substituídos.



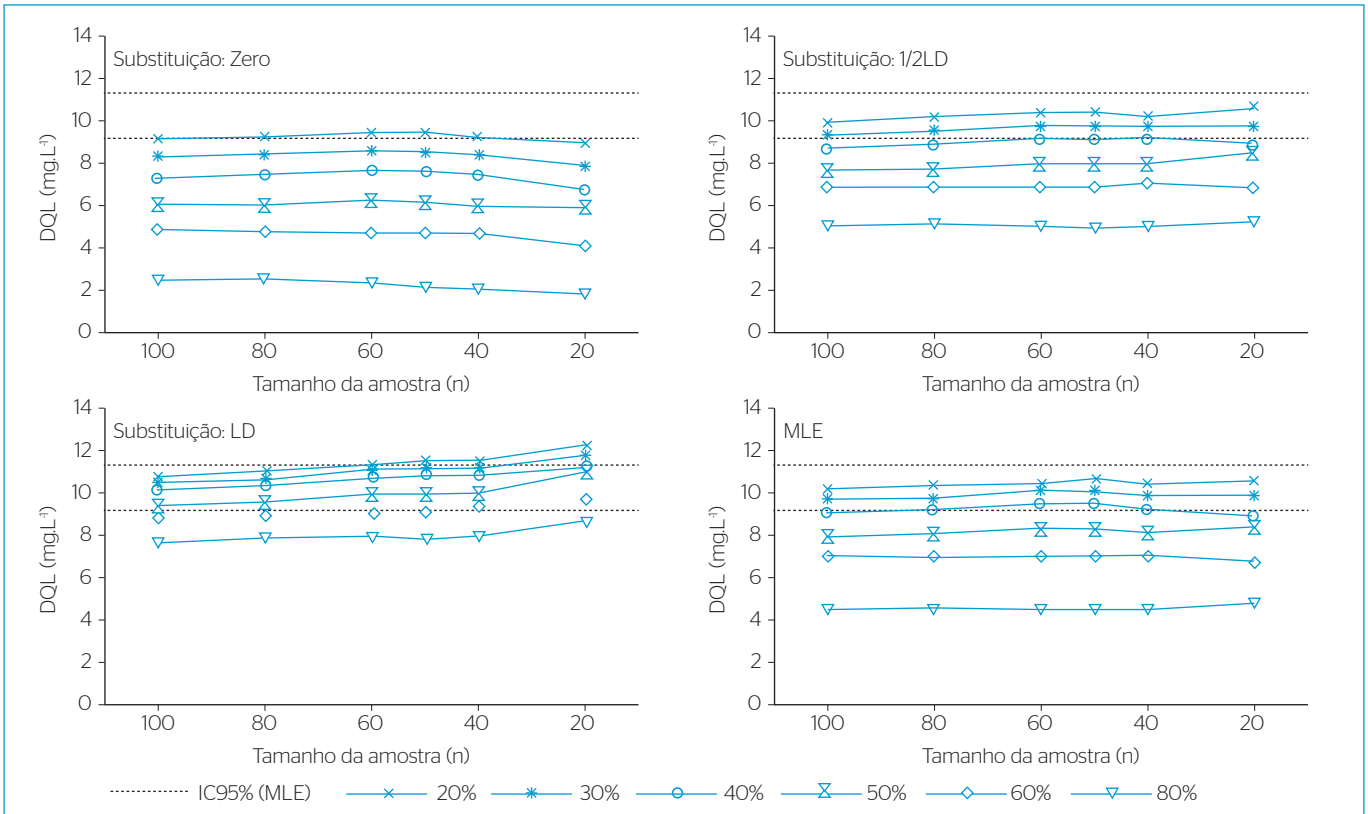
n: número total de dados da amostra; n_{obs}: número de observações; n_{cens}: número de valores sensurados; LD: limite de detecção; DQO: demanda química de oxigênio.

Figura 4 – Ajuste da concentração média observada com a função de distribuição lognormal utilizando o método *maximum likelihood estimator*.

Tabela 2 – Valor p do teste de hipótese para a aderência a distribuição teórica.

Variáveis	Valor p (½LD)	Média (mg.L^{-1})	DP (mg.L^{-1})	Mediana (mg.L^{-1})	Valor p (MLE)	Média (mg.L^{-1})	DP (mg.L^{-1})	Mediana (mg.L^{-1})
Distribuição normal								
DQO	<0,005	9,914	5,955	9,00	<0,005	11,22	4,32	12,22
Fosfato	<0,005	0,1127	0,1185	0,08	<0,005	0,38	0,105	0,91
Distribuição lognormal								
DQO	<0,005	9,914	5,955	9,00	<0,005	10,42	5,866	9,12
Fosfato	<0,005	0,1127	0,1185	0,08	<0,005	0,118	0,121	0,078

Teste de hipótese verificado com o teste de normalidade de Anderson-Darling; DP: desvio padrão; MLE: *maximum likelihood estimator*; DQO: demanda química de oxigênio



LD: limite de detecção; DQO: demanda química de oxigênio; MLE: *maximum likelihood estimator*; IC95%: intervalo de confiança de 95%.

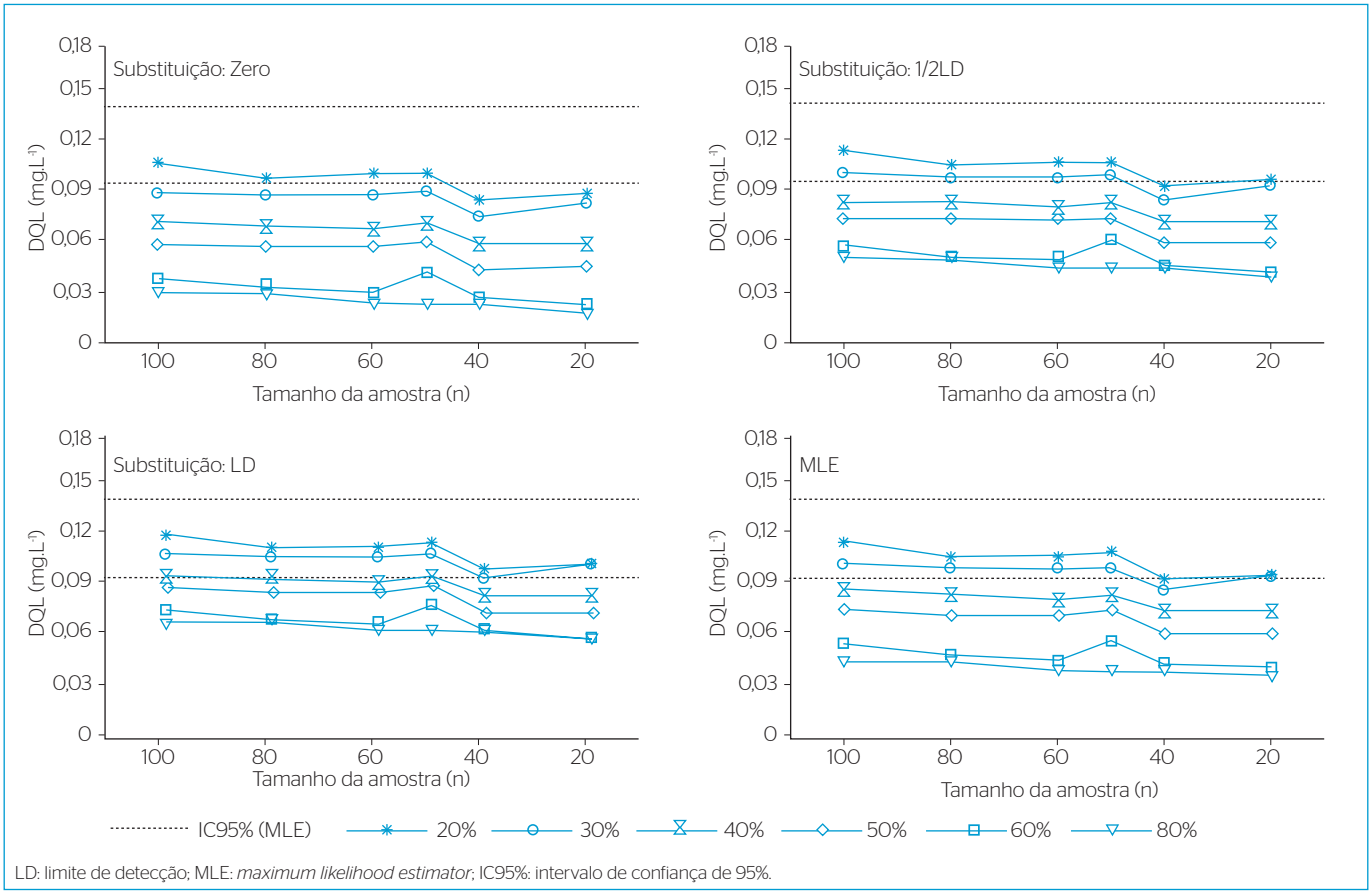
Figura 5 – Tamanho da amostra e percentual de valores censurados em função da concentração média calculado com o método substituição.

Ao substituir os valores censurados da variável DQO por zero, somente as amostras com 20% de valores censurados, exceto para n=20, encontram-se dentro do intervalo de confiança calculado através do método MLE. Em todos os demais casos, as médias amostrais encontram-se fora do intervalo de confiança, mesmo quando n=100, evidenciando que a substituição por zero subestima a concentração média. No caso da substituição por ½LD, a concentração amostral calculada está dentro do intervalo de confiança para todas as amostras que contenham até 30% de valores censurados. Para as amostras com 40% de valores censurados, as médias amostrais estão dentro do intervalo de confiança calculado apenas para n=60, 50 e 40. Embora as amostras com n=60, 50 e 40 contendo 40% de valores censurados estejam dento do intervalo de confiança, ressalta-se que as médias encontram-se no limite inferior do intervalo de confiança. A média amostral está dentro do intervalo de confiança para qualquer tamanho de amostra com pvc igual a 40 e 50%. Um aumento de 10% no pvc faz com que as concentrações médias calculadas estejam fora dos intervalo de confiança para n≥60. Para n=50, a concentração média de DQO está dentro do intervalo de confiança. Para n≤40 as concentrações médias estão dentro dos intervalos de confiança.

Na substituição por LD, apenas as amostras com pvc igual a 80% mantiveram-se fora dos intervalos de confiança, independentemente do tamanho da amostra. Ressalta-se que para pvc igual a 20 e 30% e com tamanhos de amostra pequenos (n=40 e 20), a concentração média calculada encontra-se acima dos intervalos de confiança, evidenciando a superestimação do valor da média quando os valores censurados são substituídos por LD.

Analisando a influência dos valores censurados e do tamanho das amostras no valor da concentração média de fosfato (Figura 6), verifica-se o mesmo comportamento da variável DQO, ou seja, há uma tendência de diminuição das concentrações médias com o aumento do pvc, independentemente do critério de substituição adotado (zero, LD ou ½LD). No caso da substituição por zero, as concentrações médias calculadas encontram-se dentro dos intervalos de confiança apenas nas amostras com n≥50 e com pvc=30%. Um aumento de 10% no pvc, independentemente do tamanho da amostra, faz com que todas as concentrações médias fiquem fora dos intervalos de confiança. Além disso, nos casos em que as concentrações médias estão dentro dos intervalos de confiança essas se encontram bastante próximas do intervalo de confiança inferior.

Quando os valores censurados são substituídos por ½LD, as concentrações médias de fosfato permanecem dentro dos intervalos de confiança para pvc ≤ 40% para tamanhos de amostra com n≥50. Com pvc≥50, para qualquer tamanho da amostra as concentrações médias encontram-se fora dos intervalos de confiança. Na substituição dos valores censurados por LD,



LD: limite de detecção; MLE: *maximum likelihood estimator*; IC95%: intervalo de confiança de 95%.

Figura 6 – Valores da concentração média em função do tamanho da amostra e percentual de valores censurados calculados com o método *maximum likelihood estimator*.

para qualquer tamanho da amostra com pvc=30% as concentrações médias encontram-se dentro dos intervalos de confiança. Com pvc=40% e n=40, a concentração média encontra-se dentro dos intervalos de confiança calculados. Ao observar a Figura 6, nota-se que as concentrações médias são bastante influenciadas pelo pvc, enquanto o tamanho da amostra exerce uma influência menor no cálculo da média.

CONCLUSÕES

Este estudo mostra a influência exercida pelos valores censurados, tamanho da amostra e método de cálculo no valor estimado das concentrações médias das variáveis DQO e fosfato. Os resultados mostram que as concentrações médias amostrais calculadas com o método de substituição são imprecisas, pois se posicionam fora do intervalo de confiança e não podem ser reproduzidas. Esses resultados estão de acordo com os estudos de Gleit (1985), Kuttatharmmakul *et al.* (2001), Singh e Nocerino (2002) e Helsel (2004).

Os valores censurados representam um sério problema na análise e interpretação de dados. Por exemplo, o cumprimento dos padrões de descarga de efluentes geralmente é verificado através da comparação entre a concentração média observada de uma amostra em um intervalo de tempo e o valor regulamentado por legislação. No entanto, quando valores censurados estão presentes, a concentração média da amostra não deve ser calculada com o método de substituição, especialmente para amostras com poucas observações. O devido tratamento estatístico dos valores censurados é de vital importância para a qualidade da informação produzida. Outra observação pertinente aos resultados se deve ao método de armazenamento de dados no sistema *Hidroweb*. A arquitetura do banco de dados não permite distinguir valores observados de valores censurados. Logo, utilizar dados das variáveis de qualidade que contenham valores censurados sem verificar os boletins de registro original pode, na maioria das vezes, acarretar em inferências e conclusões errôneas. Esse aspecto foi abordado por Nascimento (2010), apontando um elevado percentual de valores censurados e falhas de dados existentes nas séries históricas das 14 variáveis de qualidade monitoradas ao longo do Rio Cuiabá e Rio das Garças, ambos no Estado de Mato Grosso. Hoje deve-se evitar o uso do método de substituição, porque devido à mudança da técnica de medição ao longo do tempo, possuem múltiplos limites de detecção. Por outro lado, já existem métodos alternativos (paramétricos e não paramétricos) para estatísticas descritivas, testes de hipóteses e regressão, os quais estão disponíveis nos pacotes computacionais *NADA* (HELSEL, 2004), MiniTab e SAS. Os métodos de análise utilizados neste estudo podem ser aplicados como ferramenta de auxílio no processo de gestão dos programas de monitoramento da qualidade da água, com o intuito de reduzir problemas de interpretação e inferências errôneas sobre amostras que contenham dados censurados.

REFERÊNCIAS

AHN, H. (1998) Estimating the mean and variance of censored phosphorus concentrations in Florida rainfall. *Journal of the American Water Resources Association*, v. 34, n. 3, p. 583-593.

AITCHINSON, J. & BROWN, I.A.C. (1957) *The lognormal distribution*. Cambridge: Cambridge University Press.

GILBERT, R.O. (1987) *Statistical methods for environmental pollution and monitoring*. Nova York: Wiley. 336 p.

GILLIOM, R.J. & HELSEL, D.R. (1986) Estimation of distributional parameters for censored trace level water quality data: 1. Estimation techniques. *Water Resources Research*, v. 22, n. 1, p. 135-146.

GLEIT, A. (1985) Estimation for small normal data sets with detection limits. *Environmental Science & Technology*, v. 19, n. 12, p. 1201-1206.

HELSEL, D.R. (2004) *Nondetects and Data Analysis:* statistics for censored environmental data. New York: John Wily and Sons. 250 p.

HELSEL, D.R. & COHN, T.A. (1988) Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research*, v. 24. n. 12, p. 1997-2004.

HELSEL, D.R. & HIRSCH, R.M. (2002) *Statistical Methods in Water Resources*. Techniques of Water Resources Investigations of the United States Geological Survey. Livro 4, Capítulo A3. Disponível em: <http://water.usgs.gov/pubs/twri/twri4a3/>. Acesso em: 12 mar. 2012.

KUTTATHARMMAKUL, S.; MASSART, D.L.; COOMANS, D.; SMEYERS-VERBEK, J. (2001) Comparison of methods for the estimation of statistical parameters of censored data. *Analytica Chimica Acta*, v. 441, n. 2, p. 215-229.

NASCIMENTO, O.C. (2010) *Análise de qualidade dos dados gerados pelos programas de monitoramento da qualidade das águas das bacias hidrográficas do Rio Cuiabá e Rio das Garças*. Dissertação (Mestrado em Recursos Hídricos) – Universidade Federal de Mato Grosso, Cuiabá.

NEWMAN, M.C. & DIXON, P.M. (1990). UNCENSOR: a program to estimate means and standard deviations for data sets with below detection limit observations. *American Environmental Laboratory*, v. 4, n. 90, p. 26-30.

SHE, N. (1997) Analyzing censored water quality data using a non-parametric approach. *Journal of the American Water Resources Association*, v. 33, n. 3, p. 615-624.

SINGH, A. & NOCERINO, J. (2002) Robust estimation of mean and variance using environmental data sets with below detection limits observations. *Chemometrics and Intelligent Laboratory Systems*, v. 60, n. 1-2, p. 69-86.

WENDELBERGER, J. & CAMPBELL, K. (1994) Non-detect data in environmental investigations. In: *American Statistical Association Conference*. Toronto.