Quim. Nova, Vol. 37, No. 1, 104-110, 2014

Revisão

TRATAMENTO DE DADOS CENSURADOS EM ESTUDOS AMBIENTAIS

Cristiano Christofaro[a,*] e Mônica M. D. Leão[b]

[a]Departamento de Engenharia Florestal, Faculdade de Ciências Agrárias, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Rodovia MGT 367 - Km 583, 5000 Diamantina – MG, Brasil

[b]Departamento de Engenharia Sanitária e Ambiental, Escola de Engenharia, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte – MG, Brasil

Recebido em 11/04/2013; aceito em 21/08/2013; publicado na web em 24/09/2013

TREATMENT OF CENSORED DATA IN ENVIRONMENTAL STUDIES. Due to the inherent limitations of the analytical methods of measurement, environmental exposure data often present observations described as below a certain detection limit, also called left-censored data. Censored data directly interferes in almost all types of statistical analyzes, including descriptive parameters, hypothesis testing, confidence intervals, correlations and regressions. In this work, we investigated the performance of the main classes of methods from major publications available in the literature, considering their advantages and limitations. Some criteria for selecting the best method of dealing with censored data are presented.

Keywords: censored data; limit of detection; Kaplan-Meier.

INTRODUÇÃO

Devido às inerentes limitações dos métodos analíticos de mensuração, dados de exposição ambiental freqüentemente apresentam observações descritas como abaixo de um certo limite de detecção. Do ponto de vista estatístico, dados com registros abaixo de um certo limite são denominados "censurados à esquerda" ou simplesmente censurados.[1] Outros termos aplicados para esses limites na literatura incluem "valor crítico", "limite de detecção do método" ou apenas "limite de detecção".[2]

Dados censurados são um problema comum a várias disciplinas do conhecimento, sendo encontrada em estudos da qualidade do ar,[3,4] qualidade da água superficial,[3,5,6] exposição humana a toxinas,[7,8] segurança do trabalho,[9] dentre outras. No entanto, no âmbito das ciências ambientais, relativamente poucos estudos utilizam técnicas de tratamento da censura propostas por outras disciplinas.[2,10]

A censura de dados interfere diretamente em quase todos os tipos de análises estatísticas, incluindo: parâmetros estatísticos básicos (e.g. média, desvio-padrão, etc.),[11–13] intervalos de confiança;[14,15] testes de hipóteses,[16] ajuste de distribuições de probabilidade,[17,18] correlações,[19] análises de regressão e tendências.[20] Dependendo do método utilizado no seu tratamento, os resultados podem sofrer alterações consideráveis, tendo sua interpretação prejudicada.

Apesar desses problemas, os dados censurados não devem ser eliminados da série estudada pois, nessas situações, distorções ainda piores podem ser geradas.[2] Assim, uma vez que a presença de dados censurados prejudica a utilização dos testes estatísticos, técnicas específicas devem ser utilizadas para minimizar a interferência negativa das observações censuradas.[1,2,16]

O objetivo desse trabalho é apresentar uma revisão dos principais métodos de tratamento de dados abaixo do limite de detecção disponíveis na literatura, considerando suas abrangências de aplicações, bem como suas principais limitações, a fim de auxiliar a escolha dos métodos mais adequados de acordo com as características de dados a serem avaliados. Uma aplicação dos métodos apresentados será executada em dados do monitoramento da concentração de chumbo em cursos d'água,[21] buscando demonstrar uma situação prática de seleção dos métodos.

*e-mail: cristiano.christofaro@ufvjm.edu.br

DADOS CENSURADOS À ESQUERDA

O Comitê de Melhoria Ambiental da Sociedade Americana de Química (ACSCEI) define o limite de detecção (LD) como "o menor nível de concentração que pode ser determinado como estatisticamente diferente de uma amostra branca".[2] Uma vez determinado, o limite de detecção pode ser utilizado como um nível de censura para medições subseqüentes, sendo as concentrações observadas abaixo desse limite descritas, simplesmente, como abaixo do limite de detecção.

Os limites de detecção podem ser simples ou múltiplos. No primeiro caso, apenas um valor censurado é utilizado durante todo o estudo. No entanto, devido a alterações metodológicas ou progressos tecnológicos, ocorrem situações em que mais de um limite de detecção é verificado em um mesmo estudo.[22] Nesses casos, considera-se que a série estudada apresenta censura múltipla, também denominada "censura complexa".[9]

MÉTODOS DE TRATAMENTO DOS DADOS CENSURADOS

Atualmente, os métodos de tratamento da censura podem ser divididos em pelo menos quatro classes:[2,23] substituição, métodos paramétricos, métodos robustos e métodos não-paramétricos. Cada um desses métodos apresenta vantagens e limitações, assim como situações e critérios específicos para sua aplicação.

Métodos de substituição

O método mais comumente utilizado para tratamento da censura consiste na simples substituição dos valores não detectados por um valor constante abaixo do limite de detecção. Após essa substituição, as análises estatísticas usuais são feitas, considerando que os dados substituídos correspondem a dados reais. Cada disciplina apresenta sua própria tradição para a escolha do valor para substituição. Algumas recomendam a utilização de metade do limite de detecção,

outras indicam o uso da raiz quadrada do dobro do limite de detecção, zero ou o próprio valor do limite de detecção.[24,25]

No entanto, qualquer valor entre zero e o limite de detecção pode levar a desvios nas estimativas das estatísticas descritivas proporcionais aos valores escolhidos. 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. Apesar dessas substituições serem consideradas métodos "não-paramétricos", estudos recentes têm demostrado que a substituição por valores constantes presume que os dados abaixo do limite de detecção seguem uma distribuição normal (no caso da substituição por metade do limite de detecção) ou triângular (no caso da substituição por LD/$\sqrt{2}$).[23] Por distorcer também os valores do desvio-padrão das amostras, a substituição interfere em todos os testes paramétricos de hipóteses que utilizam essa estatística.[25]

Ao longo de mais de 20 anos, estudos vêm demonstrando que a substituição consiste em um método inadequado para o cálculo de estatísticas descritivas.[12,16,25–28] A agência de proteção ambiental dos EUA recomenda que o método de substituição não seja utilizado para séries de dados com censuras maiores do que 15%, pois podem gerar grandes desvios no cálculo da média e do desvio-padrão, sendo a piora na performance diretamente relacionada ao percentual de censura.[29] Contudo, apesar das críticas, o método da substituição continua sendo amplamente utilizado em estudos ambientais, dada sua simplicidade de implementação.[25]

Estudo testando o desempenho de nove métodos de tratamento em dados com 13,7% a 94,5% de censura, a partir dos resultados da mensuração simultânea em equipamentos com sensibilidades distintas, demonstram que o percentual de censura foi a variável que mais afetou os resultados e que a substituição pelo valor zero e pelo limite de detecção resultou nas piores performances no cálculo da estatística descritiva.[13] Estudos baseados na análise de séries de dados simulados recomendaram que a substituição por metade do limite de detecção não seja utilizada para o cálculo de intervalos de confiança, mesmo para casos de baixo percentual de censura.[14]

O desempenho dos métodos de substituição está intimamente relacionado ao valor escolhido para a substituição, sendo essa escolha desse valor um critério arbitrário. Métodos alternativos para o cálculo de estatísticas descritivas, correlações e regressão, na presença de censura, sem a substituição por valores arbitrários, são comumente utilizados na estatística médica e industrial.[25] Esses métodos podem ser facilmente adaptados para as ciências ambientais, sendo apresentados nos tópicos seguintes.

Métodos paramétricos

Uma outra classe de métodos baseia-se na hipótese de que as observações seguem uma certa distribuição de probabilidade (e.g. lognormal). Os métodos de estimativa baseiam-se principalmente no Estimador de Máxima Verossimilhança (EMV). Nesse caso, os parâmetros de uma distribuição de probabilidades ajustada aos dados são estimados com base nos dados observados acima dos limites de detecção e o percentual de dados abaixo do limite de detecção. Ao considerar o percentual de censura dos dados, o método garante o aproveitamento das informações associadas aos dados censurados.[16,25] A técnica do EMV tende a gerar pouco desvio nos casos em que as observações não-censuradas apresentam um ajuste satisfatório à distribuição presumida e para amostras suficientemente grandes.[2]

No entanto, essas condições são raras em dados obtidos no mundo real e, nos casos em que tais premissas não são satisfeitas, o método pode acabar resultando em estimativas imprecisas. Além disso, a técnica apresenta grande sensibilidade a *outliers*, condição comum em dados ambientais.[10] A condição mais importante a ser cumprida para se obter um desempenho satisfatório consiste no ajuste dos dados à distribuição presumida e, mesmo nesses casos, desvios e baixa precisão podem ser verificados em casos de amostras de tamanho pequeno (n= 5, 10 e 15), consideradas comuns em estudos ambientais.[1,10]

Avaliações baseadas em dados reais consideraram o desempenho das técnicas paramétricas de Máxima Verossimilhança insatisfatórios, superando apenas a substituição por zero e pelo limite de detecção.[13] Já estudos realizados a partir da análise de dados sintéticos gerados por distribuições de probabilidade lognormais consideraram os métodos paramétricos mais adequados para tratamento de dados censurados.[9] No entanto, os autores ressaltam que o método de geração das séries sintéticas pode ter influenciado os resultados. Assim, a interpretação dos resultados da avaliação do desempenho do tratamento da censura por métodos paramétricos nos casos em que dados sintéticos são gerados a partir de distribuições de probabilidade conhecida deve ser feita com cautela.

Contudo, mesmo estudos baseados em análises de séries sintéticas consideram as técnicas paramétricas de máxima verossimilhança inadequadas para os casos em que se verifica múltiplos limites de detecção.[14] Além disso, esses autores consideram as técnicas paramétricas excessivamente sensíveis à presença de *outliers* nas amostras.

Análises de séries de dados artificialmente censurados indicaram um melhor desempenho dos métodos paramétricos de máxima verossimilhança para o cálculo de quantis.[30] No entanto, o referido estudo considerou o desempenho dos métodos de imputação (vide próximo item) melhor do que os métodos paramétricos para o cálculo da média e desvio-padrão.

Métodos robustos

Os métodos denominados "robustos", ou de imputação, consistem no preenchimento dos valores censurados ou perdidos sem a determinação de valores repetidos.[2,16,27] Nessa abordagem, uma distribuição de probabilidade assimétrica é ajustada ao logarítimo dos dados acima do limite de detecção, por meio de técnicas de plotagem (gráfico QQ). Essa distribuição é então utilizada para extrapolar os valores censurados, que passam a fazer parte da amostra. Diferente dos métodos paramétricos que passam a utilizar a distribuição ajustada aos dados diretamente no cálculo dos parâmetros estatísticos, a distribuição é utilizada apenas para a geração de valores das amostras censuradas (Figura 1).

Os valores extrapolados devem ser utilizados apenas para o cálculo de parâmetros e testes estatísticos que consideram a amostra como um todo, não podendo ser utilizados em estimativas que considerem especificamente cada amostra (e.g. testes não-paramétricos).[16] Assim, métodos de tratamento da censura por imputação não devem anteceder a aplicação de correlações, tendências e testes de hipóteses não-paramétricos, uma vez que tendem a criar um ordenamento artificial nas amostras censuradas, podendo alterar significativamente os resultados.

O método é considerado robusto por enfatizar os dados observados.[1] No entanto, a robustez desse método foi verificada apenas nos casos de dados ajustáveis a distribuições com assimetria moderada e baixos valores do coeficiente de variabilidade.[14] Desse modo, pode não ser adequada uma generalização automática para distribuições que não apresentam tais características.

Os métodos robustos apresentam algumas vantagens em relação aos métodos paramétricos no cálculo de médias e desvios-padrão. Uma delas consiste na menor sensibilidade à distribuição de probabilidade ajustada aos dados observados. Dessa forma, apresentam melhores resultados quando aplicados a pequenas amostras. A outra

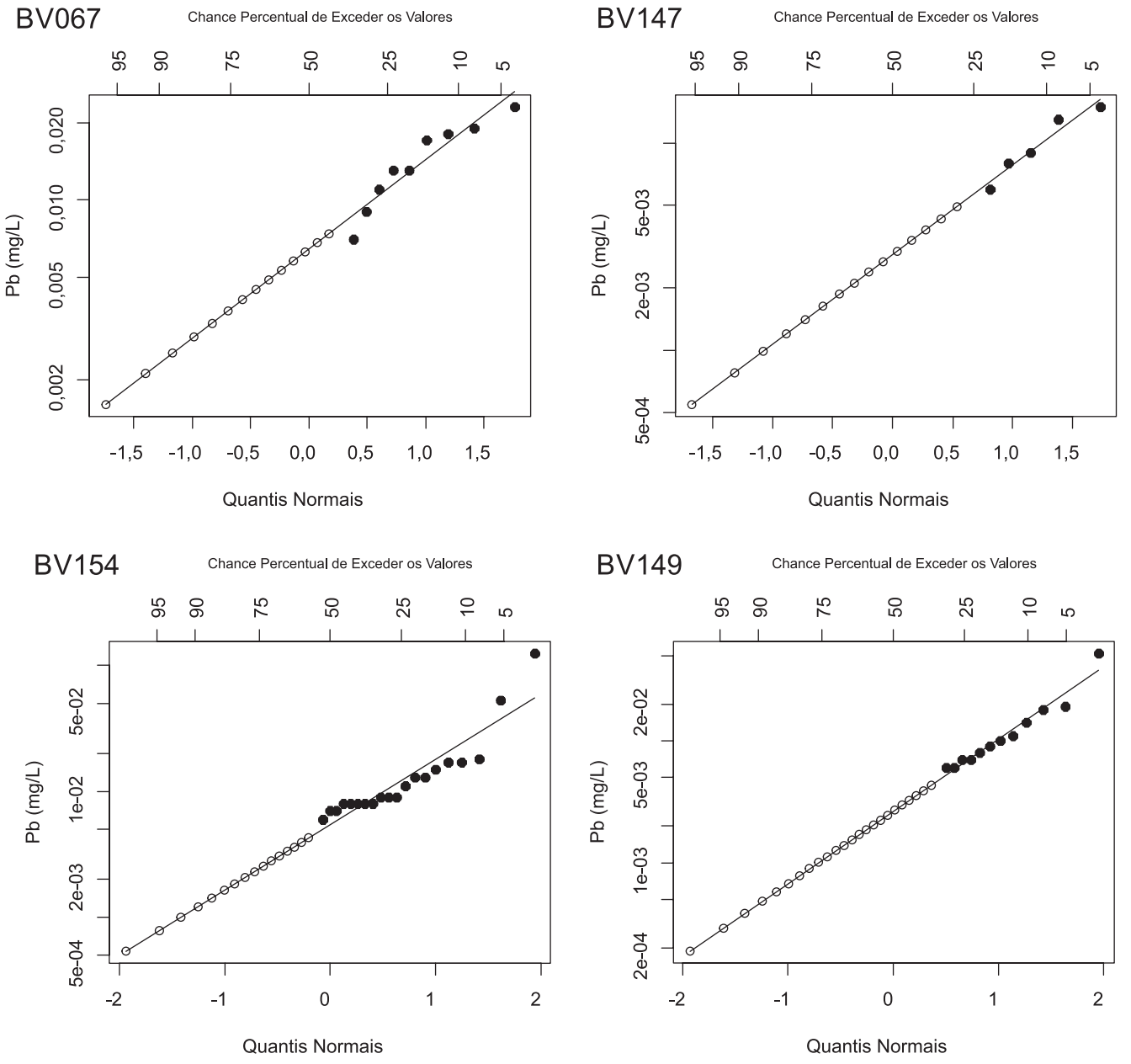


Figura 1. *Aplicação de um método de imputação (ROS) a partir de um gráfico normal QQ em quatro estações de monitamento da concentração de chumbo em cursos d'água da bacia do rio das Velhas - MG. Os valores censurados (pontos brancos menores) foram extrapolados a partir da distribuição normal melhor ajustada ao logarítimo dos dados acima do limite de detecção (pontos pretos). Em BV149 verifica-se um ajuste satisfatório dos dados não censurados à distribuição normal. Em BV154 e BV067 nota-se um ajuste pouco satisfatório dos dados. A estação BV147 apresenta elevado percentual de censura (75%) e os dados extrapolados são gerados a partir de poucas amostras, reduzindo a confiabilidade do método*

está associada ao fato dos valores extrapolados serem diretamente utilizados no cálculo da estatística descritiva, sem desvios oriundos da transformação de unidades.[16] Outra vantagem desse método consiste no fato dos valores extrapolados não apresentarem valores negativos, uma vez que são gerados a partir de uma distribuição assimétrica.[14]

Estudos demonstram que os métodos robustos podem gerar estatísticas básicas com elevada acurácia, mesmo em situações em que 60-70% dos dados apresentem censura.[16,28,30] As médias e variâncias calculadas a partir de métodos robustos foram consideradas satisfatórias quando os *outliers* foram eliminados.[12] No entanto, testes do desempenho de métodos robustos gerados por técnicas gráficas na estimativa de intervalos de confiança, sob diversas condições de assimetria e variabilidade, consideraram seu desempenho muito inferior ao apresentado pela técnica não-paramétrica Kaplan-Meier (KM) (descrita no próximo tópico), sendo seu uso não recomendando para esse fim.[14]

Métodos não-paramétricos

Os métodos não paramétricos recebem esse nome pelo fato de não envolverem o cálculo de parâmetros que descrevam analiticamente um modelo distributivo populacional presumido *a priori* como verdadeiro. Ao invés disso, esses métodos consideram a ordem de classificação dos dados. Métodos não-paramétricos são especialmente úteis para dados censurados pelo fato de otimizarem a utilização das informações disponíveis, já que requerem apenas a ordenação e as posições relativas dos valores dentro de uma série.

Essa abordagem permite a execução, sem qualquer tratamento

adicional, de testes de hipóteses, correlações e análises de tendências não-paramétricas, sendo também imune à presença de *outliers*. Essas características fazem com que esses testes apresentem grande utilidade em estudos ambientais.[16,31,32]

Para o cálculo das estatísticas descritivas das amostras, a abordagem não-paramétrica mais recomendada é conhecida como Kaplan-Meier (KM). O método foi formulado para incluir dados com múltiplos limites de detecção, não requerendo a especificação de uma distribuição de probabilidade.[33] Na abordagem KM, um percentil é gerado a partir da ordenação dos valores sem censura (Figura 2). A geração desses percentis considera o número de valores não detectados acima e abaixo de cada observação. Assim, apesar de não serem calculados percentis para os dados censurados, o percentual de censura nos dados influencia os percentis gerados para as observações detectadas.[16]

O método, originalmente utilizado em análises de sobrevivência, com censura à direita, pode ser facilmente adaptado para tratamento da censura à esquerda a partir da "inversão" dos dados. Essa inversão consiste na subtração de cada amostra por uma constante arbitrária, pouco maior que o valor máximo da amostra. O resultado é uma amostra invertida, ou seja, os valores mínimos da amostra original passam a ser os valores máximos na nova série e vice-versa.[2] Apesar dessa análise não fazer parte da codificação atual dos programas comerciais, o programa R[34] dispõe de pacote específico para o tratamento de dados censurados que realiza os cálculos para censura à esquerda.[32,35]

Os exemplos mais antigos da utilização de análises de sobrevivência para tratamento de dados censurados em estudos ambientais ocorreram na década de 1980 e 1990.[36,37] Desde então, outros trabalhos vêm utilizando essa técnica.[13,14,16,25,32] Avaliações de séries de dados com até 70% de censura demonstraram um melhor desempenho do método não-paramétrico de Kaplan-Meier.[13] Os estudos citados consideram o KM a técnica mais adequada para estimativa da média, variância e intervalos de confiança associados a séries de dados com censura.

Outras vantagens do KM incluem o fato do método poder lidar, sem nenhuma adaptação adicional, com limites múltiplos,[10] além de ser imune à presença de *outliers*, como todo teste não-paramétrico.[31] Um cuidado a ser enfatizado consiste no fato de que, dependendo do percentual de censura dos dados, a média, mas não os percentis, pode apresentar grandes desvios quando calculada por esse método.[16]

Algumas críticas ao desempenho do método KM no cálculo de estatísticas descritivas de dados sintéticos em relação ao método de máxima verossimilhança são verificadas na literatura.[9] Esses autores consideraram o desempenho do tratamento KM comparável ao de métodos de substituição simples. No entanto, é importante ressaltar que os dados desses estudo foram gerados a partir de distribuições de probabilidade previamente definidas, o que pode ter favorecido o desempenho da técnica de máxima verossimilhança.

ESTUDO DE CASO

A bacia hidrográfica do Rio das Velhas, localizada na região central do Estado de Minas Gerais, compreende uma área de 29.173 Km², onde estão localizados 51 municípios, que abrigam uma população de cerca de 5 milhões de habitantes. Os cursos d'água dessa bacia têm sido monitorados desde 1997 no âmbito do programa "Águas de Minas".[21] Nesse estudo, 29 (vinte e nove) estações de monitoramento, distribuídas ao longo da bacia, são utilizadas para a amostragem de diversos parâmetros de qualidade das águas, com freqüências trimestral ou semestral. Dentre os diversos poluentes detectados nos cursos d'água merece destaque o chumbo, que apresenta grande relevância ambiental e à saúde humana.[38]

As características desses dados de monitoramento quanto ao tamanho da amostra e percentual de censura, para os dados obtidos entre 1998 e 2007, em onze locais distintos, são apresentados na Tabela 1. Considerando o limite de detecção de 0,005 mg/L, verifica-se a ocorrência de proporções variadas de concentrações abaixo do limite de detecção (44 a 87%) nos pontos de monitoramento. Essa variação na proporção entre os pontos consiste em uma dificuldade adicional para a seleção do método de tratamento da censura.

Para ilustrar a aplicação dos métodos discutidos ao longo do texto, são apresentados na Tabela 1 o cálculo da média, mediana e desvio padrão do chumbo nessas onze estações de monitoramento, após a aplicação de cindo métodos de tratamento de censura: (1) substituição pela metade do limite de detecção (LD/2); (2) substituição por valor igual ao limite de detecção (LD), (3) método de imputação (ROS); (4) Estimador de Máxima Verossimilhança (EMV); (5) método não paramétrico de Kaplan-Meier (KM). As análises foram realizadas no programa R utilizando o pacote NADA (*Nondetects And Data Analysis*).[35]

De acordo com a Tabela 1 percebe-se que o percentual de censura consiste em uma importante variável a ser avaliada quando da escolha

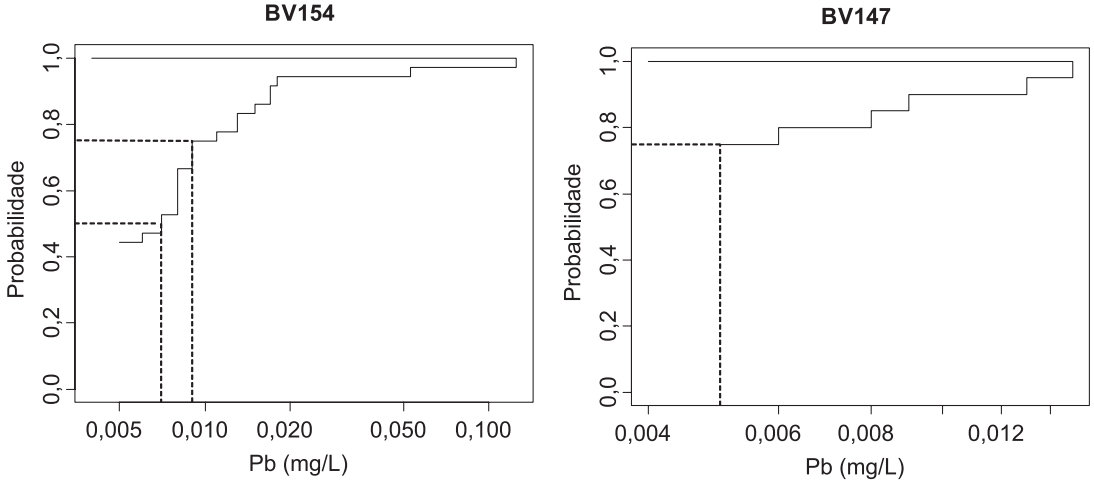


Figura 2. *Probabilidade de sobrevivência gerada pelo método Kaplan-Meier após inversão dos dados de concentração de chumbo (subtração de todos os valores por uma constante maior que o máximo da amostra) em duas estações de monitamento dos cursos d'água da bacia do rio das Velhas – MG (limite de detecção do método igual 0,005 mg/L). As linhas pontilhadas demonstram o cálculo dos percentis. Em BV154 (com 45% de censura) é possível calcular a mediana e percentis superiores (linhas pontilhadas nos percentis 50 e 75). Em BV147 (75% de censura) só é possível calcular percentis iguais ou maiores que 75*

Tabela 1. Média, mediana e desvio padrão das concentrações de chumbo em onze estações de amostragem da bacia do rio das Velhas/MG, entre 1998 a 2007,[21] após cinco métodos distintos de tratamento da censura (KM = Kaplan-Meier, ROS = Imputação, EMV = Estimativa de Máxima Verossimilhança, LD = substituição por valor igual ao limite de detecção e LD/2 = substituição por valor igual à metade do limite de detecção)

Parâmetro	Método	Estações de Amostragem										
		BV154	BV137	BV105	BV037	BV083	BV013	BV067	BV149	BV147	BV162	BV161
n	-	36	20	36	20	36	36	23	36	20	15	15
% Censura	-	44	45	47	50	56	58	61	67	75	80	87
Média (mg/L)	KM	0,013	0,016	0,016	0,014	0,031	0,015	0,010	0,009	0,007	0,005	0,008
	ROS	0,011	0,014	0,015	0,013	0,028	0,011	0,008	0,006	0,004	0,003	0,003
	EMV	0,010	0,017	0,012	0,015	0,016	0,012	0,008	0,006	0,004	0,004	0,003
	LD*	0,012	0,015	0,016	0,013	0,030	0,013	0,009	0,008	0,006	0,005	0,006
	LD/2	0,011	0,014	0,015	0,012	0,029	0,012	0,007	0,006	0,004	0,003	0,004
Mediana (mg/L)	KM	0,007	0,007	0,005	NA	NA	NA	NA	NA	NA	NA	NA
	ROS	0,007	0,008	0,006	0,008	0,003	0,004	0,006	0,003	0,003	0,002	0,001
	EMV	0,006	0,007	0,005	0,006	0,004	0,004	0,004	0,003	0,003	0,003	0,001
	LD	0,007	0,008	0,006	0,006	0,005	0,005	0,005	0,005	0,005	0,005	0,005
	LD/2	0,007	0,008	0,006	0,005	0,003	0,003	0,003	0,003	0,003	0,003	0,003
DP	KM	0,021	0,014	0,042	0,011	0,130	0,022	0,005	0,008	0,003	0,001	0,002
	ROS	0,022	0,015	0,042	0,011	0,128	0,023	0,006	0,009	0,004	0,002	0,004
	EMV	0,016	0,038	0,027	0,036	0,067	0,036	0,013	0,011	0,005	0,002	0,005
	LD	0,021	0,014	0,042	0,011	0,128	0,022	0,006	0,008	0,003	0,001	0,002
	LD/2	0,021	0,015	0,042	0,012	0,128	0,023	0,007	0,009	0,004	0,002	0,003

* LD = 0,005 mg/L.

do método de tratamento dos parâmetros estatísticos apresentados. O método KM na maior parte das vezes resultou nas maiores médias nos casos de censura acima de 50% dos dados. A substituição por valor igual à metade do limite de detecção gerou médias similares ao método ROS nos pontos. Já o método EMV apresentou um comportamento menos previsível, com situações onde alcançou a maior média entre os métodos aplicados (BV137) e a menor média (BV083). Percebe-se a diferença da média quando da escolha do valor a ser utilizado na substituição (LD ou LD/2) acrescenta uma subjetividade extra à análise, já que o responsável pela análise passa a poder escolher um valor para substituição que sabiamente contribua para o aumento ou para a redução da média da amostra.

A mediana foi o parâmetro que apresentou a menor discrepância entre os métodos. Essa menor influência era esperada, uma vez que esse parâmetro é calculado a partir da ordenação dos dados. Contudo, nos casos em que a proporção de censura é maior que 50%, o cálculo da mediana passa a ser drasticamente afetado, principalmente no método KM, que passa a ser incapaz de calcular esse parâmetro estatístico. Para os métodos de substituição verifica-se que censuras acima de 50% dos dados resultam em medianas iguais aos valores de substituição escolhidos (LD ou LD/2). Assim, nessas situações, a utilização do método KM e de substituição são prejudicadas. Ressalta-se que, para censuras acima de 80%, apenas o cálculo de percentis mais elevados (ex. percentil 90) é recomendado.[1,16]

Dentre os métodos utilizados para o cálculo do desvio-padrão, verificou-se uma acentuada discrepância para o método EMV em relação aos demais (Tabela 1). Essa diferença pode ser explicada pelo fato de dados de qualidade da água geralmente não apresentarem ajuste satisfatório à distribuição normal,[31] um dos critérios mais importantes para a utilização do método EMV,[2,31] bem como pela relativamente baixa amostragem ($n < 50$).[2] Assim, para esses dados, o tratamento da censura pelo método EMV deve ser feito com cautela quando se pretende obter o desvio padrão dos dados. Alterações no desvio padrão podem interferir em diversos outros métodos, inclusive em testes de hipótese e avaliações do risco.[31] Consequentemente, quaisquer outras análises estatísticas baseadas no desvio padrão podem sofrer grande influência pela seleção do método EMV. Os demais métodos apresentaram pouca diferença em relação aos valores de desvio padrão calculados.

O exemplo apresentado ilustra o potencial de influência dos métodos de tratamento da censura no cálculo de parâmetros estatísticos básicos. De uma forma geral, percebe-se que existe alguma relação com o percentual de censura apresentado pela amostra e a discrepância entre os métodos utilizados. Ressalta-se que a situação apresentada consiste em um caso particular à série de dados apresentada, não permitindo uma generalização para a seleção de critérios de tratamento da censura. No próximo tópico serão apresentadas sugestões de critérios de escolha do método de tratamento da censura, a partir das características básicas das amostras a serem analisados.

CRITÉRIOS PARA SELEÇÃO DOS MÉTODOS

Estudos do desempenho de técnicas de substituição e de imputação no tratamento de dados de exposição a dioxinas com 56% de censura demonstraram uma variação de 22,8% a 329,6% nos valores estimados para a média.[23] A avaliação do uso da análise de sobrevivência em dados com percentual de censura variado e múltiplos limites de detecção constatou uma diferença significativa entre dois aquíferos a partir das séries analisadas.[36] No entanto, uma nova análise dos mesmos dados, com a substituição dos valores censurados por metade do limite de detecção, resultou em ausência de diferença significativas entre aqüíferos, demonstrando como o tratamento da censura pode interferir na análise dos resultados.[2]

O cálculo da estatística descritiva de compostos orgânicos pelo método Kaplan-Meier em dados com 20% de censura e com oito limites de detecção, resultou em valores considerados satisfatórios.[37]

Tabela 2. Métodos de tratamento da censura indicados de acordo com o percentual de censura e o número de amostras[2]

Percentual de Censura	Amostragem	
	< 50 observações	> 50 observações
< 50% não detectados	Kaplan-Meier	Kaplan-Meier
50% - 80% não detectados	Métodos Robustos	Métodos paramétricos –verossimilhança
> 80% não detectados	Reportar apenas o percentual acima de um limite importante	Reportar percentis elevados (90º, 95º)

Uma nova análise com a substituição por valores constantes demonstrou uma alteração substancial nos valores desses parâmetros.[2] Assim, considera-se que as técnicas paramétricas, de imputação e o Kaplan-Meier são as mais acuradas para a análise estatística de dados que apresentam valores censurados.[2,14,25]

No entanto, os métodos paramétricos requerem uma quantidade suficiente de dados para validar o uso de um modelo de distribuição específico. O modelo de maxima verossimilhança geralmente só apresenta resultados satisfatórios quando a amostragem apresenta 50 ou mais valores não censurados, bem com exige que os dados possam ser satisfatoriamente ajustados a uma distribuição de probabilidade conhecida (e.g. normal, lognormal, gamma, etc.).[39]

Um resumo de critérios passíveis de utilização para a seleção dos métodos mais adequandos[2] é apresentado na Tabela 2. De acordo com essa tabela, o método de Kaplan-Meier é recomendado em todas as situações em que a censura seja inferior a 50% das amostras. Outros autores indicam a aplicação do método KM para o cálculo do intervalo de confiança da média mesmo nos casos em que a censura atinge 70% dos dados.[14]

No entanto, mesmo os critérios apresentados na Tabela 2 são passíveis de questionamentos. Estudos baseados em dados reais consideraram o desempenho das técnicas paramétricas de Máxima Verossimilhança para dados com menos de 70% da censura insatisfatórios, sendo comparáveis à substituição por zero e pelo limite de detecção.[13] Esses resultados insatisfatórios podem estar relacionados à dificuldade de ajuste da amostra a uma distribuição de probabilidade conhecida.

A Tabela 2 restringe a escolha do método apenas em relação ao tamanho da amostra e percentual de censura. No entanto, outros critérios importantes devem ser considerados, como número de *outliers* e presença de múltiplos limites de detecção. A ocorrência dessas situações tende à escolha de métodos não-paramétricos. Desse modo, os critérios da tabela acima devem ser aplicados com cautela na seleção da técnica mais adequada.

Atualmente, poucos esforços vêm sendo verificados na literatura especializada para a criação de novos métodos para o tratamento da censura de dados ambientais. Uma abordagem alternativa para a estimativa dos valores censurados considera a dependência entre as variáveis, quantificada via cadeia de Markov.[40] No entanto, a eficiência dessa abordagem não foi testada em comparação às abordagens tradicionais. Mesmo as abordagens tradicionais apresentam desempenho afetado por características específicas dos dados (e.g. percentual de censura, *outliers,* censura múltipla, qualidade do ajuste a distribuições de probabilidade) bem como fatores relacionados à origem da série testada (dados reais ou séries sintéticas). Verifica-se, de uma forma geral, que a abordagem paramétrica tende a ter resultados melhores nos estudos baseados em séries sintéticas[9] e a abordagem não-paramétrica em estudos baseados em dados reais.[13]

CONCLUSÕES

A partir da análise dos trabalhos que abordaram os principais métodos de tratamento da censura, verifica-se que a seleção do método de tratamento dos dados abaixo do limite de detecção é uma importante questão para reduzir as incertezas em estudos ambientais. A escolha da técnica a ser adotada deve ser baseada nas características dos dados. Especificamente, essa escolha das técnicas dependerá do tamanho da amostra, do percentual de censura, da presença de *outliers* e da qualidade do ajuste a alguma distribuição paramétrica.

Em geral, os métodos mais utilizados de tratamento da censura consistem naqueles descritos ao longo desse texto, principalmente os métodos de substituição por valores constantes, considerados mal fundamentados e com pior desempenho.[10,13,25] A aparente praticidade da aplicação dos métodos de substituição não deve ser utilizada como justificativa para a seleção desses métodos pois, atualmente, existem pacotes estatísticos gratuitos para a aplicação de métodos mais adequados.[32,34,35,39]

Apesar da falta de um consenso, verifica-se que os métodos não-paramétricos tendem a apresentar melhor desempenho no cálculo de estatísticas descritivas em situações frequentemente verificadas em dados ambientais, tais como presença de *outliers*, múltiplos limites de detecção e ajuste insatisfatório a distribuições de probabilidade, mesmo com mais de 50% de censura nos dados. Considerando o relativo sucesso verificado para outros métodos em relação aos métodos tradicionalmente utilizado de substituição, esforços devem ser direcionados para garantir a seleção adequada, bem como para o desenvolvimento de métodos ainda mais confiáveis de tratamento da censura, evitando assim desvios indevidos nas análises de dados em estudos ambientais.

MATERIAL SUPLEMENTAR

Um breve tutorial para a análise de dados com o pacote NADA (*Nondetects And Data Analysis*) no programa R está disponível em http://quimicanova.sbq.org.br, na forma de arquivo PDF, com acesso livre.

AGRADECIMENTOS

Ao Instituto Mineiro de Gestão das Águas – IGAM pela livre disponibilização dos dados de monitoramento da bacia do Rio das Velhas/MG na internet.

REFERÊNCIAS

1. Helsel, D. R.; *Environ. Sci. Technol*. **1990**, *24*, 1766.
2. Helsel, D. R.; *Nondetects and Data Analysis. Statistics for Censored Environmental Data*, Wiley: New York, 2005.
3. Frey, H. C.; Zhao, Y.; *Environ. Sci. Technol*. **2004**, *38*, 6094.
4. Alves, C.; Pio, C.; Gomes, P.; *Quim. Nova* **2006**, *29*, 477.
5. Christofaro, C.; Leão, M. M. D.; *Journal of Water and Environment Technology* **2009**, *7*, 317.
6. Farias, J. dos S.; Milani, M. R.; Niencheski, L. F. H.; Paiva, M. L.; *Quim. Nova* **2012**, *35*, 1401.
7. Perkins, J. L.; Cutter, G. N.; Cleveland, M. S.; *Am. Ind. Hyg. Assoc. J.* **1990**, *51*, 416.

8. Almeida, F. V.; Centeno, A. J.; Bisinoti, M. C.; Jardim, W. F.; *Quim. Nova* **2007**, *30*, 1976.
9. Hewett, P.; Ganser, G. H.; *Ann. Occup. Hyg.* **2007**, *51*, 611.
10. Field, M. S.; *Water Res*. **2011**, *45*, 3107.
11. Kroll, C. N.; Stedinger, J. R.; *Water Resour. Res*. **1996**, *32*, 1005.
12. Singh, A.; Nocerino, J.; *Chemom. Intell. Lab. Syst.* **2002**, *60*, 69.
13. Antweiler, R. C.; Taylor, H. E.; *Environ. Sci. Technol.* **2008**, *42*, 3732.
14. Singh, A.; Maichle, R.; Lee, S. E.; Sibert, C.; *On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations*, USEPA: Washington, 2006.
15. Zhao, Y.; Frey, H. C.; *Risk Anal.* **2004**, *24*, 1019.
16. Helsel, D. R.; *Environ. Sci. Technol.* **2005**, *39*, 419A.
17. Govaerts, B.; Beck, B.; Lecoutre, E.; Bailly, C.; Vanden Eeckaut, P.; *Atmos. Environ.* **2005**, *16*, 109.
18. Zolezzi, M.; Cattaneo, C.; Tarazona, J. V.; *Environ. Sci. Technol.* **2005**, *39*, 2920.
19. Newton, E.; Rudel, R.; *Environ. Sci. Technol.* **2007**, *41*, 221.
20. Hopke, P. K.; Liu, C.; Rubin, D. B.; *Biometrics* **2001**, *57*, 22.
21. Instituto Mineiro de Gestão das Águas – IGAM; *Monitoramento Das Águas Superficiais Na Bacia Do Rio Das Velhas 1998-2007*, IGAM: Belo Horizonte, 2008.
22. Neerchal, N. K.; Brunenmeister, S. L. Em *Environmental Statistics, Assessment and Forecasting*; Cothern, R. C.; Ross, N. P, eds.; Lewis Publishers: Boca Raton, 1993.
23. Baccarelli, A.; Pfeiffer, R.; Consonni, D.; Pesatori, A. C.; Bonzini, M.; Patterson, D. G., Jr; Bertazzi, P. A.; Landi, M. T.; *Chemosphere* **2005**, *60*, 898.
24. Sanford, R. F.; Pierson, C. T.; Crovelli, R. A.; *Math. Geol.* **1993**, *25*, 59.
25. Helsel, D. R.; *Chemosphere* **2006**, *65*, 2434.
26. Gilliom, R. J.; Helsel, D. R.; *Water Resour. Res.* **1986**, *22*, 135.
27. Helsel, D. R.; Cohn, T. A.; *Water Resour. Res*. **1988**, *24*, 1997.
28. Lubin, J. H.; Colt, J. S.; Camann, D.; Davis, S.; Cerhan, J. R.; Severson, R. K.; Bernstein, L.; Hartge, P.; *Environ. Health Perspect*. **2004**, *112*, 1691.
29. Suter, G. W. I.; Vaughan, D. S.; Gardner, R. H.; *Ecological Risk Assessement Issue Paper*, USEPA: Washington, 1994.
30. Huybrechts, T.; Thas, O.; Dewulf, J.; Langenhov, H. Van; *J. Chromatogr*. **2002**, *975*, 123.
31. Helsel, D. R.; Hirsch, R. M.; *Statistical Methods in Water Resources*, U.S. Geological Survey: Washington, 2002.
32. Lee, L.; Helsel, D.; *Comput. Geosci.* **2007**, *33*, 696.
33. Kleinbaum, D. G.; Klein, M.; *Survival Analysis: A Self-Learning Text*, 2th ed., Springer: New York, 2005.
34. R Development Core Team; *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria. http://www.R-project.org/, acessada em Junho 2013.
35. Lee, L.; *NADA: Nondetects And Data Analysis for Environmental Data; R package version 1.5-5*. http://CRAN.R-project.org/package=NADA, acessada em Junho 2013.
36. Millard, S. P.; Deverel, S. J.; *Water Resour. Res*. **1988**, *24*, 2087.
37. She, N.; *J. Am. Water Resour. Assoc.* **1997**, *33*, 615.
38. Chen, C. Y.; Folt, C. L.; *Environ. Sci. Technol.* **2000**, *32*, 117.
39. Lee, L.; Helsel, D.; *Comput. Geosci.* **2005**, *31*, 1241.
40. Zhu, Z. J. Y.; *J. Environ. Inf.* **2004**, *4*, 48.

TRATAMENTO DE DADOS CENSURADOS EM ESTUDOS AMBIENTAIS

Cristiano Christofaro[a,*] e Mônica M. D. Leão[b]
[a]Departamento de Engenharia Florestal, Faculdade de Ciências Agrárias, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Rodovia MGT 367 - Km 583, 5000 Diamantina – MG, Brasil
[b]Departamento de Engenharia Sanitária e Ambiental, Escola de Engenharia, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte – MG, Brasil

Análises de dados censurados no pacote NADA (Nondetects And Data Analysis)

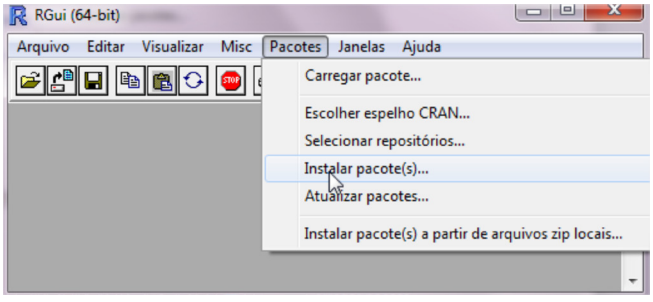
R é uma linguagem e um ambiente utilizado para análises estatísticas e na produção de gráficos. O usuário pode realizar estatísticas a partir de funções existentes (já incluídas no pacote básico ou via pacotes extras) ou pela programação de novas rotinas/extensões. Para a análise de dados abaixo do limite de detecção, o pacote mais conhecido e utilizado é denominado NADA (*Nondetects And Data Analysis for environmental data*)[1] que realiza análises estatísticas mais comumente aplicadas em pesquisas nesse campo.

Abaixo os principais passos necessários para rodar o pacote NADA no programa R pela primeira vez em ambiente Windows:

1 – Vá ao site do R e instale a versão mais atual do programa, conforme intruções no site http://www.r-project.org/

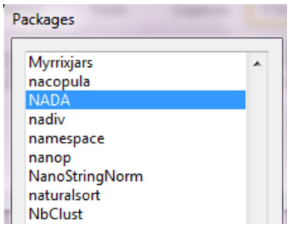
2 – Instalação do pacote NADA (*Nondetects And Data Analysis for environmental data*): O programa R já vem com pacotes básicos. Análises mais especializadas requrem pacotes especiais. Esses pacotes precisam ser instalados apenas uma vez. Para realizar essa instalação vá em:

Pacotes -> Instalar Pacote(s)...



Selecione agora sua localização (ou qualquer outro local) e clique OK.

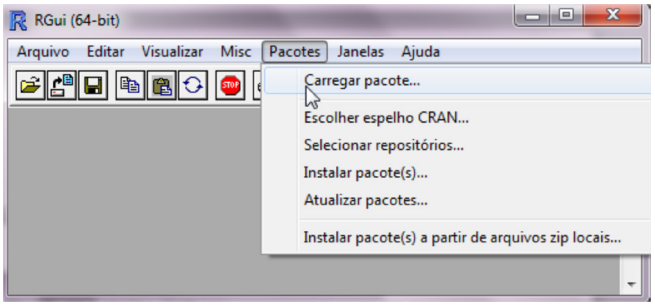
Selecione na lista o pacote para instalar, nesse caso NADA, e clique OK.



*e-mail: cristiano.christofaro@ufvjm.edu.br

5. Carregue o pacote NADA, necessário para os dados censurados:

Pacotes -> Carregar Pacotes...



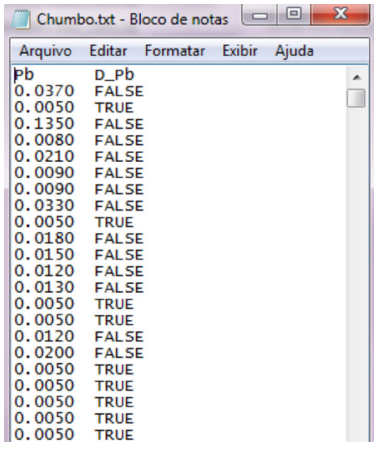
Selecione na lista NADA e clique OK.

Outra forma de carregar o pacote NADA é escrevendo o comando abaixo no R console:

```
> library(NADA)
```

Obs: O carregamento do pacote, por qualquer um dos dois métodos, deve ser feito toda vez que o R for aberto.

6. Leitura dos dados. Para uso no pacote NADA os dados precisam apresentar uma formatação específica: devem haver duas colunas para cada variável, uma com os resultados das amostras e outra com o código TRUE para indicar os valores censurados e FALSE para os não censurados. No exemplo abaixo, a coluna 'Pb' apresenta as concentrações de chumbo na água e a coluna 'D_Pb' indica presença (TRUE) ou ausência (FALSE) de censura para a respectiva amostra.



Após gravar os dados no formato 'txt' o comando abaixo deve ser utilizado para a leitura dos dados pelo R:

```
dados.Pb = read.table("C:/Chumbo.txt", header = TRUE)
```

Obs:

- Nesse exemplo, os dados foram originalmente salvos com o nome "Chumbo.txt" e foram gravados no endereço 'C:/'. O objeto 'dados.Pb' passa a ser a referência do R para essas variáveis. O usuário pode escolher qualquer nome para esse objeto e deve alterar o endereço do arquivo de acordo com a pasta onde o arquivo '.txt' foi salvo.
- O programa diferencia letras maiúsculas e minúsculas. Então para o R "Chumbo.txt" é diferente de "chumbo.txt" e "dados.Pb" é diferente de "Dados.Pb".

7. A partir de então é possível rodar as analises disponibilizadas no pacote NADA. Os nomes após o caracter '$' correspondem à coluna de interesse no objeto com os dados (nesse caso, dados.Pb). Alguns comandos importantes desse pacote são:

```
# criar boxplot
> cenboxplot(dados.Pb$Pb, dados.Pb$D_Pb)
#calculo do mínimo, máximo, número de dados censurados e percentual de censura
> censummary(dados.Pb$Pb, dados.Pb$D_Pb)
# cálculo da mediana, média e desvio padrão (por KM, ROS e MLE)
> censtats(dados.Pb$Pb, dados.Pb$D_Pb)
```

Outras opções de configuração e questões para entendimento dos comandos acima podem ser visualizadas utilizando o comando help('nome do comando').

Exemplos: help(cenboxplot), help(censtats)

8 - Outras informações e textos sobre o R podem ser visualizadas em documentos *online* disponíveis site oficial: www.r-project.org e http://cran.r-project.org/web/packages/NADA/index.html

REFERÊNCIAS

1. Lee, L.; *NADA: Nondetects And Data Analysis for Environmental Data*, R package version 1.5-5. http://CRAN.R-project.org/package=NADA, 2013