# APLICAÇÃO DO MODELO DE DOIS COMPONENTES PARA CÁLCULO DE ERRO, À ANÁLISE DE DADOS DE MONITORAÇÃO AMBIENTAL

Maria Angélica G.Carvalho e Goro Hiromoto

Instituto de Pesquisas Energéticas e Nucleares - IPEN-CNEN/SP
Av. Lineu Prestes 2.242
05508-900 Butantã, São Paulo, SP, Brasil

## RESUMO

A análise e interpretação de resultados de um Programa de Monitoração Ambiental (PMA) envolvem estimativas de médias representativas do conjunto de dados que são fortemente afetadas pelos erros associados a cada medida. O modelo de dois componentes decompõe o erro analítico associado ao resultado de medição em um erro aditivo e um erro multiplicativo e descreve seu comportamento no domínio pleno das concentrações. Este trabalho descreve a aplicação deste modelo à análise de resultados gerados em um programa de monitoração ambiental. Os parâmetros do modelo, estimados através de ajuste, são usados para recalcular os desvios e as médias representativas são estimadas novamente. Estas médias incorporam, através da estimativa dos dois parâmetros do modelo, informações referentes a variações e desvios do protocolo analítico que ocorreram ao longo do período de aquisição de dados e também corrigem a distorção do uso de um modelo não adequado na estimação dos erros analíticos. Os resultados mostram que estas médias, estimadas considerando os erros originais, são menores que aquelas estimadas a partir dos erros previstos pelo modelo, indicando que os erros publicados foram subestimados.

Keywords: environmental data, statistical analysis, two-component model, uncertainty.

## I. INTRODUÇÃO

A análise e interpretação de resultados de um Programa de Monitoração Ambiental (PMA) envolvem estimativas de médias representativas do conjunto de dados. Estas estimativas são fortemente afetadas pelos erros analíticos associados às medidas.

Dentre os modelos existentes para estimar estes erros analíticos, dois são mais usados. Um assume erro relativo constante e é adequado à faixa de concentrações mais altas porém falha nas concentrações mais baixas; o outro modelo, que prevê desvio constante, descreve bem o comportamento do erro analítico para concentrações próximas de zero porém não descreve a realidade para concentrações intermediárias e altas. O modelo proposto por Rocke & Lorenzato [1] decompõe o erro analítico associado a um resultado de medição em dois componentes, um aditivo e um multiplicativo, e descreve o comportamento desses erros no domínio inteiro das concentrações.

Neste trabalho, o modelo de dois componentes para erros de medição, é aplicado na análise de um conjunto de resultados analíticos gerados na execução de um programa de monitoração ambiental. Os valores a serem analisados, por terem sido coletados ao longo de vários anos, reúnem resultados com precisões diferentes, afetados que são por variações em calibrações, mudanças de operador, em equipamentos, ou alterações no protocolo analítico. Os parâmetros do modelo de dois componentes irão refletir estas variações e permitir o cálculo de uma estimativa do desvio para cada medida incorporando estas informações à média representativa do conjunto de resultados.

O uso de um modelo não adequado na estimativa dos erros analíticos pode acarretar graves distorções na estimativa das médias das concentrações quando o intervalo em que estas se apresentam extrapola àquele ao qual o modelo se aplica. Por exemplo, quando o modelo utilizado assume desvio relativo constante (adequado para concentrações mais altas), os erros associados às concentrações próximas de zero são subestimados. Desta forma as médias estimadas sofrem uma ponderação não condizente com a realidade. A modelagem dos erros através do modelo de dois componentes irá minimizar esta tendenciosidade na estimativa das médias, uma vez que ele se aplica ao intervalo pleno das concentrações.

É apresentado um caso onde o modelo é aplicado aos erros analíticos associados às medidas publicadas ao longo de 10 anos de execução de um PMA. Após a estimativa dos parâmetros, os desvios são recalculados e é feita a estimativa da média representativa da amostra.

## II. MODELO DE DOIS COMPONENTES

Modelo. O melhor meio de avaliar a incerteza de uma medida é, sem dúvida, a análise estatística de uma série de observações repetidas (sob as mesmas condições) desta medida [2]. Esta é, entretanto, uma alternativa pouco plausível quando se trata de análises executas em rotina, como é o caso de análises realizadas para atender a PMA's. A avaliação da incerteza da medida é realizada, na prática, usando um modelo matemático e a lei de propagação de erros.

Para medidas realizadas através de um determinado método analítico específico, o desvio padrão apresenta-se proporcional à concentração medida na faixa das concentrações mais altas, porém ao atingir a região das baixas o sistema de detecção tem dificuldade de distinguir diferenças muito pequenas num sinal fraco e o desvio padrão para estas concentrações tende a se tornar constante [3].

Vários modelos têm sido propostos para descrever a relação entre a medida da concentração e seu desvio padrão. Um modelo tradicionalmente usado é o modelo linear de primeira ordem que pode ser expresso pela expressão:

$$x = \mu + \varepsilon \qquad (1)$$

onde x é o resultado da medição, $\mu$ é o valor verdadeiro e $\varepsilon$ é a incerteza associada à medida. Neste modelo os erros são considerados independentes, normalmente distribuídos com média 0 e variância constante. Este modelo se ajusta bem para valores de concentrações muito baixas, porém, assumindo que o valor absoluto do erro não é relacionado com a concentração, ele contradiz evidências experimentais já observadas em muitos métodos analíticos [4].

Outros modelos foram propostos considerando o erro proporcional à concentração, ou com comportamento exponencial ou quadrático. Entretanto, os erros analíticos previstos por estes modelos tendem rapidamente a zero para baixas concentrações.

A expressão analítica do modelo proposto por Rocke & Lorenzato e que descreve o comportamento do desvio padrão numa faixa ampla do espectro das concentrações é

$$x = \mu^2 e^{\eta} + \varepsilon \qquad (2)$$

onde $\eta$ é o erro proporcional com distribuição normal, $N(0,\sigma_\eta)$ ,exibido em concentrações relativamente altas, e $\varepsilon$ é o erro apresentado primariamente nas baixas concentrações também com distribuição normal; $N(0,\sigma_\varepsilon)$.

Este modelo aproxima um desvio padrão constante para as concentrações baixas e um desvio relativo constante para concentrações mais altas descrevendo mais adequadamente o comportamento do erro analítico observado experimentalmente que os modelos já citados. Uma grande vantagem deste modelo é sua aplicação a concentrações intermediárias, onde os outros modelos falham.

Metodologia de Tratamento do Erro Analítico. O conjunto de dados gerados pela execução de um Programa de Monitoração Ambiental (PMA) contém resultados de medidas de uma mesma grandeza e de uma mesma matriz, realizadas ao longo de vários anos em locais pré-selecionados. Os resultados destas medidas costumam ser expressos na forma:

$$(x_i; \delta_{i)}) \, L \qquad (3)$$

onde $x_i$ é um número que representa a grandeza medida, $\delta_i$ representa o erro analítico e L é a unidade de medida no qual o resultado foi expresso. Este erro analítico apresenta grande importância na análise e interpretação dos resultados do PMA uma vez que vão influenciar fortemente a média representativa das concentrações a serem referenciadas nos estudos ambientais, sejam eles descritivos ou para cumprimento de legislação específica.

Na interpretação e tratamento destes erros é necessário levar em consideração, não só o método de determinação e modelo empregado como também o fato de que, ao se tratar de medidas geradas durante um longo período de tempo, incertezas de natureza diferente daquelas previstas e consideradas nos programas de qualidade podem acontecer no processo analítico. São pequenas alterações que ocorrem na rotina diária do laboratório e que eventualmente escapam aos mecanismos de controle de qualidade estabelecidos no protocolo analítico, como substituição do operador, alterações no procedimento analítico, substituição de reagentes, etc. Entretanto, quando o conjunto de resultados é gerado no mesmo laboratório, segundo o mesmo protocolo analítico, estes números guardam uma coerência quanto a sua natureza e modelo utilizado.

A maioria dos resultados publicados usa um dos dois modelos tradicionais citados na literatura. Estes dois modelos funcionam bem para valores próximos a zero ou para valores mais altos de concentração respectivamente, porém falham ao serem usados em um intervalo mais amplo de concentrações. Os resultados analíticos que são tratados em PMA's são tipicamente valores baixos e intermediários e se situam na região de descontinuidade entre os dois modelos. O modelo proposto por Rocke & Lorenzato ajusta-se ao intervalo pleno onde se espera encontrar os resultados de medidas de monitoração ambiental.

Para interpretar estatisticamente os conjuntos de dados gerados em um período longo de tempo, a proposta deste trabalho é fazer uma transformação nos erros, modelando-os sob o modelo proposto.

A expressão para a variância da medida neste modelo é :

$$V(x) = s_e^2 + m^2 e^{s_h^2}(e^{s_h^2} - 1) \qquad (4)$$

Considerando $x_i = \mu_i$ e $d_i = \sqrt{V(x_i)}$, temos:

$$d_i = \sqrt{s_e^2 + x_i^2 e^{s_h^2}(e^{s_h^2} - 1)} \qquad (5)$$

Esta expressão nos permite estimar os parâmetros $\sigma_\varepsilon$ e $\sigma_\eta$.

Cada incerteza reportada será então transformada para:

$$d_i^{'} = \sqrt{\hat{s}_e^{\,2} + x_i^2 e^{\hat{s}_h^2}(e^{\hat{s}_h^2} - 1)} \qquad (6)$$

onde $\hat{s}_e$ e $\hat{s}_h$ são as estimativas dos parâmetros obtidas em (5).

O conjunto de dados a ser analisado estatisticamente passará a ser:

$$(x_i;\ d_i^{'}) \qquad (7)$$

onde $x_i$ é o valor da concentração como foi reportado pelo laboratório e o valor $d_i^{'}$ é o desvio modificado que representa a incerteza da medida segundo o modelo de dois componentes para fins de tratamento estatístico. Isto significa, na prática, uma ponderação das concentrações que deverá ser considerada na estimação dos parâmetros que irão caracterizar o conjunto de dados. A interpretação das análises estatísticas a serem aplicadas deverá sempre estar atenta para o fato de estar tratando com dados transformados e comparar seus resultados como os dados originais.

Os parâmetros $\hat{s}_e$ e $\hat{s}_h$ fornecem um indicador da qualidade dos dados contidos no conjunto original dos dados, uma vez que eles expressam globalmente os desvios nos resultados oriundos da modelagem incorreta dos erros e também do afastamento do protocolo analítico ocorrido na execução individual de cada análise. O valor de $\hat{s}_e$ está relacionado ao limite de detecção, uma vez que ele corresponde ao erro analítico constante que o método apresenta para as concentrações próximas ao zero. O valor de $\hat{s}_h$, por seu lado, demonstra a precisão alcançada pelo método no período de tempo considerado.

## III. RESULTADOS

A metodologia apresentada foi aplicada a um conjunto de resultados de análises de alfa total na matriz aerossol. O protocolo analítico empregado nestas análises inclui coleta em filtros de fibra de vidro de dimensões de 20cm por 25cm, a uma vazão média de 1,1 $m^3$/min num período de 24 horas; amostrador de grande volume de ar a um metro de altura do solo; contagem alfa total de uma fração do filtro [5].

São três locais de amostragem cobrindo um período de cerca de 11 anos num total de 234 registros. Os resultados foram publicados em seu valor numérico (evitando a forma “menor que”) de modo que aparecem resultados negativos e positivos no conjunto estudado.

As Figs. 1 e 2 mostram o comportamento do erro analítico em relação à concentração alfa total para o conjunto de resultados relativos aos três pontos de aerossol.

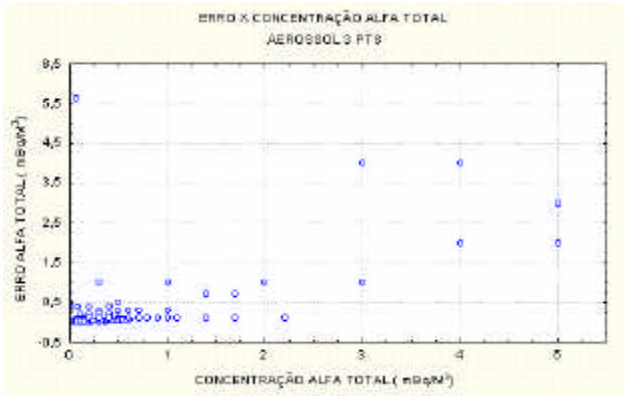


Figura 1. Erro x Concentração Alfa Total.

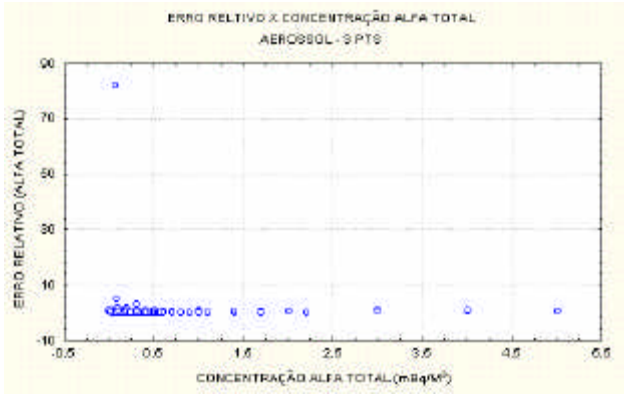


Figura 2. Erro relativo x concentração Alfa Total.

Dados discrepantes. O estudo dos dados discrepantes através do diagrama de caixa dimensionado pela média, desvio padrão e critério 1,5 σ [6] é ilustrado na Fig.3. O ponto identificado como valor extremo foi excluído do conjunto na estimação dos parâmetros.

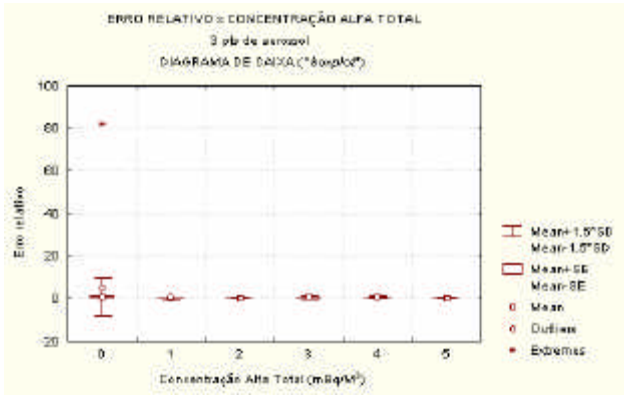


Figura 3. Estudo dos dados discrepantes.

Estimativas Iniciais para o Ajuste. O significado físico de cada parâmetro foi utilizado para encontrar as estimativas dos valores iniciais a serem usadas nos procedimentos de ajuste. Estas estimativas foram feitas através do gráfico erro x concentração mostrado na Fig 1. A estimativa de $\sigma_\varepsilon$ é

estimada pela média dos valores correspondentes às concentrações próximas de zero , região onde este erro deveria ser constante. A estimativa para $\sigma_\eta$ é obtida através da inclinação da reta na região das concentrações mais altas, quando o erro analítico deveria ser proporcional à concentração.

Ajuste. A estimação destes parâmetros foi feita pelo método dos mínimos quadrados não lineares usando o software Statistica [7] e seus resultados são apresentados na TABELA 1.

TABELA 1. Resultado dos ajustes

| | Estimativa | |
|---|---|---|
| | inicial | final |
| $\hat{s}_e$ (mBq/m3) | 0,02 | 0,0289 |
| $\hat{s}_h$ | 0,5 | 0,427 |
| SQRa | 20,45 | |

a . Soma dos quadrados dos resíduos

A qualidade das estimativas iniciais pode ser observada pela rapidez na convergência do ajuste, alcançado em 3 iterações, e na proximidade ao valor ajustado. O resultado do ajuste é mostrado graficamente na Fig. 4 onde os resultados experimentais são apresentados juntamente com a função prevista pelo modelo.

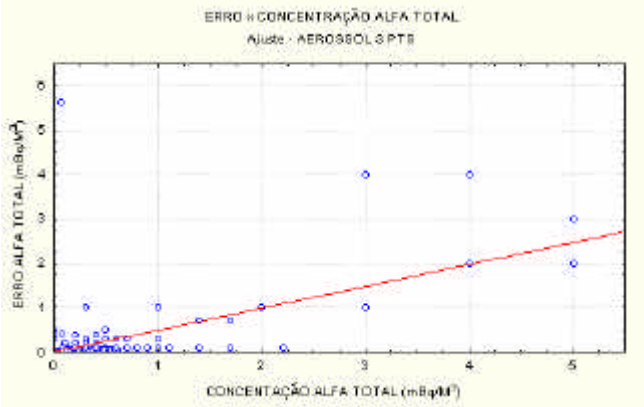


Figura 4. Visualização do ajuste.

O significado físico dos parâmetros previstos pelo modelo permite calcular o limite de detecção e a precisão, projetados para o conjunto de dados e que representam características do protocolo analítico no período analisado.

Estudo das médias. A utilização do valor estimado para os parâmetros $\hat{s}_e$ e $\hat{s}_h$ no cálculo das médias das concentrações para cada local de amostragem, inclui nesta média mais informações que a média calculada diretamente dos resultados publicados; esta média seria menos afetada pelos erros provenientes de desvios pontuais do protocolo analítico como aqueles causados por problemas no equipamento, mudança de operador, etc e que não foram detectados pelos programas de controle de qualidade.

Para o estudo dos valores das médias para cada um dos três pontos foram calculadas a média aritmética, a média ponderada pelo inverso da variância e a média ponderada pelo inverso da variância prevista pelo modelo.

A média aritmética é calculada apenas como referência, uma vez que tratando se de resultados de medição onde o desvio não é constante, ela não é aplicável.

Deve-se comparar, portanto a média ponderada, calculada pela expressão abaixo [8], usando os dados publicados e a média ponderada usando os desvios previstos pelo modelo.

$$\text{Média :} \qquad m = \frac{\sum\left[\frac{x_i}{s_i^2}\right]}{\sum\left(\frac{1}{s_i^2}\right)} \qquad (8)$$

Os resultados obtidos são apresentados na TABELA 2 onde aparecem os valores das médias estimadas para os conjuntos de dados relativos a cada local amostrado. Os valores estimados para o conjunto de todos os dados (234 registros) são mostrados na tabela apenas como referência na análise estatística, uma vez que estas médias não têm significado físico.

TABELA 2. Médias individuais

| Local | 3 pts | 01 | 02 | 03 |
|---|---|---|---|---|
| $N^0$ de registros | 234 | 60 | 83 | 91 |
| Média simples | 0,497 | 0,475 | 0,557 | 0,456 |
| Média ponderada | 0,0126 | 0,0069 | 0,0107 | 0,0235 |
| Média ponderada modelo | 0,0639 | 0,0701 | 0,0639 | 0,0604 |

Os gráficos ilustrativos do ajuste para cada local amostrado são apresentados nas Figs 5, 6 e 7.

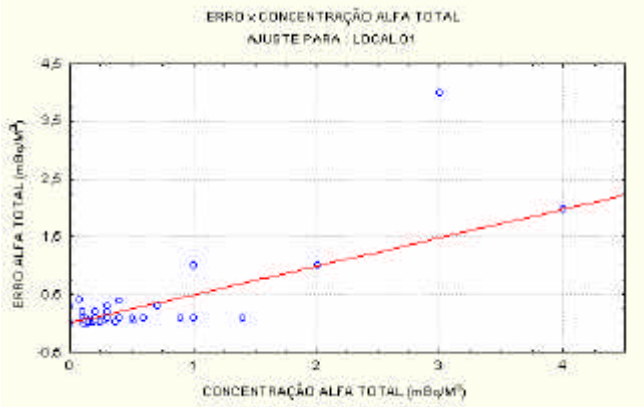


Figura 5. Visualização do Ajuste Local 01.

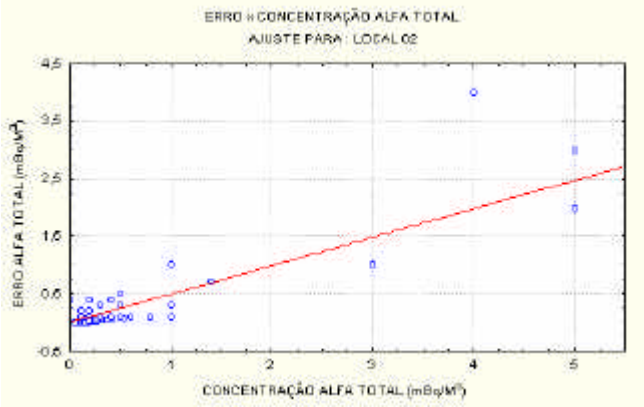


Figura 6. Visualização do Ajuste Local 02.

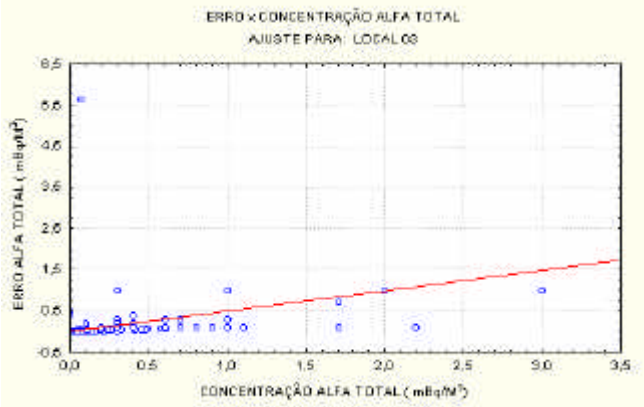


Figura 7. Visualização do Ajuste Local 03.

## IV. CONCLUSÕES

As médias estimadas pelos resultados originais apresentam-se bem menores do que aquelas alcançadas com a aplicação do modelo de dois componentes. Isto é uma indicação de que os erros analíticos associados às medidas de concentração foram subestimados, principalmente na região de concentrações mais baixas onde se encontra a maioria dos resultados apresentados. A aplicação do modelo de dois componentes, além de minimizar este efeito, incorpora na estimativa das médias correções relativas aos desvios no protocolo analítico ocorridos ao longo do processo de geração de resultados.

## AGRADECIMENTOS

Os autores agradecem ao Centro de Desenvolvimento da Tecnologia Nuclear (CDTN) pelo apoio durante o desenvolvimento deste trabalho e pela disponibilização de resultados e informações fundamentais à sua realização.

## REFERÊNCIAS

[1] Rocke, D. M., Lorenzato, S. **A Two-component model for measurement error in analytical chemistry.** Technometrics, May 1995, vol37, n 0 2, pp. 176-184, 1995.

[2] International Organization for Standartization. **Guide to the expression of uncertainty in measurement**. ISO, Switzerland, 1995.

[3] Wilson, M.; et al. **Application to Environmental Monitoring of a Two-Component Model for Chemical Analytical Error**. Application paper: University of California, Davis, California, USA, 2000.

[4] Zorn, M.E., Gibbons, R. D.; Sonzogni, W.C. **Evaluation of approximate methods for calculating the limit of detection and limit of quantification**. Environmental Science and Technology, Vol. 33, 2229-2295, 1999.

[5] Peixoto, C. M; Pêgo, V. D. **Relatório de Avaliação dos resultados analíticos do Programa de Monitoração Ambiental do Centro de Desenvolvimento da Tecnologia Nuclear, 1991 e 1992**. Publicação CDTN-827/96, 1996.

[6] Triola, M. F. **Introdução à estatística**. Livros técnicos e científicos, Rio de Janeiro, 1999.

[7] STATISTICA 5.1. Statsoft Inc., OK, USA, 1998.

[8] BEVINGTON, P. R. **Data reduction and error analysis for the Physical Sciences**. Mcgraw-Hill, USA, 1999.

## ABSTRACT

Analysis and interpretation of results of an environmental monitoring program is often based on the evaluation of the mean value of a particular set of data, which is strongly affected by the analytical errors associated with each measurement. A model proposed by Rocke & Lorenzato assumes two error components, one additive and

one multiplicative, to deal with lower and higher concentration values in a single model.

In this communication, an application of this method for re-evaluation of the errors reported in a large set of results of total alpha measurements in a environmental sample is presented. The results show that the mean values calculated taking into account the new errors is higher than as obtained with the original errors, being an indicative that the analytical errors reported before were underestimated in the region of lower concentrations.