UTILIZAÇÃO DO ALGORITMO DE BOX E HILL PARA DISCRIMINAÇÃO ENTRE MODELOS COMPETITIVOS

Ralf GIELOW[1] , Roberto Francisco Marques MENDES[2] , Gervasio A. DEGRAZIA[3]

RESUMO

O algoritmo de Box e Hill, com base no conceito de entropia da informação e no teorema de Bayes, é um procedimento para a discriminação entre M modelos que competem para representar um determinado fenômeno ou sistema N-dimensional. Apresenta-se uma versão computacional amigável desse algoritmo, asssim como os resultados de sua aplicação na comparação do desempenho, frente a dados experimentais, de três modelos para representar a concentração ao nível da superfície de um gás ou vapor em dispersão na atmosfera.

INTRODUÇÃO

Não raro, fenômenos naturais ou processos artificiais como evapotranspiração, reações químicas, acessibilidade urbana ou rendimento agrícola, podem ser representados matematicamente por mais que um modelo, de acordo com o mecanismo de funcionamento do sistema ou ajuste estatístico-matemático que o observador ou experimentador considerar. Surge assim o problema de determinar qual o modelo que melhor representa o sistema em estudo.

O algoritmo de Box e Hill (1967) constitui um procedimento para a discriminação entre modelos que competem para representar um determinado fenômeno ou processo, em que uma variável dependente é função de várias variáveis independentes e parâmetros, e do qual se tem dados observados ou medidas experimentais sob diversas condições, cobrindo todo o domínio das variáveis, e se conhece o erro observacional ou experimental. Em adição, no caso de se poder realizar observações ou experimentos adicionais após a discriminação inicial entre os modelos, o algoritmo indica quais os valores das variáveis independentes a serem examinados subsequentemente para melhorar a discriminação.

O algoritmo é baseado no conceito de entropia da informação e no teorema de Bayes; ele efetua a discriminação entre M modelos que competem entre si para representar um sistema, conhecendo-se observações ou medidas experimentais, e respectiva variância, em todo o domínio das variáveis independentes. O número de variáveis independentes e parâmetros pode ser diferente em cada modelo. Inicialmente impõe-se uma probabilidade para cada modelo, não necessariamente 1/M, a qual é corrigida após consideração dos dados iniciais. Não havendo uma clara discriminação em favor de um dos modelos, e sendo possível observações ou medidas adicionais, o algoritmo, comparando os valores observados ou medidos inicialmente com os previstos pelos diversos modelos, determina sequencialmente sob que condições (valores das variáveis independentes) proceder para melhorar a discriminação, repetindo-se sucessivamente o processo até se ter uma clara discriminação.

Wadsworth (1990) cita tão somente o algoritmo de Box e Hill para este tipo de discriminação, dito bayesiano. Os métodos bayesianos (Feigelson e Babu, 1992) consideram a probabilidade como uma medida da plausibilidade de uma hipótese (modelo), em contraposição à visão frequencial, que identifica a probabilidade com a frequência relativa de ocorrência de um resultado de uma infinidade de repetições "idênticas" de um experimento ou observação. A inferência bayesiana enfoca hipóteses alternativas, enquanto a estatística frequencial enfoca conjuntos de dados. Para avaliar uma hipótese H, o enfoque bayesiano compara a probabilidade de H com as probabilidades de outras hipóteses; já os métodos frequenciais supõem H verdadeira e comparam a probabilidade dos dados observados ou medidos com as

[1] Dr., Pesquisador Titular, Divisão de Ciências Meteorológicas, C. P. 515, 12201-970 São José dos Campos, SP. E-mail: ralf@met.inpe.br

[2] Pós-graduando em Ciência da Computação, IME/USP. E-mail: mendes@met.inpe.br

[3] Dr., Professor, Departamento de Física, UFSM, 97119-900 Santa Maria, RS

probabilidades de outros conjuntos de dados preditos por H. Pragmaticamente, há fortes evidências da superioridade dos métodos bayesianos em aplicações reais. Não obstante, em qualquer caso, a maior dificuldade matemática usualmente está na estimativa dos valores dos parâmetros que conectam não-linearmente as variáveis independentes, levando a procedimentos iterativos, ditos regressões não-lineares, que necessitam de estimativas iniciais e podem levar a resultados que dependem fortemente destas. Há muitos métodos de regressão não-linear, como os baseados na eliminação de Gauss e suas variantes (Wadsworth, 1990), sendo cada um mais indicado para certos tipos de funções.

O algoritmo de Box e Hill foi aplicado por Adeodato de Souza (1970) em problema de cinética química (equilíbrio oxigênio-hemoglobina - 8 modelos competitivos) e por Silva Filho (1976) em problema urbano (função acessibilidade entre células urbanas - 4 modelos). Por sinal, a dissertação de Silva Filho (1976), q.v., detalha muito bem o algoritmo de Box e Hill e suas fundamentações.

O presente trabalho mostra uma implementação computacional amigável do algoritmo de Box e Hill, devendo o usuário apenas digitar as expressões para os modelos, ou os resultados da aplicação destes, mais alguns dados de entrada. Apresentam-se, também, os resultados de uma aplicação do algoritmo à dispersão de um gás ou vapor na camada superficial da atmosfera.

MATERIAL E MÉTODOS

Aplicou-se o algoritmo de Box e Hill, na forma acima mencionada, a dados experimentais e três modelos para representá-los, sendo dois analíticos e um numérico, conforme detalhado em Carvalho et al. (1996).

RESULTADOS E CONCLUSÃO

Efetivou-se a discriminação entre os três modelos citados utilizando os dados constantes do trabalho de Carvalho et al. (1996), em que a distância está expressa em metros e as concentrações observadas e as calculadas, em , conforme mostrado na tabela abaixo, à qual se seguem os resultados obtidos, também reproduzidos na Figura 1:

Distância	Observação	Modelo 1	Modelo 2	Modelo 3
0.20	0.90	2.54	3.71	0.66
0.30	4.90	4.49	5.70	4.02
0.38	5.59	5.29	6.23	5.30
0.40	5.59	5.40	6.27	5.37
0.50	4.9	5.61	6.12	5.00
0.60	4.00	5.45	5.70	4.23
0.70	3.20	5.13	5.22	3.44
0.80	2.20	4.77	4.75	2.84
1.00	1.60	4.07	3.96	1.95
1.50	0.90	2.81	2.72	1.18

Variância do erro experimental = 0,1

Probabilidade

Modelo 1	0.2425
Modelo 2	0.2326
Modelo 3	0.5249

Portanto, o Modelo 3 (numérico) é o mais provável, com probabilidade em torno de 0,5, frente aos modelos analíticos 1 e 2.

AGRADECIMENTO: Ao CNPq, através do INPE, pela concessão de Bolsa de Iniciação Científica PIBIC a um dos autores, R. F. M. Mendes, de 1995 a 1997.

BIBLIOGRAFIA

ADEODATO DE SOUZA NETO, J. Discrimination among mechanistic models for oxigen-hemoglobin equilibrium. Gainesville, FL, University Of Florida, 1970. Tese de Doutorado em Engenharia Química.

BOX, G.E.; HILL, W.J. Discrimination among mechanistic models. **Technometrics**, v. 9, n.1, p.57-71, 1967.

CARVALHO, J. C.; VELHO, H. F. de C.; DEGRAZIA, G. A. Um estudo numérico da dispersão de poluentes na camada limite convectiva. In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 9. Campos do Jordão, SP. **Anais.** Sociedade Brasileira de Meteorologia, p.4-9, 1996.

FEIGELSON, E.F.; BABU G. J. **Statistical challenges in modern astronomy**. New York, Springer, 1992.

SILVA FILHO, J. F. da. Discriminação entre modelos competitivos. São José dos Campos, SP: INPE, 1977. 109p. Dissertação de Mestrado em Análise de Sistemas. (INPE - 1027- TPT/051).

Wadsworth, H. M. J. **Handbook of statistical methods for engineers and scientists**. New York, McGraw-Hill, 1990.

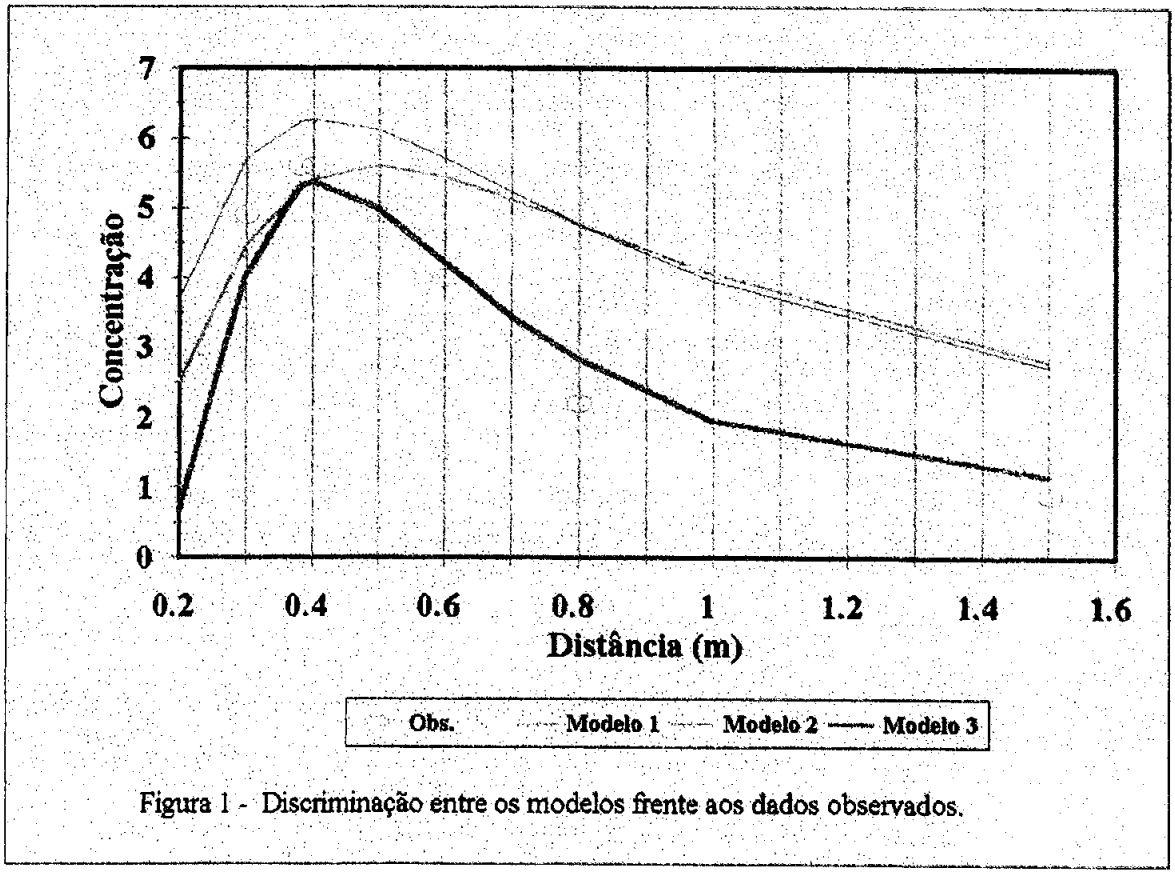

Figura 1 - Discriminação entre os modelos frente aos dados observados.