APLICAÇÃO DA TÉCNICA DE REAMOSTRAGEM BOOTSTRAP NA ESTIMAÇÃO DA PROBABILIDADE DOS ALUNOS SEREM USUÁRIOS DE TRANSPORTE PÚBLICO

Raquel Cymrot[1], Ana Lucia Tucci Rizzo [2]

Abstract *— The city of São Paulo suffers a lot with the toxic gas emission thrown daily into the atmosphere, produced by the great number of vehicles transiting in the streets. The purpose of this paper is to estimate the probability of the students using public transportation as a locomotion means to the university through the use of Bootstrap resample technique. This technique may be used even when the statistics probability distribution is unknown or when the estimators' calculation by using analytic methods is complex. A survey was conducted with the students of Escola de Engenharia of Universidade Presbiteriana Mackenzie. The students answered if they used, in most of the days, public transportation as a locomotion means to go to the university. The data analysis was made by using the Bootstrap technique.*

Index Terms — Public transportation, Bootstrap technique, Estimation

INTRODUÇÃO

Devido ao crescimento desordenado da população e a falta de recursos suficientes para suprir suas necessidades, os níveis poluentes lançados na atmosfera aumentaram de modo significante [5].

As conseqüências dessa agressão são irreversíveis e a cada dia que passa a população sente as conseqüências de viver em um ambiente inadequado. A saúde humana e animal, entretanto, é a maior vítima da poluição. A poluição pode ocasionar problemas como a irritação dos olhos, doenças crônicas do aparelho respiratório, insuficiência no transporte do oxigênio pela hemoglobina, perturbações nervosas [9].

Sabemos que entre os poluentes atmosféricos encontram-se o monóxido de carbono, o ozônio, o dióxido de carbono, o óxido de nitrogênio e os particulados [5]. Alguns destes surgem com a queima de combustíveis fósseis, particularmente das usinas elétricas a carvão e automóveis.

Além de prejudicar a qualidade do ar, a emissão desses gases pode provocar a chuva ácida [10].

As grandes cidades brasileiras apresentam graves problemas de transporte e qualidade de vida. As maiores cidades brasileiras foram adaptadas nas últimas décadas para o uso eficiente do automóvel. Por outro lado, embora tenha havido alguns investimentos importantes em sistemas de transporte público, estes foram insuficientes para atender a demanda crescente. O transporte público experimentou nos últimos anos um declínio na sua importância, na sua eficiência e na sua confiabilidade junto ao público. A falta de transporte público de qualidade, onde o sistema viário não é suficiente para garantir a circulação com eficiência, estimula a expansão do uso de transporte individual. Em setores da classe média, as grandes diferenças na qualidade do transporte resultam no uso do transporte individual, gerando um aumento da poluição atmosférica [1].

Algumas soluções podem ser estudadas para que se consiga reduzir os níveis de poluentes lançados na atmosfera.

Uma das alternativas para a redução dos índices de poluição atmosférica seria reduzir a emissão de poluentes por parte dos veículos automotores através da eletrificação de corredores de transporte público, pela aplicação de combustíveis menos poluentes e pela fiscalização dos níveis de emissão.

Outra possível alternativa a ser adotada é a melhor organização do uso das vias públicas, pelo aumento da oferta de transporte público de qualidade, pelo uso de técnicas adequadas de operação e otimização do trânsito ou pela imposição de restrições ao uso inadequado dos automóveis [6].

A conscientização da sociedade através de palestras educativas sobre as vantagens da utilização do transporte coletivo para a redução da poluição atmosférica seria um bom começo para que a população saiba que depende dela mesma melhorar a qualidade do ar.

Para que essas alternativas possam ser colocadas em prática, é preciso que alguns investimentos sejam realizados.

Este trabalho visa estimar a proporção de alunos da Escola de Engenharia da Universidade Presbiteriana Mackenzie que fazem uso do transporte público na sua locomoção para a universidade.

Em geral, quando se deseja fazer uma estimativa por intervalo de um parâmetro da população, são feitas suposições a respeito da distribuição de probabilidades deste parâmetro. Quando não é possível realizar tais suposições uma outra técnica denominada Bootstrap pode ser utilizada para esta finalidade.

[1] Raquel Cymrot, Universidade Presbiteriana Mackenzie, Rua da Consolação,930, prédio 6, 01302-907, São Paulo, SP, Brazil, raquelc@mackenzie.com.br
[2] Ana Lucia Tucci Rizzo, Bolsista PIBIC/CNPq, Universidade Presbiteriana Mackenzie, Rua da Consolação, 930, prédio 6, 01302-907, São Paulo, SP, Brazil, analurizzo@uol.com.br

A técnica de Bootstrap é uma técnica de reamostragem que consiste em sortear com reposição, dados pertencentes a uma amostra retirada anteriormente, denominada amostra mestre, de modo a formar uma nova amostra. Para realizar uma estimação através da utilização do Bootstrap é necessária a realização de um número muito grande de reamostragens e o cálculo das estatísticas de interesse para cada uma destas reamostragens, resultando no conhecimento da distribuição amostral do parâmetro a ser estimado.

Técnicas de reamostragem são úteis em especial quando o cálculo de estimadores por métodos analíticos for complicado, podendo ser aplicado em diversas situações em que se deseja estimar parâmetros na área ambiental. Muitas vezes a distribuição de probabilidade que estamos lidando é desconhecida. Nesse caso o Bootstrap é muito útil, pois é uma técnica que não exige diferentes fórmulas para cada problema e pode ser utilizada em casos gerais, não dependendo da distribuição original do parâmetro estudado.

Quando a distribuição do parâmetro a ser estimado é conhecida, a coincidência entre o intervalo paramétrico e o intervalo Bootstrap reforçam a hipótese de veracidade a respeito das suposições do modelo paramétrico.

A TÉCNICA DE BOOTSTRAP

Para realizar o teste utilizando a técnica Bootstrap é preciso colher uma amostra de tamanho n, que será denominada amostra mestre. Essa amostra deve ser coletada de maneira planejada, uma vez que se esta amostra for mal tirada e não representar bem a população, a técnica de Bootstrap não levará a resultados confiáveis.

Hesterberg et al. [2] afirmam que a amostra mestre representa a população da qual foi retirada. As reamostras desta amostra mestre representam o que se deve obter quando se retiram muitas amostras da população original. A distribuição Bootstrap da estatística, baseada em muitas reamostras, representa uma distribuição amostral desta estatística.

Para que a aplicação da técnica resulte em valores confiáveis devem ser feitas, a partir da amostra mestre, centenas ou até milhares de reamostras do mesmo tamanho n. É importante que a reamostragem seja realizada com reposição, sempre selecionando os valores de forma aleatória. Para a geração destas reamostras as técnicas computacionais são de grande utilidade. O programa Excel realiza estas reamostragens através da função de geração de números aleatórios a partir de uma distribuição discreta pré-estabelecida (distribuição da amostra mestre).

Uma vez geradas as reamostras, deve-se calcular para cada reamostra a estatística solicitada no problema. Essa técnica não altera nenhum valor da amostra mestre, ela apenas trabalha na análise da combinação dos valores iniciais com a finalidade de se obter as conclusões desejadas.

A variabilidade presente no Bootstrap é dada pela escolha da amostra mestre e pelas reamostras, sendo a variabilidade devido à escolha da amostra mestre, a mais significativa.

A distribuição Bootstrap usualmente tem aproximadamente a mesma forma e amplitude que a distribuição amostral, porém está centrada na estatística dos dados originais (amostra mestre), enquanto a distribuição amostral está centrada no parâmetro da população.

Segundo Manteiga et al. [3] uma das aplicações da metodologia Bootstrap é obter intervalos de confiança confiáveis.

Há diversas técnicas distintas para o cálculo de intervalos de confiança Bootstrap. A primeira delas é apresentada a seguir:

$$I.C._{bootstrap} = [\ \text{estatística} \pm t \times SE_{\ bootstrap}\ ] \qquad (1)$$

sendo n o tamanho da amostra mestre, t encontrado utilizando-se (n-1) graus de liberdade, N o número de reamostras realizadas e $SE_{bootstrap}$ igual ao desvio padrão das estatísticas nas N reamostras.

O intervalo de confiança Bootstrap t só funciona bem quando é conhecido que a distribuição Bootstrap é aproximadamente normal e tem pequeno vício.

Para verificar se o intervalo de confiança t calculado é confiável podemos comprá-lo com o intervalo de confiança percentil. Se o vício for pequeno e a distribuição bootstrap for aproximadamente normal, os dois intervalos irão apresentar valores muito próximos. O intervalo de confiança Bootstrap t acaba servindo mais como prova da suposição de normalidade da distribuição Bootstrap.

A segunda técnica de cálculo do intervalo de confiança Bootstrap é denominada intervalo de confiança percentil. Para uma confiança $(1 - \alpha)100\%$, encontra-se o percentil $(1 - \alpha/2)100$ % e o percentil $(\alpha/2)100\%$ da estatística nas reamostras [7].

A terceira técnica de cálculo do intervalo de confiança Bootstrap também é denominada intervalo de confiança percentil. e é calculado através dos percentis das diferenças dos valores das estatísticas das reamostras em relação ao valor médio desta mesma estatística nas reamostras [4].

Na maioria das publicações não técnicas em estatística, a forma de cálculo dos intervalos de confiança Bootstrap não costuma ser apresentada. Segundo enquête realizada por Hall [8], o método percentil é utilizado em mais da metade destas publicações.

O Bootstrap é muito genérico e devido a esta generalidade, há mais de um método Bootstrap como solução para um determinado problema [7].

Quando se deseja estimar um intervalo de confiança para a proporção, calcula-se o valor da proporção estimada para cada uma das reamostras Bootstrap $\hat{p}_i^*$ e a média dessas proporções $\overline{\hat{p}^*}$. Encontra-se então, para cada reamostra "i", a diferença entre esses valores, isto é:

$$\text{diferença} = \hat{p}_i^* - \overline{\hat{p}^*} \qquad (2)$$

Para uma confiança de 95%, encontra-se os percentis 97,5% e 2,5% destas diferenças e calcula-se o intervalo de confiança Bootstrap percentil da seguinte forma:

$$IC_{bootstrap\ percentil} = \left[\hat{p} - P_{97,5\%}diferenças;\ \hat{p} - P_{2,75\%}diferenças\right] \quad (3)$$

Uma estatística utilizada para estimar um parâmetro é viciada quando a distribuição amostral não estiver centrada no verdadeiro valor do parâmetro. A técnica Boostrap nos permite verificar o vício olhando se a distribuição Bootstrap da estatística está centrada na estatística da amostra mestre [2].

O estimador do vício da distribuição Bootstrap é: vício bootstrap = (média da estatística da distribuição boostrap – estatística dos dados originais) (4)

No caso desta estatística ser a proporção, o vício pode ser representado da seguinte forma:

$$\text{vício}_{\text{bootstrap}} = \overline{\hat{p}^*} - \hat{p} \quad (5)$$

METODOLOGIA

Foi realizada, no segundo semestre de 2005, uma pesquisa para se estimar a probabilidade de um aluno do curso Engenharia de Produção da Escola de Engenharia da Universidade Presbiteriana Mackenzie utilizar transporte público no seu deslocamento de ida e/ou volta para a universidade. Foi realizada uma amostragem por conglomerado no qual foi sorteado o sexto semestre. Foi perguntado para seus 33 alunos matriculados se eles utilizavam, na maior parte dos dias, transporte público como meio de locomoção para a universidade. Quando a resposta foi positiva a variável foi codificada como 1 e quando negativa a variável foi codificada como 0. Foi então calculada a probabilidade p de o aluno utilizar transporte público.

Os 33 dados coletados formaram a amostra mestre. Com base nesta amostra, foram realizadas 1000 reamostras de mesmo tamanho e aplicada à técnica Bootstrap a fim de calcular os intervalos de confiança Bootstrap para a proporção de respostas afirmativas. Estes resultados foram comparados com o intervalo de confiança tradicional paramétrico. Foi também calculado o intervalo de confiança Bootstrap para a variância desta proporção.

Os dados foram analisados utilizando o software MINITAB.

RESULTADOS

A tabela 1 apresenta a amostra mestre, cinco das 1000 reamostras, a proporção e a variância da proporção para a amostra mestre e reamostras.

Cada reamostra foi gerada atribuindo probabilidade igual a 1/33 para cada observação da amostra mestre.

Na figura 1 é apresentado o histograma das proporções obtidas nas 1000 reamostras no qual foi verificado que a forma da distribuição é próxima da Normal. Foram calculados os valores dos quartis Q_1 = 0,4848, Q_2 = 0,5455 e Q_3 = 0,6061 e encontrados quatro possíveis "outliers" , a saber: 0,8182; 0,7879; 0,3030; 0,2727. Estes valores foram mantidos na amostra.

TABELA 1

Amostra mestre, reamostras, proporção e variância da proporção para a amostra mestre e reamostras.

observação	amostra mestre	reamostra 1	reamostra 2	reamostra 3	reamostra 4	reamostra 5
1	1	0	0	1	1	1
2	0	0	1	0	0	1
3	0	0	1	0	0	1
4	0	1	1	1	0	0
5	1	1	1	1	0	1
6	0	0	1	1	0	0
7	1	1	1	1	0	1
8	1	1	1	1	1	1
9	1	0	1	1	0	1
10	1	0	0	1	1	1
11	1	0	0	0	0	0
12	1	0	1	0	1	1
13	0	1	1	1	1	1
14	1	0	0	1	0	0
15	0	0	1	1	1	0
16	0	1	0	0	1	1
17	0	0	0	1	1	1
18	0	1	0	0	0	1
19	1	0	0	0	0	1
20	1	1	0	0	1	0
21	0	0	0	1	1	1
22	0	0	0	1	1	1
23	0	0	1	1	1	0
24	1	1	0	0	1	1
25	1	1	0	0	0	1
26	0	0	1	1	0	1
27	0	0	0	0	0	1
28	1	0	1	1	0	1
29	1	0	1	1	1	0
30	1	0	1	1	0	1
31	0	0	0	0	1	0
32	1	1	1	0	1	0
33	1	1	1	0	0	0
proporção	0,5455	0,3636	0,5455	0,5758	0,4848	0,6667
variância da proporção	0,0075	0,0070	0,0075	0,0074	0,0076	0,0067

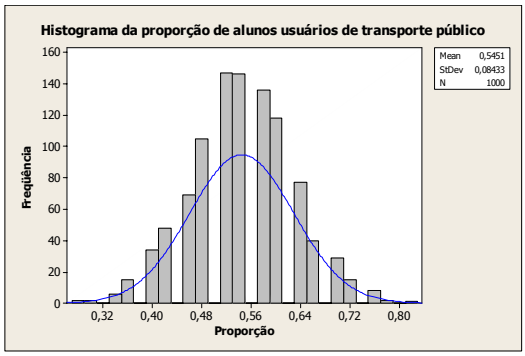

FIGURA 1

Histograma das proporções nas 1000 reamostras

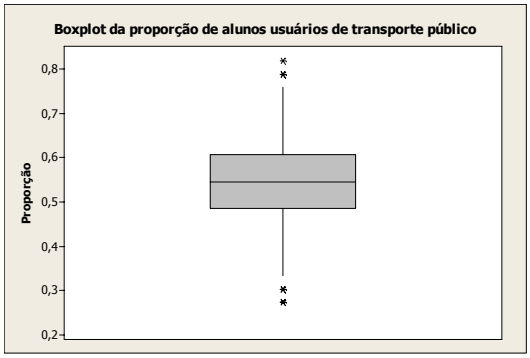

FIGURA 2

Boxplot da proporção de alunos usuários de transporte público nas 1000 reamostras.

A figura 2 apresenta o Boxplot para as proporções nas 1000 reamostras, onde é notada a simetria da distribuição.

A figura 3 apresenta o teste de aderência de Kolmogorov-Smirnov no qual foi confirmada a normalidade da distribuição das proporções estimadas nas 1000 reamostras ($p > 0,150$).

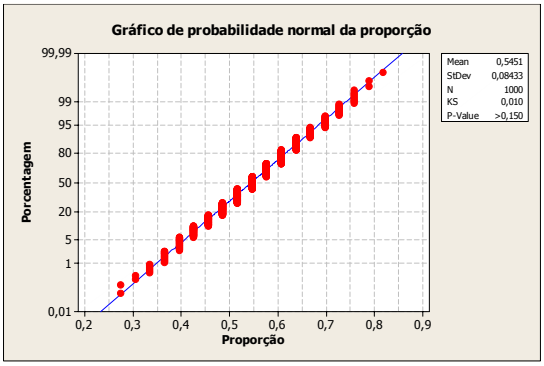

FIGURA 3

Gráfico de probabilidade normal para as proporções de alunos usuários de transporte público, nas 1000 reamostras.

Os gráficos apresentados confirmam a normalidade da distribuição das estimativas da proporção nas reamostras. Isto pode ser visto através da forma muito próxima de uma distribuição normal no histograma do gráfico 1 e do teste de aderência realizado no gráfico 3. Neste caso o intervalo de Confiança Bootstrap t pode ser utilizado e deve coincidir com os intervalos de Confiança Bootstrap Percentil.

A amostra mestre apresentou uma proporção estimada de alunos que utilizavam, na maior parte dos dias, transporte público como meio de locomoção para a universidade igual a 0,5455, com variância da proporção igual a 0,0075. As reamostras apresentaram uma média das proporções igual a 0,5451 com variância da proporção igual a 0,0071.

Os intervalos de confiança para a proporção dos alunos que utilizavam, na maior parte dos dias, transporte público como meio de locomoção para a universidade foram calculados através das três formas distintas do método Bootstrap descritas anteriormente.

Para calcular o intervalo de confiança pelo método Percentil das diferenças, foram encontrados os percentis 2,5% e 97,5% das diferenças das médias das proporções das 1000 reamostras, respectivamente iguais a –0,1519 e 0,1822.

Os intervalos de confiança para a proporção, calculados através dos três métodos revelaram-se muito próximos, a saber: Intervalo de Confiança Bootstrap Percentil = [0,3932 ; 0,7273], Intervalo de Confiança Bootstrap Percentil das Diferenças = [0,3632 ; 0,6973] e Intervalo de Confiança Bootstrap t de Student = [0,3737 ; 0,7172].

Foi também calculado o intervalo de confiança para a proporção, supondo sua distribuição conhecida e aproximadamente normal. Para este cálculo foram utilizados os dados da amostra mestre, tendo sido obtido o intervalo [0,3756 ; 0,7153], também bem próximo aos demais intervalos de confiança calculados.

O vício bootstrap foi igual a 0,5451 – 0,5455 = - 0,0004.

De modo análogo foram obtidos os intervalos de confiança para a variância da proporção dos alunos que utilizavam, na maior parte dos dias, transporte público como meio de locomoção para a universidade.

Na figura 4 é apresentado o histograma das variâncias das proporções obtidas nas 1000 reamostras no qual não foi verificado a forma normal da distribuição .

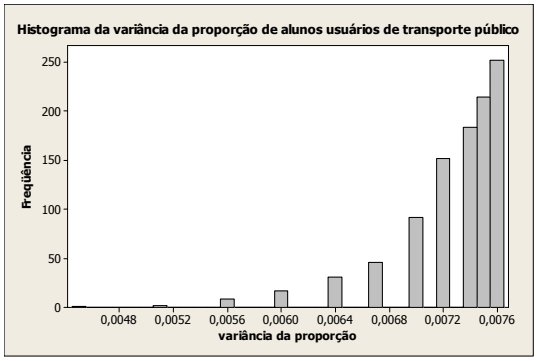

FIGURA 4

Histograma variâncias das proporções nas 1000 reamostras.

A figura 5 apresenta o teste de aderência de Kolmogorov-Smirnov no qual não foi confirmada a normalidade da distribuição das variâncias das proporções estimadas nas 1000 reamostras ($p < 0,010$).

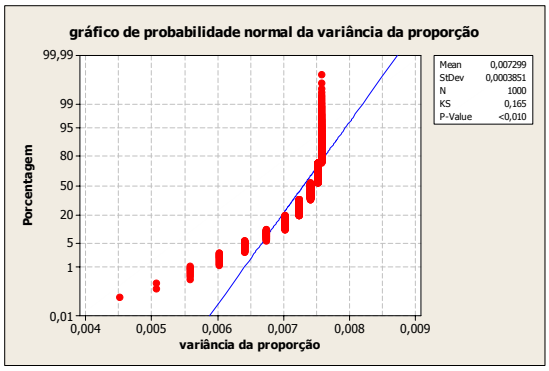

FIGURA 5

Gráfico de probabilidade normal para as variâncias das proporções de alunos usuários de transporte público, nas 1000 reamostras.

Os valores dos percentis 2,5% e 97,5% das diferenças das variâncias das proporções de cada reamostra em relação à média das variâncias das proporções obtidas pelas 1000 reamostras, foram respectivamente iguais a – 0,0013 e 0,0003.

Devido à falta de normalidade, neste caso não é indicado o uso do intervalo de confiança Bootstrap t de student.

Os intervalos de confiança para a variância da proporção, calculados através dos outros dois métodos

revelaram-se muito próximos, a saber: Intervalo de Confiança Bootstrap Percentil = [0,0060 ; 0,0076] e Intervalo de Confiança Bootstrap Percentil das Diferenças = [0,0072 ; 0,0088].

Como neste caso o estimador é tendencioso, uma melhoria conhecida como método corrigido e acelerado pela tendência poderia ser utilizada, mas a custo de complexidade adicional.

O vício bootstrap foi igual a 0,0073 – 0,0075 = - 0,0002.

CONCLUSÕES

Analisando o estudo de caso da proporção de alunos matriculados no curso de engenharia da Universidade Presbiteriana Mackenzie que utilizam o transporte público como meio de transporte, nota-se que um pouco mais da metade adota o transporte coletivo como meio de locomoção, sendo que as estimativas dos intervalos de confiança bootstrap estão compreendidas entre 0,35 e 0,70, aproximadamente.

Há, portanto, espaço para que medidas sejam tomadas de modo a aumentar o uso do transporte público, uma vez que este causa menos efeitos adversos ao ambiente quando comparados com o uso do transporte individual.

Este trabalho apresentou uma aplicação da técnica de estimação Bootstrap. Era desejado calcular um intervalo de confiança para a proporção de alunos que utilizam transporte público. Neste caso há uma fórmula fácil e conhecida para o cálculo deste intervalo de confiança. A coincidência do intervalo de confiança calculado pelo método paramétrico conhecido com os intervalos calculados pelos métodos Bootstrap atesta a eficiência destes últimos. Foram calculados também intervalos de confiança para a variância da proporção de alunos que utilizam transporte público. Neste caso em que não há fórmulas fáceis para o cálculo do intervalo de confiança, o método Bootstrap é de grande valia.

Verificou-se, portanto que o método Bootstrap de estimação permite que o cálculo do intervalo de confiança seja realizado de modo mais simples e abrangente para diversas estatísticas, mesmo quando as distribuições de probabilidades das mesmas são desconhecidas. Foi possível observar a generalidade de aplicação desta técnica de estimação através da reamostragem visto que a mesma se adequa a qualquer situação, sendo seus cálculos rápidos e seus resultados muito eficientes.

Os valores calculados para o viés nas estimações da proporção e variância da proporção foram pequenos, dando mais uma certificação de que os valores estimados devem estar próximos dos verdadeiros valores.

REFERÊNCIAS

[1] ASSOCIAÇÃO NACIONAL DE TRANSPORTES PÚBLICOS ANTP, *Transporte humano: cidades com qualidade de vida*. 2. ed. São Paulo: ANTP, 1997.

[2] HESTERBERG, T.; MOORE, D. S.; MONAGHAN, S.; CLIPSON, A.; EPSTEIN, R. Bootstrap methods and permutation tests, In: *The practice of business statistics*. New York: W. H. Freeman, 2003.

[3] MANTEIGA, W.G.; SÁNCHEZ, J.S.P.; ROMO, J. The Bootstrap – a review. *Computational Statistics.*, v.9, n. 1, p. 165-205, 1994.

[4] MONTGOMERY, D. C.; RUNGER G. C.; *Estatística Aplicada e Probabilidade para Engenheiros.* Rio de Janeiro: LTC, 1999.

[5] MATTOS, N. S. *A poluição atmosférica*. São Paulo: FTD, 1992.

[6] VASCONCELLOS, E. A. *Transporte urbano nos países em desenvolvimento:* reflexões e propostas. 3. ed. São Paulo: Annablume, 2000.

[7] EFRON, B.; TIBSHIRANI, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science* , v. 1, n. 1, p. 55–77, 1986.

[8] HALL, P. Theorical comparison of bootstrap confidence intervals, *The Annals of Statistics,* v. 16, n. 3, p. 987–953, 1988.

[9] CENTRO PEDAGÓGICO UFMG, CLUBE DE CIÊNCIAS: *Problemas causados pela poluição.* Disponível em: <http://www.cp.ufmg.br/Clube_de_Ciencias/>. Acesso em: 10 abr. 2006.

[10] BÖHM, M. G. SAÚDE TOTAL: *Doenças causadas pela poluição atmosférica*. São Paulo: Saúde total, Poluição Atmosférica. Disponível em: <http://www.saudetotal.com/artigos/meioambiente/poluicao/spdoencas.asp> . Acesso em: 17 abr. 2006.