VI CONGRESSO NACIONAL DE ENGENHARIA MECÂNICA
VI NATIONAL CONGRESS OF MECHANICAL ENGINEERING
18 a 21 de agosto de 2010 – Campina Grande – Paraíba - Brasil
August 18 – 21, 2010 – Campina Grande – Paraíba – Brazil

MODELO DE REGRESSÃO LINEAR MÚLTIPLA APLICADA À DETERMINAÇÃO DA CONCENTRAÇÃO DE OZÔNIO NO AR

Dutra, Elisete Gomides, elisete.gomides@meioambiente.mg.gov.br[(1)]
Fioravante, Edwan Fernandes, edwan.fioravante@meioambiente.mg.gov.br[(1)]
Requeijo, José Gomes, jfgr@fct.unl.pt[(2)]

[1]Fundação Estadual do Meio Ambiente; Rodovia Prefeito Américo Gianetti, snº, Prédio Minas , 1º andar, Bairro Serra Verde, Belo Horizonte, Minas Gerais, Brasil, 31630-900
[2] Faculdade de Ciência e Tecnologia da Universidade Nova de Lisboa, 2829-516, Caparica, Portugal

Resumo: *A poluição do ar por ozônio constitui problema de saúde pública em centros urbanos. O ozônio é formado na atmosfera mediante reação entre compostos orgânicos voláteis e óxidos de nitrogénio em presença de radiação solar. Os veículos automotores são grandes responsáveis pela emissão desses poluentes e, conseqüentemente, pela formação do ozônio. A média horária da concentração de poluentes, da temperatura e da radiação na Região Metropolitana de Belo Horizonte (RMBH) são obtidos em estações automáticas de monitoramento da qualidade do ar gerenciadas pela Fundação Estadual do Meio Ambiente. Apresenta-se os resultados do estudo da relação entre o valor máximo diário de ozônio e os valores de temperatura e radiação solar registrados no município de Vespasiano, localizado na RMBH, de novembro de 2008 a outubro de 2009. Foi utilizado o método dos mínimos quadrados com duas variáveis explicativas, utilizando-se o software SPSS para determinação do modelo de regressão múltipla de melhor ajuste dos dados. O modelo proposto possibilita prever a concentração de ozônio para um determinado cenário de temperatura, o que possibilita a definição de políticas públicas adequadas para evitar danos à saúde das pessoas na região em estudo. No entanto o modelo deve ser usado com cautela, pois o coeficiente de determinação encontrado pode ser considerado demasiado baixo. Esta situação indica que o modelo deveria contemplar mais variáveis independentes ou que deveria ter-se optado por uma regressão múltipla não linear. Outra abordagem possível seria utilizar concentrações médias horárias de ozônio, ao invés das concentrações máximas diárias, pois permitiria acompanhar a evolução dessas concentrações ao longo de cada dia juntamente com a evolução da temperatura média horária.*

Palavras-chaves: *Ozônio; Poluição Veicular; Radiação solar; Temperatura; Regressão linear múltipla*

1. INTRODUÇÃO

O ozônio troposférico é um poluente secundário, ou seja, não é emitido diretamente para a atmosfera. É produzido mediante reações químicas entre óxidos de nitrogênio (NOx) e compostos orgânicos voláteis (COV) em dias ensolarados (alta radiação solar), principalmente em áreas urbanas e industriais e em regiões propensas à estagnação de massas de ar. Ozônio (O_3) é um gás tóxico e em altas concentrações no ar que respiramos é um problema de saúde pública. Os sintomas, em geral, incluem tosse, dor de cabeça, náuseas, dores peitorais e falta de ar. Para concentrações superiores a 360 $\mu g/m^3$, durante uma hora, podem verificar-se danos na função pulmonar (Pereira, 1999).

Os veículos automotores são os grandes responsáveis pela emissão de compostos orgânicos voláteis (COV), óxidos de nitrogênio e, conseqüentemente, pela formação do ozônio em áreas urbanas. Acrescente-se a estes as perdas de combustíveis durante reabastecimento dos veículos e dos reservatórios de combustíveis dos postos de abastecimento, bem como as emissões específicas de alguns empreendimentos como refinaria de petróleo, envasadoras de gás e outras.

A Fundação Estadual do Meio Ambiente (FEAM) gerencia o monitoramento da qualidade do ar na Região Metropolitana de Belo Horizonte (RMBH) desde 1995, sendo que a quantificação da concentração de ozônio (O_3) iniciou-se em 1999. Essa rede é constituída por nove estações automáticas, que medem os poluentes atmosféricos regulamentados pela Resolução do Conselho Nacional de Meio Ambiente (CONAMA) Nº 3 de 1990 que são Material Particulado com granulometria inferior a 10 micrometros (MP10), Dióxido de Enxofre (SO_2), Monóxido de Carbono (CO), Ozônio (O_3) e Óxidos de Nitrogênio (NOx), durante 24 horas por dia, 365 dias por ano.

As estações são constituídas por cabines climatizadas onde estão instalados analisadores e sensores que realizam a amostragem do ar atmosférico e determinam a concentração de poluentes e dados meteorológicos de forma contínua. Os resultados são transmitidos em tempo real por modem, via linha telefônica, ao Centro Supervisório de Qualidade do Ar da FEAM.

A condição da qualidade do ar nos municípios de Belo Horizonte, Contagem e Betim, até meados de 2008, era definida principalmente em função do Material Particulado que era o poluente presente no ar em maior concentração. Nos últimos anos, esse cenário vem se modificando em função do aumento da concentração de Ozônio, tornando-se mais visível a partir de agosto de 2008.

A aplicação prática desenvolvida constou do estudo da relação entre o valor máximo diário de ozônio (μg/m^3) e os correspondentes valores máximos de temperatura (ºC) e radiação solar (W/m^2) registrados pela estação automática de monitoramento da qualidade do ar do município de Vespasiano, Região Metropolitana de Belo Horizonte, no período de novembro de 2008 a outubro de 2009.

2. ANÁLISE DE REGRESSÃO

A análise de regressão é uma técnica estatística usada para investigar e modelar a relação entre uma variável explicada (ou dependente) e uma ou várias variáveis explicativas (ou independentes). Um dos objetivos da análise de regressão consiste em estimar os parâmetros do modelo, usando um método adequado. Existem várias técnicas para a estimação dos parâmetros do modelo, como sejam o método dos mínimos quadrados (Montgomery *et al*, 2001), o método bayseano (Box & Tiao, 1973) e o método *bootstrap* (Efron & Tibshirani, 1993).

2.1. Regressão Linear Múltiplo

Quando se pretende avaliar a relação de uma variável dependente *Y* e *k* variáveis independentes X_j, $j = 1,2,\ldots,k$, o modelo toma a designação de "modelo de regressão linear múltipla", pois envolve mais do que um coeficiente de regressão. Este modelo é caracterizado por uma equação dada para o período *i* por

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \quad , \quad i = 1,2,\ldots,n \tag{1}$$

em que x_{ij} e y_i são os valores observados no período *i*, respectivamente das variáveis X_j e *Y*, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)'$ é o vetor de coeficientes de regressão (parâmetros) e ε_i é o erro aleatório no período *i*. Assume-se que os erros são independentes e identicamente distribuídos segundo uma distribuição Normal com média zero e variância σ^2 ($\varepsilon \sim NID(0;\sigma^2)$).

Usando notação matricial, a equação 1 é redefinida por

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

em que $\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)'$ e $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$.

A matriz **X** é definida por

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \ldots x_{1k} \\ 1 & x_{21} & x_{22} \ldots x_{2k} \\ 1 & \vdots & \vdots \quad \ldots \vdots \\ 1 & x_{n1} & x_{n2} \ldots x_{nk} \end{bmatrix} \tag{3}$$

A estimação dos parâmetros da regressão, usando o método dos mínimos quadrados, é dada por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{C}\mathbf{X}'\mathbf{Y} \tag{4}$$

Tanto a regressão como as estimativas dos seus parâmetros têm de ser validadas, de forma a se poder considerar adequado a aplicação do modelo.

A significância da regressão pode ser realizada através de uma Análise de Variância (ANOVA). A hipótese nula e hipótese alternativa do teste para verificar a significância da regressão são definidas por

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \quad \text{(pelo menos para um } j\text{)} \tag{5}$$

A estatística de teste F_0 é dada por

$$F_0 = \frac{MS_{Reg}}{MS_{Erro}} = \frac{\frac{SS_{Reg}}{k}}{\frac{SS_{Erro}}{n-k-1}} \tag{6}$$

As variações (*SS*) da ANOVA são determinadas a partir de

$$SS_T = SS_{Reg} + SS_{Erro} = \mathbf{Y'Y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{7}$$

$$SS_{Reg} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{8}$$

$$SS_{Erro} = \sum_{i=1}^{n} e_i^2 = \mathbf{Y'Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} \tag{9}$$

Para testar a significância da regressão compara-se a estatística de teste F_0 com o valor de $F_{crítico} = F_{\alpha;k,n-k-1}$, em que *MS* é o desvio quadrático médio ou variação média. Verifica-se a significância da regressão, para um determinado nível de significância α, se $F_0 > F_{crítico}$. A determinação da estatística de teste (F_0), da soma dos desvios quadráticos (*SS*) e dos quadrados médios (*MS*) é realizada através da aplicação da Análise de Variância. A tabela ANOVA é apresentada na Tab.(1).

Tabela 1 – Tabela ANOVA da Regressão Linear Múltipla

Fonte de Variação	*SS*	*Graus de liberdade*	*MS*	F_0
Regressão	SS_{Reg}	k	SS_{Reg}/k	MS_{Reg}/MS_{Erro}
Erro	SS_{Erro}	$n-k-1$	$SS_{Erro}/(n-k-1)$	
Total	SS_T	$n-1$		

Além de analisar a significância da regressão é necessário verificar se os coeficientes de regressão β_j são significativamente diferentes de zero. Para tal, aplica-se o teste de hipóteses da média. A hipótese nula e hipótese alternativa são definidas por

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0 \tag{10}$$

A estatística de teste é definida por

$$t_0 = \frac{\hat{\beta}_j - E\left(\hat{\beta}_j\right)}{\sqrt{Var\left(\hat{\beta}_j\right)}} \tag{11}$$

em que

$$E\left(\hat{\beta}_j\right) = \beta_j$$
$$Var\left(\hat{\beta}_1\right) = \sigma^2 C_{jj} \tag{12}$$

C_{jj} é o elemento da diagonal da matriz $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ referente à linha *j* e coluna *j*. A variância σ^2 é estimada a partir de

$$\hat{\sigma}^2 = MS_{Erro} = \frac{SS_{Erro}}{n-k-1} \tag{13}$$

Rejeita-se H_0, *i.e.*, o coeficiente de regressão β_j é significativamente diferente de zero, para um nível de significância de α, se $|t_0| > t_{crítico}$, em que $t_{crítico} = t_{\alpha/2;(n-k-1)}$.

2.2. Coeficiente de Determinação

Em geral nos modelos de regressão linear, usa-se o coeficiente de determinação R^2 para quantificar a capacidade explicativa do modelo. O valor de R^2 avalia a fracção da variação total que é explicada pelo modelo de regressão proposto, *i.e.*, a proporção da variação explicada face à variação total da variável dependente. Os melhores modelos são aqueles que apresentam maiores valores de R^2 ($0 \le R^2 \le 1$). Esta estatística é determinada a partir de

$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_{Erro}}{SS_T} \tag{14}$$

Outro indicador para quantificar a capacidade explicativa do modelo é o "coeficiente de determinação ajustado $R^2_{ajustado}$. Este coeficiente é determinado tendo por base R^2, o numero de variáveis independentes *k* e o número de dados recolhidos *n*, sendo dado por

$$R^2_{ajustado} = R^2 - \frac{k\left(1-R^2\right)}{n-k-1} \tag{15}$$

No presente artigo utilizou-se o método dos mínimos quadrados, com duas variáveis explicativas, ou seja, independentes, utilizando-se o software SPSS para determinação do modelo de regressão múltipla de melhor ajuste dos dados. Foram verificados os pressupostos do modelo: 1) resíduo com distribuição Normal, com média igual a zero e variância constante; 2) covariância nula entre as variáveis independentes e os resíduos; 3) covariância nula entre as variáveis independentes, ou seja, não há multicolinearidade entre as variáveis independentes.

3. RESULTADOS

A equação abaixo apresenta os coeficientes estimados pelo método de mínimos quadrados, cujo modelo apresentou coeficiente de determinação ajustado de 0,34, ou seja, apenas 34% da variabilidade das concentrações máximas de ozônio pode ser explicada pela combinação linear das variáveis temperatura e radiação solar:

$$Ozônio = -2{,}380 + 2{,}377\,Temperatura - 0{,}009\,Radiação \tag{16}$$

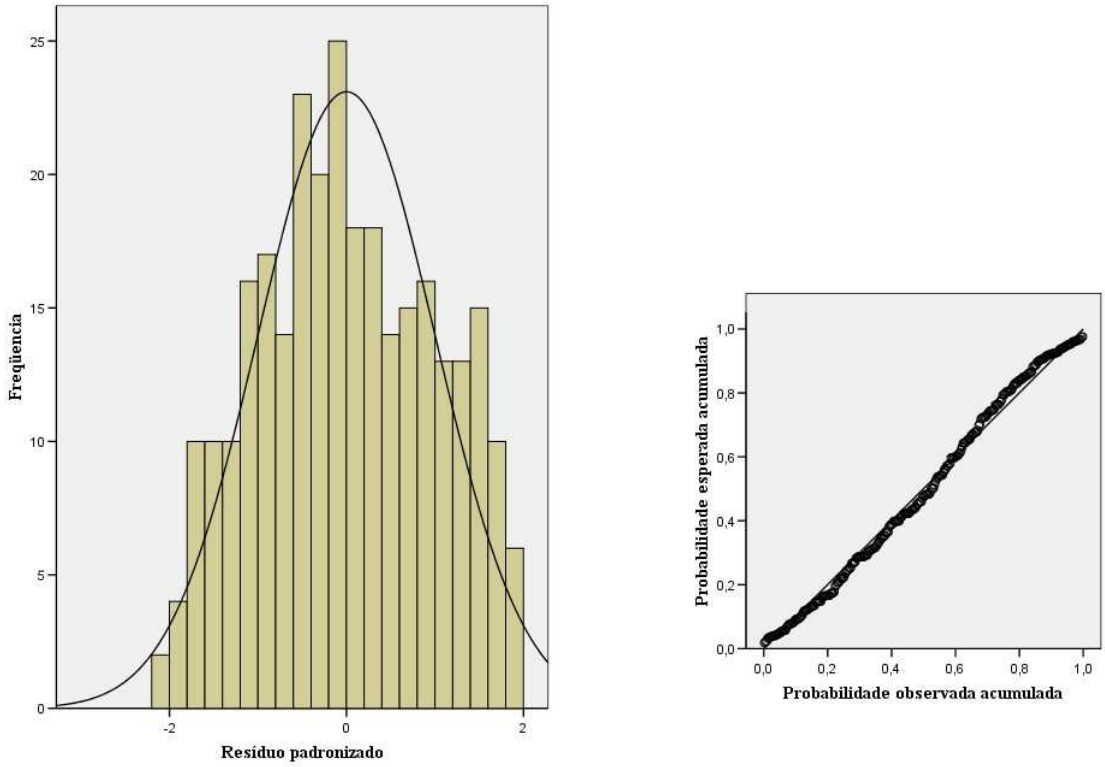
Para verificar o pressuposto de multicolinearidade, calculou-se, primeiramente, o coeficiente de correlação de Pearson da variável temperatura com a variável radiação solar, que correspondeu a 0,44. Embora esse coeficiente não seja alto, ele também não pode ser considerado baixo. O segundo passo foi calcular o valor médio do fator de inflação de variância, que é utilizado para detectar a presença de multicolinearidade entre as variáveis independentes. Segundo Neter, Kutner & Wasserman (1996, p. 387), valor médio consideravelmente maior que 1 é indicativo de sérios problemas de multicolinearidade. Para o modelo de regressão proposto, com as variáveis independentes temperatura e radiação solar, o valor médio do fator de variância corresponde a 1,24, valor próximo de 1, não indicando, assim, um sério problema de multicolinearidade

Entretanto, é provável que haja influência da correlação entre as variáveis consideradas independentes, pois o coeficiente obtido para a variável radiação solar é negativo (-0,009). Nesse modelo, as estimativas das concentrações máximas de ozônio diminuem à medida que a radiação solar aumenta; o que é contraditório, segundo a literatura, e não está de acordo com o coeficiente de correlação de Pearson obtido entre as concentrações máximas de ozônio e radiação solar (+0,153). Portanto, o modelo final terá apenas como variável independente a temperatura máxima observada:

$$Ozônio = 0{,}327 + 2{,}161\,Temperatura \tag{17}$$

O modelo acima apresenta coeficiente de determinação ajustado igual a 0,32, que é um valor bem próximo do obtido pelo modelo que incorpora, além da temperatura, a radiação solar. Conclui-se, assim, que não houve grande

redução do coeficiente de determinação após a retirada da variável radiação solar. A FIG.1 apresenta o histograma dos resíduos obtidos através da EQ. 17, com a linha que representa a distribuição normal.



(a) (b)

Figura 1: Gráficos de freqüência e de normalidade dos resíduos padronizados para o modelo de máximos de ozônio em função da temperatura

Percebe-se na FIG.1 (a) que os resíduos estão praticamente centrados em torno do valor zero, atendendo assim o pressuposto de que os mesmos apresentam valor esperado igual a zero. Para verificar o pressuposto de que eles apresentam distribuição normal, foi elaborado o gráfico de probabilidade normal, apresentado na FIG. 1(b). Pode-se supor que a distribuição dos resíduos é aproximadamente normal, havendo apenas um pequeno distanciamento da distribuição normal na cauda superior. A FIG. 2 apresenta o gráfico de dispersão dos resíduos pelos valores preditos, ambos padronizados.

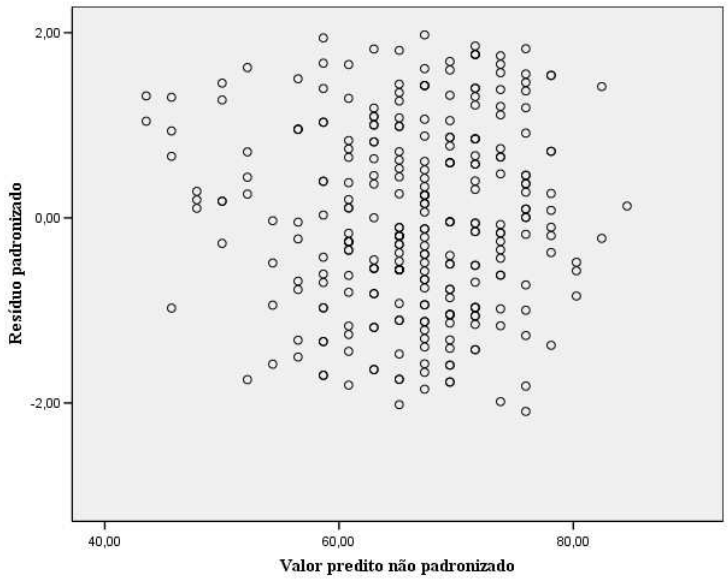


Figura 2: Gráfico dos resíduos padronizados versus valores preditos não padronizados para o modelo de máximos de ozônio em função da temperatura

Verifica-se na FIG. 3 que o pressuposto de variância constante pode ser considerado válido, uma vez que a dispersão desses resíduos assemelha-se a uma nuvem, ou seja, não havendo maior ou menor dispersão dos resíduos em função da concentração predita de ozônio.

Para verificar se alguns resíduos poderiam estar influenciando a estimativa dos coeficientes do modelo, foi utilizada a estatística DFBETA. Para grande conjuntos de dados, a regra determina que resíduos que excedam o valor de $2/\sqrt{n}$ (Neter et al., 1996), onde n correspondente ao número de observações; para n = 289, esse valor corresponde a 0,013841. A retirada das observações cujos valores de DFBETA é maior, em módulo, do que 0,013841, não influenciou grandemente o coeficiente obtido para temperatura, portanto, manteve-se o modelo estimado anteriormente.

Assim, é possível prever a concentração de ozônio para um determinado cenário de temperatura, o que permitirá, em casos gravosos, a definição de políticas públicas adequadas no sentido de evitar danos à saúde das pessoas na região em estudo. É necessário ter algum cuidado nas previsões que poderão ser realizar com base neste modelo, já que os valores dos coeficientes de determinação não são completamente satisfatórios.

4. CONCLUSÕES

No plano ambiental é de extrema importância a qualidade do ar em todas as suas vertentes, cabendo destacar uma característica especialmente nociva, o nível elevado de ozônio. A metodologia desenvolvida neste artigo mostrou-se válida para modelar as concentrações máximas diárias de ozônio em função da temperatura, tendo por base os dados obtidos nas estações de monitoramento da qualidade do ar. Pode-se enumerar as principais vantagens com a realização do presente trabalho:

- Definição de uma metodologia, com base na regressão linear múltipla, para estudar a relação entre o nível de ozônio (variável dependente) e a temperatura.
- Realização de previsões, com base no modelo obtido.
- Possibilidade de definição de metodologias semelhantes envolvendo novas variáveis independentes além das duas consideradas no presente artigo, considerando a especificidade do meio envolvente da área onde se realize o estudo.
- Com base nos resultados obtidos com a implementação proposta, que revelam um coeficiente de determinação pouco satisfatório, estudar a possibilidade de definição de uma metodologia com coeficiente de determinação mais elevado, recorrendo eventualmente a regressão múltipla não linear.

A principal desvantagem que se detecta no presente trabalho tem a ver com a validação da metodologia proposta, dado que o valor obtido do coeficiente de determinação (0,32) pode ser considerado como demasiado baixo. Esta situação pode revelar que este modelo deveria contemplar mais variáveis independentes do que as duas consideradas ou que se deveria ter optado por uma regressão múltipla não linear. Esta constatação será o ponto de partida para trabalhos futuros no sentido de obter um modelo mais robusto, ajustado aos dados em estudo.

Outra abordagem possível é utilizar as concentrações médias horárias de ozônio, ao invés das concentrações máximas diárias de ozônio, como variável dependente do modelo de regressão linear múltiplo, pois permitirá acompanhar a evolução dessas concentrações ao longo de cada dia juntamente com a evolução da temperatura média horária. Além da variável temperatura, pretende-se incluir também as medições horárias de dióxido de nitrogênio e hidrocarbonetos totais que são poluentes precursores para a formação do poluente ozônio.

REFERÊNCIAS

BOX, G. E. P.; TIAO, G. C. **Bayesian Inference in Statistical Analysis**. Addison-Wesley Publishing Co., Mass.-London-Don Mils, Ont. Addison-Wesley Series in Behavioral Science: Quantitative Methods, 1973.

EFRON, B.; TIBSHIRANI, R. J. An Introduction to the Bootstrap. In: **Monographs on Statistics and Applied Probability**. Chapman and Hall, New York, v.57, 1993.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. Wiley, New York, 2001.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN, W. **Applied linear statistical models**. McGraw-Hill, Boston, 1996.

PEREIRA, A. P. **O que deve saber sobre ozono**. Agência portuguesa do ambiente. Ministério do Ambiente, do Ordenamento do Território e do Desenvolvimento Regional: set. 1999. Disponível em <http://www.apambiente.pt/divulgacao/Publicacoes/ outrossuportes /Documents/ozono.pdf>. Acesso em: 30 abr 2009.

5. DIREITOS AUTORAIS

MULTIPLE LINEAR REGRESSION MODEL APPLIED TO DETERMINATE THE OZONE CONCENTRATION IN THE AIR

Dutra, Elisete Gomides, elisete.gomides@meioambiente.mg.gov.br(1)
Fioravante, Edwan Fernandes, edwan.fioravante@meioambiente.mg.gov.br(1)
Requeijo, José Gomes, jfgr@fct.unl.pt(2)

[1]Fundação Estadual do Meio Ambiente; Rodovia Prefeito Américo Gianetti, snº, Prédio Minas , 1º andar, Bairro Serra Verde, Belo Horizonte, Minas Gerais, Brasil, 31630-900
[2] Faculdade de Ciência e Tecnologia da Universidade Nova de Lisboa, 2829-516, Caparica, Portugal

Summary: *The air pollution related to ozone is a source of public health problem in many downtowns around the world. Ozone comes from atmosphere reactions between organic volatile* compositions *and nitrogen oxides with the sun radiation action. The automotive vehicles are the biggest responsible of emitting pollution and so by ozone formation. The ozone concentration, temperature and radiation in Belo Horizonte and nearest regions are captured through automatic stations managed by Fundação Estadual do Meio Ambiente – FEAM, the company in charge of the environment area in the Minas Gerais state. Data are shown here about the relation between the maximum ozone diary value, temperature and radiation was caught in Vespasiano in November 2009. It was used the Minimum Square method with two explicable variables, using SPSS software to determinate multiple linear regression of the best data fitting. The proposed model makes it possible to preview the ozone concentration for specific temperature scenery and in this way to define public politics that fit the better to avoid health damage to the region people. Though this model should be carefully used as the coefficient found can be lower than an useful value. This situation shows that the model would have taken in consideration more independent variables than those were used or it should be done by a non-linear regression. Another possible way should be obtained through using average ozone concentration by period of time instead of maximum diary ones. It would allow the concentrations to be followed day by day as the average temperature is changing.*

Key words: *Ozone; Vehicle pollution; Sun radiation; Temperature; Multiple linear regression.*