Mestrado Integrado em Engenharia Química

Desenvolvimento de Modelos Estatísticos Para Previsão das Concentrações de Ozono

Tese de Mestrado

desenvolvida no âmbito da disciplina de

Projecto de Desenvolvimento em Ambiente Académico

Vítor Casimiro Abreu Carneiro

Universidade do Porto
Faculdade de Engenharia
FEUP

Departamento de Engenharia Química

Orientador na FEUP: **Fernando Gomes Martins**

Co-orientador na FEUP: **Maria da Conceição Machado Alvim Ferraz**

Julho de 2008

Agradecimentos

Gostaria de agradecer a todos os docentes, amigos e familiares que me acompanharam ao longo de todo o meu percurso académico e que de alguma forma influenciaram na minha formação pessoal e/ou intelectual, permitindo-me alcançar uma etapa de enorme concretização pessoal como esta.

Um particular e sincero agradecimento ao Prof. Fernando Martins e à Prof. Conceição Alvim pelo acompanhamento e tempo dispendido, por toda a ajuda prestada, paciência e compreensão, factores sem os quais este projecto não estaria concluído. Agradeço também todos os conselhos e conhecimento que me foram transmitidos ao longo deste trabalho, que de certeza me acompanharão em trabalhos futuros.

Resumo

Devido ao impacto negativo que o ozono pode ter na saúde humana, animais ou plantas, a sua previsão em tempo útil revela-se cada vez mais importante.

Os desempenhos de três modelos estatísticos na previsão das concentrações de ozono foram testados. Os modelos criados foram baseados em regressões lineares múltiplas, modelos do tipo caixa cinzenta e aprendizagem modificada baseada em exemplos. Para o estudo desenvolvido foram utilizados valores dos anos 2003, 2004 e 2005, referentes à estação de medição dos poluentes das Antas, da Área Metropolitana do Porto.

Os resultados mostraram que a aprendizagem modificada através de exemplos obteve melhores índices de desempenho, portanto é o método mais eficaz de prever as concentrações de ozono.

Palavras-chave: Qualidade do ar, ozono, modelos estatísticos, previsão.

Abstract

Due to the negative impact of ozone in human health, animals or plants, the prediction of ozone concentrations in air ambient became a very important subject.

The main objective of this thesis is to evaluate the performances of three statistical models to predict ozone concentrations. The models created were based on multiple linear regression models, grey modeling and modified learning from examples. The studies were performed for the years 2003, 2004 and 2005, for an urban site (Antas) situated in Oporto Metropolitan Area, Portugal.

The results showed that the model modified learning from examples obtained better performance indexes. Therefore, this model is an effective method to predict concentrations of ozone.

Keywords: Quality of air, ozone, statistical models, forecasting.

Índice

Índice de Tabelas

Índice de Figuras

Notação e Glossário

d_2	Índice de concordância
n	Número de dados totais
R	Coeficiente de correlação

Regressão linear múltipla

β_i	Coeficiente i da regressão
$\hat{\beta}_i$	Coeficiente i da regressão previsto pelo modelo
h	Número total de variáveis exploratórias
x_i	Variável exploratória i
y	Variável resposta
$\hat{y}$	Variável resposta prevista pelo modelo
SSE	Somatório dos quadrados dos resíduos
Sxx_i	Somatório das diferenças quadráticas

Modelos do tipo caixa cinzenta unicomponente

$\boldsymbol{\varepsilon}^{(0)}$	Vector residual
a	Parâmetro da equação diferencial
b	Parâmetro da equação diferencial
k_a	Frequência mínima das séries de Fourier
T	Período do vector residual
$\mathbf{x}^{(0)}$	Vector variável resposta
$\bar{x}^{(0)}$	Previsão do modelo
$\mathbf{x}^{(1)}$	Vector cumulativo
$\hat{\mathbf{x}}^{(1)}$	Vector cumulativo calculado pelo modelo

Modelos do tipo caixa cinzenta multicomponente

α	Coeficiente de desenvolvimento do sistema
$\boldsymbol{\beta}$	Vector dos parâmetros do sistema
h	Número total de variáveis
T	Período do vector residual
$\mathbf{x_k}^{(0)}$	Vector da variável exploratória k
$\mathbf{x_1}^{(0)}$	Vector variável resposta

Aprendizagem modificada baseada em exemplos

$\boldsymbol{\sigma}_\mathbf{i}$	Vector linha correspondente aos valores dos desvios padrões da regra i

b_i	Valor de saída da regra i
$\mathbf{c_i}$	Vector linha correspondente aos valores centrais da regra i
e_f	Tolerância
r	Número total de regras
w	Valor de sobreposição
x_i	Variável exploratória i
y	Variável resposta
y_{modelo}	Variável resposta obtida pelo modelo

Índices

i	Índice ou contador
j	Índice ou contador
k	Índice ou contador

Lista de Siglas

AMP	Amplitude	
GM	Modelos do tipo caixa cinzenta	
GM(1,1)	Modelos do tipo caixa cinzenta unicomponente	
GM(1,h)	Modelos do tipo caixa cinzenta multicomponente	
MAE	Média dos erros absolutos	$\mu g/m^3$
MAX	Máximo	
MBE	Grau de parcialidade do erro	$\mu g/m^3$
MIN	Mínimo	
MLFE	Aprendizagem modificada baseada em exemplos	
MLR	Regressão linear múltipla	
RMSE	Raiz quadrada da média dos erros quadrados	$\mu g/m^3$

1 Introdução

1.1 Enquadramento e Apresentação do Projecto

Os fumos industriais, alguns incêndios florestais e a emissão de gases produzidos pela combustão de petróleo, carvão ou gás natural são consequência da actividade humana e produzem mudanças no ar atmosférico alterando a sua capacidade de protecção da vida.

Poluição define-se como a adição de qualquer substância ou forma de energia (por exemplo: calor, som, radioactividade) para o meio ambiente a um ritmo mais rápido do que o ambiente pode acomodá-la por dispersão, desagregação, reciclagem ou armazenagem de alguma forma inofensiva (Encyclopædia Britannica, 2008).

A poluição atmosférica define-se como a presença de um ou mais contaminantes colocados na natureza, em quantidades que podem causar danos ao homem, animais ou plantas; ou que possam alterar significativamente o equilíbrio de um ecossistema (Allen, D., 2002).

São conhecidos como principais poluentes do ar o dióxido de enxofre (SO_2), partículas em suspensão (PM), óxidos de azoto (NO_x), óxido de carbono (CO), chumbo (Pb), ozono (O_3), compostos orgânicos voláteis (COV), dioxinas, metais pesados entre outros.

Devido ao impacto negativo que estes compostos, designados por poluentes, têm no meio ambiente e na saúde humana torna-se pertinente medir, controlar e prever as concentrações destes compostos, de modo a inferir sobre a qualidade do ar que se respira.

Esta tese baseia-se na procura de modelos estatísticos de previsão das concentrações de ozono.

O ozono (O_3), molécula com três átomos de oxigénio, é o composto responsável pelo distinto odor no ar depois de uma tempestade ou no ar que rodeia um equipamento eléctrico. O O_3 é um gás azul pálido, irritante, explosivo e tóxico, mesmo em baixas concentrações. Sob certas condições, reacções fotoquímicas entre óxidos de azoto e hidrocarbonetos na atmosfera podem produzir O_3, em concentrações altas que permitem causar irritação dos olhos e mucosas (Encyclopædia Britannica, 2008).

A procura de modelos estatísticos de previsão das concentrações deste composto, tem como objectivo principal alertar para situações em que as taxas estejam próximas e/ou ultrapassem os valores limites legislados para este composto. O Decreto-Lei n.º 320/2003, de 20 de Dezembro, estabelece como objectivos a longo prazo valores alvo, um limiar de alerta e um limiar de informação ao público para as concentrações do O_3 no ar ambiente, bem como as regras de gestão da qualidade do ar aplicáveis a esse poluente, transpondo para a ordem

jurídica nacional a Directiva europeia n.º 2002/3/CE, de 12 de Fevereiro, relativa ao O_3 no ar ambiente, onde é dito:

"Tendo em vista a protecção da população em geral, deve estabelecer-se um limiar de alerta para o O_3. Deve estabelecer-se um limiar de informação destinado a alertar e proteger elementos sensíveis da população.

Devem sistematicamente divulgar-se ao público informações actualizadas sobre as concentrações de O_3 no ar ambiente."

De acordo com a mesma directiva, uma ferramenta eficaz de previsão também reduziria de forma significativa o número de zonas de medição.

"Deve ser obrigatório efectuar medições nas zonas em que são excedidos os objectivos a longo prazo. O recurso a meios complementares permitirá reduzir o número de pontos de amostragem fixos necessário."

1.2 Contributos do Trabalho

No decorrer deste trabalho foram desenvolvidos e testados vários modelos estatísticos com vista a se dispor de ferramentas capazes, de em tempo útil, poderem prever concentrações de O_3 no ar ambiente.

O trabalho desenvolvido consistiu na aplicação de três modelos distintos: a regressão linear múltipla (MLR), modelos do tipo caixa cinzenta (GM) e a uma técnica de aprendizagem modificada baseada em exemplos (MLFE).

Enquanto que vários estudos baseados em modelos de MLR têm vindo a ser realizados nos últimos anos, tanto quanto se sabe, a aplicação de modelos do tipo GM a este tipo de problemas, nunca foi testada anteriormente, e relativamente à aplicação da MLFE apenas se conhecem dois estudos realizados por Heo e Kim (2004) e Mintz et al. (2005).

No que diz respeito ao desenvolvimento de modelos baseados em MLR, este trabalho teve como objectivo principal avaliar o impacto da estrutura de dados nos desempenhos dos modelos.

1.3 Organização da Tese

A tese encontra-se dividida em 6 capítulos.

No primeiro capítulo é apresentado o tema desta tese, a necessidade do seu estudo e a metodologia usada na sua abordagem.

No Capítulo 2 é feito um enquadramento mais centrado da necessidade deste projecto e são introduzidos os principais métodos de resolução do problema objecto.

No Capítulo 3, os modelos utilizados para a construção das previsões são descritos detalhadamente. Os índices que efectuaram as medidas dos seus desempenhos são também enumerados.

No Capítulo 4 encontra-se uma descrição global das técnicas e variantes usadas para a construção dos modelos. Os resultados finais, bem como as observações mais relevantes podem também ser encontrados nesta secção.

No Capítulo 5, os resultados obtidos são enquadrados com os objectivos enunciados e são retiradas as conclusões mais importantes.

Finalmente, no Capítulo 6, é feita uma análise auto-crítica, bem como possíveis futuras melhorias do trabalho realizado.

2 Estado da Arte

Um tópico suficientemente imergente no meio científico é o estudo do ar que se respira e em particular a poluição atmosférica (Schlink et al., 2006).

São vários os efeitos negativos de elevadas concentrações de O_3 no ar ambiente. A área de maior estudo durante os últimos anos tem sido no sentido de comprovar e identificar os efeitos adversos do O_3 no ar ambiente para a saúde humana (Kelsall et al., 1997; Hoek et al., 1997; Zmirou et al., 1998; Schwartz, 2000). Os indicadores de saúde humana considerados incluem medições da função pulmonar, sintomas respiratórios, registos de admissões hospitalares e de mortalidade. Em contraste, outros trabalhos indicaram que não encontraram quaisquer efeitos ou que estes são muitos reduzidos (Borja-Aburto et al., 1997; Lee et al., 1999). Assim, atendendo à diversidade de conclusões, há uma divergência na avaliação do impacto do O_3 na saúde humana.

Tendo em conta a possível ameaça do O_3 para a saúde humana, uma ferramenta capaz de efectuar a sua previsão com sucesso, permitirá desenvolver mecanismos que possam evitar elevadas exposições.

Tal como é referido em Sousa et al. (2006), os modelos de previsão das concentrações do O_3 dividem-se em 2 tipos principais: i) os modelos do tipo 1 que se baseiam nas leis físicas; e ii) modelos do tipo 2, baseados em técnicas estatísticas. Os modelos do tipo 1 são bastantes complexos devido à necessidade de conhecer aspectos particulares das reacções, como as equações de cinética, que ainda não são bem conhecidas (Soja e Soja, 1999; Ballester et al., 2002). Em contraste, a simplicidade dos modelos estatísticos quando comparados com os do tipo 1, é uma vantagem para a sua implementação. Contudo, devido ao facto de se basearem em informações dos locais, tornam-se bastante específicos.

Na literatura há uma grande diversidade de aproximações estatísticas para a previsão das concentrações de O_3. Os principais modelos usados baseiam-se em redes neuronais artificiais (Sousa et al., 2006; Sousa et al., 2007), regressões lineares (Millonis e Davies, 1994; Hubbard e Cobourn, 1998; Sousa et al., 2006; Sousa et al., 2007; Pires et al., 2008), função de transferência (Arroyo-Lopez et al., 1999), aprendizagem modificada baseada em exemplos (Heo e Kim, 2004; Mintz et al., 2005), entre outros.

Em Sousa et al. (2006) é realizada uma comparação entre três modelos distintos: séries temporais, regressão linear múltipla e redes neuronais artificiais. Para a aplicação destes modelos foram usados dados relativos ao dia anterior e de dois dias antes para efectuar a previsão. Os resultados foram favoráveis às redes neuronais artificiais, obtendo melhores

índices de desempenho, sendo então o método mais eficiente de prever as concentrações de O_3. Estudo idêntico realizado por Chaloulakou et al. (2003) obteve os mesmos resultados. Sousa et al. (2007) efectuaram um novo estudo de previsão. Desta vez focaram-se em regressões lineares múltiplas e redes neuronais artificiais, baseadas na técnica de análise por componentes principais. As previsões para o dia seguinte foram efectuadas através de dados relativos ao dia anterior usando uma nova metodologia baseada em componentes principais. Os resultados mostraram que o uso desta nova técnica melhorou ambos os modelos reduzindo a sua complexidade e eliminando a problemas de colinearidade.

Na mesma área de estudo, Pires et al. (2008) apresentaram vários métodos de selecção e validação de parâmetros em regressões lineares múltiplas e de componentes principais. Dados relativos às condições meteorológicas e ambientais, do dia anterior, foram usados para a previsão. Os resultados mostraram que, usando as variáveis originais, cada método escolhido resultava num modelo diferente e que o processamento prévio de fazer a transformação para componentes principais permitiria resolver este problema.

Uma inter-comparação entre quinze modelos estatísticos diferentes é apresentada por Schlink et al. (2003), demonstrando que as redes neuronais artificiais e os modelos aditivos generalizados deveriam ser usados na maior parte das situações, realçando a sua capacidade de manipular associações não-lineares.

Mintz et al. (2005) implementou um novo método adaptativo de aprendizagem, o qual designou por aprendizagem modificada baseada em exemplos (modified learning from examples, MLFE), dando boas indicações da sua capacidade de previsão das concentrações de O_3. Segundo estes autores, trata-se de modelos do tipo caixa preta devido às suas variáveis serem muitas vezes imprecisas e incertas, mas produzem soluções eficazes para sistemas não-lineares e em sistemas parcialmente desconhecidos.

3 Modelos

Nas secções seguintes faz-se uma descrição detalhada dos modelos utilizados neste trabalho.

3.1 Regressão Linear Múltipla (MLR)

Regressões lineares múltiplas modelam a relação entre duas ou mais variáveis exploratórias e a variável de resposta, através do ajuste de uma equação linear aos dados observados.

Para a construção do modelo devem ser seguidos os seguintes passos:

Passo 1: A variável de resposta é calculada por:

$$\mathrm{y} = \beta_0 + \sum_{\mathrm{i}=1}^{\mathrm{h}} \beta_{\mathrm{i}} \mathrm{x}_{\mathrm{i}} + \varepsilon \quad , i = 1,2,\ldots,h \qquad (1)$$

Onde x_i representa as variáveis exploratórias independentes, β_i os coeficientes da regressão e ε o erro associado à regressão. O número total de variáveis exploratórias é representado por h.

Passo 2: O método mais usado para estimar os parâmetros da regressão é a minimização do somatório dos quadrados dos resíduos. A equação é a seguinte:

$$\hat{\beta}_{\mathrm{i}} = \arg\min \sum_{\mathrm{i}=1}^{\mathrm{n}} (y_i - \hat{y}_i)^2 \, , i = 1,2,\ldots,n \qquad (2)$$

O número total de pontos utilizados na fase de construção (conjunto de treino) do modelo é dado por n, sendo $\hat{y}$ o valor previsto pelo modelo.

Passo 3: O valor previsto pelo modelo é calculado através:

$$\hat{\mathrm{y}} = \hat{\beta}_0 + \sum_{\mathrm{i}=1}^{\mathrm{h}} \hat{\beta}_{\mathrm{i}} \mathrm{x}_{\mathrm{i}} \qquad (3)$$

Passo 4: As significâncias dos coeficientes da regressão são avaliadas através dos seus intervalos de confiança. Um parâmetro $\hat{\beta}_i$ é válido se:

$$\left|\hat{\beta}_i\right| > \frac{t_{n-h-1}^{\alpha/2}\hat{\sigma}}{\sqrt{Sxx_i}} \qquad (4)$$

Onde t é a distribuição de t de Student, α é o nível de significância, $\hat{\sigma}$ é o desvio padrão dado por $\sqrt{SSE/(n-h-1)}$, sendo *SSE* o somatório dos quadrados dos resíduos e Sxx_i o somatório das diferenças quadráticas relacionadas com x_i, calculado por $\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2$.

3.2 Modelos do Tipo Caixa Cinzenta (GM)

O grau de conhecimento da informação sobre o modelo de um dado processo leva a que este possa ser designado por modelo de caixa branca, modelo de caixa cinzenta e modelo de caixa preta.

Os modelos designados por modelos do tipo caixa branca caracterizam-se pela sua natureza adaptável e capacidade de extrapolação. A sua principal desvantagem relaciona-se com o facto de os modelos encontrados não se ajustarem adequadamente aos dados existentes.

Contrariamente a estes modelos, os modelos do tipo caixa preta apresentam uma elevada capacidade de ajuste, mas são muito menos robustos em zonas de extrapolação.

Os modelos do tipo caixa cinzenta funcionam como uma combinação dos dois tipos de modelos referidos anteriormente.

3.2.1 Modelo do Tipo Caixa Cinzenta Unicomponente, GM(1, 1)

A implementação mais usual do modelo do tipo caixa cinzenta é através da utilização de apenas uma variável (variável de resposta) e da aplicação de uma equação diferencial de primeira ordem para correlacionar a informação existente dessa variável.

Seja um vector de dados da variável resposta designado por $\mathbf{x^{(0)}}$, em que $\mathbf{x^{(0)}} = x^{(0)}(1), x^{(0)}(2), x^{(0)}(3),\ldots,x^{(0)}(j),\ldots,x^{(0)}(n)$. O elemento $x^{(0)}(j)$ corresponde ao valor da variável no instante de tempo j e n é o número de dados utilizados. Os principais passos do desenvolvimento do modelo são:

Passo 1: Criação um vector cumulativo, $\mathbf{x^{(1)}}$, através da Equação 5:

$$\mathrm{x}^{(1)}(\mathrm{k}) = \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{x}^{(0)}(\mathrm{j}) \tag{5}$$

Obtendo-se o vector $\mathbf{x^{(1)}} = x^{(1)}(1), x^{(1)}(2), x^{(1)}(3),\ldots,x^{(1)}(j),\ldots,x^{(1)}(n)$.

Passo 2: Definição dos parâmetros a e b.

O vector $\mathbf{x^{(1)}}$ é ajustado a uma equação diferencial de primeira ordem da seguinte forma:

$$\frac{d\mathrm{x}^{(1)}(t)}{dt} + a\mathrm{x}^{(1)}(t) = \mathrm{b} \tag{6}$$

Passo 3: Estimação dos valores de a e b.

$$\begin{bmatrix} a \\ b \end{bmatrix} = (\mathbf{B^T B})^{-1}\mathbf{B^T Y_N} \tag{7}$$

onde,

$$\mathbf{B} = \begin{bmatrix} -0.5(x^{(1)}(1) + x^{(1)}(2)) & 1 \\ -0.5(x^{(1)}(2) + x^{(1)}(3)) & 1 \\ \vdots & \vdots \\ -0.5(x^{(1)}(n-1) + x^{(1)}(n)) & 1 \end{bmatrix} \tag{8}$$

e

$$\mathbf{Y_N} = (\mathrm{x}^{(0)}(2), \mathrm{x}^{(0)}(3), \ldots, \mathrm{x}^{(0)}(\mathrm{n}))^{\mathrm{T}} \tag{9}$$

Passo 4: Obtenção da componente tendencial do modelo.

O cálculo dos valores de *a* e de *b* da Equação 6 permite determinar a componente tendencial do modelo. A condição inicial é definida a partir *de* $x^{(1)}(n)$ e não de

$x^{(1)}(0)$. Este procedimento permite dar mais relevo à tendência dos últimos valores, uma vez que estes encontram-se mais próximos, em termos temporais, da solução que se pretende encontrar. A solução geral da Equação 6 é dada por:

$$\hat{x}^{(1)}(k) = \left(x^{(1)}(n) - \frac{b}{a} \right) e^{-a(k-n)} + \frac{b}{a} \tag{10}$$

onde $\hat{x}^{(1)}(k)$ corresponde ao valor de $x^{(1)}(k)$ obtido pelo modelo. Através da substituição $k=(n+1)$ na Equação 6, faz-se a previsão de $x^{(1)}(k)$ para o passo seguinte, sendo obtido por:

$$\hat{x}^{(1)}(n+1) = \left(x^{(1)}(n) - \frac{b}{a} \right) e^{-a} + \frac{b}{a} \tag{11}$$

Finalmente, obtém-se a previsão de $x^{(0)}(k+1)$, através da Equação 12:

$$x^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) \tag{12}$$

Desta forma, fica definida a componente tendencial da previsão.

A determinação da componente periódica engloba uma série de passos, descritos de seguida:

Passo 1: Estimação do vector residual $\boldsymbol{\varepsilon}^{(0)}$ a partir da Equação 13:

$$\varepsilon^{(0)}(k) = x^{(0)}(k) - \hat{x}^{(0)}(k) \tag{13}$$

para $\mathrm{k} = 2,3,\ldots,\mathrm{n}$

Resultando no vector $\boldsymbol{\varepsilon}^{(0)} = (\varepsilon^{(0)}(2), \varepsilon^{(0)}(3), \ldots \varepsilon^{(0)}(n))^T$ (14)

Séries de Fourier são usadas na determinação da componente periódica inerente ao vector residual.

Passo 2: O vector residual é modelado usando séries de Fourier através de:

$$\hat{\varepsilon}^{(0)}(k) = \frac{1}{2} a_0 + \sum_{i=1}^{k_a} \left(a_i \cos\left(\frac{i2\pi}{T} k \right) + b_i \sin\left(\frac{i2\pi}{T} k \right) \right) \tag{15}$$

para $k = 2,3,\ldots,n$.

onde T indica o período do vector residual, sendo igual a n-1. k_a é a frequência mínima das séries de Fourier, e corresponde ao valor inteiro de $[(n-1)/2-1]$.

Passo 3: Os parâmetros a_0, a_i e b_i são estimados através da Equação 16:

$$\mathbf{C} = \left(\mathbf{P}^{\mathbf{T}}\mathbf{P}\right)^{-1}\mathbf{P}^{\mathbf{T}}\boldsymbol{\varepsilon}^{(\mathbf{0})}, \tag{16}$$

sendo $$\mathbf{C} = \left(a_0, a_1, b_1, a_2, b_2, ..., , a_{k_a}, b_{k_a}\right)^T \tag{17}$$

e

$$\mathbf{P} = \begin{bmatrix} 1/2 & \cos\left(\frac{2\pi\times 1}{\mathrm{T}}2\right) & \sin\left(\frac{2\pi\times 1}{\mathrm{T}}2\right) & \cos\left(\frac{2\pi\times 2}{\mathrm{T}}2\right) & \sin\left(\frac{2\pi\times 2}{\mathrm{T}}2\right) & \dots & \cos\left(\frac{2\pi\times \mathrm{k_a}}{\mathrm{T}}2\right) & \sin\left(\frac{2\pi\times \mathrm{k_a}}{\mathrm{T}}2\right) \\ 1/2 & \cos\left(\frac{2\pi\times 1}{\mathrm{T}}3\right) & \sin\left(\frac{2\pi\times 1}{\mathrm{T}}3\right) & \cos\left(\frac{2\pi\times 2}{\mathrm{T}}3\right) & \sin\left(\frac{2\pi\times 2}{\mathrm{T}}3\right) & \dots & \cos\left(\frac{2\pi\times \mathrm{k_a}}{\mathrm{T}}3\right) & \sin\left(\frac{2\pi\times \mathrm{k_a}}{\mathrm{T}}3\right) \\ 1/2 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1/2 & \cos\left(\frac{2\pi\times 1}{\mathrm{T}}\mathrm{n}\right) & \sin\left(\frac{2\pi\times 1}{\mathrm{T}}\mathrm{n}\right) & \cos\left(\frac{2\pi\times 2}{\mathrm{T}}\mathrm{n}\right) & \sin\left(\frac{2\pi\times 2}{\mathrm{T}}\mathrm{n}\right) & \dots & \cos\left(\frac{2\pi\times \mathrm{k_a}}{\mathrm{T}}\mathrm{n}\right) & \sin\left(\frac{2\pi\times \mathrm{k_a}}{\mathrm{T}}\mathrm{n}\right) \end{bmatrix} \tag{18}$$

Os novos valores do vector residual formam o vector $\hat{\boldsymbol{\varepsilon}}^{(\mathbf{0})} = (\hat{\varepsilon}^{(0)}(2), \hat{\varepsilon}^{(0)}(3), ..., \hat{\varepsilon}^{(0)}(n))^T$, resultando na componente periódica da previsão.

Passo 4: Por último, a tendência e a componente periódica residual do sistema são adicionadas, originando a previsão para o passo seguinte:

$$\bar{x}^{(0)}(n+1) = \hat{x}^{(0)}(n+1) + \hat{\varepsilon}^{(0)}(n+1) \tag{19}$$

3.2.2 Modelo Tipo Caixa Cinzenta Multicomponente, GM(1, h)

Este modelo é bastante idêntico ao anterior, constituído pelas duas componentes já referidas anteriormente – componente tendencial e componente periódica. Enquanto que a segunda obtém-se da mesma maneira, a primeira sofre algumas alterações, descritas de seguida.

Este modelo difere do modelo unicomponente pelo facto de este possuir, para além do vector com os dados principais (variável de resposta), um ou mais vectores com os factores influenciadores (variáveis exploratórias).

Passo 1: Define-se o vector $\mathbf{x_1^{(0)}} = \mathrm{x_1}^{(0)}(1), \mathrm{x_1}^{(0)}(2), ..., \mathrm{x_1}^{(0)}(\mathrm{n})$, como sendo a sequência dos dados principais, onde *n* representa o número total de dados utilizados.

Os vectores $\mathbf{x_k^{(0)}} = \mathrm{x_k}^{(0)}(1), \mathrm{x_k}^{(0)}(2), ..., \mathrm{x_k}^{(0)}(\mathrm{n}), \mathrm{k} = 2, 3, ..., \mathrm{h}$ são os vectores das variáveis exploratórias, sendo $(h-1)$ o número total dessas variáveis.

Passo 2: O vector cumulativo de cada uma das variáveis exploratórias calcula-se através da Equação 20:

$$\mathrm{x_k}^{(1)}(\mathrm{i}) = \sum_{\mathrm{j}=1}^{\mathrm{i}} \mathrm{x_k}^{(0)}(\mathrm{j}) \quad , i = 1,2,...,n \quad e \quad k = 1,2,...,h \tag{20}$$

Passo 3: A equação diferencial do modelo cinzento, GM(1,h) toma agora a forma:

$$x_1^{(0)}(i) + \alpha z_1^{(1)}(i) = \sum_{j=2}^{h} \beta_j x_j^{(1)}(i) \quad , i = 1,2,...,n \tag{21}$$

onde α é o coeficiente de desenvolvimento do sistema, $\boldsymbol{\beta}$ é o vector dos parâmetros do sistema e $z_1^{(1)}(i) = \frac{1}{2}\left(x_1^{(1)}(i) + x_1^{(1)}(i-1)\right) \quad , i = 1,2,...,n$ (22)

Resultando em, $\frac{dx_1^{(1)}}{dt} + a x_1^{(1)} = \sum_{j=2}^{h} b_j x_j^{(1)}$ (23)

Passo 4: A integração da Equação 23 pelo método de Euler resulta na Equação 24, admitindo o passo de integração igual a 1.

$$x_1^{(1)}(i+1) = a x_1^{(1)}(i) + \sum_{j=2}^{h} b_j x_j^{(1)}(i) \quad , i = 1,2,...,n-1 \tag{24}$$

Passo 5: Através da minimização da diferença do quadrado dos resíduos entre os valores experimentais e os obtidos pela Equação 24 determina-se os parâmetros $[a, b_2, b_3, ..., b_h]$ através de:

$$\mathbf{B}\underline{\mathbf{a}} = \mathbf{Y_N} \tag{25}$$

sendo

$$\underline{\mathbf{a}} = \left[a, b_2, b_3, \ldots, b_h\right]^T \tag{26}$$

$$\mathbf{B} = \begin{bmatrix} \sum_{i=1}^{n-1}(x_1^{(1)}(i) \times x_1^{(1)}(i)) & \sum_{i=1}^{n-1}(x_1^{(1)}(i) \times x_2^{(1)}(i)) & \cdots & \sum_{i=1}^{n-1}(x_1^{(1)}(i) \times x_h^{(1)}(i)) \\ \sum_{i=1}^{n-1}(x_2^{(1)}(i) \times x_1^{(1)}(i)) & \sum_{i=1}^{n-1}(x_2^{(1)}(i) \times x_2^{(1)}(i)) & \cdots & \sum_{i=1}^{n-1}(x_2^{(1)}(i) \times x_h^{(1)}(i)) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n-1}(x_h^{(1)}(i) \times x_1^{(1)}(i)) & \sum_{i=1}^{n-1}(x_h^{(1)}(i) \times x_2^{(1)}(i)) & \cdots & \sum_{i=1}^{n-1}(x_h^{(1)}(i) \times x_h^{(1)}(i)) \end{bmatrix} \tag{27}$$

$$\mathbf{Y_N} = \begin{bmatrix} \sum_{i=1}^{n-1}(x_1^{(1)}(i) \times x_1^{(1)}(i+1)) \\ \sum_{i=1}^{n-1}(x_2^{(1)}(i) \times x_1^{(1)}(i+1)) \\ \vdots \\ \sum_{i=1}^{n-1}(x_h^{(1)}(i) \times x_1^{(1)}(i+1)) \end{bmatrix} \tag{28}$$

Passo 6: Através dos parâmetros calculados no passo anterior e da Equação 29 é possível efectuar a previsão da componente tendencial para o instante $n+1$.

$$x_1^{(1)}(n+1) = a x_1^{(1)}(n) + \sum_{j=2}^{h} b_j x_j^{(1)}(n) \tag{29}$$

A determinação da componente periódica segue a metodologia já descrita na Secção 3.1.1 para o modelo tipo caixa cinzenta unicomponente.

3.3 Aprendizagem Modificada baseada em Exemplos (MLFE)

A técnica MLFE permite criar modelos através de um conjunto de regras desenvolvidas a partir de dados disponíveis.

As regras são constituídas a partir de duas matrizes e um vector, em que cada linha dessas matrizes e vector correspondem a desvios padrões, valores centrais e valor de saída para cada regra usada, respectivamente:

$$\boldsymbol{\sigma}_{\mathbf{i}} = [\sigma_{i1}, \sigma_{i2}, ..., \sigma_{ih}] \quad (30)$$

$$\mathbf{c}_{\mathbf{i}} = [c_{i1}, c_{i2}, ..., c_{ih}] \quad (31)$$

$$\mathbf{b}_{\mathbf{i}} = [b_1, b_2, ..., b_h]^T \quad (32)$$

Onde i varia de 1 até r, sendo este último o numero total de regras criadas. h representa o número total de variáveis exploratórias.

O modelo e as regras desenvolvem-se usando os passos seguintes:

Passo 1: Definição da tolerância (e_f) admitida e do valor de sobreposição (w).

Passo 2: Definição da primeira regra do modelo através do primeiro exemplo de dados.

Sendo a matriz correspondente às variáveis exploratórias representada por:

$$\mathbf{x}_{\mathbf{i}} = [x_{i1}, x_{i2}, ..., x_{ih}], \quad i = 1,2,..., n \quad (33)$$

onde *n* representa o número total de dados e o vector da variável resposta é:

$$\mathbf{y} = [y_1, y_2, ..., y_r]^T \quad (34)$$

Resultando na Regra 1 com:

$$\boldsymbol{\sigma}_{\mathbf{1}} = [\sigma_{11}, \sigma_{12}, ..., \sigma_{1h}] = \left[\frac{abs(x_{11})}{10}, \frac{abs(x_{12})}{10}, ..., \frac{abs(x_{1h})}{10}\right]$$

$$\mathbf{c}_{\mathbf{1}} = [c_{11}, c_{12}, ..., c_{1h}] = [x_{11}, x_{12}, ..., x_{1h}]$$

$$\mathbf{b}_{\mathbf{1}} = y_1$$

Passo 3: Leitura dos exemplos de aprendizagem seguintes, $\mathbf{x}_{\mathbf{i}} = [x_{i1}, x_{i2}, ..., x_{ih}]$ e y_i.

Passo 4: Avaliação de y pelo modelo. Obtida por:

$$\mu_j = \prod_{i=1}^{h} e^{-\frac{1}{2}\left(\frac{x_i - c_{ji}}{\sigma_{ji}}\right)}, \; j = 1,2,..., r \quad (35)$$

$$y_{\text{modelo}=} \frac{\sum_{i=1}^{r}(b_i \mu_i)}{\sum_{i=1}^{r}(\mu_i)} \tag{36}$$

Passo 5: Critério de decisão:

Se $|y_{\text{modelo}} - y_i| > e_f$, é adicionada uma nova regra: $b_j = y_i$, $c_j = x_i$ e $\sigma_j = \frac{\max|c_j - c|}{w}$

Passo 6: Se todos os pontos de aprendizagem já tiverem sido considerados, terminar e avaliar o desempenho do modelo em pontos não considerados durante a aprendizagem. Caso contrário, voltar ao passo 3.

3.4 Medidas de Desempenho dos Modelos

As medidas de desempenho de modelos estatísticos são determinadas a partir de um conjunto de indicadores, como sejam:

$$R = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2}} \tag{37}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i) \tag{38}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{39}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{40}$$

$$d_2 = \frac{\left[\sum_{i=1}^{n}|\hat{Y}_i - Y_i|^2\right]}{\left[\sum_{i=1}^{n}|\hat{Y}_i - \overline{Y}_i|^2 + \sum_{i=1}^{n}|Y_i - \overline{Y}_i|^2\right]} \tag{41}$$

O valor de R (coeficiente de correlação) é a medida mais comum relacionada com a medida de qualidade do ajuste do modelo. No entanto, este indicador não fornece informação sobre a precisão do modelo. Com este objectivo surge o MBE (grau de parcialidade do erro), que indica se o modelo estima os valores observados por defeito ou por excesso. O MAE (média dos erros absolutos) e o RMSE (raiz quadrada da média dos quadrados dos erros) medem os erros residuais, dando uma ideia global da diferença entre os valores obtidos pelo modelo e os valores observados. O valor de d_2 permite comparar as diferenças entre a média, a previsão e os valores das concentrações observados, indicando o grau de liberdade do erro para a previsão (Gardner e dorling, 2000; chaloulakou et al., 2003; Sousa et al., 2005). Os valores de RMSE, MAE e MBE apresentam as mesmas unidades da variável de previsão.

4 Aplicação dos Modelos, Resultados e Discussão

4.1 Descrição dos dados recolhidos

O período de estudo correspondeu aos anos de 2003, 2004 e 2005, referente a valores de concentrações de poluentes da cidade do Porto para a estação das Antas. A informação meteorológica foi obtida na Serra do Pilar (Vila Nova de Gaia). Os dados foram cedidos pela *Comissão de Coordenação e Desenvolvimento Regional do Norte* e pelo *Instituto Geofísico da Faculdade de Ciências da Universidade do Porto* e dividem-se em dois grupos:

1. Meteorológicos:
    - Temperatura (ºC)
    - Humidade relativa (%)
    - Velocidade do vento (km/h)
2. Ambientais
    - Concentrações de CO ($\mu g/m^3$)
    - Concentrações de NO ($\mu g/m^3$)
    - Concentrações de NO_2 ($\mu g/m^3$)
    - Concentrações de O_3 ($\mu g/m^3$)

Os dados foram agrupados de vários modos de forma a desenvolver e avaliar modelos de diferentes tipos.

4.2 Estrutura dos modelos

4.2.1 Regressão Linear Múltipla (MLR)

Vários modelos foram desenvolvidos baseados em regressões lineares múltiplas com diferentes estruturas de dados, das quais resultaram diferentes modelos. As variáveis exploratórias foram normalizadas.

Numa primeira fase, foram desenvolvidos modelos de MLR com o objectivo de prever as concentrações de O_3 para a hora seguinte com base nos valores das concentrações de O_3 e das outras variáveis na hora anterior.

Diferentes conjuntos de dados foram criados correspondentes a diferentes períodos, de acordo com o descrito seguidamente:

Ajuste 1a – Ajuste com os dados correspondentes a um ano (1 modelo).

Ajuste 2a - Ajuste com os dados do período nocturno correspondentes a um ano (1 modelo).

Ajuste 3a - Ajuste com os dados do período diurno correspondentes a um ano (1 modelo).

Ajustes 4a a 7a – Modelos trimestrais com todos os dados correspondentes ao mesmo período (4 modelos).

Ajustes 8a a 11a - Modelos trimestrais com dados nocturnos correspondentes ao mesmo período (4 modelos).

Ajustes 12a a 15a - Modelos trimestrais com dados diurnos correspondentes ao mesmo período (4 modelos).

O período nocturno considerado foi entre as 20 e as 8 horas para o período de Inverno (Outubro - Março) e entre as 21 e as 7 horas para o Verão (Abril - Setembro).

Posteriormente fez-se uma nova abordagem, utilizando dia anterior, à mesma hora, para obter a previsão do dia seguinte à mesma hora. Por exemplo, para prever o dia 28 de Janeiro às 14 horas, são utilizados dados do dia 27 de Janeiro às 14 horas. Esta nova abordagem permite efectuar uma previsão com 1 dia de antecedência.

Os dados foram divididos em 2 conjuntos: conjunto de treino e conjunto de teste. O primeiro conjunto permite determinar o modelo e o segundo serve de verificação do desempenho do modelo (conjunto formado pelos dados dos últimos 15 dias de cada trimestre).

Desta vez as configurações de dados estudadas foram as seguintes:

Ajuste 1b – Ajuste com todos os dados correspondentes a um ano (1 modelo).

Ajustes 2b a 5b – Modelos trimestrais com todos os dados correspondentes ao mesmo período (4 modelos).

Ajustes 6b a 7b – Modelos semestrais com todos os dados correspondentes ao mesmo período (2 modelos).

Ajustes 8b a 31b – Modelos para todas as horas com os dados correspondentes a um ano (24 modelos – um modelos para cada hora).

Ajustes 32b a 127b – Modelos para todas as horas e para cada trimestre com todos os dados correspondentes ao mesmo período (96 modelos).

Ajustes 128b a 175b – Modelos para todas as horas e para cada semestre com os dados nocturnos correspondentes ao mesmo período (48 modelos).

4.2.2 Modelos do Tipo Caixa Cinzenta (GM)

Modelos do tipo GM(1,1) foram também desenvolvidos para previsão do dia seguinte à mesma hora. Foram criados modelos "hora a hora" específicos para cada hora (24 ajustes). Com este modelo definiu-se que a cada 30 novos valores consecutivos era gerado um modelo e efectuada a previsão para o dia imediatamente seguinte à mesma hora. É assim definida a fase de treino e a fase de teste respectivamente.

Devido ao GM(1,1) apenas necessitar da variável de resposta, as variáveis exploratórias (variáveis meteorológicas e os valores das concentrações dos outros poluentes) não foram aqui consideradas.

O processo de desenvolvimento dos modelos do tipo GM(1,h) é muito idêntico ao desenvolvimento de modelos do tipo GM(1,1), já descrito anteriormente. As diferenças baseiam-se na introdução das variáveis exploratórias nos modelos e no estudo do desempenho deste modelo na previsão da hora seguinte. Manteve-se o estudo "hora a hora" já relatado.

4.2.3 Aprendizagem Modificada baseada em Exemplos (MLFE)

Os modelos baseados em aprendizagem modificada baseada em exemplos foram igualmente desenvolvidos para estimar os valores das concentrações de O_3 para a hora seguinte (ajuste global com horas consecutivas) e para o dia seguinte à mesma hora (ajuste global com todas as horas para o dia seguinte).

Adicionalmente, foram desenvolvidos modelos específicos para cada hora para o dia seguinte (24 ajustes).

Foi escolhido como período de teste deste modelo, os dados correspondentes aos últimos 30 dias de cada modelo criado.

Tal como para as MLR, as variáveis exploratórias foram normalizadas.

4.3 Apresentação e discussão de resultados

4.3.1 Regressão Linear Múltipla (MLR)

Como exemplo dos ajustes referidos na Secção 4.2.1, relativa à estrutura dos modelos, apresenta-se na Tabela 1 os resultados dos modelos criados por MLR (previsão para a hora seguinte) para o ano de 2005.

As células assinaladas com "---" significam que o factor exploratório correspondente não tem significância para a regressão. Os valores apresentados são correspondentes à fase de treino.

Tabela 1 - Coeficiente de correlação, número de dados utilizados e parâmetros da regressão obtidos para os ajustes criados por MLR (previsão para a hora seguinte), relativos ao ano de 2005.

	P_0	CO	NO	NO_2	NO_2/NO	O_3	T	HR	WS	n	R
Anual	35,31	1,23	0,58	-1,83	-1,88	24,23	1,87	-0,54	---	7756	0,929
Diurno	41,11	1,10	1,31	-2,17	-4,21	26,92	3,09	---	-0,79	4488	0,933
Noct.	27,35	0,69	-1,24	1,22	---	19,28	0,73	---	2,11	3268	0,922
1º Trim	28,06	---	1,08	-1,09	-0,79	20,28	---	---	---	2031	0,924
Diurno	29,28	---	---	---	-3,46	21,74	---	0,66	---	1254	0,925
Noct.	26,10	---	-3,21	4,46	0,72	19,33	0,91	---	2,47	777	0,943
2º Trim	46,61	1,95	---	-1,42	-1,69	22,74	1,09	-0,88	0,63	2045	0,921
Diurno	57,27	1,61	1,96	-1,73	-3,39	22,10	1,68	---	---	1102	0,914
Noct.	34,16	1,78	-1,45	---	---	18,93	---	---	2,46	943	0,904
3º Trim	39,20	1,19	1,92	-2,76	-2,51	26,93	2,79	-1,16	-1,20	1987	0,930
Diurno	53,13	---	3,28	-1,94	-5,51	28,99	3,68	---	-2,01	1077	0,919
Noct.	22,71	1,05	---	---	0,81	16,78	---	---	2,13	910	0,909
4º Trim	25,79	1,69	---	-2,12	-0,72	19,66	1,60	---	0,71	1693	0,909
Diurno	26,01	2,07	---	-2,44	-2,73	21,41	1,55	---	---	1055	0,904
Noct.	25,43	---	-2,95	3,47	---	18,63	1,22	---	2,54	638	0,936

Em que:

P_0 = Ordenada na origem, ($\mu g/m^3$);

CO = Parâmetro da regressão correspondente às concentrações de CO;

NO = Parâmetro da regressão correspondente às concentrações de NO;

NO_2 = Parâmetro da regressão correspondente às concentrações de NO_2;

NO_2/NO = Parâmetro da regressão correspondente à razão entre NO_2 e NO;

O_3 = Parâmetro da regressão correspondente às concentrações de O_3;

T = Parâmetro da regressão correspondente à temperatura;

HR = Parâmetro da regressão correspondente à humidade relativa;

WS = Parâmetro da regressão correspondente à velocidade do vento;

R = Coeficiente de correlação;

n = Número de dados utilizados no ajuste

Através de uma análise sucinta da tabela anterior, é possível retirar a seguinte informação:

- O único parâmetro que se mantém em todas as regressões criadas é a concentração de O_3 da hora anterior;
- A variável menos vezes presente é a humidade relativa;
- Observando os coeficientes de correlação, verifica-se que a divisão dos dados estudados por período nocturno e período diurno não acrescenta melhorias significativas aos modelos e em alguns casos piora-os (ex: 2º e 3º trimestre);
- Observando os coeficientes de correlação, verifica-se que a divisão dos dados estudados por períodos trimestrais não acrescenta qualquer melhoria aos modelos, à excepção do 3º trimestre ($R_{3º\ Trimestre}$=0,930 e R_{Anual}=0,929) mas não de forma significativa;

Como exemplo dos ajustes criados por MLR para a previsão para o dia seguinte, é apresentado na Tabela 2 as regressões obtidas para as horas 0, 6, 12 e 18 e a regressão obtida com todas as horas do ano de 2005.

As células assinaladas com "---" significam que o factor exploratório correspondente não tem significância para a regressão. Os valores apresentados são correspondentes à fase de treino.

Tabela 2 - Coeficiente de correlação, número de dados utilizados e parâmetros da regressão obtidos para os ajustes criados por MLR (previsão para o dia seguinte), relativos ao ano de 2005, para as horas 0, 6, 12 e 18 e o conjunto de todas.

Hora	Ajuste	P_0	CO	NO	NO_2	NO_2/NO	O_3	T	HR	WS	n	R
0	Global	23,48	---	-3,32	4,49	-2,61	11,19	---	3,74	---	261	0,45
	1º Trim.	17,30	---	---	---	---	5,25	---	---	---	73	0,32
	2º Trim.	35,29	---	---	---	---	7,66	---	---	---	65	0,33
	3º Trim.	23,31	---	---	---	---	---	---	---	---	70	---
	4º Trim	17,70	---	---	---	---	---	8,39	---	6,39	53	0,67
	1º Sem.	25,78	---	---	5,34	---	15,58	---	4,50	---	138	0,50
	2º Sem.	20,89	---	---	---	---	3,55	4,88	---	---	123	0,38
6	Global	25,01	---	---	---	---	4,72	---	-3,44	---	259	0,36
	1º Trim.	25,09	---	---	---	---	---	---	-6,09	---	70	0,33
	2º Trim.	28,83	---	---	---	5,60	---	---	-9,20	---	71	0,49
	3º Trim.	21,09	---	---	---	---	---	7,85	---	---	65	0,42
	4º Trim	24,60	---	---	---	---	7,14	---	---	---	53	0,34
	1º Sem.	26,97	---	---	---	4,37	---	---	-6,94	---	141	0,42
	2º Sem.	22,67	---	---	---	-4,64	7,90	---	---	---	118	0,38
12	Global	54,28	4,27	-6,36	4,57	-8,47	13,65	11,94	---	---	246	0,76
	1º Trim.	38,75	3,89	-6,05	7,24	---	13,31	-3,54	---	---	68	0,77
	2º Trim.	65,84	---	---	---	---	---	9,56	---	6,38	64	0,58
	3º Trim.	69,84	---	---	---	-12,83	13,93	16,52	---	---	67	0,81
	4º Trim	38,83	14,91	-11,85	---	---	---	7,69	---	---	47	0,61
	1º Sem.	51,89	---	---	---	---	11,82	6,00	---	---	132	0,73
	2º Sem.	57,05	---	---	6,90	-9,15	11,84	19,47	---	---	114	0,80
18	Global	37,68	---	---	---	---	13,80	7,97	---	---	265	0,73
	1º Trim.	19,24	---	-8,98	15,12	---	16,21	-4,49	4,01	---	74	0,67
	2º Trim.	58,22	---	---	---	---	---	---	-5,06	4,13	69	0,37
	3º Trim.	53,21	---	---	---	---	---	14,48	---	---	70	0,63
	4º Trim	15,75	---	---	---	8,93	---	---	---	---	52	0,43
	1º Sem.	38,05	---	---	---	---	10,32	8,95	---	4,91	143	0,75
	2º Sem.	37,25	---	---	---	---	7,85	15,03	---	---	122	0,76
Todas	Global	34,61	---	-1,56	2,89	-1,99	15,76	5,29	0,92	1,83	6112	0,69
	1º Trim.	25,47	---	---	---	---	13,24	1,02	---	-2,60	1702	0,64
	2º Trim.	46,78	---	-2,77	3,51	-1,81	9,65	3,79	---	5,95	1590	0,62
	3º Trim.	39,94	-1,45	---	---	-1,59	9,35	16,44	3,49	2,94	1563	0,75
	4º Trim	24,97	2,02	---	---	---	13,39	5,78	3,88	---	1257	0,59
	1º Sem.	35,77	---	-2,66	2,55	-1,40	13,65	5,65	1,15	1,46	3292	0,69
	2º Sem.	33,27	---	---	2,56	-2,03	14,75	8,45	1,53	3,04	2820	0,70

Da análise Tabela 2, é possível realizar uma análise semelhante à anterior:

- Nenhum parâmetro se mantém em todas as regressões quando esta se baseia apenas no estudo de uma hora, embora continue a ser a concentração de O_3 a variável exploratória dominante;
- Não é possível definir um parâmetro notoriamente com menor significância, mas neste caso o parâmetro menos frequente nos modelos criados para o ano de 2005, é o relativo à concentração de CO;
- Observando os coeficientes de correlação, verifica-se que a divisão dos dados estudados por períodos trimestrais não acrescenta de forma significativa uma melhoria de previsão com a utilização destes modelos;
- Observando os coeficientes de correlação, verifica-se que a divisão dos dados estudados por períodos semestrais também não acrescenta de forma significativa uma melhoria aos modelos;

Fazendo uma comparação entre os coeficientes de correlação obtidos para os modelos de previsão para a hora seguinte (Tabela 1) e os obtidos para o dia seguinte à mesma hora (Tabela 2), verifica-se que são bastante mais elevados para o primeiro caso. Portanto, a regressão obtida para este modelo ajusta melhor os pontos experimentais estudados. No entanto, uma hora de antecedência de previsão não tem qualquer efeito prático.

Os índices de desempenho do modelo MRL relativos à fase de teste, para os anos 2003, 2004 e 2005 encontram-se nas Tabelas 3, 4 e 5, respectivamente.

Entende-se como modelos horários os modelos criados para uma hora específica ("hora a hora") e como modelos globais os modelos criados com todas as horas, para o dia seguinte.

Tabela 3 - Índices de desempenho do modelo MRL, do ano 2003

	Modelo Horário			*Modelo Global*		
	Ano	Semestre	Trimestre	Ano	Semestre	Trimestre
R:	0,67	r<0	0,50	0,62	0,61	0,62
MIN:	0	0	0	0	0	0
MAX:	155	155	155	155	155	155
AMP:	155	155	155	155	155	155
MBE:	-1,11	3,67	-4,29	-0,14	-0,38	-1,81
MAE:	15,01	20,51	17,39	16,09	16,15	16,08
RMSE:	19,68	27,28	22,94	20,80	21,00	20,69
D2:	0,81	0,76	0,75	0,78	0,79	0,78

* r<0 significa que o coeficiente de correlação é negativo.

Tabela 4 - Índices de desempenho do modelo MRL, do ano 2004

	Modelo Horário			*Modelo Global*		
	Ano	Semestre	Trimestre	Ano	Semestre	Trimestre
R:	0,68	0,40	0,36	0,66	0,67	0,57
MIN:	0	0	0	0	0	0
MAX:	198	198	198	198	198	198
AMP:	198	198	198	198	198	198
MBE:	0,40	5,05	-2,01	-0,05	-0,46	-1,63
MAE:	14,91	17,13	17,83	15,31	15,25	16,23
RMSE:	19,25	24,05	24,44	19,64	19,50	21,45
D2:	0,81	0,80	0,73	0,79	0,81	0,77

Tabela 5 - Índices de desempenho do modelo MRL, do ano 2005

	Modelo Horário			*Modelo Global*		
	Ano	Semestre	Trimestre	Ano	Semestre	Trimestre
R:	0,55	r<0	0,51	0,54	0,54	0,55
MIN:	0	0	0	0	0	0
MAX:	185	185	185	185	185	185
AMP:	185	185	185	185	185	185
MBE:	-1,85	2,29	-3,74	-1,21	-1,90	-5,16
MAE:	17,48	20,69	18,13	17,65	17,68	17,31
RMSE:	22,47	28,12	23,06	22,60	22,66	22,41
D2:	0,73	0,70	0,72	0,71	0,72	0,74

Mais uma vez se observa que a criação de modelos para períodos semestrais e trimestrais não introduz melhorias na previsão de O_3, tanto nos modelos criados para cada hora como nos modelos para todas as horas. Em algumas situações os modelos semestrais sobreavaliam as concentrações de O_3, no entanto os desvios médios são mais elevados.

A título de exemplo dos modelos "hora a hora" criados por MRL, apresenta-se nas Tabelas 6 e 7 os resultados obtidos para as horas 0, 6, 12 e 18, do ano 2005. Resultados relativos à fase de teste.

Tabela 6 - Índices de desempenho do modelo MRL criado para a hora 0 e hora 6 do ano de 2005

	Modelo Hora 0			*Modelo Hora 6*		
	Ano	Semestre	Trimestre	Ano	Semestre	Trimestre
R:	0,30	0,26	r<0	0,40	r<0	0,21
MIN:	0	0	0	0	0	0
MAX:	76	76	76	90	90	90
AMP:	76	76	76	90	90	90
MBE:	-3,03	-2,06	-4,15	-2,44	-2,28	-5,05
MAE:	18,40	18,29	19,33	16,61	17,86	16,12
RMSE:	22,25	22,47	23,53	20,58	22,65	22,01
D2:	0,50	0,61	0,44	0,47	0,43	0,53

Tabela 7 - Índices de desempenho do modelo MRL criado para a hora 12 e hora 18 do ano de 2005

	Modelo Hora 12			*Modelo Hora 18*		
	Ano	Semestre	Trimestre	Ano	Semestre	Trimestre
R:	r<0	r<0	0,45	0,39	r<0	0,49
MIN:	12	12	12	2	2	2
MAX:	151	151	151	170	170	170
AMP:	139	139	139	168	168	168
MBE:	0,74	1,34	-3,19	-0,81	15,39	-5,37
MAE:	17,04	21,89	15,76	20,07	30,07	18,25
RMSE:	24,43	30,57	21,25	27,25	45,55	25,74
D2:	0,37	0,46	0,42	0,61	0,61	0,67

Nota-se que para as horas indicadas do ano 2005, os modelos semestrais e trimestrais apresentam piores desempenhos. Existem algumas excepções como é o caso das horas 12 e 18 para os modelos trimestrais. Apresentam menores desvios médios mas os valores encontram-se sob estimados.

4.3.2 Modelos do Tipo Caixa Cinzenta (GM)

Neste tipo de modelo foi aplicado utilizando duas técnicas distintas: i) Modelos do Tipo Caixa Cinzenta Unicomponente, GM(1,1) e ii) Modelos do Tipo Caixa Cinzenta Multicomponente, GM(1,h).

Relativamente à técnica do tipo i) GM(1,1), criou-se um exemplo teórico de modo a testar a fiabilidade do modelo desenvolvido. Uma lista de valores foi gerada no sentido de englobar as mesmas duas componentes previstas no modelo: componente tendencial e componente periódica.

Os valores criados, (x), foram calculados através da seguinte equação:

$$x(t) = 2.3 + 2.1 * \mathrm{t} + \cos\left(2 * \mathrm{t} * \frac{\pi}{3}\right), \mathrm{t} = 0,1,2,...,32 \qquad (42)$$

Como para este caso foi estipulado que a cada 10 dados consecutivos era efectuada uma previsão foi possível obter 23 previsões de x.

Os resultados obtidos encontram-se representados nas Figuras 1 e 2.

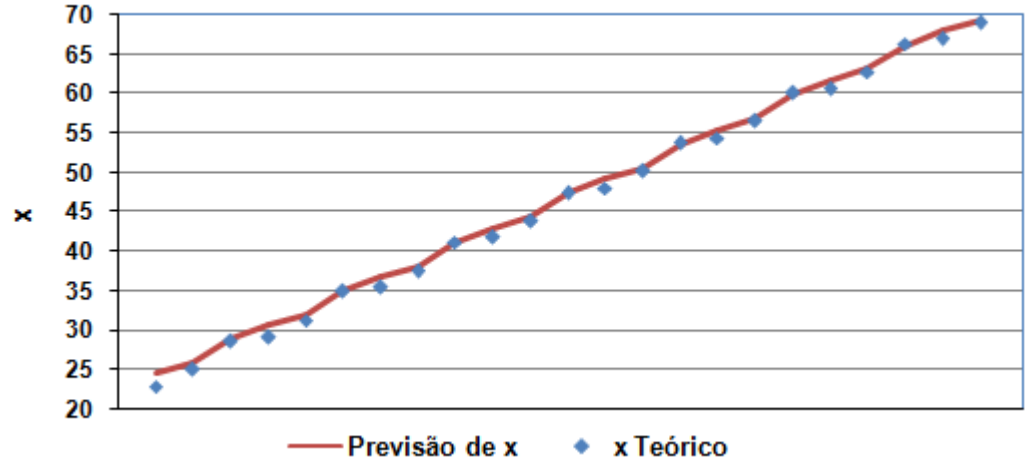

Figura 1 Valores de x teóricos e dos valores previstos pelo modelo GM(1,1)

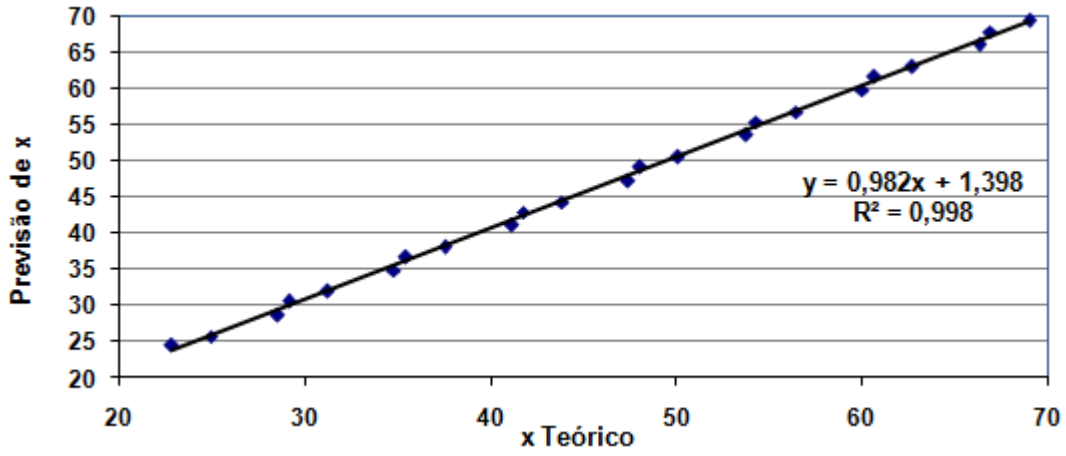

Figura 2 Valores de x previstos pelo modelo GM(1,1) em função dos valores teóricos

As Figuras 1 e 2 mostram que o modelo criado ajusta os valores pretendidos com R^2 = 0,998, um declive próximo de 1 e uma ordenada na origem próxima de 0.

Quanto à sua aplicação ao caso de estudo - previsão das concentrações de O_3 - o modelo não tem o mesmo desempenho. Na Figura 3 é apresentado, a título de exemplo, os resultados obtidos pelo modelo para a hora 23, entre os dias 17 e 28 de Setembro de 2005. São resultados que correspondem à fase de teste.

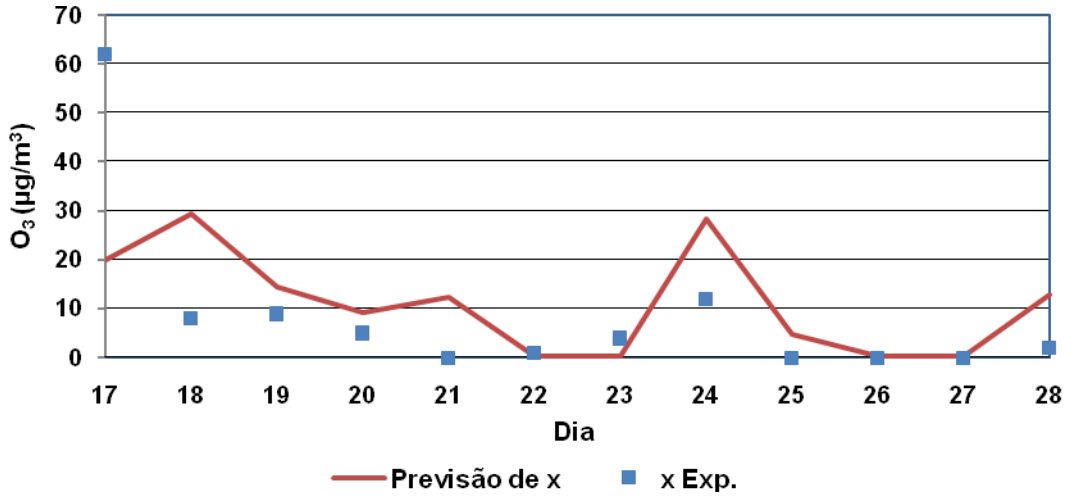

Figura 3 Valores da concentração de O_3 medidos e valores previstos pelo modelo GM(1,1) para a hora 23, entre os dias 17 e 28 de Setembro de 2005.

A título de exemplo dos modelos "hora a hora" criados por GM(1,1), apresenta-se nas Tabelas 8 e 9 os resultados obtidos para as horas 0, 3, 6, 9, 12, 15 e 18, do ano 2005. Resultados relativos à fase de teste.

Tabela 8 - Índices de desempenho do modelo GM(1,1), previsão para o dia seguinte à mesma hora, criado para as horas 0, 3, 6 e 9 do ano de 2005

	Ano 2005			
	Hora 0	Hora 3	Hora 6	Hora 9
R:	r<0	r<0	r<0	r<0
MIN:	0	0	0	4
MAX:	76	89	80	79
AMP:	76	89	80	75
MBE:	-3,52	-2,54	0,22	2,36
MAE:	20,32	24,97	26,76	19,27
RMSE:	26,17	31,53	32,24	25,14
D2:	0,42	0,35	0,32	0,53

Tabela 9 - Índices de desempenho do modelo GM(1,1), previsão para o dia seguinte à mesma hora, criado para as horas 12, 15, 18 e 21 do ano de 2005

	Ano 2005			
	Hora 12	Hora 15	Hora 18	Hora 21
R:	r<0	r<0	r<0	r<0
MIN:	38	15	3	0
MAX:	151	89	106	82
AMP:	113	74	103	82
MBE:	8,08	-2,95	5,94	-0,14
MAE:	29,11	20,86	21,52	17,06
RMSE:	37,71	25,54	27,80	22,63
D2:	0,13	0,20	0,60	0,65

Com o objectivo de obter melhores resultados quanto ao desempenho deste tipo de modelos, foi implementado uma versão com várias variáveis exploratórias - Modelos do Tipo Caixa Cinzenta Multicomponente GM(1,h).

Tal como no caso anterior, para a técnica do tipo GM(1,h), criou-se um exemplo teórico de modo a testar o modelo desenvolvido. Criou-se uma lista de valores gerados aleatoriamente, onde a uma das colunas foi atribuído o significado de variável resposta e às restantes o significado de variáveis de extrapolação. O modelo GM(1,h) foi então aplicado.

A Figura 4 apresenta o resultado do ajuste.

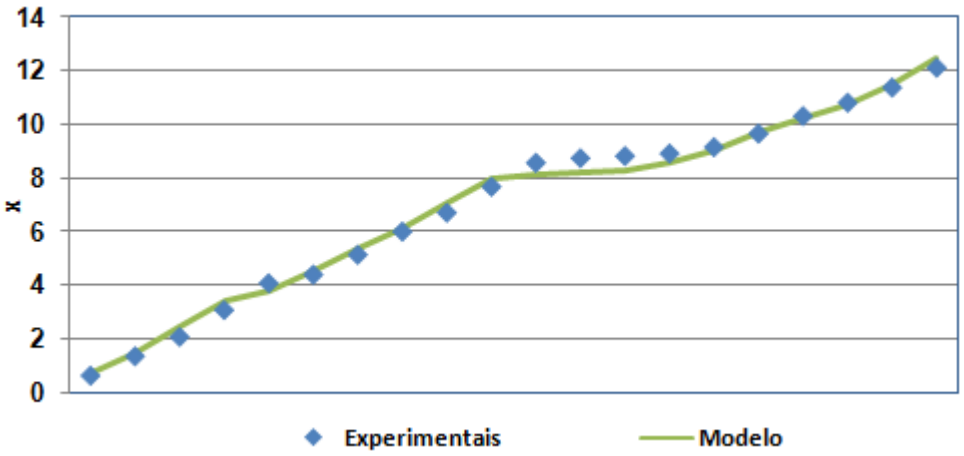

Figura 4 Ajuste efectuado com o modelo GM(1,h) ao problema teórico criado.

É possível observar que o modelo criado ajusta bem os pontos experimentais. Há uma melhoria na parte final do ajuste o que favorece a futura previsão.

A título de exemplo dos modelos "hora a hora" criados por GM(1,h), apresenta-se nas Tabelas 10 e 11 os resultados obtidos para as horas 0, 3, 6, 9, 12, 15 e 18, do ano 2005. Resultados relativos à fase de teste.

Tabela 10 - Índices de desempenho do modelo GM(1,h), previsão para o dia seguinte à mesma hora, criado para as horas 0, 3, 6 e 9 do ano de 2005

	Ano 2005			
	Hora 0	Hora 3	Hora 6	Hora 9
R:	r<0	r<0	r<0	r<0
MIN:	0	1	0	2
MAX:	76	89	75	89
AMP:	76	88	75	87
MBE:	6,79	14,10	16,45	10,18
MAE:	28,35	30,80	29,30	23,52
RMSE:	36,89	37,00	34,70	28,84
D2:	0,43	0,56	0,40	0,67

Tabela 11 - Índices de desempenho do modelo GM(1,h), previsão para o dia seguinte à mesma hora, criado para as horas 12, 15, 18 e 21 do ano de 2005

	Ano 2005			
	Hora 12	Hora 15	Hora 18	Hora 21
R:	r<0	r<0	r<0	r<0
MIN:	49	17	3	0
MAX:	65	143	78	75
AMP:	16	126	75	75
MBE:	45,24	38,65	7,55	6,58
MAE:	45,24	43,26	19,94	22,10
RMSE:	47,50	56,40	30,86	30,54
D2:	0,18	0,42	0,51	0,55

Através da observação das tabelas anteriores é possível verificar que os modelos criados por GM(1,h), para o dia seguinte, não constituem modelos adequados para a previsão das concentrações de O_3 para o dia seguinte. Apresentam maus resultados nos seus coeficientes de correlação e valores elevados de desvios médios.

O desempenho do mesmo modelo também foi testado para a previsão da hora seguinte. Tal como no caso do modelo MRL. A previsão para a hora seguinte obteve melhores índices de desempenho. Os resultados deste método são apresentados na Tabela 12.

Tabela 12 - Índices de desempenho do modelo GM(1,h), previsão para a hora seguinte, correspondente aos anos de 2003, 2004 e 2005

	2003	*2004*	*2005*
R:	0,43	0,57	0,53
MIN:	0	0	0
MAX:	219	204	219
AMP:	219	204	219
MBE:	-13,09	-10,18	-11,22
MAE:	18,23	15,00	16,52
RMSE:	25,47	21,34	22,40
D2:	0,83	0,85	0,85

Melhores índices de desempenho, como valores mais elevados dos coeficientes de correlação e índices de concordância e valores mais baixos nos desvios médios comprovam que o GM(1,h) obteve um melhor comportamento para a previsão da hora seguinte. No entanto, os resultados demonstram através da observação dos MBE que as previsões foram, em média, sub-estimadas.

4.3.3 Aprendizagem Modificada baseada em Exemplos (MLFE)

Também para este modelo foram estudadas as duas abordagens já enunciadas anteriormente – previsão para a hora seguinte e previsão para o dia seguinte à mesma hora. Para o último caso criaram-se modelos específicos para cada hora e outro global para todas as horas juntas.

Este tipo de modelos necessita da introdução tolerância (e_f) admitida e do valor de sobreposição (w). Foram testadas várias conjugações destas duas variáveis não se encontrando a conjugação óptima aplicável a todos os casos. Portanto, os resultados abaixo apresentados foram obtidos com w e e_f óptimos para cada situação.

De seguida serão apresentados os índices de desempenho dos modelos criados correspondentes à fase de teste.

Tabela 13 - Índices de desempenho do modelo MLFE, previsão para a hora seguinte, correspondente aos anos de 2003, 2004 e 2005

	2003 *W=10 e_f=5*	*2004* *W=10 e_f=5*	*2005* *W=15 e_f=2*
R:	0,84	0,87	0,64
MIN:	18	2	31
MAX:	57	45	72
AMP:	39	43	41
MBE:	-0,75	-0,01	-6,53
MAE:	5,48	4,02	8,64
RMSE:	6,37	5,84	10,31
D2:	0,91	0,91	0,85

Tabela 14 - Índices de desempenho do modelo MLFE, previsão para o dia seguinte, correspondente aos anos de 2003, 2004 e 2005

	2003 *W=15 e_f=2*	*2004* *W=15 e_f=2*	*2005* *W=15 e_f=5*
R:	0,49	0,52	0,43
MIN:	0	2	0
MAX:	58	45	70
AMP:	58	43	70
MBE:	4,69	1,56	2,41
MAE:	12,27	8,01	13,77
RMSE:	15,62	10,02	19,04
D2:	0,70	0,63	0,64

Para este modelo verifica-se, através da observação das tabelas anteriores, que as previsões efectuadas para a hora seguinte obtiveram melhores índices de desempenho do que as previsões efectuadas para o dia seguinte. Os valores de MBE são a única excepção observada.

A título de exemplo dos modelos "hora a hora" criados por MLFE, apresenta-se nas Tabelas 15 e 16 os resultados obtidos para as horas 0, 3, 6, 9, 12, 15 e 18, do ano 2005. Resultados relativos à fase de teste.

Tabela 15 - Índices de desempenho do modelo MLFE, previsão para o dia seguinte à mesma hora, criado para as horas 0, 3, 6 e 9 do ano de 2005

	Ano 2005			
	Hora 0 *W=15 e_f=2*	Hora 3 *W=15 e_f=2*	Hora 6 *W=15 e_f=2*	Hora 9 *W=15 e_f=2*
R:	0,13	r<0	r<0	r<0
MIN:	2	0	1	2
MAX:	67	72	72	47
AMP:	65	72	71	45
MBE:	4,27	-24,92	-20,54	-6,02
MAE:	16,50	26,47	21,91	12,98
RMSE:	19,23	31,94	27,40	17,40
D2:	0,48	0,44	0,50	0,36

Tabela 16 - Índices de desempenho do modelo MLFE, previsão para o dia seguinte à mesma hora, criado para as horas 12, 15, 18 e 21 do ano de 2005

	Ano 2005			
	Hora 12 *W=15 e_f=2*	Hora 15 *W=15 e_f=2*	Hora 18 *W=15 e_f=2*	Hora 21 *W=15 e_f=2*
R:	r<0	r<0	r<0	r<0
MIN:	12	3	1	2
MAX:	63	64	54	71
AMP:	51	61	53	69
MBE:	-12,72	4,39	-7,58	-10,60
MAE:	16,77	13,47	15,65	18,21
RMSE:	20,72	16,24	22,02	23,13
D2:	0,42	0,33	0,44	0,48

Para uma melhor comparação entre os diferentes modelos são apresentados os valores das previsões efectuadas pelos dois modelos onde se obtiveram melhores índices de desempenho na previsão do dia seguinte – MLR e MLFE. O período apresentado é do dia 16 de Dezembro até ao dia 30 do mesmo mês, à hora 23. A escolha deste período deve-se ao facto de corresponder à fase de teste para os dois modelos analisados.

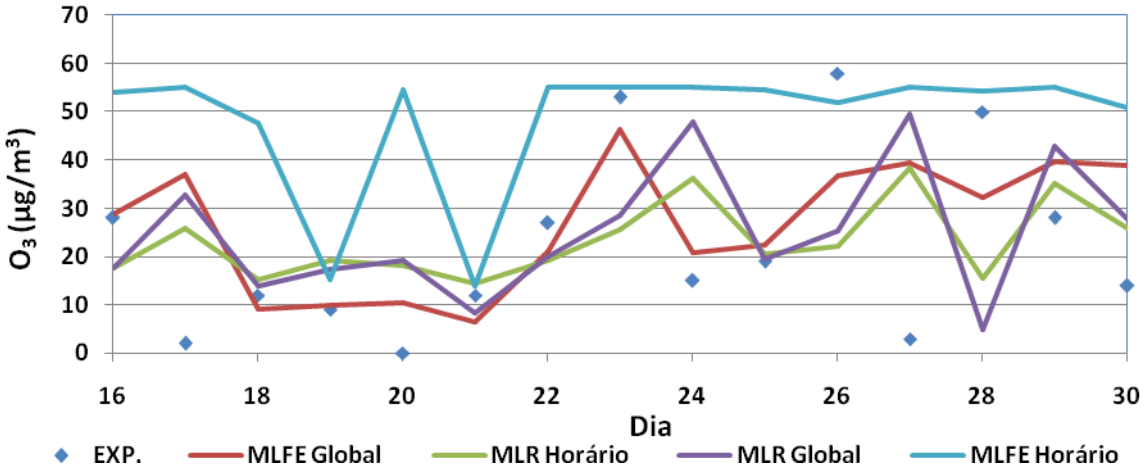

Figura 5 Ajustes dos modelos criados por MLFE e MLR para o período de 16 a 30 de Dezembro, à hora 23, do ano de 2003.

Tabela 17 - Índices de desempenho dos modelos MLFE e MLR, previsão para o dia seguinte à mesma hora, para o período de 16 a 30 de Dezembro, à hora 23, do ano de 2003

	Ano 2003			
	MLFE Global	MLFE Horário	MLR Horário	MLR Global
R:	0,33	r<0	r<0	r<0
MIN:	0	0	0	0
MAX:	58	58	58	58
AMP:	58	58	58	58
MBE:	4,60	26,47	1,29	3,02
MAE:	12,60	27,29	16,72	19,48
RMSE:	17,03	32,94	20,51	24,39
D2:	0,69	0,48	0,23	0,26

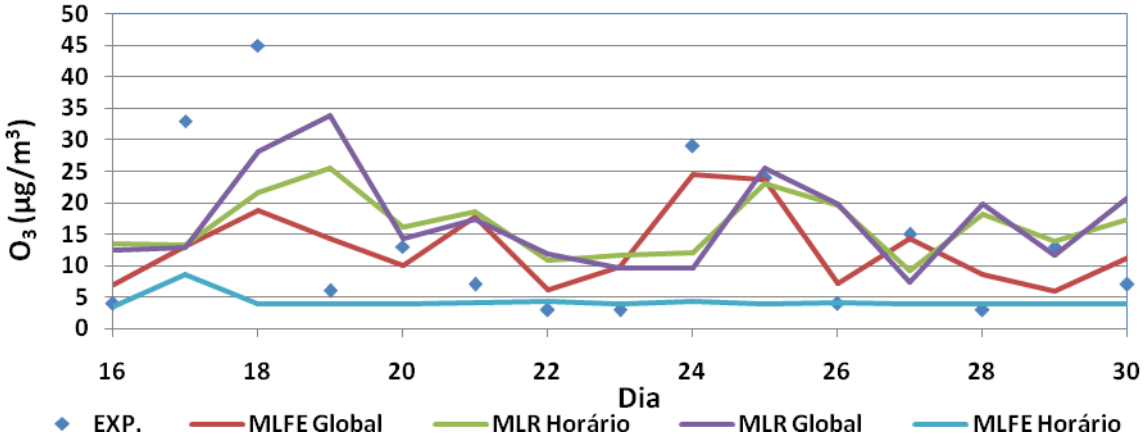

Figura 6 Ajustes dos modelos criados por MLFE e MLR para o período de 16 a 30 de Dezembro, à hora 23, do ano de 2004.

Tabela 18 - Índices de desempenho dos modelos MLFE e MLR, previsão para o dia seguinte à mesma hora, para o período de 16 a 30 de Dezembro, à hora 23, do ano de 2004

	Ano 2004			
	MLFE Global	MLFE Horário	MLR Horário	MLR Global
R:	0,62	r<0	r<0	r<0
MIN:	3	3	3	3
MAX:	45	45	45	45
AMP:	42	42	42	42
MBE:	-1,12	-9,59	2,38	3,05
MAE:	7,11	10,05	11,29	11,78
RMSE:	9,89	15,48	13,16	13,97
D2:	0,68	0,46	0,38	0,47

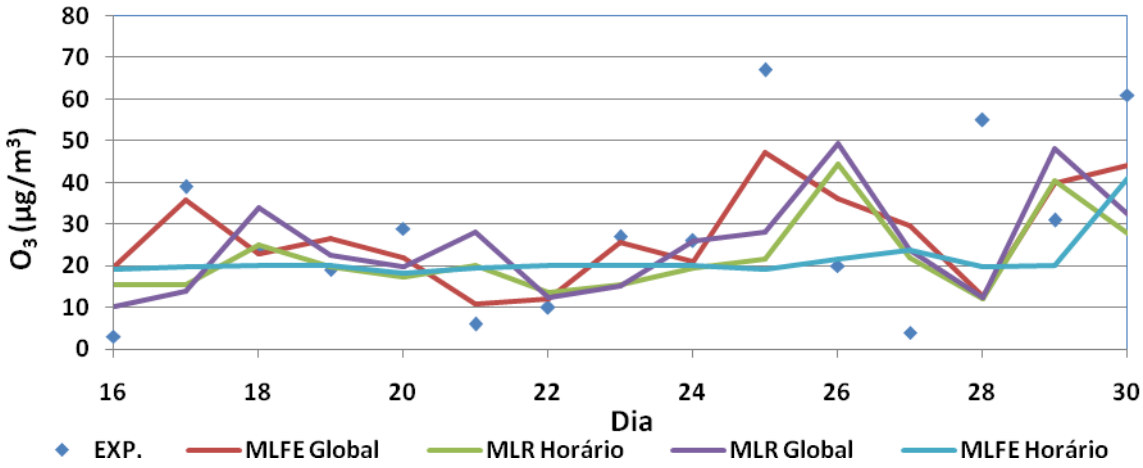

Figura 7 Ajustes dos modelos criados por MLFE e MLR para o período de 16 a 30 de Dezembro, à hora 23, do ano de 2005.

Tabela 19 - Índices de desempenho dos modelos MLFE e MLR, previsão para o dia seguinte à mesma hora, para o período de 16 a 30 de Dezembro, à hora 23, do ano de 2005

	Ano 2005			
	MLFE Global	MLFE Horário	MLR Horário	MLR Global
R:	0,55	0,12	r<0	r<0
MIN:	3	3	3	3
MAX:	67	67	67	67
AMP:	64	64	64	64
MBE:	-1,00	-6,57	-6,05	-2,97
MAE:	11,89	14,86	17,25	17,82
RMSE:	16,16	19,27	21,98	21,89
D2:	0,70	0,48	0,37	0,38

Através da análise das figuras e tabelas anteriores, é possível retirar a seguinte informação:

- O modelo que melhor ajusta os dados é o MLFE criado com todas as horas.
- Os resultados obtidos por MLR para os modelos "hora a hora" e pelos modelos globais foram bastante semelhantes. No entanto, os modelos globais obtiveram menores desvios médios.
- Relativamente aos modelos "hora a hora" e globais por MLFE as diferenças foram maiores. As previsões pelos modelos globais aproximaram-se mais dos valores experimentais do que pelos modelos "hora a hora".
- As previsões "hora a hora" criadas por MLFE praticamente mantiveram um valor constante. Isto deve-se ao menor número de dados disponíveis para criar o modelo.
- De uma maneira geral, os modelos criados com todas as horas obtiveram melhores índices de desempenho do que os criados especificamente para cada hora.

5 Conclusões

Modelos MLR, GM e MLFE foram aplicados para a previsão das concentrações de ozono para a hora seguinte e para o dia seguinte.

Diferentes estruturas de dados foram testadas nos modelos por MLR, concluindo-se que as divisões dos dados anuais por períodos semestrais e trimestrais e divisões por períodos nocturnos e diurnos não introduziam melhorias significativas nas previsões das concentrações de O_3.

A implementação de modelos "hora a hora" (específicos para cada hora) foi também estudada. De uma maneira geral, as melhores previsões foram obtidas através de modelos globais (criados com todas as horas). No entanto, sugere-se que novos estudos sejam feitos no sentido de aperfeiçoar a técnica de "hora a hora", principalmente na quantidade de dados fornecidos para a criação dos modelos.

As previsões para a hora seguinte obtiveram melhores índices de desempenho do que as previsões efectuadas para o dia seguinte à mesma hora. No entanto, com vista a futuros alertas em caso de concentrações elevadas de O_3, as previsões efectuadas para o dia seguinte teriam maiores efeitos práticos.

Relativamente às previsões para a hora seguinte, nada se pode dizer relativamente aos modelos criados por MLR, pois não foi definido um conjunto de valores de teste. Para este modelo deu-se mais relevo ao estudo das diferentes estruturas de dados. Comparando os restantes modelos, o MLFE foi o que obteve melhores resultados.

O modelo que obteve melhores índices de desempenho para a previsão do dia seguinte à mesma hora foi o MLFE e o modelo com piores resultados foi o GM.

Concluindo, modelos de aprendizagem modificada baseada em exemplos, criados a partir de todas as horas, são os modelos mais eficazes na previsão das concentrações de O_3 para a mesma hora do dia seguinte.

6 Avaliação do trabalho realizado

6.1 Objectivos Realizados

O objectivo deste trabalho consistia em desenvolver modelos estatísticos capazes de, em tempo útil, poderem prever concentrações de ozono no ar ambiente.

Ao longo deste trabalho foram desenvolvidos três modelos distintos: a regressão linear múltipla (MLR), modelos do tipo caixa cinzenta (GM) e a uma técnica de aprendizagem modificada baseada em exemplos (MLFE). Para cada modelo foram aplicadas diferentes técnicas de abordagem ao problema, como diferentes estruturas de dados e construções de modelos específicos para cada hora.

Foram obtidos resultados com algum grau de precisão para os modelos do tipo MLFE, para a previsão do dia seguinte à mesma hora. Desta forma os objectivos propostos foram cumpridos.

6.2 Limitações e Trabalho Futuro

A aplicação de MLFE a este tipo de problemas mostrou-se eficaz. No entanto, sugere-se a utilização de um maior conjunto de dados para a construção dos modelos, especialmente para os modelos desenvolvidos para uma hora específica.

A utilização de dois dias anteriores, para efectuar a previsão, em vez de apenas um como foi tratado nesta tese, poderá introduzir melhorias à previsão.

Sugere-se também o estudo de novos modelos estatísticos aplicados à previsão das concentrações de ozono.

6.3 Apreciação final

A criação de uma ferramenta capaz de efectuar a previsão das concentrações de ozono de forma eficiente e em tempo útil revela-se cada vez mais importante. Esta permitiria lançar alertas à população e reduzir o número de locais afectados. Na opinião do autor, atendendo aos resultados obtidos o trabalho desenvolvido forneceu um bom contributo para o alcance desse objectivo.

Referências

- Allen, D. Air pollution (online). *Kirk-Othmer Encyclopedia of Chemical Technology*. Volume 1, 787-815 (2002). Data de acesso: 23 Junho 2008.

  <http://www.mrw.interscience.wiley.com/emrw/9780471238966/kirk/article/airwolf.a01/current/pdf>.

- Arroyo-Lopez, P.E., Jaramillo-Osorio, A., Gaytan-Iniestra, J., Wojcik-Rojek, A.R. Using transfer function methodology for ozone forecasting in Toluca city. *Proceedings of Second International Conference on Urban Air Quality,* Madrid, 130–131 (1999).

- Ballester, E.B., Camps i Valls, G., Carrasco-Rodríguez, J.L., Soria-Olivas, E., Valle-Tascon, S. Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks. *Ecological modelling*, **156**, 27–41 (2002).

- Borja-Aburto, V.H., Loomis, D.P., Bangdiwala, S., Shy, C.M., Rascon-Pacheco, R.A. Ozone, suspended particulates, and daily mortality in Mexico City. *American Journal of Epidemiology*, **145**, 258–268 (1997).

- Chaloulakou, A., Saisana, M., Spyrellis, N. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Science of the Total Environment*, **313**, 1-13 (2003).

- Chen, J.-L., Islam, S., Biswas, P. Nonlinear dynamics of hourly ozone concentrations: nonparametric short term prediction. *Atmospheric Environment*, **32**, 1839–1848 (1998).

- Gardner, M.W., Dorling, S.R. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment,* **34**, 21-34 (2000)

- Guo, R. A Grey Modeling of Covariate Information in Repairable System. *Proceedings of the 4th International Conference on Quality and Reliability*, Beijing, China, 667-674 (2005).

- Heo, J.-S., Kim, D.-S. A new method of ozone forecasting using fuzzy expert and neural network systems. *Science of the Total Environment,* **325**, 221– 237 (2004).

- Hoek, G., Schwartz, J.D., Groot, B., Eilers, P. Effects of ambient particulate matter and ozone on daily mortality in Rotterdam. *Archives of Environmental Health*, **52**, 455–463 (1997).
- Hubbard, M.C., Cobourn, W.G. Development of a regression model to forecast ground-level ozone concentration in Louisville, KY. *Atmospheric Environment*, **32**, 2637–2647 (1998).
- Kelsall, J.E., Samet, J.M., Zeger, S.L., Xu, J. Air pollution and mortality in Philadelphia, 1974–1988. *American Journal of Epidemiology* **146**, 750–762 (1997).
- Kocak, K., Saylan, L., Sen, O. Nonlinear time series prediction of O3 concentration in Istanbul. *Atmospheric Environment*, **34**, 1267–1271 (2000).
- Lee, J.T., Shin, D., Chung, Y. Air pollution and daily mortality in Seoul and Ulsan, Korea. *Environmental Health Perspectives*, **107**, 149–154 (1999).
- Lin, Y.H., Lee, P.C. Novel high-precision grey forecasting model. *Automation in Construction*, **16**, 771-777 (2007)
- Milionis, A.E., Davies, T.D. Regression and stochastic models for air pollution – I. Review, comments and suggestions. *Atmospheric Environment*, **28**, 2801–2810 (1994).
- Mintz, R., Young, B.R., Svrcek, W.Y. Fuzzy logic modeling of surface ozone concentrations. *Computers & Chemical Engineering*, **29**, 2049–2059 (2005).
- Ozone. *Encyclopædia Britannica Online (*2005). Data de acesso: 23 Junho 2008 <http://search.eb.com/eb/article-9057878>.
- Pires, J.C.M., Martins, F.G., Sousa, S.I.V., Alvim-Ferraz, M.C.M., Pereira M.C. Selection and validation of parameters in multiple linear and principal component regressions. *Environment Modelling & Software,* **23**, 50-55 (2008).
- Pollution. *Encyclopædia Britannica Online* (2008). Data de acesso: 23 Junho 2008 <http://search.eb.com/eb/article-9109632>.
- Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., Greig, A., Foxall, R., Eben, K., Chatterton, T., Vondracek, J., Richter, M., Dostal, M., Bertucco, L., Kolehmainen, M., Doyle, M. A rigorous inter-comparison of ground-level ozone predictions. *Atmospheric Environment*, **37**, 3237–3253 (2003).
- Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G., Pelikan, E. Statistical models to assess the health effects and to forecast ground-level ozone. *Environmental Modelling & Software,* **21**, 547–558 (2006).

- Schwartz, J. Harvesting and long term exposure effects in the relation between air pollution and mortality. *American Journal of Epidemiology*, **151**, 440–451 (2000).
- Soja, G., Soja, A.-M. Ozone indices based on simple meteorological parameters: potentials and limitations of regression and neural network models. *Atmospheric Environment*, **33**, 4299–4307 (1999).
- Sousa, S.I.V., Martins, F.G. Pereira M.C., Alvim-Ferraz, M.C.M. Prediction of ozone concentrations in Oporto city with statistical approaches. *Chemosphere*, **64**, 1141-1149 (2006).
- Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M., Pereira M.C. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environment Modelling & Software*, **22**, 97-103 (2007).
- Takens, F. Detecting strange attractors in turbulence. In: Rand, D.A., Young, L.S. (Eds.), Dynamical systems and turbulence, lecture notes in mathematics, vol. 898. Springer, Berlin, Germany, pp. 366–381 (1981).
- Zmirou, D., Schwartz, J., Saez, M., Zanobetti, A., Wojtyniak, B., Touloumi, G., Spix, C., Ponce de Leon, A., LeMoullec, Y., Bacharova, L., Schouten, J., Ponka, A., Katsouyanni, K. Time-series analysis of air pollution and cause-specific mortality. *Epidemiology*, **9**, 495–503 (1998).