REDES NEURAIS ARTIFICIAIS: UMA APLICAÇÃO NO ESTUDO DA POLUIÇÃO ATMOSFÉRICA E SEUS EFEITOS ADVERSOS À SAÚDE

Emília Matos do NASCIMENTO[1]
Basilio de Bragança PEREIRA[1,2]
José Manoel de SEIXAS[3]

- RESUMO: Há uma enorme necessidade em buscar a associação entre condições climáticas e poluição atmosférica com mortalidade ou internações hospitalares por doenças respiratórias. Este trabalho propõe a utilização das redes neurais como metodologia alternativa para avaliar tal associação. Os dados utilizados referem-se ao número de internações hospitalares na cidade de Paris, por bronquiolite infantil, entre 1997 e 2000. Os modelos neurais foram avaliados em termos de descrição dos dados e da sua capacidade de generalização. Os melhores resultados foram obtidos através do pré-processamento dos dados, com remoção de ciclos e uso de filtro de média móvel. Um estudo de relevância das variáveis explicativas foi também desenvolvido. Os resultados obtidos foram compatíveis aos encontrados através dos modelos aditivos generalizados, apontando o material particulado (PM_{10}) como principal responsável no número de internações hospitalares.
- PALAVRAS-CHAVE: Redes neurais artificiais; poluição atmosférica; doenças respiratórias.

1 Introdução

A cada dia aumenta a preocupação do homem com a degradação ambiental, que se expressa através de mudanças climáticas, poluição do ar, contaminação das águas e do solo, com conseqüências desastrosas para a fauna e a flora, causando doenças e afetando negativamente a qualidade da vida humana (Bates et al., 2008).

A poluição do ar tem sido apontada como um dos principais responsáveis por doenças relacionadas ao sistema respiratório, especialmente em crianças, idosos e pessoas com problemas respiratórios. A exposição aos poluentes pode provocar ou agravar doenças tais como asma, bronquite crônica, enfisema pulmonar, infecções pulmonares, rinofaringites e irritação nas vias respiratórias, entre outras, conduzindo a um elevado número de internações hospitalares, óbitos e aumentando a procura de atendimento nas salas de emergência (Singer et al., 2002; Šrám et al., 2005).

[1] Programa de Engenharia de Produção, COPPE/UFRJ, Caixa Postal: 68507, CEP. 21941-972 - Rio de Janeiro - RJ – Brasil. E-mail: *emilia@pep.ufrj.br*
[2] Faculdade de Medicina, HUCFF/UFRJ, Universidade Federal do Rio de Janeiro – UFRJ, Caixa Postal 68507, CEP: 21941-972, Rio de Janeiro, RJ, Brasil. E-mail: *basílio@hucff.ufrj.br*
[3] Laboratório de Processamento de Sinais, Programa de Engenharia Elétrica, COPPE/Poli-UFRJ, Caixa Postal: 68504, CEP. 21941-972, Rio de Janeiro, RJ, Brasil. E-mail: *seixas@lps.ufrj.br*

Vários episódios têm sido registrados em grandes centros urbanos em virtude das altas concentrações de poluentes na atmosfera, com trágicas conseqüências para a população (Braga et al., 2002; Conceição et al., 2002). A preocupação causada pelos efeitos decorrentes da degradação do ar tem sido motivo de discussão entre líderes de diversos países, levando-os a firmarem acordos para controle e redução da emissão de gases de efeito estufa (UNFCCC, 1992; UNFCCC, 2007).

A poluição atmosférica é um problema de saúde pública, constituindo um grande desafio à gestão. A compreensão dos seus efeitos pode contribuir para o planejamento e nortear a adoção de medidas públicas visando proteger a população. O conhecimento científico é de grande utilidade no auxílio à gestão da informação relativa à saúde coletiva, subsidiando a elaboração de medidas voltadas à prevenção e à redução da poluição do ar.

Este trabalho tem como objetivo propor a utilização das redes neurais artificiais (Haykin, 2008; Pereira e Rodrigues, 1998), como metodologia alternativa, no estudo da poluição atmosférica e seus efeitos adversos à saúde. As redes neurais artificiais foram utilizadas, para reproduzir a análise de Willems et al. (2007), que utilizaram os modelos aditivos generalizados (Hastie e Tibshirani, 1990) para estimar os efeitos causados pela poluição atmosférica e condições climáticas, tendo em vista o número de internações hospitalares na cidade de Paris, França, motivadas por bronquiolite infantil, doença respiratória causada frequentemente pelo vírus *syncethial respiratory virus* (RSV). O contato com este vírus causa, normalmente, um resfriado. Mas em crianças e, em algumas circunstâncias, especialmente no início do inverno, o vírus pode ser responsável por uma grave doença respiratória, conduzindo a um elevado número de consultas e internações hospitalares (Everard et al., 1994; Farhat et al., 2002; Gutiérrez et al., 2003).

A base de informação utilizada no presente estudo é apresentada na seção 2. A etapa de normalização e o pré-processamento dos dados para o projeto da rede neural são descritos na seção 3. A modelagem neural e a metodologia utilizada para identificação das variáveis relevantes são apresentadas na seção 4. Finalmente, na seção 5, são apresentados os resultados obtidos e, em seguida, as respectivas conclusões.

2 Base de dados

Este estudo foi desenvolvido sobre a mesma base de dados utilizada por WILLEMS et al. (2007), obtida no ERBUS (*Epidémiologie et Recueil des Bronchiolites en Urgence pour Surveillance*), referente a 43 hospitais localizados em Paris, no período compreendido entre 15 de outubro a 15 de janeiro dos anos de 1997 a 2000, além de dados meteorológicos e de poluição atmosférica medidos pelas estações de AIRPARIF - *Surveillance de la Qualité de l'Air en Ile-de-France*, nos anos de 1997 a 2001. O procedimento, adotado pelos referidos autores para evitar problemas com dados faltantes, foi utilizado também neste trabalho. Assim, foram considerados, ao todo, 419 dias de observação dos 34 hospitais, cujos dados encontravam-se completos. A base de informação é composta por 17 variáveis explicativas (12 variáveis climáticas e 5 de poluição atmosférica). A série histórica do número de internações hospitalares, motivadas por bronquiolite infantil, representa o alvo (variável dependente). A Figura 1 mostra essa série, indicando claramente a existência de variações sazonais. Quatro períodos podem ser observados, cada um dos quais representa um período incluindo o inverno europeu.

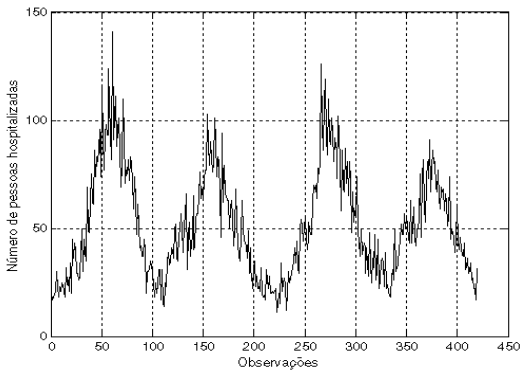


Figura 1 - Número de internações hospitalares.

3 Normalização e pré-processamento dos dados

Os modelos neurais foram projetados com a finalidade de se obter uma descrição dos dados ou uma generalização para o conjunto total dos dados. Nos modelos projetados para descrição, a avaliação foi feita em termos de interpretação dos dados, tendo sido utilizada toda a informação disponível no seu desenvolvimento. Os modelos que visam à generalização são restritos a um conjunto de desenvolvimento (treino), sendo a avaliação da generalização do modelo obtida através do teste de desempenho no conjunto restante de dados. Assim, nestes modelos, a base de dados foi subdividida em dois conjuntos: treinamento e teste. O conjunto de teste recebeu um terço das observações, tomadas de três em três, a partir da terceira, num total de 139 observações, cabendo ao conjunto de treinamento as 280 observações restantes, referentes aos índices 1, 2, 4, 5 da base de dados e assim, por diante.

O projeto das redes neurais artificiais consistiu em uma etapa inicial de normalização, que tem como finalidade adaptar os dados de entrada à faixa dinâmica das funções de ativação da rede neural, neste caso, tangente hiperbólica (na camada intermediária) e função linear (na camada de saída). O processo de normalização dos dados foi feito da seguinte forma:

$$x^* = \frac{x_0}{\overline{x} + n_s s} \tag{1}$$

onde x^* é a variável normalizada; x_0 é a variável original; $\overline{x}$ é a média obtida a partir da amostra que forma o conjunto de treinamento; s é o desvio-padrão da variável original, também obtido a partir do conjunto de treinamento; e n_s representa o número de desvios-padrão a considerar. O valor de n_s foi obtido empiricamente, respeitando a faixa dinâmica da rede neural, que deve conter o resultado encontrado no cálculo de x^*.

A normalização foi aplicada, inicialmente, ao conjunto de treinamento e, em seguida, estendida ao conjunto de teste, onde foi adotado o mesmo fator de normalização.

A característica sazonal das variáveis deve ser levada em consideração. Após a normalização, os dados devem ser dessazonalizados a fim de revelar o resíduo da série a ser modelada, permitindo assim, a detecção das características escondidas. Nelson et al. (1999) observaram, através da utilização de um modelo neural para previsão, que os resultados obtidos a partir de dados dessazonalizados são, significativamente, mais acurados do que os obtidos sem o pré-processamento. Calôba et al. (2002), Alekseev e Seixas (2008) procederam a dessazonalização dos dados através da remoção da tendência, seguida da retirada dos ciclos.

A tendência da série pode ser estimada pelo método dos mínimos quadrados, através do seguinte cálculo:

$$s_2[n] = s_1[n] - (a + b * n) \tag{2}$$

onde $s_1[n]$ é a série original normalizada; $s_2[n]$ é a série obtida após a eliminação da tendência; a é o intercepto; e b é o coeficiente angular da reta de regressão. Observando que a série alvo (número de internações hospitalares) não apresenta tendência (tal fato pode ser visualizado através da Figura 1), considerou-se desnecessária a retirada da tendência.

Os eventuais efeitos das concentrações dos poluentes e de possíveis variáveis de confundimento não incidem, necessariamente, no mesmo dia em que foi observado o evento (internação ou óbito). Assim, é também comum o uso de modelos com defasagem ou médias móveis das variáveis meteorológicas e dos poluentes (Zanobetti et al., 2000 apud Lima et al., 2001). Willems et al. (2007) observaram que a utilização de um filtro de médias móveis de seis dias forneceu um efeito mais significativo em seu modelo do que o uso de um filtro de médias móveis de dois dias, por exemplo. O filtro de médias móveis foi utilizado, também, neste trabalho.

Levando-se em consideração a característica sazonal das variáveis, avaliou-se a necessidade do pré-processamento dos dados. Esta avaliação foi feita após a implementação de diversas redes neurais, tendo, como critério, o maior poder de generalização do modelo. A generalização está associada à capacidade de predição de novos casos, indicando se o modelo neural conseguiu extrair as características principais da informação.

Os melhores resultados foram obtidos por um modelo no qual o pré-processamento dos dados foi feito através da retirada dos ciclos, utilizando a análise de Fourier, seguida do uso de um filtro de média móvel de seis dias.

A análise de Fourier (Cooley e Tukey, 1965) permite que sejam modelados os componentes que refletem os ciclos presentes na série. A retirada dos ciclos identificados pela análise de Fourier possibilita que a rede neural seja alimentada com a série residual, livre dos componentes já modelados. A remoção dos componentes senoidais de uma série é feita mediante a aplicação da transformada de Fourier e posterior observação das frequências que apresentem um pico significativo de amplitude, relativamente às demais frequências que compõem o espectro da série. Desta forma, identificados os coeficientes dos senos e co-senos que compõem a informação da frequência em questão, procede-se à retirada do pico correspondente.

Alekseev e Seixas (2008) removeram os ciclos observados na variável alvo do conjunto de treinamento, e estenderam esse procedimento às variáveis de entrada, quando a mesma frequência foi identificada. Calôba et al. (2002) utilizaram o valor obtido no

cálculo da raiz quadrada da média dos quadrados dos resíduos (RMSE) para decidir se um componente senoidal deveria, ou não, ser retirado. Desta forma, caso a remoção do ciclo provocasse um aumento no valor da RMSE, o componente senoidal não seria retirado.

Neste trabalho, os ciclos identificados na série alvo de treino foram retirados, deterministicamente, por senóides. Os mesmos ciclos, porventura presentes nas variáveis explicativas do conjunto de treinamento, foram também removidos. Após a retirada dos componentes senoidais das variáveis de entrada, uma nova análise espectral foi feita para verificar a existência de ciclos remanescentes. Caso tais ciclos fossem encontrados, seriam também removidos. De forma análoga ao procedimento adotado por Calôba et al. (2002), a decisão pela remoção, ou não, de um determinado componente senoidal ficou condicionada ao valor da RMSE. Assim, quando o aumento da RMSE foi observado, o componente senoidal não foi retirado. Esse procedimento foi estendido aos dados pertencentes ao conjunto de teste, retirando-se os mesmos componentes identificados no conjunto de treinamento.

4 Modelagem neural e análise de relevância

A implementação dos modelos foi feita através de redes neurais completamente conectadas, modelo MLP – *Multilayer Perceptron* (Rumelhart et al., 1986), com uma camada intermediária e sem realimentação. Os neurônios na camada escondida foram do tipo tangente hiperbólica. Na camada de saída, foi utilizado um único neurônio, com função de ativação linear. Os pesos da rede neural podem ser atualizados de duas formas: (1) modo batelada, onde a atualização dos pesos se dá quando todos os pares (entrada e saída) do conjunto de treinamento forem apresentados; e (2) modo instantâneo, onde a atualização ocorre cada vez que uma amostra do conjunto de treinamento for apresentada. Neste trabalho, o treinamento foi feito, no modo batelada, através do algoritmo *Resilient Backpropagation*, desenvolvido por Riedmiller e Braun (1993) e Riedmiller (1994), que tem a vantagem de convergir mais rapidamente.

A Figura 2 mostra a arquitetura do modelo neural, composto de 17 variáveis de entrada (12 relacionadas aos fatores climáticos e 5 referentes às concentrações de poluentes) e uma variável de saída (número de internações hospitalares).

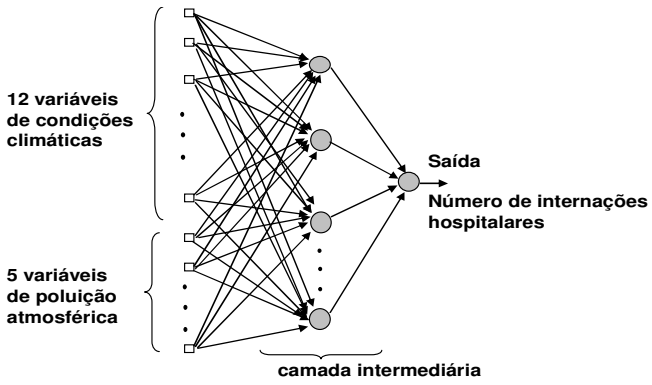


Figura 2 - Arquitetura do modelo neural.

Como medida de desempenho no processo de treinamento da rede neural, adotou-se o MSE (erro médio quadrático) cometido à saída da rede, em relação ao valor observado no alvo de treinamento. A função objetivo teve, como finalidade, a minimização do MSE no conjunto de treinamento.

O conjunto de teste não participa da fase de treinamento da rede neural, sendo usado para avaliar o poder de generalização da rede neural. O MAPE (erro percentual absoluto médio), calculado no conjunto de teste, foi adotado, tanto como critério de parada do treinamento, quanto para avaliar o poder de generalização do modelo. Esta medida é definida como:

$$mape = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

onde y é o valor observado (variável resposta) e $\hat{y}$ é a saída da rede.

Após o treinamento do modelo neural, procedeu-se à recomposição dos alvos (número de internações hospitalares), adicionando-se, à saída obtida, os valores retirados na fase do pré-processamento (remoção de ciclos e sazonalidades) e multiplicando-se o resultado obtido pelo fator de normalização. O poder de generalização da rede foi medido no conjunto de teste através do cálculo do MAPE, avaliado sobre os alvos originais e os alvos recompostos.

A seleção das variáveis explicativas a serem utilizadas na rede neural deve ser feita de forma criteriosa, de modo a se obter um modelo mais rico em informações. Essa escolha é importante, uma vez que a presença de variáveis não relevantes pode afetar o treinamento da rede, comprometendo o seu poder de generalização.

Seixas et al. (1996) apresentaram uma proposta para identificação das variáveis explicativas mais relevantes, servindo como método para seleção das variáveis de entrada do modelo neural. O estudo da relevância de uma variável x_j estabelece uma comparação entre a saída obtida pela rede neural treinada (modelo final) e a resposta da rede neural obtida ao se manter fixa, esta mesma variável x_j, no seu valor médio, calculado sobre as amostras que compõem o conjunto de treino. Esse cálculo é feito para cada variável, segundo a equação:

$$R_j = \frac{1}{Npat}\sum_{i=1}^{Npat}\left[y(x_i, w) - \left(y(x_i, w) \mid x_{j,\,i} = \bar{x}_j\right)\right]^2 \tag{4}$$

onde *Npat* é o número de padrões e $(y(x_i, w))$ corresponde à saída da rede. Quanto maior o valor obtido no cálculo da estatística Rj, maior será a relevância da variável.

A etapa seguinte do desenvolvimento do modelo consiste no treinamento da rede neural utilizando, como informação de entrada, apenas as variáveis indicadas como relevantes na etapa anterior e identificando, após o treinamento, o novo conjunto de variáveis que exerce maior influência na variável resposta. Com isto, pretende-se obter um modelo eficiente e mais compacto.

Neste trabalho, a análise de relevância foi feita adotando-se um limiar correspondente a, aproximadamente, 10% do maior valor encontrado na estatística Rj. Assim, foram consideradas relevantes as variáveis que tivessem apresentado Rj superior a este ponto de corte.

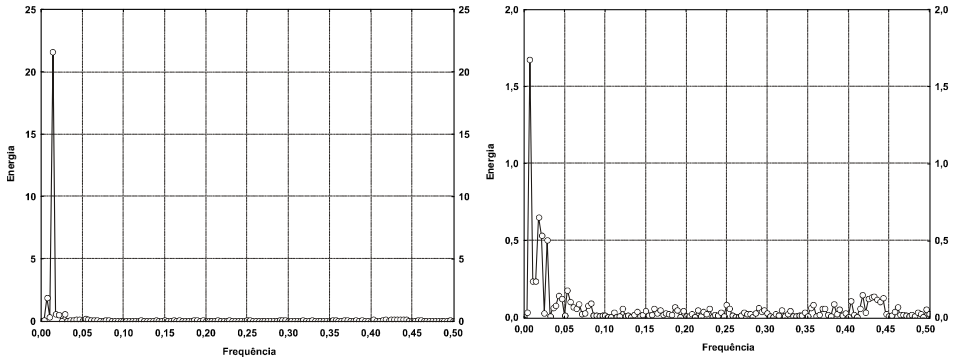
5 Resultados

A análise espectral da variável correspondente ao número de internações hospitalares (alvo de treino) foi realizada sobre a série normalizada. A Tabela 1 apresenta as frequências (normalizadas) mais relevantes da série alvo de treino, a energia correspondente e o valor resultante do cálculo da RMSE após a retirada do componente senoidal. Como se pode observar, a frequência de maior importância, na primeira análise espectral, foi 0,0143. Para avaliar a necessidade da retirada de algum componente senoidal remanescente, fez-se uma segunda análise espectral da série. As frequências mais relevantes foram 0,0071, 0,0179 e 0,0214. Os ciclos foram removidos mediante a observação do valor obtido no cálculo da RMSE. Notou-se também a presença de um componente senoidal com frequência 0,0286. Entretanto, a tentativa de remoção deste componente provocou o aumento do valor da RMSE (0,6889), motivo pelo qual este componente não foi retirado. Após a retirada dos componentes senoidais, observou-se a redução no valor da energia, inicialmente 21,6135, passando a 0,3539 ao final do processo.

Tabela 1 - Série das internações hospitalares – análise espectral

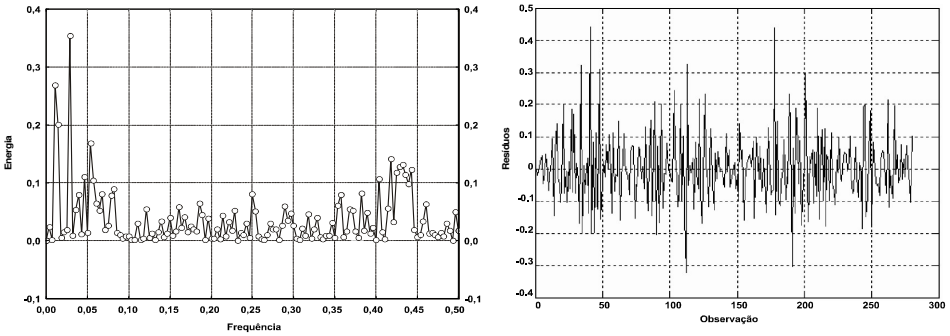
	Frequência	Energia	RMSE
1ª análise espectral	0,0143	21,6135	0,6949
	0,0071	1,6703	0,6896
2ª análise espectral	0,0179	0,6511	0,6882
	0,0214	0,3539	0,6870

A Figura 3 apresenta o espectro resultante da primeira análise e a Figura 4 mostra o espectro após a segunda análise. A Figura 4 mostra, ainda, a série de resíduos, que irá alimentar o modelo neural.



(a) (b)

Figura 3 - Série das internações hospitalares – análise espectral: espectro original (a) e resultante após a primeira análise (b).

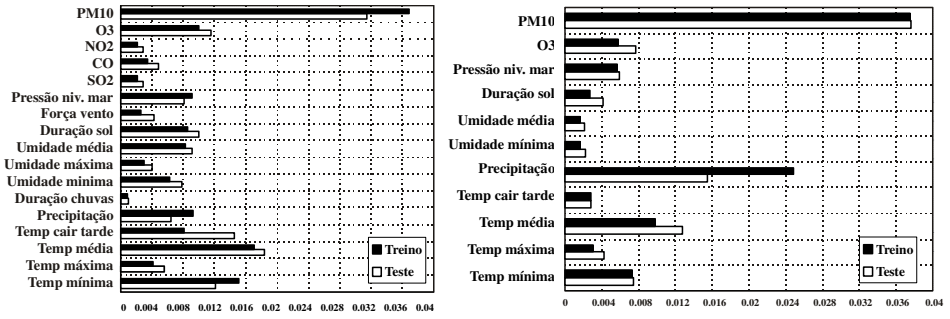


(a) (b)

Figura 4 - Série das internações hospitalares após a retirada dos componentes senoidais: espectro resultante (a) e série de resíduos (b).

Após as etapas de normalização e de pré-processamento, deu-se início à modelagem neural. A rede neural foi implementada sobre os resíduos obtidos após as etapas de normalização e pré-processamento dos dados, consistindo na retirada dos ciclos e, em seguida, das médias móveis de seis dias. O número de neurônios que compuseram a camada intermediária foi determinado de acordo com o menor valor calculado para o MAPE no conjunto de teste, após o pré-processamento dos dados. Foram desenvolvidos modelos neurais, com números distintos de neurônios na camada escondida, número esse que varia de 2 a 10 neurônios. O menor MAPE observado no conjunto de teste foi de 0,138. Assim, o modelo neural foi implementado com a topologia 17-7-1, representando 17 variáveis de entrada (12 relacionadas aos fatores climáticos e 5 referentes às concentrações de poluentes), 7 neurônios na camada intermediária e uma variável de saída (número de internações hospitalares).

Após a implementação do modelo neural, o critério Rj de relevância foi utilizado para reduzir a dimensionalidade da informação relativa aos fatores climáticos e de poluição atmosférica, indicando a informação mais relevante. A Figura 5(a) mostra o resultado da análise de relevância aplicada à saída do primeiro modelo neural implementado, onde de um total de 17 variáveis explicativas, 11 foram selecionadas.



(a) (b)

Figura 5 - Análise de relevância: (a) modelo completo e (b) variáveis selecionadas após a implementação do modelo completo.

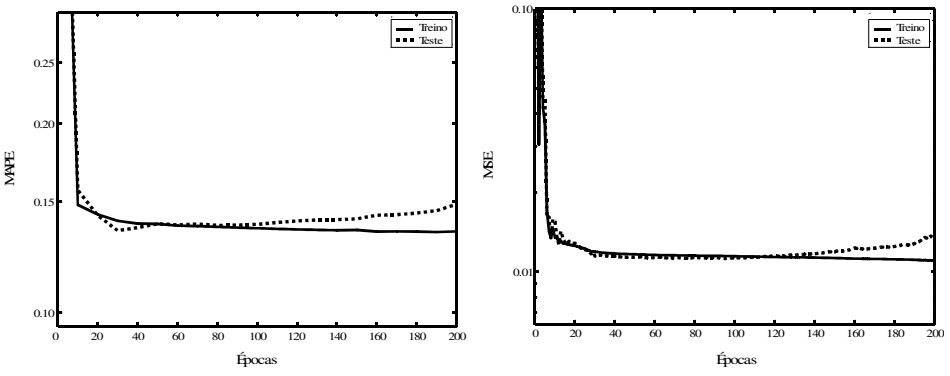
Uma nova redução de dimensionalidade foi obtida partindo-se desse modelo, cuja base de entrada era composta por essas 11 variáveis selecionadas, implementado com a topologia 11-5-1 (11 neurônios referentes às variáveis selecionadas pelo critério Rj; 5 neurônios na camada intermediária; e um neurônio na saída). Após uma nova aplicação do critério Rj sobre o modelo neural desenvolvido, 6 variáveis foram consideradas relevantes, sendo apresentadas na Figura 5(b): Material particulado com diâmetro inferior a 10μm (PM_{10}); Emissão de O_3; Pressão ao nível do mar; Precipitação diária; Temperatura média; e Temperatura mínima do dia. O modelo neural foi, então, projetado com a estrutura 6 – 4 – 1, representando 6 variáveis de entrada, 4 neurônios na camada intermediária e 1 na saída.

A Tabela 2 apresenta as medidas de desempenho dos modelos neurais implementados com suas respectivas topologias, onde se observa uma boa capacidade de generalização, com MAPE avaliado em torno de 0,13 no conjunto de teste, e uma boa qualidade no ajuste, com o MSE próximo de 1% no conjunto de treinamento.

Tabela 2 - Desempenho dos modelos neurais após o treinamento

Topologia	MSE (Treino)	MSE (Teste)	MAPE (Treino)	MAPE (Teste)
17 - 7 - 1	0,0087	0,0120	0,1355	0,1380
11 - 5 - 1	0,0105	0,0121	0,1342	0,1396
6 - 4 -1	0,0110	0,0112	0,1364	0,1377

A Figura 6 mostra o processo de aprendizagem durante o treinamento desta última rede neural. Observa-se a ocorrência do *overtraining* (Haykin, 2008) nas proximidades da época 100, que foi evitado através da parada prematura do treinamento na época 96.



(a) (b)

Figura 6 - Medidas de desempenho do modelo neural: MAPE (a) e MSE (b).

O estudo de relevância das variáveis explicativas deste modelo apontou o PM_{10} e a variável climática precipitação diária, como os principais responsáveis pelo número de internações nos hospitais de Paris, motivadas por bronquiolite infantil. A Figura 7 apresenta o resultado desta análise, mostrando a relevância das variáveis, tanto no conjunto de treinamento quanto no de teste.

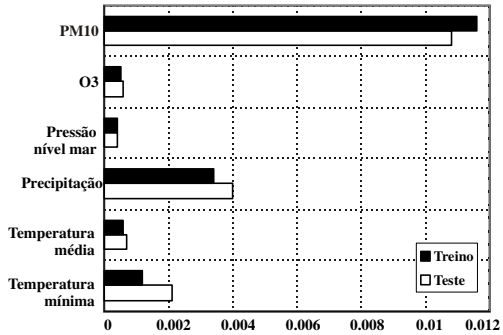


Figura 7 - Análise de relevância do modelo compacto.

Com a finalidade de desenvolver modelos neurais com a maior informação relevante possível, algumas redes neurais, com diferentes topologias, foram projetadas. Nesses novos modelos, foi adotado o mesmo tipo de pré-processamento que forneceu o menor valor calculado para o MAPE no conjunto de teste: retirada dos ciclos e, em seguida, a remoção das médias móveis de seis dias. As variáveis explicativas que compõem cada um desses modelos são as seguintes:

- Modelo 1: Somente variáveis climáticas;
- Modelo 2: Variáveis relevantes obtidas pelo critério Rj, para o Modelo 1;
- Modelo 3: Variáveis relevantes obtidas pelo critério Rj, para o Modelo 1, acrescentando-se os poluentes;
- Modelo 4: Variáveis relevantes obtidas pelo critério Rj, para o Modelo 1, acrescentando-se o PM_{10}, como única variável de poluição;
- Modelo 5: Variáveis climáticas selecionadas por Willems et al. (2007), através da análise de componentes principais: temperatura mínima do dia, precipitação diária, umidade média do dia, força do vento e pressão ao nível do mar;
- Modelo 6: Variáveis climáticas consideradas no Modelo 5, adicionando-se as variáveis relacionadas à poluição atmosférica;
- Modelo 7: Variáveis climáticas consideradas no Modelo 5, juntando-se o PM_{10}.

O estudo de Willems et al. (2007) apontou o PM_{10} como único responsável pelo número de internações hospitalares, conforme mencionado anteriormente, motivo pelo qual o poluente foi incluído como única variável de poluição nos modelos 4 e 7.

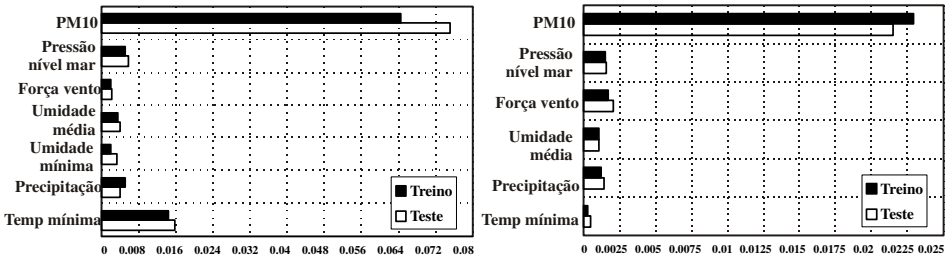
Após, cada implementação, o poder de generalização das redes neurais foi avaliado e a análise de relevância, segundo a estatística Rj, foi desenvolvida.

Nessas implementações, destacaram-se duas redes neurais (modelos 4 e 7), que diferem apenas pela inclusão da variável explicativa umidade mínima do dia na entrada de dados do modelo 4. A Tabela 3 apresenta as medidas de desempenho de ambos os modelos neurais, com suas respectivas topologias, onde se pode observar um bom poder de generalização e uma boa qualidade do ajuste, com valores próximos de 0,14 e 0,01, associados, respectivamente, ao MAPE (avaliado no conjunto de teste) e ao MSE (medido no conjunto de treinamento).

Tabela 3 - Desempenho das redes neurais para os modelos nos quais as bases de entrada diferem apenas pela presença da variável umidade mínima do dia

Modelo	Topologia	MSE (Treino)	MSE (Teste)	MAPE (Treino)	MAPE (Teste)
4	7 - 7 - 1	0,0103	0,0120	0,1434	0,1490
7	6 - 5 -1	0,0105	0,0115	0,1391	0,1478

A Figura 8 mostra a análise de relevância após o treinamento das referidas redes neurais, apontando o PM_{10} como variável mais relevante em ambos os modelos. Observa-se que quando a variável correspondente à umidade mínima do dia está presente, a rede neural identifica duas variáveis relevantes: PM_{10} e Temperatura mínima do dia. Observa-se, ainda, que na ausência da referida variável, o modelo neural aponta o PM_{10} como única variável relevante.



(a) (b)

Figura 8 - Análise de relevância: (a) variáveis climáticas obtidas pelo critério Rj, acrescentando-se o PM_{10} e (b) variáveis climáticas selecionadas pela análise de componentes principais, adicionando-se o PM_{10}.

Nos ensaios clínicos, há muitas fontes de variação que causam impacto no desfecho. As eventuais variações, que não puderem ser identificadas e controladas adequadamente, poderão interferir no efeito do tratamento em questão (Chow e Liu, 1995 apud Chow e Liu, 2004). No estudo da relação entre exposição e doença, um verdadeiro confundidor está associado à exposição em questão, sendo, ao mesmo tempo, um fator determinante da doença. Ou seja, ele confunde a relação entre a exposição e a doença (Vandenbroucke, 2002).

A inclusão da variável umidade mínima revelou, portanto, uma relação entre o número de internações hospitalares e a temperatura mínima do dia. Neste caso, a umidade mínima do dia pode ser considerada como uma variável de "anti-confundimento" (ou de confundimento negativo).

Conclusões

Os resultados obtidos através das redes neurais foram compatíveis com os encontrados por Willems et al. (2007), através da utilização dos modelos aditivos generalizados, apontando o material particulado com diâmetro inferior a 10μm (PM_{10})

como principal responsável no número de internações infantis nos hospitais de Paris devido a bronquiolite.

A metodologia utilizada neste estudo não é dependente das doenças do sistema respiratório, nem fica restrita a apenas uma determinada região geográfica, podendo ser estendida para outras aplicações.

Agradecimentos

Ao Professor Mounir Mesbah da Universidade Pierre et Marie Curie (Paris – França) por ter fornecido a base de dados que serviu como alicerce para o desenvolvimento deste estudo e à FAPERJ, ao CNPq e à CAPES pelo apoio, viabilizando a realização deste projeto.

NASCIMENTO, E. M.; PEREIRA, B. de B.; SEIXAS, J. M. Artificial neural networks: an application in the study of air pollution and its adverse health effects. *Rev. Bras. Biom.*, São Paulo, v.27, n.1, p.37-50, 2009.

- *ABSTRACT: There is a great need to assess the association between weather and air pollution with mortality or hospital admissions due to respiratory diseases. This paper proposes neural networks as alternative methodology to evaluate that association. The data refer to the number of hospitalizations in the city of Paris due to infant bronchiolitis, between 1997 and 2000. The neural models were evaluated for data description and to measure their capacity of generalization. The best results were obtained through data pre-processing, with removal of cycles and use of a moving average filter. A relevance study of the explanatory variables was also carried out. The results were consistent to those found through generalized additive models pointing out the particulate matter (PM_{10}) as the main responsible for the number of hospital admissions.*
- *KEYWORDS: Artificial neural network; air pollution; respiratory diseases.*

Referências

ALEKSEEV, K. P. G.; SEIXAS, J. M. A multivariate neural forecasting modeling for air transport – preprocessed by decomposition: a Brazilian application, *J. Air Transp. Manag.*, Disponível em: <http://dx.doi.org>, p.1-5, 2008.

BATES, B. C. ET al.(Ed.). Climate change and water. Technical paper of the Intergovernmental Panel on Climate Change. IPCC Secretariat. Geneva, 2008. 210p. Disponível em: <http://www.ipcc.ch/pdf/technical-papers/climate-change-water-en.pdf>. Acesso em: 19 mar. 2009.

BRAGA, A.; PEREIRA, L. A. A.; SALDIVA, P. H. N. Poluição atmosférica e seus efeitos na saúde humana. In: SUSTENTABILIDADE NA GERAÇÃO E USO DE ENERGIA NO BRASIL: OS PRÓXIMOS VINTE ANOS. Campinas: UNICAMP, 2002.

CALÔBA, G. M.; CALÔBA, L. P.; SALIBY, E. Cooperação entre redes neurais artificiais e técnicas 'clássicas' para previsão de demanda de uma série de vendas de cerveja na Austrália. *Pesqui. Oper.*, Rio de Janeiro, v.22, n.3, p.345-358, 2002.

CHOW, S. C.; LIU, J. P. *Design and analysis of clinical trials*: concepts and methodologies, 2.ed. New York: Wiley, 2004.

CHOW, S. C.; LIU, J. P. *Statistical design and analysis in pharmaceutical science*. New York: Dekker, 1995.

CONCEIÇÃO, G. M. S.; SALDIVA, P. H. N.; SINGER, J. M. Associação entre séries de mortalidade / morbidade e concentrações de poluentes atmosféricos: uma estratégia para análise estatística. *Rev. Bras. Estat*., Rio de Janeiro, v.63, n.219, p.75-98, 2002.

COOLEY, J. W.; TUKEY, J. W. An algorithm for the machine calculation of complex Fourier series. *Math. Comput*. v.19, p.297-301, 1965.

EVERARD, M. L. et al. Analysis of cells obtained by bronchial lavage of infants with respiratory syncytial virus infection. *Arch. Dis. Childhood*, v.71, p.428-432, 1994.

FARHAT, C. K.; CINTRA, O. A. L.; TREGNAGHI, M. W. Vacinas e o trato respiratório – o que devemos saber? *J. Pediatr.*, Rio de Janeiro, v.78, supl.2, S195-S204, 2002.

GUTIÉRREZ, A. M. et al. Frecuencia de niños hospitalizados por el virus sincitial respiratorio en tres periodos invernales. *Rev. Mex. Pediatr.*, v.70, n.4, p.167-170, 2003.

HASTIE, T. J.; TIBSHIRANI, R. J. *Generalized additive models*. London: Chapman & Hall, 1990.

HAYKIN, S. *Neural networks and learning machines*. Prentice Hall, 2008.

LIMA, L. P.; ANDRÉ, C. D. S.; SINGER, J. M. Modelos aditivos generalizados: metodologia e prática. *Rev. Bras. Estat*., Rio de Janeiro, v.62, n.217, p.37-69, 2001.

NELSON, M. et al. Times series forecasting using neural networks: should the data be deseasonalized first? *J. Forecast*., v.18, p.359-367, 1999.

PEREIRA, B. B.; RODRIGUES, C. V. S. *Redes neurais em estatística*. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA - SINAPE, 13., 1998, Caxambu. *Resumo*...p.13-20.

RIEDMILLER, M. *RPROP – description and implementation details*. Technical Report. Universitat Karlsruhe, 1994. Disponível em: <http://citeseer.ist.psu.edu/riedmiller94rprop.html>. Acesso em 20 mar. 2009.

RIEDMILLER, M.; BRAUN, H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In: INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, 1993, San Francisco. *Proceedings*… p.586-591.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R.J. Learning internal representations by error propagation. In: RUMELHART, D. E.; McCLELLAND, J. L. (Ed.) *Parallel distributed processing*. Cambridge: MIT Press, v.1., chap.8, p.317-362. 1986.

SEIXAS, J. M.; CALÔBA, L. P.; DELPINO, I. *Relevance criteria for variance selection in classifier designs*. In: INTERNATIONAL CONFERENCE ON ENGINEERING APPLICATIONS OF NEURAL NETWORKS, 1996. *Proceedings*...p.451-454.

SINGER, J. M. et al. *Association between atmospheric pollution and mortality in São Paulo, Brazil: regression models and analysis strategy*. In: INTERNATIONAL

CONFERENCE ON STATISTICAL DATA ANALYSIS BASED ON THE L1 NORM AND RELATED METHODS, 4., 2002, Neuchâtel. *Proceedings...* p.429-450.

ŠRÁM R. J. et al. *Intrauterine growth retardation, low birth weight, prematurity and infant mortality*. In: OMS. Organização Mundial da Saúde. Special. Effects of air pollution on children's health and development: a review of the evidence. Bonn: European Centre for Environment and Health, 2005. p.14-27.

UNFCCC - United Nations Framework Convention on Climate Change. 1992. Disponível em: <http://unfccc.int/resource/docs/convkp/conveng.pdf>. Acesso em: 12 dez. 2008.

UNFCCC - United Nations Framework Convention on Climate Change. *Kyoto protocol reference manual on accounting of emissions and assigned amounts*. 2007. Disponível em: <http://unfccc.int/files/national_reports/accounting_reporting_and_review_under_the_kyoto_protocol/application/pdf/rm_final.pdf>. Acesso em: 12 dez. 2008.

VANDENBROUCKE, J. P.The history of confounding. Soz. - Und Präventivmed. *Soc. Prev. Med.*, v.47, n.4, p.216-224, 2002.

WILLEMS, S. ET al. *Longitudinal analysis of short-term bronchiolitis air pollution association using semiparametric models*. In: Advances in Statistical Methods for the Health Sciences, 2007. BALAKRISHNAN, N. et al. Boston: Springer, 2007. p.467-487.

ZANOBETTI, A. et a.. Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, New York, v.1, n.3, p.279-292, 2000.

Recebido em 15.12.2008.

Aprovado após revisão 06.04.2009.