Prediction of Hospital Admissions from Air Pollutant Data

Marcelo Laendle Junior[1], **Simone Andréa Pozza**[1], **Guilherme P. Coelho**[1]

[1]Faculdade de Tecnologia (FT)
Universidade Estadual de Campinas (UNICAMP)
Limeira – SP – Brasil

laendle1999@gmail.com, {simone.pozza, guilherme}@ft.unicamp.br

Abstract. *Several papers in the literature indicate that air pollutants are harmful to health. Thus, in this work we sought to verify the possibility of predicting, based on atmospheric pollutants concentration and 24 h in advance, the number of hospital admissions associated with respiratory diseases. For this, Extreme Learning Machines (ELMs) were used as predictors and experiments were made with data from the city of Campinas (SP). The results showed that, although the use of some specific pollutants leads to smaller prediction errors, the best results were still obtained using only the historical series of hospitalizations as input to the ELMs.*

Resumo. *Diversos trabalho na literatura indicam que poluentes atmosféricos são nocivos à saúde. Sendo assim, neste trabalho buscou-se verificar a possibilidade de prever, a partir de dados de concentração de poluentes atmosféricos e com 24 h de antecedência, o número de internações hospitalares associadas a doenças respiratórias. Para isso, foram utilizados Extreme Learning Machines (ELMs) como preditores e experimentos foram realizados com dados referentes à cidade de Campinas (SP). Os resultados mostraram que, apesar do uso de dados de alguns poluentes específicos levarem a menores erros de previsão, os melhores resultados ainda foram obtidos utilizando-se apenas a série histórica de internações hospitalares como entrada para as ELMs.*

1. Introdução

Os níveis de poluição atmosférica têm preocupado agências ambientais, a comunidade científica e profissionais de saúde, já que estima-se que 4,2 milhões de pessoas morram por ano devido a problemas respiratórios causados por exposição à poluição [WHO, 2016a]. Entende-se por poluição atmosférica a presença, na atmosfera, de uma ou mais substâncias ou partículas em concentração suficiente para causar danos a seres humanos, animais, plantas ou materiais [Braga et al., 2005]. Dentre os diferentes poluentes atmosféricos, o foco do presente trabalho será no estudo dos poluentes que estão relacionados a doenças respiratórias, tais como o Material Particulado (MP) e o Ozônio (O_3) [CETESB, 2019].

Material particulado é um termo genérico usado para classificar pequenas partículas que se mantêm suspensas na atmosfera. O MP pode ser constituído por fumaça, poeira ou quaisquer outras partículas, tanto sólidas quanto líquidas. Tais partículas são normalmente oriundas da queima de biomassa, emissão de veículos movidos a combustão interna e produção industrial, dentre outras [CETESB, 2019, Seinfeld and Pandis, 2016].

O MP pode ser classificado a partir do diâmetro aerodinâmico de suas partículas, sendo três as principais categorias: *Partículas Totais em Suspensão* (PTS), que correspondem às partículas com diâmetro aerodinâmico menor ou igual a 50 μm; *Partículas Inaláveis* (MP_{10}), que correspondem àquelas com diâmetro aerodinâmico menor ou igual a 10 μm; e as *Partículas Inaláveis Finas* ($MP_{2,5}$), que são aquelas com diâmetro menor ou igual a 2,5 μm [CETESB, 2019]. Como MP_{10} e $MP_{2,5}$ podem causar mais danos à saúde, uma vez que não ficam retidas nas vias aéreas superiores e podem chegar até ao trato respiratório inferior (sendo depositados nos alvéolos e penetrando até na corrente sanguínea, no caso de $MP_{2,5}$) [Seinfeld and Pandis, 2016], estes dois tipos de material particulado foram considerados neste trabalho.

Vários trabalhos na literatura mostram que altos níveis de concentração dos poluentes atmosféricos são prejudicais à saúde humana. Dentre eles, Anderson et al. [2012] verificaram que o MP está relacionado à inflamação dos brônquios, levando ao desenvolvimento de asma e de outras doenças relacionadas ao bloqueio do fluxo de ar. Já Godish [2004] indica que o O_3 causa uma série de efeitos adversos à saúde humana, tais como redução da função pulmonar e agravamento de doenças pré-existentes como asma. Além disso, altas concentrações de O_3 podem levar a um aumento no número de internações hospitalares.

Entre 1993 e 1995, na Cidade do México, Loomis et al. [1999] identificaram um crescimento de 5% na mortalidade infantil cerca de 5 dias após a exposição a altas concentrações de poluentes atmosféricos. Já na Califórnia, nove cidades foram estudadas e observou-se um aumento no número de mortes diárias associado ao aumento da concentração de $MP_{2,5}$ na atmosfera [Ostro et al., 2006]. Resultados similares foram observados para oito cidades canadenses, onde Burnett et al. [2000] identificaram uma correlação positiva e estatisticamente significativa entre variações na concentração de poluentes atmosféricos e nas taxas de mortalidade.

Diante das evidências do risco que poluentes atmosféricos trazem para a saúde humana, instituições de pesquisa e órgãos governamentais buscam monitorar constantemente a concentração destes poluentes. No estado de São Paulo, a Companhia Ambiental de Estado de São Paulo (CETESB) realiza este monitoramento em diversas cidades, disponibilizando os dados obtidos por estações automáticas, de forma horária, em um sistema on-line denominado QUALAR [QUALAR, 2019].

Com relação às internações hospitalares, o Sistema Único de Saúde (SUS) brasileiro disponibiliza dados diários. Tais informações podem ser obtidas através do sistema TABNET, desenvolvido pelo Departamento de Informática do SUS [DataSUS, 2019]. Tanto os dados de internações hospitalares quanto os dados de concentração de poluentes atmosféricos podem ser considerados séries temporais [Morettin and Toloi, 2006], uma vez que são avaliados periodicamente. Neste contexto, ferramentas clássicas de análise de séries temporais e técnicas de aprendizagem de máquina [Han et al., 2012] podem ser utilizadas para predição de valores futuros de tais séries, tendo sido amplamente empregadas na literatura [Barajas, 2018, Tadano et al., 2017, Zhan et al., 2017].

Buscando prever valores futuros de concentração de $MP_{2,5}$, Zhan et al. [2017] propuseram uma nova abordagem baseada em aprendizagem de máquina, denominada *Geographically-Weighted Gradient Boosting Machine* (GW-GBM). O algoritmo GW-

GBM foi aplicado a dados de diferentes regiões da China e seus resultados de previsão se mostraram superiores aos obtidos pelo algoritmo original em que foi baseado.

Já Bisht and Seeja [2018] propuseram uma estratégia de predição de poluição atmosférica baseada no algoritmo de aprendizado de máquina conhecido como *Extreme Learning Machine* (ELM) [Huang et al., 2006]. Esta estratégia foi aplicada na previsão, um dia à frente, de índices de qualidade do ar associados a cinco poluentes (MP_{10}, $MP_{2,5}$, NO_2, CO, O_3), e os resultados se mostraram melhores que os apresentados por sistemas de previsão da poluição do ar já existentes. Barajas [2018] também se baseou em ELMs para previsão de concentração de MP_{10} e $MP_{2,5}$, tendo proposto um sistema para previsão *on-line* capaz de se adaptar a mudanças no comportamento dos dados ao longo do tempo.

Tadano et al. [2017] fizeram um estudo comparativo do desempenho de diferentes Redes Neurais Artificiais (RNAs) [Haykin, 2008] na avaliação do impacto da poluição atmosférica na saúde. Mais especificamente, analisaram o impacto que a concentração de MP_{10} e variáveis meteorológicas (temperatura e umidade do ar) têm no número de internações hospitalares por doenças respiratórias ocorridas na cidade de Campinas (SP). Dentre as RNAs comparadas, as ELMs apresentaram os melhores resultados.

Em uma linha similar à adotada por Tadano et al. [2017], neste trabalho buscou-se verificar a possibilidade de se prever internações hospitalares, 24 h à frente, a partir da concentração diária de poluentes atmosféricos. No entanto, além de MP_{10} este trabalho também considerou outros poluentes atmosféricos que estão diretamente relacionados à ocorrência de doenças respiratórias ($MP_{2,5}$ e O_3). Para isso, foram utilizadas ELMs como preditores e analisados dados referentes à cidade de Campinas (SP).

Este artigo está organizado da seguinte forma. Na Seção 2 são discutidos os principais aspectos relacionados à fundamentação teórica das técnicas empregadas neste trabalho. Na Seção 3 os dados utilizados e a metodologia experimental adotada são descritos. Os resultados experimentais são apresentados e discutidos na Seção 4 e, por fim, as conclusões finais e possíveis trabalhos futuros são apresentados na Seção 5.

2. Fundamentação Teórica

Para o desenvolvimento deste trabalho tomou-se como base as *Extreme Learning Machines* (ELMs) [Huang et al., 2006], que são um tipo especial de redes neurais artificiais (RNAs) do tipo *feedforward*, densas e com uma única camada oculta. As ELMs foram escolhidas por notadamente apresentarem um bom desempenho na previsão de séries temporais, inclusive em aplicações como a desenvolvida aqui [Tadano et al., 2017].

As RNAs [Haykin, 2008, Gardner and Dorling, 1998] podem ser consideradas uma das estruturas mais tradicionais de aprendizado de máquina [Han et al., 2012]. Dentre os diversos tipos de RNAs existentes, as multicamadas do tipo *feedforward*, como a ilustrada na Figura 1, têm sido muito utilizadas. Estas redes, inspiradas em mecanismos do cérebro dos animais vertebrados, são formadas por unidades simples de processamento (os neurônios artificiais) interconectadas em camadas. Matematicamente, cada neurônio artificial multiplica as entradas recebidas por um certo *peso*, antes de agregá-las e aplicar a chamada *função de ativação* ao resultado, gerando sua saída. De maneira geral, é possível afirmar que este tipo de RNA realiza um mapeamento não-linear de suas entradas em suas saídas, sendo que este mapeamento é utilizado para resolver diversas tarefas, tais como estimação e classificação de dados [Haykin, 2008].

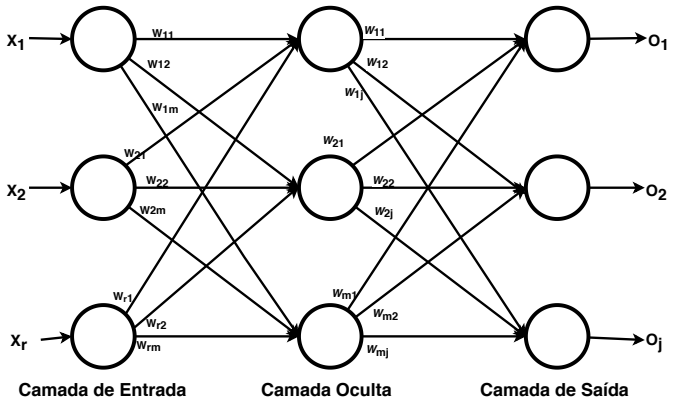


Figura 1. Representação gráfica de uma rede *feedforward* multicamadas. As entradas da rede são representadas por X_i, as saídas por O_j e os pesos de cada conexão entre os neurônios por W_{kl}.

O *treinamento* de uma RNA *feedforward* multicamadas consiste no ajuste dos pesos das conexões entre os neurônios, o que pode ser feito por diferentes algoritmos [Haykin, 2008]. De maneira geral, tal ajuste se dá de forma a minimizar, para uma determinada base de dados (conhecida como base de *treinamento*), o erro entre as saídas da RNA e as saídas reais esperadas para cada amostra dos dados. Como este treinamento exige esta base de dados com as saídas conhecidas para cada amostra, o que permite avaliar o erro apresentado pela RNA a cada etapa, trata-se de um processo de *aprendizado supervisionado* [Han et al., 2012].

As diferenças entre ELMs e as RNAs descritas até o momento estão (*i*) na existência obrigatória de uma única camada oculta nas ELMs e (*ii*) nos pesos da camada de entrada, que são definidos aleatoriamente nas ELMs. Dessa forma, o treinamento das ELMs se torna muito mais rápido, uma vez que apenas os pesos da camada de saída devem ser ajustados [Huang et al., 2006]. A estrutura geral de uma ELM é dada na Figura 2.

Considerando um conjunto de dados $D = [(\vec{x_t}, y_t)|t = 1, \cdots, N]$ com N amostras, a saída de uma ELM para a $t-$ésima amostra dos dados ($\vec{x_t}$) pode ser escrita matematicamente como na Equação (1).

$$f_L(\vec{x_t}) = \sum_{j=1}^{L} \beta_j g(\vec{a_j}, b_j, \vec{x_t}) = y_t, \tag{1}$$

onde L é o número de neurônios na camada oculta da rede, β_j é o peso da camada de saída associado ao $j-$ésimo neurônio na camada oculta, $g(\cdot)$ é a função de ativação dos neurônios da camada oculta, $\vec{a_j} = [a_{j1}, a_{j2}, ..., a_{jr}]^T$ é o vetor de pesos associados ao $j-$ésimo neurônio da camada oculta e b_j é o seu *bias*.

Sendo assim, considerando todas as amostras do conjunto de dados D, a ELM pode ser representada pela Equação (2).

$$\mathbf{H}\vec{\beta} = \vec{y}, \tag{2}$$

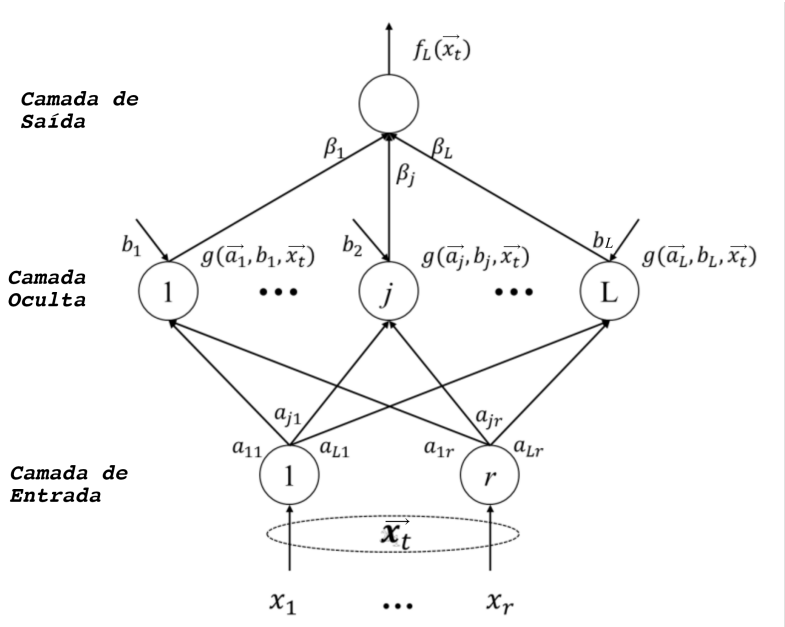


Figura 2. Representação gráfica de uma ELM com r entradas ($x_1 \cdots x_r$), uma saída ($f_L(\vec{x_t})$) e L neurônios na camada oculta. Os pesos e *bias* da camada de entrada são dados por a_{ij} e b_k, respectivamente, e os pesos da camada de saída por β_l. Adaptado de [Barajas, 2018].

onde $\mathbf{H}$ é a matriz de saídas da camada oculta, dada pela Equação (3), $\vec{\beta}$ é o vetor de pesos da camada de saída, dado por $\vec{\beta} = [\beta_1, ..., \beta_L]^T$, e $\vec{y}$ é o vetor de saídas da ELM para todas as N amostras dos dados, dado por $\vec{y} = [y_1, ..., y_N]^T$. Na matriz $\mathbf{H}$, a $j-$ésima coluna representa o vetor de saída do $j-$ésimo neurônio da camada oculta para todas as N amostras dos dados, enquanto que a $i-$ésima linha corresponde ao vetor de saída da camada oculta para $\vec{x_i}$ ($i-$ésima amostra dos dados).

$$\mathbf{H} = \begin{bmatrix} g(\vec{a_1}, b_1, \vec{x_1}) & \cdots & g(\vec{a_L}, b_L, \vec{x_1}) \\ \cdots & \cdots & \cdots \\ g(\vec{a_1}, b_1, \vec{x_N}) & \cdots & g(\vec{a_L}, b_L, \vec{x_N}) \end{bmatrix}. \tag{3}$$

Como os pesos da camada de entrada de uma ELM são gerados aleatoriamente, o treinamento destas redes consiste apenas em definir os pesos da camada de saída ($\vec{\beta}$) de forma que a Equação (2) seja satisfeita. Isto pode ser feito como descrito na Equação (4).

$$\vec{\beta} = \mathbf{H}^{\dagger}\vec{y}, \tag{4}$$

onde $\mathbf{H}^{\dagger}$ é a pseudo-inversa de $\mathbf{H}$, que pode ser calculada pela Equação (5) caso a inversa de $\mathbf{H}^T\mathbf{H}$ exista.

$$\mathbf{H}^{\dagger} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T. \tag{5}$$

Substituindo a Equação (5) na Equação (4), temos que $\vec{\beta}$ pode ser obtido por:

$$\vec{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\vec{y}. \tag{6}$$

Como o treinamento de uma ELM consiste apenas na resolução da Equação (6), sem a necessidade de aplicação de métodos iterativos como ocorre em outras RNAs, o processo de aprendizado acaba sendo muito mais rápido. Além disso, as ELMs têm apresentado boa capacidade de generalização em diversas aplicações, o que fez com que elas se tornassem muito utilizadas na literatura [Huang et al., 2011]. Sendo assim, as ELMs foram empregadas aqui como os preditores-base nos experimentos realizados.

3. Metodologia

O desenvolvimento deste trabalho envolveu três passos principais: (*i*) a obtenção dos dados de concentração de poluentes atmosféricos ($MP_{2,5}$, MP_{10} e O_3) e de internações hospitalares; (*ii*) a preparação dos dados coletados; e (*iii*) a realização de experimentos comparativos, utilizando ELMs, de previsão do número de internações hospitalares usando diferentes atributos como entrada. A metodologia empregada nestes passos será descrita nesta seção.

3.1. Obtenção dos Dados

Esta etapa foi dividida entre a obtenção dos dados de internações hospitalares, disponibilizados pelo SUS no sistema TABNET [DataSUS, 2019], e a obtenção dos dados de concentração de poluentes atmosféricos, disponibilizados pela CETESB no sistema QUALAR [QUALAR, 2019]. O período de estudo definido para este trabalho foi de 01-jan-2015 à 31-dez-2017, correspondente aos dois últimos anos completos de dados no momento de início deste trabalho.

A cidade de Campinas (SP) foi escolhida como caso de estudo. Esta escolha se deu com base no número total de internações hospitalares e no número de estações automáticas de monitoramento do ar instaladas. Estes critérios foram escolhidos para que se tivesse um número significativo de internações diárias em um município cujos dados coletados pelas estações de monitoramento do ar representem, da melhor maneira possível, a situação em todo o município. Sendo assim, dentre as cidades monitoradas pela CETESB, Campinas (SP) foi a que melhor atendeu a estes critérios.

O município de Campinas, localizado no interior do estado de São Paulo, tem cerca de 1 milhão de habitantes distribuídos em uma área de 797,6 km^2. A cidade é monitorada por 3 estações da CETESB espalhadas pela cidade. Dentre os poluentes monitorados pela CETESB, foram escolhidos $MP_{2,5}$, MP_{10} e O_3, por serem aqueles com mais evidências na literatura de seu impacto negativo na saúde humana. O sistema QUALAR disponibiliza os dados com atualizações horárias para cada estação de monitoramento. No entanto, como os dados de internações hospitalares são diários, os dados horários de concentração de cada poluente, em cada dia do período de estudos, foram convertidos em três atributos:

- Média aritmética diária;
- Máxima concentração no dia;
- Mínima concentração no dia.

Além disso, foi feita uma média entre os valores indicados por cada uma das estações.

Já os dados de internações hospitalares foram obtidos do sistema TABNET disponibilizado pelo SUS. Os dados presentes no TABNET correspondem a uma série de

informações, atualizadas diariamente para cada cidade brasileira, sobre as internações hospitalares realizadas na rede pública de saúde.

Para manipulação dos dados retornados pelo TABNET foi criado um *script*, em Python, para selecionar apenas as internações ocorridas na cidade de Campinas (que possui código 350950 no TABNET) e relacionadas a doenças respiratórias. Para isso, foi utilizado o código J da CID10 (Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde) como critério de seleção, que indica a classe de doenças do sistema respiratório. A CID10, desenvolvida pela Organização Mundial da Saúde [WHO, 2016b], é um catálogo de doenças e sintomas que associa códigos alfanuméricos a cada item cadastrado.

Todos os dados coletados e filtrados foram salvos em arquivos CSV (*comma-separated values*).

3.2. Preparação dos Dados

O primeiro passo após a recuperação dos dados das bases QUALAR e TABNET foi o tratamento dos valores faltantes. Para isso, como os dados deste trabalho são séries temporais, optou-se pela estratégia de imputação pela média entre o último valor existente e o próximo valor não faltante na série [Enders, 2013].

Na sequência, foi feita uma normalização Min-Max [Han et al., 2012] dos valores de cada série para o intervalo $[-1; +1]$. Este intervalo foi definido de tal forma pois as ELMs utilizadas neste trabalho foram configuradas com função de ativação do tipo tangente hiperbólica (vide Seção 3.3).

Por fim, os dados preparados foram salvos em novos arquivos do tipo CSV. Na sequência, foram criados *scripts*, também em Python, para transformar os dados das séries temporais em amostras, com atributos de entrada e valor desejado de saída, para treinamento das ELMs. Estes *scripts* recebem como entrada o número de dias anteriores da série que serão considerados como entrada para a ELM, e preparam novos arquivos de dados para serem usados nas etapas de treinamento e teste dos experimentos (vide Seção 3.3). Foram criados *scripts* para tratar tanto os dados de concentração de poluentes e internações hospitalares isoladamente quanto de forma combinada.

3.3. Metodologia Experimental

Todos os experimentos realizados neste trabalho tiveram como base a implementação de *Extreme Learning Machines* da biblioteca *Python-ELM*, disponível no GitHub[1]. As ELMs foram configuradas com função de ativação do tipo tangente hiperbólica em seus neurônios e tanto o número de neurônios na camada oculta quanto o número de dias anteriores utilizados como entrada[2] foram definidos empiricamente. Nesta etapa foram consideradas ELMs com 20, 40, 60 e 80 neurônios na camada oculta e dados de entrada referentes a 5, 7, 10 e 20 dias anteriores ao instante de previsão. Em todos os experimentos, o objetivo foi sempre obter o número total de internações hospitalares, causadas por doenças respiratórias, no dia seguinte ao instante atual ($t + 1$).

[1]Disponível em: `https://github.com/dclambert/Python-ELM`

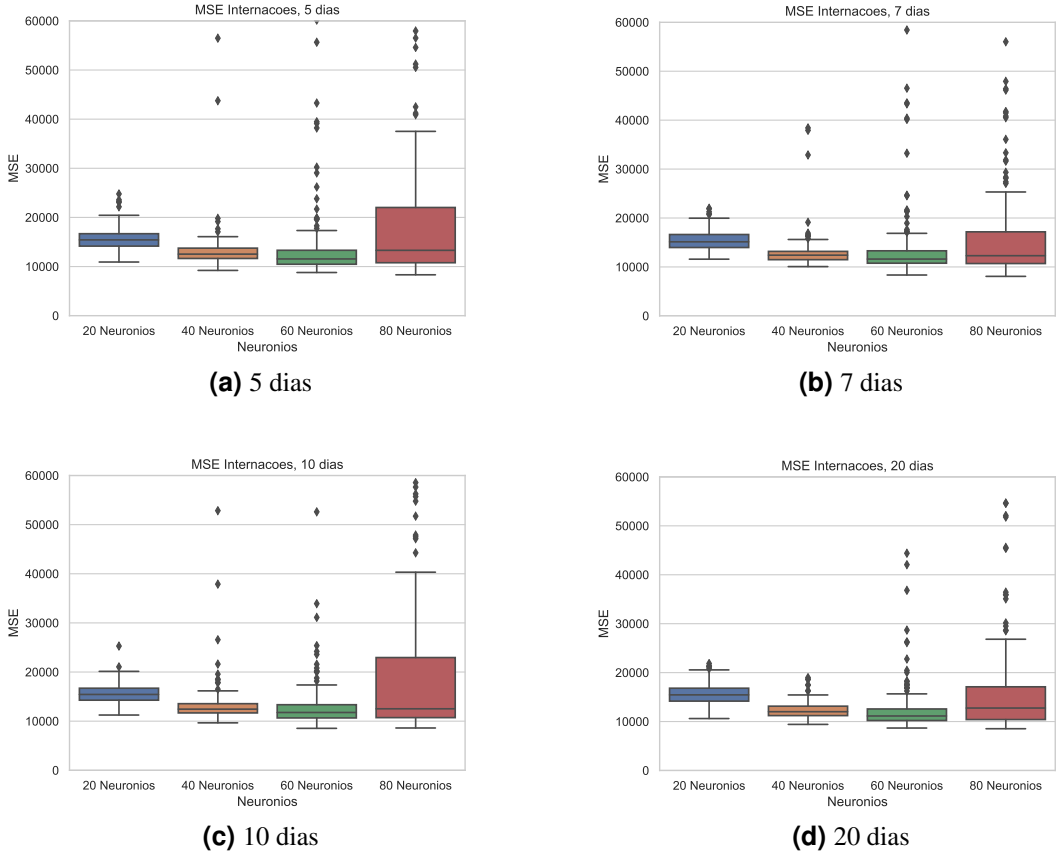
[2]Para cada atributo (poluente ou dados de internações), o número de dias anteriores da série que devem ser considerados na entrada das ELMs.

Todas as comparações foram feitas tendo como critério o erro quadrático médio (MSE, do inglês *Mean Squared Error*), calculado entre a previsão do número de internações retornada pelas ELMs e o valor real constante na base de dados.

Cada experimento foi repetido 200 vezes, sendo que em cada uma as amostras do conjunto de dados, previamente organizadas no formato entrada-saída, foram divididas aleatoriamente em 80% para treinamento e 20% para testes. Ao final, os resultados foram avaliados através de *box-plots*, criados a partir dos resultados obtidos nas 200 repetições de cada experimento.

4. Resultados

Os *box-plots* obtidos para as 200 repetições de cada experimento de ajuste dos parâmetros das ELMs, realizados com dados da própria série temporal de internações hospitalares como entrada, são dados na Figura 3. Como é possível observar, a melhor configuração das ELMs foi obtida com 60 neurônios na camada oculta e entradas compostas por dados dos 20 dias anteriores ao instante de previsão, parâmetros estes que foram adotados nas próximas etapas.



(a) 5 dias

(b) 7 dias

(c) 10 dias

(d) 20 dias

Figura 3. *Box-plots* dos erros quadráticos médios (MSE) para diferentes números de neurônios na camada oculta das ELMs e diferentes entradas (números de dias anteriores).

Definidos os parâmetros das ELMs, passou-se à etapa de verificação da qualidade das previsões do número de internações hospitalares a partir de diferentes dados de entrada. Para isso, foram feitos experimentos com: (*i*) apenas dados de concentração de

$MP_{2,5}$ como entrada (valores médios, máximo e mínimo diários para os 20 dias que antecedem a previsão); (*ii*) apenas dados de concentração de MP_{10} (mesma configuração que a adotada para $MP_{2,5}$); e (*iii*) apenas dados de concentração de O_3 (mesma configuração que a adotada para $MP_{2,5}$). Os valores médios, mediana, quartil inferior e quartil superior do erro quadrático médio (MSE), obtidos nas 200 repetições destes experimentos, em conjunto com os resultados para os experimentos que utilizaram apenas a série original de dados de internações hospitalares como entradas para as ELMs, são dados na Tabela 1.

Tabela 1. Resultados obtidos (MSE) para os experimentos que utilizaram os dados originais de internações hospitalares, concentração de MP_{10} ($[MP_{10}]$), de $MP_{2,5}$ ($[MP_{2,5}]$) e de O_3 ($[O_3]$) como entrada das ELMs.

Dados de Entrada	Média	Mediana	Quartil Inferior	Quartil Superior
Internações	$3{,}0\times10^4$	$1{,}1\times10^4$	$1{,}0\times10^4$	$1{,}3\times10^4$
$[MP_{10}]$	$7{,}0\times10^4$	$2{,}5\times10^4$	$2{,}4\times10^4$	$2{,}5\times10^4$
$[MP_{2,5}]$	$1{,}8\times10^4$	$1{,}6\times10^4$	$1{,}5\times10^4$	$1{,}7\times10^4$
$[O_3]$	$2{,}5\times10^4$	$2{,}2\times10^4$	$2{,}1\times10^4$	$2{,}2\times10^4$

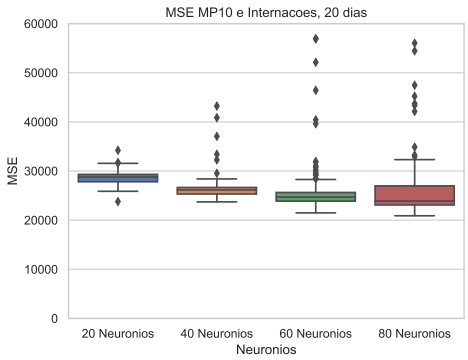
Como é possível observar na Tabela 1, após 200 repetições dos experimentos os melhores resultados (mediana e quartis inferior e superior) foram obtidos para as previsões realizadas com os dados originais de internações hospitalares. No entanto, as previsões feitas com concentração de $MP_{2,5}$ e de O_3 apresentaram um valor médio de MSE menor, o que pode indicar a existência de *outliers* nos resultados obtidos com a série original de internações hospitalares.

Na sequência foram feitos experimentos combinando, como entrada das ELMs, tanto dados da série original de internações hospitalares quanto de concentração de cada poluente atmosférico considerado aqui (ainda considerando dados dos 20 dias que antecedem cada instante de previsão como entrada das ELMs). Como neste caso o número de entradas nas ELMs aumenta significativamente, novas análises para definir o melhor número de neurônios foram feitas e os resultados são apresentados na Figura 4. Como é possível observar, em todos os casos o menor valor da mediana do MSE foi obtido com 80 neurônios na camada oculta, valor que foi adotado nos experimentos restantes.

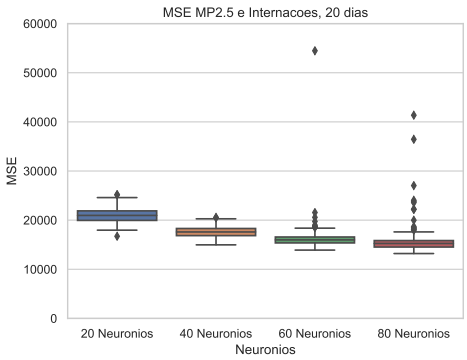
Na Tabela 2 são apresentados os valores médios, mediana, quartil inferior e quartil superior do erro quadrático médio (MSE), obtidos nas 200 repetições dos experimentos que combinaram dados de concentração de cada poluente atmosférico com a série original de internações hospitalares na entrada das ELMs. Como é possível perceber, no caso da combinação de concentração de $MP_{2,5}$ e O_3 com dados de internações houve uma redução em todos os critérios avaliados (exceto na média, para $[O_3]$), quando comparado aos resultados obtidos para as previsões realizadas apenas com os dados de concentração de cada poluente (vide Tabela 1). No entanto, esta redução não foi capaz de superar os resultados de previsão obtidos com os dados da série original de internações isoladamente.

Por fim, foram feitos também experimentos que combinaram, como entrada para as ELMs, dados de concentração de diferentes poluentes atmosféricos. No entanto, tais resultados (apresentados na Tabela 3) não se mostraram promissores, uma vez que o MSE foi, na maioria dos casos, significativamente maior ao observado anteriormente.

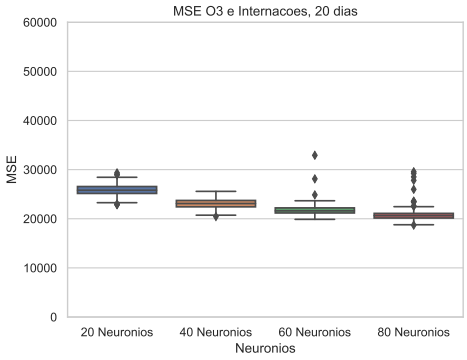
Diante dos resultados obtidos pode-se concluir que, para os dados da cidade de



(a) MP_{10} e Internações Hospitalares



(b) $MP_{2,5}$ e Internações Hospitalares



(c) O_3 e Internações Hospitalares

Figura 4. *Box-plots* dos erros quadráticos médios (MSE) para diferentes números de neurônios na camada oculta das ELMs e para diferentes combinações de dados de entrada.

Tabela 2. Resultados obtidos (MSE) para os experimentos que combinaram dados de concentração de cada poluente atmosférico com a série original de internações hospitalares na entrada das ELMs.

Dados de Entrada	Média	Mediana	Quartil Inferior	Quartil Superior
$[MP_{10}]$ e Internações	$1{,}8\times10^5$	$2{,}4\times10^4$	$2{,}3\times10^4$	$2{,}7\times10^4$
$[MP_{2,5}]$ e Internações	$1{,}8\times10^4$	$1{,}5\times10^4$	$1{,}5\times10^4$	$1{,}6\times10^4$
$[O_3]$ e Internações	$2{,}8\times10^4$	$2{,}1\times10^4$	$2{,}0\times10^4$	$2{,}1\times10^4$

Tabela 3. Resultados obtidos (MSE) para os experimentos que combinaram dados de concentração de diferentes poluentes atmosféricos como entrada das ELMs.

Dados de Entrada	Média	Mediana	Quartil Inferior	Quartil Superior
$[MP_{10}]$ e $[MP_{2,5}]$	$1{,}6\times10^7$	$2{,}4\times10^4$	$2{,}3\times10^4$	$2{,}6\times10^4$
$[MP_{10}]$ e $[O_3]$	$2{,}6\times10^5$	$2{,}4\times10^4$	$2{,}3\times10^4$	$2{,}8\times10^4$
$[MP_{2,5}]$ e $[O_3]$	$2{,}8\times10^4$	$2{,}1\times10^4$	$2{,}0\times10^4$	$2{,}1\times10^4$

Campinas (SP), os menores erros de previsão do número de internações hospitalares 24 h à frente são obtidos quando a própria série histórica é utilizada para treinar as *Extreme Learning Machines*. Em todos os casos em que os dados de concentração de MP_{10}, $MP_{2,5}$ e O_3 foram utilizados, os erros de predição observados foram maiores.

5. Conclusões e Trabalhos Futuros

Neste trabalho verificou-se a possibilidade de prever, a partir de dados de concentração de poluentes atmosféricos e com 24 h de antecedência, o número de internações hospitalares associadas a doenças respiratórias. Foram utilizadas *Extreme Learning Machines* como preditores e diversos experimentos foram feitos com dados referentes à cidade de Campinas (SP). Avaliou-se a possibilidade de se prever o número de internações hospitalares tanto a partir dos dados de concentração de MP_{10}, $MP_{2,5}$ e O_3 isoladamente quanto a partir de diferentes combinações de dados, incluindo a série histórica de internações.

Os resultados mostraram que, apesar do uso de dados de alguns poluentes específicos levar a menores erros de previsão, quando comparados às previsões feitas com dados de outros poluentes, os melhores resultados ainda foram obtidos utilizando-se apenas a série histórica de internações hospitalares como entrada para as ELMs.

Como trabalhos futuros, pretende-se avaliar se estes resultados se mantêm para outras cidades, principalmente para aquelas que apresentam concentrações mais altas de poluentes. Além disso, será verificado também se a inclusão de outros parâmetros ambientais, tais como temperatura e umidade do ar, podem contribuir para a melhoria da qualidade da previsão do número de internações hospitalares.

Agradecimentos

Agradecemos ao SAE/UNICAMP, pela concessão da bolsa de Iniciação Científica, e à Profa. Dra. Yara S. Tadano (UTFPR), pelo auxílio na coleta dos dados via TABNET.

Referências

Anderson, J. O., Thundiyil, J. G., and Stolbach, A. (2012). Clearing the air: A review of the effects of particulate matter air pollution on human health. *Journal of Medical Toxicology*, 8(2):166–175.

Barajas, J. A. B. (2018). Dynamic ensemble mechanisms to improve particulate matter forecasting. Master's thesis, Faculdade de Tecnologia (FT), Universidade Estadual de Campinas (UNICAMP).

Bisht, M. and Seeja, K. R. (2018). Air pollution prediction using Extreme Learning Machine: A case study on Delhi (India). In *Proceedings of the First International Conference on Smart System, Innovations and Computing*, pages 181–189.

Braga, B., Hespanhol, I., Conejo, J. G. L., Mierzwa, J. C., de Barros, M. T. L., Spencer, M., Porto, M., Nucci, N., Juliano, N., and Eiger, S. (2005). *Introdução à engenharia ambiental – O desafio do desenvolvimento sustentável*. Prentice-Hall Brasil.

Burnett, R. T., Brook, J., Dann, T., Delocla, C., Philips, O., Cakmak, S., Vincent, R., Goldberg, M. S., and Krewski, D. (2000). Association between particulate- and gas-phase components of urban air pollution and daily mortality in eight Canadian cities. *Inhalation Toxicology*, 12(sup4):15–39.

CETESB (2019). Qualidade do ar – Poluentes. Disponível em: `https://cetesb.sp.gov.br/ar/poluentes/`. Acessado em 12-Jul-2019.

DataSUS (2019). Informações de saúde (TABNET). Disponível em: `http://datasus.saude.gov.br/informacoes-de-saude/tabnet`. Acessado em 12-Jul-2019.

Enders, C. K. (2013). *Applied Missing Data Analysis*. Guilford Press.

Gardner, M. W. and Dorling, S. R. (1998). Artificial neural networks (the Multilayer Perceptron) — a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14–15):2627—2636.

Godish, T. (2004). *Air Quality*. CRC Press, 4th. edition.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers.

Haykin, S. (2008). *Neural Networks and Learning Machines*. Pearson, 3rd. edition.

Huang, G.-B., Wang, D. H., and Lan, Y. (2011). Extreme learning machines: A survey. *International Journal of Machine Learning and Cybernetics*, 2(2):107–122.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1–3):489–501.

Loomis, D., Castillejos, M., Gold, D. R., McDonnell, W., and Borja-Aburto, V. H. (1999). Air pollution and infant mortality in Mexico City. *Epidemiology*, 10(2):118–123.

Morettin, P. A. and Toloi, C. M. C. (2006). *Análise de Séries Temporais*. Edgard Blucher.

Ostro, B., Broadwin, R., Feng, S. G. W. Y., and Lipsett, M. F. (2006). Particulate air pollution and mortality in nine California counties: Results from CALFINE. *Environmental Health Perspectives*, 144(1):29–33.

QUALAR (2019). Qualidade do ar. Disponível em: `https://cetesb.sp.gov.br/ar/qualar/`. Acessado em 12-Jul-2019.

Seinfeld, J. H. and Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. Wiley.

Tadano, Y. S., Alves, T. A., Silva, N. S. S., and Valadares, H. S. (2017). Impacto da poluição atmosférica e das alterações climáticas na saúde populacional utilizando redes neurais artificiais. *Mecânica Experimental – Revista da Associação Portuguesa de Análise Experimental de Tensões*, 29:35–42.

WHO (2016a). Ambient air pollution: A global assessment of exposure and burden of disease. Technical report, World Health Organization. Disponível em: `https://www.who.int/phe/publications/air-pollution-global-assessment/en/`. Acessado em: 12-jul-2019.

WHO (2016b). International statistical classification of diseases and related health problems 10th revision. Disponível em: `https://icd.who.int/browse10/2016/en`. Acessado em 12-Jul-2019.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., Zhu, L., and Zhang, M. (2017). Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across china using a spatially explicit machine learning algorithm. *Atmospheric Environment*, 55:129–139.