Impacto da poluição atmosférica na saúde: Máquinas de Aprendizado Extremo

Health impact of air pollution: Extreme Learning Machines

João Luiz Miranda Meyer*, Yara de Souza Tadano†,
Hugo Valadares Siqueira‡, Thiago Antonini Alves§

RESUMO

A poluição atmosférica é responsável por aproximadamente 7 milhões de mortes anualmente. Neste sentido, realizar previsões dos impactos à saúde causados pelo material particulado com diâmetro aerodinâmico menor que 10 micrometros (MP_{10}), um dos poluentes mais prejudiciais à saúde, é de suma importância. Assim, o objetivo do presente trabalho será prever o número de internações por doenças respiratórias devido à exposição ao MP_{10} utilizando uma rede neural artificial, denominada Máquinas de Aprendizado Extremo (ELM – do inglês *Extreme Learning Machine*). O estudo de caso foi realizado para a cidade de São Paulo, datados de 01 de janeiro de 2014 até 31 de dezembro de 2016. As variáveis consideradas foram: concentração de MP_{10}, temperatura média, umidade relativa, o dia da semana e o dia ser feriado ou não. Os resultados obtidos foram satisfatórios, mostrando um poder de previsão da ELM para o problema em questão, mas com limitações, por possuir uma solução analítica e não iterativa como outras arquiteturas de rede, possuindo erros consideráveis. Trabalhos futuros podem avaliar a concentração de outros poluentes e variáveis, testar novos parâmetros da rede e outras arquiteturas de rede.

Palavras-chave: Rede Neural Artificial, Doenças Respiratórias, Previsão, Máquinas Desorganizadas, Material Particulado.

ABSTRACT

Air pollution is responsible for approximately 7 million deaths annually. In this sense, forecasting the health impacts caused by particulate matter with an aerodynamic diameter smaller than 10 micrometers (PM_{10}), one of the most harmful pollutants to health, is of paramount importance. Thus, the objective of this study will be to predict the number of hospitalizations for respiratory diseases due to exposure to MP_{10} using an artificial neural network, called Extreme Learning Machine (ELM). The case study was carried out for the city of São Paulo, dated from January 1, 2014 to December 31, 2016. The variables considered were: PM_{10} concentration, average temperature, relative humidity, day of the week and day to be holiday or not. The results obtained were satisfactory, showing ELM's predictive power for the problem in question, but with limitations, as it has an analytical and non-iterative solution like other network architectures, with considerable errors. Future work may assess the concentration of other pollutants and variables, test new network parameters and other network architectures.

Keywords: Artificial Neural Network, Respiratory Diseases, Prediction, Disorganized Machines, Particulate Matter.

*Engenharia Química, Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil; joaomeyer@alunos.utfpr.edu.br
†Universidade Tecnológica Federal do Paraná, Campus Ponta Grossa; yaratadano@utfpr.edu.br
‡Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil; hugosiqueira@utfpr.edu.br
§Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil; antonini@utfpr.edu.br

1 INTRODUÇÃO

Segundo a Organização Mundial da Saúde (OMS 2020), aproximadamente 7 milhões de pessoas morrem todos os anos devido à poluição atmosférica. Um dos principais poluentes que impacta diretamente na saúde humana é o material particulado (POLEZER et al., 2018). Neste sentido, é importante estudos que possam avaliar e prever os impactos causados pela poluição atmosférica na saúde populacional.

Como há uma relação visível entre a poluição atmosférica e o malefício causado na saúde humana, é possível prever o número de internações hospitalares causadas por doenças respiratórias ou cardiovasculares devido exclusivamente à poluição atmosférica. Há uma vasta quantidade de publicações que utilizam a regressão estatística como forma para realizar essa previsão (TADANO et al., 2012; LAZARRI, 2013; VANOS, HEBBERN e CAKMAK, 2014). Entretanto, estudos recentes têm mostrado que as Redes Neurais Artificiais (RNA) são uma ótima ferramenta para classificação e previsão de problemas complexos que envolvem muitas variáveis (POLEZER et al., 2018; ARAUJO et al., 2020).

Desta forma, o objetivo do presente trabalho será realizar a previsão de internações por doenças respiratórias causadas pelo material particulado com até 10 µm de diâmetro aerodinâmico (MP_{10}), na cidade de São Paulo, Estado de São Paulo, Brasil, utilizando uma RNA denominada Máquinas de Aprendizado Extremo (ELM – do inglês *Extreme Learning Machine*). Seria possível a utilização da ELM para previsão de doenças respiratórias causadas pela poluição atmosférica, com um certo nível de confiabilidade?

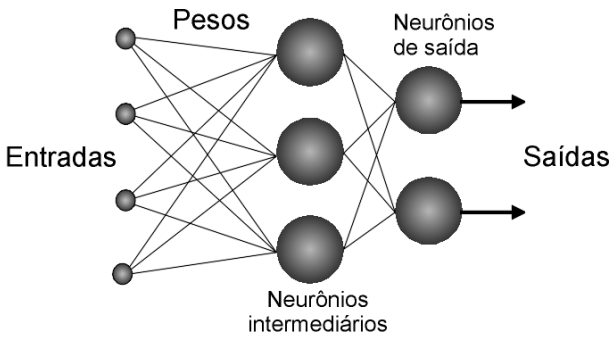
1.1 Material particulado

O material particulado (MP) é um poluente atmosférico que consiste em partículas sólidas ou líquidas suspensas no ar, podendo ser de origem natural (vulcões, queimadas naturais) como antropogênica (combustíveis fósseis, processos industriais, ressuspensão de poeiras ou outras origens) (FREITAS e SOLCI, 2009). A classificação do MP pode ser feita devido ao seu tamanho, sua composição química ou tipo (primário ou secundário). Quanto ao tamanho, é frequentemente classificado como: Partículas Inaláveis Finas ($MP_{2,5}$) - cujo diâmetro aerodinâmico é menor ou igual a 2,5 µm, penetram profundamente no sistema respiratório, podendo atingir os alvéolos pulmonares; Partículas Inaláveis (MP_{10}) - diâmetro aerodinâmico é menor ou igual a 10 µm, e podem ficar retidas na parte superior do sistema respiratório ou penetrar mais profundamente, alcançando os alvéolos pulmonares; Partículas Totais em Suspensão (PTS) - diâmetro aerodinâmico é menor ou igual a 50 µm; uma parte dessas partículas é inalável e pode causar problemas à saúde, outra parte pode afetar desfavoravelmente a qualidade de vida da população, interferindo nas condições estéticas do ambiente e prejudicando as atividades normais da comunidade (CETESB, 2021).

1.2 Redes neurais artificiais

As RNA são modelos computacionais capazes de relacionar dados de entrada e de saída, com o intuito de fornecer uma generalização de padrões e assim ser possível realizar previsões ou então classificações. Dentre as tarefas passíveis de sua aplicação, destacam-se também reconhecimentos faciais, reconhecimento de padrões e detecção de anomalias (HAYKIN, 2015). As RNA são inspiradas no sistema nervoso de seres superiores, possuindo neurônios artificiais e conexões neurais (pesos) que fazem as comunicações entre os neurônios. São divididas em três camadas: entrada, escondida (podendo possuir mais de uma) e de saída. A Figura 1 representa uma arquitetura de uma RNA.

Figura 1 - Estrutura de uma RNA



Fonte: TAFNER (1998).

A RNA que será usada para este trabalho será a ELM. A sua estrutura possui apenas uma camada escondida. O método de treinamento é feito de forma analítica, necessitando de um baixo poder computacional para o seu treinamento. Os seus pesos são gerados inicialmente de forma aleatórias e apenas os pesos entre a camada oculta e a camada de saída são atualizados (POLEZER et al., 2018).

2 MÉTODO

Como estudo de caso, considerou-se os dados de 01 de janeiro de 2014 à 31 de dezembro de 2016 para a cidade de São Paulo, SP, Brasil, consistindo em 1.009 dados diários. As variáveis utilizadas neste trabalho foram: concentração de MP_{10}, temperatura média, umidade relativa, o dia da semana e o dia ser feriado ou não. Incluir dias da semana e feriado como variáveis qualitativas é primordial, já que os dados de internações frequentemente possuem um padrão definido, com menos internações sendo observadas em dias de semana e dias de não feriado (TADANO et al., 2016; ARAUJO et al. 2020).

Pelo website da Companhia Ambiental do Estado de São Paulo (CETESB) foi possível obter os dados meteorológicos e de concentração de MP_{10} (CETESB 2018). Quanto aos dados de saúde, o website do DataSUS, disponibiliza de forma pública e acesso aberto dados de internações hospitalares, não sendo necessário aprovação de comitê de ética em pesquisa para a utilização dos dados. Neste caso, foram utilizados dados de internações por doenças respiratórias (CID-10: J00- J99) (DATASUS, 2018).

Outra análise importante é considerar o efeito na saúde da poluição atmosférica após alguns dias de exposição, sendo comum trabalhar com defasagens de até sete dias (TADANO et al., 2012; POLEZER et al., 2018; ARAUJO et al., 2020). Portanto, neste trabalho, foram realizadas análises considerando o efeito na saúde devido à exposição ao MP_{10} no mesmo dia da exposição (lag 0) até 7 dias após a exposição (lag 7).

Os 85% dos dados iniciais foram utilizados para treinamento, enquanto os 15% restantes para teste, o que possibilitou uma comparação da resposta da ELM com os dados reais observados. Essa divisão foi escolhida por ser bastante usada na literatura (SILVA et al., 2010). A ordem temporal dos dados foi respeitada para realizar esta divisão.

Para a implementação da ELM, utilizou-se o *software* da programação Python, versão 3.7. Além disso, foi usado a biblioteca Numpy (PYTHON, 2019), o que facilitou as multiplicações de matrizes e possibilitou a obtenção da matriz inversa generalizada de Moore-Penrose, necessária para a solução da rede. A tangente hiperbólica foi utilizada como função de ativação, bem como 20 neurônios na camada escondida.

3 RESULTADOS

Métricas de erro entre a saída desejada (dados reais) e a saída calculada pela rede é frequentemente utilizada para avaliar o desempenho de uma RNA. Neste estudo, utilizou-se o Erro Absoluto Médio (do inglês *Mean Absolute Error* – MAE), que informa o quão distante os resultados ficaram dos valores reais, em média e; a Raiz Quadrada do Erro Quadrático Médio (do inglês *Root Mean Squared Error* – RMSE), que é um erro mais sensível para quando os resultados são distantes dos valores reais, por estar elevado ao quadrado. As Equações 1 e 2 apresentam o cálculo do MAE e do RMSE, respectivamente:

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N} |d_t - r_t| \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} (d_t - r_t)^2} \quad (2)$$

sendo que d_t é o valor real, r_t é a saída da rede e N é o número total de dados utilizados na fase de teste. O Quadro 1 mostra os resultados para as melhores 30 execuções da ELM em relação aos dados de teste.

Quadro 1 – Melhores MAE e RMSE obtidos pela rede

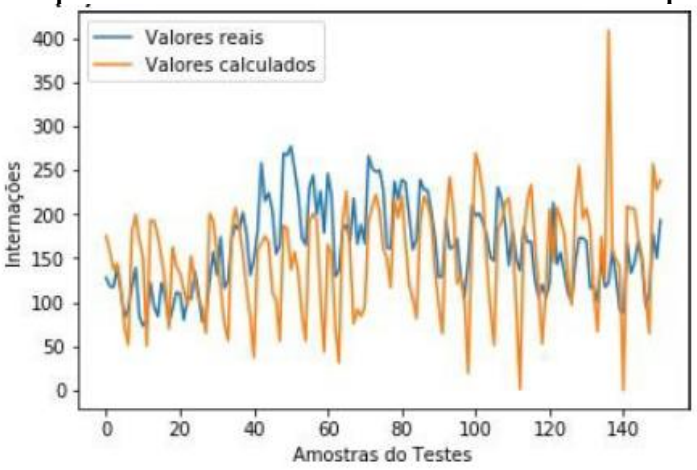
	MAE	RMSE
Lag 0	54,95	70,00
Lag 1	62,39	79,32
Lag 2	63,62	78,97
Lag 3	55,62	70,17
Lag 4	59,07	72,76
Lag 5	59,41	73,48
Lag 6	**52,68**	**66,94**
Lag 7	54,06	68,85

Fonte: Autoria própria (2021).

É possível observar, pelo Quadro 1, que os melhores resultados obtidos foram para o lag 6, ou seja, os resultados indicam uma maior relação entre a exposição ao PM_{10} e o aumento do número de internações 6 dias após o contato. É importante enfatizar que muitos outros fatores que não foram considerados para este estudo podem influenciar o resultado final (idade, gênero, condição social, genética, alimentação), isso explica a existência de erros consideráveis no estudo (TADANO et al., 2016; POLEZER et al., 2018; ARAUJO et al., 2020).

A Figura 2 mostra, de forma gráfica, a diferença entre os resultados observados e os calculados pela rede. Esse gráfico representa o melhor resultado obtido para as 30 execuções da rede.

Figura 2 – Comparativo entre os valores reais e os valores calculados pela rede



Fonte: Autoria própria (2021).

É possível observar regiões com valores próximos entre os reais e os calculados enquanto outras regiões uma maior diferença. É importante observar que a rede possui mais facilidade de prever os valores que estão mais próximos à média.

4 CONCLUSÃO

A ELM mostrou ser uma ferramenta robusta para o presente objetivo, sendo capaz de realizar previsões de internações por doenças respiratórias causadas pela poluição atmosférica, para este caso específico o MP_{10}. Isso possibilita medidas preventivas por parte de governantes com o intuito de desafogar os sistemas de saúde em períodos críticos de poluição do ar ou outras situações, como a da atual pandemia da COVID-19 (TADANO et al. 2021). Salienta-se, porém, que a ELM, por possuir uma solução analítica e não iterativa como outras RNA, pode possuir limitações para os resultados, nem sempre atingindo erros pequenos. Trabalhos futuros podem avaliar a concentração de outros poluentes e variáveis, testar novos parâmetros da rede e outras arquiteturas de RNA.

5 AGRADECIMENTOS

Os autores agradecem o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa concedida durante o desenvolvimento do projeto, fazendo parte do Programa Institucional de Bolsas de Iniciação Tecnológica e Inovação (PIBITI 2020 – 2021).

REFERÊNCIAS

ARAUJO, Lilian Nascimento. et al. *Ensemble method based on Artificial Neural Networks to estimate air pollution health risks*. **Environmental Modelling and Software**, vol. 123. 2020.

CETESB - Companhia Ambiental do Estado de São Paulo. 2021 Disponível em: <https://cetesb.sp.gov.br/ar/poluentes/>. Acesso em: 12 de ago. 2021.

DATASUS - Departamento de Informática do Sistema Único de Saúde. 2018. Disponível em: <http://www2.datasus.gov.br/DATASUS/%20index.php?area=02>. Acesso em: 7 de ago. 2021.

FREITAS, Adriana de Marques; SOLCI, Maria Cristina. Caracterização do MP_{10} e $MP_{2,5}$ e distribuição por tamanho de cloreto, nitrato e sulfato em atmosfera urbana e rural de Londrina**. Química Nova**, vol. 32, p. 1750-1754. 2009.

HAYKIN, Simon. *Neural Networks and Learning Machines*, **Prentice Hall, 3rd edition**. 2015.

IBGE - Instituto Brasileiro de Geografia e Estatística. 2019. Disponível em: <https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao/9103-estimativas-de populacao.html?1⁄4&t1⁄4resultados>. Acesso em: 8 de ago. 2021.

LAZZARI, Angela Radünz. Comparação de técnicas estatísticas para analisar a relação entre doenças respiratórias e concentrações de poluentes atmosféricos. **Ciência e Natura**, v. 35, n. 1, p. 98-105, 2013.

OMS - Organização Mundial da Saúde. ***Air Pollution***. 2020**.** Disponível em: <https://www.who.int/health-topics/air-pollution#tab=tab_1>. Acesso em: 12 de ago. 2021.

PYTHON. 2019. Disponível em: < https://phyton.org>. Acesso em: 12 de ago. 2021.

POLEZER, Gabriela. et al. *Assessing the impact of PM 2.5 on respiratory disease using artificial neural networks*. **Environmental Pollution**, n. 235, páginas 394-403. 2018.

SILVA, Ivan Nunes Da. et al. **Redes Neurais Artificiais para Engenharia e Ciências Aplicadas**, Artliber, 1a edição. 2010.

TADANO, Yara de Souza. et al. *Methodology to assess air pollution impact on human health using the Generalized Linear Model with Poisson regression*. **Air Pollution – Monitoring, Modelling and Health**, England. 2012.

TADANO, Yara de Souza. et al. *Unorganized machines to predict hospital admissions for respiratory diseases*. In: IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2016, Cartagena. **Anais…** Cartagena, IEEE Xplore, 2016.

TADANO, Yara de Souza. et al. *Dynamic model to predict the association between air quality, COVID-19 cases, and level of lockdown*. **Environmental Pollution**, n. 268. 2021.

TAFNER, Malcon Anderson. 1998. O Que São as Redes Neurais Artificiais**. Cérebro & Mente**. Disponível em: <https://cerebromente.org.br/n05/tecnologia/rna_i.htm>. Acesso em: 19 ago. 2021.

VANOS, Jennifer; HEBBERN, Christopher; CAKMAK, Sabit. *Risk assessment for cardiovascular and respiratory mortality due to air pollution and synoptic meteorology in 10 Canadian cities*. **Environmental Pollution**, n. 185, p. 322–332. 2014.