Comparação entre Redes Neurais Artificiais para Avaliação do Impacto do MP_{10} na Saúde Humana

João Luiz M. Meyer[1], Hugo Siqueira Valadares[2], Thiago Antonini Alves[3], Yara de Souza Tadano[4]

[1]Departamento de Engenharia Química – Universidade Tecnológica Federal do Paraná (UTFPR) – 84.017-220 – Ponta Grossa – PR – Brasil

[2]Departamento de Engenharia Elétrica – Universidade Tecnológica Federal do Paraná (UTFPR) – 84.017-220 – Ponta Grossa – PR – Brasil

[3]Departamento de Engenharia Mecânica – Universidade Tecnológica Federal do Paraná (UTFPR) – 84.017-220 – Ponta Grossa – PR – Brasil

[4]Departamento de Matemática – Universidade Tecnológica Federal do Paraná (UTFPR) – 84.017-220 – Ponta Grossa – PR – Brasil (orientadora)

joao_lmm@hotmail.com, {hugosiqueira,antonini,yaratadano}@utfpr.edu.br

Abstract. *The high emission of air pollutants in large urban centers causes several harms for population health. In this work, the proposal was to estimate the number of hospital admissions for respiratory diseases due to the concentration of particulate matter with an aerodynamic diameter less than or equal to 10 micrometers (PM_{10}) using two Artificial Neural Networks, Multilayer Perceptron (MLP) and Extreme Learning Machines (ELM). The considered models' inputs are PM_{10} concentration and meteorological variables. The computational results showed that the MLP was able to achieve better performances.*

Resumo. *A alta emissão de poluentes atmosféricos em grandes centros urbanos acarreta diversos malefícios para a saúde populacional. Neste trabalho, a proposta foi estimar o número de internações hospitalares por doenças respiratórias devido à concentração de material particulado com diâmetro aerodinâmico menor ou igual à 10 micrometros (MP_{10}) utilizando duas Redes Neurais Artificiais, Perceptron de Múltiplas Camadas (MLP) e Máquinas de Aprendizado Extremo (ELM). As entradas dos modelos consideradas são a concentração de MP_{10} e variáveis meteorológicas. Os resultados computacionais mostraram que a rede MLP foi capaz de alcançar melhores desempenhos.*

1. Introdução

A emissão de poluentes atmosféricos em grandes centros urbanos costuma ser a causa do elevado número de internações por doenças respiratórias ou cardiovasculares [Polezer et al. 2018]. De acordo com a Organização Mundial da Saúde [OMS 2020], há cerca de 7 milhões de mortes anuais no mundo devido à poluição do ar. Além disso, 9 em cada 10 pessoas respiram o ar com altas concentrações dos diversos

poluentes [OMS 2020]. Isso gera a necessidade de um entendimento adequado sobre impacto dos poluentes atmosféricos na saúde humana.

A literatura mostra que tal problema é tratado frequentemente com o uso de regressão estatística, porém também podem ser modelado usando Redes Neurais Artificiais (RNA), com melhor poder de previsão [Araujo et al. 2020]. Uma instância deste problema é estimar o número de internações hospitalares causadas por doenças respiratórias tendo como entradas do modelo, variáveis climáticas e de concentração de poluentes [Polezer et al. 2018], [Araujo et al. 2020].

Este trabalho utilizará duas arquiteturas de RNA, a Perceptron de Múltiplas Camadas (do inglês *Multilayer Perceptron* - MLP) e a Máquina de Aprendizado Extremo (*Extreme Learning Machine* - ELM) para a estimação de internações por doenças respiratórias na cidade de São Paulo.

2. Poluentes Atmosféricos

Poluentes atmosféricos são quaisquer substâncias presentes no ar e que possam de alguma forma causar danos à saúde, materiais, fauna e flora. Vários deles podem ser causadores de doenças respiratórias, dentre eles, o material particulado que consiste em partículas sólidas e líquidas suspensas no ar [CETESB 2021]. Podem ser de origem natural (vulcões, queimadas naturais) como antropogênica (combustíveis fósseis, processos industriais, ressuspensão de poeiras ou outras origens [Freitas e Solci 2009]. Este trabalho tem como foco o material particulado com diâmetro aerodinâmico menor ou igual a 10 μm (MP_{10}).

3. Redes Neurais Artificiais

As RNA são ferramentas computacionais capazes de mapear de maneira não-linear padrões de entrada em uma saída desejada, com capacidade de generalização elevada para dados desconhecidos. Dentre as tarefas passíveis de sua aplicação, destacam-se classificações, previsões, reconhecimentos faciais, reconhecimento de padrões e detecção de anomalias [Haykin 2015].

A MLP é a rede neural mais tradicional devido a seu processo sistemático de treinamento (calibração) e aos bons resultados apresentados na literatura. Sua estrutura é composta pela camada de entrada, uma ou mais camadas escondidas e a camada de saída. Em cada uma delas, neurônios artificiais estarão dispostos de modo que em camadas disjuntas, eles são totalmente conectados, mas não se comunicam quando na mesma camada [Haykin 2015], [Polezer et al. 2018].

A estrutura da ELM é similar à da MLP e apresenta apenas uma camada escondida. A principal diferença é a forma de treinamento da rede. Enquanto a MLP utiliza um processo iterativo, a ELM é feita de maneira analítica e apenas na camada de saída, o que acarreta grande economia de esforço computacional. Além disso, os pesos desta camada são gerados de forma aleatória e permanecem sem ajuste [Polezer et al. 2018].

4. Dados

O estudo de caso foi realizado para a cidade de São Paulo de 01 de janeiro de 2014 até 31 de dezembro de 2016. O tamanho da amostra foi de 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. As variáveis meteorológicas utilizadas consistem naquelas que mais têm influência direta no número de internações [Polezer et al. 2018]. Devido à um padrão existente nos dados de internações, que depende dos dias da semana e o dia ser feriado ou não, é importância considerar estes fatores [Tadano et al. 2016], [Araujo et al. 2020].

Os dados meteorológicos foram obtidos pela Companhia Ambiental do Estado de São Paulo (CETESB) por meio de estações de monitoramento da qualidade do ar [CETESB 2018]. Os dados de internações por doenças respiratórias (CID-10: J00-J99) são dados públicos de acesso irrestrito, por meio do site do DataSUS, não necessitando de aprovação de comitê de ética em pesquisa [DataSUS 2018].

São Paulo é o maior centro urbano da América Latina, com 12.176.866 habitantes e 8 milhões de automóveis circulando pela cidade diariamente, possuindo uma área territorial total de 1521,11 km^2 e 968,32 km^2 de área urbana [IBGE 2019].

O efeito da poluição do ar na saúde pode ocorrer após alguns dias de sua exposição, portanto é comum trabalhar com defasagem de alguns dias. Desta forma, neste trabalho, foram realizadas análises para o efeito da poluição do ar na saúde da população no mesmo dia da exposição (*lag* 0) até 7 dias após a exposição (*lag* 7) [Tadano et al. 2012], [Polezer et al. 2018], [Araujo et al. 2020].

Os primeiros 70% dos dados foram utilizados para treinamento, os 15% subsequente para validação (foi usado o tipo *holdout*) e os 15% restantes para teste. Essa divisão foi escolhida por ser comumente usada na literatura [Silva et al. 2010].

5. Implementação Computacional

As redes foram implementadas utilizando a linguagem de programação Python versão 3.7 [Python 2019]. Para facilitar a criação da MLP com uma e duas camadas escondidas, utilizou-se a biblioteca *sklearn* com a ferramenta MLPRegressor. O algoritmo Broyden-Fletcher-Goldfarb-Shanno (BFGS) foi utilizado para treinamento das MLP [Python 2019]. Os parâmetros utilizados para a rede foram definidos a partir de testes preliminares empíricos e são os seguintes: Quantidade de neurônios em cada camada escondida: 20; Número máximo de iterações: 200; Taxa de aprendizagem: 0,001; Porcentagem de validação: 0,15; Função de ativação: tangente hiperbólica; Valor do erro para o treinamento: 10^{-4}.

Para a implementação da ELM, utilizou-se principalmente a biblioteca *Numpy* [Python 2019], o que facilitou as multiplicações de matrizes e possibilitou a obtenção da matriz inversa generalizada de Moore-Penrose. Para a ELM, também foram usados 20 neurônios na camada escondida e tangente hiperbólica como função de ativação.

6. Resultados e Discussões

Uma forma de realizar uma avaliação comparativa de desempenho é utilizar métricas de erro entre a saída desejada (dados reais) e a saída da rede. Neste trabalho, foram utilizados o Erro Absoluto Médio (do inglês *Mean Absolute Error – MAE*), que diz 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 se

tratar de uma diferença elevada ao quadrado. As respectivas expressões matemáticas podem ser observadas nas equações (1) e (2):

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

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

sendo atribuídos d_t para o valor real, r_t para a saída da rede e N para o número total de dados utilizados na fase de teste.

A Tabela 1 mostra os desempenhos computacionais considerando a melhor de 30 execuções da ELM e MLP com uma (1) e duas camadas (2) escondidas, considerando o *MAE* e o *RMSE* em relação ao conjunto de testes.

Tabela 1. *MAE* e *RMSE* para MLP com uma (1) e duas camadas (2) e ELM.

	MAE			*RMSE*		
	MLP (1)	**MLP (2)**	**ELM**	**MLP (1)**	**MLP (2)**	**ELM**
lag 0	**35,12**	**34,75**	54,95	**43,09**	**42,31**	70,00
lag 1	36,95	37,12	62,39	45,56	44,57	79,32
lag 2	43,23	41,77	63,62	53,47	51,51	78,97
lag 3	39,85	39,19	55,62	48,28	48,02	70,17
lag 4	39,52	38,61	59,07	48,59	47,67	72,76
lag 5	39,89	39,94	59,41	48,17	48,43	73,48
lag 6	40,98	39,14	**52,68**	50,73	48,35	**66,94**
lag 7	40,25	39,84	54,06	49,74	49,16	68,85

A Tabela 1 permite algumas considerações relevantes. Inicialmente, percebe-se que, comparando os resultados de cada defasagem (*lag*), as MLP alcançaram menores erros para *lag* 0, enquanto a ELM no *lag* 6 para ambas as métricas de erro. Além disso, fica evidenciado que o desempenho das MLP foi superior para todas as defasagens. Isto é uma evidência de que a ideia de ajustar todos os pesos da rede leva a resultados mais precisos, devido à alta capacidade de aproximação e generalização da rede, mesmo que com um esforço computacional mais elevado.

Comparando as duas propostas de MLP, percebe-se que na maioria dos casos, a rede de camada escondida dupla alcançou melhores resultados gerais, com destaque para o *lag* 0. Considerando todos as defasagens para observar o poder de mapeamento não-linear, esta arquitetura perde apenas no *lag* 1 e no *lag* 5.

As Figuras 1 e 2 apresentam os melhores resultados das 30 execuções de cada rede neural proposta. Observando os resultados entre os dias 40 e 80 do período de teste, em que foram observados os maiores valores de internações hospitalares, pode-se notar o desempenho mais elevado da MLP com duas camadas escondidas. Alguns estudos epidemiológicos de poluição do ar também apresentaram melhor desempenho com uso da MLP, como Polezer et al. (2018), Araujo et al. (2020). Entretanto, em outros, a ELM apresentou melhores resultados [Tadano et al. 2016]. Estes resultados mostram a importância de aplicação de diferentes métodos, pois o melhor desempenho também está ligado ao problema a ser estudado.

Salienta-se que, este é um problema de mapeamento não-linear estático, em que a saída (internações por doenças respiratórias) pode depender de inúmeras outras variáveis, além das consideradas neste estudo, como idade, histórico familiar, fatores genéticos, entre outros [Araujo et al. 2020], sendo assim, é esperada uma diferença entre os valores ajustados e reais, os quais não coincidem para parte significativa das amostras [Tadano et al. 2016], [Polezer et al. 2018], [Araujo et al. 2020].

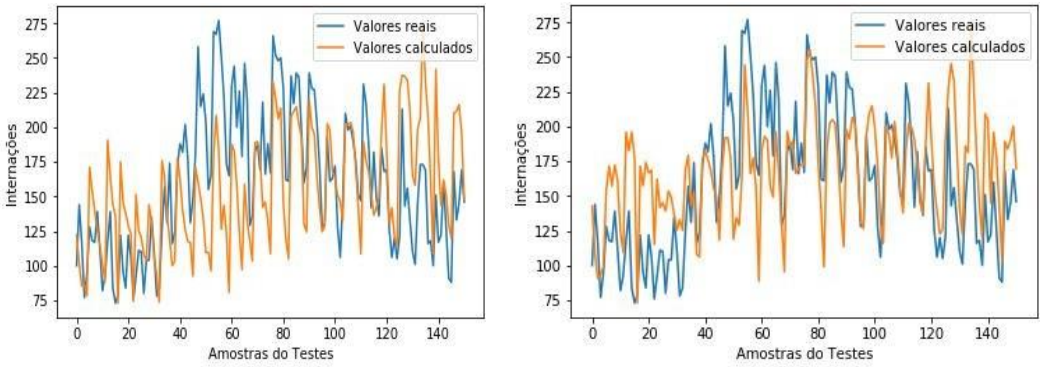

Figura 1. Valores reais *versus* valores calculados para a MLP com uma camada escondida (a) e valores reais *versus* valores calculados para a MLP com duas camadas escondidas (b).

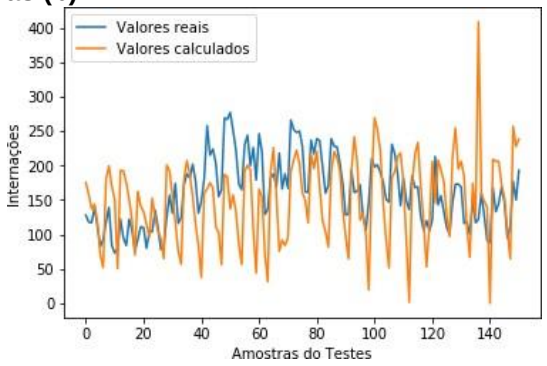

Figura 2. Valores reais *versus* valores calculados para a ELM.

7. Conclusões

Os resultados deste estudo mostram a importância de aplicação de diferentes técnicas computacionais na previsão de impactos da poluição o ar na saúde. As RNA provaram ser uma ferramenta importante de gestão para o sistema de saúde, com a vantagem de realizar previsões de cenários futuros, como um aumento significativo na concentração da poluição do ar. 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]. Trabalhos futuros podem avaliar a concentração de outros poluentes, ssim como testar outras arquiteturas de RNA e, ainda, investigar modelos capazes de prever com maior precisão eventos extremos de poluição do ar e internações.

Como trabalhos futuros, sugere-se a aplicação de RNA recorrentes (*Long Short Term Memory* – LSTM, por exemplo), o uso de outros processos de aprendizagem, como a regularização no ajuste da ELM, métodos de 2ª ordem para ajuste da MLP e validação cruzada tipo *k-fold* para aprimorar a capacidade de generalização das redes.

8. Referências

Araujo, L. N. et al. (2020) Ensemble method based on Artificial Neural Networks to estimate air pollution health risks. *Environmental Modelling and Software*, vol. 123, 104567.

CETESB - Companhia Ambiental do Estado de São Paulo (2021) https://cetesb.sp.gov.br/ar/poluentes/, março.

CETESB - Companhia Ambiental do Estado de São Paulo (2018) https://cetesb.sp.gov.br/ar/qualar/, outubro.

DataSUS - Departamento de Informática do Sistema Único de Saúde (2018) http://www2.datasus.gov.br/DATASUS/%20index.php?area=02, dezembro.

Freitas, A. M. e Solci, M. C. (2009) 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, páginas 1750-1754.

Haykin, S. (2015), Neural Networks and Learning Machines, Prentice Hall, 3rd edition.

IBGE - Instituto Brasileiro de Geografia e Estatística (2019) https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao/9103-estimativas-de-populacao.html?1⁄4&t1⁄4resultados, julho.

OMS - Organização Mundial da Saúde (2020) "Air Pollution", https://www.who.int/health-topics/air-pollution#tab=tab_1, maio.

Python (2019) https://phyton.org, agosto.

Polezer et al. (2018) Assessing the impact of PM 2.5 on respiratory disease using artificial neural networks. *Environmental Pollution*, vol. 235, pages 394-403.

Silva, I. D. et al. (2010), Redes Neurais Artificiais para Engenharia e Ciências Aplicadas, Artliber, 1ª edição.

Tadano, Y. S. et al. (2021) Dynamic model to predict the association between air quality, COVID-19 cases, and level of lockdown. *Environmental Pollution*, vol. 268, 115920.

Tadano, Y. S. et al. (2012) "Methodology to assess air pollution impact on human health using the Generalized Linear Model with Poisson regression", In: Air Pollution – Monitoring, Modelling and Health, Edited by D. M. Khare, Intech, England.

Tadano, Y. S. et al. (2016) Unorganized machines to predict hospital admissions for respiratory diseases. *2016 IEEE Latin American Conference on Computational Intelligence* (*LA-CCI*).