

Método baseado em inteligência computacional para estimativa de internações por doenças respiratórias

Computational intelligence-based method for estimating hospital admissions for respiratory diseases

RESUMO

Diego Solak Castanho
diegocastanho@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil

Hugo Valadares Siqueira
hugosiqueira@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil

Estudos que tratam da saúde humana e que utilizam técnicas inovadoras e avançadas e que possam trazer melhorias na qualidade de vida das pessoas são de fundamental importância e geram impactos em todas as áreas da sociedade. O presente estudo apresenta o processo de calibração de um modelo linear (GLM) por meio desta técnica de inteligência computacional (PSO), que tem por objetivo realizar a previsão do número de internações relacionadas a doenças respiratórias na cidade de Campinas - SP. Como resultado obteve-se um MSE de previsão próximo de 15, que configura uma previsão extremamente relevante quando comparada às técnicas disponíveis para previsão.

PALAVRAS-CHAVE: Previsão Doenças Respiratórias. GLM. PSO.

ABSTRACT

Studies that deal with human health and that use innovative and advanced techniques and that can bring improvements in the quality of life of people are of fundamental importance and generate impacts in all areas of society. This study presents the calibration process of a linear model (GLM) using this computational intelligence technique (PSO), which aims to forecast the number of hospitalizations related to respiratory diseases in the city of Campinas - SP. As a result, a forecast MSE close to 15 was obtained, which represents an extremely relevant forecast when compared to the available forecasting techniques.

KEYWORDS: Respiratory Tract Disease. GLM. PSO.

Recebido: 19 ago. 2020.

Aprovado: 01 out. 2020.

Direito autorial: Este trabalho está licenciado sob os termos da Licença Creative Commons-Atribuição 4.0 Internacional.



INTRODUÇÃO

Questões de saúde relacionadas a doenças respiratórias são assuntos tratados mundialmente pelo grande impacto que têm no cotidiano das pessoas. Estudos que tratam da saúde humana e possam trazer melhorias na qualidade de vida das pessoas são de fundamental importância e vêm sendo pesquisados a décadas (WHO, 2017).

O Modelo Linear Generalizado (GLM) apresentado pela primeira vez em 1972, é uma classe capaz que representa respostas contínuas entre uma ou mais variáveis de previsão em uma grande variedade de respostas que não são distribuídas de forma normal (LAZZARIN, 2019), (MCCULLAGH, 1983).

No GLM, os coeficientes livres podem ser estimados a partir de uma base de dados com auxílio de uma técnica de inteligência computacional, como a otimização por enxame de partículas (PSO do inglês, Particle Swarm Intelligence).

Este é inspirado em teorias sociopsicológicas, que ao contrário de algoritmos evolutivos, mantém suas partículas durante todo o processo de otimização, possuindo uma espécie de “memória” que lhe possibilita chegar a solução desejada (CLERC, 2002), (KENNEDY, 2010).

O presente estudo mostra a processo de calibração de um modelo linear por meio desta técnica de inteligência computacional, que tem por objetivo realizar a previsão do número de internações relacionadas a doenças respiratórias na cidade de Campinas - SP.

METODOLOGIA

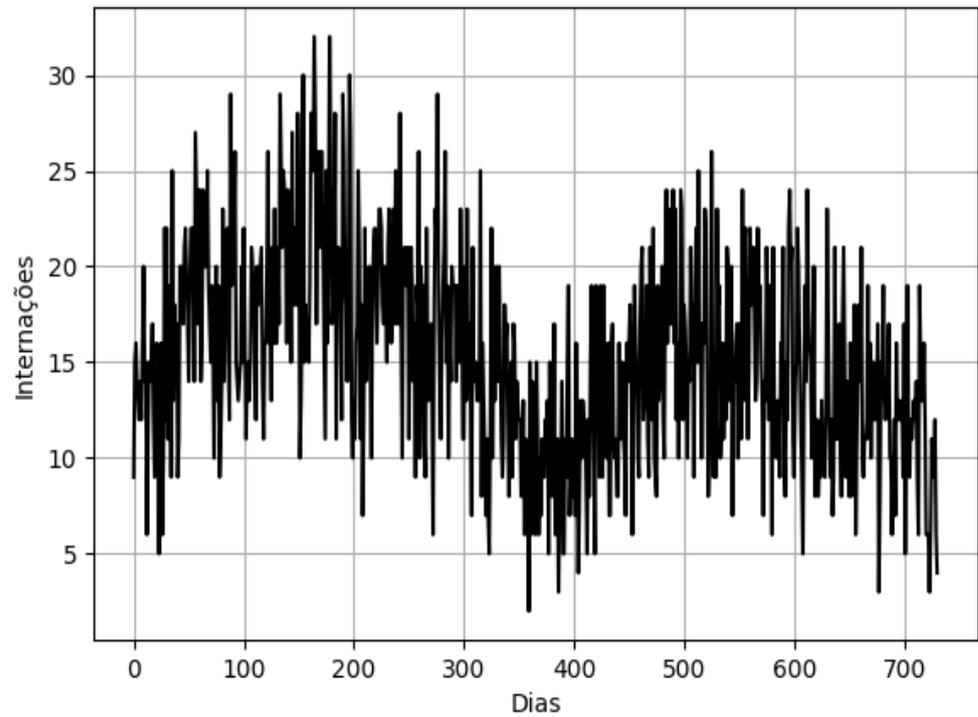
A metodologia consiste primeiramente na realização de um estudo de caso, envolvendo a estimação do número de internações hospitalares relacionadas ao contato com material particulado (MP10) para a cidade de Campinas, localizada no estado de São Paulo.

Como variáveis de entrada foram consideradas temperatura média, umidade relativa do ar, dia da semana em formato binário, e se o dia em que a internação registrada era feriado.

A base de dados de Campinas empregue para ajuste e teste dos modelos de estimação possui um total de 732 amostras. Os dados foram coletados em 2007 e 2008, em que foram identificados um total de 11.121 internações.

Para calibração do modelo foram utilizados 85% dos valores presentes na base de dados e para teste utilizou-se os 15% restantes. A Figura 1 apresenta as internações ocorridas a cada dia presentes na base de dados, que inicia-se em janeiro de 2007 e encerra-se em dezembro de 2008.

Figura 1 – Número de internações por doenças respiratórias em 2007 e 2008 (Campinas - SP).



Fonte: Autoria própria (2020)

Os dados das entradas relacionados às internações por doenças respiratórias foram obtidos através da base do Sistema Único de Saúde Brasileiro (DATASUS, 2009). As entradas referentes a poluição atmosférica e temperatura ambiente foram obtidos junto a Companhia Ambiental do estado de São Paulo (DATASUS, 2009).

MÉTRICAS DE AVALIAÇÃO E FUNÇÃO CUSTO

Neste trabalho foi considerado um modelo linear, o GLM, em conjunto com um algoritmo bio-inspirado, o PSO, que tem por objetivo realizar o ajuste dos respectivos coeficientes do modelo. A métrica de avaliação utilizada foi o MSE, do inglês, *Mean Square Error*, quanto menor o valor do MSE, mais próxima a curva de treinamento está da curva real de treinamento. A Equação 1 apresenta o cálculo do MSE.

$$MSE = \frac{1}{N} \sum_{t=1}^N (d_t - y_t)^2 \quad (1)$$

Para obtenção da função fitness basta dividir 1 sobre o valor de MSE, quanto mais próximo de 1, melhor o resultado da otimização, como apresentado na Equação 2.

$$fitness = \frac{1}{MSE} \quad (2)$$

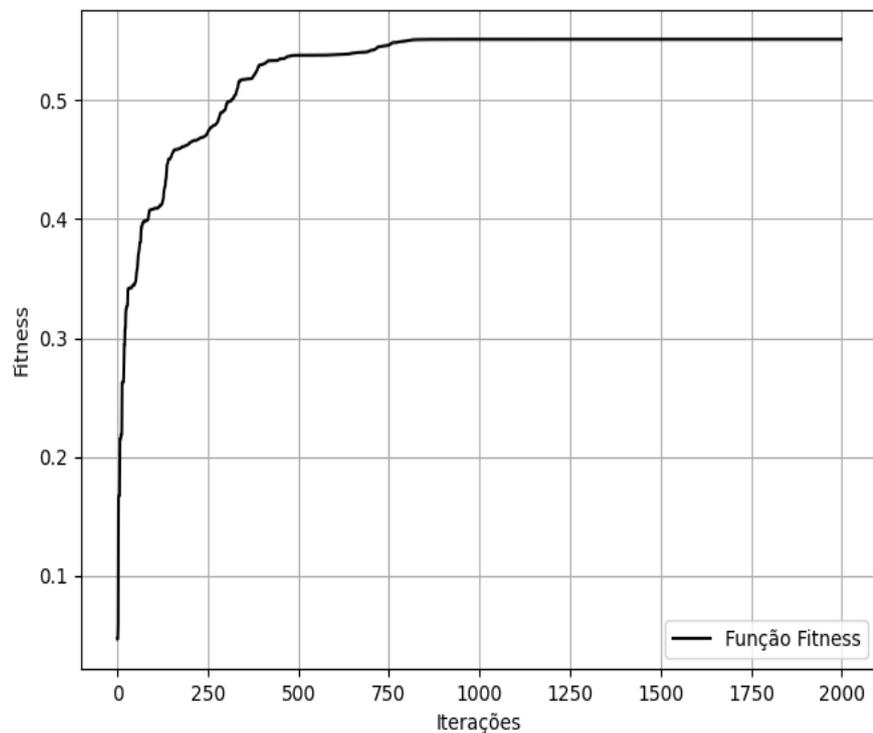
CONFIGURAÇÃO DA SIMULAÇÃO

A fim de se obter uma amostragem estatística de desempenho, o algoritmo foi executado 30 vezes, de modo que a melhor execução foi a escolhida. Para a primeira variação, o cálculo de velocidade usado foi o clássico, com a constante de inércia (W) fixa em 0,8 e as constantes social e cognitiva em 2.

RESULTADOS E DISCUSSÕES

A Figura 2 apresenta a evolução da função fitness ao longo do processo de otimização para 1 dia de atraso. A função de fitness é obtida a partir da divisão de 1 sobre o MSE, em que quanto mais próximo de 1 melhor é o resultado da otimização.

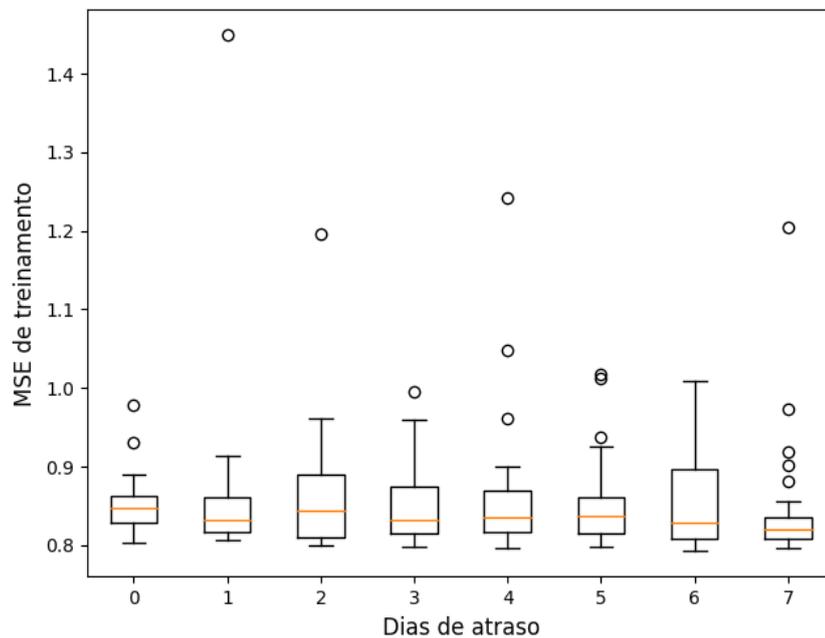
Figura 2 – Evolução da função fitness GLM-PSO



Fonte: Autoria própria (2020)

A Figura 3 apresenta um *boxplot* que apresenta a dispersão obtida nos treinamentos para os 8 dias após a exposição aos poluentes. Como se observa, o pior resultado ocorreu no com um dia de atraso, em que em uma das iterações o algoritmo chegou a um MSE de treinamento próximo a 1,4. Pode-se perceber que em todos os dias de atraso, obteve-se o melhor resultado próximo a 0,8.

Figura 3 – BoxPlot função custo (MSE) 30 iterações, para 8 dias de atraso.



Fonte: Autoria própria (2020).

A Quadro 1 apresenta os resultados obtidos a partir do processo de otimização e os resultados da previsão. Os dias de atraso são apresentados na primeira coluna, em seguida são apresentados os valores relativos ao treinamento para os 7 dias de atraso, quanto mais próximo de zero, melhor foi o processo de otimização. Em todos os dias, a geração de convergência foi próxima da geração limite definida. Também, obteve-se um MSE de previsão próximo de 15, MAE de previsão próximo de 3 e uma MAPE próximo a 35.

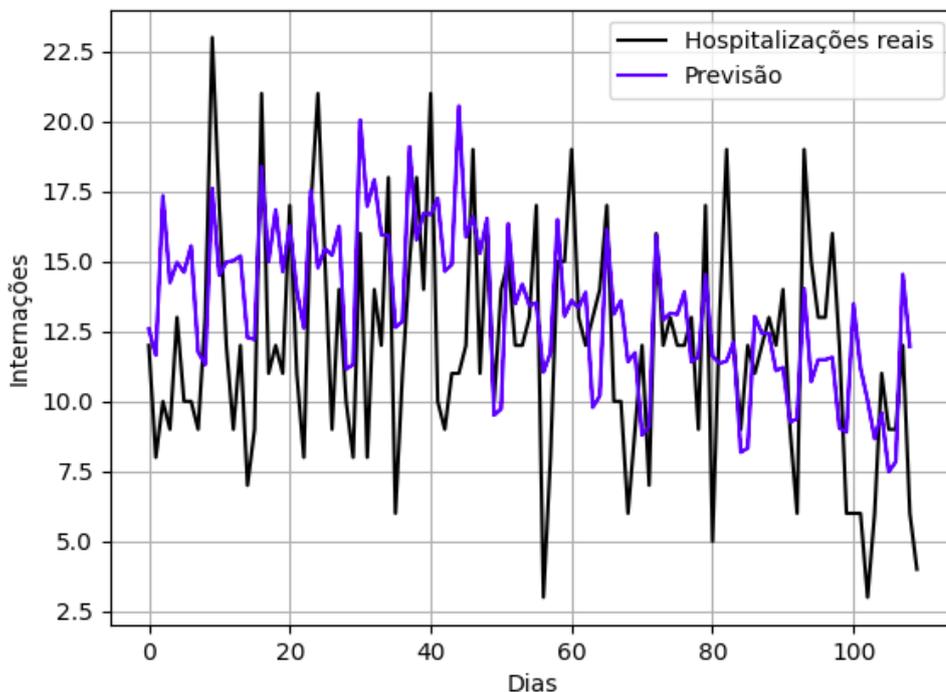
Quadro 1 – Resultados da simulação

Dia de Atraso	Melhor MSE Treinamento	Geração de Convergência	MSE Previsão	MAE previsão	MAPE previsão
0	0,82	1999	15,10	3,25	35,91
1	0,81	1999	15,32	3,25	36,04
2	0,80	1999	15,18	3,24	36,28
3	0,80	1997	15,34	3,23	36,07
4	0,80	1999	15,25	3,26	36,18
5	0,84	1999	15,20	3,24	35,91
6	0,80	1999	14,80	3,22	35,75
7	0,81	1999	14,74	3,22	35,53

Fonte: Autoria própria (2020).

A Figura 4 apresenta o resultado gráfico da previsão para 1 dia de atraso, em que a linha preta representa o número real de hospitalizações e a linha em azul apresenta a previsão obtida pelo GLM-PSO.

Figura 4 – Melhor previsão para 1 dia de atraso GLM-PSO



Fonte: Autoria própria (2020).

CONCLUSÕES

No presente estudo, foram estimados o número de internações relacionadas a doenças respiratórias ocorridas entre 2007 e 2008 na cidade de Campinas, localizada no estado de São Paulo. Esta pesquisa correlacionou a concentração de material particulado, fatores meteorológicos, temperatura média, dia da semana e o se o dia era feriado ou não, realizando a previsão de internações relacionadas a doenças respiratórias por até 7 dias após exposição aos poluentes.

Como trabalhos futuros, pretende-se testar mais variações de algoritmos de otimização, somado a variações do próprio PSO. Mais modelos poderiam ser testados, bem como diferentes métricas de função custo, que poderiam melhorar significativamente os resultados das previsões.

AGRADECIMENTOS

Os autores agradecem a UTFPR, ao CNPQ e a CAPES pelo apoio no desenvolvimento deste projeto.

REFERÊNCIAS

CETESB, Companhia de Tecnologia de Saneamento Ambiental. **Qualidade do ar no estado de São Paulo** 2007, 2008 e 2009.

CLERC, M.; KENNEDY, J. **The particle swarm-explosion, stability, and convergence in a multidimensional complex space.** IEEE transactions on Evolutionary Computation, v. 6, p. 58–73, 2002.

DATASUS, Departamento de Informática do Sistema único de Saúde. **Informações de Saúde.**

KENNEDY, J. **Particle swarm optimization.** In: SAMMUT, C.; WEBB, G. I. (Ed.). Encyclopedia of Machine Learning. [S.l.]: Springer US, 2010. p. 760–766

LAZZARIN, L. A. et al. **Método ensemble baseado em redes neurais artificiais para estimação de interações por doenças respiratórias.** 2019. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2019.

MCCULLAGH, P. Quasi-likelihood functions. **The Annals of Statistics**, v. 11, n. 1, p. 59–67, 1983.

WHO. **Evolution of WHO Air Quality Guidelines: Past, Present and Future.** WHO Regional Office for Europe, Copenhagen (2017). 2017.