Predição temporal e espaço-temporal dos parâmetros da qualidade da água

Anderson Almeida[1], **Marcos Amaris**[1], **Bruno Merlin**[1]

[1]Universidade Federal do Pará (UFPA) - Tucuruí - PA - Brasil

{andersonchico, marcos.amaris,bruno.merlin}@gmail.com

Abstract. *The quality of water is directly related to its level of pollution caused by anthropic and industrial actions. Therefore, limnological monitoring of the basic parameters of water quality is carried out, as a way of obtaining data that guide the decision-making of water resources management bodies. In this context, this study aims to analyze datasets and the performance of linear regression, random forest, MLP and LSTM neural networks algorithms in the temporal and space-time prediction. Models are evaluated using MAPE and RMSE metrics. Therefore, in temporal prediction, the LSTM technique presents the lowest average MAPE, 4.66% and the MLP the lowest average RMSE, 2.47. However, in the spatio-temporal prediction, the MLP has the lowest average result of MAPE and RMSE, respectively, 5.94% and 1.34.*

Resumo. *A qualidade da água está diretamente relacionada com o seu nível de poluição causada pelas ações antrópicas e industriais. Por isso, são realizados os monitoramentos limnológicos dos parâmetros básicos da qualidade da água, como forma de obtenção de dados que norteiam as tomadas de decisão dos órgãos gestores de recursos hídricos. Neste contexto, o presente estudo tem o objetivo de analisar o conjunto de dados e o desempenho dos algoritmos regressão linear, random forest, redes neurais MLP e LSTM na predição temporal e espaço-temporal. Os modelos são avaliados através das métricas MAPE e RMSE. Portanto, na predição temporal a técnica LSTM apresenta o menor MAPE médio, 4.66% e o MLP o menor RMSE médio, 2.47. Porém, na predição espaço-temporal, o MLP tem o menor resultado médio de MAPE e RMSE, respectivamente, 5.94% e 1.34.*

1. Introdução

A água é um recurso natural essencial para o consumo humano, para as atividades industriais e agrícolas, assim como para os ecossistemas vegetal e animal. A oferta de água é determinada pela dinâmica hídrica e socioeconômica das bacias dos rios, além das condições da qualidade da água. E o conhecimento dessa oferta é produzido pelo monitoramento limnológico que coleta periodicamente dados dos parâmetros básicos da qualidade da água como potencial hidrogeniônico (pH), temperatura, oxigênio dissolvido (OD), turbidez e condutividade elétrica em rios e lagos, permitindo o planejamento e a gestão dos recursos hídricos [Marotta et al. 2008]. Desse modo, as análises preditivas sobre uma serie temporal contribui para a qualidade da tomada de decisão [Solanki et al. 2015], por isso, o aprendizado de máquina, que é uma sub-área da inteligência artificial aplicada que trata de análises preditivas para adquirir conhecimento de

forma automática [Monard and Baranauskas 2003]. Assim, este artigo pretende-se analisar o conjunto de dados e o desempenho dos algoritmos regressão linear, random forest, das redes neurais *Long-Short Term Memory* (LSTM) e *Multilayer Perceptron* (MLP) nas predições temporal e espaço-temporal dos parâmetros da qualidade de água, avaliando o ajuste dos modelos através das médias de MAPE e RMSE. Portanto, este trabalho visa contribuir com o processo de monitoramento da qualidade das águas, avaliando as suas tendências e diagnósticos, tornando-se um subsídio para o planejamento da gestão hídrica em manter o equilíbrio entre o desenvolvimento econômico, demográfico e a disponibilidade hídrica para diversos tipos de usos. Além disso, as modelagens e as predições podem suprir as demandas de áreas onde o monitoramento não é viável [Fu et al. 2019].

2. Conceitos

A qualidade da água têm características físicas, químicas e biológicas, representadas por parâmetros, de acordo com o valor padrão estabelecido em lei para cada finalidade de uso [Derisio 2016]. Na regressão linear a saída de um problema é um número com erro de predição ajustado. O Random Forest consiste no treinamento de árvores de decisão considerando a média dos resultados das árvores para melhorar a predição e o controle do sobreajuste [Breiman 2001] O MLP possui no mínimo uma camada oculta entre as camadas de entrada e saída, com treinamento supervisionado [Da Silva et al. 2010]. O LSTM tem o objetivo de reparar a falta de memória de longo prazo da rede neural, retendo as principais informações dos sinais de entrada [Bandara et al. 2020], persistindo essas através dos laços de repetição.

3. Metodologia

Neste artigo são utilizados dados dos monitoramentos dos parâmetros da qualidade da água nos pontos de coleta TIET02050 e TIET04150 na UGRHI 06 - ALTO TIÊTE, localizadas em áreas industrializadas [1] do Estado de São Paulo, referente ao período de 1978 à 2019. Para tanto, na seleção dos dados, é identificada a UGRHI com maior quantidade de amostras, após isso, são destacados os seus pontos de coleta com maior quantidade de amostras e menor quantidade de valores ausentes, nesta ordem, pois estes podem interferir na qualidade dos resultados.

Além disso, há a análise do resumo estatístico e visualização de *outliers*. No pré-processamento os dados são: transformados para a frequência mensal, limpos, tratados os valores ausentes, transformados em logarítmico e normalizados. Os modelos das redes neurais têm 100 épocas de iteração, monitoramento da métrica de erro (*EarlyStopping*), função de ativação *ReLu* nas camadas facilitando o treinamento. Também, há 50 neurônios nas camadas de entrada e intermediária, pois têm os melhores desempenhos na predição temporal desses parâmetros [Almeida et al. 2020], e 1 neurônio saída. A camada intermediária no LSTM é implementada com *Dropout* para evitar o ajuste excessivo (*overfitting*), compilado com a métrica MAE e otimizador *adam* para ajustar os pesos de acordo com MAE. Os dados são divididos em 70% para treinar e 30% para testar.

Ademais, os modelos têm um tempo (lags) definido para predizer o valor do parâmetro e determinado pela quantidade de observações imediatamente anteriores ao

[1]CETESB (2019). Infoáguas .https://sistemainfoaguas.cetesb.sp.gov.br/. [Online; De-zembro 7 de 2019]

valor predito, transformando a série histórica em problemas de aprendizado supervisionado para a tarefa de regressão. Desse modo, são atribuídas de 1 a 10 observações como lags e nas redes neurais, que são técnicas não determinísticas, são obtidas as médias dos resultados de 5 simulações para cada lags. Por fim, os resultados são comparados com a média de MAPE e RMSE dos baselines a partir das séries temporais dos parâmetros.

4. Resultados

A Tabela 1 apresenta o resumo estatístico dos pontos de coleta mais adequados para os experimentos. Assim, a maioria dos valores estão distribuídos assimetricamente, exceto os valores de pH. Além disso, nos pontos, há uniformidade nos valores de pH, DBO, sólido e temperatura, diferentemente dos valores do coliformes e fósforo. Por fim, os valores de turbidez estão uniformes apenas no ponto TIET04150. Nos valores de fósforo no ponto TIET02050 e DBO no ponto TIET04150 não têm *outliers*.

Na predição temporal com os dados de TIET02050, dos 8 parâmetros preditos, apenas DBO, sólido, temperatura e pH têm MAPE médio $<= 15\%$, conforme a Figura 1, e que com 2 lags há o menor MAPE médio. Nos desempenhos das técnicas por parâmetros, o LSTM destaca-se nas predições do DBO (2.33%), temperatura (8.08%) e pH (2.89%) de MAPE médio, enquanto, o MLP tem o melhor MAPE médio na predição do Sólido Total (4.98%). No desempenho por técnica, o LSTM possui o menor MAPE médio e o MLP tem o menor RMSE médio, conforme a Tabela 2.

Na predição espaço-temporal com os dados de TIET02050 e TIET04150, dos 8 parâmetros preditos, somente temperatura e pH têm MAPE médio $<= 15\%$, conforme a Figura 1, visto que, na predição da temperatura, isto ocorre a partir do 3º lags. Diante disso, com as técnicas o menor MAPE médio é com 10 lags e com Baseline o menor MAPE médio é com 5 lags. Nos desempenhos das técnicas por parâmetros, o MLP destaca-se nas predições da temperatura (9.49%) e pH (1.94%) de MAPE médio. No desempenho por técnica, o MLP têm os melhores resultados médio de MAPE e RMSE, Tabela 2.

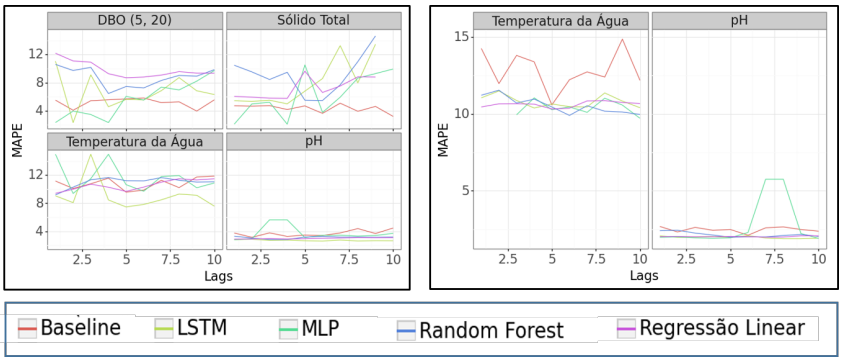
Tabela 1. Resumo estatístico dos dados dos pontos de coleta.

	TIET02050						TIET04150					
Parameters	**Samples**	**Mean**	**STD**	**Min**	**50%**	**Máx**	**Samples**	**Mean**	**STD**	**Min**	**50%**	**Máx**
Coliformes	171.00	447.05	1685.88	0.00	36.00	17000.00	170.00	1053989.23	3483274.16	1.00	270000.00	32033833.33
DBO	206.00	3.11	2.11	0.10	3.00	19.00	206.00	25.17	20.57	2.00	18.00	93.00
Fósforo	208.00	0.06	0.06	0.00	0.05	0.49	208.00	0.86	0.97	0.04	0.52	8.70
OD	208.00	4.62	1.92	0.00	4.87	10.20	197.00	0.74	0.88	0.00	0.40	4.50
pH	208.00	6.38	0.39	5.40	6.38	7.72	207.00	6.77	0.50	3.70	6.90	7.50
Sólido	208.00	69.08	55.82	8.00	56.00	684.00	208.00	350.07	160.57	100.00	325.50	1490.00
Temperatura	208.00	20.91	2.37	15.00	21.00	29.00	207.00	21.39	2.68	14.00	21.90	27.00
turbidez	206.00	8.56	10.45	0.00	5.04	70.00	204.00	39.49	38.74	0.37	29.00	270.00

Tabela 2. Desempenho médio por técnica.

	Temporal		Espaço-temporal	
Técnicas	**MAPE(%)**	**RMSE**	**MAPE(%)**	**RMSE**
Baseline	5.49	3.48	6.53	1.62
LSTM	**4.66**	2.49	6.42	1.50
MLP	5.32	**2.47**	**5.94**	**1.34**
Random Forest	8.14	3.32	6.48	1.50
Regressão Linear	7.48	2.71	6.34	1.47

Figura 1. MAPE $<= 15$% nas predições à esq. temporal e a dir. espaço-temporal.



5. Conclusões

Neste artigo, os dados dos pontos de coleta TIET02050 e TIET04150 da UGRHI 06 - ALTO TIÊTE são mais adequados para as predições, caracterizados como assimétricos na distribuição dos valores, possui uniformidade e *outliers*. O LSTM e MLP são adequados na predição temporal, dependendo da métrica erro. Porém, na predição espaço-temporal, o MLP é mais adequado, independente da métrica de erro. Por fim, os resultados das predições podem viabilizar a implementação de um Sistema de Apoio à decisão (SAD) para auxiliar a tomada de decisão dos gestores de recursos hídricos.

Referências

Almeida, A., Amaris, M., Merlin, B., and Veras, A. (2020). Modelagem e predição temporal de parâmetros de qualidade de água usando redes neurais profundas. In *Anais do XI Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 121–130. SBC.

Bandara, K., Bergmeir, C., and Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, 140:112896.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Da Silva, I. N., Spatti, D. H., and Flauzino, R. A. (2010). *Redes NEurais Artificiais para Engenharia e Ciências Aplicadas: Curso Prático*. Artliber.

Derisio, J. C. (2016). *Introdução ao controle de poluição ambiental*. Oficina de Textos.

Fu, B., Merritt, W. S., Croke, B. F., Weber, T. R., and Jakeman, A. J. (2019). A review of catchment-scale water quality and erosion models and a synthesis of future prospects. *Environmental modelling & software*, 114:75–97.

Marotta, H., Santos, R. O. d., and Enrich-Prast, A. (2008). Monitoramento limnológico: um instrumento para a conservação dos recursos hídricos no planejamento e na gestão urbano-ambientais. *Ambiente & sociedade*, 11(1):67–79.

Monard, M. C. and Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, 1(1):1.

Solanki, A., Agrawal, H., and Khare, K. (2015). Predictive analysis of water quality parameters using deep learning. *International Journal of Computer Applications*, 125(9):0975–8887.