MODELAGEM DE PARÂMETROS DE UM SISTEMA DE LAGOAS

Miguel Feliciano Mota Alves[1]

Nathalia Seixas Trotte Motta[2]

Resumo

O presente trabalho consiste na apresentação de uma modelagem de dados de uma lagoa de estabilização. Após uma análise exploratória de dados extraídos tanto na entrada da lagoa quanto na saída, é apresentado, pelo método de regressão linear múltipla, uma possibilidade de prever as Demandas químicas e bioquímicas do Oxigênio a partir de outros fatores medidos, como pH, temperatura, amônia, entre outros.

Palavras-chave: Demanda Bioquímica de Oxigênio (DBO); Demanda Química de Oxigênio (DQO); dados; análise exploratória; outliers; modelagem.

Introdução

As lagoas de estabilização são excelentes meios de tratamento biológico. A partir de processos naturais, compostos orgânicos putrescíveis são transformados em compostos orgânicos mais estáveis ou minerais. Nesse sentido, o controle e a supervisão dos parâmetros dela é essencial para um bom funcionamento e longevidade desse sistema.

Fatores como Demanda Bioquímica de Oxigênio (DBO), Demanda Química de Oxigênio (DQO), pH, quantidade de nutrientes e a temperatura são, então, importantíssimos para isso. Entretanto, o tempo e os recursos necessários para uma análise detalhada de cada um desses parâmetros são grandes.

Principalmente quando se é levado em consideração a DBO e a DQO, as quais dependem bastante desses outros indicadores e os mecanismos utilizados para conseguir os dados são ainda mais complexos.

Com isso, a predição das variáveis através da modelagem permite uma previsibilidade desses componentes por meio de uma análise estatística contribuindo para a tomada de decisões de maneira mais efetiva e otimizada.

Objetivos

Nesse contexto, esse trabalho tem como propósito a obtenção de um modelo de predição de DBO e de DQO por meio de métodos lineares.

[1] Universidade Federal da Bahia (UFBA), miguelfma23@gmail.com

[2] Universidade Federal da Bahia (UFBA), nathaliatrotte@gmail.com

Material e Método

Para tanto, dados - os quais estão disponíveis no Anexo - de um sistema de lagoas aeradas de uma Indústria de Papel e Celulose (P & C) foram utilizados. Nele, alguns parâmetros dividem-se entre entrada (in) e saída (out) e são:

1. Datas: $Date_{in}$ e $Date_{out}$;
2. DBO_{in} e DBO_{out};
3. DQO_{in} e DQO_{out};
4. Vazão do Efluente: FR_{in} e FR_{out};
5. Sólidos em Suspensão: SS_{in} e SS_{out};
6. pH;
7. Nitrogênio Amoniacal: NAm;
8. Nitrogênio Nitrato: NN;
9. Fósforo: P;
10. Cor: Col_{in} e Col_{out};
11. Condutividade: $Cond_{in}$ e $Cond_{out}$;
12. Temperatura: T_{in} e T_{out};
13. Precipitação: RF;
14. Produção de Papel: Pap;
15. Produção de Celulose: Pulp;

Em um primeiro momento, foi feita a análise exploratória, a qual consistiu em excluir colunas vazias e identificar o tamanho do documento. Após isso, alterou-se alguns dos tipos de variáveis a fim de ajustar-se ao formato. Obteve-se, em seguida, um resumo estatístico dos dados.

Em um segundo momento, o método gráfico *box plot* foi usado para a verificação de outliers e, com isso, poderem ser removidos. Foi reconhecido, porém, que havia muitas variáveis sem informação. Por conseguinte, a técnica de interpolação linear foi aplicada para estipular esses fatores e valores que não podiam ser ajustados foram excluídos da manipulação.

Com os dados tratados, pôde-se começar com a modelagem do sistema. Assim como já está disposto, os parâmetros foram divididos em entrada e saída. Primeiro, lidou-se com um modelo com todas as variáveis e esse número foi diminuindo com intuito de encontrar uma predição mais simples e com menor viés (*bias*).

Por fim, construiu-se uma tabela com os erros de cada um dos modelos. Mais especificamente, o coeficiente de determinação (R^2), o erro quadrático médio (MSE) e o erro absoluto médio (MAE).

Resultados e Discussão

A priori, foi feita uma análise exploratória dos dados, onde cada variável foi submetida a uma remoção de outliers utilizando o método *box plot*. No mais, a base de dados bruta possui muitos valores faltantes, necessitando também de um tratamento.

Através da interpolação, é possível estimar esses valores faltantes. Primeiramente, foi utilizado uma interpolação polinomial, contudo os dados não responderam bem. Por isso, foi decidido utilizar a interpolação linear. Assim, os dados podem ser comparados através das seguintes tabelas: Tabela 1 - Dados Brutos. Tabela 2 - Dados Tratados.

Tabela 1

Resumo Estatístico dos Parâmetros antes da remoção de outliers e preenchimento dos valores faltantes

Parâmetros	Quantidade	Média	Mínimo	Máximo	Valores Vazios (%)
$Date_{in}$	1430	–	–	–	–
$Date_{out}$	1428	–	–	–	–
DBO_{in}	1341	245.12	41	449	6.22
DBO_{out}	1343	85.19	16	187	6.08
DQO_{in}	1341	561.39	136	925	6.22
DQO_{out}	1345	315.44	105	865	5.94
FR_{in}	1430	67358.61	4474	97850	0
FR_{out}	1428	67657.37	4474	106942	0.14
SS_{in}	568	149.20	12	591	60.28
SS_{out}	192	52.16	10	130	86.57
pH	1377	7.45	0.85	12.53	3.71
NAm	660	2.45	0	20	53.85
NN	279	1.43	0.03	7.39	80.49
P	260	1.41	0.05	14	81.82
Col_{in}	1379	464.40	41	1317	3.57
Col_{out}	1387	418.20	32	1138	3.01
$Cond_{in}$	1374	1530.46	79	5810	3.91
$Cond_{out}$	1385	1550.84	222	2890	3.15
T_{in}	963	45.45	28	50.5	32.66
T_{out}	1234	34.70	25	39	13.71

RF	1175	4.82	0	175.40	17.83
Pap	1338	1042.71	382.4	1304.80	6.43
Pulp	1327	884.74	-48	1112.09	7.20

Tabela 2

Resumo Estatístico dos Parâmetros depois da remoção de outliers e preenchimento dos valores faltantes

Parâmetros	Quantidade	Média	Mínimo	Máximo	Valores Vazios (%)
DBO_{in}	625	239.35	150	326	0
DBO_{out}	624	85.13	26.72	141	0.16
DQO_{in}	625	552.75	374	752	0
DQO_{out}	624	313.32	202	418	0.16
FR_{in}	625	69374.31	51309	86115	0
FR_{out}	624	69415.61	51309	86509	0.16
SS_{in}	617	113.08	16	286	1.28
SS_{out}	617	52.81	10	93	1.28
pH	625	7.03	6.26	7.72	0
NAm	614	1.88	0	5.2	1.76
NN	623	1.29	0.03	3.28	0.32
P	613	1.17	0.47	1.9	1.92
Col_{in}	625	490.69	306	684	0
Col_{out}	624	465.22	312	676	0.16
$Cond_{in}$	625	1433	885	1945	0
$Cond_{out}$	624	1478.72	1062	2860	0.16
T_{in}	623	45.93	41	50	0.32
T_{out}	623	34.95	30	39	0.32
RF	625	1.34	0	10	0.32
Pap	625	1069.17	924.1	1219	0
Pulp	625	917.13	762.87	1046.03	0

Mesmo com a remoção de outliers e a interpolação linear, algumas linhas continuaram a apresentar valores vazios, e por isso foram retiradas do dataframe [Linha de código 55], totalizando 613 linhas de dados de cada fator.

Após as tratativas, utilizando o modelo de regressão linear múltipla, foram realizadas algumas modelagens para o melhor entendimento do desempenho do modelo. A primeira modelagem, para o DQO, foi realizada considerando todas as variáveis preditoras (excluindo o DBO). Depois, o modelo foi treinado com as variáveis preditoras (excluindo o DBO), exceto Pap, FR, Cond. Na modelagem seguinte, além da retirada das variáveis anteriores, foram retiradas também as variáveis SS, Pulp, Col. Na quarta modelagem, foram desconsideradas as variáveis NAm, NN, P, SS, RF, Col, Cond.

Já para a modelagem do DBO, a primeira modelagem foi feita considerando somente o DQO. A segunda modelagem foi realizada considerando todas as variáveis preditoras (incluindo o DQO). A terceira modelagem para o DQO foi realizada com as variáveis preditoras, exceto SS, P, NN, NAm. A quarta modelagem para o DQO foi feita desconsiderando SS, P, NN, NAm, RF, Col, Cond. A quinta modelagem foi feita retirando SS, P, NN, NAm, Col, Cond e Pap.

Para uma melhor análise do resultado desses modelos, as tabelas 3 e 4 foram construídas, trazendo as informações do R^2, MAE, MSE nas colunas e cada modelagem nas linhas. A tabela 3 refere-se ao DQO(in and out) e a tabela 4 ao DBO(in and out).

Tabela 3 - Resumo estatístico dos modelos para DQO (in and out)

Número do modelo	R^2	MAE	MSE
1(in)	0.53181	32.952	1721.22
1(out)	0.4999	21.14299	686.490
2(in)	0.499773	32.995945	1849.67978
2(out)	0.474812	21.637416	720.990589
3(in)	0.298688	39.890694	2593.226415
3(out)	0.226332	25.811161	1062.111174
4(in)	0.341315	37.77328	2435.606586
4(out)	0.194487	26.332505	1105.828315

Tabela 4 - Resumo estatístico dos modelos para DBO (in and out)

Número do modelo	R^2	MAE	MSE
1(in)	0.039858	22.675368	835.534459
1(out)	0.074979	12.59911	246.581025

2(in)	0.202142	20.759488	694.312175
2(out)	0.300106	10.856396	186.569423
3(in)	0.195332	20.462396	700.237739
3(out)	0.301501	11.032444	186.197515
4(in)	0.192572	20.508698	702.639891
4(out)	0.291142	11.228111	188.958845
5(in)	0.173639	20.58716	719.115346
5(out)	0.294498	11.117818	188.064347

Durante a execução do projeto, foi desconsiderada a exclusão de outliers para o pH, o que gerou uma grande diferença na resposta do modelo (tabelas 5 e 6 representam os erros sem a exclusão de outliers para o pH). Após perceber que essa variável não havia recebido o tratamento adequado, foi realizada a remoção dos outliers para a mesma. Antes o número de linhas do Data Frame tinha sido de 726 sem remoção de outliers do pHin.

Com o tratamento adequado para o pHin, o Data Frame ficou com 613. Ademais, o R^2 para o modelo 1 do DQO, sem tratamento de outliers para o pH foi de 47%. Com o devido tratamento, o modelo apresentou um R^2 de 53%. Isso confirma que tão importante quanto modelar, é necessário realizar a análise exploratória dos dados de maneira correta.

Tabela 5 - Resumo estatístico dos modelos para DQO sem a exclusão de outliers do pH (in and out)

Número do modelo	R^2	MAE	MSE
1(in)	0.473513	37.741529	2252.456473
1(out)	0.326904	23.350246	879.306795
2(in)	0.418905	39.285855	2486.085481
2(out)	0.344683	23.310738	856.08117
3(in)	0.217786	45.473331	3346.526412
3(out)	0.09892	27.674449	1177.136121
4(in)	0.338021	41.859127	2832.126798
4(out)	0.060021	27.964049	1227.952744

Tabela 6 - Resumo estatístico dos modelos para DBO sem a exclusão de outliers do pH (in and out)

Número do modelo	R^2	MAE	MSE
1(in)	0.047725	26.058726	1080.245551
1(out)	0.070978	14.008729	296.633093
2(in)	0.223346	24,020996	881.024358
2(out)	0.319346	11.593473	217.330179
3(in)	0.210632	24.171912	895.446821
3(out)	0.304944	11.663762	221.928783
4(in)	0.229827	23.991937	873.672409
4(out)	0.313623	11.646951	219.157555
5(in)	0.212079	24.164848	893.805003
5(out)	0.308819	11.729604	220.691594

Para o modelo do presente trabalho, a análise temporal foi desconsiderada, pois os dados de entrada foram separados dos dados de saída e criou-se uma predição para cada etapa em específico. Sendo assim, as informações de data foram implicitadas nas variáveis, afinal cada valor de uma variável é obtido em um dia diferente.

Conclusão

Percebe-se então que os modelos treinados com todas as variáveis preditoras possuíram um melhor desempenho, pois o modelo 1 para o DQO (todas as variáveis) apresentou o maior R^2, assim como o modelo 2 para o DBO (todas as variáveis), apresentou o maior R^2. Isso implica dizer que a diminuição de parâmetros no modelo acarreta um maior viés-variância.

Desta maneira, encontra-se quatro modelos preditivos. As suas funções são demonstradas abaixo, sendo:

- A função 1 para o DQO_{in};
- A função 2 para o DQO_{out};
- A função 3 para o DBO_{in};
- A função 4 para o DBO_{out};

$$DQO_{in} = 459 + 3.41 \,.\, NAm + 3.74 \,.\, NN + 21.37 \,.\, P + 0.21 \,.\, SS_{in} - 2.63 \,.\, T_{in} - 0.05 \,.\, Pap -$$
$$- 0.002 \,.\, FR_{in} + 0.200 \,.\, Pulp - 1.973 \,.\, RF + 0.344 \,.\, Col_{in} - 6.506 \,.\, pH + 0.036 \,.\, Cond_{in} \quad (1)$$

$$DQO_{out} = 100.8 + 0.5 \,.\, NAm + 2.29 \,.\, NN + 11.81 \,.\, P - 0 + 034 \,.\, SS_{out} - 0.75 \,.\, T_{out} - 0.05 \,.\, Pap$$

$$- 0.0002 . FR_{out} + 0.1 . Pulp - 0.6 . RF + 0.3 . Col_{out} + 4.8 . pH + 0.04 . Cond_{out} \quad (2)$$

$$DBO_{in} = 293.817 - 0.583 . NAm - 3.895 . NN + 9.645 . P - 0.019 . SS_{in} - 3.973 . T_{in} -$$
$$- 0.045 . Pap - 0.43 . RF + 0.009 . Col_{in} - 3.18 . pH - 0.01 . Cond_{in} + 0.19 . DQO_{in} \quad (3)$$

$$DBO_{out} = - 10.4 - 1.25 . NAm - 0.56 . NN - 2.6 . P + 0.24 . SS_{out} - 0.83 . T_{out} - 0.03 . Pap +$$
$$0.001 . FR_{out} + 0.01 . Pulp - 0.03 . Col_{out} - 0.3 . pH - 0.003 . Cond_{out} + 0.3 . DQO_{out} \quad (4)$$

Para os próximos estudos, existem outras formas de modelagem como modelos baseados em árvore, entre outros que podem ser aplicados a fim de comparação.

Outro direcionamento que pode ser dado ao estudo é uma análise temporal explícita, onde os dados de entrada são comparados aos dados de saída após um certo intervalo de dias.

Referências

DETERMINAÇÃO da Demanda Bioquímica de Oxigênio – DBO. **Portal de Tratamento de Água**, 2017. Disponível em <https://tratamentodeagua.com.br/artigo/determinacao-da-demanda-bioquimica-de-oxigenio-dbo/>. Acesso em 13 de nov. de 2022.

FUSATI. O Que é Demanda Bioquímica de Oxigênio (DBO)?. **Fusati**, 2021. Disponível em <https://www.fusati.com.br/o-que-e-demanda-bioquimica-de-oxigenio-dbo/>. Acesso em 13 de nov. de 2022.

HARRIS, C.R.; MILLMAN, K.J.; VAN DER WALT, S.J. et al. Array programming with NumPy. **Nature,** v. 585, n. 7825, p. 357–362, Setembro, 2020. Disponível em <https://www.nature.com/articles/s41586-020-2649-2>. Acesso em 13 de nov. de 2022.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, [*s.l*], v. 9,n. 3, p. 90-95, Junho, 2007. Disponível em <https://ieeexplore.ieee.org/document/4160265>. Acesso em 13 de nov. de 2022.

IERVOLINO, Luís Fernando. **Lagoas de estabilização**. Portal de Tratamento de Água, 2019. Disponível em <https://tratamentodeagua.com.br/artigo/lagoas-estabilizacao/>. Acesso em 13 de nov. de 2022.

ESQUERRE, Karla P. O.; SEBORG, Dale E.; BRUNS, Roy E.; MORI, Milton. Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part I. Linear approaches. **Chemical Engineering Journal**, Campinas (SP/BR), v. 104, p. 73–81, Novembro,2004

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, [*s.l.*], v. 12, p. 2825–2830, Outubro, 2011. Disponível em <https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf>. Acesso em 13 de nov. de 2022.

PYTHON SOFTWARE FOUNDATION. Python 3.11.0 Documentation, 2022. Página de documentação da Linguagem de Programação Python. Disponível em <https://docs.python.org/3/>. Acesso em 13 de nov. de 2022.

THE PANDAS DEVELOPMENT TEAM. pandas-dev/pandas: Pandas. **Zenodo**, [*s.l.*], versão 1.5.1, Outubro, 2022. Disponível em <https://zenodo.org/record/7223478#.Y3OgCuTMLIU>. Acesso em 13 de nov. de 2022.

WASKOM, M. L. seaborn: statistical data visualization. **The Journal of Open Source Software**, New York (EUA), v. 6, n. 60, p. 1-4, Abril, 2021. Disponível em <https://doi.org/10.21105/joss.03021>. Acesso em 13 de nov. 2022.