ANÁLISE EXPLORATÓRIA E MODELAGEM PARA SISTEMAS DE LAGOAS AERADAS DE UMA INDÚSTRIA P&C.

Marcus Victor Pires Matos[1]

Claudio Veloso Texeira[2]

Resumo

Esse trabalho busca desenvolver uma análise exploratória e preditora, visando o comportamento das variáveis de um sistema de tratamento em uma indústria. Nesse processo, é analisado parâmetros de qualidade como Demanda Bioquímica de Oxigênio (BOD) e variáveis que os influenciam, como pH, temperatura, cor, entre outros. Na indústria a ser trabalhada, existe o problema da demora para a análise do parâmetro citado. Dessa forma, o trabalho explora formas de construir um modelo, através da regressão linear, para predição do comportamento do BOD.

Palavras-chave: Análise, BOD, Preditora, Regressão, Modelo

Introdução.

O descarte industrial menos agressor ao planeta é uma ideia que ganha força, anualmente, nos governos mundiais. Hoje existem leis mais rigorosas e que tentam controlar o impacto ambiental da eliminação desses resíduos. Contudo, ainda falta uma longa caminhada até que os efeitos para o planeta sejam minimizados.

Uma indústria de papel e celulose possui um sistema de tratamento com lagoas aeradas. Através da coleta de dados por aproximadamento 4 anos, foi possível montar um conjunto de dados com variáveis importantíssimas para o processo. Tendo em mãos essas medições, será possível investigar, pela estatística, comportamentos que podem, ou não, estar influenciando no sistema de tratamento. Após essa análise, cria-se a esperança de possuir resultados favoráveis que ajudem a indústria a melhorar a eficiência dos seus processos.

[1] Universidade Federal da Bahia (UFBA), marcus.victor@ufba.br

[2] Universidade Federal da Bahia (UFBA), claudio.veloso@ufba.br

Objetivo.

Entender e interpretar os comportamentos das variáveis do tratamento do efluente. Também objetiva esse trabalho, investigar um possível modelo que explique com eficiência o comportamento do BOD (Parâmetro de qualidade), buscando a aprimoração do tempo para controle de qualidade desse parâmetro, já que a análise laboratorial demanda 5 dias.

Material e Método.

No tratamento dos dados, foi definido que só seriam avaliadas as variáveis que contivessem mais de 60% dos dados coletados. Dessa forma, a análise esteve pautada nas variáveis:

- Vazão de Efluente de Entrada (FR); Demanda Bioquímica de Oxigênio de entrada (BODin); Demanda Química de Oxigênio de Entrada (CODin); Pontencial Hidrogeônico de Entrada (pHin); Cor de Entrada (COLin); Temperatura de Entrada (Tin); Condutividade de Entrada (CONDin); Precipitação (RF); Produção de Celulose (Pulp); Produção de Papel (Pap); Demanda Bioquímica de Oxigênio de saída (BODout); Demanda Química de Oxigênio de Saída (CODout); Cor de Saída (COLout); Temperatura de Saída (Tout); Condutividade de Saída (CONDout), Vazão de Efluente de Saída (FRout).

Para a análise exploratória, foi verificado medidas de tendência central e de dispersão. Ou seja, média, mediana, desvio padrão, mínimo, máximo e quartis. Para análise gráfica foi usado gráficos de barra e séries temporais mensais para análise de comportamento geral, dispersão diária e histograma para concentração e tendência, além do boxplot para análise de variação e identificação de outliers.

Como verificação da veracidade e regularidade dos dados, usou-se a resolução nº 430/2011 do Conselho Nacional do Meio Ambiente – CONAMA como referência.

Após a análise exploratória de todas as variáveis, foi investigado a distribuição de cada variável, buscando entender a confiabiliadade do dados através de concentração, simetria e probabilidade. Para isso foi usado histogramas, curvas de Estimativa de Densidade de Kernel e gráficos Q-Q (Quantile - Quantile). A ideia de usar testes como Kolmogorov-Smirnov, correção por Lilliefors ou Shapiro-Wilk foi descartada após verificar ineficiência com grande quantidade de dados ruidosos durante o decorrer do trabalho.

A inferência estatística, através do intervalo de confiança para média, foi usada como uma forma de aumentar a confiabilidade das variáveis trabalhadas. Para verificar a correlação entre variáveis, aplicou-se uma matriz de correlação, baseada no método de Pearson e uma matriz de dispersão das variáveis. Dessa forma, foi possível identificar comportamentos lineares ou não, entre as variáveis de interesse e as variáveis influenciadoras.

Com as variáveis devidamente escolhidas, criou-se o modelo de regressão linear múltipla, por stepwise (passo a passo), analisando cada estatística F e escolhendo o modelo de maior valor. Após isso, verificou-se a qualidade do modelo e adequação dos resíduos. Para a qualificação do modelo, usou-se o R^2 ajustado devido a confiabilidade da mudança do mesmo; teste t para verificação individual da significância das variáveis; p-value como auxílio para significância individual e teste F para significância global do modelo. Para os resíduos usou-se Durbin-Watson para verificar se a variação dos resíduos é constante (homocedasticidade) e probabilidade de Jarque-Bera para normalidade dos resíduos.

Resultados e Discussão.

Para a análise exploratória, foi possível tirar algumas conclusões a respeito das variáveis e do sistema de tratamento. Foi possível verificar que que a variável cor teve uma queda brusca a partir de fevereiro de 2000. Uma conjectura a ser feita, é que existiu alguma situação que obrigou a fábrica a melhorar sua eficiência de resíduos, dessa forma, diminuindo a produção da substância que fortalecia a cor. A segunda grande observação está no comportamento da temperatura, que obedece uma sazonalidade padronizada. Nesse caso, a temperatura aumenta durante o verão e diminui durante o inverno, o que leva ao entendimento direto que pode ser uma influência das estações do ano. A ideia anterior, portanto, aponta para uma ineficiência do controle de temperatura das lagoas. Por fim, é possível supor que a coleta dos dados também possui falhas, pois ao verificar os dados de produção de papel e celulose, encontra-se um completo vazio entre novembro de 1996 e fevereiro de 1997.

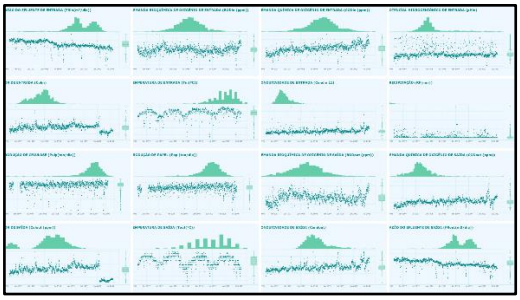

Figura 1 – Gráficos de Análise Exploratória (Cada gráfico composto de histograma, dispersão e boxplot – Para melhor visualização, acessar Colaboratoy)

Fonte: Próprio Autor

Segundo o CONAMA, os parâmetros de qualidade do tratamento de efluentes precisam estar dentro de alguns critérios. Para BOD, é necessário que exista uma redução de pelo menos 60% no processo. Analisando a redução nos dados da indústria, verificou-se que em ∼73,3% dos dias, o BOD foi reduzido seguindo a régua do CONAMA. O que é um apontamento positivo para a eficiência do sistema no caso desse parâmetro de qualidade, mas que ainda pode melhorar. Para o pH, a delimitação está entre 6,0 e 9,0. Ao verificar os dados, em 1146 dos 1324 dias analisados, o pH está dentro do esperado. Logo, por mais que existam variações altas de pH, o parâmetro, em sua maioria, é controlado. A temperatura ideal é definida como 40 °C. Verificando a temperatura de entrada, a mesma apresenta apenas ∼5,4% dos dados dentro da norma. Em contraste, a temperatura de saída apresenta 100% dos dados respeitando a resolução do CONAMA. Isso pode implicar na ideia de que o efluente, quando passado para o sistema de tratamento, está entrando com a temperatura muito alta. Ou então, que o controle de temperatura da lagoa II é mais eficiente do que a da lagoa I. Cada gráfico tem retas vermelhas tracejadas, indicando as delimitações do CONAMA.

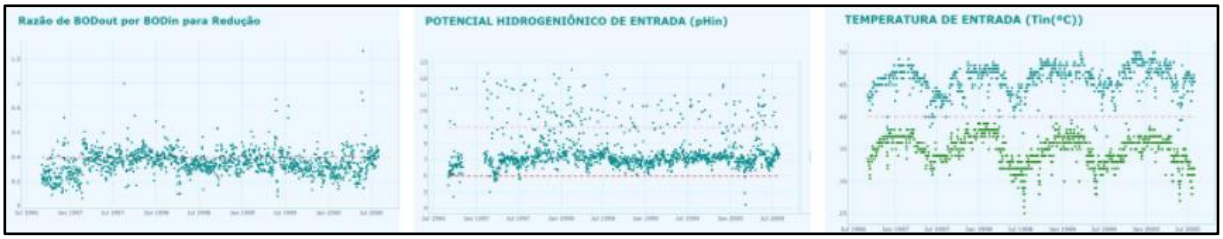

Figura 2 – Gráficos de Análise de Regulariade (Cada gráfico composto de histograma, dispersão e boxplot. O primeiro gráfico para razão de BOD, o segundo para pH e o terceiro para temperatura de entrada e saída. – Para melhor visualização, acessar Colaboratoy)

Fonte: Próprio Autor

Após isso, buscou-se entender o comportamento de distribuição da variável, classificando a partir da normalidade.

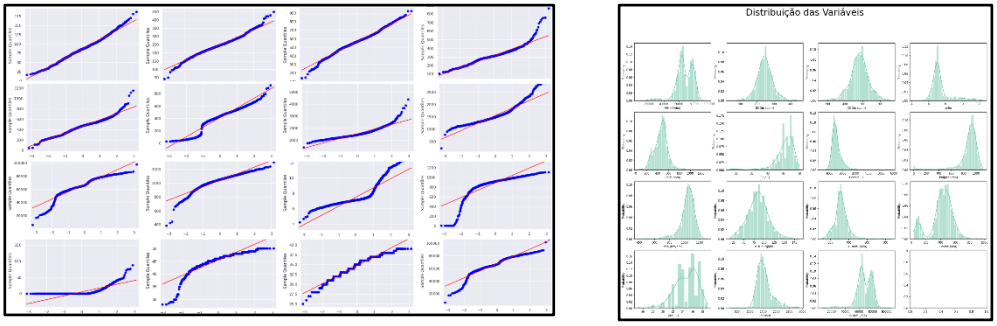

Figura 3 – Gráficos QQ-Plot na esqueda e histogramas com curvas KDE à direita (Para melhor visualização, acessar Colaboratoy)

Fonte: Próprio Autor

Após a análise, verificou-se que, temperatura de entrada e saída, produção de celulose, precipitação, pH e cor de saída não seguem a distribuição normal. Ou seja, 37% das variáveis estão fora dessa distribuição, o que é relativamente preocupante para a qualidade da análise e significânica da mesma para a predição. Para completude da análise, foi calculado o intervalo de confiança para as variáveis com normalidade.

Após isso, buscou-se entender a correlação entre as variáveis estudadas e o comportamento das mesmas. Assim, foi realizado uma matriz de correlação pelo método de Pearson, e uma matriz de dispersão.

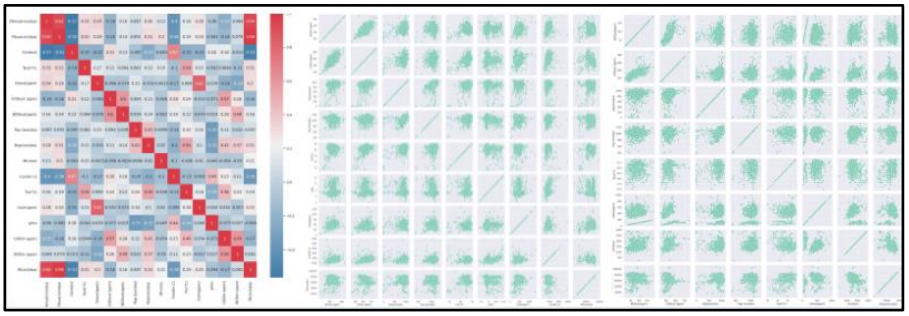

Figura 4 – Á esquerda tem a matriz de correlação, ao centro a matriz de dispersão para variáveis de entrada e à direita a matriz de dispersão das variáveis de saída. (Para melhor visualização, acessar Colaboratoy)

Fonte: Próprio Autor

Analisando os maiores valores de correlação, e os comportamentos lineares pela dispersão para BODin e BODout, as variáveis escolhidas para o modelo serão:

- Entrada: CODin, Pulp, Tin, CONDin e FR
- Saída: CODout, Pulp, Pap, CONDout, FRout, CODin

CODin é uma variável de entrada, mas foi incluída na modelagem de saída, pois, ao analisar o diagrama do processo, conclui-se que variáveis na entrada podem influenciar a saída, além de ter uma análise laboratorial mais rápida que BOD. O BODin não foi incluído no modelo de saída, pois é uma variável que também será analisada preditivamente.

Assim, através de stepwise (passo a passo), teoria físico-química, R^2 ajustado, p valor, valor da estatística F, determinou-se esses modelos como sendo os melhores:

$$BODin = 74{,}822 + 0.24577 \cdot CODin + 0.036641 \cdot Pulp$$

$$BODout = 21{,}165 - 0.024452 \cdot CODout + 1.0015 \cdot CODin$$

Segundo o R^2 ajustado para o primeiro modelo, ~33,5% do processo é explicado pelas variáveis independentes. Ao olhar a probabiliade de F($probabilidade\ de\ F\ (6{,}49 \times 10^{-109})$), é possível rejeitar a hipótese nula de que todos os coeficientes de regressão são nulos, pois ela é muito próxima de 0. Isso conta com um ponto a favor da significânica global do modelo. Ao olhar os p-valores, existe uma corroboração quanto a afirmação anterior, já que são bem pequenos e menores que 0,5. Pelo teste t, para um grau de liberade de 1217 e 95% de confiança, cria-se o intervalo de -1,96 < t < 1,96. Verificando os valores de t para cada variável (21,203; 4,982), entende-se que de fato, as variáveis tem significância para o modelo. Durbin-Watson implica em uma variação dos erros constante (homocedasticidade), e é interessante que seu valor esteja entre 1 e 2. Para o caso do modelo I, o valor de Durbin-Watson de 0,994 é bem próximo do ideal, tendo diferença apenas na 3ª casa decimal, aumentando a confiabilidae do modelo. Avaliando a normalidade dos resíduos através da probabilidade de Jarque-Bera, encontra-se um valor muito pequeno ($1{,}56 \times 10^{-46}$), o que indica uma possível normalidade nos dados.

Para o segundo modelo, pelo R^2 ajustado, 36% do processo é explicado. Pela probabilidade da estatística F ($Probabilidade\ de\ F(9{,}24 \times 10^{-119})$), verifica-se uma significância global muito boa. Pelos p-valores muito próximos de 0, o t de cada variável (22,624; -3,396), entende-se também uma significância individual também confiável. Para o teste de Durbin-Watson, o valor obtido de 0,699, está bem abaixo do intervalo ideal, o que implica em uma possível heterocedasticidade. Para o teste de Jarque-Bera, o valor é de 8,38 x 10^{-6}, um valor relativamente grande quando comparado ao teste do primeiro modelo, o qual leva a crer uma falta de normalidade nos resíduos.

Conclusão

Por fim, é possível notar uma ruidosidade muito grande nos dados, o que afetou muito a escolha das variáveis físico-químicas para os modelos. É fato que 33,5% e 36% são porcentagens baixíssimas, para que toda uma indústria se baseie nessas equações para uso cotidiano. Porém, a análise abre portas para que a indústria melhore a variação dos parâmetros e a ruidosidade de coleta, assim, podendo encontrar um modelo superior, e então ter seu sistema baseado nele. Dessa forma, mesmo com uma análise básica, foi possível ter resultados satisfatórios e que iluminam um caminho promissor para o uso de regressão na indústria.

Referências

BRASIL. Ministério do Meio Ambiente. Conselho Nacional do Meio Ambiente. **Resolução Nº 430, de 13 de Maio de 2011**. Brasília, 2011.

Esquerre K., Seborg D., Bruns R., Mori M. **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. 140, p. 73–81, 2004.

Esquerre K., Seborg D., Bruns R., Mori M. **Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part II. Nonlinear approaches**. Chemical Engineering Journal. 105, p. 61–69, 2004.

ESQUERRE, Karla. ENGD02 – **Estatística na Engenharia**: Regressão Linear Simples – ENGD02. 43 slides. Disponível em: <https://ava.ufba.br/pluginfile.php/1751248/mod_resource/content/1/Regressa%CC%83o%20linear%20simples.pdf>

ESQUERRE, Karla. ENGD02 – **Estatística na Engenharia**: Regressão Linear Múltipla – ENGD02. 28 slides. Disponível em: <https://ava.ufba.br/pluginfile.php/1751251/mod_resource/content/1/Regressa%CC%83o%20linear%20mu%CC%81ltipla.pdf>

ESQUERRE, Karla. **Roteiro:** Modelagem – SISTEMA DE LAGOAS AERADAS DE UMA INDÚSTRIA DE PAPEL E CELULOSE (P & C).

KORSTANJE, Joos. 6 ways to test for a Normal Distribution — which one to use. **Towards Data Science**. 2019. Disponível em: <https://towardsdatascience.com/6-ways-to-test-for-a-normal-distribution-which-one-to-use-9dcf47d8fa93>. Acesso em: 27 de nov 2021.

Montgomery D. C., Runger G. C. (2003). **Applied Statistics and Probabilities for Engineers**. John Wiley & Sons, 3ª edição.

Pedregosa et al. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research. 12, p. 2825 – 2830. 2011.

Plotly. **Plotly Python Open Source Graphing Library**. 2021. Disponível em: <https://plotly.com/python/>. Acesso em: 27 nov 2021.

PROCÓPIO, Aline. **Avaliação da Eficiência do Sistema de Tratamento de Efluentes tipo Mizumo Business em Canteiro de Obras em Sabará-MG**.Orientadora: LANA, Msc. Lívia. 2014. 61f. TCC (Graduação) – Curso de Engenharia Ambiental e Sanitarista, Departamento de Ciência e Tecnologia Ambiental, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2014.

Python Software Foundation. **Documentação Python 3.9.7**. Versão 3.9.7, 2021. Disponível em: <https://docs.python.org/pt-br/3.9/>. Acesso em: 24 nov 2021.

SciPy. **SciPy documentation**. Versão 1.8. 2021. Disponível em: < https://scipy.github.io/devdocs/index.html>.

The Matplotlib development team. **Matplotlib 3.5.0 Documentation**. Versão 3.5.0. 2021. Disponível em: < https://matplotlib.org/stable/index.html>. Acesso em: 27 nov 2021.

The NumPy community. **NumPy v1.21 Manual**. Versão 1.21. 2021. Disponível em: < https://numpy.org/doc/stable/>

The pandas development team. **Pandas documentation**. Versão 1.3.3, 2021. Disponível em: < https://pandas.pydata.org/docs/>. Acesso em: 27 nov 2021.

WASKOM, Michael. **Seaborn: statistical data visualization**. Journal of Open Source Software. V6, nº 60, p. 3021, 2021.

YADAV, Jyoti. Statistics: How Should I interpret results of OLS? Medium. 2019. Disponível em: <https://jyotiyadav99111.medium.com/statistics-how-should-i-interpret-results-of-ols-3bde1ebeec01>. Acesso em: 27 nov 2021.

Anexo

https://colab.research.google.com/drive/12uNiBYlg8XKSEPY55OCQrdH7vveCMmVN?usp=sharing