Estudo estatístico da relação funcional entre os parâmetros DBO5 e COT em corpos d´agua do Estado de São Paulo

Interessados:

Nelson Menegon Jr. – Divisão de qualidade das águas e do solo,

Antonio da Costa Rugue Jr. - Divisão de Laboratório de Limeira

Estatístico Responsável:

Antonio de Castro **Bruni** (PhD)

CONRE 6363/A – 3ª. Região

Dez/2019

Objetivo

- ✓ Verificar a existência de relação funcional entre os parâmetros DBO_5 (demanda bioquímica de oxigênio) e COT (carbono orgânico total) através de modelagem estatística;
- ✓ Identificar variáveis correlacionadas com COT e DBO_5 e avaliar suas influências na modelagem;
- ✓ Testar se o modelo ajustado é aplicável a todo o Estado de São Paulo ou se existem diferenças de comportamento para diferentes UGRHIs (Unidade de Gerenciamento de Recursos Hídricos);
- ✓ Avaliar a qualidade do modelo ajustado quanto à explicação do comportamento da DBO_5 em função do COT e analisar o comportamento dos resíduos do modelo.

Fonte de dados

Todos os dados do presente estudo foram gerados pelos Laboratórios da CETESB que pertencem à Rede Brasileira de Laboratórios de Ensaio (**RBLE**) e à *Red de Laboratórios de Ambiente y Salud de América Latina y El Caribe* (**RELAC**), possuindo sistema de qualidade laboratorial com acreditação junto ao Instituto Nacional de Metrologia, Qualidade e Tecnologia - INMETRO, pela NBR ISO/IEC 17.025 para mais de 800 ensaios na área ambiental englobando análises inorgânicas, orgânicas, microbiológicas, parasitológicas, ecotoxicológicas, hidrobiológicas e de mutagenicidade, bem como amostragem de águas, sedimentos, efluentes e comunidades aquáticas. Possui adicionalmente Certificados de Qualidade em Biossegurança, o **CQB** (CQB 286/09).

Metodologia estatística

Dados do estudo

Os dados que serviram de base para a realização deste estudo são aqueles obtidos da rede de monitoramento da qualidade das águas da CETESB no estado de São Paulo. No total são 392 pontos que monitoram os parâmetros DBO_5 e COT. Utilizamos os dados do período de 2014 a 2019 provenientes das coletas bimestrais realizadas no monitoramento da rede.

Os parâmetros pesquisados foram: pH, oxigênio dissolvido - OD, Fósforo total, *E. coli*, COT, nitrogênio Kjeldahl - NTK, Ni amoniacal e DBO_5.

O número total de registros no período de estudo foi de 13711.

Na Figura 1 apresentamos o mapa do Estado de São Paulo com os pontos de monitoramento que fizeram parte do estudo.

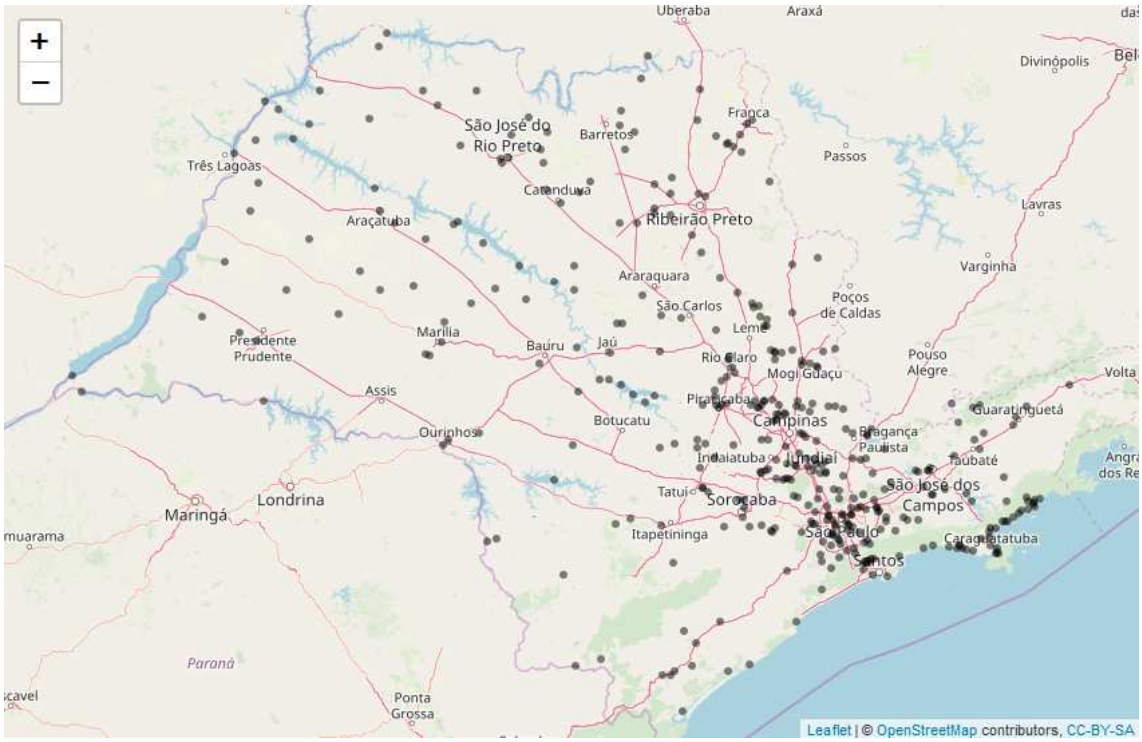

Figura 1 – Mapa com a localização dos pontos de monitoramento de DBO_5 e COT em São Paulo

Software estatístico

Todas as análises estatísticas e todos os gráficos foram feitos utilizando a linguagem R (R Core Team, 2017).

Matriz de correlação

Para compreendermos a estrutura de dependência entre as variáveis do estudo, elaboramos a matriz de correlação entre as variáveis quantitativas. De maneira visual ela apresenta o grau de correlação e o sentido da mesma, se direto ou inverso. Essa matriz é apresentada na Figura 2. Nela observamos que:

- ✓ pH não se correlacionou significativamente com as demais variáveis pesquisadas;
- ✓ OD se correlacionou de maneira inversa com as demais variáveis, mas sua magnitude foi baixa;
- ✓ Fósforo total não se correlacionou significativamente com as demais variáveis;
- ✓ *E.coli* se correlacionou de maneira direta com COT e DBO_5;
- ✓ COT se correlacionou com DBO_5 e com a série de Nitrogênio, assim, somente o COT deve estar presente no modelo para DBO_5;
- ✓ NTK se correlacionou de maneira direta com Ni amoniacal e DBO_5;
- ✓ A razão DBO_5/COT se correlacionou diretamente com o DBO_5.

A modelagem então será feita considerando COT como variável preditiva (ou independente) e a variável DBO_5 como variável resposta, ou seja, $DBO_5 = f(COT)$

O desafio então é descobrir a que família de funções “f” pertence.

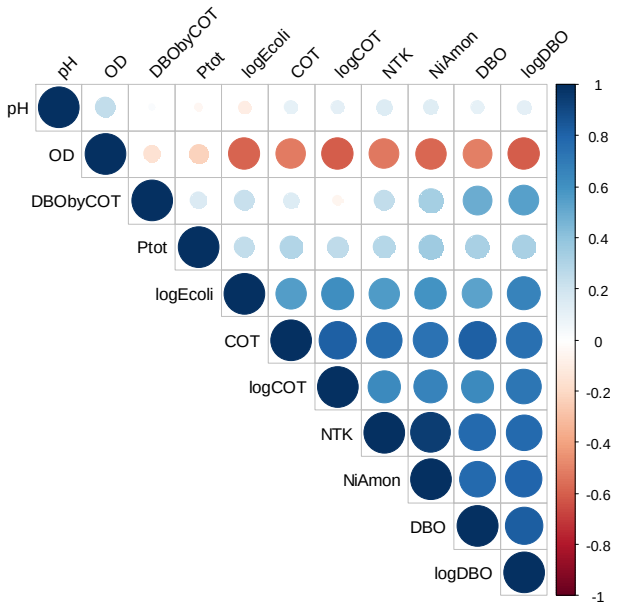

Figura 2 – Matriz de correlação entre as variáveis quantitativas do estudo

Modelagem Estatística

A modelagem estatística tem por objetivo procurar a lei que rege o comportamento das variáveis envolvidas. Neste estudo desejamos determinar qual é a relação funcional que descreve o comportamento do parâmetro DBO_5 em função do COT.

Ajustamos então um modelo utilizando a metodologia LOESS - *Locally Estimated Scatterplot Smoothing*, que ajusta um modelo por partes (locais) visando obtermos o comportamento geral da curva, a metodologia não assume nenhuma família de funções à priori. Na Figura 3 temos o resultado do ajuste observado.

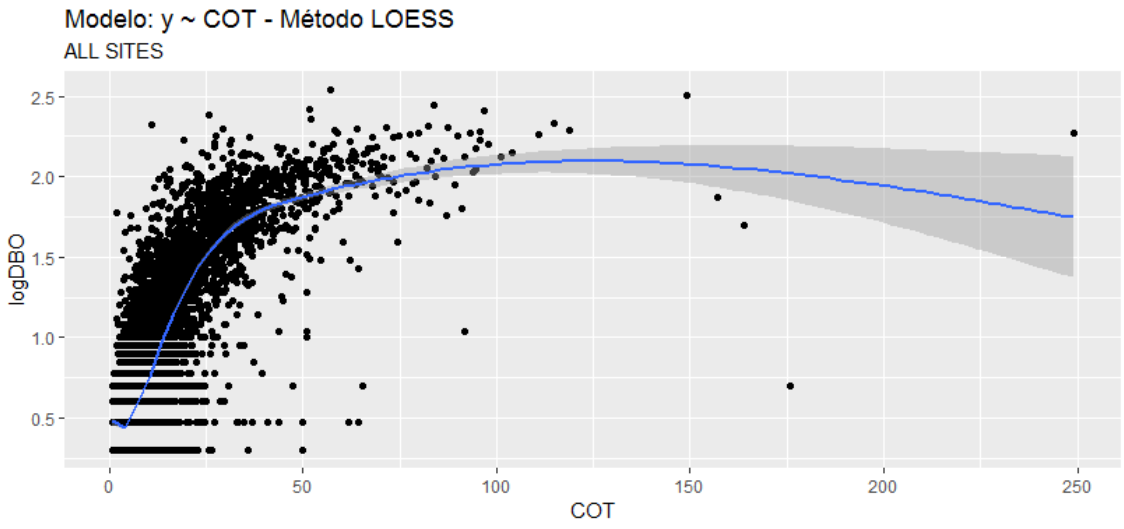

Figura 3 – Regressão LOESS para avaliar o comportamento da DBO_5 em função da COT

Podemos constatar a partir da Figura 3 que o comportamento envolvendo as variáveis não é linear. De fato, outros autores tiveram a mesma constatação em seus trabalhos, Roopali et.al., 2003 e Constable & McBean, 1979.

A Figura 3 aponta ainda para a necessidade de retirada de pontos que destoam do comportamento esperado, critérios não subjetivos para retirada desses pontos devem ser definidos a partir da totalidade dos dados.

Critério de aceitação de dados

O critério de aceitação de dados para o estudo de correlação entre DBO_5 e COT foi estabelecido com base nos resultados observados nos 392 pontos pesquisados. Na Figura 3 apresentamos o histograma dos valores de DBO_5/COT, nela fica claro que a quase totalidade dos resultados encontra-se abaixo do valor 3 para a razão DBO_5 / COT, assim sendo estabelecemos como critério de aceitação dos dados caso a razão fosse menor que 3. A Figura 5 apresenta os diagramas de *Boxplot* para a razão DBO/COT, apontando as diferenças entre as diversas UGRHIs.

Do total de 13711 registros, apenas 451 (3,28%) foram excluídos por este critério, veja na Figura 4 o histograma da razão DBO_5/COT. Na Figura 5 apresentamos o diagrama de Boxplot da razão, segundo a UGRHI, nela já foi aplicado este critério.

Valores observados de COT acima de 150 mg/L foram retirados dos ajustes dos modelos, são apenas 4 dados que não atenderam este critério. Veja que mantendo estes pontos temos uma queda na curva após o valor de 150 mg/L de COT, o que não faz sentido.

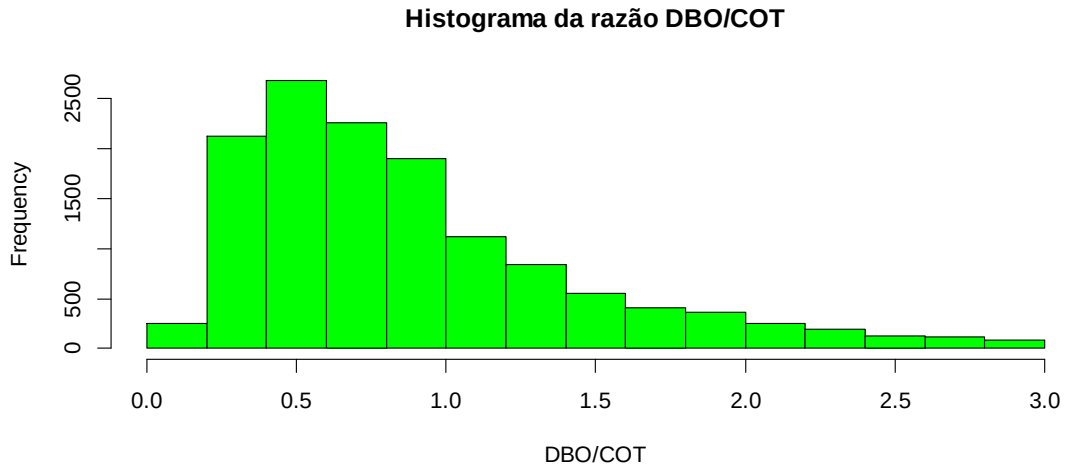

Figura 4 – Histograma da razão DBO_5/COT

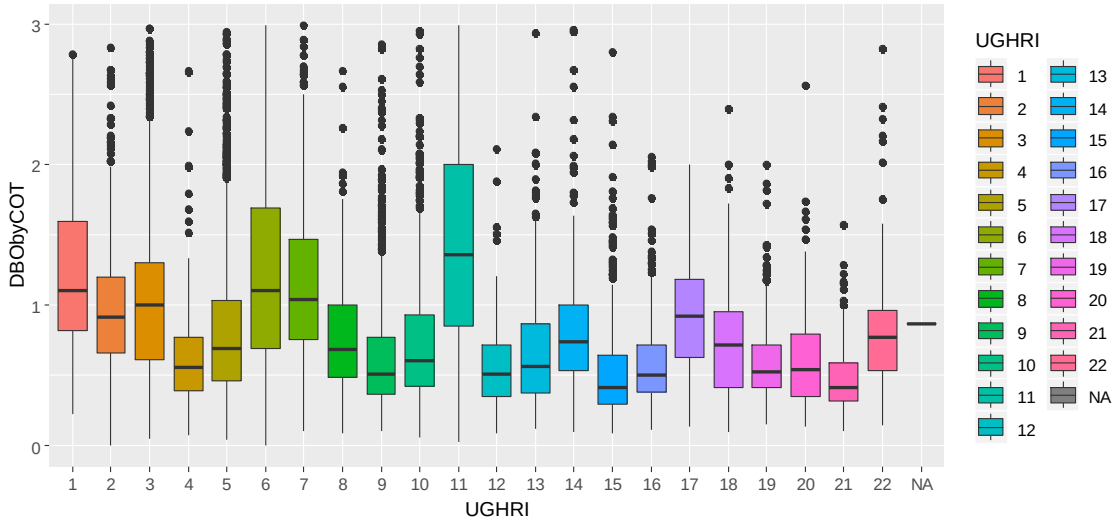

Figura 5 – Diagramas de Boxplot para a razão DBO/COT por UGRHI

Resultados

Na Figura 6 apresentamos a distribuição conjunta dos dados de DBO_5 e COT, ambos em escala logarítmica. Esse gráfico aponta para uma correlação positiva (direta) entre estes dois parâmetros.

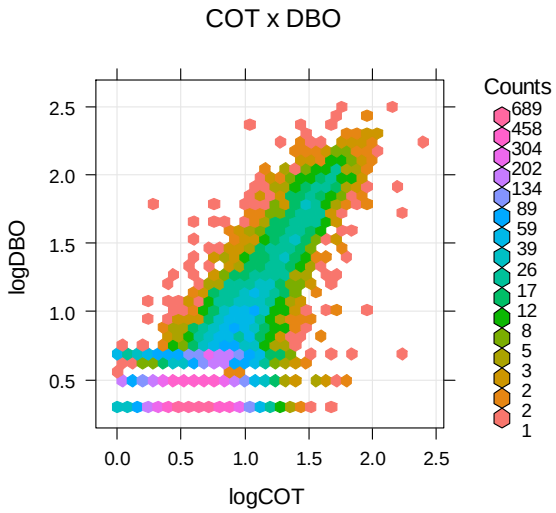

Figura 6 – Distribuição conjunta de COT e DBO_5 – Toda a rede de monitoramento do Estado de São Paulo

Na Figura 7 apresentamos os diagramas de Boxplot por UGRHI e Ano, nela constatamos que o comportamento não foi similar em todas as UGRHIs. Iremos então ajustar um modelo linear geral – MLG (Chambers, 1992) para estimar as diferenças entre elas. Ano e Mês da coleta igualmente serão introduzidos no modelo para verificação de sua contribuição na variabilidade dos dados de DBO_5.

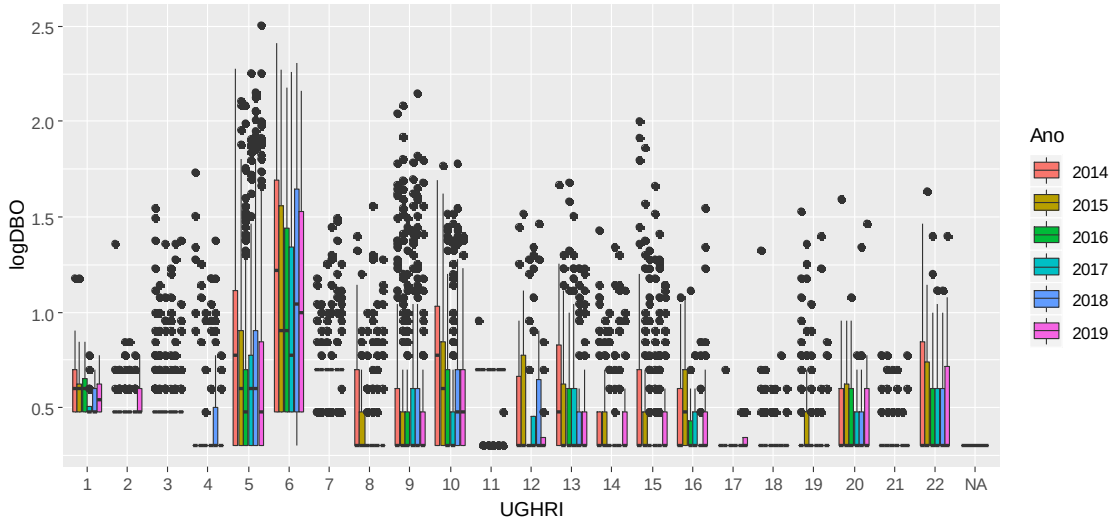

Figura 7 – Boxplot do DBO_5 [log] em função da UGRHI e do Ano

Na Figura 8 apresentamos de maneira visual os resultados do ajuste do modelo linear geral.

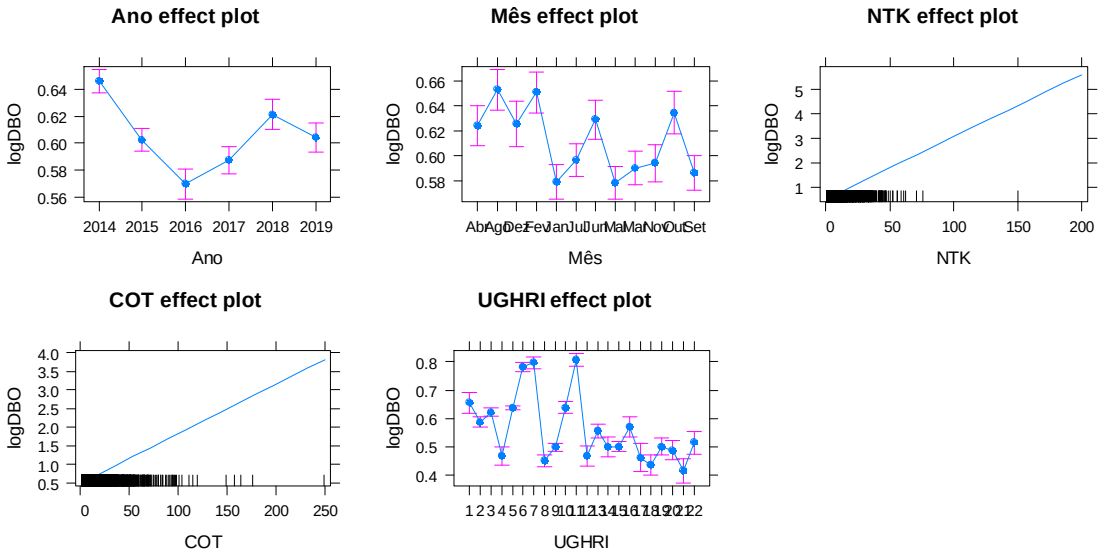

Figura 8 – Estimativas de efeitos obtidas pelo Modelo Linear Geral

A partir de sua observação, identificamos alguns grupos de médias similares para o Ano, Mês e UGRHI. Criamos então variáveis indicadoras para esses grupos e ajustaremos novo modelo para testarmos a significância das diferenças entre os grupo. Esses agrupamentos foram confrontados com os resultados apresentados na Figura 7. Definimos então os seguintes agrupamentos:

- Para o Ano: 1=[2014], 2=[2015,2018,2019] e 3=[2016,2017]

- Para os meses: 1=[Abr, Ago, Dez, Fev, Jun, Out] e 2=[Jan, Jul, Mai, Mar, Nov, Set].

- Para as UGRHIs: 1=[5,6], 2=[1,2,3,4,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22]

Estes números dos grupos são os expostos na Figura 9.

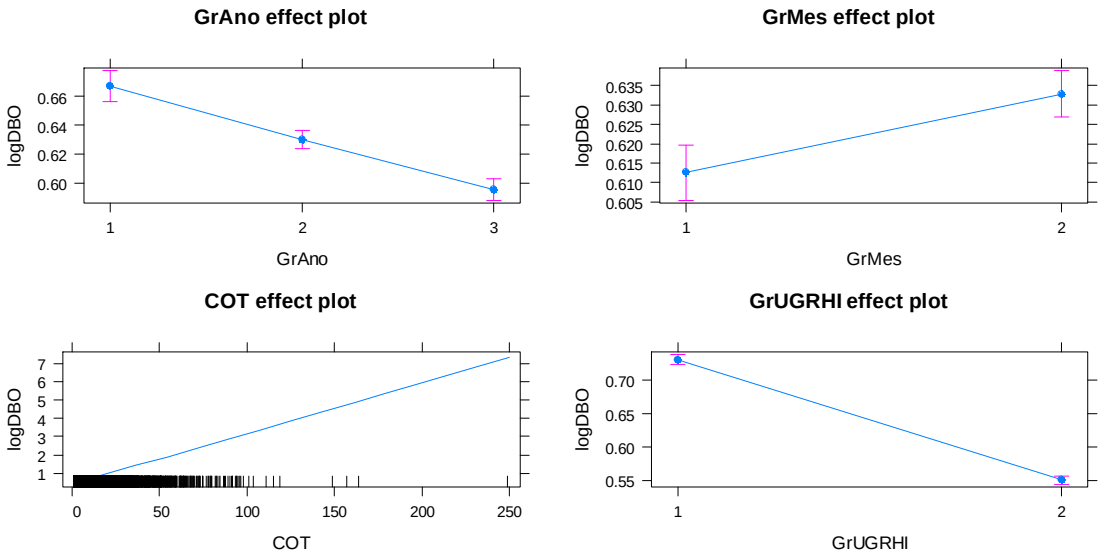

Figura 9 – Efeitos estimados pelo Modelo Linear Geral para os Grupos

A magnitude do efeito máximo observado entre os grupos de Anos foi igual a 3,88 mg/L, entre os grupos de meses foi igual a 1,66 mg/L, entre os grupos de UGRHIs foi de 2,07 mg/L. Resumidamente, apesar de significativas as diferenças, em magnitude não causam efeito significativo, o modelo então será simplificado quanto ao seu uso e abrangência.

Várias famílias de funções foram testadas, a que apresentou melhor ajuste foi a família do seguinte tipo:

$$DBO_5 = \beta * COT^{\lambda}$$

Para exemplificar essa conclusão, ajustamos modelos para os dois grupos de UGRHIs, os resultados estão apresentados nas Figuras 10 e 11 que seguem.

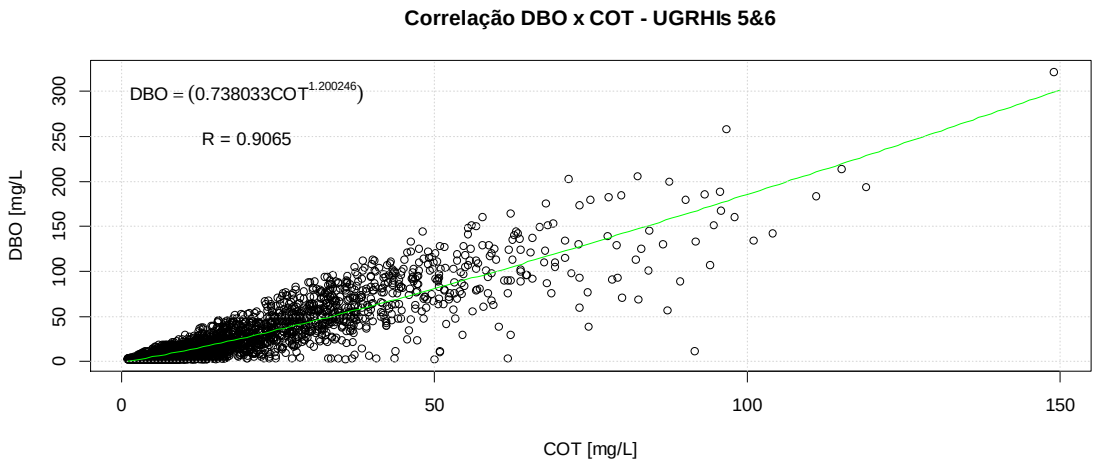

Figura 10 – Modelo ajustado para obter DBO a partir de COT nas UGRHIs 5 e 6

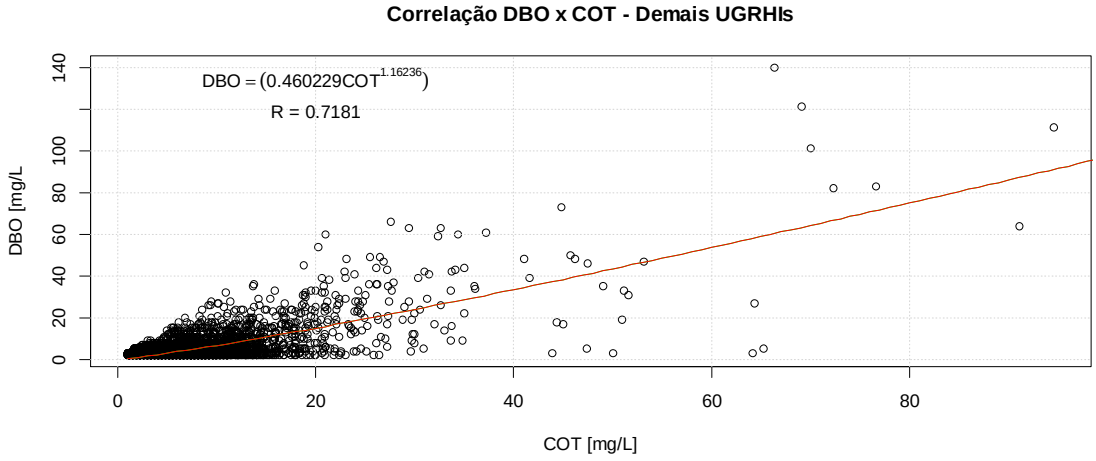

Figura 11 – Modelo ajustado para obter DBO a partir de COT nas Demais UGRHIs

Constatamos que há perda na explicação do modelo das Demais UGRHIs, se comparamos com o resultado observado com os dados dos pontos das UGRHIs 5 e 6.

Se ajustarmos um único modelo para todas as UGRHIs o resultado é superior (R=0,8895) àquele observado especificamente para as Demais UGRHIs (R=0,7181).

Na Figura 12 apresentamos o resultado do modelo ajustado para Todas as UGRHIs.

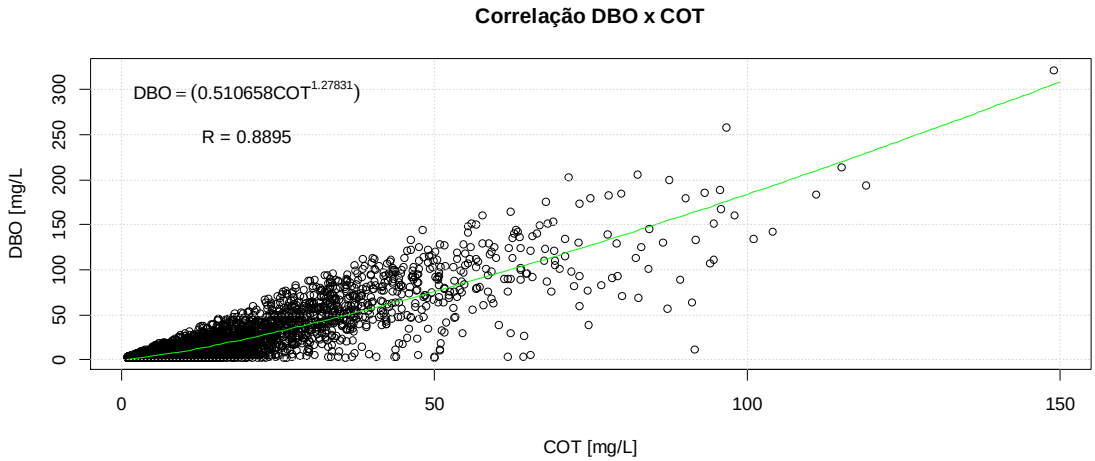

Figura 12 – Modelo ajustado para previsão de DBO em função de COT para todas as UGRHIs

Análise dos resíduos do Modelo geral para previsão de DBO em função de COT

São duas as hipóteses do modelo ajustado: os resíduos são independentes (não correlacionados) e os resíduos do modelo seguem a distribuição Normal (Gaussiana) de probabilidades.

Na Figura 13 apresentamos os gráficos que confirmam essas duas hipóteses.

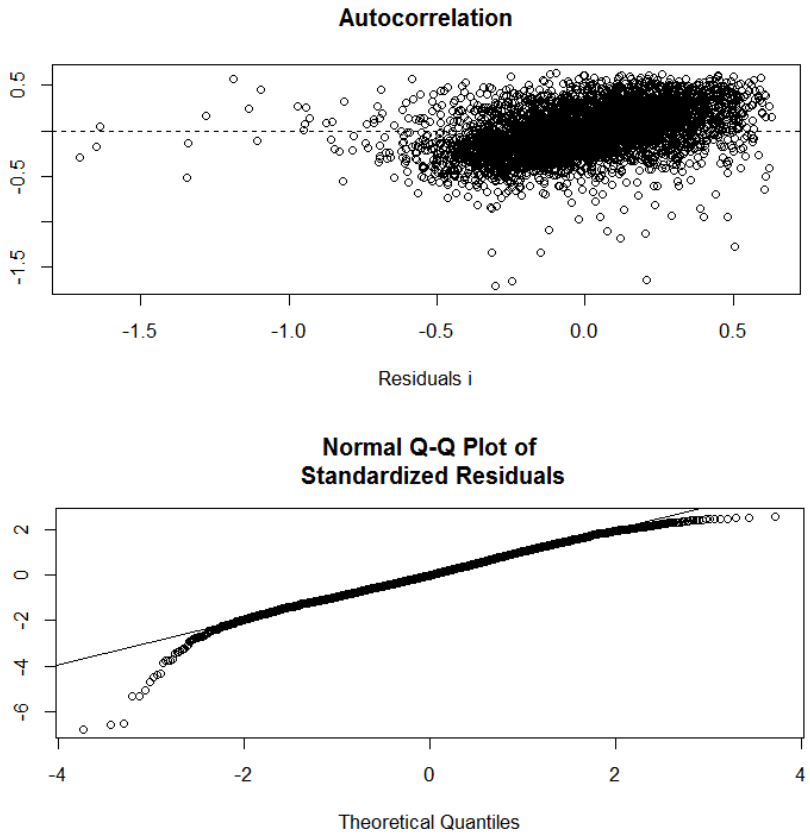

Figura 13 – Comportamento dos resíduos do modelo geral para DBO em função de COT

Podemos constatar a partir da Figura 13 que os resíduos aderem perfeitamente a distribuição Normal no intervalo [-2 +2] e a hipótese de média igual a zero e aleatoriedade foram confirmadas.

Conclusões

- ✓ Identificamos a relação funcional que relaciona as medições de COT com as de DBO_5, essa relação funcional é da seguinte forma: $\mathbf{DBO_5 = \beta * COT^{\lambda}}$
- ✓ As variáveis Ni Amoniacal, NTK e E.coli se correlacionam significativamente e de maneira direta com o COT, fato que desaconselha seu emprego no modelo pois levariam à presença de multicolinearidade;
- ✓ O modelo ajustado é aplicável a todas as UGRHIs do Estado de São Paulo, seus parâmetros são β**= 0,510658 e** λ**=1,27831.** O coeficiente de explicação ficou em **88,95%;**
- ✓ A condição de contorno para aplicação do modelo é que **COT < 150 mg/L**, na faixa de 150 mg/L a 250 mg/L ainda pode ser usado, mas não é uma faixa ideal para aplicação pois poucas amostras tiveram valor nesta faixa;
- ✓ O modelo ajustado atendeu as hipóteses de resíduos com distribuição Normal com média zero e não correlacionados. A explicação, em termos estatísticos, faculta o seu emprego para estimar as concentrações de DBO_5 a parir do valor observado de COT nos pontos de monitoramento da qualidade das águas;
- ✓ Esse modelo não é aplicável para emissões de fontes industriais.

Discussão e recomendações

Conforme podemos observar, o comportamento do DBO_5 em função do COT não apresenta uma ruptura quando separados em dois grupos de UGRHIs, de fato, eles seguem o mesmo perfil de curva. Na Figura 14 a seguir apresentamos o *scatterplot* distinguindo os dados dos dois grupos de UGRHIs, um formado somente pelas unidades 5 e 6 e o outro com as Demais UGRHIs.

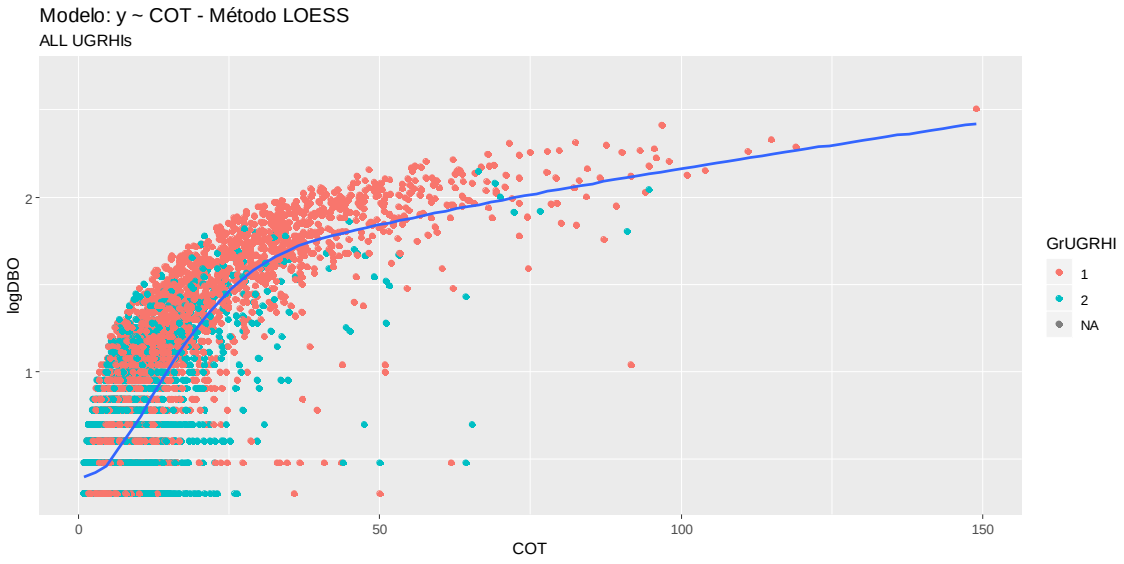

Figura 14 – Diagrama de dispersão dos dados de COT e DBO5 segundo os grupos de UGRHIs

Recomendo que pelo menos 10% dos pontos de monitoramento das UGRHIs 5 e 6 continuem monitorando o parâmetro DBO_5, pelo período de um ano, para confrontar com as previsões do modelo.

Caso em algum ponto de monitoramento da rede, valores de COT > 150 mg/L comecem a se tornar frequentes, recomendamos que a DBO_5 seja então monitorada nestes pontos.

Referências

Roopali MRB, Hiremath S and Kulkarni VR (2003**)** *Correlation between BOD, COD and TOC, Journal of Industrial Pollution Control* **19** (2), pp. 187-191.

Constable TW and McBean ER (1979) BOD/TOC correlations and their application to water quality evaluation. *Water, Air and Soil pollution* (**11**), pp.363-375.

Chambers JM (1992) *Linear models.* Chapter 4 of *Statistical Models in S.* Eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.