



Programa de Pós-Graduação Lato Sensu
Especialização em Gestão Ambiental
Campus Nilópolis

Paulo Sérgio De Oliveira Cezário

**AVALIAÇÃO DO DESEMPENHO DE MÉTODOS DE
REGRESSÃO MULTIVARIADOS PARA ESTIMAR A DEMANDA
QUÍMICA DE OXIGÊNIO DE UMA ESTAÇÃO DE TRATAMENTO DE
EFLUENTES**

Nilópolis – RJ

2018

Paulo Sérgio De Oliveira Cezário

**AVALIAÇÃO DO DESEMPENHO DE MÉTODOS DE
REGRESSÃO MULTIVARIADOS PARA ESTIMAR A DEMANDA
QUÍMICA DE OXIGÊNIO DE UMA ESTAÇÃO DE TRATAMENTO DE
EFLUENTES**

Trabalho de Conclusão de Curso
realizado no Programa de Pós-Graduação
do Instituto Federal do Rio de Janeiro,
como parte dos requisitos para a obtenção
do título de Especialista em Gestão
Ambiental.

Orientador: Prof. Dr. Marco Aurélio Passos Louzada

Nilópolis – RJ

2018

CIP - Catalogação na Publicação

C387a

Cezário, Paulo Sérgio de Oliveira
AVALIAÇÃO DO DESEMPENHO DE MÉTODOS DE REGRESSÃO
MULTIVARIADOS PARA ESTIMAR A DEMANDA QUÍMICA DE
OXIGÊNIO DE UMA ESTAÇÃO DE TRATAMENTO DE EFLUENTES
/ Paulo Sérgio de Oliveira Cezário. -- Nilópolis, 2018.
84 f. : il. ; 30 cm.

Orientação: Marco Aurélio Passos Louzada.

Trabalho de Conclusão de Curso (especialização) --Instituto
Federal de Educação, Ciência e Tecnologia do Rio de Janeiro,
Especialização em Gestão Ambiental, 2018.

1 . Estação de Tratamento de Efluentes. 2. Regressão Linear
Múltipla. 3. Regressão por Componentes Principais. 4. Mínimos
Quadrados Parciais. 5. Máquina de Vetor de Suporte. I. Título.

Elaborado pelo Módulo Ficha Catalográfica do Sistema Intranet do
IFRJ - Campus Volta Redonda e Modificado pelo Campus
Nilópolis/LAC, com os dados fornecidos pelo(a) autor(a).

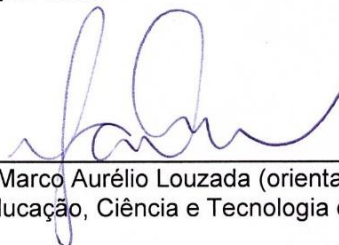
Bibliotecário: Elon F. Lima CRB-7/5783

Paulo Sérgio de Oliveira Cezário

**AVALIAÇÃO DO DESEMPENHO DE MÉTODOS DE REGRESSÃO MULTIVARIADOS
PARA ESTIMAR A DEMANDA QUÍMICA DE OXIGÊNIO DE UMA ESTAÇÃO DE
TRATAMENTO DE EFLUENTES**

Trabalho de Conclusão de Curso no
Programa de Pós-Graduação do Instituto
Federal do Rio de Janeiro, como parte
dos requisitos para a obtenção do título
de Especialista em Gestão Ambiental.

Data de aprovação: 11 de 12 de 2018.



Prof. Dr. Marco Aurélio Louzada (orientador)
Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro



Prof. Dr. Alexandre Mendes
Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro



Prof. MSc. Sérgio Henrique Silva Junior
Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro

Nilópolis – RJ
2018

AGRADECIMENTOS

Gostaria de neste momento agradecer a todas as pessoas que de alguma forma colaboraram para a execução deste estudo, em especial:

À Deus.

À minha família.

À Luana, meu amor.

Ao Professor Dr. Aderval Severino Luna.

Ao Professor Dr. Marco Aurélio Passos Louzada

Ao Manuel Poch, Javier Bejar e Ulises Cortes por disponibilizarem o banco de dados na UCI Machine Learning Repository.

Da sua posição orbital pode ver-se que, indubitavelmente, alguma coisa ocorreu mal. Os organismos dominantes, quaisquer que sejam – que se deram ao imenso trabalho de modificar a superfície -, estão a destruir simultaneamente a sua camada de ozônio e as suas florestas, a deixar erodir o seu solo fértil e a realizar experiências maciças e descontroladas sobre o clima de seu planeta. Não terão dado pelo que se passa? Terão esquecido o que os aguarda? Serão incapazes de trabalhar em conjunto em nome do meio ambiente que os sustenta a todos?

Talvez, pensa você, esteja na altura de reavaliar a conjectura de que existe vida inteligente na Terra.

Carl Sagan, O Ponto-Azul Claro.

RESUMO

CEZÁRIO, Paulo Sérgio de Oliveira. Avaliação do desempenho de métodos de regressão multivariados para estimar a demanda química de oxigênio de uma estação de tratamento de efluentes. 84 p. Trabalho de Conclusão de Curso. Programa de Pós-Graduação do Instituto Federal do Rio de Janeiro. Nilópolis, RJ. 2018.

A operação e o desempenho de uma estação de tratamento de efluentes (ETE) é essencial para a adequação do efluente gerado pelo processo aos padrões de lançamento no corpo hídrico receptor, mantendo assim um equilíbrio entre a produção e a qualidade ambiental. Neste cenário, a gestão ambiental pode fazer o uso das técnicas estatísticas multivariadas para acompanhar a operação e a previsão de cenários futuros de uma planta de tratamento de águas residuais.

Neste trabalho, são observados os comportamentos das variáveis e a correlação entre estas, para assim empregar técnicas de regressão: Linear Múltipla (RLM), por Componentes Principais (PCR) e por Mínimos Quadrados Parciais (PLS), assim como uma técnica não-linear, a Máquina de Vetores de Suporte (SVM) para a previsão da Demanda Química de Oxigênio (DQO) na saída da ETE. Estas são comparadas e avaliadas por meio dos parâmetros de mérito e da relação entre os valores previstos pelos modelos e os valores reais. Apesar de existir um comportamento linear entre algumas variáveis, isto torna-se irrelevante quando o objetivo é a previsão da DQO. Os métodos lineares foram ineficazes quando comparados em relação à técnica não-linear SVM utilizando o *kernel* de base radial para a previsão de DQO.

Palavras-Chaves: ETE, RLM, PCR, PLS, SVM

ABSTRACT

CEZÁRIO, Paulo Sérgio de Oliveira. Avaliação do desempenho de métodos de regressão multivariados para estimar a demanda química de oxigênio de uma estação de tratamento de efluentes. 84 p. Trabalho de Conclusão de Curso. Programa de Pós-Graduação do Instituto Federal do Rio de Janeiro. Nilópolis, RJ. 2018.

The operation and performance of a Wastewater Treatment Plant (WWT) are essential for the adequacy of the effluent generated by the process to the standards of release in the receiving water body, thus maintaining a balance between production and environmental quality. In this scenario, environmental management can make use of multivariate statistical techniques to monitor the operation and prediction of future scenarios of a wastewater treatment plant. In this work, we observe the behavior of the variables and the correlation between them, to use regression techniques: Multiple Linear (RLM), Principal Components (PCR) and Partial Least Squares (PLS), as well as a nonlinear technique, the Support Vector Machine (SVM) for the prediction of Chemical Oxygen Demand (COD) at the exit of the WWT. These are compared and evaluated using figures of merit and the relationship between the values predicted by the models and the actual values. Although there is a linear behavior among some variables, this becomes irrelevant when the objective is the COD forecast. Linear methods were ineffective when compared to the nonlinear SVM technique using the radial base *kernel* for the prediction of COD.

Keywords: WWT, MLR, PCR, PLS, SVM

LISTA DE FIGURAS

Figura 1 Mapeamento do espaço de entrada em um espaço de características (Chen <i>et al</i> , 2004).....	21
Figura 2 Função de base radial (RBF) (Adaptado de Poppi, 2015).....	21
Figura 3 Transformação de um problema de regressão linear em classificação binária (Adaptado de Poppi, 2015).....	22
Figura 4 Geometria do hiperplano (Adaptado de Poppi, 2015)	22
Figura 5 Diagrama do processo da ETE (Adaptado de Béjar <i>et al</i> , 1993)	26
Figura 6 Fluxograma de análise dos dados	29
Figura 7 Gráfico de linhas para a vazão.....	30
Figura 8 Gráfico de linhas para o pH	31
Figura 9 Gráfico de correlação para o pH	31
Figura 10 Gráfico de linhas para a condutividade	32
Figura 11 Gráfico de correlação para a condutividade.....	33
Figura 12 Gráfico de correlação para a DBO	34
Figura 13 Gráfico de correlação para a DBO	34
Figura 14 Gráfico de linhas para a DQO	35
Figura 15 Gráfico de correlação para a DQO.....	36
Figura 16 Gráfico de linhas para os SS	37
Figura 17 Gráfico de linhas para os SED.....	37
Figura 18 Gráfico de correlação para os SS	38
Figura 19 Gráfico de correlação para os SED.....	38
Figura 20 Gráfico de linhas para os SSV	39
Figura 21 Gráfico de correlação para os SSV.....	40

Figura 22 Gráfico de correlação das variáveis	41
Figura 23 Gráficos de caixa das variáveis originais	43
Figura 24 Gráficos de caixa das variáveis autoescaladas	44
Figura 25 Gráfico de diagnóstico (PCA clássico)	46
Figura 26 Gráfico de diagnóstico (PCA robusto)	47
Figura 27 Outliers sobrepostos nas variáveis autoescaladas.....	48
Figura 28 Gráfico dos conjuntos de treinamento (vermelho) e de teste (preto) em componentes principais.....	49
Figura 29 Valores preditos x Valores reais (Treinamento).....	50
Figura 30 Análise dos resíduos do modelo linear.....	51
Figura 31 Valores preditos x Valores reais (Teste)	52
Figura 32 Seleção do N° de componentes (onesigma)	53
Figura 33 Seleção do N° de componentes (randomization)	53
Figura 34 Valores preditos x Valores reais (Treinamento).....	54
Figura 35 Valores preditos x Valores reais (Teste)	55
Figura 36 Seleção do N° de componentes (onesigma)	56
Figura 37 Seleção do N° de componentes (randomization)	56
Figura 38 Valores preditos x Valores reais (Treinamento).....	57
Figura 39 Valores preditos x Valores reais (Testes).....	57
Figura 40 Gráfico de ajuste do modelo SVM.....	59
Figura 41 Valores preditos x Valores reais (Treinamento).....	59
Figura 42 Análise dos resíduos (Treinamento).....	60
Figura 43 Valores preditos x Valores reais (Teste)	61
Figura 44 Análise dos resíduos (Teste)	61

LISTA DE TABELAS

Tabela 1 Parâmetros de mérito (RLM)	52
Tabela 2 Parâmetros de mérito (PCR)	55
Tabela 3 Parâmetros de mérito (PLSR)	58
Tabela 4 Parâmetros de mérito (SVM)	62
Tabela 5 Parâmetros de mérito dos modelos	62

SUMÁRIO

RESUMO.....	7
ABSTRACT	8
LISTA DE FIGURAS	9
LISTA DE TABELAS.....	11
SUMÁRIO	12
1. INTRODUÇÃO	14
2. OBJETIVOS	15
2.1. Objetivos Específicos	15
3. REVISÃO BIBLIOGRÁFICA.....	16
4. INTRODUÇÃO TEÓRICA	18
4.1. Regressão Linear Múltipla	18
4.1.1. Colinearidade.....	18
4.1.2. AIC.....	18
4.2. Regressão por Componentes Principais.....	18
4.3. Regressão por Mínimos Quadrados Parciais.....	20
4.4. Regressão por Máquina de Vetor de Suporte.....	20
4.5. Parâmetros de mérito.....	23
4.5.1. MAE	23
4.5.2. MSE	23
4.5.3. RMSE	23
4.5.4. R ²	24
4.6. Análise dos resíduos.....	24
4.7. Validação.....	24
4.7.1. Validação externa	25
4.7.2. Validação interna	25
5. BANCO DE DADOS	26
5.1. Visualização dos dados	29
5.1.1. Vazão.....	29
5.1.2. Potencial de Hidrogênio	30
5.1.3. Condutividade	31
5.1.4. Demanda Bioquímica de Oxigênio	33
5.1.5. Demanda Química de Oxigênio	35
5.1.6. Sólidos em Suspensão e Sólidos Sedimentáveis	36
5.1.7. Sólidos Suspensos Voláteis.....	39

5.1.8. Pré-processamento dos dados.....	42
6. DETECÇÃO DE OUTLIERS.....	45
7. PARTIÇÃO DAS AMOSTRAS	49
8. MODELOS DE REGRESSÃO	50
8.1. Regressão Linear Múltipla	50
8.2. Regressão por Componentes Principais	52
8.3. Regressão por Mínimos Quadrados Parciais	55
8.4. Máquina de Vetor de Suporte	58
9. CONCLUSÃO.....	63
10. REFERÊNCIAS BIBLIOGRÁFICAS.....	64
11. ANEXOS	67
Anexo A - Comandos no R	67
Anexo B - Sumário do modelo de regressão linear múltipla.....	81
Anexo C - Sumário do modelo de regressão linear múltipla ajustado	82
Anexo D - Sumário do modelo de regressão por componentes principais	83
Anexo E - Sumário do modelo de regressão por mínimos quadrados parciais	84

1. INTRODUÇÃO

A estatística na gestão ambiental possui o objetivo de descrever, o melhor possível, as informações sobre o ambiente, com as mudanças que ocorrem ao longo do tempo e as variáveis que influenciam, melhorando assim o nosso conhecimento sobre o ambiente, facilitando as tomadas de decisões e fornecendo informações para o público em geral e usuários específicos.

A química ambiental atual dispõe de diversas informações altamente complexas sobre os processos ocorridos no meio ambiente. Estes processos ocorrem em sistemas abertos, sendo irreversíveis e influenciados por fenômenos físicos e biológicos. Os avanços da química, da eletrônica e da computação forneceram ferramentas poderosas para trabalhar com mais detalhes as informações sobre processos ambientais. Neste contexto, a quimiometria é uma área da química, relativamente jovem, que tem por objetivo a utilização de métodos matemáticos, de lógica e estatísticos, para avaliar e interpretar os dados químicos e analíticos, otimizando e modelando os processos (Pérez-Benedito e Rubio, 1999).

O desempenho de uma estação de tratamento de efluentes (ETE) é um desafio da engenharia, com diversos tipos de controles que devem estar em perfeito estado de funcionamento para que atuem corretamente e garantam que o efluente estará adequado para o lançamento no corpo hídrico receptor. Uma água residuária imprópria para a emissão pode causar sérios problemas às pessoas e ao meio ambiente e com isso pesadas multas à instituição geradora.

2. OBJETIVOS

Utilizar ferramentas da estatística multivariada para a análise dos dados de uma Estação de Tratamento de Efluentes, visualizando as variáveis do banco de dados de uma maneira geral e criando um modelo para a predição dos valores da Demanda Química de Oxigênio (DQO) na saída da ETE, logo que a DQO é um parâmetro mais rápido e fácil de determinar que a Demanda Bioquímica de Oxigênio (DBO) e pode ser facilmente correlacionado com a DBO.

2.1. Objetivos Específicos

- Visualizar as variáveis do banco de dados e a relação entre as mesmas;
- Gerar modelos de Regressão Linear Múltipla, por Componentes Principais, por Mínimos Quadrados Parciais e por Máquina de Vetor de Suporte;
- Comparar os modelos e avaliar qual o melhor na predição dos dados.

3. REVISÃO BIBLIOGRÁFICA

Sistema de tratamento de esgoto é um processo biológico complexo, não-linear e multivariado, com diversas reações químicas e elevada variabilidade na carga de entrada, ou seja, é um processo difícil de descrever matematicamente. Portanto, a previsão da qualidade do efluente da estação de tratamento de esgoto através de um modelo matemático é um desafio.

Pérez-Benedito e Rubio (1999) comentaram a necessidade da análise multivariada dos dados ambientais devido aos processos naturais envolverem alterações multidimensionais e as substâncias possuírem diferentes comportamentos. Além dos dados ambientais serem de natureza variável, ou seja, há muita variabilidade espacial e temporal na distribuição do poluente.

No estudo sobre a modelagem de uma estação de tratamento de lodos ativados de Teppola *et al.* (1997), foram feitos modelos de regressão linear múltipla, porém a colinearidade entre as variáveis levou a problemas no modelo, enquanto para a regressão por mínimos quadrados parciais (do inglês, Partial Least Squares Regression – PLS) isso não foi um problema. Os resultados por PLS foram utilizados para obter uma visão mais profunda do processo, podendo isolar facilmente a perturbação e descobrir quais variáveis que causaram o desvio das condições normais de operação.

Teppola *et al.* (1998) fizeram um estudo combinando as técnicas PLS e fuzzy, para o monitoramento de uma estação de tratamento de águas residuais com lodo ativado. O PLS foi utilizado para extrair as informações úteis das variáveis de controle do processo, a fim de prever a variável resposta, neste caso, o índice de volume de lodo. Os valores de pontuação foram usados na fuzzy com o objetivo de classificação, pois as variáveis originais apresentavam colinearidade, e o uso dos dados brutos diretamente pela técnica fuzzy eram comprometidos. Deste modo, a transformação das variáveis originais em variáveis latentes foi crucial para a classificação dos dados.

Pons, Wu e Potier (2005) estimaram a DQO, os teores de amônia e nitrogênio orgânico em águas residuárias para o controle on-line das plantas de tratamento, utilizando PCR e PLS aplicados aos dados espectrométricos, como a turbidez, espectrometria de fluorescência sincronizada e espectrofotometria UV-visível. Os modelos baseados na espectrometria de UV-visível foram satisfatórios. Os espectros de fluorescência sincronizada tiveram que ser divididos em regiões de interesse que podem ser relacionadas a urina e aos ácidos húmicos e fúlvicos. Neste estudo, nenhuma diferença foi encontrada entre os modelos PCR e PLS quando aplicados aos dados espectrométricos UV-visíveis e turbidez. No entanto, os modelos precisaram ser adaptados para levar em consideração as variações da atividade humana.

Li-Juan e Chao-Bo (2008) utilizaram a regressão por vetores de suporte (do inglês, Support Vector Machine Regression - SVR) para definir um modelo de previsão dos

indicadores de qualidade DQO (Demanda Química de Oxigênio) e SS (Sólidos em suspensão) na saída de uma ETE que utiliza o sistema de lodos ativados. Foi usado a função de base radial como *kernel* do modelo de previsão sendo os seus parâmetros determinados por validação cruzada. Este estudo mostrou que o SVR foi uma poderosa ferramenta de previsão apesar do pequeno número de amostras.

4. INTRODUÇÃO TEÓRICA

Neste capítulo serão discutidos brevemente os modelos abordados neste trabalho, assim como os parâmetros de mérito utilizados para a avaliação dos modelos.

4.1. Regressão Linear Múltipla

A regressão múltipla é uma extensão da regressão linear simples, onde a variável resposta é modelada a relação à combinação linear de duas ou mais variáveis preditoras simultaneamente. A adição de mais variáveis independentes ao modelo de regressão possui o propósito de encontrar um melhor modelo preditivo em comparação aos modelos simples e investigar os efeitos individuais de cada variável independente (Logan, 2010). Abaixo é mostrado a equação do modelo linear múltiplo aditivo (4.1), onde β_0 é o valor da variável resposta (y) quando todas as inclinações parciais ($\beta_1, \beta_2, \dots, \beta_j$) são iguais a zero. As variáveis independentes são representadas por (x_1, x_2, \dots, x_j) e o erro aleatório ou resíduo é representado por ε_i .

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_j x_{ij} + \varepsilon_i \quad (4.1)$$

4.1.1. Colinearidade

A colinearidade ou a multicolinearidade, ocorre quando as variáveis preditoras do modelo estão correlacionadas entre si, causando efeitos prejudiciais no ajuste do modelo, como a instabilidade dos coeficientes de regressão, maiores erros padrão e intervalos de confiança, aumentando a chance de ocorrer o erro do tipo II, quando o teste de hipótese não consegue rejeitar uma hipótese falsa (Logan, 2010).

4.1.2. AIC

O Critério de Informação de Akaike (do inglês *Akaike's information criterion* –AIC) é baseado no equilíbrio entre a menor Soma dos Quadrados do Erro ou Resíduos (SQ_{res}) penalizando a quantidade de variáveis (p), ou seja, a sua complexidade (Sheater, 2009). A medida que o valor de AIC reduz, a soma de quadrados do erro diminui, melhorando o ajuste. A equação de AIC (4.2) está indicada abaixo, onde n é o número de amostras:

$$AIC = n \log \left(\frac{SQ_{res}}{n} \right) + 2p \quad (4.2)$$

4.2. Regressão por Componentes Principais

A regressão por componentes principais (do inglês, *Principal Component Regression* – PCR) utiliza as pontuações, ou escores, em vez dos dados originais, como variáveis

independentes na regressão. Este artifício matemático contorna os problemas relacionados como a colinearidade entre as variáveis preditoras, pois as pontuações são ortogonais, ou seja, não são correlacionadas. Além disso, este recurso reduz o número de variáveis originais, sendo o modelo final muito mais reduzido em comparação à regressão linear múltipla, e por consequência, aumenta o número de graus de liberdade para a estimativa de erros (Wehrens, 2011).

O algoritmo mais comum para calcular os componentes principais (PCs) é o método da decomposição em valores singulares (do inglês, *singular value decomposition* - SVD) (4.3), que decompõe uma matriz de dados X em três partes:

$$X = UDV^T \quad (4.3)$$

Onde U e V são matrizes quadradas e ortonormais, com dimensões $(n \times n)$ e $(p \times p)$, respectivamente. Suas colunas são ortogonais e normalizadas, sendo $UU^T = U^T U = I_n$ e $VV^T = V^T V = I_p$, no qual I é a matriz identidade. A matriz D é retangular $(n \times p)$ com todos os valores fora da diagonal iguais a zero (Ferreira, 2015).

Deste modo, utilizando a terminologia do PCA (do inglês, *Principal Component Analysis*), temos:

$$X = (UD)V^T = TL^T \quad (4.4)$$

No qual T é a matriz dos escores (UD) e L é a matriz dos pesos (V^T) (do inglês, *loadings*). O modelo de regressão é construído utilizando a matriz reconstruída, onde as matrizes E , e ε representam as informações não explicadas pelos modelos, ou seja, os resíduos. Em (4.5) temos a matriz reconstituída. Com a substituição de (4.4) em (4.6) temos o modelo de regressão. A matriz dos coeficientes das pontuações está em (4.7). A matriz dos coeficientes das variáveis originais é fornecida em (4.8).

$$\tilde{X} = TL^T + E \quad (4.5)$$

$$Y = \tilde{X}B + \varepsilon = T(L^T B) + \varepsilon = TA + \varepsilon \quad (4.6)$$

$$A = (T^T T)^{-1} T^T Y = L^T B \quad (4.7)$$

$$B = LA = L(T^T T)^{-1} T^T Y \quad (4.8)$$

A matriz T é ortogonal e diagonal, uma vez que é o produto entre U e D , sendo D uma matriz diagonal e U uma matriz ortogonal e substituindo T em B , temos as pontuações (4.9). E o cálculo da matriz dos coeficientes das variáveis originais (4.10) é fornecida com a substituição de (4.4) em (4.8).

$$T = UD \quad (4.9)$$

$$B = L(DU^TUD)^{-1}DU^TY = LD^{-2}DU^TY = LD^{-1}U^TY \quad (4.10)$$

A técnica do PCR tem o problema adicional de estimar o número de PCs que devem ser retidos. Um dos critérios utilizados é a quantidade de variância de Y , variável resposta, que é explicada. No entanto, um valor mínimo da raiz quadrada do erro médio quadrático (do inglês, *Root Mean Squared Error* – RMSE), que será descrito a seguir, é usado com maior frequência (Wehrens, 2011).

4.3. Regressão por Mínimos Quadrados Parciais

A regressão por mínimos quadrados parciais (do inglês, *Partial Least Squares Regression* - PLSR), assim como o PCR, cria fatores ortogonais para compactar as informações, sendo este o método mais popular na quimiometria. No PLS, os fatores são designados como variáveis latentes e diferentemente do PCR, não são obtidos pela decomposição SVD, mas são definidos de modo a manter um compromisso entre a variância explicada em X e a previsão da variável resposta (Y) (Ferreira, 2015).

As equações gerais do modelo de regressão PLS são as mesmas usadas para construir o modelo PCR, são as equações 4 e 5. Mas, no PLSR a variável latente, que correlaciona X e Y , é obtida maximizando a covariância entre os escores T . Deste modo, é otimizada a decomposição em relação à previsão da variável resposta. O termo “parcial”, referente ao nome deste método, é devido ao fato desta solução por mínimos quadrados não ser aplicada a qualquer conjunto de pesos, mas somente àquele satisfaz a maximização da covariância entre os escores.

4.4. Regressão por Máquina de Vetor de Suporte

As máquinas de vetores de suporte (do inglês, *Support Vector Machine* – SVM) foram desenvolvidas nos anos 60, na Rússia, por Vapnik, Lerner e Chervonenkis, no entanto, a forma mais atual foi desenvolvida por Vapnik no final dos anos 90 (Vapnik, 1999). Esta ferramenta é uma metodologia de aprendizagem de máquina que foi criada para classificação de amostras de duas classes diferentes, porém com as adaptações recentes (Vapnik, 1999), o SVM pode ser aplicado para regressão.

As SVMs usam um mapeamento implícito (Φ) dos dados de entrada, transformando-os no “espaço de características” com dimensão maior do que a original, onde os dados são resolvidos por um problema linear. Isto é feito pela função *kernel*, uma função retornando o produto interno ($\Phi(x), \Phi(x')$) entre as imagens de dois pontos de dados (x, x') no espaço de característica (Karatzoglou *et al.*, 2006). Neste estudo a função *kernel* utilizada foi a função de base radial (do inglês, *Radial Basis Function*).

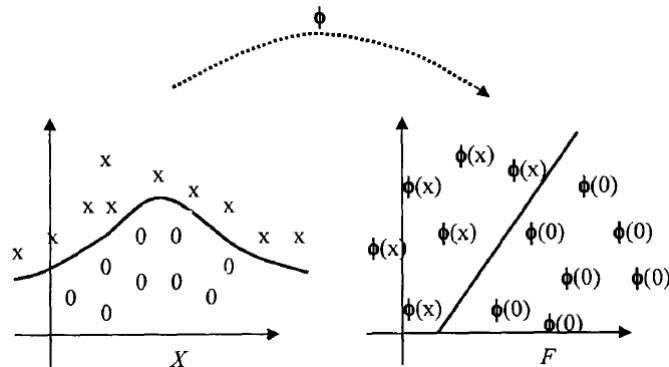


Figura 1 Mapeamento do espaço de entrada em um espaço de características (Chen *et al.*, 2004)

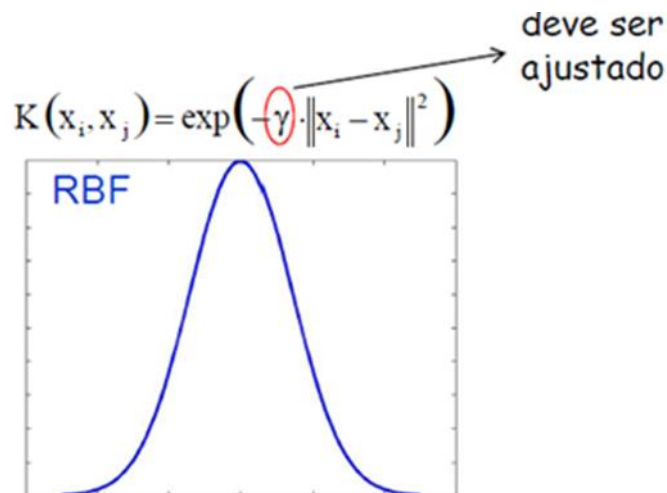


Figura 2 Função de base radial (RBF) (Adaptado de Poppi, 2015)

Para cada amostra (x_i) dos dados, o valor de resposta correspondente (y_i) é adicionado por um número positivo ϵ para produzir uma nova amostra ($f(x)+\epsilon$), que pertencente ao Grupo 1. Da mesma forma, o y_i também é subtraído por ϵ para produzir outra nova amostra ($f(x)-\epsilon$), que pertencente ao Grupo 2. Repetindo este procedimento, as N amostras da regressão são duplicadas e classificadas em dois grupos, e o hiperplano ótimo de separação pode ser obtido pelo ajuste $f(x)=wx+b=0$. Deste modo, o problema de regressão é transformado em um problema de classificação binária (Li *et al.*, 2009).

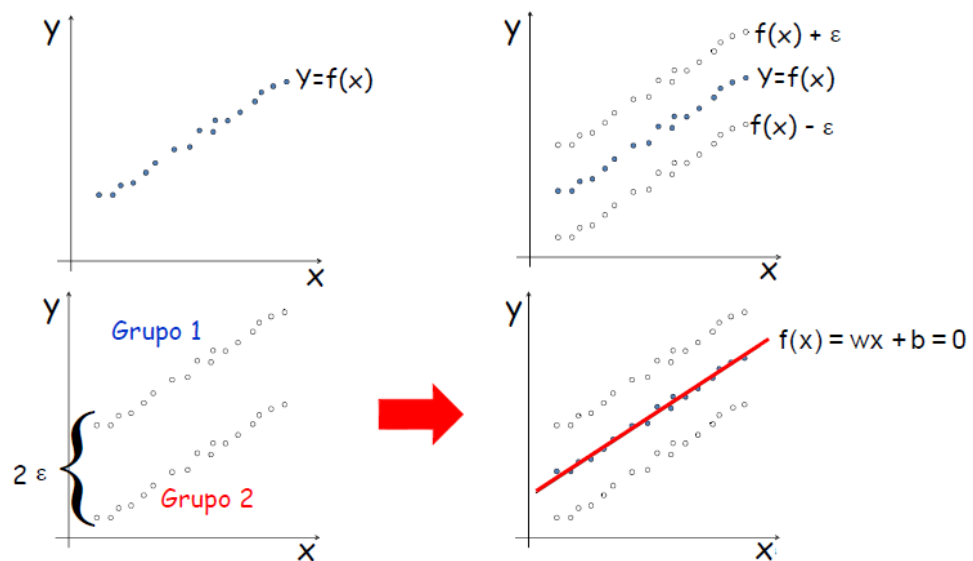


Figura 3 Transformação de um problema de regressão linear em classificação binária
(Adaptado de Poppi, 2015)

Observe que a superfície de separação é dada pela equação $f(x)=(w^T x)+b=0$, onde $w \in \mathbb{R}^m$ é o vetor de pesos, e $b \in \mathbb{R}$ é o intercepto. Deste modo, a separação entre o hiperplano e o dado de entrada mais próximo é chamada de margem de separação, ε .

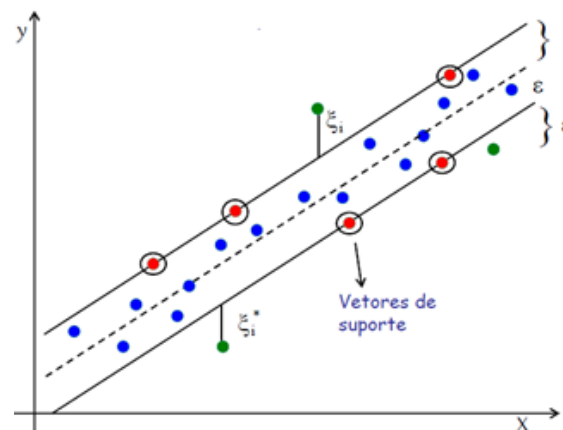


Figura 4 Geometria do hiperplano (Adaptado de Poppi, 2015)

Os dados para que se encontram a uma distância ε do hiperplano são chamados de vetores de suporte (x^{sv}) e possuem um papel crucial na localização do hiperplano. Observe que quando $(w^T x^{sv+})+b=+1$, o vetor de suporte pertence à classe +1 (Grupo 1) e quando $(w^T x^{sv-})+b=-1$, o vetor de suporte pertence à classe -1 (Grupo 2). Em nosso caso, para a regressão, devemos minimizar a margem de separação e o somatório dos erros, $\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*)$, para que o hiperplano esteja o mais próximo possível dos dados, literalmente passando pelos dados. Observe que C é o custo, ou fator de penalização, que influencia diretamente no desempenho da regressão por vetores de suporte (do inglês, *Support Vector Regression* –

SVR). O aumento do fator de penalização reduz o erro de predição, enquanto a redução deste fator aumenta a margem de separação (Salcedo-Sanz, 2014).

4.5. Parâmetros de mérito

Um determinado modelo de regressão é preciso somente se possuir um erro padrão baixo em relação à variável resposta. Para isto, são usados alguns parâmetros para a estimativa do erro.

4.5.1. MAE

O erro absoluto médio (do inglês, *Mean Absolute Error* – MAE) é dado pelo somatório das diferenças entre os valores preditos pelo modelo e os valores reais da resposta. Este parâmetro é menos sensível a valores elevados dos erros, sendo assim um parâmetro mais robusto. Além de possuir a mesma unidade da variável resposta.

$$MAE = \frac{\sum_{t=1}^n |\hat{y}_t - y_t|}{n} \quad (4.11)$$

4.5.2. MSE

O erro médio quadrático (do inglês, *Mean Squared Error* – MSE) é dado pela média dos quadrados dos erros, deste modo é um estimador de erro que sempre fornece um valor positivo. Este parâmetro também é sensível a pequenas variações no erro de predição.

$$MSE = \frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n} \quad (4.12)$$

4.5.3. RMSE

A raiz quadrada do erro médio quadrático (do inglês, *Root Mean Squared Error* – RMSE) é dado pela raiz quadrada da média dos quadrados dos erros, este estimador de erro, do mesmo modo que o anterior, sempre fornece um valor positivo, porém possui unidade igual ao da variável resposta, além de ser menos sensível a variações no erro de predição. Segundo Ferreira (2015) e Esbensen e Geladi (2010), o RMSE é o parâmetro mais utilizado para avaliar a qualidade de predição e a escolha do número ótimo de fatores em um determinado modelo.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}} \quad (4.13)$$

4.5.4. R^2

O coeficiente de determinação (R^2) é uma medida que diz o quanto as variáveis preditoras do modelo explicam a variável resposta. Este valor é dado pela divisão entre a soma dos quadrados da diferença entre os valores preditos pelo modelo (\hat{y}_i) e a média dos valores observados (\bar{y}_i), e a soma total dos quadrados, que é a soma dos quadrados das diferenças entre os valores observados (y_i) e a média dos valores observados. Este parâmetro não é uma medida do erro, mas do grau de ajuste obtido entre as variáveis preditoras e a variável resposta, ou seja, o R^2 é uma medida de correlação e não de precisão (Kuhn e Johnson, 2013).

$$R^2 = \frac{SQ_{exp}}{SQ_{tot}} = \frac{SQ_{exp}}{SQ_{exp} + SQ_{res}} = 1 - \frac{SQ_{res}}{SQ_{tot}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.14)$$

4.6. Análise dos resíduos

A análise dos resíduos é uma ferramenta importante de diagnóstico do modelo de regressão. Esses gráficos nos permitirão avaliar visualmente se um modelo foi ajustado aos dados, não importando quantas variáveis independentes sejam usadas. Os resíduos não podem produzir uma variabilidade heterogênea (heterocedasticidade) ou um padrão reconhecível na plotagem dos resíduos contra os valores ajustados, ou seja, os resíduos devem se distribuir de forma aleatória e com variância constante (homocedasticidade) em torno do zero na horizontal, indicando assim que não há erros de tendência no modelo e de falta de ajuste (Sheather, 2009).

Os *Leverages*, termo comum na estatística utilizado para “pontos influentes” na língua inglesa, são resíduos que indicam a presença de um ponto que pode prejudicar o modelo, ou seja, este tem o potencial para ser um valor atípico (*outlier*). Deste modo, o gráfico dos resíduos padronizados contra o *Leverage* podem indicar a presença de pontos influentes e atípicos, porém mais testes são feitos para a confirmação da presença de valores atípicos.

O gráfico normal Q-Q é uma ferramenta visual que permite observar se a distribuição dos resíduos segue uma distribuição normal. Isto ocorre quando os resíduos estão mais próximos da reta normal. Caso haja desvio de um ou mais pontos reta normal, pode ser o indicativo de um ponto influente (*Leverage*), sendo que estes pontos influentes podem afetar o modelo (Sheather, 2009).

4.7. Validação

A validação, em regressão multivariada, expressa avaliar ou comprovar que o desempenho da previsão é válido, ou seja, o objetivo da validação é confirmar que um determinado modelo de previsão funcionará de acordo com o seu propósito. Não se referindo

apenas ao comportamento na situação de modelagem, como também ao comportamento em relação ao desempenho futuro com novos "dados similares" (Esbensen e Geladi, 2010).

4.7.1. Validação externa

O banco de dados, com n amostras, é particionado em dois, o de calibração (amostras C) que é usado para construir o modelo, e o de testes (amostras T) para validação, onde o total de amostras é dado por $n = C + T$. O termo em inglês para este tipo de validação é *Hold-out* (Esbensen e Geladi, 2010). Este método é mais conservador e robusto, pois o modelo é testado em um conjunto de teste independente e representativo de tamanho suficiente (Westad e Marini, 2015).

4.7.2. Validação interna

Na validação interna, as amostras do banco de calibração são utilizadas para a validação do modelo a partir de técnicas de reamostragem, com o objetivo de estimar o número de variáveis latentes e a incerteza de variáveis individuais para encontrar quais são relevantes dentre muitas variáveis (Westad e Marini, 2015). Deste modo, neste trabalho foram utilizadas as seguintes técnicas:

- *Leave-one-out* (LOO) ou *Jackknife* – cada amostra é deixada de lado uma vez, assim, o modelo é construído com $n-1$ amostras e são feitos n modelos.
- *k-fold* – o conjunto de dados é particionado em partes n/k , sendo construídos n/k modelos. Esta abordagem também é chamada de validação segmentada, logo que é deixado de fora uma fração ou segmento de amostras (Esbensen e Geladi, 2010).

5. BANCO DE DADOS

Os dados são provenientes de uma ETE que utiliza o processo biológico de lodos ativados para a remoção da matéria orgânica. O processo de lodos ativados faz uso dos microrganismos presentes no próprio efluente para consumir a matéria orgânica, agregando-a em sua própria biomassa. Devido a elevada taxa de crescimento microbiano é necessária a adição de oxigênio dissolvido por meio de aeração externa.

A figura abaixo é um esquema da planta estudada, que foi adaptada do artigo escrito por Béjar *et al.* (1993). Como descrito no periódico, o pré-tratamento é composto por gradeamento, que é seguido de um decantador primário, dispondo do lodo primário como rejeito. Temos em seguida o reator biológico por lodos ativados, com sistema de aeração forçada, sendo sucedido por um decantador secundário para a remoção do lodo decantado, com sistema de reciclo de parte do lodo para o reator biológico, mantendo um nível adequado de material biológico no reator para a oxidação da matéria orgânica.

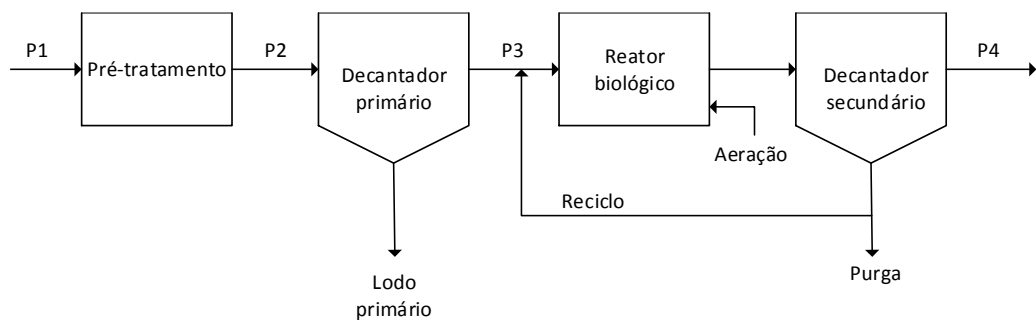


Figura 5 Diagrama do processo da ETE (Adaptado de Béjar *et al.*, 1993)

A planta estudada está situada em Manresa, um município na Espanha na província de Barcelona, comunidade autônoma da Catalunha, com 100.000 habitantes. A estação trata 35.000 m³/dia de esgoto doméstico e efluentes de indústrias localizadas na região.

Os dados originais são constituídos por 38 variáveis numéricas e contínuas, medidas na entrada (P1), após o pré-tratamento (P2), na entrada do reator biológico (P3) e na saída da planta (P4), sendo monitoradas por 527 dias. As variáveis são:

- 1 Q-E (vazão de entrada)
- 2 ZN-E (entrada de zinco)
- 3 PH-E (pH na entrada)
- 4 DBO-E (demanda bioquímica de oxigênio de entrada)
- 5 DQO-E (demanda química de oxigênio de entrada)
- 6 SS-E (sólidos suspensos na entrada)
- 7 SSV-E (sólidos suspensos voláteis na entrada)
- 8 SED-E (sedimentáveis na entrada)

- 9 COND-E (condutividade na entrada)
- 10 PH-P (pH na entrada do decantador primário)
- 11 DBO-P (DBO na entrada do decantador primário)
- 12 SS-P (SS na entrada do decantador primário)
- 13 SSV-P (SSV na entrada do decantador primário)
- 14 SED-P (SED na entrada do decantador primário)
- 15 COND-P (condutividade na entrada do decantador primário)
- 16 PH-D (pH no decantador secundário)
- 17 DBO-D (DBO na entrada do decantador secundário)
- 18 DQO-D (DQO na entrada do decantador secundário)
- 19 SS-D (SS na entrada do decantador secundário)
- 20 SSV-D (SSV na entrada do decantador secundário)
- 21 SED-D (SED na entrada do decantador secundário)
- 22 COND-D (condutividade na entrada do decantador secundário)
- 23 PH-S (pH na saída)
- 24 DBO-S (DBO na saída)
- 25 DQO-S (DQO na saída)
- 26 SS-S (SS na saída)
- 27 SSV-S (SSV na saída)
- 28 SED-S (SED na saída)
- 29 COND-S (condutividade na saída)
- 30 RD-DBO-P (rendimento de remoção da DBO no decantador primário)
- 31 RD-SS-P (rendimento de remoção de SS no decantador primário)
- 32 RD-SED-P (rendimento de remoção dos SED no decantador primário)
- 33 RD-DBO-S (rendimento de remoção da DBO no decantador secundário)
- 34 RD-DQO-S (rendimento de remoção da DQO no decantador secundário)
- 35 RD-DBO-G (rendimento global de remoção da DBO)
- 36 RD-DQO-G (rendimento global de remoção da DQO)
- 37 RD-SS-G (rendimento global de remoção de SS)
- 38 RD-SED-G (rendimento global de remoção de SED)

O conjunto de dados foi disponibilizado por Javier Bejar e Ulises Cortes (1992) no banco de dados internacional *UCI Machine Learning Repository*. Belanche *et al.* (1992) desenvolveram uma ferramenta de diagnóstico e controle da ETE baseado no conhecimento operacional e em ferramentas de classificação, assim auxiliando no gerenciamento da estação. O trabalho de Béjar *et al.* (1993) classificou os dados de acordo com o estado operacional da planta utilizando para este fim as variáveis de estado e o conhecimento dos operadores em cada uma das etapas do processo, sendo declarado como um domínio mal

estruturado. A distribuição das classes foi feita pelo algoritmo conceitual de agrupamento, e foram dadas por:

- Classe 1: Situação Normal (275 dias)
- Classe 2: Problemas no decantador secundário-1 (1 dias)
- Classe 3: Problemas no decantador secundário -2 (1 dias)
- Classe 4: Problemas no decantador secundário -3 (1 dias)
- Classe 5: Situação Normal com eficiência acima da média (116 dias)
- Classe 6: Sobrecarga de sólidos-1 (3 dias)
- Classe 7: Problemas no decantador secundário -4 (1 dias)
- Classe 8: Tempestade-1 (1 dias)
- Classe 9: Situação Normal com pouco efluente (69 dias)
- Classe 10: Tempestade-2 (1 dias)
- Classe 11: Situação Normal (53 dias)
- Classe 12: Tempestade -3 (1 dias)
- Classe 13: Sobrecarga de sólidos -2 (1 dias)

Foi feito o download dos arquivos no domínio da *UCI Machine Learning* no formato de bloco de notas. Este foi importado para o formato da planilha do Excel e foram atribuídas as classes do conjunto descrito. O conjunto consiste nas 38 variáveis descritas, com 527 observações coletadas entre o período de 01/01/1990 e 30/10/1991, sendo estas ordenadas de acordo com a data e com 13 classes.

A análise dos dados, assim como todos os gráficos produzidos neste estudo, foram feitos no *software* R, com a interface RStudio, que é um *software* livre e desenvolvido pelo R Core Team (2017). O pacote *rmarkdown* (Allaire *et al.*, 2017) foi utilizado para a geração dos documentos no formato “.docx”, para a leitura em texto, e também no formato “.html” para melhor visualização dos dados e gráficos gerados pelos códigos de comando. Todas as linhas de comando escritas neste trabalho estão dispostas no Anexo A.

Para a predição da DQO na saída da estação em situação de operação normal, foram selecionados somente as observações onde não há situações atípicas, como tempestades, sobrecarga de sólidos e problemas no decantador secundário. As variáveis imputadas nos modelos foram consideradas até o decantador secundário, sendo as demais variáveis na saída não contabilizadas para a predição da DQO. A variável de entrada Zinco também não foi computada no modelo, pois esta foi tomada somente na entrada do processo, sendo ignorada nos demais pontos de coleta.

O banco possui valores faltantes que foram omitidos, assim reduzindo a quantidade de observações de 527 para 395. A omissão de valores faltantes foi escolhida, pois a inserção destes dados na estatística multivariada pode levar a uma estimativa tendenciosa, como foi o resultado da pesquisa do Gorelick (2004) onde a imputação de valores faltantes levou a

estimativas enviesadas dos modelos preditivos. Na investigação de Yadav e Roychoudhury (2018), verificou-se que o efeito da imputação de valores faltantes resultou na alteração da variância dos dados, assim como no desempenho preditivo dos modelos desenvolvidos, além disso o desempenho da modelagem, assim como a variância dos dados mudou conforme a quantidade de informações faltantes.

A sequência de etapas feitas no presente trabalho está disposta no fluxograma abaixo, que facilita a compreensão do trabalho como um todo.

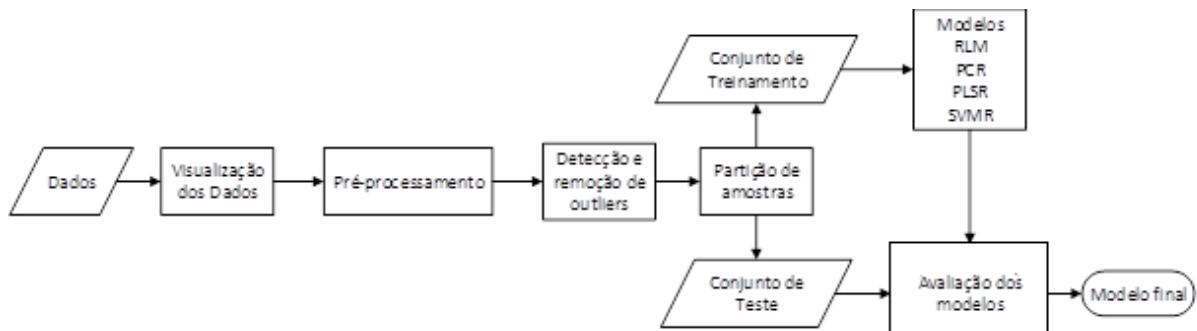


Figura 6 Fluxograma de análise dos dados

5.1. Visualização dos dados

Para iniciar o tratamento dos dados é importante ter uma visão geral destas informações. Deste modo as variáveis foram agrupadas de acordo com as suas características, sendo feito um gráfico de linhas para cada grupo. Os gráficos de correlação foram feitos utilizando como métrica a correlação de Spearman (5.1), pois esta é mais robusta às variações da distribuição normal. Nos gráficos, o valor da correlação está na parte superior, na diagonal temos o histograma da variável, e na parte inferior, temos os gráficos de dispersão entre as variáveis.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^n (x_i - \bar{x})^2][\sum_{i=1}^n (y_i - \bar{y})^2]}} \quad (5.1)$$

5.1.1. Vazão

A medida da vazão é feita somente na entrada da estação, deste modo, deve ser considerado o regime em estado estacionário do processo de tratamento de efluentes, logo a tomada de vazão na entrada do processo é considerada constante por todo o tratamento. Observando o gráfico da vazão, percebe-se que há uma grande variabilidade, mas como não foi informado no artigo se há um tanque de específico para a equalização da vazão, podemos presumir que o tanque de pré-tratamento pode ser utilizado para este fim, nota-se que este parâmetro pode ter uma grande influência sobre os demais.

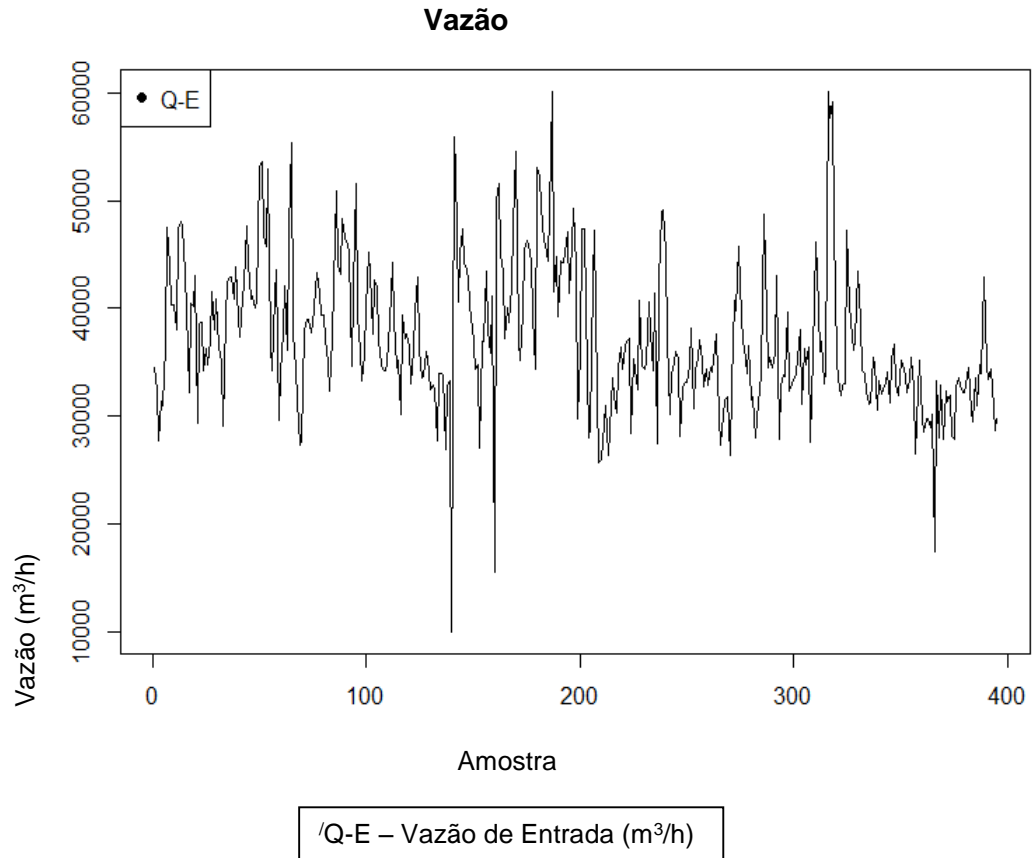
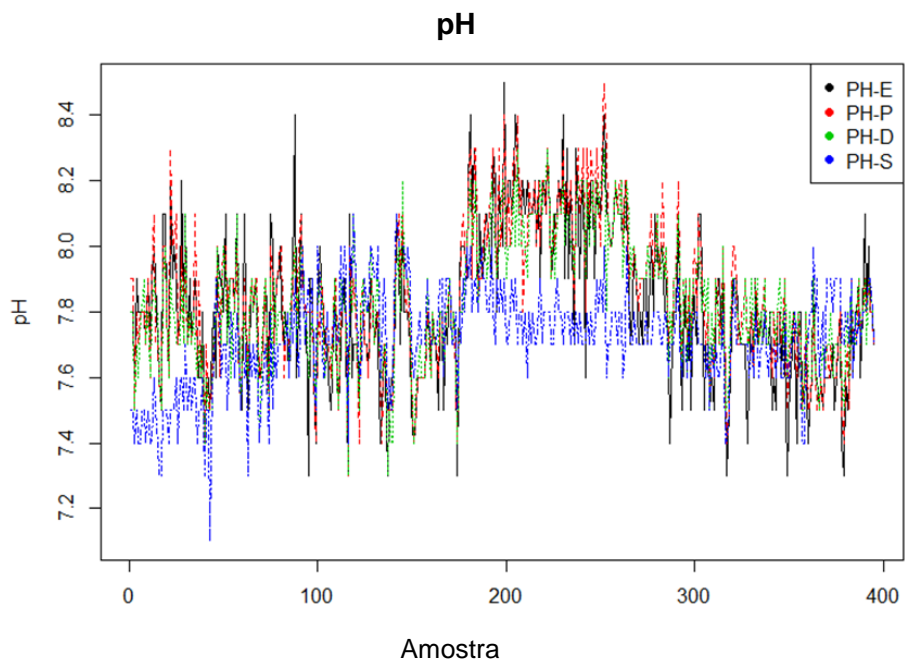


Figura 7 Gráfico de linhas para a vazão

5.1.2. Potencial de Hidrogênio

O gráfico do pH que este variou entre 7 e 8,5 nos quatro pontos de coleta na estação. Não houve uma alteração brusca do pH na saída, o que não é surpresa pois esta faixa de pH não é agressiva para os lodos ativados.



PH-E – pH na entrada
PH-P – pH na entrada do decantador primário
PH-D – pH na entrada do decantador secundário
PH-S – pH na saída

Figura 8 Gráfico de linhas para o pH

Há uma forte correlação para o parâmetro pH entre os pontos de coleta na entrada, no decantador primário e secundário, isto deve-se ao fato de não ocorrer nenhuma transformação química nas substâncias presentes no efluente nas etapas anteriores, e somente na saída há mudança, o que pode ser atribuída ao tratamento biológico do efluente.

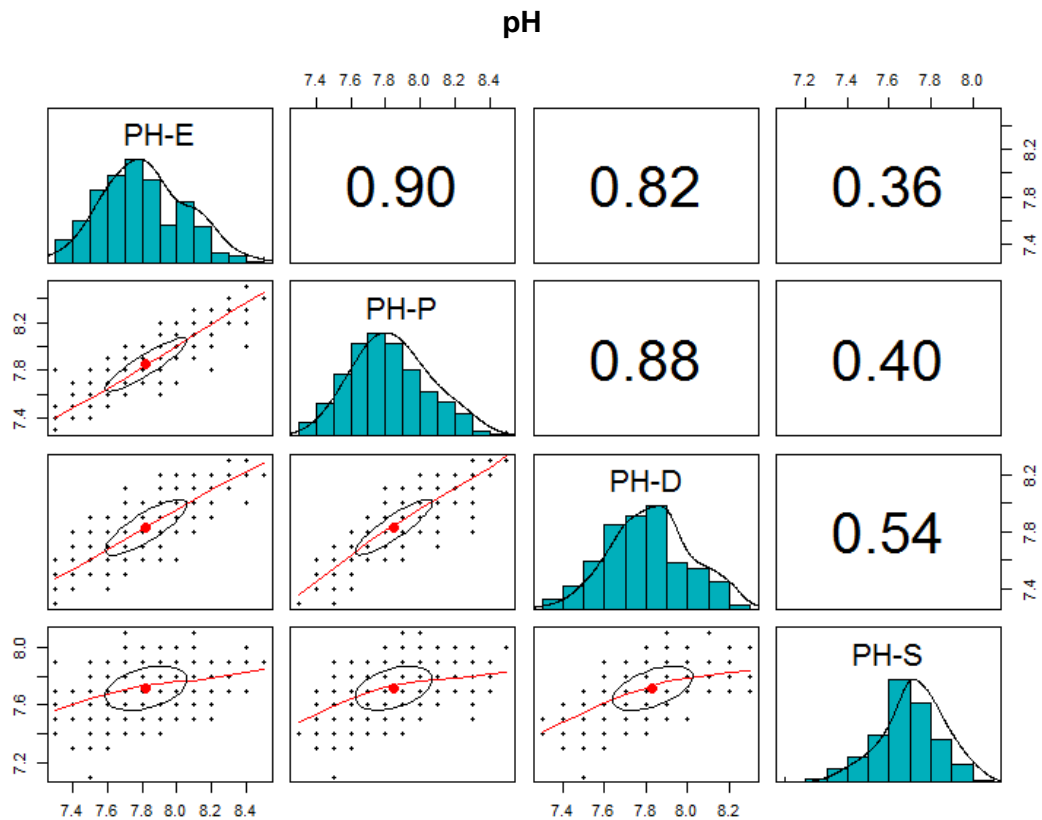


Figura 9 Gráfico de correlação para o pH

5.1.3. Condutividade

A Condutividade (COND) é mantida praticamente constante, com poucas alterações bruscas ao longo dos pontos de coleta na ETE, e percebe-se uma forte sobreposição das linhas ao longo dos pontos de amostragem. Logo, podemos considerar que a parcela de sais dissolvidos é praticamente constante ao longo do processo e o tratamento biológico não é capaz de lidar com a remoção destes sais.

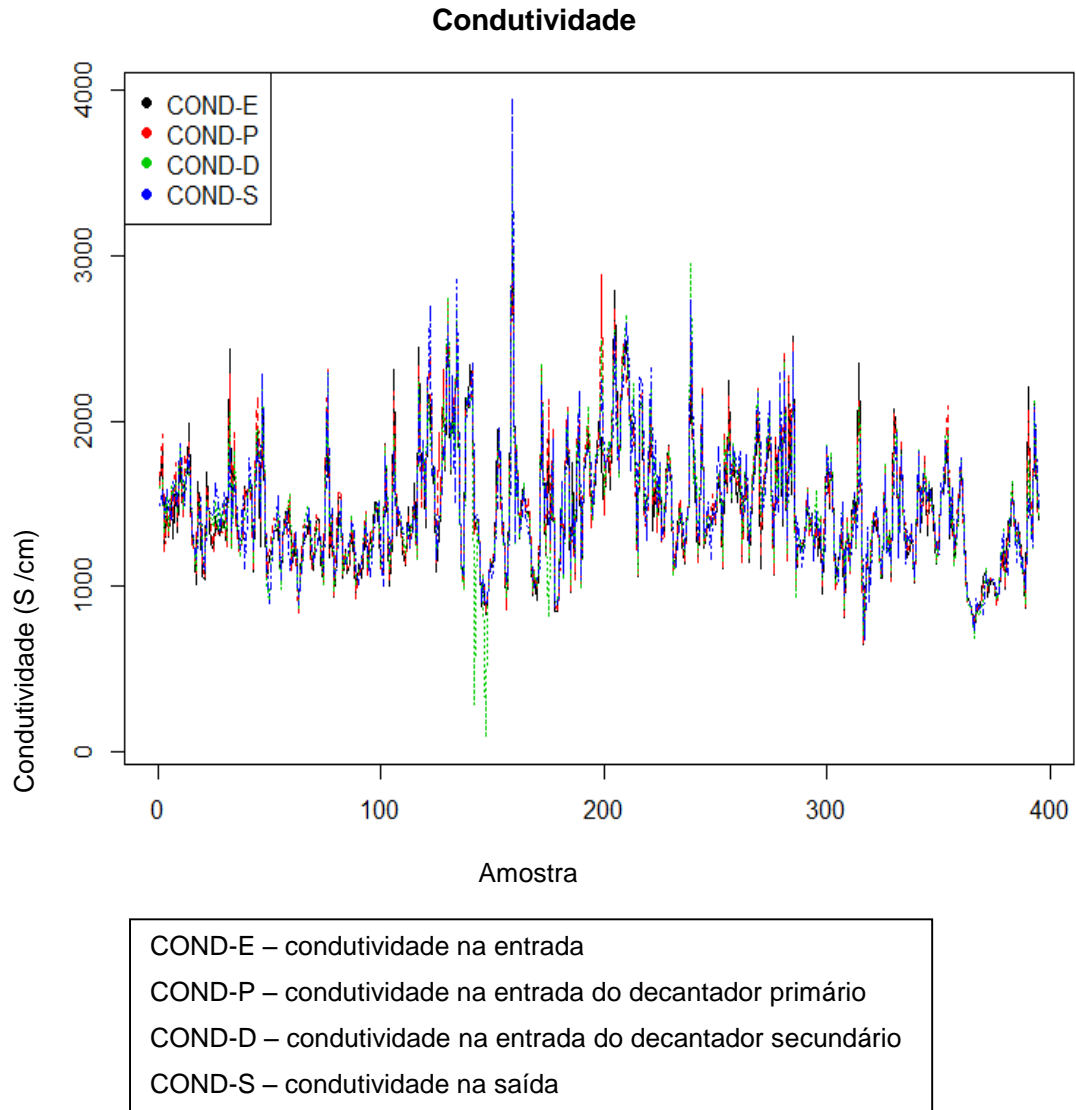


Figura 10 Gráfico de linhas para a condutividade

O gráfico de correlação para a condutividade confirma uma forte correlação deste parâmetro entre os pontos de coleta para ao longo do processo na ETE. Também é possível observar alguns valores atípicos, que não estão dispostos ao longo da dispersão de pontos na reta.

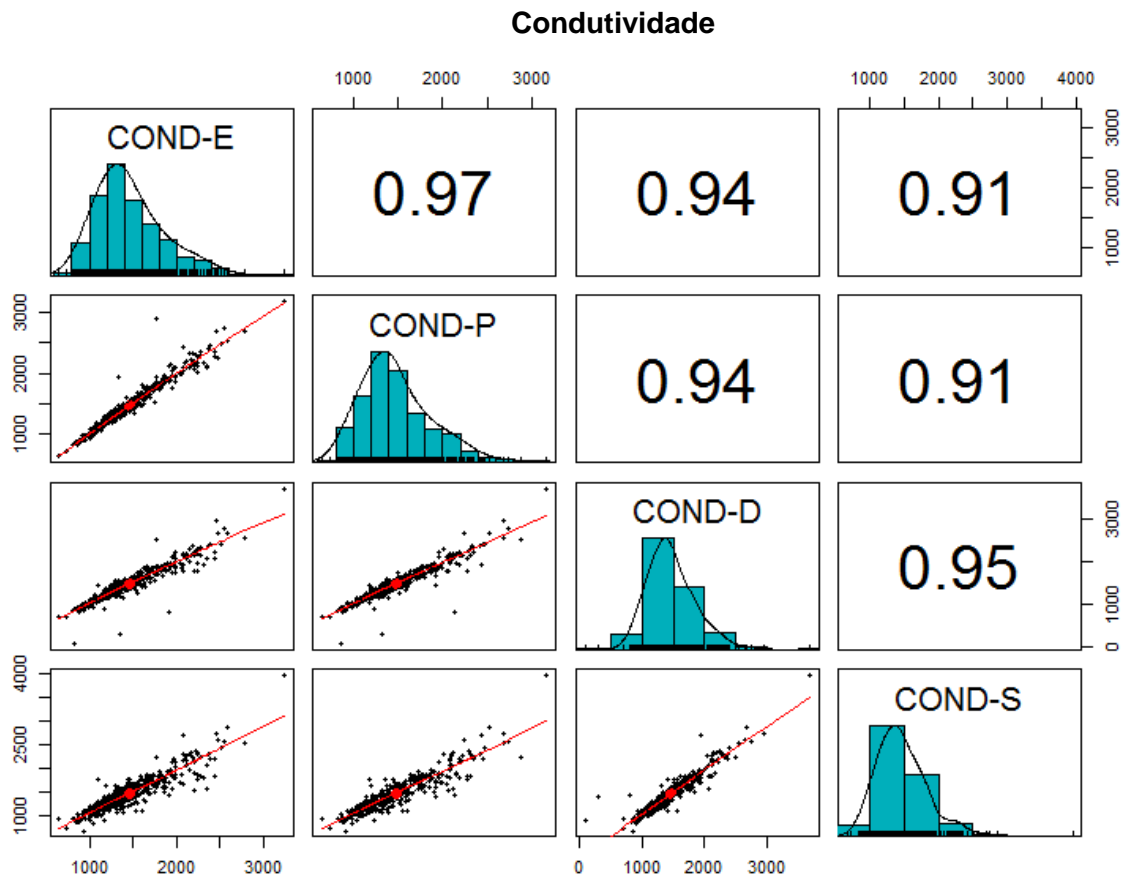
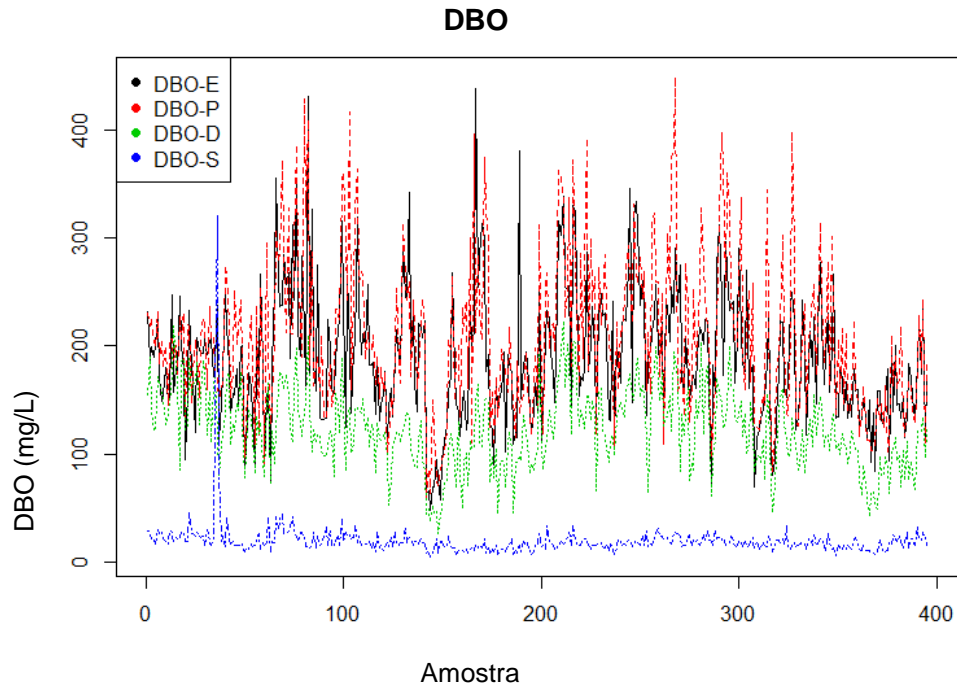


Figura 11 Gráfico de correlação para a condutividade

5.1.4. Demanda Bioquímica de Oxigênio

Observando o gráfico, podemos ter uma sequência de redução das linhas ao longo dos pontos de coleta na ETE. A DBO de entrada e na entrada do decantador primário são praticamente os mesmos valores, logo a maior parte da matéria orgânica biodegradável está suspensa ou dissolvida no efluente. Na entrada do decantador secundário podemos observar uma pequena queda, e um valor baixo de DBO na saída da estação, sendo assim, a soma das linhas da DBO na saída e no decantador secundário terá como resultado um valor próximo à DBO na entrada e no decantador primário, devido à sua correspondência com o balanço de massa. Portanto, podemos perceber pelo gráfico que a maior parte da matéria orgânica biodegradável foi convertida em biomassa e retida no decantador secundário.



DBO-E – DBO na entrada
 DBO-P – DBO na entrada do decantador primário
 DBO-D – DBO na entrada do decantador secundário
 DBO-S – DBO na saída

Figura 12 Gráfico de correlação para a DBO

Pode-se observar uma boa correlação para a DBO nas primeiras etapas de tratamento da estação, porém na saída há uma alteração abrupta deste parâmetro, devido há remoção de parte do lodo no decantador secundário, com isso tem-se uma correlação fraca entre a saída e os demais pontos da estação. Além disso, nota-se a presença de alguns valores atípicos desta variável na saída da estação.

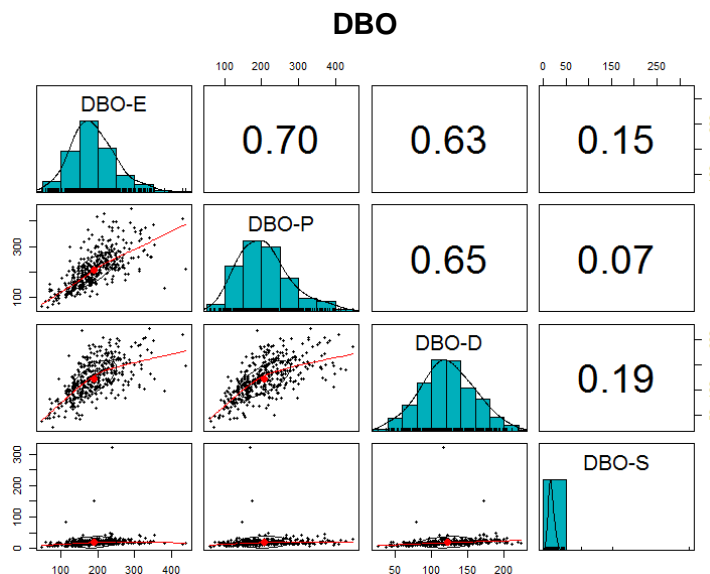


Figura 13 Gráfico de correlação para a DBO

5.1.5. Demanda Química de Oxigênio

A DQO possui um comportamento semelhante à DBO ao longo da ETE. Porém, a DQO possui uma faixa mais ampla de variabilidade, alcançando valores acima de 800 mg/L de O₂, enquanto a DBO está limitada em 500 mg/L de O₂, logo este efluente possui uma boa parcela de material não-biodegradável em sua composição, isto pode ser atribuído ao fato de haver indústrias que destinam suas águas residuárias para esta planta de tratamento. Comparando também a DBO e a DQO na saída, podemos observar que a DBO flutua em valores próximos a 10 mg/L de O₂, enquanto a DQO oscila próximo a 100 mg/L de O₂. Pelo fato de a planta operar com o sistema de lodos ativados, a degradação da matéria orgânica é muito maior quando comparada com a parcela não orgânica.

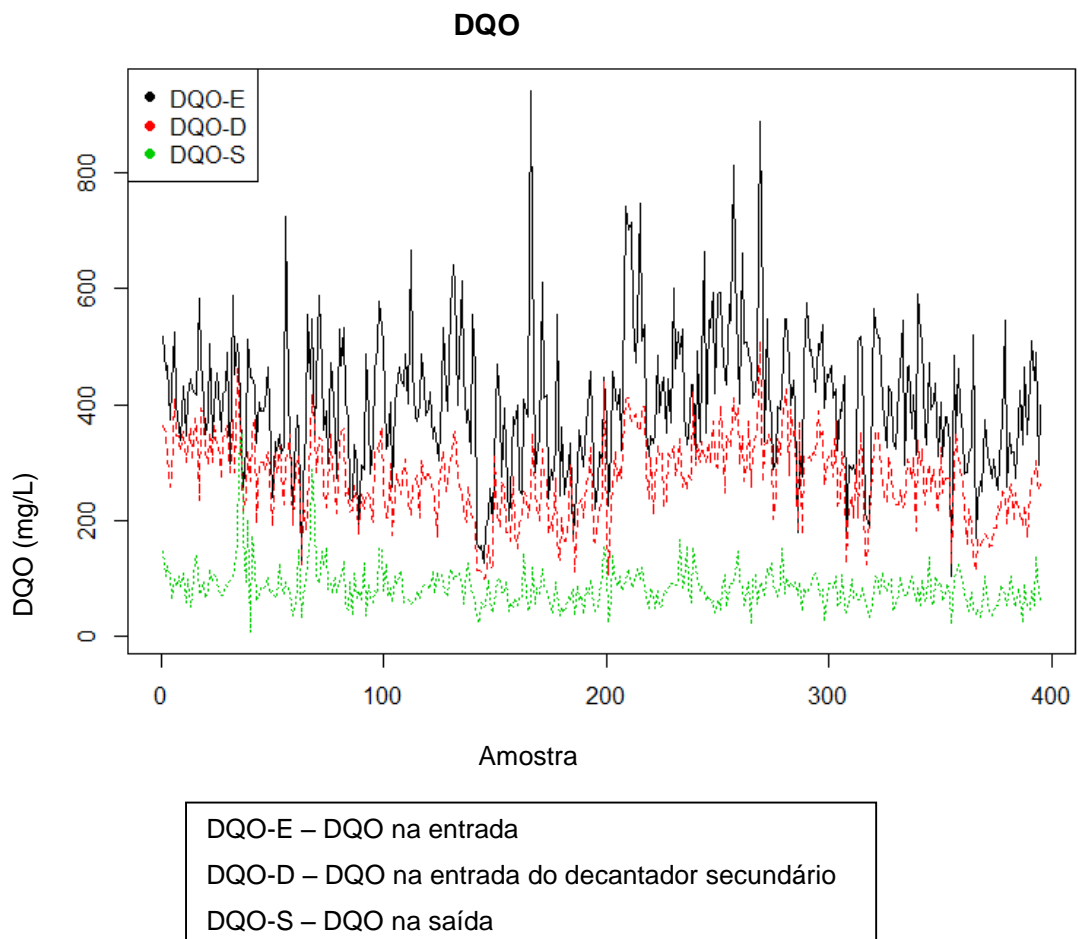
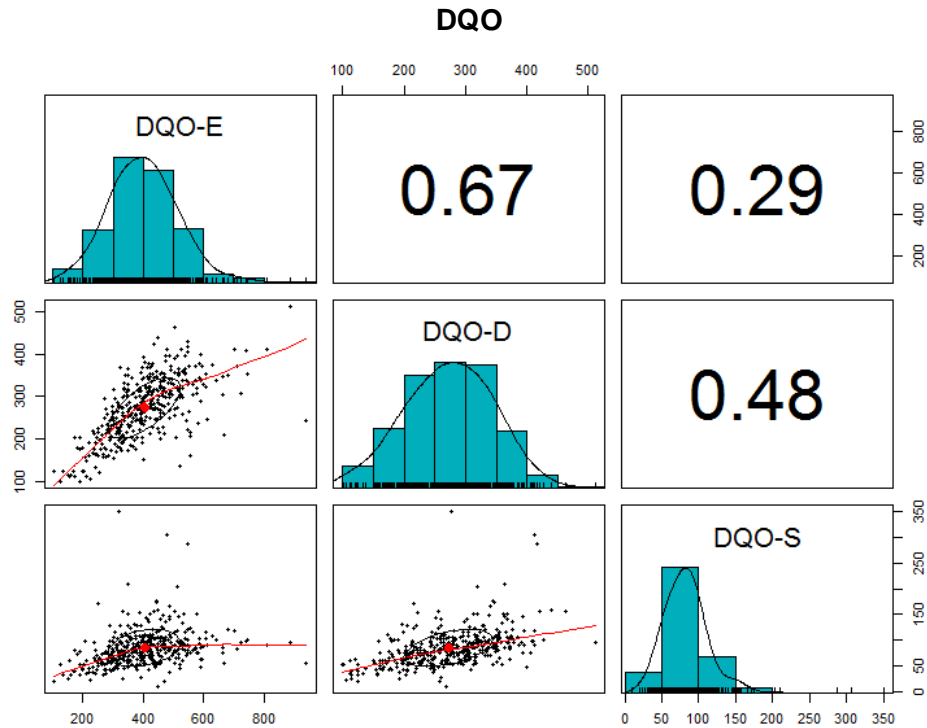


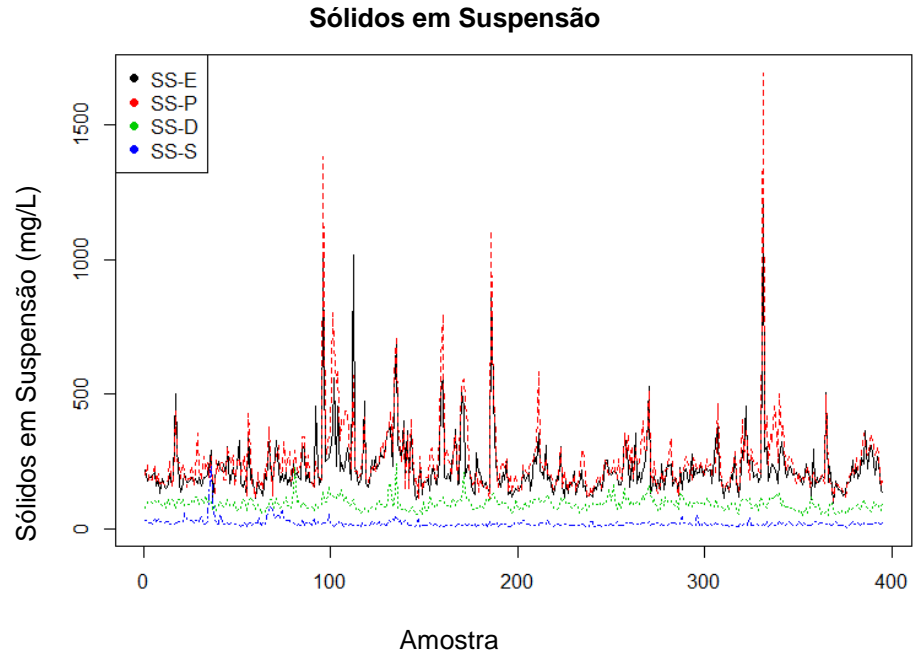
Figura 14 Gráfico de linhas para a DQO

De modo semelhante à DBO, temos uma boa correlação para a DQO nas primeiras etapas, e na saída há uma diminuição brusca deste parâmetro, mas esta redução é menor quando comparada à DBO. Assim como no caso anterior, temos valores atípicos na saída da estação.



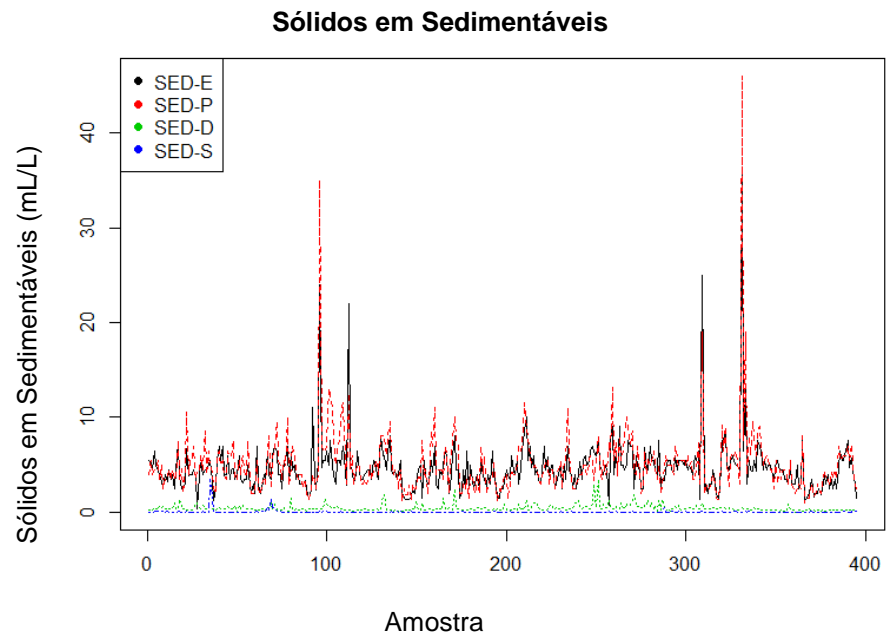
5.1.6. Sólidos em Suspensão e Sólidos Sedimentáveis

Os Sólidos em Suspensão (SS) e os Sólidos Sedimentáveis (SED) possuem comportamento semelhantes ao longo do processo, ocorrendo picos semelhantes nos mesmos instantes. Além de haver uma sequência de quedas no decurso do tratamento. Porém, a queda na entrada do decantador secundário é maior para os sedimentáveis, sendo assim o decantador primário atua bem na remoção dos sólidos sedimentáveis. Enquanto a maior redução dos sólidos suspensos ocorre na saída da estação, assim inferindo que grande parte de sua composição é matéria orgânica.



SS-E – Sólidos Suspensos na entrada
 SS-P – SS na entrada do decantador primário
 SS-D – SS na entrada do decantador secundário
 SS-S – SS na saída

Figura 16 Gráfico de linhas para os SS



SED-E – sedimentáveis na entrada
 SED-P – sedimentáveis na entrada do decantador primário
 SED-D – sedimentáveis na entrada do decantador secundário
 SED-S – sedimentáveis na saída

Figura 17 Gráfico de linhas para os SED

Observando as correlações dos SS, temos uma boa correlação na entrada da estação e após o gradeamento. Porém, com reduções nas etapas subsequentes. Para os SED a redução é maior após o decantador primário. Em ambas as variáveis temos a presença de muitos valores atípicos, onde estes são valores elevados que geram um desvio da distribuição das variáveis à esquerda.

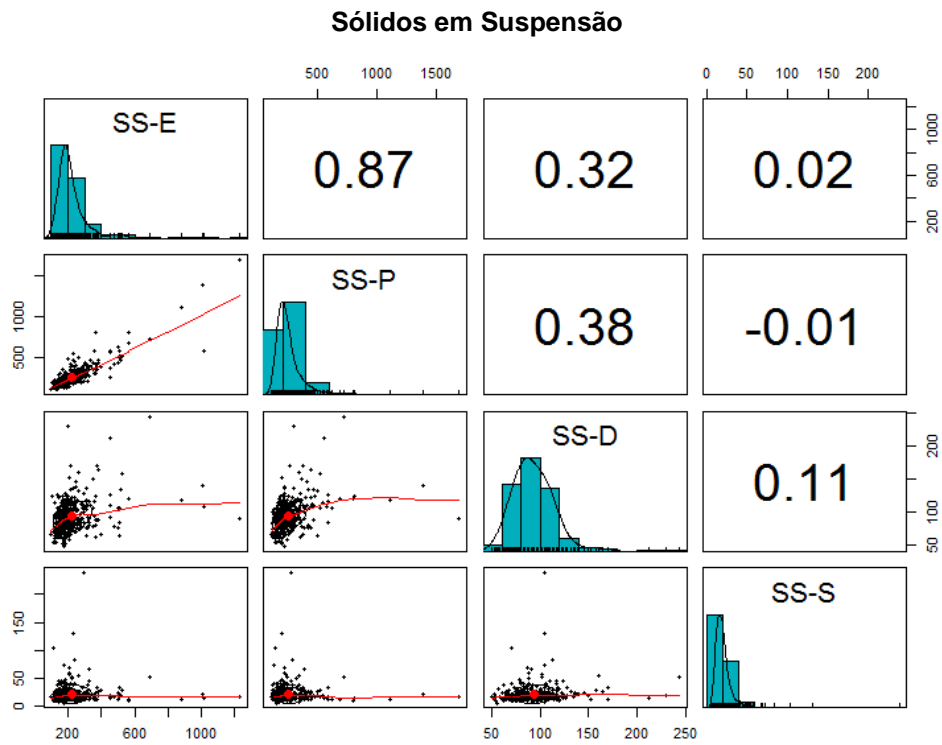


Figura 18 Gráfico de correlação para os SS

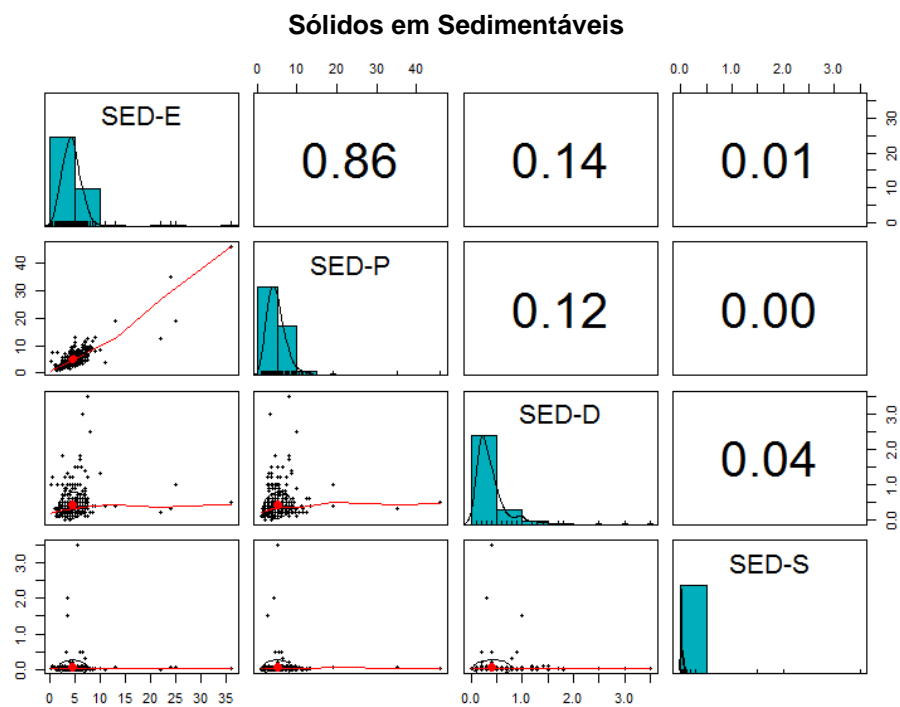


Figura 19 Gráfico de correlação para os SED

5.1.7. Sólidos Suspensos Voláteis

Para os Sólidos Suspensos Voláteis (SSV), temos um aumento de sua concentração ao longo da estação, com um gradual aumento desde a entrada até um máximo na saída. Isto deve-se ao fato de os microrganismos do lodo ativado degradarem a matéria orgânica e parte deste material tornar-se dissolvido na água, sendo a sua remoção dificultada pelo método de separação utilizado, a decantação.

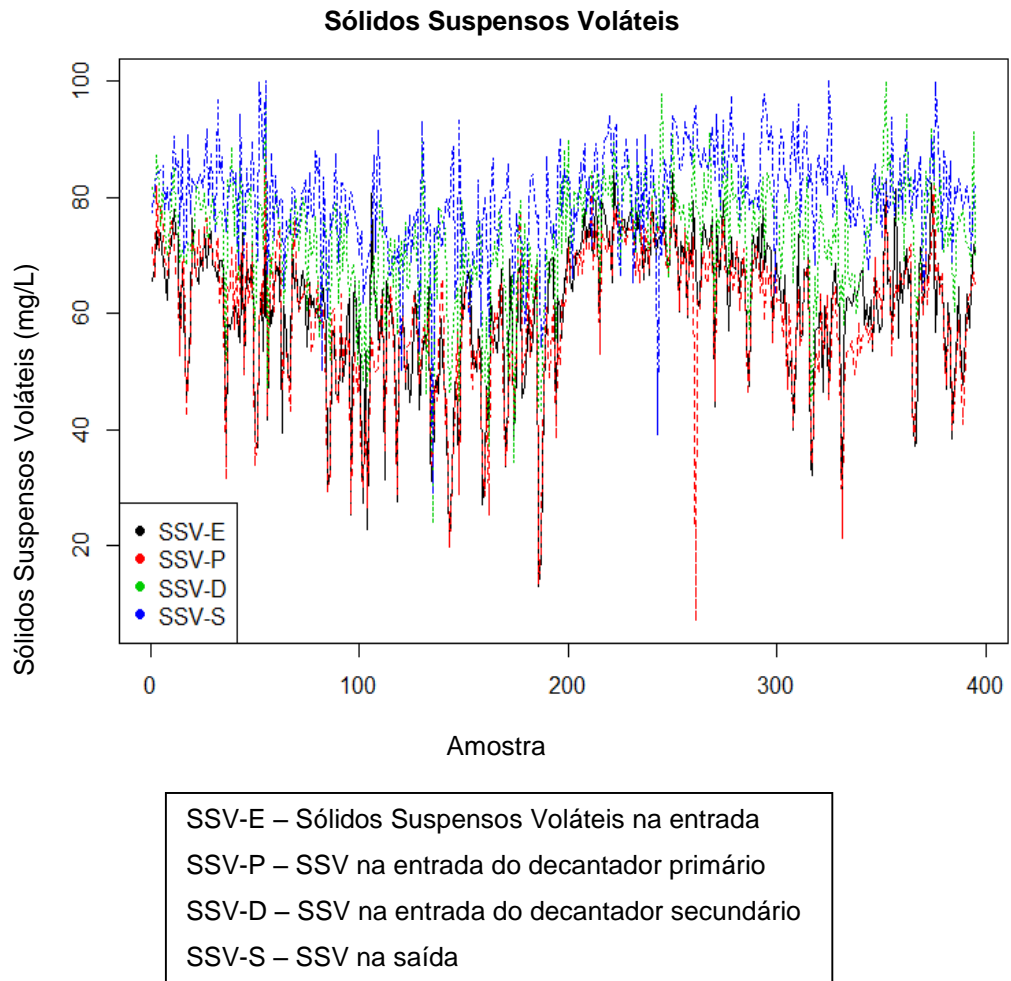


Figura 20 Gráfico de linhas para os SSV

Observe que da entrada do efluente da estação até o ponto de amostragem antes do decantador secundário, há boas correlações entre os pontos de coleta, entretanto na saída da estação os valores de SSV atingem picos mais elevados e com uma variabilidade maior, devido à ação dos microrganismos na decomposição da matéria orgânica.

Sólidos Suspensos Voláteis

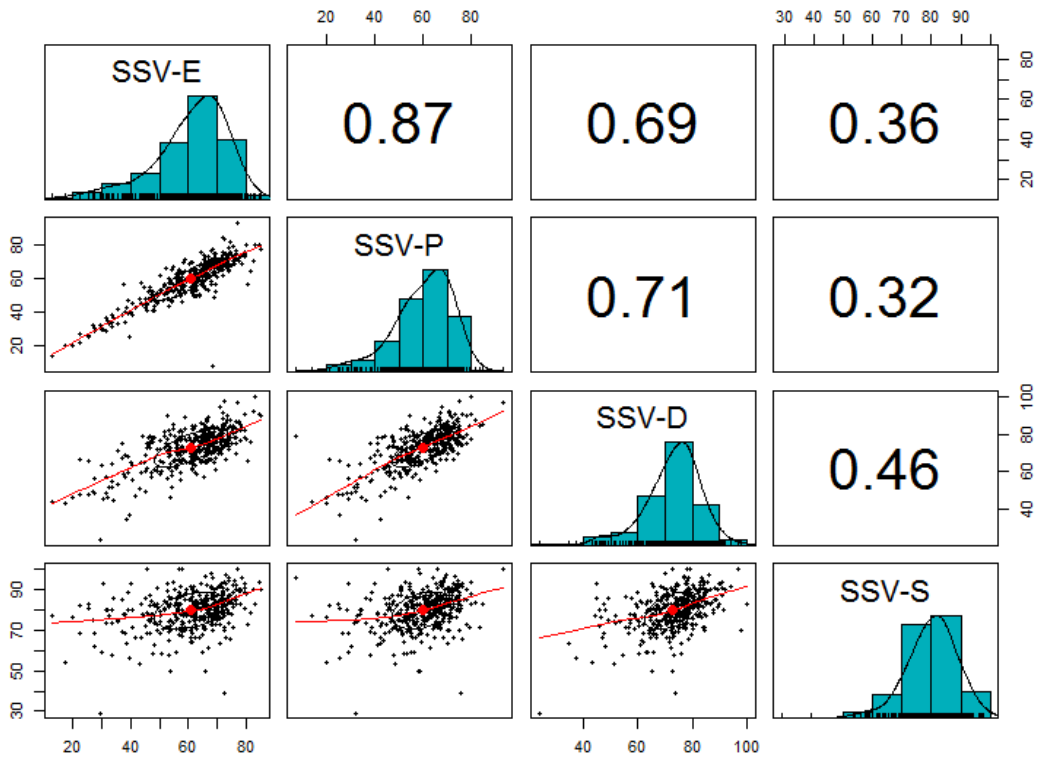


Figura 21 Gráfico de correlação para os SSV

Na Figura 22 Gráfico de correlação das variáveis, temos que quanto mais intensa é a cor azul, maior é a correlação positiva. E quanto mais intensa é a cor vermelha, maior é a correlação negativa. Seguindo esta lógica, quanto mais fechada é o círculo, maior é a correlação entre as variáveis e quando o oposto acontece, menor é a correlação.

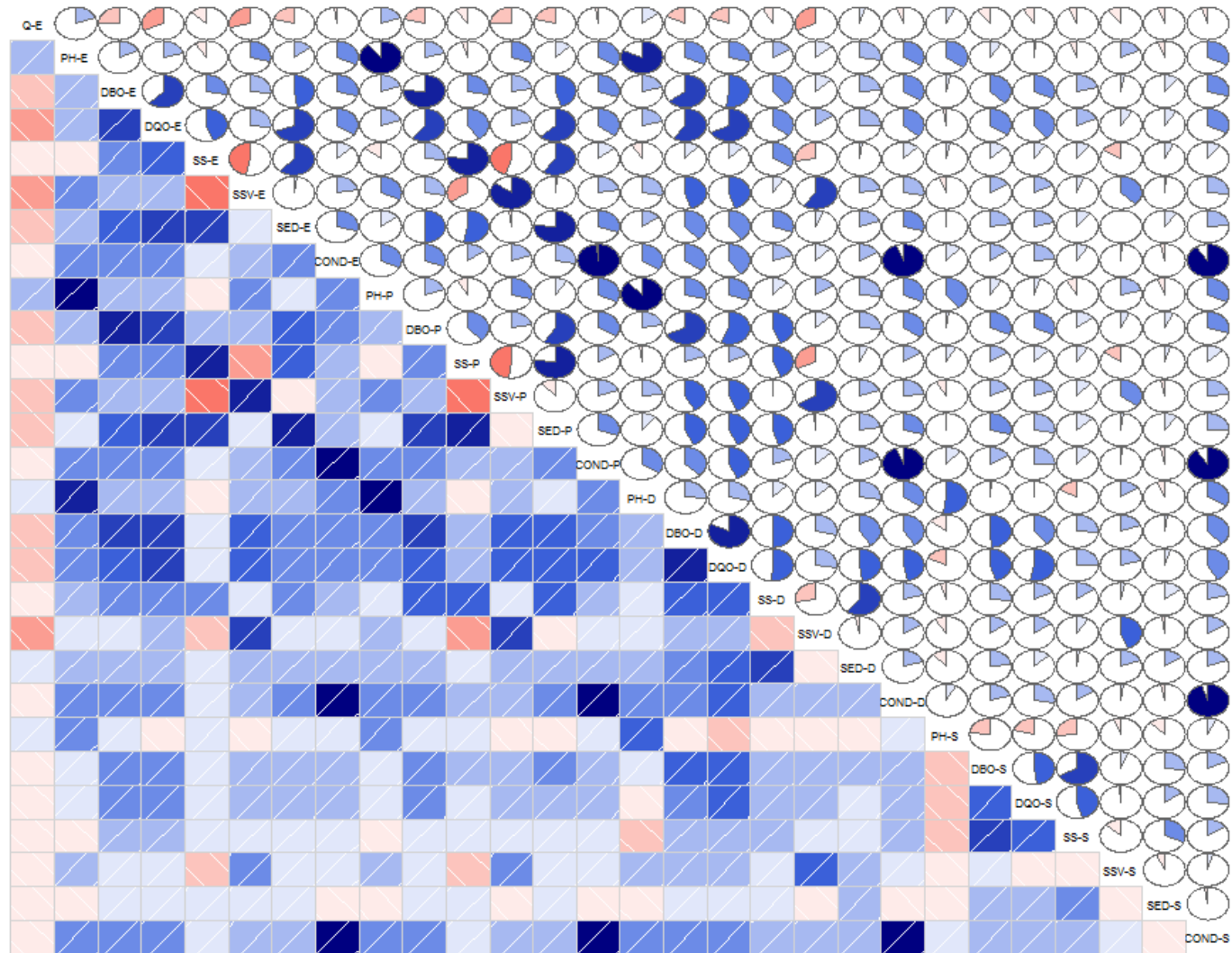


Figura 22 Gráfico de correlação das variáveis

5.1.8. Pré-processamento dos dados

É uma transformação dos dados aplicados às variáveis com o objetivo de melhorar o processamento dos dados. Neste caso, foi escolhido o autoescalamento, que é a subtração dos dados do valor médio da variável e a divisão pelo desvio-padrão da respectiva variável. Este tipo de pré-processamento, torna as variáveis adimensionais, e permite a comparação entre variáveis com grandezas diferentes, fazendo assim a ponderação das variáveis (Ferreira, 2015).

$$s_j^2 = \frac{1}{I-1} \sum_{i=1}^I (x_{ij} - \bar{x}_j)^2 \quad (5.2)$$

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (5.3)$$

Para comparação entre os dados sem e com processamento, temos a Figura 23 Gráficos de caixa das variáveis e Figura 24 Gráficos de caixa das variáveis autoescaladas. Assim, podemos observar que sem processamento a variável vazão sobrepõe as demais variáveis devido à sua grandeza numérica, e após o autoescalamento todas as variáveis estão na mesma escala, sendo permitido compará-las.

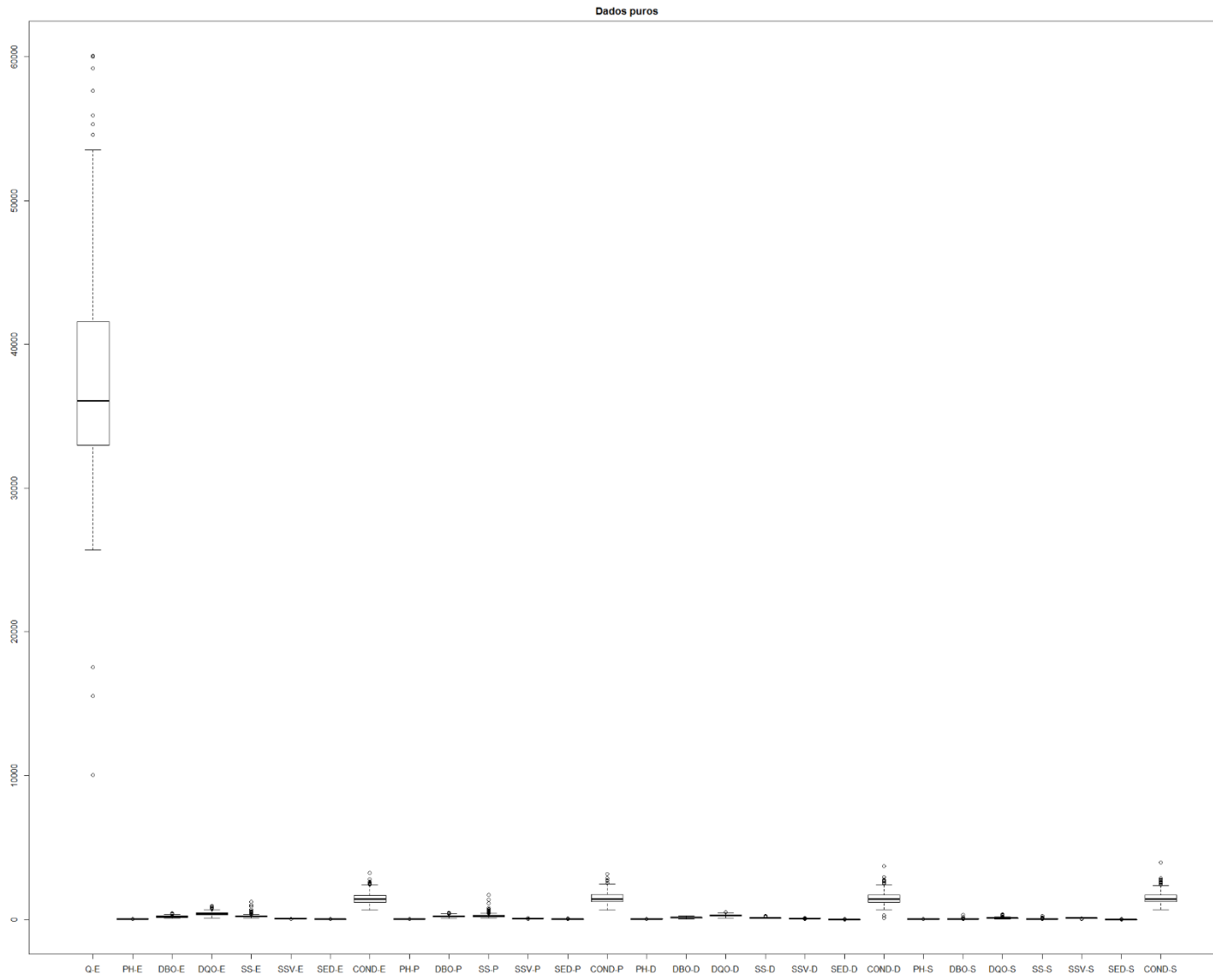


Figura 23 Gráficos de caixa das variáveis originais

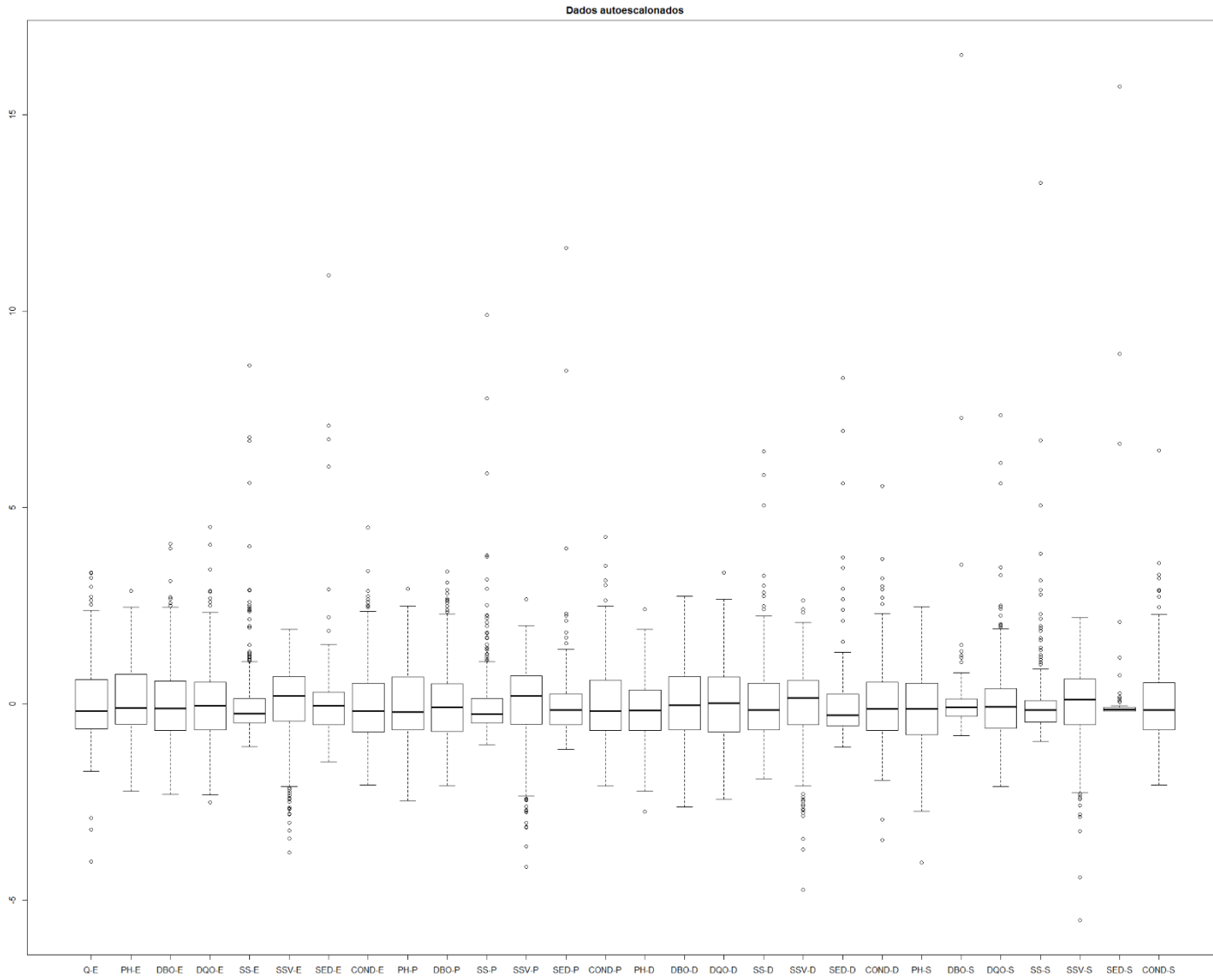


Figura 24 Gráficos de caixa das variáveis autoescaladas

6. DETECÇÃO DE OUTLIERS

A detecção de *outliers* no espaço com muitas dimensões pode ser feita utilizando a técnica do PCA, porém quando este método é aplicado a um banco de dados com a possibilidade de conter muitos valores atípicos, os componentes principais podem ser influenciados por estes dados, fazendo com que dados comuns sejam considerados *outliers* de maneira errada. Deste modo é necessária uma técnica mais robusta ao ruído e aos valores atípicos.

A análise robusta por componentes principais (Robust Principal Component Analysis – ROBPCA) foi desenvolvida por Hubert *et al.* (2005) com o objetivo de obter componentes que não são muito influenciáveis por *outliers*, diferentemente da abordagem clássica que é muito sensível a *outliers*, a abordagem robusta permite um diagnóstico rápido e visual dos *outliers* em um espaço multidimensional e com presença de multicolinearidade.

Os dados originais são armazenados em uma matriz $n \times p$, onde n é o número de amostras e p são as variáveis originais. O método do ROBPCA é dado em três etapas. Primeiro, os dados são pré-processados sendo transformados em um subespaço com dimensão máxima $n-1$. Em seguida, uma matriz de covariância preliminar (S_0) é construída e usada para selecionar o número de componentes (k) que serão retidos na sequência, produzindo um subespaço com k -dimensões. Então, os dados são projetados neste subespaço onde sua localização e matriz de dispersão são estimados de modo robusto, do qual seus k autovalores não nulos (l_1, \dots, l_k) são computados. Os autovetores correspondentes são os k componentes principais robustos (Hubert *et al.*, 2005).

A distinção entre as amostras regulares e os *outliers* é feito pelo gráfico de diagnóstico. Onde no eixo vertical está a distância do escore robusto (SD_i) e no eixo horizontal está a distância ortogonal (OD_i).

$$SD_i = \sqrt{\sum_{j=1}^k \frac{t_{ij}^2}{l_j}} \quad (6.1)$$

$$OD_i = \|x_i - \hat{\mu} - P_{p,k} t_i\| \quad (6.2)$$

A aplicação do algoritmo do ROBPCA para a detecção de outliers foi feita utilizando o pacote *rrcov*, o algoritmo do PCA clássico foi realizado em concomitante para permitir a comparação entre as metodologias.

O gráfico de diagnóstico obtido pelo método do PCA clássico resultou em um valor de limite crítico das distâncias das pontuações igual a zero, e um limite crítico da distância

ortogonal igual a 5,82 (linha vermelha). Logo o PCA clássico foi muito sensível aos valores atípicos, classificando erroneamente outras amostras como *outliers*.

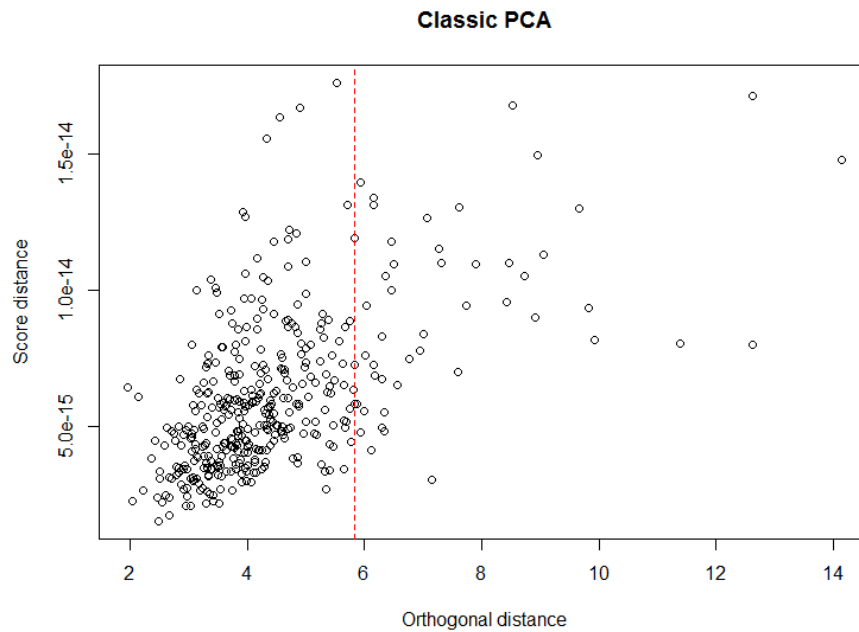


Figura 25 Gráfico de diagnóstico (PCA clássico)

Para o PCA robusto, o limite crítico das distâncias dos escores foi igual a 3,75 e o limite crítico da distância ortogonal igual a 2,42 (linha vermelha). Neste método, temos um número menor de valores atípicos em comparação com o método anterior, evidenciando a importância dos estimadores robustos para a detecção de *outliers*, evitando assim o descarte inadequado de amostras.

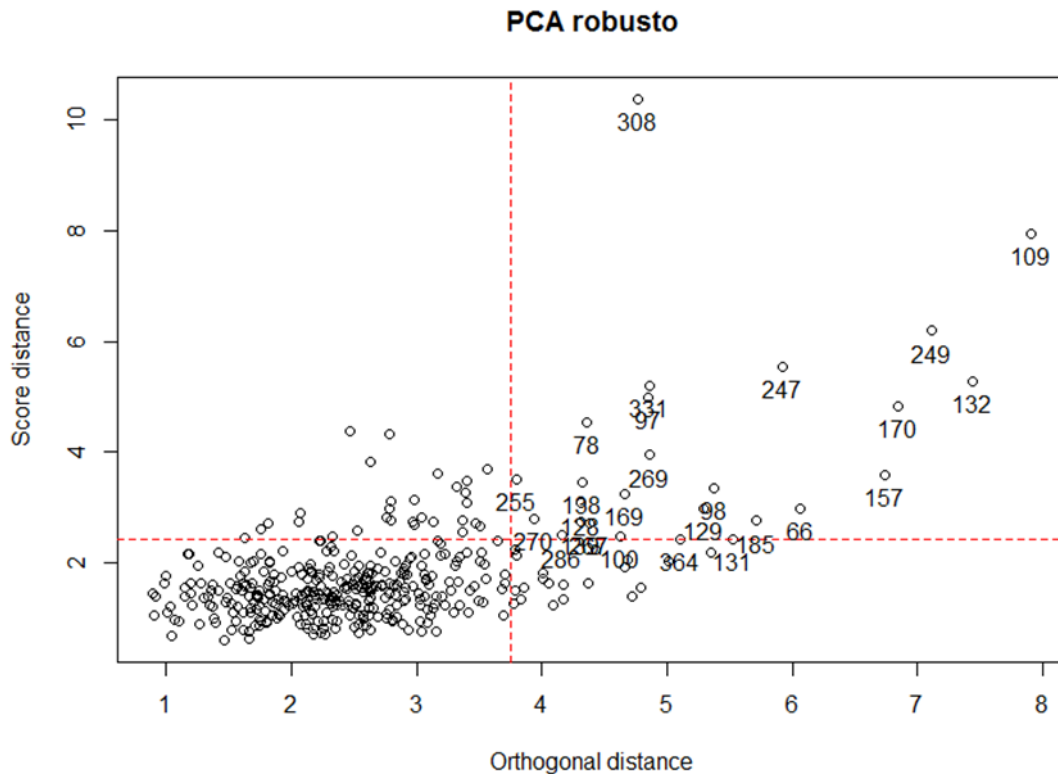


Figura 26 Gráfico de diagnóstico (PCA robusto)

Os valores atípicos detectados e descartados do banco de dados foram: "66" "78" "97" "98" "100" "109" "128" "129" "131" "132" "138" "139" "157" "169" "170" "185" "247" "249" "255" "267" "269" "270" "286" "308" "331" "364". Estes valores foram sobrepostos sobre os gráficos de caixa das variáveis autoescaladas, permitindo assim avaliar quais variáveis mais contribuíram para as amostras atípicas. Com a remoção dos outliers o número de amostras reduziu de 395 para 369.

Pelo gráfico abaixo, pode-se observar que os grupos de variáveis dos sólidos suspensos e dos sólidos suspensos voláteis foram os que mais colaboraram para os *outliers*, pois são onde os pontos vermelhos (*outliers*) mais se concentram sobre os valores discrepantes. Vale ressaltar que enquanto um ponto é um valor atípico para uma variável, este mesmo não influencia outras variáveis, por isso é necessário um método de detecção de *outliers* no espaço multidimensional.

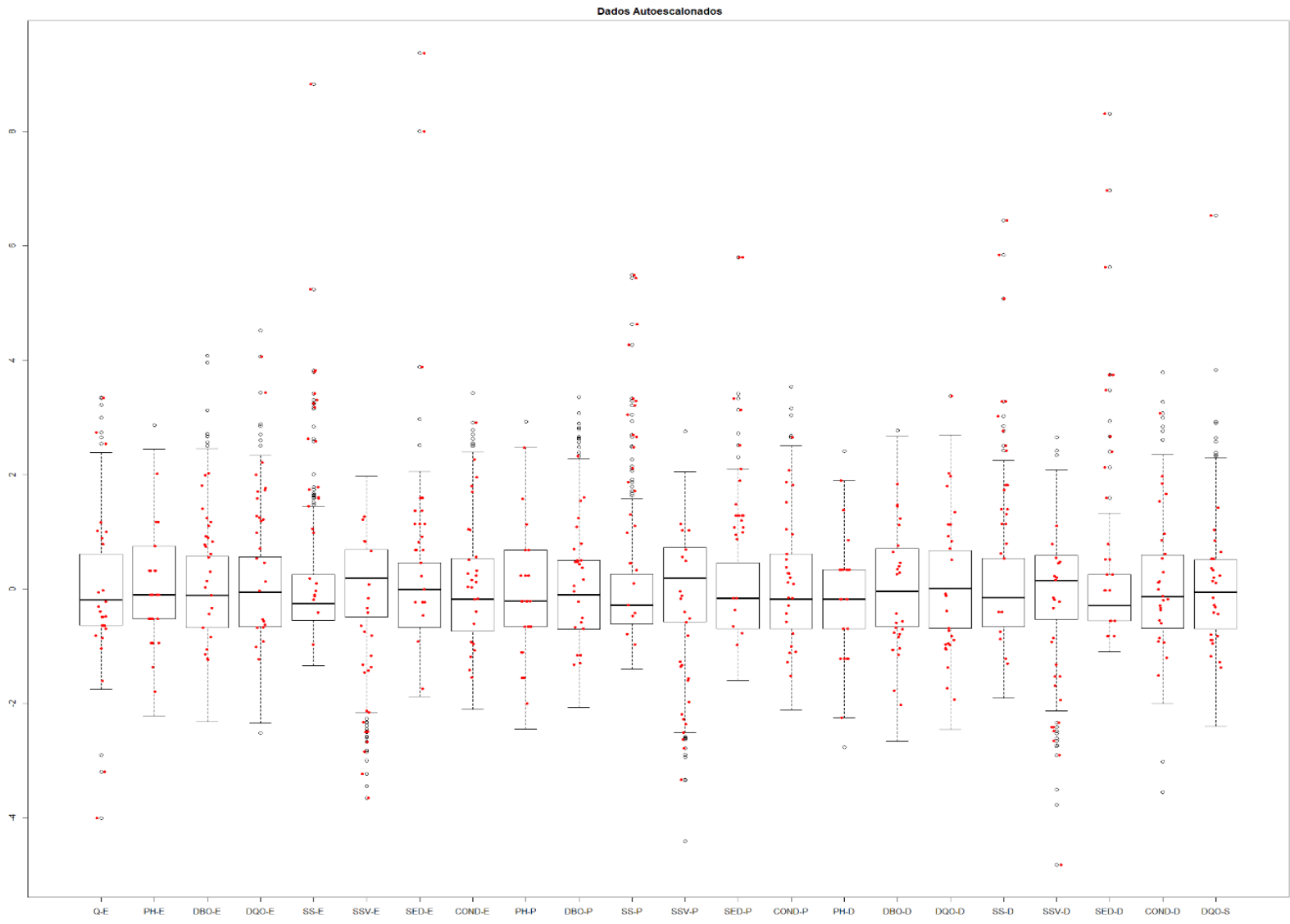


Figura 27 Outliers sobrepostos nas variáveis autoescaladas

7. PARTIÇÃO DAS AMOSTRAS

A partição das amostras é uma etapa importante na modelagem dos dados, logo que é a fase onde ocorre a divisão das amostras em dois conjuntos, o de treinamento, utilizado na construção do modelo, e o de teste, que avalia o desempenho do modelo. As amostras foram divididas em conjunto de calibração (70%), com 258 amostras, e conjunto de teste (30%) para validação, com 111 amostras.

O algoritmo de Kennard e Stone (1969) foi usado para a partição de amostras nos conjuntos de treinamento e de teste. Este algoritmo seleciona duas amostras com a maior distância Euclidiana entre si, em sequência, por meio de um planejamento experimental e com a restrição de que estejam distribuídas uniformemente no espaço, até que todas as amostras do conjunto de treinamento sejam selecionadas (Ferreira, 2015).

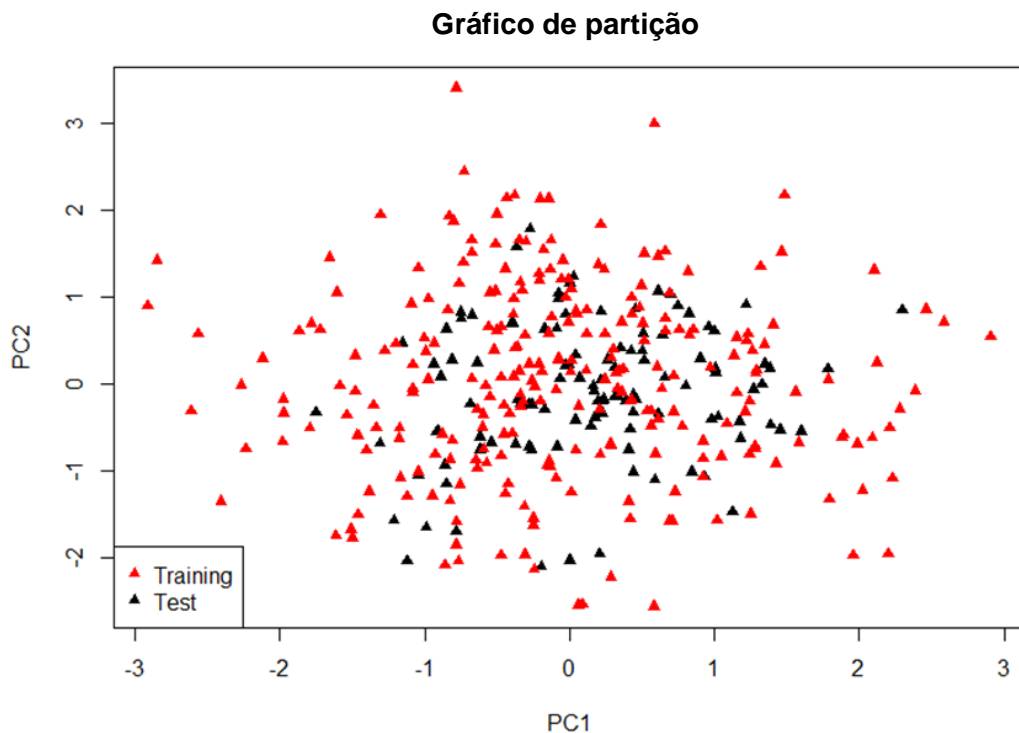


Figura 28 Gráfico dos conjuntos de treinamento (vermelho) e de teste (preto) em componentes principais

8. MODELOS DE REGRESSÃO

A Demanda Química de Oxigênio é um parâmetro que determina a carga orgânica total do efluente, a biodegradável e a não-biodegradável, além de ser mais fácil e rápida de realizar a sua análise em laboratório. Esta também pode ser facilmente relacionada com a DBO, a partir de dados históricos do efluente, para predições futuras. Logo a DQO é um parâmetro mais atrativo para este estudo de predição.

8.1. Regressão Linear Múltipla

A regressão linear múltipla foi feita utilizando a função “lm”, e considerou inicialmente com todas as variáveis. O sumário do modelo inicial está disposto no Anexo B.

No modelo linear considerando todas as variáveis, observa-se que das 21 variáveis, apenas algumas possuem o valor-p abaixo de 0,05, considerando um nível de confiança de 95%, logo, foi utilizada a função “step” que seleciona as variáveis do modelo linear baseado no critério de AIC (Akaike). Como resultado, temos o modelo ajustado que está disposto no Anexo C. Este possui 9 das 21 variáveis originais, sendo assim um modelo mais simples que o modelo inicial.

Abaixo está o gráfico dos valores preditos pelo modelo contra os valores reais para o conjunto de treinamento, a linha vermelha não é a linha de tendência do modelo, mas uma linha de 45°, que passa pela a origem (0,0) e reflete a situação de melhor ajuste dos dados. Nota-se que há uma grande dispersão dos valores preditos pelo modelo linear, com valores muito mal ajustados.

Gráfico de predição

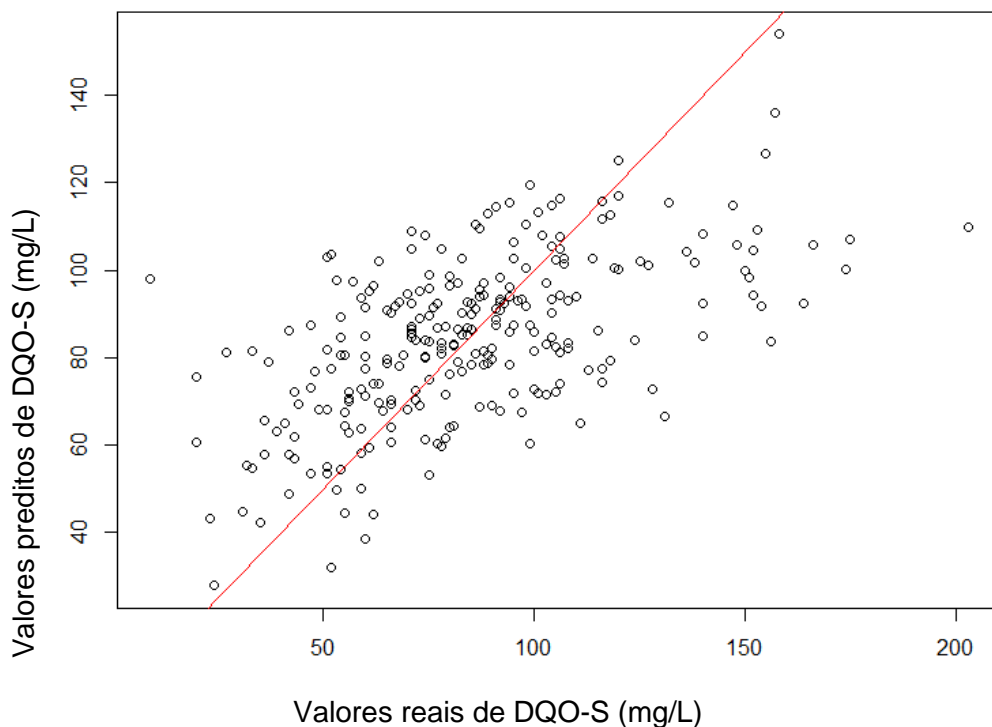


Figura 29 Valores preditos x Valores reais (Treinamento)

Na figura abaixo, temos quatro gráficos para análise dos resíduos, logo percebe-se que há uma boa dispersão dos resíduos em torno da reta horizontal (*Residuals vs Fitted*). Estes também possuem um bom ajuste na reta normal (Normal Q-Q). Não há *outliers* para o modelo (*Residuals vs Leverage*), logo que não possuem resíduos com valor de alavancagem (Leverage) maior que 0,5.

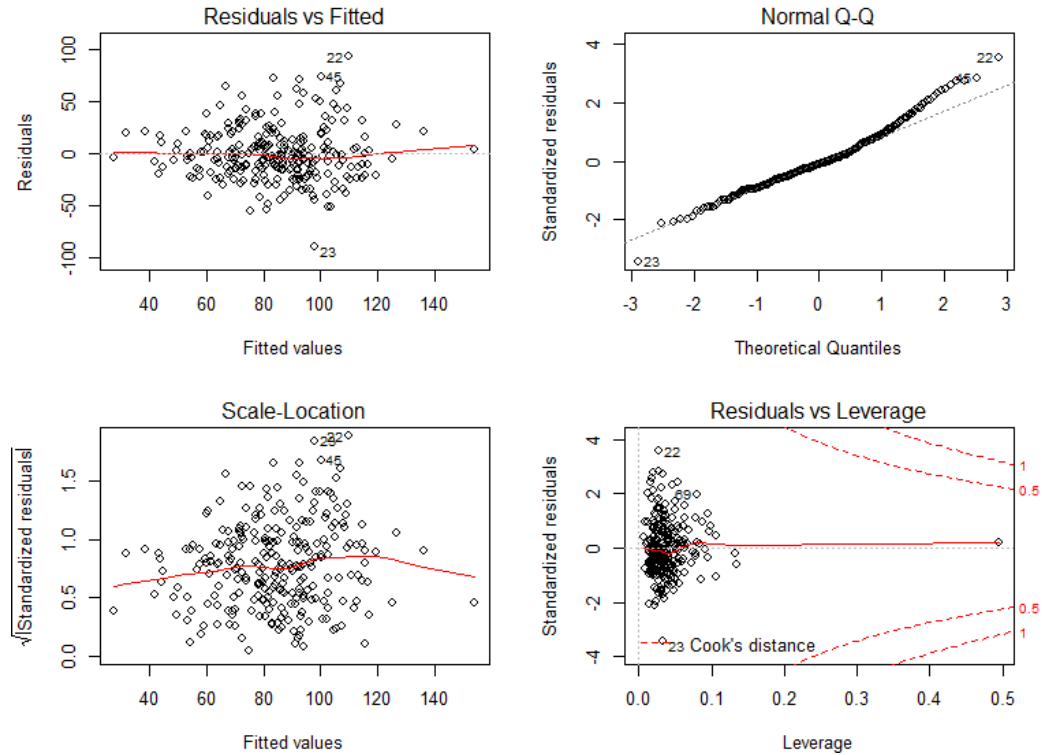


Figura 30 Análise dos resíduos do modelo linear

Para o conjunto de testes, os dados ficaram muito dispersos, com poucos pontos sobre a linha de referência ($x=y$), refletindo assim uma baixa capacidade de precisão do modelo.

Gráfico de predição

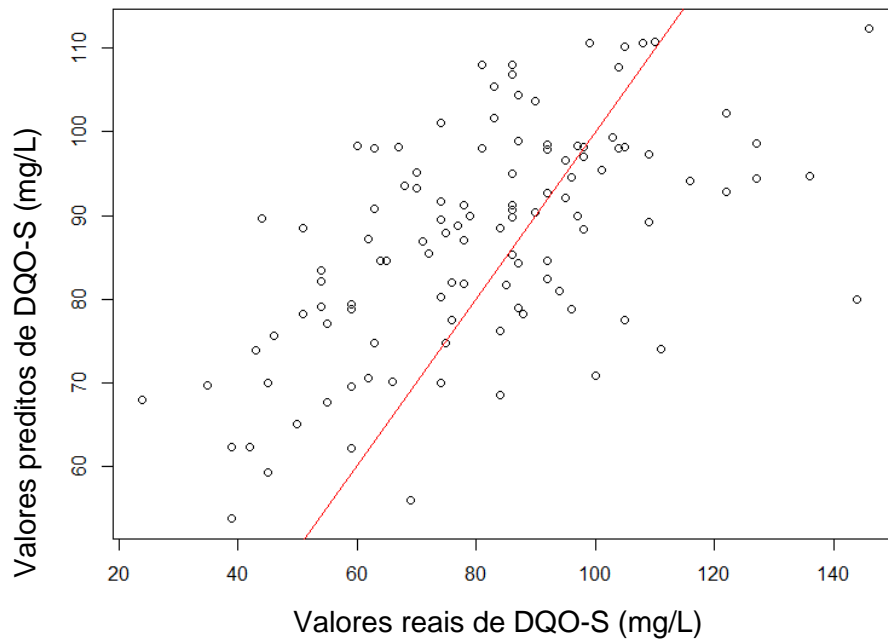


Figura 31 Valores preditos x Valores reais (Teste)

Na tabela abaixo, os parâmetros de mérito dos conjuntos de treinamento e de testes estão dispostos para a comparação do ajuste do modelos entre os conjuntos de dados. O conjunto de treinamento possui um R^2 um pouco maior, mas isto não é refletido nos parâmetros que mensuram o erro, logo que estes foram maiores para o conjunto de treinamento.

Tabela 1 Parâmetros de mérito (RLM)

Parâmetros de mérito	Conjunto de Treinamento	Conjunto de Teste
R^2	0,34	0,27
MAE	19,63	16,08
MSE	675,43	406,49
RMSE	25,99	20,16

8.2. Regressão por Componentes Principais

A regressão por componentes principais foi criada utilizando a função `pcr` do pacote `pls` (Mevik, Wehrens e Liland, 2016), a qual o sumário do modelo está disposto abaixo. Onde RMSEP (do inglês, *Root Mean Squared Error of Prediction*), ou em português, raiz quadrada do erro médio quadrático da predição, que foi calculado utilizando a validação cruzada (*leave-one-out*), varia de 29,68 a 27,85, o que é uma redução pequena considerando os 21 componentes principais desenvolvidos pelo modelo. A variância explicada para as 21 variáveis regressoras originais é iniciada em 34,76% no primeiro componente e alcança o acumulado de 100% nos 21 componentes principais do modelo. Para a variável resposta, o primeiro componente explica 15,28% de sua variância e ao final com os 21 componentes, explica somente 35,7% de sua variabilidade, limitando assim a previsão do modelo. O sumário do modelo de regressão por componentes principais está disposto no Anexo D.

Para a escolha do número “ótimo” de componentes foi utilizada a função `selectNcomp`, que dispõe de duas estratégias para tal. O primeiro é baseado na heurística `onesigma` (Mevik, Wehrens e Liland, 2016), que consiste em escolher o modelo com menos componentes com o menor erro padrão no modelo em geral. A segunda estratégia emprega a abordagem de permutação, e basicamente testa se a adição de um novo componente beneficia o modelo.

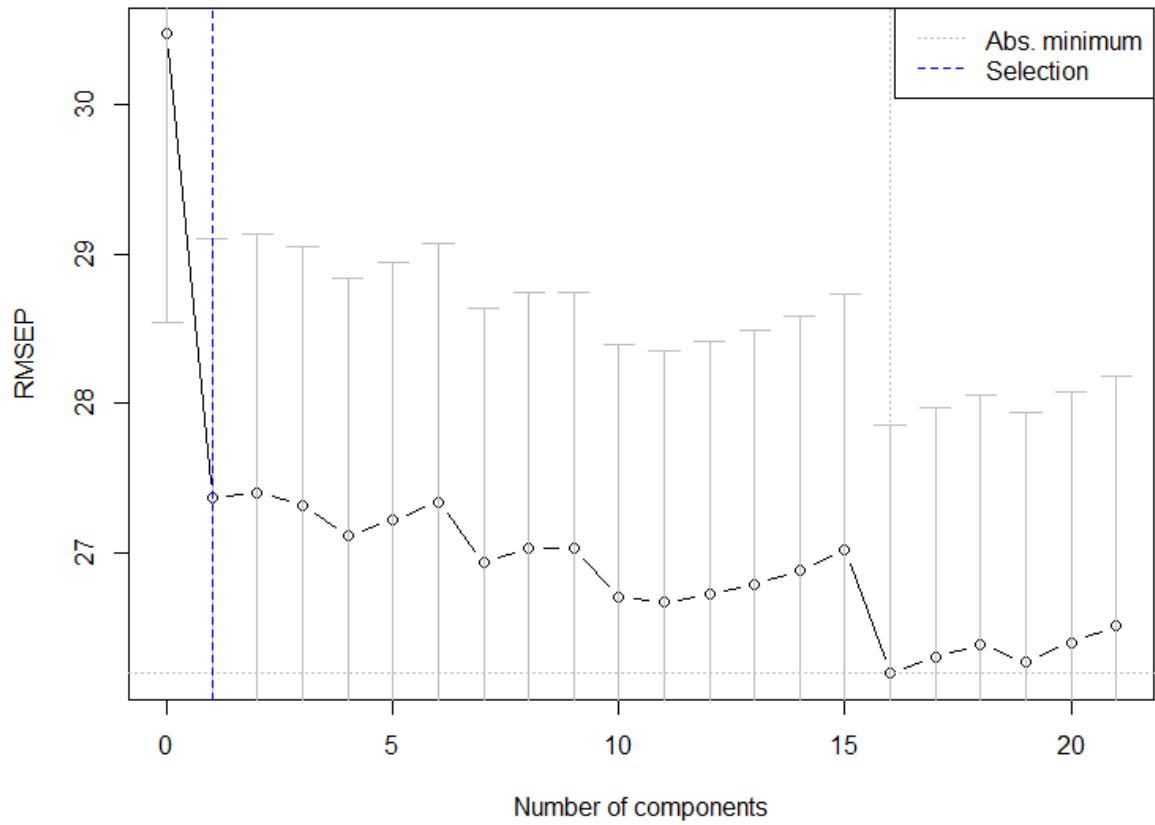


Figura 32 Seleção do N° de componentes (onesigma)

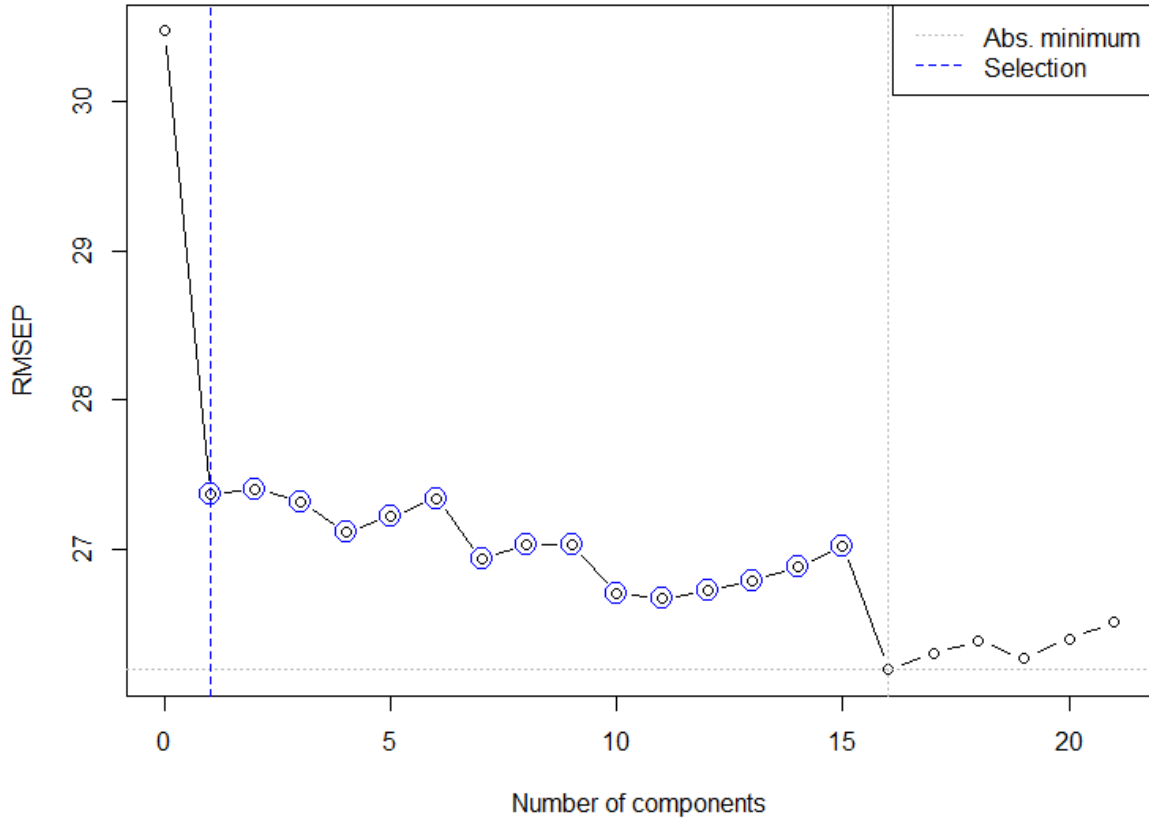


Figura 33 Seleção do N° de componentes (randomization)

Temos que para as duas estratégias, o número ótimo de componentes foi retido em 1 componente principal, apesar do RMSEP mínimo ser em 16 componentes. Deste modo, os demais componentes adicionam mais ruído ao modelo do que contribuem para uma melhoria significativa na previsão.

Logo abaixo, está o gráfico dos valores preditos contra os valores reais, onde os pontos possuem uma grande dispersão, sendo há diversos pontos os quais os valores preditos ultrapassam a faixa de variação dos valores reais.

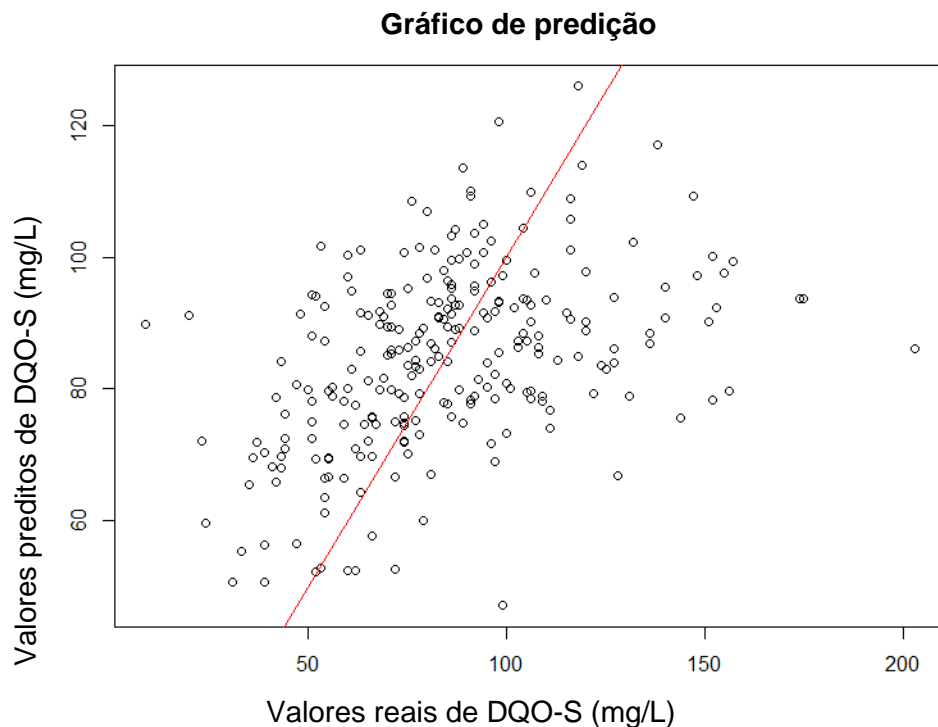


Figura 34 Valores preditos x Valores reais (Treinamento)

Para o conjunto de testes a variação é ainda maior, deste modo o modelo PCR, possui uma previsão muito limitada, logo que submetido a um conjunto de teste, este possui uma previsão com erros maiores do que o conjunto de treinamento, quando na verdade o conjunto de treinamento possui uma variabilidade maior que o de teste (vide o gráfico de seleção de amostras).

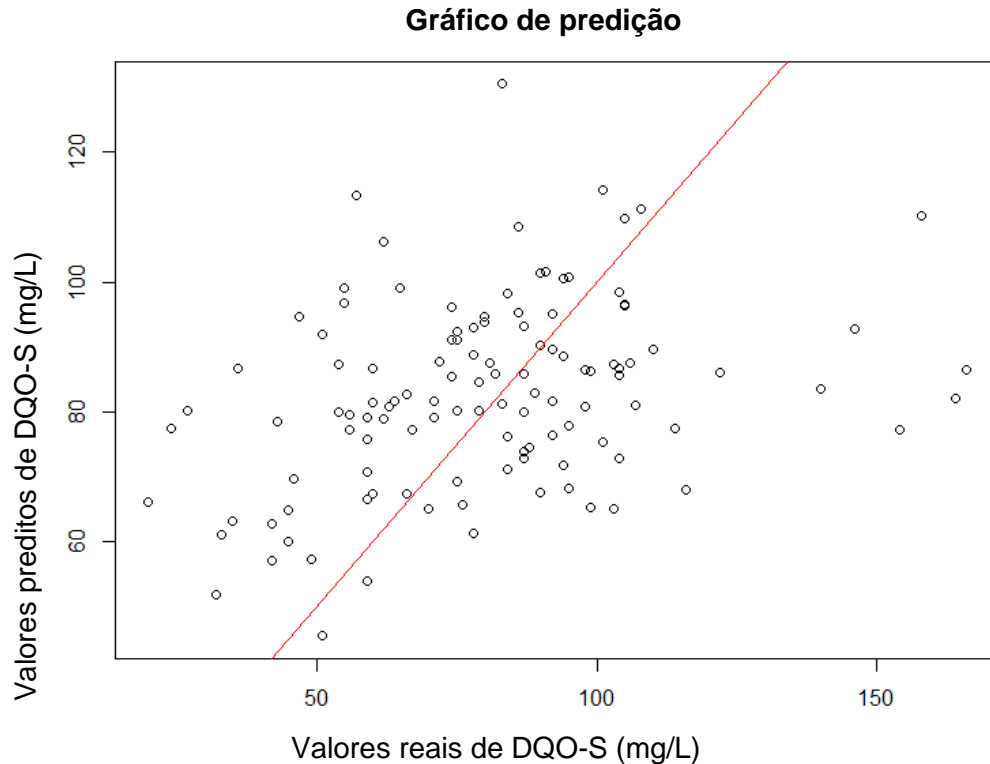


Figura 35 Valores preditos x Valores reais (Teste)

Apesar do modelo PCR reduzir as variáveis originais em 1 componente principal, não foi obtido um bom resultado de previsão, pois o modelo conseguiu explicar somente 35,7% da variabilidade contida na variável resposta, isto é refletido nos parâmetros de mérito, logo que possuem valores elevados. Como o conjunto de teste obteve um coeficiente de correlação pior do que o de treinamento, este modelo não é robusto quanto à aplicação de amostras externas.

Tabela 2 Parâmetros de mérito (PCR)

Parâmetros de mérito	Conjunto de Treinamento	Conjunto de Teste
R^2	0,20	0,07
MAE	20,62	20,82
MSE	738,06	721,10
RMSE	27,16	26,85

8.3. Regressão por Mínimos Quadrados Parciais

A regressão por mínimos quadrados parciais foi criada utilizando a função pls do pacote pls (Mevik, Wehrens e Liland, 2016), onde temos o sumário do modelo abaixo. O RMSEP, do mesmo modo que no PCR, foi calculado utilizando a validação cruzada leave-

one-out, variou entre 29,26 a 27,85. A variância explicada para as 21 variáveis originais é iniciada em 34,10% na primeira variável latente e alcança o acumulado de 100% nas 21 variáveis latentes. Para a variável resposta, a primeira variável latente explica 18,97% de sua variância e alcançando 35,7% ao total das 21 variáveis latentes. O sumário do modelo de regressão por mínimos quadrados parciais está disposto no Anexo E.

O número “ótimo” de variáveis latentes foi escolhido utilizando a mesma função do caso PCR. E para as duas estratégias disponíveis na função selectNcomp do pacote pls, foi retida apenas uma variável latente.

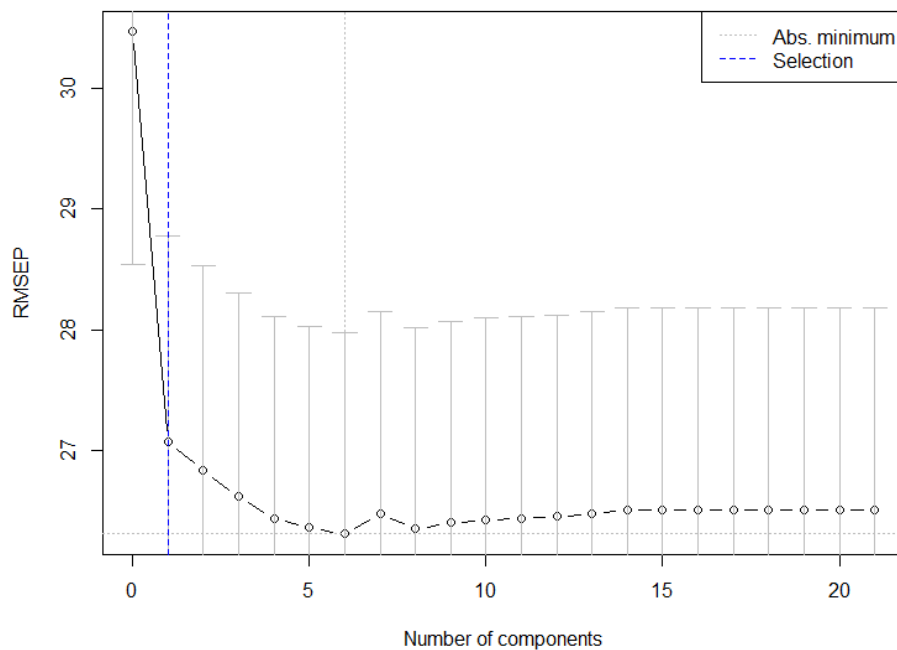


Figura 36 Seleção do N° de componentes (onesigma)

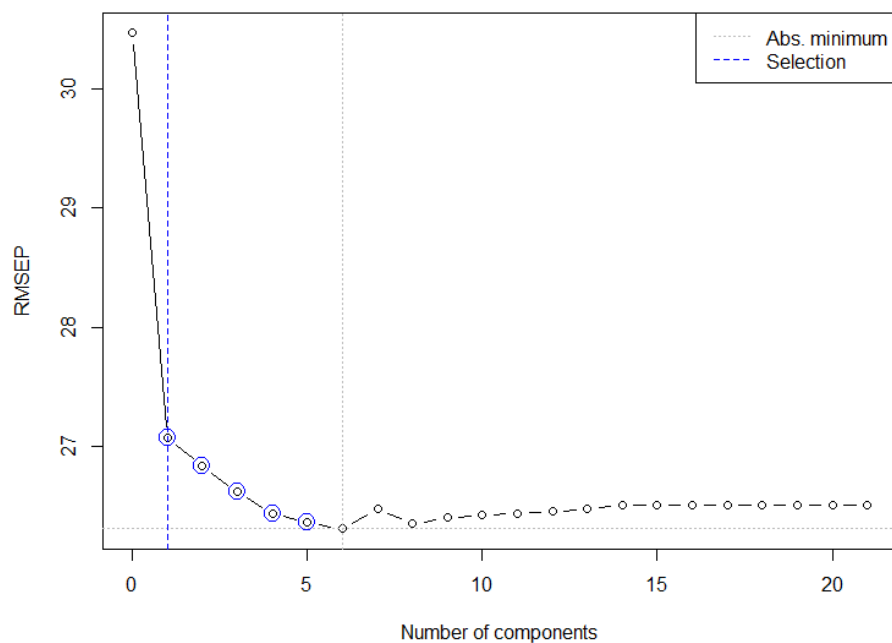


Figura 37 Seleção do N° de componentes (randomization)

No gráfico dos valores preditos contra os valores reais, apesar dos pontos possuírem uma grande dispersão, houve uma melhora em relação ao PCR. Porém, ainda existem valores de previsão que ultrapassam a faixa de variação dos valores reais.

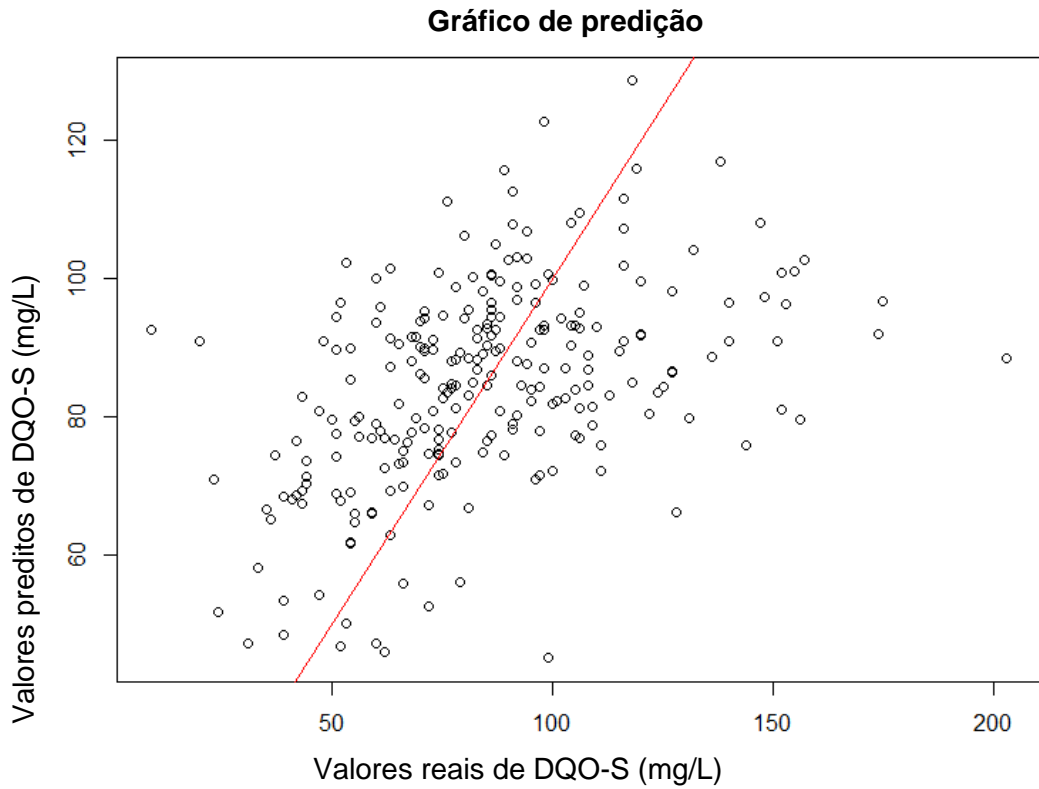


Figura 38 Valores preditos x Valores reais (Treinamento)

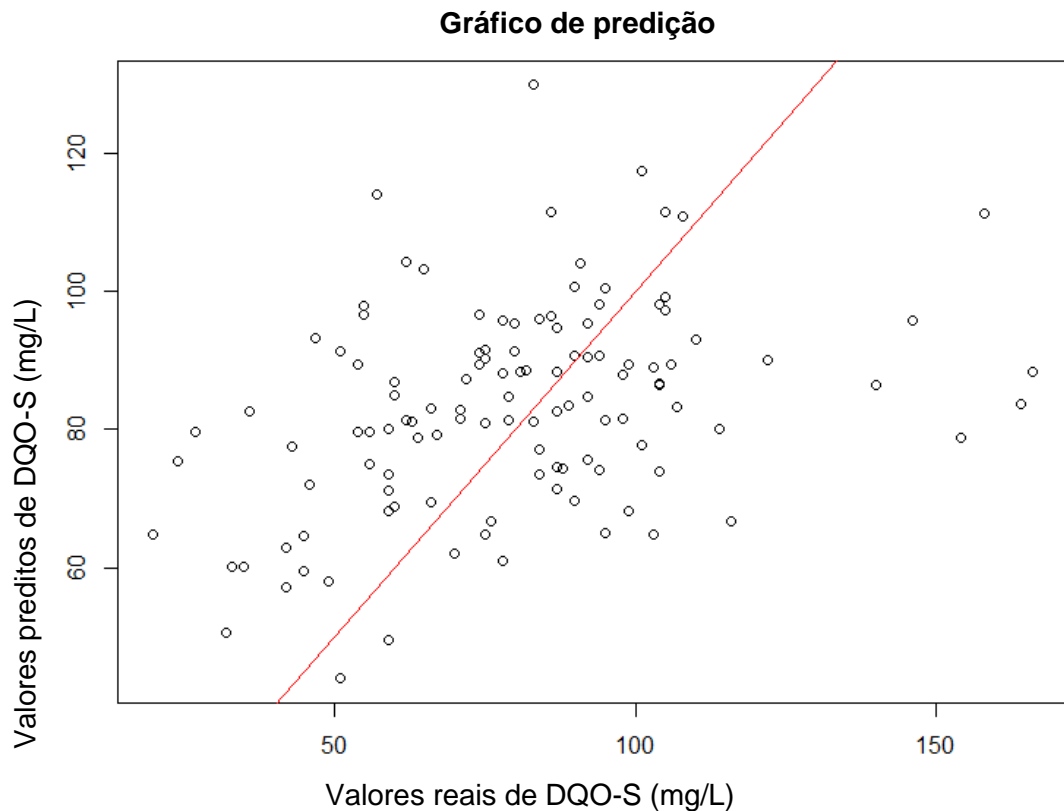


Figura 39 Valores preditos x Valores reais (Testes)

No conjunto de testes a variação entre os valores previstos e os reais é ainda maior que no conjunto de treinamento, refletindo em um coeficiente de correlação menor para o conjunto de teste. Além dos parâmetros de mérito de erro serem elevados para os conjuntos de treinamento e de teste.

Tabela 3 Parâmetros de mérito (PLSR)

Parâmetros de mérito	Conjunto de Treinamento	Conjunto de Teste
R ²	0,23	0,10
MAE	20,22	20,70
MSE	710,03	699,15
RMSE	26,65	26,44

Deste modo os modelos PCR e PLSR, neste estudo, foram bastante semelhantes, logo que obtiveram resultados análogos para os conjunto de treinamento e de teste, devido ao fato da variância explicada da variável resposta pelo modelo ser bastante inferior ao esperado, sendo de 35,7% para ambos os modelos. A regressão linear múltipla obteve um resultado um pouco melhor que o PCR e o PLSR, logo que esta não faz o uso da redução de dimensionalidade para a construção do modelo.

8.4. Máquina de Vetor de Suporte

A regressão por máquina de vetor de suporte foi desenvolvida utilizando a função `tune`, disponível no pacote `e1071` (Meyer *et al*, 2018). O *kernel* de base radial foi escolhido como uma alternativa a um método não-linear, pois os modelos anteriores que são lineares obtiveram um resultado abaixo do esperado. O parâmetro ϵ foi variado entre 0 e 1 ao passo de 0,1, e o custo de 4 a 512 em potências de 2. A validação interna utilizada foi a validação segmentada, com $k=10$. Abaixo estão os parâmetros que obtiveram um modelo com o melhor desempenho.

Parameter tuning of 'svm':

```
- sampling method: 10-fold cross validation
- best parameters:
  epsilon cost
    0      32
```

Abaixo está o gráfico de ajuste dos parâmetros de custo e ϵ , onde a região mais escura é que obteve melhor desempenho do modelo.

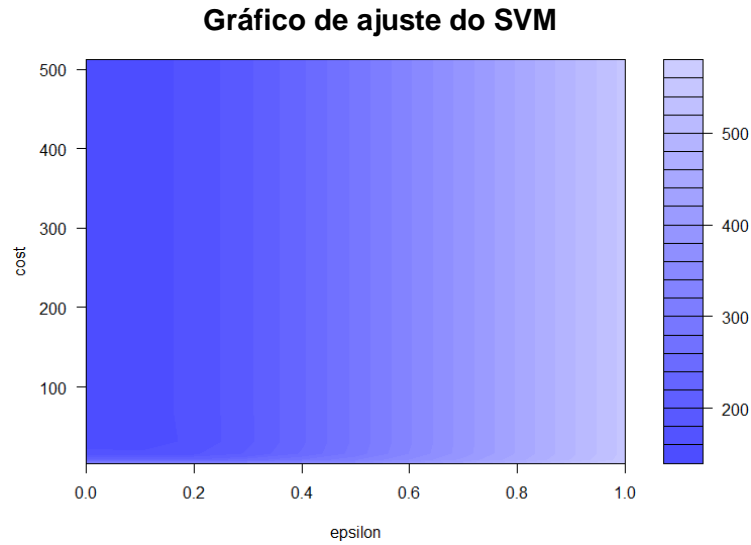


Figura 40 Gráfico de ajuste do modelo SVM

O modelo de vetores de suporte utilizou nas 21 variáveis de entrada 110 vetores de suporte, ou seja, o hiperplano para o ajuste do modelo utilizou 110 planos para definir a sua localização.

No gráfico dos valores preditos contra os valores reais, pode-se observar um excelente ajuste, onde quase que todos os pontos estão dispostos sobre a linha de referência, há somente três pontos com erros maiores, porém isto é irrelevante frente a grande capacidade de previsão do modelo.

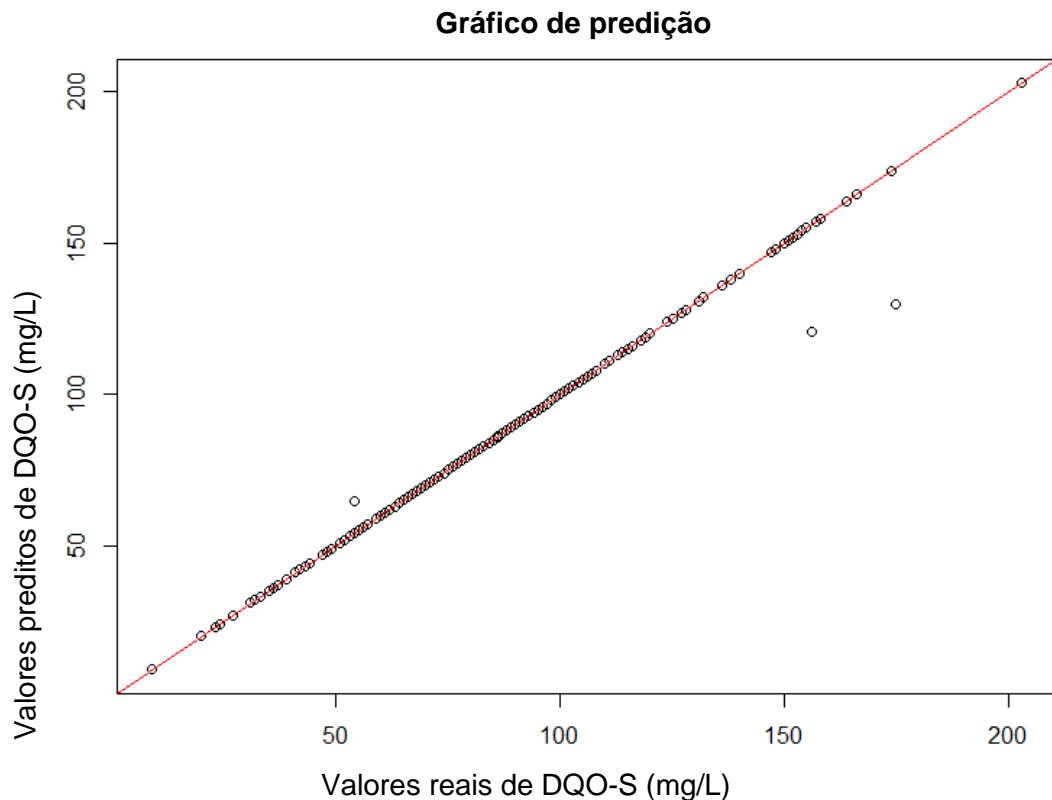


Figura 41 Valores preditos x Valores reais (Treinamento)

Na análise dos resíduos, estes três pontos fora do ajuste ficam mais evidentes, enquanto a grande maioria dos pontos estão com o erro praticamente nulo, estes estão foram da linha de tendência. Também estão fora da linha de distribuição normal. O gráfico dos resíduos vs Leverage mostra que o ponto 24 poderia ser considerado um outlier, porém os outliers foram removidos com o PCA robusto, e para a comparação com os demais modelos, não foram removidos outliers após esta etapa.

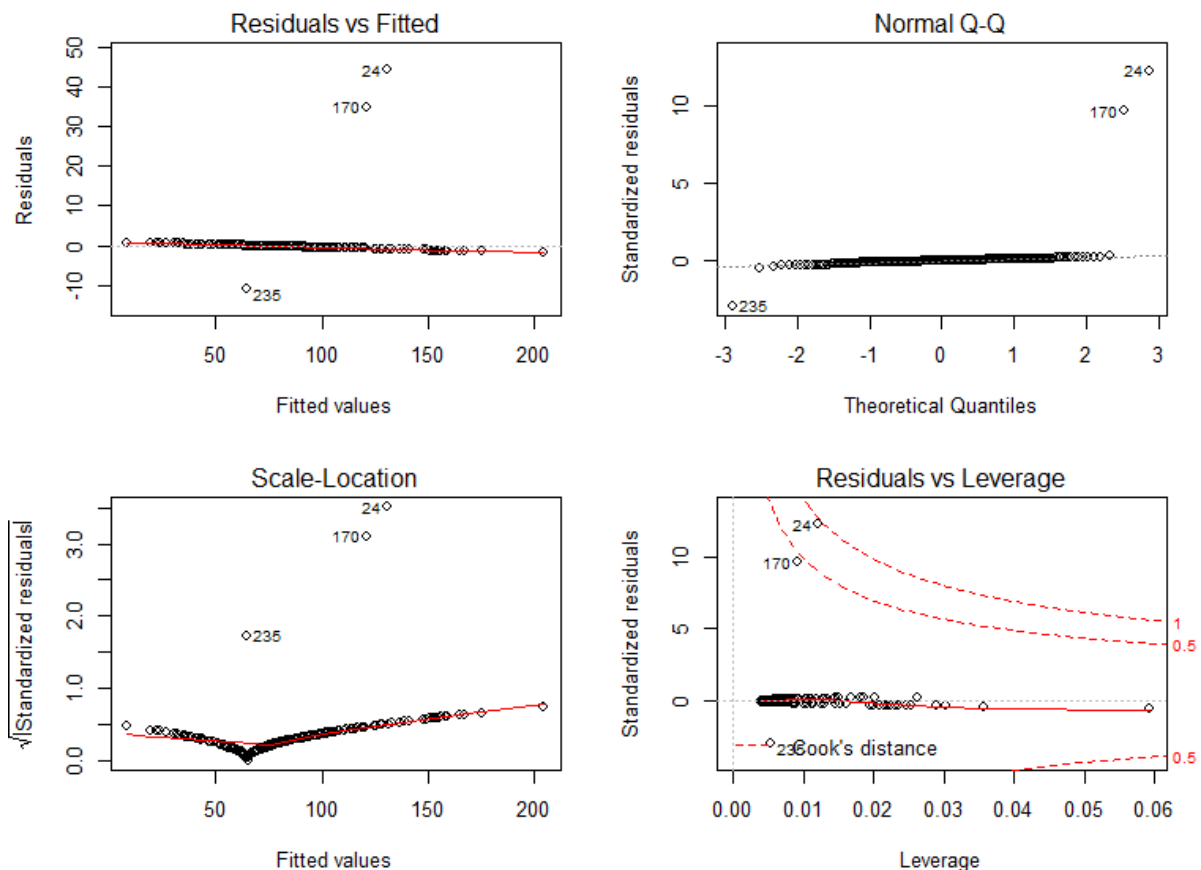


Figura 42 Análise dos resíduos (Treinamento)

Para o conjunto de testes o ajuste está exatamente sobre a reta de referência, indicando um perfeito ajuste entre os valores preditos e reais. Deste modo, o modelo SVM além de ser o mais preciso, é também o mais robusto em comparação com os modelos anteriores. Este fato é devido aos modelos anteriores serem em sua essência modelos lineares, enquanto o SVM é um modelo não-linear e que desenvolve mais dimensões na procura de um melhor ajuste, enquanto o PCR e o PLSR usam da redução da dimensionalidade para a compressão dos dados, o que neste caso foi um impicílio pois estes modelos conseguiram explicar somente 35,7% da variância contida da variável resposta, afetando assim a previsão do PCR e do PLSR.

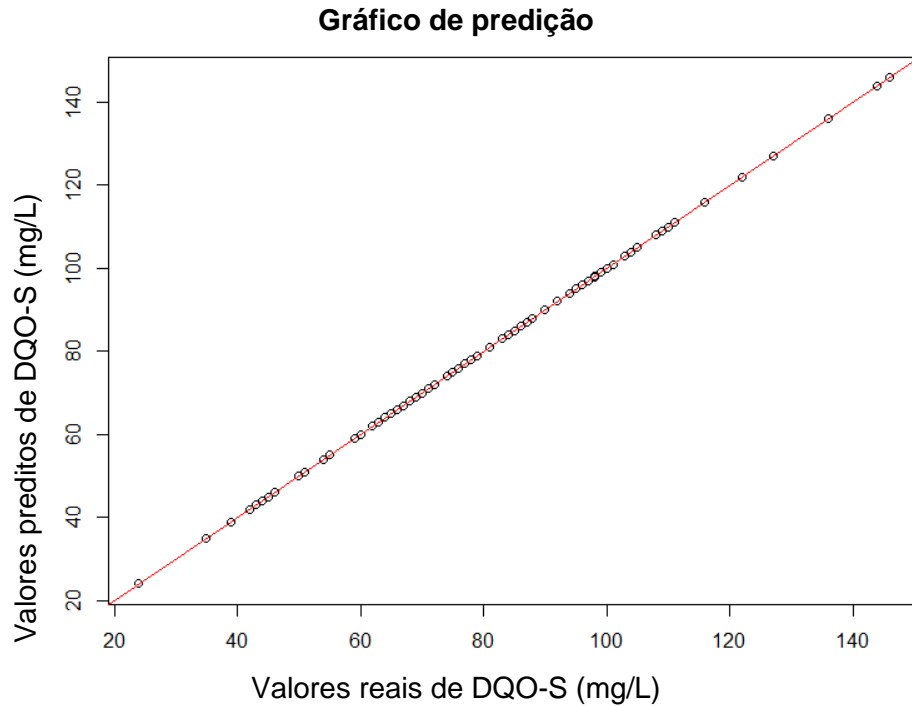


Figura 43 Valores preditos x Valores reais (Teste)

O gráfico abaixo mostra a análise dos resíduos da aplicação do conjunto de teste no modelo de SVM. Como pode-se observar, há uma boa distribuição dos resíduos em torno da origem, além de seguirem a distribuição normal. Estes também não apresentam *outliers* para o modelo.

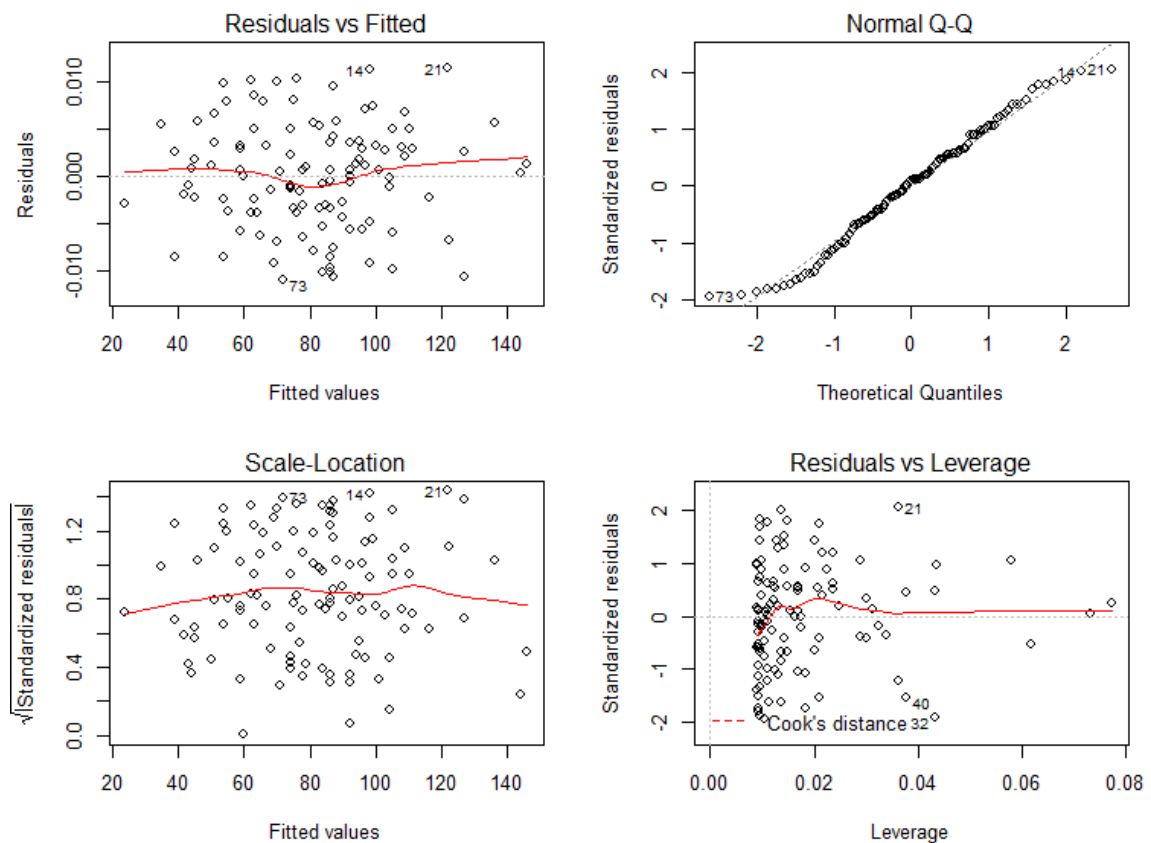


Figura 44 Análise dos resíduos (Teste)

A tabela abaixo mostra os parâmetros de mérito para o modelo SVM, os quais, evidentemente, são superiores aos modelos apresentados. Observe novamente o quanto o modelo é robusto em relação há aplicação do conjunto de teste, logo que este possui parâmetros de resíduos muito inferiores em relação ao banco de treinamento.

Tabela 4 Parâmetros de mérito (SVM)

Parâmetros de mérito	Conjunto de Treinamento	Conjunto de Teste
R ²	0,99	1,00
MAE	0,36	0,4.10 ⁻⁴
MSE	13,20	0,3.10 ⁻⁶
RMSE	3,63	0,6.10 ⁻⁴

Na seguinte tabela, estão reunidos os parâmetros de mérito dos modelos gerados, assim, podemos observar com clareza que o modelo de regressão por máquina de vetor de suporte foi muito superior aos demais modelos, logo que este obteve valor baixos para os parâmetros de medida de resíduos e um excelente ajuste.

Tabela 5 Parâmetros de mérito dos modelos

Modelos	Parâmetros de mérito	Conjunto de Treinamento	Conjunto de Teste
Regressão Linear Múltipla	R ²	0,34	0,27
	MAE	19,63	16,08
	MSE	675,43	406,49
	RMSE	25,99	20,16
Regressão por Componentes Principais	R ²	0,20	0,07
	MAE	20,62	20,82
	MSE	738,06	721,10
	RMSE	27,16	26,85
Regressão por Mínimos Quadrados Parciais	R ²	0,23	0,10
	MAE	20,22	20,70
	MSE	710,03	699,15
	RMSE	26,65	26,44
Regressão por Máquina de Vetor de Suporte	R ²	0,99	1,00
	MAE	0,36	0,4.10 ⁻⁴
	MSE	13,20	0,3.10 ⁻⁶
	RMSE	3,63	0,6.10 ⁻⁴

9. CONCLUSÃO

Neste trabalho, foi apresentado como o uso de ferramentas estatísticas avançadas, que são de uso habitual na quimiometria, podem contribuir para a operação de uma ETE, que é uma parte vital para toda a indústria que faça o descarte do efluente em um corpo hídrico, possibilitando a previsão de situações futuras, contribuindo assim para a gestão ambiental da organização que porte deste tipo de tratamento.

Como a DQO é uma medida rápida e fácil para a carga orgânica total, e a sua relação com a DBO pode ser facilmente estabelecida com base em dados históricos. A DQO foi a melhor escolha para este estudo de predição. Lembrando que neste estudo foram utilizados os dados históricos de 21 variáveis ao longo das etapas da ETE, deste modo os modelos gerados neste estudo são aplicados somente na estação específica deste estudo, devido às variações operacionais e os equipamentos específicos desta ETE. Entretanto, as ferramentas de regressão utilizadas neste estudo servem como base para a criação de outros modelos para outras estações de tratamento.

Portanto, podemos concluir que as relações entre as variáveis estudadas como um todo não são lineares, apesar dos grupos de variáveis pH, condutividade e sólidos em suspensão voláteis apresentarem relações lineares entre si, isto é minimizado quando relacionado com todas as variáveis do banco de dados, pois os modelos que são basicamente lineares, como a regressão linear múltipla, o PCR e o PLSR resultaram em baixos desempenhos de previsão da DQO, enquanto o SVM obteve um excelente desempenho de previsão.

Temos que ressaltar também que a similaridade entre os modelos PCR e PLSR foram evidentes neste estudo, pois ambos resultaram em um componente principal e uma variável latente, além de explicarem somente 35,7% da variabilidade da variável resposta, o que acarretou em um baixo índice de previsão da DQO na saída da estação.

A máquina de vetores de suporte utilizando o *kernel* de base radial foi a alternativa em relação aos métodos lineares, que resultou em uma excelente previsão no modelo final, confirmando assim que as interações entre as variáveis nas estações de tratamento seguem um padrão de interação não-linear.

10. REFERÊNCIAS BIBLIOGRÁFICAS

- ADLER, D.; MURDOCH, D. rgl: 3D Visualization Using OpenGL. R package version 0.98.1. 2017. <https://CRAN.R-project.org/package=rgl>
- ALLAIRE, J. J.; CHENG, J.; XIE, Y.; Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, Rob Hyndman and Ruben Arslan (2017). rmarkdown: Dynamic Documents for R. R package version 1.6. <https://CRAN.R-project.org/package=rmarkdown>
- BELANCHE, L.; CORTES, U.; SANCHEZ, M. "A knowledge-based system for the diagnosis of waste-water treatment plant". Proceedings of the 5th international conference of industrial and engineering applications of AI and Expert Systems IEA/AIE-92. Ed Springer-Verlag. Paderborn, Germany, June 1992.
- BEJAR, J.; CORTES, U.; POCH, M. "LINNEO+: A Classification Methodology for Ill-structured Domains". Research report RT-93-10-R. Dept. Llenguatges i Sistemes Informatics. Barcelona. 1993.
- CHEN, N.; LU, U.; VANG, J.; LI, G. Support vector machine in chemistry. World Scientific Publishing, 2004.
- ESBENSEN, K. H.; GELADI, P.; Principles of Proper Validation: use and abuse of re-sampling for validation. *Journal of Chemometrics*, v. 24: 168–187, 2010. DOI: 10.1002/cem.1310
- FERREIRA, M. M. C.; *Quimiometria – Conceitos, Métodos e Aplicações*. Campinas, SP; Editora Unicamp, 2015.
- GORELICK, M. H. Bias arising from missing data in predictive models. *Journal of Clinical Epidemiology*, v. 59 p. 1115-1123, 2006.
- HUBER, M.; ROUSSEEUW, P. J.; BRANDEN, K. V. ROBPCA: A New Approach to Robust Principal Component Analysis. *American Statistical Association and the American Society for Quality. TECHNOMETRICS*, February 2005, VOL. 47, NO. 1
- KARATZOGLOU, A.; MEYER, D.; HORNIK, K. Support Vector Machines in R. *Journal of Statistical Software*, Vo 15, April 2006.
- KENNARD, R. W.; STONE, L. A. Computer aided design of experiments. *Technometrics* 1969, 11, 137-148.
- KUHN, M.; KJELL, J. *Applied Predictive Modeling*. New York: Springer, 2013.
DOI 10.1007/978-1-4614-6849-3 1
https://vuquangnguyen2016.files.wordpress.com/2018/03/applied-predictive-modeling-max-kuhn-kjell-johnson_1518.pdf
- LE, S.; JOSSE, J.; HUSSON, F.; FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1-18. 2018. 10.18637/jss.v025.i01
- LI, H.; LIANG, Y.; XU, Q. Support Vector Machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems*, vol. 95, p. 188–198, 2009.

- LI-JUAN, W.; CHAO-BO, C.; Support Vector Machine Applying in the Prediction of Effluent Quality of Sewage Treatment Plant with Cyclic Activated Sludge System Process. IEEE International Symposium on Knowledge Acquisition and Modeling Workshop. 2008.
- LOGAN, M. Biostatistical Design and Analysis Using R: A Practical Guide. 1st ed. Wiley-BlackWeel, Oxford: 2010.
- MAECHLER, M.; ROUSSEUW, P.; CROUX, C.; TODOROV, V.; RUCKSTUHL, A.; BARRERA, M. S.; VERBEKE, T.; KOLLER, M.; PALMA, M. C. A. robustbase: Basic Robust Statistics R package version 0.92-8. 2017. URL <http://CRAN.R-project.org/package=robustbase>
- MEYER, D.; DIMITRIADOU, E.; HORNIK, K.; WEINGESSEL A.; Leisch, F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071). R package version 1.7-0. 2018. <https://CRAN.R-project.org/package=e1071>
- MEVIK, B.; WEHRENS, R.; LILAND, R. H.; pls: Partial Least Squares and Principal Component Regression. R package version 2.6-0, 2016. <https://CRAN.R-project.org/package=pls>
- PÉREZ-BENEDITO, D.; RUBIO, S. Environmental analytical chemistry. Elsevier, Amsterdam, 1999.
- PONS, M. N.; WU, J. POTIER, O. Chemometric estimation of wastewater composition for the on-line control of treatment plants. 16th Triennial World Congress, Prague, Czech Republic.
- POPPI, R. Palestra sobre “Máquinas de Vetor de Suporte” apresentada no V Workshop de Quimiometria, Ilhéus – BA, 2015.
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL <https://www.R-project.org/>.
- REVELLE, W. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, 2017. <https://CRAN.R-project.org/package=psych> Version = 1.7.5.
- SALCEDO-SANZ, S.; ROJO-ÁLVAREZ, J. L.; MARTÍNEZ-RAMÓN, M.; CAMPS-VALLS, G. Support vector machines in engineering: an overview. WIREs Data Mining Knowl Discov 2014, 4:234–267. doi: 10.1002/widm.1125
- SANTOS-FERNANDEZ, E. (2013). Multivariate Statistical Quality Control Using R. Springer, 14. URL <http://www.springer.com/statistics/computational+statistics/book/978-1-4614-5452-6>.
- SHEATHER, S. J. A Modern Approach to Regression with R. New York, NY: Springer, 2009. DOI: 10.1007/978-0-387-09608-7_3
- STEVENS, A.; RAMIREZ-Lopez, L. An introduction to the prospectr package. R package Vignette R package version 0.1.3, 2013.
- TANG, Y.; HORIKOSHI, M.; LI, W. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal, p. 478-489, 2016. ggfortify: Data Visualization Tools for

Statistical Analysis Results. <https://CRAN.R-project.org/package=ggfortify>

TEPPOLA, P.; MUJUNEN, S.-P.; MINKKINEN, P. Partial least squares modeling of an activated sludge plant: A case study. *Chemometrics and Intelligent Laboratory Systems*, vol 38, p. 197-208, 1997.

TEPPOLA, P.; MUJUNEN S.-P.; MINKKINEN, P. A combined approach of partial least squares and fuzzy c-means clustering for the monitoring of an activated-sludge waste-water treatment plant. *Chemometrics and Intelligent Laboratory Systems*, vol 41, p. 95–103. 1998.

TODOROV, V.; FILZMOSE, P. An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, 32(3), 1-47, 2009. URL <http://www.jstatsoft.org/v32/i03/>.

UCI Machine Learning repository.

<<https://archive.ics.uci.edu/ml/datasets/water+treatment+plant>> Acessado em: 30/10/2017

VAPNIK, V. *The Nature of Statistical Learning Theory*, Second Edition, Springer, New York, 1999.

WEHRENS, R. *Chemometrics with R - Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Springer, 2011. DOI 10.1007/978-3-642-17841-2

WESTAD, F.; MARINI, F. Validation of chemometric models - a tutorial, *Analytica Chimica Acta*, 2015, doi: 10.1016/j.aca.2015.06.056.

YADAV, M. L., ROYCHOUDHURY, B. *Handling Missing Values: A study of Popular Imputation Packages in R*, *Knowledge-Based Systems*, 2018. Doi: 10.1016/j.knosys.2018.06.012

YAN, Y. *MLmetrics: Machine Learning Evaluation Metrics*. R package version 1.1.1. 2016. <https://CRAN.R-project.org/package=MLmetrics>

11. ANEXOS

Anexo A - Comandos no R

title: "Modelagem da DQO de uma ETE"

author: "Paulo Cezario"

date: "24 de julho de 2018"

output:

html_document:

df_print: paged

```
```{r setup, include=FALSE}
```

```
knitr::opts_chunk$set(
```

```
 message = FALSE,
```

```
 warning = FALSE,
```

```
 comment = NA
```

```
)
```

```
````
```

```
# Dados
```

```
```{r data, include=FALSE}
```

```
getwd()
```

```
library(readxl)
```

```
Dados <- read_excel("D:/PAULO/Google Drive/Mestrado - PPG EQ UERJ/Tese/Water
Treatment Plant Data Set/WWT_IFRJ/WWT data.xlsx",
```

```
 sheet = "Dados", na = "?")
```

```
Seleção das classes
```

```
normal<-c("Normal situation-1","Normal situation-2",
```

```
 "Normal situation with performance over the mean",
```

```
 "Normal situation with low influent")
```

```
Dat<-Dados[Dados$OrigClass %in% normal,]
```

```
Dat<-Dat[,-c(2,23,24,26:29)]
```

```
WWT<-na.omit(Dados[,1:29]);WWT<-WWT[,-2] # Omisão de valores faltantes e da variável
Zinco
```

```
Dat<-na.omit(Dat)
```

```
dat.sc<-scale(Dat[,1:22], scale=T)
```

```
...
```

```
Visualização dos dados
```

```
```{r boxplots, echo=FALSE, fig.width=25, fig.height=20}
```

```
# Gráficos para comparação
```

```
boxplot(WWT,main='Dados puros')
```

```
boxplot(scale(na.omit(WWT), scale=T), main='Dados autoescalonados')
```

```
...
```

```
```{r corr, echo=FALSE, fig.width=10, fig.height=8}
```

```
In <- WWT[,1:8] # Entrada na estação de tratamento de efluentes
```

```
Out <- WWT[,22:27] # Saída na estação de tratamento de efluentes
```

```
pH<-WWT[,c(2,9,15,22)] # pH
```

```
DBO<-WWT[,c(3,10,16,23)]
```

```
DQO <- WWT[,c(4,17,24)]
```

```
SS <- WWT[,c(5,11,18,25)]
```

```
SSV <- WWT[,c(6,12,19,26)]
```

```
SED <- WWT[,c(7,13,20,27)]
```

```
COND <- WWT[,c(8,14,21)]
```

```
library(corrgram)
```

```
corrgram(cor(na.omit(WWT), method = c("spearman")),
```

```
 type = "cor",
```

```
 lower.panel = panel.shade,
```

```
 upper.panel = panel.pie,
```

```
 diag.panel=panel.density)
```

```
heatmap
```

```
col<- colorRampPalette(c("blue", "white", "red"))(20) # Get some colors
```

```
heatmap(cor(na.omit(WWT)), col = col, symm = TRUE)
```

```
HCPC
```

```
Practical Guide to Cluster Analysis in R - Book p.144
```

```
library("factoextra")
```

```
library("magrittr")
```

```
library("FactoMineR")
```

```

Compute hierarchical clustering
res.hc <- t(na.omit(WWT)) %>%
 get_dist(method = "pearson") %>% # Compute dissimilarity matrix
 hclust(method = "ward.D2") # Compute hierachical clustering
Visualize using factoextra
fviz_dend(res.hc)
...

Gráficos das variáveis da ETE
```{r matplots, echo=FALSE, fig.width=8, fig.height=6}
# Gráficos dos dados
for(i in list(WWT[,1], COND, DBO, DQO, SS, SED)) {
  matplot(i, type = c("l"),col = 1:4, ylab="")
  legend("topleft", legend = names(i), col=1:4, pch=19)
}
matplot(pH, type = c("l"),col = 1:4)
legend("topright", legend = names(pH), col=1:4, pch=19)
matplot(SSV, type = c("l"),col = 1:4)
legend("bottomleft", legend = names(SSV), col=1:4, pch=19)
...

# Correlogramas das variáveis da ETE
```{r Corrplots, echo=FALSE, fig.width=8, fig.height=6}
Correlogramas
for(i in list(pH, COND, DBO, DQO, SS, SED, SSV)) {
 library(psych)
 pairs.panels(i,
 method = "spearman", # correlation method
 hist.col = "#00AFBB",
 density = TRUE, # show density plots
 ellipses = TRUE) # show correlation ellipses
}
...

Análise dos grupos
```{r pca, echo=FALSE, fig.width=8, fig.height=6}
# Principal component analysis

```

```

library(FactoMineR)
dat.pca <- PCA(Dat[,1:22], scale.unit = TRUE, graph = FALSE)
library(factoextra)
fviz_pca_ind(dat.pca,
             geom.ind = "point",
             col.ind = Dat$OrigClass,
             palette= "jco",
             addEllipses=TRUE,
             ellipse.level=0.95,
             legend.title="Groups")

fviz_pca_biplot(dat.pca,
               col.var = "black",
               col.ind = Dat$OrigClass,
               palette= "jco",
               addEllipses=TRUE, label = "var",
               ellipse.level=0.95,
               legend.title="Groups")

mod.gen <- lm(Dat$`DQO-S` ~., data=Dat[,1:23])
summary(mod.gen)
par(mfrow=c(2,2));plot(mod.gen)

mod.gen2 <- step(mod.gen, direction = "backward")
summary(mod.gen2)
par(mfrow=c(2,2));plot(mod.gen2)

library("ggfortify")
autoplot(mod.gen2, which = 1:6, label.size = 3, data = Dat[,1:23] ,colour = 'OrigClass')

par(mfrow=c(2,2))
boxplot(scale(Dat[which(Dat$OrigClass == 'Normal situation-1'), 1:22], scale=T),
        main="Normal situation-1" autoescalonados)
boxplot(scale(Dat[which(Dat$OrigClass == 'Normal situation-2'), 1:22], scale=T),
        main="Normal situation-2" autoescalonados)
boxplot(scale(Dat[which(Dat$OrigClass == 'Normal situation with performance over the
mean'), 1:22], scale=T), main="Normal situation with performance over the mean"

```

```

autoescalonados')
boxplot(scale(Dat[which(Dat$OrigClass == 'Normal situation with low influent'), 1:22],
scale=T), main="Normal situation with low influent" autoescalonados')
```

Controle Estatístico Multivariado de Qualidade
```{r MSQC, echo=FALSE, fig.width=8, fig.height=6}
# MSQC
library(MSQC)
#Multivariate Control Chart
mt2DQO<-mult.chart(type="t2", na.omit(as.matrix(DQO)))
# The following(s) point(s) fall outside of the control limits [1] 36 49 53 58 59 95 96 142
213 227 309 320 333 434 442 444 445
out.t2<-c(36, 49, 53, 58, 59, 95, 96, 142, 213, 227, 309, 320, 333, 434, 442, 444, 445)
Dados[out.t2,30] # Classes dos pontos fora do limite de controle baseado no T2 de Hotelling
DQO.sc<-scale(na.omit(DQO), scale=T)
DQO.sc<-as.data.frame(DQO.sc)
DQO.sc[out.t2,]
boxplot(DQO.sc, main=" Dados de DQO Autoescalonados")
stripchart(DQO.sc[out.t2,], vertical = T,
           method = "jitter", add=T, pch=20, col= "red")
```

Regiões do processo
```{r proc reg, echo = FALSE, results = TRUE}
proc.reg(na.omit(DQO), type ="t2", alpha=0.05) # Process region
```

```{r MSQC 3d, echo=FALSE, fig.width=8, fig.height=6}
# Gráfico 3d das variáveis DQO
library(rgl)
library(rglwidget)
# Elipse de confiança das variáveis DQO
plot3d(ellipse3d(cov(na.omit(DQO)), centre=colMeans(na.omit(DQO)), level=.95),
       xlab="DQO-E", ylab="DQO-D", zlab="DQO-S",type="wire", col="red")
# Pontos das variáveis DQO
points3d(na.omit(DQO), size=4, cex=2, add=TRUE)

```

```

# Prisma com os Limites Superior e Inferior do Processo
prism(proc.reg(na.omit(DQO), alpha=0.05)$LPL,
      proc.reg(na.omit(DQO), alpha=0.05)$UPL, add=TRUE,col=3)
text3d(na.omit(DQO),texts=row.names(na.omit(DQO)), col=4)
rglwidget()
rgl.clear()
#movie3d(spin3d(axis = c(0, 0, 1)), duration = 9)
#movie3d(spin3d(axis = c(0, 1, 0)), duration = 9)
...

# Exclusão de outliers: PcaHubert
``{r PcaHubert1, results=T, echo=F, fig.width=8, fig.height=6}
# Outlier Detection with Robust PCA
# Chemometrics With R - Book pg 235
require(rrcov)
set.seed(123)
dat.HubPCA<-PcaHubert(Dat[,1:22], scale=T, alpha= 0.75, crit.pca.distances=0.95)
dat.CPCA<-PcaClassic(Dat[,1:22], scale=T, crit.pca.distances=0.95)
# Sumérios dos PCA's
summary(dat.HubPCA)
summary(dat.CPCA)
# Gráficos dos PCA's
rrcov::screeplot(dat.HubPCA, type="lines", main="Screeplot: Robust PCA")
rrcov::screeplot(dat.CPCA, type="lines", main="Screeplot: Classic PCA")
biplot(dat.HubPCA, main="Biplot: Robust PCA")
biplot(dat.CPCA, main="Biplot: Classic PCA")
scorePlot(dat.HubPCA)
scorePlot(dat.CPCA)
# Outliers
plot(dat.HubPCA)
dat.HubPCA@cutoff.od
dat.HubPCA@cutoff.sd
plot(dat.CPCA@sd, dat.CPCA@od, xlab="Orthogonal distance",
      ylab="Score distance", main="Classic PCA")
abline(h= dat.CPCA@cutoff.od, col="red", lty=2)
abline(v= dat.CPCA@cutoff.sd, col="red", lty=2)
dat.CPCA@cutoff.od

```



```

dat.CPCA@cutoff.sd
# Remoção de outliers baseado no PCA robusto (sd e od)
dat.HubPCA@cutoff.sd
dat.HubPCA@cutoff.od

outliers1 <- subset(cbind(dat.HubPCA@sd,dat.HubPCA@od),
                    dat.HubPCA@sd > dat.HubPCA@cutoff.sd &
                    dat.HubPCA@od > dat.HubPCA@cutoff.od);row.names(outliers1)
plot(dat.HubPCA@sd, dat.HubPCA@od, xlab="Orthogonal distance",
     ylab="Score distance", main="PCA robusto")
abline(h= dat.HubPCA@cutoff.od, col="red", lty=2)
abline(v= dat.HubPCA@cutoff.sd, col="red", lty=2)
text(x= outliers1[,1], y= outliers1[,2],
     labels = row.names(outliers1), pos = 1)
# identify(dat.HubPCA@sd, dat.HubPCA@od)

Dat[c(as.numeric(row.names(outliers1))),"OrigClass"]

dat.sc<-as.data.frame(dat.sc)
boxplot(dat.sc, main=" Dados Autoescalonados")
stripchart(dat.sc[c(as.numeric(row.names(outliers1))),], vertical = T,
           method = "jitter", add=T, pch=20, col= "red")

Dat2<-Dat[-c(as.numeric(row.names(outliers1))),]
dat2.sc<-dat.sc[-c(as.numeric(row.names(outliers1))),]
...

```{r PcaHubert1boxplot, results=T, echo=F, fig.width=25, fig.height=20}
dat.sc<-as.data.frame(dat.sc)
boxplot(dat.sc, main=" Dados Autoescalonados")
stripchart(dat.sc[c(as.numeric(row.names(outliers1))),], vertical = T,
 method = "jitter", add=T, pch=20, col= "red")
...

Seleção de amostras de treinamento e de teste
```{r prospectr, results=T, echo=F, fig.width=8, fig.height=6}

```

```

# Selection of samples for calibration and validation sets
library("prospectr")
# Using the Kennard-Stone algorithm to select the samples for calibration set for each class
# k = number of desired calibration samples
# pc = explained variance
set.seed(123)
dat.ks <- kenStone(dat2.sc, k=round(.7*(dim(dat2.sc)[1])), pc=.99)

# Create the graph to show the selected samples
plot(dat.ks$pc[,1:2], xlab="PC1", ylab="PC2", pch=17)
# The samples selected for the calibration set for class 1 (model= modelo de calibração)
points(dat.ks$pc[dat.ks$model, 1:2], pch=17, col=2)
legend("bottomleft",c('Training','Test'),col=c('red','black'),pch=17)

# Extraction of the selected samples to create the training and test sets for class 1 or C73
tra <- dat.ks$model # training
tst <- dat.ks$test # test

dat.tra<-Dat2[-tst,] # Training data
dat.tst<-Dat2[-tra,] # Test data
...

# Regressão Linear Múltipla
``{r GLM, results=T, echo=F, fig.width=8, fig.height=6}
# Alteração da variável resposta para DQO-D (decantador secundário)
# e variáveis de saída até o decantador primário
mod.lm <- lm(dat.tra$`DQO-S`~., data=dat.tra[,1:21])
summary(mod.lm)
par(mfrow=c(2,2));plot(mod.lm)
kable(summary(mod.lm)$coef, format.args = list(decimal.mark = ",", big.mark = ""))

mod.lm2 <- step(mod.lm, direction = "backward")
summary(mod.lm2)
library(knitr)
kable(summary(mod.lm2)$coef, format.args = list(decimal.mark = ",", big.mark = ""))
par(mfrow=c(1,1))
plot(dat.tra$`DQO-S`, predict(mod.lm2));abline(0,1, col='red')

```

```

par(mfrow=c(2,2));plot(mod.lm2)

anova(mod.lm, mod.lm2)

library(MLmetrics)
print("R2");R2_Score(predict(mod.lm2), as.matrix(dat.tra$`DQO-S`))
print("MAE");MAE(predict(mod.lm2), as.matrix(dat.tra$`DQO-S`))
print("MSE");MSE(predict(mod.lm2), as.matrix(dat.tra$`DQO-S`))
print("RMSE");RMSE(predict(mod.lm2), as.matrix(dat.tra$`DQO-S`))

#library("ggfortify"); autoplot(mod.lm2, which = 1:6, label.size = 3, data = dat.tra ,colour =
'OrigClass')

# Test data
pred.tst.lm<-predict(mod.lm2, dat.tst[,1:21])
par(mfrow=c(1,1))
plot(dat.tst$`DQO-S`, pred.tst.lm);abline(0,1, col='red')

print("R2");R2_Score(pred.tst.lm, as.matrix(dat.tst$`DQO-S`))
print("MAE");MAE(pred.tst.lm, as.matrix(dat.tst$`DQO-S`))
print("MSE");MSE(pred.tst.lm, as.matrix(dat.tst$`DQO-S`))
print("RMSE");RMSE(pred.tst.lm, as.matrix(dat.tst$`DQO-S`))
...

# Regressão Linear Robusta
```{r lmrob, results=T, echo=F, fig.width=8, fig.height=6}
set.seed(123)
library(robustbase)
summary(m1 <- lmrob(dat.tra$`DQO-S` ~., data=dat.tra[,1:21], setting = "KS2011"));plot(m1)

summary(m2 <- lmrob(dat.tra$`DQO-S` ~., data=dat.tra[,1:21], setting = "KS2014"));
par(mfrow=c(2,3));plot(m2); plot(residuals(m2) ~ weights(m2, type="robustness"));abline(h=0,
lty=3)

Test data
pred.tst.lmrob<-predict(m2, dat.tst[,1:21])
par(mfrow=c(1,1)); plot(pred.tst.lmrob, dat.tst$`DQO-S`);abline(0,1, col='red')

```

...

```

Regressão por Componentes Principais (PCR)
```{r PCR, results=T, echo=F, fig.width=8, fig.height=6}
library(pls)
# Alteração da variável resposta para DQO-D (decantador secundário)
# e variáveis de saída até o decantador primário
dat.pcr <- pcr(as.matrix(dat.tra[,22]) ~ as.matrix(dat.tra[,1:21]) ,
              ncomp = 21, data = dat.tra,
              validation="LOO", scale = TRUE, jackknife = TRUE)
summary(dat.pcr)

selectNcomp(dat.pcr, "randomization", plot = TRUE, alpha = 0.05)
selectNcomp(dat.pcr, "onesigma", plot = TRUE, alpha = 0.05)

jack.test(dat.pcr, ncomp = 1)

plot(dat.pcr, "validation", estimate = "CV", main="DBO-S")
plot(dat.pcr, "validation", val.type = "MSEP")
validationplot(dat.pcr, val.type="R2", main="DBO-S")

R2(dat.pcr, ncomp=1)
MSEP(dat.pcr, ncomp=1)
RMSEP(dat.pcr, ncomp=1)

pred.pcr.tra<- as.matrix(predict(dat.pcr, ncomp=1))

mod.lm.pcr<-lm(dat.tra$`DQO-S` ~ pred.pcr.tra)
summary(mod.lm.pcr)
par(mfrow=c(2,2));plot(mod.lm.pcr)

library(MLmetrics)
print("R2");R2_Score(pred.pcr.tra, as.matrix(dat.tra$`DQO-S`))
print("MAE");MAE(as.matrix(pred.pcr.tra), as.matrix(dat.tra$`DQO-S`))
print("MSE");MSE(pred.pcr.tra, as.matrix(dat.tra$`DQO-S`))
print("RMSE");RMSE(pred.pcr.tra, as.matrix(dat.tra$`DQO-S`))

```

```

coefplot(dat.pcr, ncomp = 1:4, main="DBO-S", legendpos = "bottomright")
scoreplot(dat.pcr, comps = 1:4)
loadingplot(dat.pcr, comps = 1:4)
corrplot(dat.pcr, comps = 1:4)

#Gráfico dos valores previstos pelo modelo PCR versus valores observados
predplot(dat.pcr, main="DQO-S",line=T, line.col="red",
         which = c("validation"), ncomp=2)
#legend("topleft",expression("R^2*=45,77 %"))

# Conjunto de Teste
pred1.pcr<-predict(dat.pcr, newdata = as.matrix(dat.tst[,c(1:21)]), ncomp=1)
plot(dat.tst$`DQO-S`, pred1.pcr); abline(0,1, col='red')
mod.pred1<-lm(dat.tst$`DQO-S` ~ pred1.pcr)
summary(mod.pred1)

print("R2");R2_Score(as.matrix(pred1.pcr), as.matrix(dat.tst$`DQO-S`))
print("MAE");MAE(as.matrix(pred1.pcr), as.matrix(dat.tst$`DQO-S`))
print("MSE");MSE(as.matrix(pred1.pcr), as.matrix(dat.tst$`DQO-S`))
print("RMSE");RMSE(as.matrix(pred1.pcr), as.matrix(dat.tst$`DQO-S`))
...

# Regressão por Mínimos Quadrados Parciais (PLSR)
```{r PLSR, results=T, echo=F, fig.width=8, fig.height=6}
PLSR
library(pls)
Alteração da variável resposta para DQO-D (decantador secundário)
e variáveis de saída até o decantador primário
dat.plsr <- pls(as.matrix(dat.tra[,22]) ~ as.matrix(dat.tra[,1:21]) , ncomp = 21, data =
dat.tra, scale = TRUE,
 validation="LOO", jackknife = TRUE)
summary(dat.plsr)

selectNcomp(dat.plsr, "randomization", plot = TRUE, alpha = 0.05)
selectNcomp(dat.plsr, "onesigma", plot = TRUE, alpha = 0.05)

```

```

plot(dat.plsr, "validation", estimate = "CV", main="DBO-S")
plot(dat.plsr, "validation", val.type = "MSEP")
validationplot(dat.plsr, val.type="R2", main="DBO-S")

jack.test(dat.plsr, ncomp = 1)

RMSEP(dat.plsr, ncomp = 1)
R2(dat.plsr, ncomp = 1)
MSEP(dat.plsr, ncomp = 1)

pred.pls.tra<- as.matrix(predict(dat.plsr, ncomp=1))

mod.lm.pls<-lm(dat.tra`DQO-S` ~ pred.pls.tra)
summary(mod.lm.pls)
par(mfrow=c(2,2));plot(mod.lm.pls)

library(MLmetrics)
print("R2");R2_Score(pred.pls.tra, as.matrix(dat.tra`DQO-S`))
print("MAE");MAE(as.matrix(pred.pls.tra), as.matrix(dat.tra`DQO-S`))
print("MSE");MSE(pred.pls.tra, as.matrix(dat.tra`DQO-S`))
print("RMSE");RMSE(pred.pls.tra, as.matrix(dat.tra`DQO-S`))

coefplot(dat.plsr, ncomp = 1:6, main="DBO-S", legendpos = "bottomright")
scoreplot(dat.plsr, comps = 1:6)
loadingplot(dat.plsr, comps = 1:6)
corrplot(dat.plsr, comps = 1:6)

Prediction plot for training
predplot(dat.plsr, main="DQO-S",line=T, line.col="red", ncomp=1)
#legend("topleft",expression("R^2*=45,77 %"))

Prediction plot for test
pred.tst.plsr<-predict(dat.plsr, newdata = as.matrix(dat.tst[,c(1:21)]), ncomp=1)
plot(dat.tst`DQO-S`, pred.tst.plsr); abline(0,1, col='red')

print("R2");R2_Score(as.matrix(pred.tst.plsr), as.matrix(dat.tst`DQO-S`))
print("MAE");MAE(as.matrix(pred.tst.plsr), as.matrix(dat.tst`DQO-S`))

```

```

print("MSE");MSE(as.matrix(pred.tst.plsr), as.matrix(dat.tst$`DQO-S`))
print("RMSE");RMSE(as.matrix(pred.tst.plsr), as.matrix(dat.tst$`DQO-S`))
...

Máquina de Vetor de Suporte (SVM)
``{r SVM, results=T, echo=F, fig.width=8, fig.height=6}
library(e1071)
set.seed(123)
svm.model <- tune(svm, as.matrix(dat.tra$`DQO-S`) ~ as.matrix(dat.tra[,1:21]), data= dat.tra,
ranges = list(epsilon = seq(0,1,0.1), cost = 2^(2:9)))

print(svm.model)
summary(svm.model)
str(svm.model)
Draw the tuning graph
plot(svm.model)

svm.pred.tra <- predict(svm.model$best.model)
plot(dat.tra$`DQO-S`, svm.pred.tra); abline(0,1, col='red')
mod.svm<-lm(dat.tra$`DQO-S` ~ svm.pred.tra)
summary(mod.svm)
par(mfrow=c(2,2));plot(mod.svm)

library(MLmetrics)
R2_Score(predict(svm.model$best.model), as.matrix(dat.tra$`DQO-S`))
MAE(predict(svm.model$best.model), as.matrix(dat.tra$`DQO-S`))
MSE(predict(svm.model$best.model), as.matrix(dat.tra$`DQO-S`))
RMSE(predict(svm.model$best.model), as.matrix(dat.tra$`DQO-S`))

svm.test <- svm(as.matrix(dat.tst$`DQO-S`)~as.matrix(dat.tst[,1:21]), data= dat.tst, epsilon =
svm.model$best.model$epsilon, cost=svm.model$best.model$cost)

par(mfrow=c(1,1));plot(dat.tst$`DQO-S`, predict(svm.test)); abline(0,1, col='red')
mod.svm.tst<-lm(dat.tst$`DQO-S` ~ predict(svm.test))
summary(mod.svm.tst)
par(mfrow=c(2,2));plot(mod.svm.tst)

```

```
R2_Score(predict(svm.test), as.matrix(dat.tst`DQO-S`))
MAE(predict(svm.test), as.matrix(dat.tst`DQO-S`))
MSE(predict(svm.test), as.matrix(dat.tst`DQO-S`))
RMSE(predict(svm.test), as.matrix(dat.tst`DQO-S`))

#svm.pred.tst <- predict(svm.model, as.matrix(Dat3[-tra,c(1:14)]))

...

Referências
``{r citations, results=T, echo=F, }
citation()
citation(package = "rmarkdown")
citation(package = "readxl")
citation(package = "psych")
citation(package = "MSQC")
citation(package = "rgl")
citation(package = "rrcov")
citation(package = "fitdistrplus")
citation(package = "prospectr")
citation(package = "car")
citation(package = "MASS")
citation(package = "pls")
citation(package = "e1071")
citation(package = "MLmetrics")
...

```



## Anexo B - Sumário do modelo de regressão linear múltipla

Abaixo pode-se observar o sumário do modelo, onde estão os valores do intercepto e dos coeficientes das variáveis (*Estimate*), assim como o erro padrão (*Std. Error*), o valor de *t* (*t value*), o valor-p (*Pr*) e o código de significância (\*) que identificado pela quantidade de asteriscos.

Call:

```
lm(formula = dat.tra$`DQO-S` ~ ., data = dat.tra[, 1:21])
```

Residuals:

	Min	1Q	Median	3Q	Max
	-84.303	-17.294	-2.932	14.801	90.069

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.730e+02	7.332e+01	3.723	0.000246	***
`Q-E`	7.258e-05	3.045e-04	0.238	0.811820	
`PH-E`	-1.054e+01	1.609e+01	-0.655	0.512952	
`DBO-E`	8.204e-02	4.081e-02	2.010	0.045571	*
`DQO-E`	2.568e-02	2.570e-02	0.999	0.318794	
`SS-E`	-9.349e-02	5.223e-02	-1.790	0.074757	.
`SSV-E`	-5.020e-01	3.542e-01	-1.417	0.157770	
`SED-E`	-1.261e+00	1.753e+00	-0.719	0.472637	
`COND-E`	-3.682e-02	1.889e-02	-1.949	0.052423	.
`PH-P`	-1.073e+00	2.009e+01	-0.053	0.957455	
`DBO-P`	-5.554e-02	3.811e-02	-1.457	0.146329	
`SS-P`	3.320e-03	4.277e-02	0.078	0.938203	
`SSV-P`	2.737e-01	2.893e-01	0.946	0.345186	
`SED-P`	2.621e+00	1.576e+00	1.663	0.097587	.
`COND-P`	3.303e-02	1.956e-02	1.689	0.092578	.
`PH-D`	-1.691e+01	1.778e+01	-0.951	0.342537	
`DBO-D`	-3.213e-02	9.010e-02	-0.357	0.721708	
`DQO-D`	2.559e-01	5.075e-02	5.042	9.19e-07	***
`SS-D`	-2.656e-01	1.539e-01	-1.726	0.085699	.
`SSV-D`	-3.892e-02	2.852e-01	-0.136	0.891563	
`SED-D`	-5.577e+00	7.930e+00	-0.703	0.482591	
`COND-D`	9.229e-03	1.223e-02	0.755	0.451086	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.85 on 236 degrees of freedom

Multiple R-squared: 0.357, Adjusted R-squared: 0.2997

F-statistic: 6.238 on 21 and 236 DF, p-value: 9.894e-14

### Anexo C - Sumário do modelo de regressão linear múltipla ajustado

Call:

```
lm(formula = dat.tra$`DQO-S` ~ `DBO-E` + `SS-E` + `COND-E` +
 `DBO-P` + `SED-P` + `COND-P` + `PH-D` + `DQO-D` + `SS-D`,
 data = dat.tra[, 1:21])
```

Residuals:

	Min	1Q	Median	3Q	Max
	-89.076	-16.023	-2.504	14.798	93.102

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	277.27413	68.15752	4.068	6.37e-05	***
`DBO-E`	0.07396	0.03642	2.030	0.043378	*
`SS-E`	-0.05522	0.02804	-1.969	0.050064	.
`COND-E`	-0.03591	0.01757	-2.044	0.041995	*
`DBO-P`	-0.05438	0.03608	-1.507	0.133072	
`SED-P`	1.95074	1.06221	1.837	0.067480	.
`COND-P`	0.03936	0.01731	2.274	0.023807	*
`PH-D`	-30.15045	8.74701	-3.447	0.000666	***
`DQO-D`	0.24518	0.03305	7.419	1.87e-12	***
`SS-D`	-0.31531	0.11023	-2.860	0.004592	**

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.51 on 248 degrees of freedom

Multiple R-squared: 0.3414, Adjusted R-squared: 0.3174

F-statistic: 14.28 on 9 and 248 DF, p-value: < 2.2e-16

## Anexo D - Sumário do modelo de regressão por componentes principais

Data: X dimension: 258 21

Y dimension: 258 1

Fit method: svdpc

Number of components considered: 21

VALIDATION: RMSEP

Cross-validated using 258 leave-one-out segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
CV	32.15	29.68	29.77	29.43	29.25	29.35	29.44
adjCV	32.15	29.68	29.77	29.43	29.24	29.35	29.44
	7 comps	8 comps	9 comps	10 comps	11 comps	12 comps	13 comps
CV	29.41	29.52	29.82	28.44	28.22	28.38	28.34
adjCV	29.41	29.52	29.90	28.43	28.22	28.38	28.34
	14 comps	15 comps	16 comps	17 comps	18 comps	19 comps	
CV	28.25	28.49	27.69	27.70	27.76	27.88	
adjCV	28.24	28.48	27.68	27.69	27.76	27.87	
	20 comps	21 comps					
CV	27.96	27.85					
adjCV	27.96	27.85					

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps
X	34.76	51.83	62.33	70.83	77.82
as.matrix(dat.tra[, 22])	15.28	15.44	17.92	19.73	19.80
	6 comps	7 comps	8 comps	9 comps	10 comps
X	81.68	84.91	87.85	89.90	91.92
as.matrix(dat.tra[, 22])	19.92	20.82	21.08	22.01	27.72
	11 comps	12 comps	13 comps	14 comps	15 comps
X	93.46	94.72	95.77	96.73	97.49
as.matrix(dat.tra[, 22])	28.94	28.94	29.75	30.54	30.54
	16 comps	17 comps	18 comps	19 comps	20 comps
X	98.20	98.73	99.16	99.55	99.89
as.matrix(dat.tra[, 22])	33.97	34.46	34.67	34.68	34.91
	21 comps				
X	100.0				
as.matrix(dat.tra[, 22])	35.7				

## Anexo E - Sumário do modelo de regressão por mínimos quadrados parciais

Data: X dimension: 258 21

Y dimension: 258 1

Fit method: kernelpLS

Number of components considered: 21

VALIDATION: RMSEP

Cross-validated using 258 leave-one-out segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
CV	32.15	29.26	28.74	28.15	27.96	27.86	27.77
adjCV	32.15	29.26	28.74	28.14	27.96	27.86	27.76
	7 comps	8 comps	9 comps	10 comps	11 comps	12 comps	13 comps
CV	27.9	27.91	27.91	27.92	27.95	27.93	27.91
adjCV	27.9	27.91	27.90	27.92	27.94	27.93	27.91
	14 comps	15 comps	16 comps	17 comps	18 comps	19 comps	
CV	27.85	27.85	27.85	27.85	27.85	27.85	
adjCV	27.84	27.85	27.84	27.84	27.85	27.85	
	20 comps	21 comps					
CV	27.85	27.85					
adjCV	27.85	27.85					

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps
X	34.10	43.66	50.73	63.10	70.67
as.matrix(dat.tra[, 22])	18.97	26.13	31.51	33.13	33.91
	6 comps	7 comps	8 comps	9 comps	10 comps
X	74.6	77.83	83.44	85.51	88.02
as.matrix(dat.tra[, 22])	34.8	35.15	35.24	35.39	35.50
	11 comps	12 comps	13 comps	14 comps	15 comps
X	89.36	91.25	92.95	93.69	95.07
as.matrix(dat.tra[, 22])	35.61	35.66	35.68	35.69	35.70
	16 comps	17 comps	18 comps	19 comps	20 comps
X	96.24	97.15	97.97	98.84	99.24
as.matrix(dat.tra[, 22])	35.70	35.70	35.70	35.70	35.70
	21 comps				
X	100.0				
as.matrix(dat.tra[, 22])	35.7				