



Licenciatura em Matemática Aplicada à Tecnologia e à Empresa



Assessoria Técnica e Qualidade de Produtos

Modelação estatística empírica e análise de componentes principais em
acompanhamento de lubrificantes em serviço de manutenção proactiva

GALP MODe AIML

Relatório de estágio curricular

Rafaela Couchinho Lino

Orientador GALP: Paulo Miguel Ferreira

Carlos Miguel Mota

Orientadoras ISEL: Sandra Aleixo

Iola Pinto

Lisboa, Julho de 2019

RESUMO

O acompanhamento de lubrificantes em serviço é uma ferramenta essencial para a manutenção preditiva e pró-ativa de equipamentos críticos, no sector industrial. O seu *output* é um diagnóstico qualitativo, da condição do lubrificante, com evidente associação com o estado do equipamento.

Existem vários fatores que influenciam a vida útil e a fiabilidade de qualquer equipamento mecânico crítico, entenda-se como crítico, um equipamento cuja função não pode ser interrompida, caso contrário obriga à paragem da produção da fábrica. Muitos desses fatores podem ser supervisionados através da análise do lubrificante em serviço. Obviamente, é importante detetar o desgaste anormal da máquina ou a degradação do estado físico-químico do lubrificante, mas o objetivo fundamental é evitar a avaria.

Os lubrificantes são um componente essencial nos equipamentos com partes móveis, evitando atrito entre superfícies metálicas, sendo capazes de suportar cargas elevadas, oferecendo proteção contra o desgaste e corrosão.

O diagnóstico das análises a uma amostra de refletem a condição do estado do lubrificante, à data da sua recolha. Frequentemente não é possível indicar a estimativa do ponto em que se quer atuar, sem recorrer a uma maior frequência e/ou ampliação das análises ao lubrificante em serviço. Situar no tempo o momento preciso em que o lubrificante deixa de estar em condições para proteger o equipamento é uma limitação.

A partir dessa necessidade de previsão do tempo ótimo de utilização do lubrificante surge como objetivo revisar a condição da máquina baseada no acompanhamento do lubrificante em serviço e relacioná-la com um modelo.

Assim, ao ajustar um modelo cuja variável resposta traduza de forma quantitativa o diagnóstico a um tipo de lubrificante, torna-se possível prever futuras ações de manutenção, preservação da condição do lubrificante e do equipamento.

Deste modo, neste trabalho foram utilizados os modelos de regressão linear múltipla, o modelo de regressão logística e também se perspetivou a utilização da análise de sobrevivência, designadamente o modelo de regressão de Cox. A seleção de preditores para cada modelo foi efetuada em diversas etapas, tendo em conta os resultados da análise exploratória, as correlações existentes entre as diversas variáveis, os resultados dos testes de normalidade e Análise de Componentes Principais (ACP).

Por fim, é importante salientar que o *output* do modelo foi sempre a previsão de variável com informação capaz de se traduzir em ações de manutenção.

AGRADECIMENTOS

Com a finalização deste Relatório de Estágio não posso deixar de agradecer a algumas pessoas que, direta ou indiretamente, me ajudaram nesta caminhada tão importante na minha vida pessoal e profissional.

Em primeiro lugar, agradeço a orientação que as professoras Sandra Aleixo e Iola Pinto me disponibilizaram, todo o apoio prestado e a disponibilização de ferramentas que se revelaram essenciais para a realização deste estágio.

Aqui presto também o meu agradecimento ao meu orientador da Galp Engenheiro Paulo Ferreira por toda a sua orientação, disponibilidade e ensinamentos.

Ao Departamento da Assessoria Técnica e Qualidade de Produtos da GALP de Lisboa, em especial, ao Engenheiro Carlos Mota e Engenheiro Fernando Cavaco agradeço pela simpatia que tiveram comigo.

À Galp Lisboa, S.A. pela oportunidade de realização deste estágio.

Por fim, mas não menos importante, agradeço à minha família que sem ela dificilmente conseguiria chegar até aqui. Obrigada aos meus pais, à irmã e ao meu cunhado, por estarem sempre lá para me amparar, para me criticar e para me congratular. Obrigada ao Fábio, por me ter feito sentir capaz de superar qualquer adversidade estando ao meu lado e obrigada aos meus amigos por todo o apoio e incentivo dado ao longo deste meu percurso académico.

ÍNDICE

RESUMO	i
AGRADECIMENTOS	ii
ÍNDICE	iii
ÍNDICE DE TABELAS	v
ÍNDICE DE FIGURAS	vi
1. Introdução	1
1.1. Lubrificantes e Lubrificação	1
1.2. Modelação Estatística Empírica	4
1.3. Objetivos	9
2. Apresentação de dados	11
2.1. Variáveis de Entrada	11
2.1.1. Estado Físico-Químico	11
2.1.2. Desgaste	14
2.1.3. Contaminação.....	16
2.1.4. Variáveis Externas ao Lubrificante	17
2.2. Variáveis de saída	18
2.2.1. Horas de funcionamento do lubrificante	18
2.2.2. Atesto Acumulado.....	19
3. Caso de Estudo Preliminar	20
3.1. Análise Exploratória dos dados	21
3.1.1. Medidas descritivas	21
3.1.2. Teste de Normalidade	24
3.1.3. Correlações entre Variáveis	24
3.1.4. Análise de Componentes Principais	26
3.2. Modelo de Regressão Linear	28
3.2.1. Análise de Regressão Linear Univariável	29
3.2.2. Seleção de Preditores	30
3.2.3. Modelo de regressão Linear Múltipla.....	31
3.2.4. Análise de Resíduos.....	32
3.2.5. Previsão.....	33
3.3. Modelo de Regressão Logística	35

3.3.1.	Análise de Regressão Logística Univariável.....	36
3.3.2.	Seleção de Preditores	37
3.3.3.	Modelo de Regressão Logística	38
3.3.4.	Análise de Resíduos.....	40
3.3.5.	Previsão.....	41
3.4.	Análise de Sobrevida – Modelo de Cox.....	42
3.4.1.	Análise de Modelo de Cox Univariável	43
3.5.	Discussão de Resultados do Caso de Estudo Preliminar	45
4.	Modelação de um Caso de Estudo.....	46
4.1.	Prensa 1	48
4.2.	Prensa 3	56
4.3.	Discussão de Resultados Caso de Estudo	58
5.	Conclusão e Sugestões de Trabalho Futuro	59
BIBLIOGRAFIA		61
ANEXO A- Boxplots (Caixa de Bigodes) dos dados do Caso de Estudo Preliminar.....		63
ANEXO B-Correlações entre Variáveis dos dados do Caso de Estudo Preliminar.....		65
ANEXO C- Gráficos de dispersão entre Variáveis dos dados do Caso de Estudo Preliminar.....		66

ÍNDICE DE TABELAS

Tabela 1-Variáveis do agrupamento Estado Físico-Químico presente no universo de dados.....	12
Tabela 2- Variáveis do agrupamento Desgaste presente no universo de dados.	14
Tabela 3- Variáveis do agrupamento Contaminação presente no universo de dados.	16
Tabela 4- Variáveis Externas ao Lubrificante presente no universo de dados.	17
Tabela 5 - Medidas Descritivas das Variáveis.....	22
Tabela 6 - Exemplo de resultados dos testes à significância dos coeficientes de correlação.....	25
Tabela 7 - Resultados da dos modelos de regressão univariável.	29
Tabela 8 - Resultados finais da modelação com regressão Linear Múltipla.....	34
Tabela 9 - Resultados dos modelos de regressão logística univariável.	36
Tabela 10 - Análise dos modelos de Cox univariável.	43
Tabela 11 - Resultados finais para o modelo do rolo fixo LA, da prensa 1.	48
Tabela 12 - Resultados finais para o modelo do rolo fixo LOA, da prensa 1....	50
Tabela 13 - Resultados finais para o modelo do rolo móvel LA, da prensa 1. .	52
Tabela 14 - Resultados finais para o modelo do rolo móvel LOA, da prensa 1.	54
Tabela 15-Previsão para modelo do rolo fixo LOA, da prensa 3.....	56

ÍNDICE DE FIGURAS

Figura 1-Boletim de análise de amostras de lubrificante.....	3
Figura 2- Representação da vida útil do lubrificante. a) quando ciclicamente é totalmente renovado; b) quando é gradualmente renovado.....	18
Figura 3-Boxplot dos dados da variável Sulfatação	23
Figura 4-Gráficos análise de resíduos de modelo de regressão linear múltipla: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.....	32
Figura 5-Curva ROC modelo de regressão logística.....	40
Figura 6-Gráficos de análise de resíduos do modelo de regressão logística: Nitração; b) TAN; c) Ni; d) Horas; e)Valores ajustados	40
Figura 7 – Esquema de Prensa de Papel.....	46
Figura 8-Análise de resíduos modelo de regressão linear múltipla do rolo fixo LA, Prensa 1: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.....	49
Figura 9-Análise de resíduos modelo de regressão linear múltipla do rolo fixo LOA, Prensa 1: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.....	51
Figura 10-Análise de resíduos modelo de regressão linear múltipla do rolo móvel LA, Prensa 1: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.....	53
Figura 11-Análise de resíduos modelo de regressão linear múltipla do rolo móvel LOA, Prensa 1: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.....	55
Figura 12-Análise de resíduos modelo de regressão linear múltipla do rolo fixo LOA, Prensa 3: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.....	57

1. Introdução

Este relatório traduz o trabalho desenvolvido no âmbito do estágio curricular da Licenciatura em Matemática Aplicada à Tecnologia e à Empresa, no Instituto Superior de Engenharia de Lisboa (ISEL), que decorreu no Departamento de Assessoria Técnica e Qualidade de Produtos na GALP, em Lisboa, entre 2 de fevereiro e 19 de julho de 2019, no âmbito do projeto GALP MODe AIML.

A equipa de Assessoria Técnica e Qualidade de Produtos tem como missão dar suporte técnico a todas as áreas comerciais do sector da Refinação e Marketing da Galp. No portefólio de produtos comercializados, os lubrificantes são a tipologia de produto mais intensiva em suporte técnico, porque se trata de um componente rotável que incorpora equipamentos e não um consumível, como por exemplo os combustíveis. Um rotável é um elemento da estrutura do equipamento que é dimensionado para uma tarefa de sacrifício para preservar os restantes elementos definitivos. O lubrificante faz essa tarefa para proteger o motor de um veículo ligeiro, por exemplo, como os pneus o fazem para proteger as engrenagens de transmissão.

Este trabalho incide exclusivamente sobre a análise dos lubrificantes.

O projecto GALP MODe AIML teve a sua primeira divulgação científica no congresso *Lubmat* 2018, onde foi apresentada a modelação estatística empírica de dados da análise de lubrificantes em serviço. Este estágio curricular incide precisamente na sequência desse estudo. [1]

1.1. Lubrificantes e Lubrificação

Os lubrificantes são substâncias (hidrocarbonetos), formulados com óleo base e aditivos, colocados entre duas superfícies em movimento, formando uma película protetora, que tem como principal função reduzir o atrito e o desgaste. A lubrificação é fundamental na fiabilidade e manutenção de máquinas industriais e deve, portanto, ser entendida como um elemento que visa contribuir para aumentar a vida útil dos componentes. [2] [3]

De notar que o lubrificante começa a degradar-se logo a partir do momento que é adicionado ao equipamento, por ventura até antes se não for bem acondicionado. O serviço de acompanhamento de lubrificantes, através do envio de amostras para laboratório, é a forma de determinar se a sua condição está apta, ou não, para continuar em serviço, i.e. determinar a vida útil remanescente do lubrificante (RUL). O lubrificante deve, portanto, ser parcial ou totalmente renovado antes de se tornar prejudicial, ou seja, quando os resultados de um ou mais parâmetros de análise, presentes na rotina, violam os limites estabelecidos.

A degradação/envelhecimento das cargas de lubrificante é acompanhada pela análise suportada em três grandes vertentes:

- Estado Físico-Químico – avalia as propriedades do lubrificante face à sua formulação inicial, acompanha as reações de oxidação que degradam o óleo base e consumo o *package* de aditivos, com o objetivo de definir a vida útil remanescente (RUL);
- Contaminação – substâncias provenientes do ambiente exterior ao equipamento que se solubilizam ou são transportadas pelo lubrificante, por exemplo poeiras/sujidade;
- Desgaste – partículas metálicas que são libertadas pelo equipamento geradas pela falta de proteção/escassez de película de lubrificante, que acabam por ser arrastadas pelo mesmo. [4]

Um acompanhamento de lubrificante em serviço começa com um plano de amostragem, onde cada amostra é submetida a um programa de análise que consiste numa série de testes de acordo com os métodos padrão. Cada programa pode facilmente fornecer mais de vinte resultados numéricos diferentes. Esses resultados são comparados com os alertas estabelecidos, pelo fabricante e/ou entidades especializadas em lubrificação, e resumidos num diagnóstico qualitativo.

Existem uma infinidade de possíveis diagnósticos diferentes, contudo estes podem ser resumidos em três categorias:

- Normal – quando todas as variáveis controladas apresentam resultados dentro dos parâmetros considerados normais;
- Aviso – Quando uma ou mais variáveis apresentam variação significativa num sentido não desejável, contudo sem ultrapassar os limites estabelecidos, representando algum risco para o funcionamento do equipamento a médio prazo;
- Perigo – quando pelo menos uma das variáveis controladas viola o limite, colocando em risco o funcionamento do equipamento no imediato.

A Figura 1 apresenta um exemplo de um boletim de resultados do acompanhamento de lubrificantes em serviço.

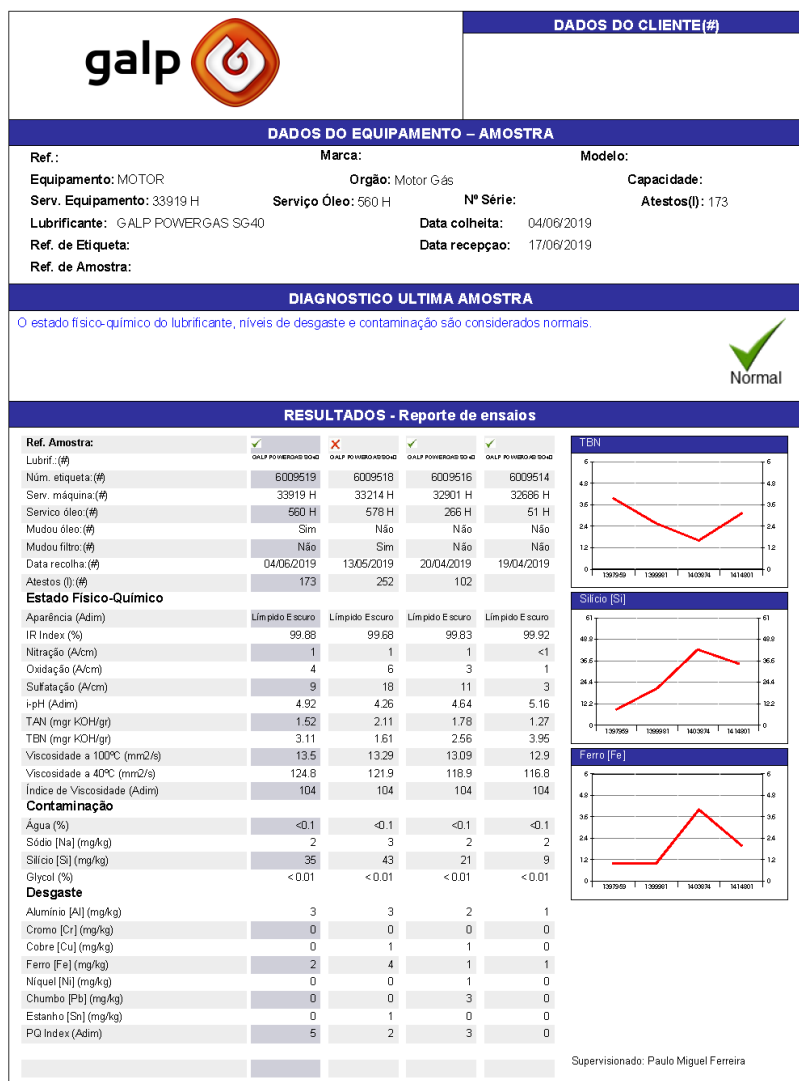


Figura 1-Boletim de análise de amostras de lubrificante.

Os resultados e respetivo diagnóstico que os enquadra são arquivados num registo histórico do equipamento. Com o acumular destes dados, também as evoluções registadas podem ter impacto no diagnóstico que é emitido. Por exemplo, uma série de amostras determinam que um contaminante específico nunca é detetado, exceto na amostra mais recente. Tal pode suscitar alerta, mesmo que esta única ocorrência não ultrapasse o limite estipulado pelo fabricante, porque teme-se que esta diferença na condição normal, que se registou ao longo do tempo, seja precursora de uma evolução crescente desse mesmo contaminante. Isto é por definição, aprendizagem do equipamento.

Principalmente quando nos situamos num diagnóstico 'Vigiar', os recetores destes, tipicamente equipas de manutenção, questionam o tempo que se estima até à falha que o diagnóstico aponta como possível. Frequentemente não é possível indicar essa estimativa, sem recorrer a uma maior frequência e/ou ampliação das análises ao lubrificante em serviço.

O universo de dados, gravados em histórico, podem permitir uma projeção da evolução provável do estado do lubrificante, se o aliarmos à modelação estatística empírica. Assim será possível converter diagnósticos qualitativos numa variável de saída numérica, capaz de se traduzir em ações de manutenção pró-ativa.

1.2. Modelação Estatística Empírica

Ferreira et al (2018), utilizaram modelos de regressão linear multivariável. Para essa modelação foram criadas variáveis dependente (LSC-*Lubrificant State Condition* e ESC-*Equipmant State Condition*) que é determinada em função de todas as outras variáveis, provenientes das análises dos lubrificantes em serviço, dando uma classificação sobre o estado do lubrificante e do equipamento. [1]

À imagem daquilo que se faz em exames de ensino, cada variável é considerada como uma questão com determinada ponderação. Consoante o resultado, essa ponderação é assimilada como pontuação, totalmente se satisfatória, parcialmente se tolerável ou nula se violar os limites estabelecidos. [1]

$$(LSC)_i = vw_i \times ff_i \quad (1)$$

A observação LSC é a soma de todas as contribuições de variáveis:

$$LSC = \sum_{i=1}^{n^{\circ} \text{ variáveis}} (LSC)_i \quad (2)$$

onde:

- vw_i é um determinado peso variável (*vw-variable weight*), atribuído a cada variável, i ;
- ff_i traduz a conformidade da especificação relativa à variável, i , atendendo aos lubrificantes em serviço. O ff_i (*fulfil factor*) tem cinco níveis de classificação:

$$ff_i = \begin{cases} 1 & \text{se satisfaz a condição } \checkmark\checkmark\checkmark \\ 0.75 & \text{se satisfaz a condição } \checkmark\checkmark \\ 0.5 & \text{se satisfaz a condição } \checkmark \\ 0.25 & \text{se satisfaz a condição } \Delta \\ 0 & \text{se satisfaz a condição } \times \end{cases}$$

Pode-se afirmar que a variável de resposta utilizada não é apropriada, pois em vez de ter sido observada, foi calculada como uma combinação linear das variáveis independentes. Por essa razão, durante o estágio foi utilizado como variável de resposta as horas de serviço de lubrificante (Capítulo 2.2.1) e em

alguns casos o atesto acumulado (Capítulo 2.2.2), ou seja, os litros/gramas de lubrificante adicionado ao equipamento.

Antes de modelar, é importante realizar uma análise exploratória dos dados, de modo a saber que tipo de dados estamos a utilizar e de que forma cada uma das potenciais variáveis preditoras se relaciona com a variável resposta.

O rácio entre observações e variáveis é um indicador de maturação do modelo, sendo que:

- Se for inferior a 1, o modelo não é válido;
- Se for superior a 1 e inferior a 10, é um modelo em maturação que deve ser atualizado a cada nova observação;
- Se for superior a 10, é um modelo maduro.

Para realizar essa análise estudaram-se:

- Medidas Descritivas

As medidas descritivas são utilizadas para descrever, organizar e resumir as características principais dos dados.

Algumas das medidas que são normalmente usadas para descrever um conjunto de dados, são as medidas de tendência central e as medidas de variabilidade ou dispersão. Medidas de tendência central incluem média e mediana. Medidas de variabilidade incluem o desvio padrão, a variância, o valor máximo e o mínimo. [5]

O estudo das medidas descritivas fornece resumos simples sobre a amostra e sobre as observações que foram registadas. Este resumo pode ser quantitativo ou visual, através de, por exemplo, *boxplots* (ou vulgarmente também chamados de Caixas de Bigodes). (Vd. Anexo A)

- Teste de Normalidade

Os testes de normalidade são utilizados para verificar se a distribuição de probabilidade associada a um conjunto de dados de uma variável aleatória pode ser aproximada pela distribuição normal. Os testes mais utilizados são do Kolmogorov-Smirnov e o de Shapiro-Wilk, que é o utilizado neste trabalho. [5]

Para além desta utilização, estes testes também são uma forma de verificar a qualidade dos modelos, fazendo um teste à normalidade dos resíduos.

- Matriz de Correlações

A matriz de correlações é composta pelas correlações entre as diversas variáveis e é utilizada para iniciar a análise estatística dos dados históricos dos diversos universos estudados, identificando visualmente as variáveis envolvidas no estudo que se relacionam entre si.

Informalmente, correlação é sinónimo de dependência. Em termos técnicos, correlação refere-se a qualquer um dos vários tipos específicos de relação entre os valores médios.

O coeficiente de correlação linear, r^2 , é um indicador de existência de associação linear, varia entre 0 e 1, i.e. quando não há correlação é 0 e quando a correlação é perfeita é 1.

Existem diferentes coeficientes de correlação para medir o grau de correlação e que devem ser utilizados de acordo com a tipologia das variáveis que se pretendem analisar. Um dos coeficientes de correlação mais conhecido, e utilizado neste trabalho, é o coeficiente de correlação de *Pearson*, obtido pela divisão da covariância de duas variáveis pelo produto dos seus desvios padrão. Entretanto, há outros coeficientes de correlação mais robustos, como o coeficiente de correlação de *Spearman*, isto é, mais sensíveis às relações não lineares. [5] [6]

- Análise de Componentes Principais (ACP):

Muitas vezes, os conjuntos de dados que se pretendem analisar são caracterizados por inúmeros atributos. A existência de um elevado número de dimensões associada a estes conjuntos de dados faz com que visualizá-los graficamente seja impossível. Neste cenário, o ideal seria poder reduzir o número de dimensões dos dados sem que haja uma perda significativa de informação. Com essa redução os dados podem ser visualizados graficamente, facilitando, assim, a sua compreensão e exploração, de modo a que seja mais fácil encontrar padrões nos dados em análise.

O ACP é um procedimento matemático que utiliza uma transformação ortogonal para converter um conjunto de observações de variáveis possivelmente correlacionadas num conjunto de valores de variáveis linearmente não correlacionadas chamadas de componentes principais. O número de componentes principais é menor ou igual ao número de variáveis originais. Esta transformação é definida de forma a que a primeira componente principal tenha a maior variância possível (ou seja, é responsável pela explicação máxima de variabilidade nos dados), e cada componente seguinte, por sua vez, tenha a máxima variância sob a restrição de ser ortogonal às componentes anteriores.

As componentes principais são garantidamente independentes quando os dados forem normalmente distribuídos. [7]

Assim sendo, com o objetivo de se selecionar o modelo mais indicado para este tipo de dados, foram analisados três tipos de modelos:

- Modelo de Regressão Linear Múltipla (Vd. Capítulo 3.2)

Em diversos problemas em áreas científicas é de grande interesse verificar se duas ou mais variáveis estão relacionadas de alguma forma. Para expressar esta relação é muito importante estabelecer um modelo matemático. Este tipo de modelação é chamado de regressão, e ajuda a entender como determinadas variáveis influenciam outra variável, ou seja, verifica como o comportamento de uma variável pode mudar o comportamento de outra(s).

A análise de regressão possibilita encontrar uma relação razoável entre a variável dependente e as variáveis independentes, por meio de relações empíricas, i.e. porque não existe ainda uma solução determinada que os possa selecionar.

Para a construção de um modelo, é necessário realizar uma seleção de variáveis, estimar os parâmetros e concluir acerca do ajuste do modelo, através de, por exemplo, testes de hipóteses e intervalos de confiança.

Os principais problemas que podem ocorrer ao realizar o ajustamento de um modelo de regressão são:

- Atipicidade – deve-se ao facto de uma variável independente não apresentar nenhum padrão nos seus valores;
- Multicolinearidade – refere-se a uma correlação inaceitavelmente alta entre preditores (dependência entre variáveis independentes). Quando a multicolinearidade aumenta, os coeficientes mantêm-se centrados mas os erros padrões aumentam e a verosimilhança do modelo diminui.
- Heterocedasticidade – apresenta-se como uma forte dispersão dos resíduos em função de uma variável. O seu oposto é a homocedasticidade, que implica que a variância da condicional das variáveis independentes com o erro seja constante;
- Autocorrelação – é uma medida que informa o quanto uma variável aleatória é correlacionada com ela mesma, ou seja, mede para cada variável independente a correlação entre uma observação e a observação seguinte;
- Endogeneidade – ocorre quando um dos preditores do modelo é correlacionado com o erro.

Uma vez que esperamos que grande parte da variação da variável dependente seja explicada pelas variáveis independentes, podemos utilizar o modelo para obter valores da variável resposta correspondentes a valores dos preditores que não estavam entre os dados. Esse procedimento é chamado de predição e, em geral, usamos valores para as variáveis que estão dentro do intervalo de variação estudado. A utilização de valores fora desse intervalo recebe o nome de extrapolação e deve ser usada com muito cuidado, pois, o modelo adotado pode não ser correto fora do intervalo estudado. Acredita-se que a predição seja a aplicação mais comum dos modelos de regressão. [6] [8]

- Modelo de Regressão Logística (Vd. Capítulo 3.3)

A regressão logística é uma regressão múltipla, utilizada quando a variável resposta é categórica dicotômica e as variáveis independentes podem ser contínuas ou categóricas.

Na regressão logística as variáveis dependentes estão dispostas em categorias (sim/não; 0/1), enquanto na regressão linear estas variáveis são contínuas.

A resposta na regressão logística é expressa através de uma probabilidade de ocorrência, enquanto na regressão simples, obtém-se um valor numérico.

Quando se pretende explicar uma variável binária ou *dummy*, os modelos de escolha binária mais adequados são os modelos com função de ligação *logit* e *probit*. Neste caso, usamos a função de ligação *logit*. [9] [10]

- Análise de Sobrevivência – Modelo de Cox (Vd. Capítulo 3.4)

O tempo de sobrevivência entre um evento e a falha pode ser influenciado por diversos fatores. Nos indivíduos em estudo pretende-se determinar as características do óleo que têm capacidade explicativa do tempo até à necessidade de mudança do óleo.

Na análise de sobrevivência a variável dependente é sempre o tempo até ocorrência de determinado evento, ao contrário da análise estatística “clássica” a variável dependente é a própria ocorrência de determinado evento. [11]

Para a elaboração deste trabalho e no decorrer do estágio como ferramentas de programação foram:

- R Studio (versão 1.1.463)
- Matlab (R2019a)

1.3. Objetivos

Assim sendo, este trabalho tem o objetivo de gerar modelos estatísticos, aproveitando dados do acompanhamento de lubrificantes em serviço, com o intuito de projetar a condição de um lubrificante ou equipamento, de modo a evitar avarias.

Consequentemente, tem-se a seguinte matriz de dados, em que cada amostra é uma observação, cada coluna é uma variável $[x_1, \dots, x_m]$, sendo a última coluna a variável resposta, y :

$$\begin{matrix} & x_1 & x_2 & x_3 & \cdots & x_m & y \\ \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,m} & y_1 \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,m} & y_2 \\ x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,m} & y_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,m} & y_n \end{bmatrix} \end{matrix} \quad (3)$$

Este documento está organizado em quatro capítulos, sendo o Capítulo 1, a Introdução. O Capítulo 2, a Apresentação de dados, onde são explicados os diferentes tipos de variáveis e a sua importância na análise de lubrificantes em serviço.

O Capítulo 3, Caso de Estudo Preliminar, é destinado à preparação para o projeto principal do estágio, por essa razão, foi proposto o estudo de um motor de uma empresa de Tratamento e Valorização de Resíduos Sólidos Urbanos. Esta empresa é uma referência no setor Ambiental e de Tratamento e Valorização de Resíduos em Portugal e é responsável por assegurar o Tratamento e Valorização de Resíduos, de forma sustentável.

Esta preparação tem como objetivo determinar o procedimento indicado para tratar estes tipos de dados, e assim escolher o melhor modelo que preveja o diagnóstico de um lubrificante em serviço.

Foi feita uma análise exploratória aos dados do mesmo, usando medidas básicas da estatística. De facto, antes de formular o modelo é necessário estudar as correlações entre variáveis, características amostrais de cada uma (por exemplo média, mediana, variância, dispersão, assimetria, entre outras métricas) e como impactam a variável de saída. Entre o universo de variáveis, torna-se necessário identificar aquelas que recolhem melhores condições de serem promovidas a preditores para integrar o modelo.

De seguida, foram estudados três possíveis modelos, regressão linear múltipla, regressão logística e modelo de Cox (Análise de Sobrevida).

No Capítulo 4, Modelação de um Caso de Estudo é analisada a base de dados das prensas de papel do caso de estudo do projeto, através das conclusões retiradas do capítulo anterior. Por fim, são realizados modelos para os mesmos, retirando conclusões acerca da qualidade de cada um.

O último capítulo, Conclusão, é dedicado à apresentação das ideias finais e sugestões de trabalho futuro deste estágio.

A Bibliografia apresenta as obras citadas neste trabalho, por ordem de surgimento.

Por fim, o Anexo A contem os *boxplots* dos dados utilizados no caso de estudo preliminar, o Anexo B apresenta as correlações entre variáveis desses mesmos dados e o Anexo C os gráficos de dispersão correspondentes às correlações apresentadas no anexo anterior.

2. Apresentação de dados

Por forma a conhecer os dados dividiram-se os mesmos em dois grupos: Variáveis de Entrada (2.1) e Variáveis de Saída (2.2).

As Variáveis de Entrada são resultados numéricos de análises laboratoriais de lubrificantes em serviço.

As Variáveis de Saída traduzem um diagnóstico qualitativo, da condição do lubrificante à data da recolha de amostra, com correlação muito forte com o estado do equipamento.

Nota: Todas as variáveis em estudo são contínuas, apesar de na base de dados algumas serem apresentadas arredondadas às unidades, por serem fornecidas pelo laboratório dessa forma.

2.1. Variáveis de Entrada

Considera-se como variáveis independentes/explicativas aquelas que estão registadas, de forma numérica, nos dados de análise dos lubrificantes em serviço do equipamento. Neste universo pretende-se selecionar um conjunto reduzido de variáveis preditoras com bom ajustamento à variável de saída, que por sua vez seja capaz de traduzir em ações de manutenção, que preservam o estado do equipamento.

Para facilitar a análise e seleção de variáveis, estas foram divididas em três grupos: Estado Físico-químico (2.1.1), Desgaste (2.1.2), Contaminação (2.1.3) e ainda Variáveis Externas (2.1.4) que não provém da análise do lubrificante em serviço.

2.1.1. Estado Físico-Químico

As variáveis que pertencem a este agrupamento inicial são as que representam a composição físico-química do lubrificante, e que variam ao longo do envelhecimento do mesmo.

Uma correta interpretação destes resultados direcionará à mudança do lubrificante, que deixam de cumprir as suas funções.

Tabela 1-Variáveis do agrupamento Estado Físico-Químico presente no universo de dados.

Identificação	Unidade	Descrições
IR Index (IRidx) - Percentagem de transmissão de radiação de infravermelhos face a óleo novo.	%	A variável examina o envelhecimento do lubrificante e identifica eventual mistura com outros tipos de óleo. A variável traça a correspondência entre o espectro infravermelho do lubrificante novo com o espectro de amostra em serviço. [12]
Nitração (Nitr) - reação do lubrificante com Azoto, N.	Abs/cm	A nitração é uma reação química que degrada moléculas do óleo base do lubrificante, que também consome Azoto (N) presente no ar. Esta variável só é acompanhada em óleos de motor, porque necessita de elevada energia de ativação, concedida pela combustão.
Oxidação (Oxid) - reação do lubrificante com Oxigénio, O.	Abs/cm	A oxidação é o principal mecanismo de degradação do lubrificante, e o óleo oxidado não pode efetivamente lubrificar as máquinas. O aumento da viscosidade e de compostos insolúveis, o entupimento do filtro, a acumulação de depósitos e sedimentos no motor, o desgaste corrosivo e a redução da vida útil do óleo são as causas mais frequentes provocadas pela oxidação. [2]
pH	Adim	Como é do senso comum, o valor de pH demonstra se uma substância tem caracter neutro (=7), base (>7) ou ácido (<7). A degradação do lubrificante torna-o cada vez mais ácido. Este método, especialmente útil para óleos de motores estacionários que consomam biogás, identifica quando o lubrificante é demasiado agressivo para as superfícies metálicas que se desejam proteger.
Sulfatação (Sulf) - reações do lubrificante com Enxofre, S.	Abs/cm	O Enxofre (S) é um contaminante presente no crude que é transmitido para a formulação de muitos lubrificantes, sendo que potencia reação de sulfatação que também formam compostos ácidos. Esta variável também é exclusivo de óleos de motor

Tabela 1-Variáveis do agrupamento *Estado Físico-Químico* presente no universo de dados.(continuação)

Identificação	Unidade	Descrições
TAN, Total Acid Number (Número Ácido)	mg KOH/g	Determina o número de acidez, um indicador dos constituintes ácidos de um lubrificante. O método consiste em medir a massa de hidróxido de potássio necessária, em mgKOH, necessária para neutralizar 1g de lubrificante. O valor do número de acidez é um indicador da oxidação do lubrificante porque o envelhecimento gera compostos ácidos. Ou seja, é seguro que o valor de TAN só poderá aumentar ao longo da vida útil do lubrificante. [3] Nota: No acompanhamento em serviço de lubrificantes, esta variável é muito valorizada.
Teor de Fósforo (P)	ppm	Indica a presença de aditivos antioxidantes, inibidores de corrosão, anti-desgaste e de extrema-pressão.
Teor de Enxofre (S)	ppm	Este composto pertence à composição físico-química do lubrificante, por ser um contaminante típico no crude.
Viscosidade a 40°C (Visc40) e Viscosidade a 100°C (Visc100)	cSt	A viscosidade é a propriedade dos lubrificantes que corresponde à quantidade de movimento por difusão molecular. Ou seja, quanto maior a viscosidade, menor será a velocidade em que o fluido se movimenta numa determinada área. A viscosidade do lubrificante é responsável por separar fisicamente as superfícies metálicas em movimento, num regime hidrodinâmico, e afeta o calor gerado nos rolamentos das chumaceiras das turbinas. Em termos de lubrificação é considerada a mãe de todas as grandezas. [13]

Onde:

% - percentagem.

Adim – número adimensional.

Abs/cm – absorção por centímetro do comprimento de onda.

mg KOH/g - massa de hidróxido de potássio (KOH) em miligramas que é necessária para neutralizar um grama de substância química.

ppm – partes por milhão (quanto de soluto em mg há em 1kg de solução).

cSt – centistokes. unidade física da viscosidade cinemática.

2.1.2. Desgaste

As variáveis que compõem este grupo são os elementos que apresentam a condição das peças do equipamento, ou seja, são capazes de indicar desgastes prematuros em peças como rolamentos, bombas hidráulicas, anéis de segmentos e outros.

A combinação das altas concentrações destes elementos é capaz de identificar desgastes prematuros específicos de um compartimento. E uma correta interpretação destes resultados direcionará à intervenção do equipamento para a causa da anormalidade.

Neste agrupamento de desgaste, todos os teores metálicos são determinados através do mesmo método de análise espectrométrica de emissão atômica, capaz de identificar a concentração de elementos químicos metálicos, onde se destacam o Teor de Ferro, o Teor de Cobre e o Teor de Alumínio, bem como outros elementos que têm origem no *package* de aditivos como: o Teor de Fósforo, o Teor de Zinco e o Teor de Cálcio, entre outros.

Tabela 2- Variáveis do agrupamento Desgaste presente no universo de dados.

Identificação	Unidade	Descrição
PQ Index (PQidx) - Teor de partículas ferromagnéticas, superiores a 14 microns.	Adim	O índice PQ quantifica a presença de partículas ferromagnéticas de dimensão significativa. Apesar de ser capaz de detetar partículas a partir dos 6 µm de diâmetro, considera-se que capta a totalidade dos que são iguais e superiores a 14µm.
Teor de Alumínio (Al)₁	ppm	Indica desgaste nos pistões, rolamentos, bombas, rotores, poeiras.
Teor de Boro (B)	ppm	Furos no sistema de resfriamento (inibidor refrigerante), agente de limpeza.
Teor de Bário (Ba)	ppm	Aditivo do óleo do motor, aditivo do combustível, graxa.
Teor de Crómio (Cr)	ppm	Pode indicar desgaste nos anéis de revestimento, aço inoxidável, rolamentos, cubos de freio, cilindros e partes de sistemas hidráulicos.
Teor de Estanho (Sn)	ppm	Pode indicar desgaste nos rolamentos.

Tabela 2- Variáveis do agrupamento Desgaste presente no universo de dados. (continuação)

Identificação	Unidade	Descrição
Teor de Cobre (Cu)¹	ppm	O metal cobre não é adicionado ao óleo lubrificante como aditivo, e a presença deste metal em amostras de óleos lubrificantes usados geralmente ocorre devido ao desgaste de guias de válvula, anéis de pistão, pontos de apoio, etc. Sendo assim, ao longo do tempo, observa-se um aumento na concentração deste metal. Os teores de Cu podem diferenciar tendo em conta o ano de fabrico, assim como o modelo do equipamento.
Teor de Ferro (Fe)¹	ppm	O Ferro (Fe) é o elemento químico que minoritariamente constitui as superfícies metálicas que o lubrificante protege. Uma concentração elevada deste elemento significa falta de proteção e é um precursor de ocorrência de falha mecânica.
Teor de Níquel (Ni)	ppm	Apresentam-se abaixo do limite de deteção do método. O aumento da presença de níquel em pequenas quantidades pode ter origem no desgaste de pontos de apoio, válvulas, engrenagens (revestida com esse metal), etc.
Teor de Molibdênio (Mo)	ppm	Indica a presença de aditivos modificadores de atrito, que facilita o deslizamento entre as peças e desgaste de ligas metálicas de aço.
Teor de Chumbo (Pb)	ppm	Elemento químico associado ao desgaste do equipamento. Também existem compostos orgânicos com chumbo que desempenham o papel de aditivos, contudo estão obsoletos por razões ambientais.

Onde:

Adim – número adimensional.

ppm – partes por milhão (quanto de soluto em mg há em 1kg de solução).

¹ No caso de estudo preliminar, também foram consideradas variáveis de rácios de teores em metais para cada 100 horas de serviço do lubrificante, em ppm/100.

2.1.3. Contaminação

Estas variáveis identificam contaminações externas ou internas que podem causar grandes alterações na composição do lubrificante e prejudicar o seu desempenho.

Tabela 3- Variáveis do agrupamento Contaminação presente no universo de dados.

Identificação	Unidade	Descrição
Porcentagem de água no lubrificante	%	A água é o principal contaminante a poder surgir dissolvido em lubrificantes. É importante o lubrificante permanecer isento de água, porque esta é precursora de oxidação para o óleo e de corrosão para o equipamento. [14]
Teor de Sódio (Na)	ppm	Pode indicar furos no sistema de refrigeração. Contaminação externa (altas concentrações de Sódio).
Teor de Silício (Si)	ppm	O Silício que se deteta em lubrificante pode ter várias origens diferentes. Entre elas: poeiras por contaminações com o ambiente (sem filtração), detritos de borracha, por degradação de vedantes e contaminação de combustível, no caso de queima de biogás, entre outros.

Onde:

% - percentagem.

ppm – partes por milhão (quanto de soluto em mg há em 1kg de solução).

2.1.4. Variáveis Externas ao Lubrificante

Como referido no Capítulo 1 - Introdução, este trabalho de modelação estatística, tem como base os dados as análises laboratoriais que se fazem ao lubrificante, contudo não se limita só a esse tipo de dados.

Nas variáveis de entrada podem ser somadas variáveis que se acreditam ter impacto indireto na lubrificação do equipamento.

Por exemplo, é sabido que a temperatura tem impacto na vida útil do lubrificante, esta pode ser adicionada ao universo de dados e, eventualmente, selecionada como preditor permitindo gerar cenários de impacto de temperatura e na durabilidade da vida útil do lubrificante.

Estas variáveis foram inseridas na base de dados para o caso de estudo de uma prensa de papel (Capítulo 4), este equipamento tem como objetivo espremer água e fibras menos resistentes do fluxo que forma o produto final.

Sendo classificadas como fatores externos ao lubrificante, ou seja, não se encontram no boletim de análise dos lubrificantes em serviço, estas variáveis apresentam grande importância no diagnóstico do lubrificante em serviço e do equipamento.

Tabela 4- Variáveis Externas ao Lubrificante presente no universo de dados.

Identificação	Unidade	Descrição
Binário	MPa	O binário reflete a força exercida pelos rolamentos da prensa sobre a pasta de papel com o objetivo de expelir água e fibras de baixa resistência mecânica na pasta.
Velocidade de Rotação	rpm	Representa a velocidade a que a pasta passa pela prensa.

Onde:

MPa – megapascal é a unidade de pressão e tensão.

rpm - rotações por minuto, é uma unidade de velocidade angular.

Estas duas variáveis são registadas de 5 em 5 minutos, assim sendo, o número de observações das mesma é muito superior ao número de amostras do lubrificante em serviço.

Nos casos de estudo presentes neste documento, as recolhas de observações têm uma frequência média de uma semana.

Por essa razão, e para poder incluir estas variáveis ao modelo, determinou-se a média e o desvio padrão, para cada uma, referentes ao intervalo de tempo decorrido entre a amostra mais recente e a sua antecessora.

2.2. Variáveis de saída

A variável de saída, de resposta ou dependente para os modelos deste trabalho, tem de ser capaz de se relacionar com operações de manutenção, como já referido no Capítulo 1 - Introdução.

No que toca à lubrificação, os responsáveis de manutenção desejam saber se necessitam de substituir ou remover o lubrificante nos seus equipamentos.

Como tal, as variáveis de saída presentes neste trabalho, das duas uma, ou definem o intervalo de tempo ótimo que o lubrificante deve servir o equipamento antes de ser totalmente substituído, ou define qual a quantidade acumulada de lubrificante que deve ser introduzido.

Cada um dos tipos de variáveis de resposta adequa-se à maneira como o lubrificante é substituído ou renovado.

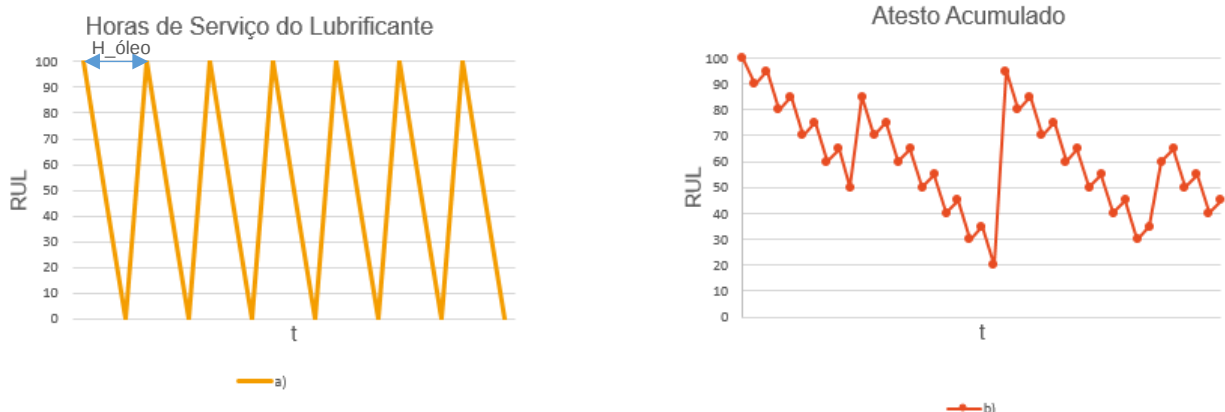


Figura 2- Representação da vida útil do lubrificante. a) quando ciclicamente é totalmente renovado; b) quando é gradualmente renovado.

2.2.1. Horas de funcionamento do lubrificante

Como variável dependente foi definido o tempo ótimo do lubrificante em serviço, i.e para cada amostra retirada do equipamento em funcionamento, também é registado o número de horas em que a carga de lubrificante já está em atividade.

Através desses registos é possível estudar as correlações entre diversos os resultados laboratoriais e o número de horas de funcionamento do lubrificante.

Assim, depois de selecionados os preditores que melhor explicam esta variável resposta, é possível prever, através do modelo estimado, o número de horas que restam ao óleo até à próxima mudança.

2.2.2. Atesto Acumulado

Para tipos de dados diferentes, foi definido como variável dependente o valor acumulado de lubrificante adicionado.

Esta variável é uma simulação do acumulado de lubrificante adicionado ao sistema, quando este é renovado quase exclusivamente por pequenos atestos periódicos que preserva o estado do lubrificante e o estado do equipamento dentro dos limites normais de funcionamento, informação registada em massa ou em volume (gramas/litros).

3. Caso de Estudo Preliminar

Inicialmente, foram analisados os dados provenientes de um motor estacionário a biogás (motores de combustão interna cujo conjunto não se desloca).

Biogás é o nome comum dado a uma mistura de gases que foi produzida pela decomposição biológica da matéria orgânica na ausência de oxigênio. Normalmente consiste em uma mistura gasosa composta principalmente de gás metano (CH_4) e dióxido de carbono (CO_2), com pequenas quantidades de gás sulfídrico (H_2S) e humidade.

Para este tipo de motor a biogás é sugerido a utilização do lubrificante *GALP POWER GAS NGB 40*, indicado para motores estacionários, que operem a gás natural ou biogás. Formulado com aditivos de com baixo teor de cinzas e bases hidrocracadas que garantem elevados períodos de vida e asseguram a limpeza e a proteção do motor. [15]

Considerando como variáveis explicativas aquelas que estão registadas nos dados das análises dos lubrificantes em serviço do equipamento, pretende-se encontrar um conjunto reduzido de variáveis que auxiliem a prever a necessidade de mudança do lubrificante. (Vd. *Tabela 1, Tabela 2, Tabela 3 e Tabela 4*)

Neste caso de estudo têm-se 42 observações, 28 variáveis explicativas que se dividem pelos três tipos. (Vd. Capítulo 2.1, exceto Capítulo 2.1.4) Serão consideradas duas variáveis resposta para este caso: o número de horas de serviço do lubrificante e a variável binária que indica se o lubrificante foi ou não substituído.

Por forma a caracterizar os dados realizou-se uma análise exploratória, que incluiu as estatísticas descritivas básicas (Capítulo 3.1.1), testes de normalidade (Capítulo 3.1.2), os coeficientes de correlação (Capítulo 3.1.3) e ainda análise em componentes principais (Capítulo 3.1.4).

Depois da análise exploratória de dados, tentaram-se ajustar três tipos de modelos: Modelo de Regressão Linear (Capítulo 3.2), Modelo de Regressão Logística (Capítulo 3.3) e Modelo de Cox (Capítulo 3.4). A seleção dos preditores é baseada nas análises realizadas e o tipo de modelo. Por fim, fez-se a análise dos resíduos para cada um dos modelos, concluindo-se qual o mais adequado (Capítulo 3.2.4, para a regressão linear múltipla e o Capítulo 3.3.4, para a regressão logística).

3.1. Análise Exploratória dos dados

Como já foi dito anteriormente, antes de modelar o conjunto de dados realizou-se uma análise detalhada de cada uma das variáveis.

Para além das medidas descritivas para as variáveis e do valor-p associado ao teste de normalidade que constam na *Tabela 5*, obtiveram-se os gráficos *boxplot* para todas as variáveis, cujas figuras se encontram no Anexo A.

3.1.1. Medidas descritivas

Considerando n como o número de observações de uma amostra, para realizar esta análise foram utilizadas as seguintes medidas:

- Média:
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (4)$$

- Mediana:
$$\begin{cases} Md = x_{\frac{n+1}{2}}, \text{ se } n \text{ for ímpar} \\ Md = \frac{x_{\frac{n+2}{2}} + x_{\frac{n}{2}}}{2}, \text{ se } n \text{ for par} \end{cases} \quad (5)$$

- Desvio Padrão:
$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (6)$$

- 1ºQuartil:
$$Q_1 = x_{\frac{n+3}{4}} \quad (7)$$

- 3ºQuartil:
$$Q_3 = x_{\frac{3n+1}{4}} \quad (8)$$

- Intervalo Interquartil:
$$IQR = Q_3 - Q_1 \quad (9)$$

- Coeficiente de Assimetria:
$$CA = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3} \quad (10)$$

- Coeficiente de Achatamento:

$$CK = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} \quad (11)$$

Tabela 5 - Medidas Descritivas das Variáveis

Variável	$\bar{x}$	Md	s	Q_1	Q_3	IQR	CA	CK	Valor-p teste de normalidade
IRídx	99.827	99.835	0.075	99.790	99.880	0.090	-0.832	0.879	0.060
Nitr	2.262	2.000	1.149	2.000	3.000	1.000	-0.507	-0.426	0.001 ²
Oxid	6.262	6.000	2.400	5.000	8.000	3.000	-0.095	-1.013	0.106
pH	4.974	4.780	0.833	4.303	5.638	1.335	0.181	-1.172	0.121
PQidx	2.048	2.000	1.268	1.250	3.000	1.750	-0.297	-0.918	0.001
Sulf	3.119	3.000	2.559	1.000	4.000	3.000	0.672	-0.188	0.007 ¹
TAN	1.227	1.290	0.474	0.890	1.558	0.668	-0.131	-0.950	0.499
TBN	4.235	4.190	0.516	3.810	4.620	0.810	-0.167	-0.737	0.456
Al	7.238	7.000	3.161	5.000	10.000	5.000	0.379	-0.894	0.032 ¹
B	0.762	0	1.394	0	0.750	0.750	1.473	0.777	<0.001 ¹
Ca	1589.071	1586.000	125.743	1501.250	1640.500	139.250	1.242	2.242	0.003 ¹
Cr	1.167	1.000	0.762	1.000	2.000	1.000	0.048	-0.717	<0.001 ¹
Cu	0.833	1.000	0.581	0.250	1.000	0.750	0.013	-0.336	<0.001 ¹
Fe	7.095	6.500	3.931	4.000	9.750	4.250	0.424	-0.732	0.086
Mo	0.286	0	0.457	0	1.000	1.000	0.915	-1.189	<0.001 ¹
Na	2.929	3.000	2.005	1.250	3.750	1.500	1.532	2.495	<0.001 ¹
Ni	0.310	0	0.468	0	1.000	1.000	0.795	-1.400	<0.001 ¹
P	267.929	267.000	13.560	258.250	274.000	15.750	1.286	2.900	0.003 ¹
Pb	0.452	0	1.656	0	0	0	4.157	16.916	<0.001 ¹
S	3555.881	3487.000	291.828	3354.250	3734.250	380.000	0.664	-0.478	0.021 ¹

² Valores p inferiores ao nível de significância de 5%. Nestes casos não se rejeita a hipótese nula do teste de normalidade.

Tabela 5 - Medidas Descritivas das Variáveis(continuação)

Variável	$\bar{x}$	Md	s	Q_1	Q_3	IQR	CA	CK	Valor-p teste de normalidade
Si	142.095	140.000	68.311	83.250	186.500	103.250	0.222	-1.054	0.113
Sn	1.214	1.000	0.645	1.000	2.000	1.000	-0.212	-0.768	<0.001
Zn	354.429	354.500	22.636	336.000	372.500	36.500	0.214	-0.710	0.599
Visc100	13.433	13.470	0.225	13.288	13.570	0.292	-0.240	-0.763	0.316
Visc40	121.217	121.050	2.575	119.150	122.750	3.600	0.224	-0.844	0.551
Hora	704.167	663.000	407.628	381.750	1009.250	607.500	0.209	-1.064	0.164
Mudança de óleo	0.190	0	0.397	0	0	0	1.521	0.321	<0.001 ¹

¹ Valores p inferiores ao nível de significância de 5%. Nesses casos não se rejeita a hipótese nula do teste de normalidade.

A título de exemplo, observando a *Tabela 5* pode dizer-se que a variável Sulfatação, varia entre 0 A/cm e 10 A/cm, tem uma média de aproximadamente 3.1 A/cm e apresentando estes dados uma forte variabilidade, calculada através da divisão do desvio padrão com a média (CV=82.1%).

Cerca de metade das observações da Sulfatação, são inferiores ou iguais a 3 A/cm.

O *boxplot* fornece uma análise visual da localização, dispersão, assimetria, caudas da dispersão e valores discrepantes (*outliers*) do conjunto de dados.

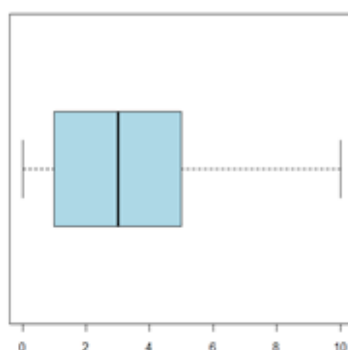


Figura 3-Boxplot dos dados da variável Sulfatação

Relativamente ao *boxplot* da Sulfatação, quanto à assimetria observa-se que a mediana se encontra muito próxima do centro da caixa, logo indica simetria, mas no entanto a cauda da direita é maior que a esquerda o que leva a uma ligeira assimetria positiva.

A dispersão dos dados pode ser avaliada pelo intervalo interquartil, que é a diferença entre o terceiro quartil e o primeiro quartil (corresponde ao comprimento da caixa), que neste caso é igual a 3.

Por fim, para estes dados não temos nenhum *outlier*.

3.1.2. Teste de Normalidade

Como se pode ver na *Tabela 5*, realizou-se o teste de hipóteses de *Shapiro-Wilk*, teste de normalidade, usando a estatística de teste *W*, para cada uma das variáveis. [16]

Neste teste tem-se como hipóteses:

H_0 : A amostra provém de uma população com distribuição normal (12)

H_1 : A amostra não provém de uma população com distribuição normal

Decisão: Se o *valor-p* for maior do que o nível de significância de 5%, não rejeitamos a hipótese nula (H_0).

Conclusão: Nesse caso, a amostra provém de uma população que tem distribuição normal.

No caso da variável Sulfatação, por exemplo, como o $\text{valor-p} = 0.007 < 0.05$, rejeita-se a hipótese H_0 , pelo que os dados relativos a essa variável não parecem ser provenientes de uma população com distribuição normal.

3.1.3. Correlações entre Variáveis

Com o intuito de averiguar a existência de possíveis associações entre as variáveis, foi calculada a matriz de correlações entre as variáveis, encontrando-se no Anexo B, bem como todos os diagramas de dispersão entre todos os pares de variáveis e os respectivos índices de correlação, o que poderá ser analisado no Anexo C. Para além do cálculo das correlações de *Pearson* calcularam-se também os coeficientes de correlação de *Spearman*, tendo-se verificado que os resultados produzidos foram semelhantes para ambos os coeficientes. [5]

De seguida, realizou-se o teste de hipóteses à significância do coeficiente de correlação, para avaliar quais das variáveis independentes são mais importantes para o modelo. [16]

Sendo assim, este teste tem como hipóteses:

$$H_0: p=0 \quad \text{vs} \quad H_1: p \neq 0 \quad (13)$$

Decisão: Se o valor-p for menor do que o nível de significância de 5%, rejeita-se a hipótese nula (H_0).

Conclusão: Nesse caso, se o coeficiente de correlação é significativamente diferente de zero, então existe associação linear entre as duas variáveis.

Tabela 6 - Exemplo de resultados dos testes à significância dos coeficientes de correlação

Variável	Valor-p	Variável	Valor-p
Oxid e pH	<0.001 [#]	pH e hora	<0.001 [#]
IRidx e pH	<0.001 [#]	Al e hora	<0.001 [#]
pH e Fe	<0.001 [#]	Fe e hora	<0.001 [#]
pH e TAN	<0.001 [#]	Si e hora	<0.001 [#]
TAN e Fe	<0.001 [#]	Zn e hora	0.008 [#]
TAN e Cu	<0.001 [#]	Mudança de óleo e pH	0.015 [#]
TAN e Si	<0.001 [#]	Mudança de óleo e S	0.189
Ca e Fe	0.741	Mudança de óleo e Si	0.002 [#]
Cu e Fe	<0.001 [#]	Mudança de óleo e Cu	0.003 [#]
Si e Fe	<0.001 [#]	Mudança de óleo e Fe	0.012 [#]
Visc40 e Si	<0.001 [#]		
Al e Fe	<0.001 [#]		
Si e Zn	0.005 [#]		
Si e Al	<0.001 [#]		
pH e Si	<0.001 [#]		
pH e Zn	0.031 [#]		
pH e Al	<0.001 [#]		
pH e Fe	<0.001 [#]		
pH e hora	<0.001 [#]		

A partir da Tabela 6, pode-se afirmar que as variáveis que têm uma correlação estatisticamente significativa com as variáveis resposta (horas de funcionamento do lubrificante e mudança do lubrificante) são o pH, o Teor de Alumínio, o Teor de Ferro e o Teor de Silício, pois no teste à significância dos coeficientes têm um valor-p inferior ao nível de significância de 5% (Capítulo 3.1.3). Estes resultados indicam que estas variáveis podem ser possíveis preditores para os modelos, pois apresentam uma associação linear forte com o número de horas de serviço do lubrificante.

É fundamental que para que uma variável seja selecionada como preditor esta tenha uma forte correlação com a variável resposta e com as variáveis

[#] valores p menores que o nível de significância de 5%

independentes que não são selecionadas. Os preditores selecionados não devem ser correlacionados entre si ou se forem, devem ter uma fraca correlação.

3.1.4. Análise de Componentes Principais

Frequentemente, o número de variáveis a analisar é muito elevado, sendo bastante superior ao número de fontes de variabilidade que afetam o processo. Em lugar de analisar todas as variáveis, analisam-se as componentes principais, que explicam praticamente a mesma variabilidade, mas são em número bastante mais reduzido.

O ACP é matematicamente definido como uma transformação linear ortogonal que transforma os dados num novo sistema de coordenadas de forma a que a maior variância por qualquer projeção dos dados fica ao longo da primeira coordenada (primeiro componente), a segunda maior variância fica ao longo da segunda coordenada, e assim por diante. Assim sendo, o ACP determina as direções que maximizam a variabilidade dos dados e esta análise é sensível à escala na qual os dados são apresentados. Uma solução comum para contornar o problema da escala, de uma forma prática e automática, consiste em dividir as variáveis, previamente centradas por subtração pela sua média, pelo seu desvio padrão (*Autoscaling*), desta forma todas as variáveis têm igual importância à partida.

Seja a matriz de dados, $\mathbf{X}_{(n \times m)}$, onde cada uma das n linhas representa uma observação de uma dada unidade, e cada uma das m colunas representa uma variável, com vetor de valores médios $\boldsymbol{\mu}$ e matriz de variâncias covariâncias $\boldsymbol{\Sigma}$. Pretende-se encontrar um novo conjunto de variáveis $Y_1, Y_2, \dots, Y_m$, não correlacionadas entre si e cujas variâncias decresçam da primeira para a última.

Cada nova variável Y_j é uma combinação linear das m variáveis populacionais originais correlacionadas $X_1, X_2, \dots, X_m$:

$$Y_j = \mathbf{a}'_j \mathbf{X} \quad (14)$$

Sendo $\mathbf{a}'_j = [a_{1j} \ a_{2j} \dots \ a_{mj}]$ um vetor de componentes tal que $\mathbf{a}'_j \mathbf{a}_j = \sum_{i=1}^m a_{ij}^2 = 1$ e $\mathbf{a}'_j \mathbf{a}_r = 0$, para $j \neq r$, $j=1,2,\dots,m$ e $r=1,2,\dots,m$.

Encontra-se a primeira componente principal Y_1 escolhendo o vetor de constantes $\mathbf{a}_1$ de modo a que Y_1 tenha a máxima variância possível.

Tal como para todas as componentes, tem-se a restrição de que Y_1 e Y_2 , as duas primeiras componentes principais, não poderão estar correlacionadas:

$$\text{cov}[Y_1, Y_2] = \text{cov}[\mathbf{a}_2' \mathbf{X}, \mathbf{a}_1' \mathbf{X}] = E[\mathbf{a}_2' (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \mathbf{a}_1] = \mathbf{a}_2' \boldsymbol{\Sigma} \mathbf{a}_1 = 0 \quad (15)$$

Por $\boldsymbol{\Sigma}$ ser uma matriz simétrica, a restrição de normalidade de vetores próprios diferentes assegura, desde logo, que esses vetores sejam ortogonais entre si. Assim, a segunda componente principal Y_2 , será encontrada a partir de $\mathbf{a}_2$, o vetor próprio associado ao segundo valor próprio mais elevado e, em termos gerais, a j -ésima componente será encontrada a partir do vetor próprio associado ao j -ésimo valor próprio da matriz de variâncias covariâncias $\boldsymbol{\Sigma}$.

Procede-se da mesma forma para obter as restantes componentes principais.

Pode-se dizer que os *scores* $\mathbf{a}_{ij}$ são estimados de modo a que:

- a primeira componente principal explique a maior proporção de variância total das variáveis originais.
- a componente seguinte explique a maior proporção da variância não explicada pela primeira componente e esta é independente da primeira. Esta condição pode ser matematicamente dada por:

$$\mathbf{a}_{1i} \mathbf{a}_{1j} + \mathbf{a}_{2i} \mathbf{a}_{2j} + \dots + \mathbf{a}_{mi} \mathbf{a}_{mj} = 0 \quad (16)$$

Com $i=1, \dots, m$.

A obtenção das componentes principais depende somente da matriz de variâncias covariâncias, não requerendo a assunção de que as m variáveis originais apresentem distribuição normal multivariada.

Com esta transformação, os resultados mais importantes a que se chega são:

$$E[Y_j] = 0; \quad (17)$$

$$\text{var}[Y_j] = \lambda_j; \quad (18)$$

$$\text{cov}[Y_i, Y_j] = 0, \text{ se } i \neq j; \quad (19)$$

$$\text{var}[Y_1] \geq \text{var}[Y_2] \geq \dots \geq \text{var}[Y_m] \geq 0; \quad (20)$$

$$\sum_{j=1}^m \text{var}[Y_j] = \text{tr}(\boldsymbol{\Sigma}); \quad (21)$$

$$\prod_{j=1}^m \text{var}[Y_j] = |\boldsymbol{\Sigma}|. \quad (22)$$

Assim sendo, o ACP divide-se em três conceitos principais. Os *Scores*, que são ortogonais e contêm a informação sobre a relação entre as observações, os *Loadings* que são ortonormais e contêm informação sobre a relação entre as variáveis e os valores-próprios, informação sobre a variabilidade explicada pelo modelo ACP.

A variância explicada por cada componente principal (variância dos seus *scores*) reflete a importância desta componente na explicação da variabilidade total dos dados. Esta variância é dada pelo valor próprio que lhe é associado. [7]

3.2. Modelo de Regressão Linear

A equação de regressão linear múltipla define-se como:

$$y \approx \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \dots + \beta_n \cdot X_n \quad (23)$$

Onde:

- $\beta_1, \beta_2, \beta_3$ e β_n são constantes multiplicadas às variáveis ou coeficientes de regressão;
- β_0 é uma constante que representa a interceção com o eixo das ordenadas (variável resposta);
- X_1, X_2, X_3 e X_n são as variáveis preditoras;
- y é a variável de saída, neste trabalho será sempre uma de duas possíveis (horas – serviço acumulado do lubrificante medido em horas; atesto acumulado – quantidade de lubrificante adicionado ao sistema medido em gramas).

Para este caso de estudo, tem-se como *input* as variáveis presentes nas análises efetuadas aos lubrificantes em serviço, como *output* tem-se o diagnóstico do lubrificante traduzido numa variável de saída, neste caso, as horas ótimas de funcionamento do lubrificante.

Ao ajustar um modelo de qualidade, podemos simular valores dos preditores que possam representar situações limite, entenda-se cenários de falha, que se pretende evitar. A variação da resposta do modelo, indica as ações de manutenção necessárias para evitar esse evento.

O modelo pode mudar e resultar numa aproximação diferente com a introdução de uma ou mais novas observações/amostras. O modelo tende a estabilizar com a sua maturação, i.e. número de observações dez vezes superior ao número de variáveis de entrada. O coeficiente de determinação ajustado, R^2_{ajust} , é uma medida de variação explicada, mas independente do aumento no número de variáveis, tornando-se uma medida mais útil de determinar a qualidade do modelo, sendo a métrica utilizada neste trabalho para avaliar a qualidade de ajuste dos mesmos.

O ajuste do modelo pode ser classificado como satisfatório quando apresenta um R^2_{ajust} superior a 50% ou 60%, mesmo assim, a qualidade de um modelo só é verificada depois desse realizar uma previsão adequada. [6] [8]

3.2.1. Análise de Regressão Linear Univariável

Com o objetivo de ver como é que cada variável se comporta num modelo que explique as horas de serviço do lubrificante, foi ajustado um modelo de regressão linear simples para cada variável.

Através destes modelos e das análises realizadas anteriormente segue-se a escolha das melhores variáveis a inserir no modelo pretendido.

Tabela 7 - Resultados da dos modelos de regressão univariável.

Variável	$\hat{\beta}_0$	$\hat{\beta}_1$	IC 95% ($\hat{\beta}_1$)	Valor-p	Correlação com horas
IRidx	434286	-4343	] -5378.22, -3308.455[	<0.001**	-0.800
Nitr	-5.059	313.553	] 260.496, 366.609[	<0.001**	0.880
Oxid	-303.231	160.877	] 143.460, 178.295[	<0.001**	0.950
pH	2949.34	-451.390	] -511.713, -391.073[	<0.001**	-0.920
PQidx	732.080	-13.630	] -116.281, 89.019[	0.790	0.042
Sulf	339.330	116.970	] 82.404, 151.535[	<0.001**	0.730
TAN	-254.360	781.11	] 666.705, 895.511[	<0.001**	0.910
TBN	3377.920	-631.38	] -782.775, -479.992[	<0.001**	-0.800
Al	-93.340	110.180	] 88.765, 131.598[	<0.001**	0.850
B	725.950	-28.590	] -121.619, 64.438[	0.538	0.098
Ca	-1128.480	1.153	] 0.185, 2.121[	0.021**	0.360
Cr	279.770	363.770	] 238.563, 488.976[	<0.001**	0.680
Cu	244.340	551.800	] 413.238, 690.352[	<0.001**	0.790
Fe	204.400	70.440	] 46.120, 94.754[	<0.001**	0.680
Mo	669.30	122.03	] -160.178, 404.245[	0.387	0.140
Na	256.750	152.780	] 109.902, 195.652[	<0.001**	0.750
Ni	563.280	455.190	] 217.813, 692.558[	<0.001**	0.520
P	1840.612	-4.242	] -13.752, 5.269[	0.373	-0.140
Pb	684.21	44.12	] -33.282, 121.514[	0.256	0.180
S	1166.404	-0.130	] -0.574, 0.314[	0.558	0.093
Si	-130.573	5.875	] 5.540, 6.209[	<0.001**	0.980
Sn	134.740	468.940	] 333.681, 604.201[	<0.001**	0.750
Zn	-1894.592	7.332	] 2.076, 12.588[	0.007**	0.410
Visc100	-17095.400	1325.000	] 931.845, 1718.254[	<0.001**	0.730
Visc40	-15312.890	132.140	] 104.292, 159.980[	<0.001**	0.830

** Valore-p menores que o nível de significância de 5%, levam à rejeição da hipótese nula, ou seja, corresponde aos modelos que são estatisticamente significativos.

A negrito encontram-se as correlações mais significativas, pois assumem valores superiores a 0.7.

3.2.2. Seleção de Preditores

Para a seleção dos preditores deste modelo foram adotados os seguintes passos:

1. Exclusão das variáveis que não apresentaram correlação significativa com a variável de resposta (horas do lubrificante em serviço).
2. Por forma a identificar para cada grupo quais as variáveis mais importantes, procedeu-se à análise em componentes principais (ACP). Por exemplo, para o Estado físico-químico os resultados da aplicação do ACP indicaram que é mais relevante utilizar três componentes principais. Para cada uma dessas três componentes identificaram-se as variáveis com maior peso (scores), maiores valores absolutos dos coeficientes, tendo-se obtido neste caso, PC1 (pH, Oxid, TAN e Nitr), PC2 (Zn e Ca) e PC3 (Pb e P).
3. Com a noção de quais as variáveis com mais importância para cada grupo, realizou-se a seleção dos preditores através da análise das correlações entre as do mesmo grupo, para que o modelo não fosse constituído por preditores fortemente correlacionados. É importante salientar que para este passo de seleção de variáveis foram considerados os resultados do ACP apenas para saber quais as variáveis com mais peso em cada grupo, potenciando uma melhor escolha, ou seja, as componentes principais não foram utilizadas como variáveis independentes no modelo.

Para validar o modelo, o sinal dos coeficientes estimados tem de ser coerente com a evolução temporal dos preditores. Preditores oriundos dos agrupamentos Contaminação e Desgaste têm de ter coeficientes positivos. No agrupamento do Estado Físico-químico existem potenciais preditores que têm de ter coeficientes positivos como o TAN, negativos como o TBN ou teores de aditivos, e uma minoria que podem ter dupla tendência e não apresentam poder de exclusão, e.g. viscosidade.

3.2.3. Modelo de regressão Linear Múltipla

As variáveis Percentagem de Água, Percentagem de Glicol e Teor de Bário que apresentam o valor 0 em todas as observações, e as variáveis PQ Index, Teor de Boro, Teor de Molibdénio, Teor de Fósforo e Teor de Enxofre que não têm correlações significativas com a variável de resposta, não tendo apresentado resultados significativos na análise univariável, não foram consideradas para a construção do modelo de regressão linear múltipla.

Analísaram-se as variáveis por tipo, ou seja, por grupos de variáveis.

No agrupamento de Contaminação restavam apenas duas variáveis Teor de Sódio e Teor de Silício. De acordo com a opinião experiente do orientador da Galp considera-se mais relevante para o modelo o Teor de Silício, que apresenta maior correlação com a variável resposta, embora esteja um pouco mais correlacionada com as restantes variáveis independentes do que o Teor de Sódio.

Quanto ao grupo do Estado Físico-químico, através do ACP, analisando as variáveis com mais peso (score) e as suas correlações, tanto com as outras variáveis independentes como com a variável resposta, escolheram-se as variáveis pH e Teor de Zinco para incorporarem o modelo.

No grupo do Desgaste, foram selecionadas as variáveis Teor de Alumínio e Teor de Ferro.

Assim sendo, selecionaram-se inicialmente para o modelo as variáveis: Teor de Silício, pH, Teor de Zinco, Teor de Alumínio e Teor de Ferro. Aplicando a regressão linear múltipla obteve-se o seguinte modelo:

$$\text{Hora} \approx 287.388 + 5.6074 * \text{Si} - 62.390 * \text{pH} - 9.757 * \text{Fe} \quad (24)$$

$$R^2_{\text{adj}} = 97.7\%$$

$$\text{Valor-p} < 0.001$$

Observando o valor-p associado ao teste de hipóteses relativo à significância do modelo, valor-p < 0.001, conclui-se que o modelo é globalmente significativo. O valor elevado do R^2_{adj} de 97.7%, que representa a percentagem de variabilidade da variável resposta que é explicada através do modelo de regressão linear múltipla, conclui-se que à partida que o modelo (24) está bem ajustado.

O R^2_{adj} (coeficiente de determinação ajustado) é usado na regressão linear múltipla para ver o grau de intensidade ou eficácia das variáveis independentes na explicação da variável dependente.

Diz qual percentagem da variação da variável dependente que é explicada coletivamente por todas as variáveis independentes.

A utilização desse coeficiente justifica-se, pois à medida que adicionamos variáveis a um modelo de regressão, o coeficiente de determinação sem ajuste (R^2) tende a aumentar. Mesmo quando a contribuição marginal de cada uma das novas variáveis adicionadas não tem relevância estatística.

Portanto, ao adicionar variáveis para o modelo, o coeficiente de determinação poderia aumentar e poderia pensar-se, erradamente, que um maior número de variáveis selecionadas pode explicar mais a variação da variável dependente.

3.2.4. Análise de Resíduos

Após o ajustamento do modelo de regressão linear múltipla as suposições do modelo ajustado precisam ser validadas para que os resultados sejam confiáveis. Sendo assim, é benéfico analisar os resíduos. Esta análise contém um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos.

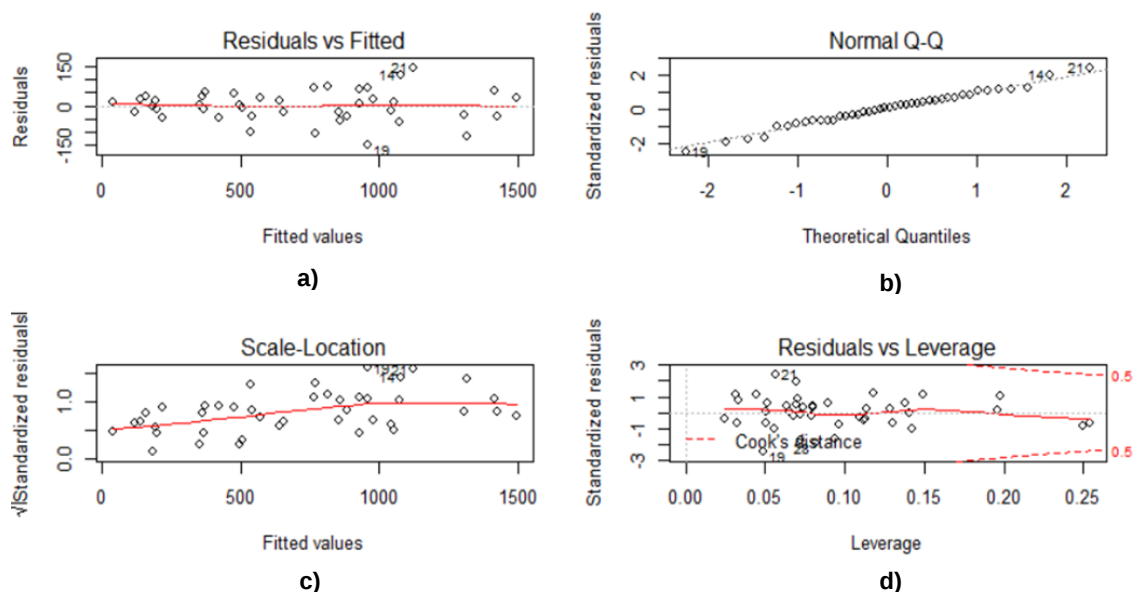


Figura 4-Gráficos análise de resíduos de modelo de regressão linear múltipla: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.

Na Figura 4. a), pretende-se avaliar se a variância dos resíduos é não constante (existência de heterocedasticidade). Neste caso, observa-se que não existe nenhum padrão distinto na comparação entre os resíduos e os valores estimados, o que indica que nos resíduos não há informação que possa ser relevante para o modelo, i.e. assume-se que os resíduos são independentes dos valores ajustados, ou seja, a correlação entre eles toma o valor 0.

A Figura 4. b) representa a normalidade dos resíduos, é benéfico que os resíduos estejam bem alinhados na linha reta tracejada, o que acontece na maioria das observações, exceto as observações 14, 19 e 21 que devem ser mantidas sob controlo como um potencial problema.

Para confirmar, o teste de *Shapiro-Wilk* para os resíduos apresenta um *valor-p* = 0.866, o que indica que não se rejeita a hipótese nula, ou seja, os resíduos têm distribuição normal. [17]

Na Figura 4. c), os resíduos são distribuídos igualmente ao longo dos intervalos de preditores, o que parece verificar a suposição de igual variância (homocedasticidade). É bom ver uma linha horizontal com pontos de igual distribuição aleatória. Neste caso não temos uma linha horizontal perfeita, mas aproxima-se muito do desejado, em que podemos observar claramente a aleatoriedade. Mais uma vez temos as observações 14, 19 e 21 que se afastam um pouco das restantes.

Por último, a Figura 4. d) mostra a existência de *outliers*. Isso poderá ajudar se esses *outliers* não contribuírem em nada para o nosso modelo, e se decidirmos retirá-los haverá modificações.

Neste caso não há casos influentes. Mal se consegue ver as linhas de distância de Cook (a linha tracejada vermelha) porque todas as observações estão dentro dessas linhas de distância. Mesmo assim, a função mostra-nos as observações que mais se afastam das restantes.

3.2.5. Previsão

Por fim, utilizou-se o modelo para simular as horas de serviço do lubrificante para valores limite dados pelo orientador da Galp para cada um dos preditores.

A Assessoria Técnica da Galp pretende utilizar os modelos para simular condições em que os preditores tomam valores limite, i.e. estão na fronteira entre uma condição de lubrificação estável e a necessidade de tomar uma intervenção no equipamento.

Sendo assim, com Si= 300 ppm, pH= 4 e Fe= 40 ppm obteve-se uma previsão de aproximadamente 1328 horas, com um intervalo de confiança entre aproximadamente 1124 e 1532 horas, o que é transmitido aos responsáveis de manutenção como o serviço ótimo que o lubrificante deve acumular, em horas, até ser completamente renovado.

A tabela seguinte apresenta o resultado final da modelação com regressão linear múltipla:

Tabela 8 - Resultados finais da modelação com regressão Linear Múltipla.

Modelo:	$Hora \approx 287.388 + 5.6074 * Si - 62.390 * pH - 9.757 * Fe$ (24)
Parâmetros de Simulação:	Si= 300 ppm, pH= 4 e Fe= 40 ppm
Resultado:	Hora= 1328 horas; IC = [1124 ,1532]

Através deste *output* afirmar-se que o modelo está a conceder valores aceitáveis, mas com um intervalo de horas muito extenso.

Como se trata de equipamentos muito delicados, a estimativa intervalar do número de horas que o lubrificante pode permanecer na máquina deveria desejavelmente apresentar uma amplitude pequena, ou seja, deveria ser mais precisa para que não haja erros que levem à ocorrência de uma possível avaria.

Em suma, este modelo parece adequado ao tipo de problema que temos, mas para que fosse mais preciso seria necessário ter um maior número de observações, para ajustar um modelo com maior confiança.

3.3. Modelo de Regressão Logística

Como já foi explicado na introdução, a regressão logística é uma técnica recomendada para situações em que a variável dependente é de natureza dicotômica ou binária. Quanto às variáveis independentes, tanto podem ser categóricas ou não.

A regressão logística é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias, ou seja, o seu *output* fica contido no intervalo de zero a um. [10] [18]

Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente. No caso da variável dependente y assumir apenas dois possíveis estados (1 ou 0) e haver um conjunto de n variáveis independentes X_1, X_2, X_3 e X_n , o modelo de regressão logística pode ser escrito da seguinte forma:

$$P(y=1) = \frac{e^{f(x)}}{1+e^{f(x)}} \quad (25)$$

Em que,

$$f(x) \approx \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \dots + \beta_n \cdot X_n \quad (26)$$

Onde:

- y – Variável binária de saída;
- $\beta_1, \beta_2, \beta_3$ e β_n são constantes multiplicadas às variáveis ou coeficientes de regressão;
- β_0 é uma constante que representa a interceção com o eixo das ordenadas (variável resposta);

Assim sendo, para este caso de estudo, temos como variável dependente a mudança de óleo e como preditor obrigatório as horas de funcionamento do lubrificante.

3.3.1. Análise de Regressão Logística Univariável

Na Tabela 9 apresentam-se os resultados do modelo de regressão logística univariável. Estes resultados permitem analisar a associação entre cada variável e a necessidade de mudança do lubrificante.

Tabela 9 - Resultados dos modelos de regressão logística univariável.

Variável	$\hat{\beta}_0$	$\hat{\beta}_1$	IC95% ($\hat{OR}$)	$\hat{OR}$	Valor-p
IRídx	998.909	-10.022	]1.981e-10, 1.192[	4.439e-05	0.069
Nitr	-7.251	2.070	]2.232, 53.142[	7.926	0.008**
Oxid	-7.359	0.803	]1.346, 4.838[	2.233	0.010**
pH	7.228	-1.859	]0.022, 5.764e-01[	0.154	0.020**
PQídx	-2.309	0.387	]0.776, 3.141[	1.472	0.265
Sulf	-2.266	0.230	]0.936, 1.738[	1.259	0.133
TAN	-6.204	3.359	]2.971, 672.797[	28.748	0.013**
TBN	8.230	-2.379	]0.008, 5.532e-01[	0.093	0.022**
Al	-5.156	0.442	]1.160, 2.329[	1.555	0.010**
B	-1.380	-0.095	]0.433, 1.558[	0.909	0.755
Ca	-9.545	0.005	]9.992e-01, 1.012[	1.005	0.100
Cr	-3.557	1.507	]1.368, 21.467[	4.511	0.027**
Cu	-4.387	2.837	]2.597, 354.640[	17.070	0.014**
Fe	-3.939	0.298	]1.086, 1.763[	1.348	0.013**
Mo	-2.197	1.861	]1.279, 38.246[	6.429	0.028**
Na	-2.383	0.288	]0.933, 1.976[	1.334	0.114
Ni	-2.603	2.449	]2.157, 92.017[	11.571	0.008**
P	11.748	-0.050	]0.872, 1.018[	0.952	0.205
Pb	-1.366	-0.309	]NA, 1.324[	0.734	0.594
S	5.415	-0.002	]0.994, 1.001[	0.998	0.237
Si	-5.530	0.024	]1.009, 1.046[	1.024	0.007**
Sn	-4.548	2.138	]1.827, 64.306[	8.480	0.015**
Zn	-3.472	0.006	]9.710e-01, 1.042[	1.006	0.745
Visc100	-58.979	4.264	]1.281, 1.381e+04[	71.101	0.064
Visc40	-56.047	0.448	]1.113, 2.409[	1.564	0.020
IRídx	-4.802	0.004	]1.001, 1.007[	1.001	0.008

** Valores-p menores que o nível de significância de 5%, ou seja, correspondem a modelos que são estatisticamente significativos.

A negrito estão as estimativas dos odds ratio (OR) com valores superiores a 1.

Neste tipo de regressão são analisadas as *odds ratio* (OR), são sempre não negativas, sendo importante distinguir que, no nosso caso, quando o seu valor é menor que 1 é menos provável ter que mudar o lubrificante, se for maior que 1 é mais provável ter que mudar o lubrificante e se for igual a 1, ter ou não de mudar o lubrificante é igualmente provável. [9]

No universo de observações destes dados não existem recolhas de amostras, i.e. observações, intervaladas com menos de 50 horas, assim sendo, faz mais sentido analisarmos a probabilidade de mudar o lubrificante quando as horas de funcionamento do lubrificante aumentam 50 horas, do que só aumentar 1 hora. Portanto analisamos as *odds ratio* como a probabilidade de ter de mudar de lubrificante para cada 50 horas de funcionamento a mais.

Neste caso, houve a necessidade de na base de dados alterar a variável horas, ou seja, dividiram-se os valores registados por 50. Sendo assim, a interpretação da *odds ratio* já faz mais sentido para esta variável, passando para uma $\widehat{OR}_{\text{horas}}$ é igual a 1.2128.

3.3.2. Seleção de Preditores

Para a seleção dos preditores deste modelo foram adotados os seguintes passos:

1. Exclusão das variáveis com pouco significado estatístico, ou seja, ajustando um modelo de regressão logística univariável, foram excluídas aquelas que apresentavam um valor-p, do teste de *Wald*, superior ao nível de significância de 5%.
2. Para cada grupo de variáveis fez-se o ACP, ou seja, por exemplo, para o Estado físico-químico as variáveis que têm maior peso são o pH, TBN, Oxidação e TAN.
3. Realizou-se a seleção dos preditores através da análise das correlações entre as variáveis selecionadas para cada grupo, para que no modelo não fosse constituído por preditores fortemente correlacionados, mas que também consigam explicar as variáveis independentes que não permaneceram no modelo.

Como referido na seleção de preditores para o modelo de regressão linear múltipla (Vd. Capítulo 3.2.2) para validar o modelo, o sinal dos coeficientes estimados tem de ser coerente com a evolução temporal dos preditores, i.e. se o preditor tiver uma tendência de crescimento com a vida da carga de lubrificante, o coeficiente também tem de ser positivo, se por outro lado, o

preditor decrescer, o seu coeficiente terá de ser negativo, estas tendências podem variar de acordo com o grupo a que o preditor pertence, explicado no capítulo referido.

Este aspeto é muito difícil de cumprir neste caso de estudo pois existem muito poucas observações para tantas variáveis o que causa um reduzido número de graus de liberdade, fazendo que existam efeitos supressores.

3.3.3. Modelo de Regressão Logística

Seguindo os passos explicados anteriormente, excluíram-se as variáveis Percentagem de Água, Percentagem de Glicol e Teor de Bário que apresentam o valor 0 em todas as observações, e as variáveis menos significativas para o modelo IR Index, PQ Index, Sulfatação, Teor de Boro, Teor de Cálcio, Teor de Sódio, Teor de Fósforo, Teor de Chumbo, Teor de Enxofre, Teor de Zinco e Viscosidade a 100°C pois não apresentaram resultados significativos na análise de correlação univariável.

De seguida, analisaram-se as variáveis por grupo. No agrupamento da Contaminação restava apenas o Teor de Silício.

Quanto ao grupo do Estado Físico-químico, através do ACP somente para as variáveis inseridas neste grupo, é aconselhada a inclusão das variáveis Nitração e TAN.

No grupo do Desgaste o procedimento é semelhante, escolhendo neste caso as variáveis Teor de Níquel e Teor de Estanho.

Sendo assim, considerando inicialmente as variáveis independentes: Teor de Silício, Nitração, TAN, Teor de Níquel e Teor de Estanho. Aplicou-se a regressão logística obteve-se o seguinte modelo estimado:

$$f(x) \approx -5.467 + 3.812 * \text{Nitr} - 7.543 * \text{TAN} + 3.596 * \text{Ni} + 0.119 * \text{hora}/50 \quad (27)$$

$$P(y=1) = \frac{e^{f(x)}}{1 + e^{f(x)}} \quad (28)$$

Em que:

- y – Mudança de lubrificante (1-Mudar, 0-Não Mudar);
- hora/50 – Número de horas que o lubrificante fica em funcionamento divididas por 50.

O modelo (27) tem um AIC de 32.001. De entre todos os modelos que se realizaram deste tipo, este é o que apresenta menor AIC, pelo que se procedeu ao estudo da qualidade e resíduos deste modelo. [10]

Para testar a qualidade e significância do modelo foi realizado o teste de *Hosmer-Lemeshow* que é muito utilizado em regressão logística com a finalidade de testar a bondade do ajuste, ou seja, o teste comprova se o modelo proposto pode explicar bem o que se observa. O teste avalia o modelo ajustado através das distâncias entre as probabilidades ajustadas e as probabilidades observadas.

Neste caso, tem-se um valor-p aproximadamente igual a 0.776, que é superior ao nível de significância de 5%, ou seja, o modelo está bem ajustado.

O teste de *Hosmer-Lemeshow* não depende do número de ensaios por linha nos dados como os outros testes de qualidade do ajuste. Quando os dados têm poucos ensaios por linha, o teste de *Hosmer-Lemeshow* é um indicador mais confiável, para avaliar o ajuste do modelo.

Na regressão logística, a qualidade do ajuste de um modelo não se analisa da mesma forma que na regressão linear múltipla, através do R^2_{adj} , por essa razão, por forma a confirmar a qualidade do modelo de regressão logística, realizou-se o cálculo de *Pseudo R²*, composto por três cálculos distintos de R^2 . Infelizmente estes valores não são muito expressivos, pois raramente se consegue valores de R^2 elevados. [10]

Neste caso tem-se o $R^2_{Mc Fadden}$ igual a 0.462, o $R^2_{Cox and Snell}$ de 0.362 e o $R^2_{Nagel Kerke}$ igual a 0.582.

Fez-se ainda a curva *Receiver Operating Characteristic* (ROC) e obteve-se a AUC, correspondente à área sob a curva ROC (Vd. Figura 5). A análise ROC é uma ferramenta que permite avaliar a qualidade do desempenho de sistemas de classificação que visam responder a que classe pertence cada indivíduo em estudo. Neste caso, tem-se a classificação de um lubrificante em duas classes, 1 se for preciso realizar a sua mudança, 0 caso contrário. [10]

A curva ROC é um gráfico em que a coordenada das abcissas corresponde a (1-especificidade) e a das ordenadas corresponde à sensibilidade. A sensibilidade é a probabilidade de que o modelo indique a necessidade de mudar o óleo e realmente o lubrificante ter sido mudado. A especificidade é a probabilidade de que o modelo não indique necessidade de mudar o óleo e realmente esse lubrificante não ter sido mudado.

É também calculada a área abaixo da curva ROC, AUC, que está associada ao poder discriminante de um teste de diagnóstico, neste caso, a estimativa da AUC para o modelo ajustado é igual a 0.915, o que é muito próximo de 100%, indicando que o modelo fez uma boa classificação.

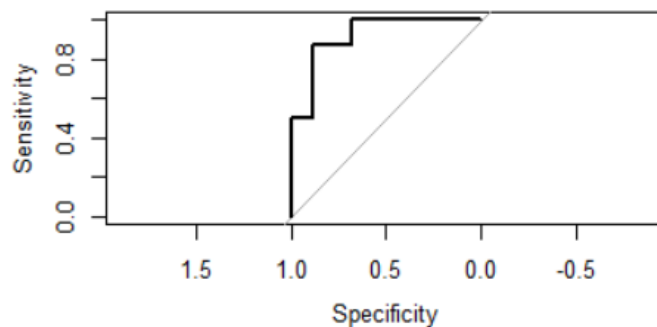


Figura 5-Curva ROC modelo de regressão logística

3.3.4. Análise de Resíduos

Para estudar os resíduos neste tipo de modelos é utilizado os resíduos de *Pearson*. Na figura seguinte são comparados os resíduos com cada um dos preditores do modelo, e por fim com o modelo. [19]

Note-se que a relação entre os resíduos de *Pearson* com os preditores TAN e a hora de funcionamento é linear, mas existe uma ligeira tendência, já para a Nitração e Teor de Níquel apesar de linear é mais complicado analisar os resíduos devido aos valores que estas duas variáveis tomam.

Para além dos gráficos, esta função realiza testes estatísticos formais (teste de falta de ajuste) para ver se uma variável tem relação com os resíduos. Para este caso, como previsto confirma-se que a variável TAN e horas não forma uma boa escolha pois têm *valores-p* inferiores a 5% no teste dos resíduos.

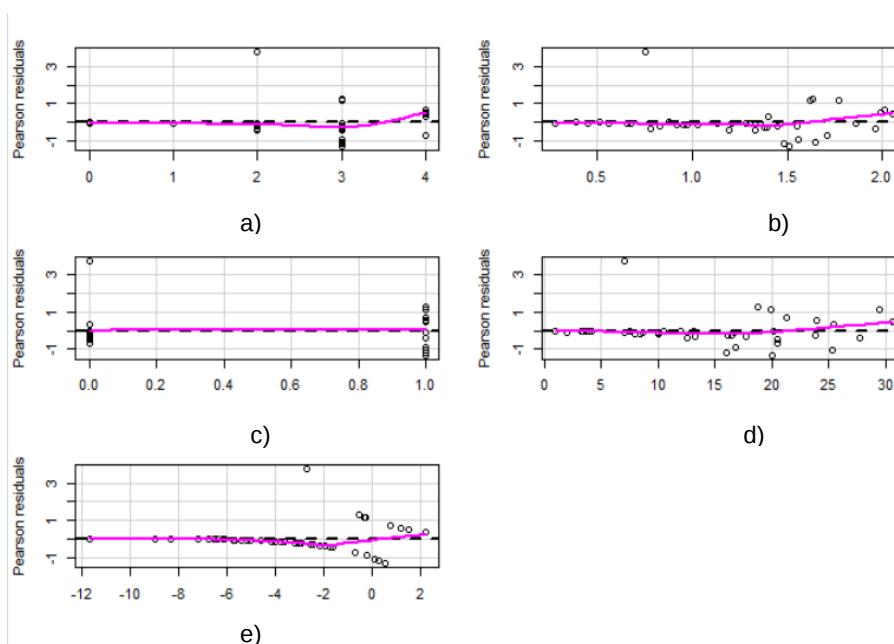


Figura 6-Gráficos de análise de resíduos do modelo de regressão logística: a) Nitração; b) TAN; c) Ni; d) Horas; e) Valores ajustados

3.3.5. Previsão

Por fim, para testar o modelo estimaram-se as probabilidades de necessidade de mudança do lubrificante para valores limite dados pelo orientador da Galp, para cada um dos preditores.

Para este modelo, foram realizadas 3 simulações distintas:

- 1) Para o óleo novo com valores de Nitração=0 A/cm, TAN=0.45 mgr KOH/gr, Ni=0 ppm e horas/50= 0 horas.
Obteve-se uma probabilidade de aproximadamente 0.0001 o que indica que o lubrificante não necessita de ser mudado.
- 2) Para um RUL DE 50%, tem-se os valores de Nitração=10 A/cm, TAN=1.2 mgr KOH/gr, Ni=5 ppm e horas/50= 14 horas.
Obteve-se uma probabilidade de aproximadamente 1 o que indica que o lubrificante necessita de ser mudado.
- 3) Para um RUL de 0%, tem-se os valores de Nitração=20 A/cm, TAN=2 mgr KOH/gr, Ni=10 ppm e horas/50= 27.72 horas.
Obteve-se uma probabilidade de aproximadamente 0 o que indica que o lubrificante não necessita de ser mudado.

Ao analisar os resultados das três simulações, pode-se afirmar que para o lubrificante novo e para um RUL de 0% o modelo está a dar resultados aceitáveis. Para um lubrificante novo, o modelo não prevê a mudança de lubrificante e para o RUL=0%, fim da sua vida útil, prevê a mudança do lubrificante.

Já na segunda simulação os resultados não são aceitáveis, visto que quando as características do lubrificante indicam que este tem um RUL de 50%, ou seja, ainda não é necessário a sua troca, o modelo antecipa a sua mudança.

3.4. Análise de Sobrevivência – Modelo de Cox

Neste tipo de modelação tem-se a variável aleatória contínua T , que representa o tempo de vida de um indivíduo com vetor de covariáveis associadas, $\mathbf{x}=(x_1, \dots, x_n)'$, num determinado instante t . Sejam $\boldsymbol{\beta}=(\beta_1, \dots, \beta_n)$, os coeficientes de regressão (desconhecidos), que representam o efeito das covariáveis na sobrevivência e $h_0(t)$ a função de risco subjacente (função arbitrária não negativa), ou seja, aquela que corresponde a um indivíduo com vetor de covariáveis nulo. Então o modelo de Cox tem a seguinte expressão:

$$h(t; \mathbf{z}) = h_0(t) e^{\boldsymbol{\beta}' \mathbf{x}} = h_0(t) e^{(\beta_1 x_1 + \dots + \beta_n x_n)} \quad (29)$$

Deste modo, o efeito das covariáveis é modelado parametricamente, mas o mesmo não acontece em relação à função de risco subjacente, pelo que o modelo de Cox é um modelo de regressão semi-paramétrico.

Neste caso de estudo, temos um problema que pode dificultar a escolha deste método, pois as observações não são independentes entre si, visto que a cada observação estuda-se o mesmo lubrificante do mesmo equipamento, não havendo indivíduos diferentes e independentes entre si.

Outro problema é o número reduzido de observações e o facto de nessas poucas observações só ter sido trocado o lubrificante 7 vezes.

3.4.1. Análise de Modelo de Cox Univariável

Com o objetivo de ver como é que cada variável se comporta num modelo de Cox que indique o tempo de sobrevivência do lubrificante, fez-se o modelo para cada variável.

Tabela 10 - Análise dos modelos de Cox univariável.

Variável	$\hat{\beta}$	Hazards Ratios	Valor-p	Proportional-hazards valor-p
IRidx	30.490	1.736*10 ¹³	0.034**	0.487
Nitr	-0.232	0.793	0.750	0.599
Oxid	-0.438	0.645	0.211	0.949
pH	1.795	6.017	0.103	0.841
PQidx	0.369	1.446	0.341	0.619
Sulf	-0.285	0.752	0.094	0.751
TAN	-2.323	0.098	0.151	0.997
TBN	4.579	97.444	0.053	0.890
Al	-0.206	0.814	0.264	0.201
B	0.466	1.594	0.184	0.783
Ca	-0.0003	1.000	0.911	0.258
Cr	-0.303	0.739	0.641	0.236
Cu	-1.208	0.299	0.199	0.227
Fe	0.011	1.011	0.931	0.229
Mo	2.213	9.147	0.045**	0.416
Na	-1.157	0.315	0.018**	0.591
Ni	0.046	1.047	0.960	0.545
P	-0.056	0.945	0.324	0.450
Pb	-18.360	1.064*10 ⁻⁸	0.998	1
S	-0.004	0.996	0.142	0.786
Si	-0.059	0.943	0.029**	0.728
Sn	-0.843	0.430	0.391	0.985
Zn	-0.073	0.930	0.020**	0.342
Visc100	-5.350	0.005	0.056**	0.495
Visc40	-0.319	0.727	0.053**	0.809

** Valores-p menores que o nível de significância de 5%, ou seja, correspondem aos modelos que são estatisticamente significativos.

A negrito estão as estimativas dos odds com valores superiores a 1.

Após confirmar os resultados obtidos nestas análises verifica-se que não é possível selecionar os preditores, visto que os valores obtidos para os modelos de Cox univariável, dos *hazard ratios* e do teste de hipótese de proporcionalidade não são estatisticamente significativos. Por essa razão, não foi realizado nenhum modelo de Cox para estes dados.

3.5. Discussão de Resultados do Caso de Estudo Preliminar

Após a preparação feita, é possível tirar conclusões sobre que metodologias usar para o caso de estudo principal. Para o modelo de regressão linear múltipla (24) obtiveram-se bons resultados, tanto nos valores de indicação de qualidade, como também nas previsões realizadas.

O defeito que lhe é detetado é o facto de resultar num intervalo de confiança alargado, que será previsivelmente e gradualmente atenuado com o aumento de observações.

Após o ajuste do modelo de regressão logística conclui-se que o mesmo não é indicado para este projeto, visto que o seu *output* não satisfaz o objetivo e apresenta falhas para diferentes previsões. Pode-se considerar que a escolha dos preditores não foi muito adequada, mas ao tentar outras seleções de variáveis, o modelo (27) foi o que teve melhores resultados.

Para uma possível utilização deste modelo é necessário um maior número de observações para que o modelo apresente melhor qualidade de ajuste.

Verifica-se também que a variável de resposta indica a recomendação de 'muda=1' quando os preditores são simulados para um $RUL \approx 50\%$, quando os diagnósticos qualitativos da Assessoria Técnica e Qualidade de Produtos são ainda 'Normais'. A utilização deste modelo leva a recomendações de muda, muito antes do necessário.

No caso da análise de sobrevivência não foi ajustado o modelo de Cox, pois no processo de escolha de preditores através das análises que se realizaram os valores não são estatisticamente significativos, confirmando-se a suspeita que este modelo não é adequado para estes dados.

Visto que para este modelo é necessário que cada observação seja independente de todas as outras, o que não acontece porque todas as observações correspondem ao mesmo lubrificante do mesmo equipamento.

Assim sendo, conclui-se que o melhor modelo a ser utilizado com este tipo de dados é a regressão linear múltipla, sendo a filosofia de modelação que vai ser adensada no projeto GALP MODe AIML e no projeto principal do estágio.

4. Modelação de um Caso de Estudo

Como projeto principal do estágio, pretendem-se estudar duas prensas de uma fábrica que produz pastas branqueadas a partir de biomassa.

A biomassa utilizada é proveniente de eucalipto. Entre a entrada desta biomassa e a saída de papel branco para impressora, tem lugar o processo de branqueamento. Neste processo, a matéria-prima, após o cozimento, é colocado no seio fluido com água e químicos.

Este fluxo é sujeito a uma série de estágios de prensas que têm dois objetivos: renovar fibras com baixa resistência mecânica e gradualmente esbranquiçar a pasta de papel.

As referidas prensas consistem em dois rolos, um deles fixo, outro móvel, suportado num sistema hidráulico que controla o binário a que o mesmo está sujeito. Os rolos têm tipicamente diâmetro entre os 300 e os 500 milímetros e velocidades de rotação baixos.

A atividade destes equipamentos é determinante na qualidade da pasta de papel produzida, bem como, na continuidade do fluxo produtivo. Estes equipamentos têm regimes de produção muito intensivos e são considerados críticos.

Cada um destes rolos, fixo e móvel, têm um rolamento nas suas extremidades. Considera-se o lado do Acoplamento (LA) o lado do qual existe a transmissão de movimento, proveniente de motores elétricos acoplados a redutores, que reduzem a rotação. O lado oposto ao acoplamento (LOA) refere-se à outra extremidade.

No total perfaz 4 rolamentos por prensa.

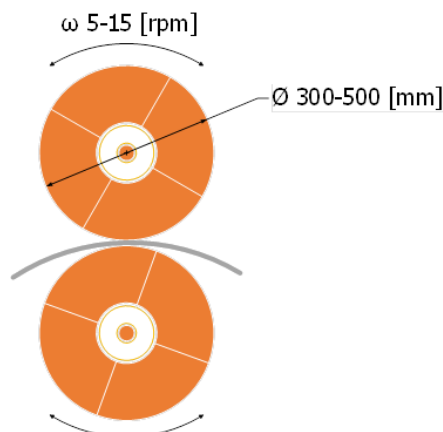


Figura 7 – Esquema de Prensa de Papel.

Para o projeto serão estudadas a Prensa 1, constituída por 4 rolos, dois fixos e dois móveis, que são lubrificados com óleo; e a Prensa 3, ao qual só se estuda um rolo lubrificado com massa.

Foi aconselhado para a prensa 1 uma lubrificação com o óleo *GALP OPS 3200*. É um lubrificante sintético de viscosidade muito elevada, destinados especialmente à lubrificação de engrenagens e rolamentos sujeitos a cargas elevadas e velocidades reduzidas. [20]

Para a prensa 3, é utilizada a massa lubrificante *GALP GEMINI PLEX 0 1*, de extrema pressão e alto desempenho, que oferece proteção ao desgaste, corrosão e à oxidação sem utilizar as formulações asfálticas convencionais. Apresenta bom desempenho a temperaturas elevadas e a tecnologia de aditivação utilizada oferece uma elevada estabilidade mecânica bem como uma grande adesividade, a qual garante o máximo de cobertura de filme por um longo período de tempo. [21]

Portanto para o projeto será feito o acompanhamento de 5 órgãos, sendo necessário realizar um modelo estatístico para cada um. Como concluído no capítulo anterior serão ajustados modelos através da regressão linear múltipla.

Para este caso de estudo, para além de se considerar como variáveis explicativas aquelas que estão registadas nos dados das análises dos lubrificantes em serviço, serão adicionadas à base de dados a velocidade de rotação dos rolos e o binário. Como variável resposta tem-se o atesto acumulado de lubrificante, ou seja, os litros adicionados, caso se esteja a analisar óleo, ou as gramas adicionadas, caso seja massa lubrificante.

Tal como no estudo de preparação, para cada órgão foi realizada uma análise exploratória, que incluiu as estatísticas descritivas básicas, os coeficientes de correlação, testes de normalidade e ainda análise em componentes principais.

A seleção dos preditores é baseada nessas análises e ainda nos valores obtidos através dos modelos de regressão linear univariável. Por fim, fez-se a análise dos resíduos e a previsão do atesto acumulado para cada um dos modelos.

4.1. Prensa 1

- Rolo Fixo LA

Para este rolo tem-se uma base de dados com 16 observações completas de 24-10-2018 a 13-03-2019 e 16 variáveis. A proporção é de uma observação por variável, o que ainda não permite um modelo maduro.

Foi criado um modelo com um $R^2_{\text{ajust}}=34.5\%$, que retorna os litros de atesto acumulado, tendo como preditor o número de acidez (TAN), indicador da reserva disponível de aditivos antioxidantes do lubrificante.

$$\text{atesto acum} \approx 118.578 - 43.632 * \text{TAN} \quad (30)$$

Para o valor de TAN=0.7 mgr KOH/gr resulta um atesto acumulado de 88.036 litros, com intervalo de confiança a 95% igual a [67.497;108.574].

Tabela 11 - Resultados finais para o modelo do rolo fixo LA, da prensa 1.

Capacidade do Cáter:	45 litros
Volume de Atesto do Lubrificante Acumulado:	102.10 litros
Volume de Atesto do Lubrificante Recomendado:	88.0.36 litros (-4.10 litros)
Intervalo de Confiança para volume de atesto recomendado:	[67.497 ; 108.574]

É de salientar que o R^2_{ajust} para este modelo é muito baixo, pois é menor que 60%. Para validar a qualidade do modelo e do seu ajuste realizou-se a análise de resíduos do modelo.

Analizando os quatro gráficos, Figura 8, pode-se observar que existem 3 possíveis *outliers* (observações 3, 15 e 16). Visto que a base de dados apresenta um reduzido número de observações, analisar os resíduos pode ser uma tarefa difícil, pois as poucas observações que se tem são muito concentradas e apresentam valores muito idênticos. Quanto à normalidade dos resíduos, segundo o teste de *Shapiro-Wilk* apresenta um valor-p = 0.012, o que indica que se rejeita a hipótese nula, ou seja, os resíduos não têm distribuição normal.

Seria aconselhado aguardar o aumento do número de observações para uma melhor modelação, e verificar se as 3 observações tomadas como possíveis

outliers se mantêm afastadas das restantes observações, antes de as excluir da base de dados.

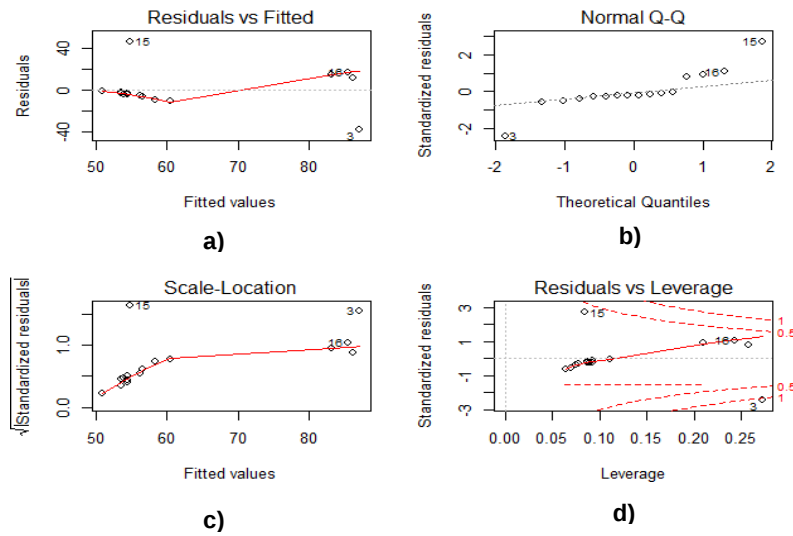


Figura 8-Análise de resíduos modelo de regressão linear múltipla do rolo fixo LA, Prensa 1: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.

- Rolo Fixo LOA

Para este rolo tem-se uma base de dados com 15 observações completas de 24-10-2018 a 13-03-2019 e 16 variáveis. A proporção é menor que uma observação por variável, o que ainda não permite um modelo maduro.

Foi criado um modelo com um $R^2_{\text{ajust}}=79.7\%$, que retorna os litros de atesto acumulado, em que os preditores são o Teor de Cálcio, indicador do estado físico-químico do lubrificante e o Teor de Ferro, indicador de desgaste mecânico.

$$\text{atesto acum} \approx 50.944 + 0.093 \cdot \text{Ca} - 0.122 \cdot \text{Fe} \quad (31)$$

Para o valor de $\text{Ca}=60$ ppm e $\text{Fe}=20$ ppm resulta um atesto acumulado de 54.057 litros, com intervalo de confiança a 95% igual a [52.691;55.424].

Tabela 12 - Resultados finais para o modelo do rolo fixo LOA, da prensa 1.

Capacidade do Cáster:	45 litros
Volume de Atesto do Lubrificante Acumulado:	56.600 litros
Volume de Atesto do Lubrificante Recomendado:	54.057 litros (-2.543 litros)
Intervalo de Confiança para volume de atesto recomendado:	[52.691 ; 55.424]

Analisando os quatro gráficos, Figura 9, pode-se observar que existem 2 observações que se afastam mais das restantes, sendo aconselhado uma maior atenção (observações 12 e 14).

Quanto à normalidade dos resíduos, segundo o teste de *Shapiro-Wilk* apresenta um valor-p = 0.171, o que indica que não se rejeita a hipótese nula, ou seja, os resíduos têm distribuição normal.

Os restantes gráficos apresentam valores aceitáveis mas é necessário uma segunda modelação com um maior número de observações para ser confirmada a qualidade do modelo.

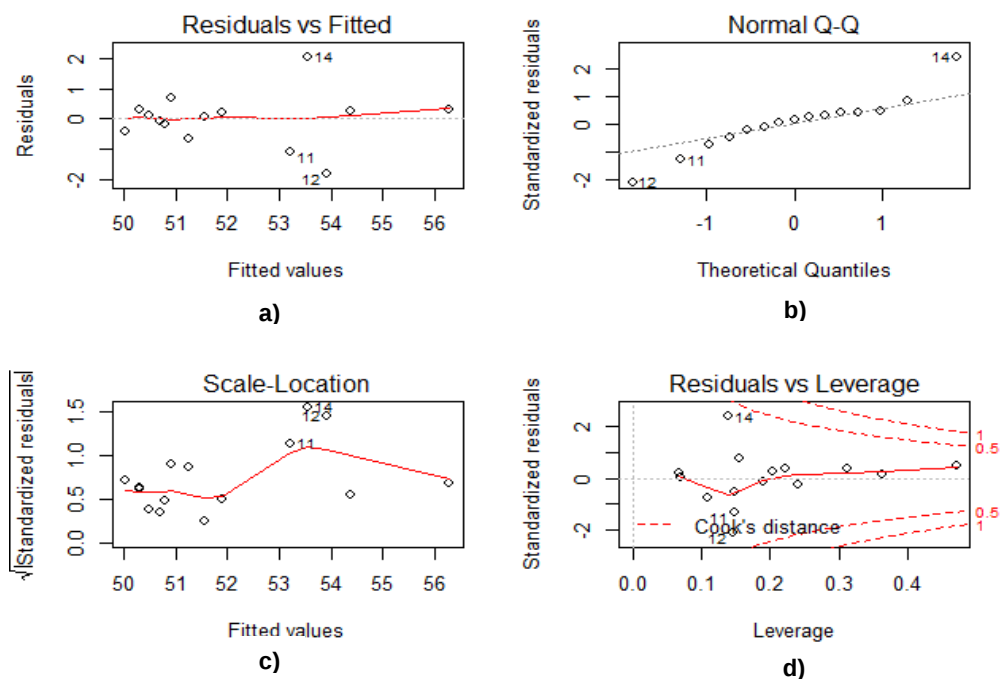


Figura 9-Análise de resíduos modelo de regressão linear múltipla do rolo fixo LOA, Prensa 1: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.

- Rolo Móvel LA

Para este rolo tem-se uma base de dados com 16 observações completas de 24-10-2018 a 13-03-2019 e 17 variáveis. A proporção é um pouco maior do que uma observação por variável, o que ainda não permite um modelo maduro.

Foi criado um modelo com um $R^2_{\text{ajust}}=96.1\%$, que retorna os litros de atesto acumulado, tendo como preditor o número de acidez (TAN), indicador da reserva disponível de aditivos antioxidantes do lubrificante, o Teor de Ferro, indicador de desgaste mecânico e o Teor de Sódio, indicador de contaminação.

$$\text{atesto acum} \approx 149.994 - 69.484 * \text{TAN} + 0.080 * \text{Fe} - 4.429 * \text{Na} \quad (32)$$

Para o valor de TAN=0.8 mgr KOH/gr, Fe= 200 ppm e Na= 0 ppm resulta um atesto acumulado de 110.455 litros, com intervalo de confiança a 95% igual a [100.735;120.175].

Tabela 13 - Resultados finais para o modelo do rolo móvel LA, da prensa 1.

Capacidade do Cáter:	45 litros
Volume de Atesto do Lubrificante Acumulado:	101.100 litros
Volume de Atesto do Lubrificante Recomendado:	110.455 litros (+9.355 litros)
Intervalo de Confiança para volume de atesto recomendado:	[100.735 ; 120.175]

Ao analisar os gráficos, Figura 10, pode-se observar que os resíduos apresentam um padrão mas também que as observações se concentram em dois grupos, não sendo visível nenhum tipo de *outlier*, estando as observações 12 e 14 de o serem.

Quanto à normalidade dos resíduos, segundo o teste de *Shapiro-Wilk* apresenta um valor-p = 0.066, o que indica que não se rejeita a hipótese nula, para um nível de significância de 5%, ou seja, os resíduos têm distribuição normal. Mesmo verificando-se a normalidade dos resíduos, este modelo é um dos que se deve ter maior atenção para uma futura modelação com uma base de dados mais extensa.

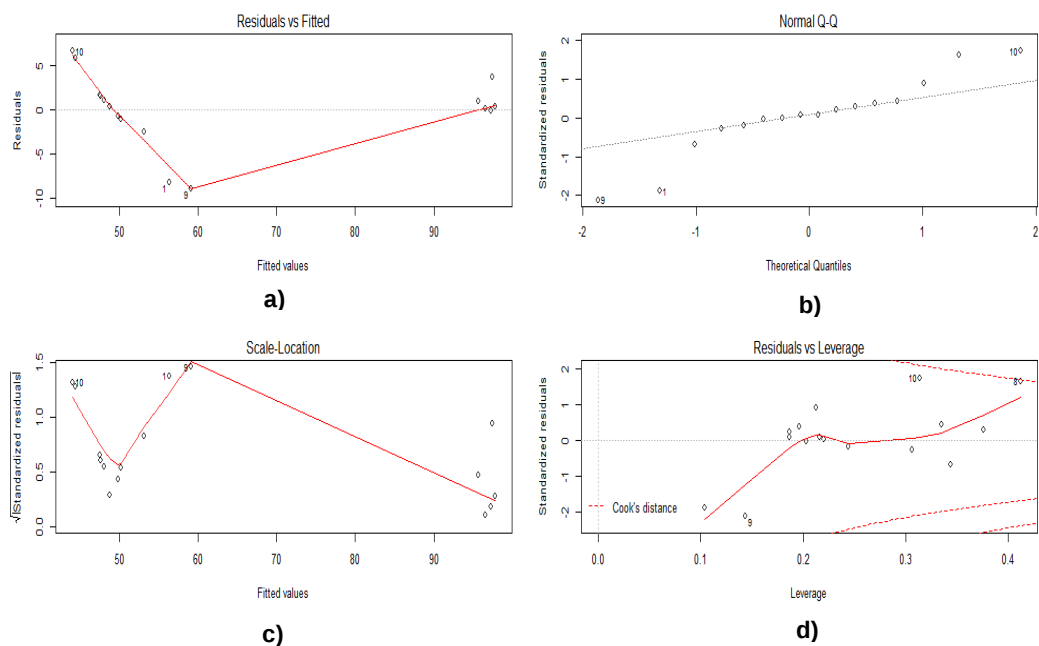


Figura 10-Análise de resíduos modelo de regressão linear múltipla do rolo móvel LA, Prensa 1: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.

- Rolo Móvel LOA

Para este rolo tem-se uma base de dados com 12 observações completas de 24-10-2018 a 14-02-2019 e 18 variáveis. A proporção é um pouco maior que uma observação por variável, o que ainda não permite um modelo maduro.

Foi criado um modelo com um $R^2_{\text{ajust}}=96.5\%$, que retorna os litros de atesto acumulado, em que os preditores são o TAN, o Teor de Alumínio, Teor de Cálcio e o Teor de Zinco:

$$\text{atesto acum} \approx 20.711 - 12.645 \cdot \text{TAN} - 0.843 \cdot \text{Al} + 0.031 \cdot \text{Ca} - 0.185 \cdot \text{Zn} \quad (33)$$

Para o valor de TAN=1 mgr KOH/gr, Al= 0 ppm, Ca=150 ppm e Zn= 20 ppm resulta um atesto acumulado de 8.950 litros, com intervalo de confiança a 95% igual a [3.866; 14.034].

Tabela 14 - Resultados finais para o modelo do rolo móvel LOA, da prensa 1.

Capacidade do Cáter:	45 litros
Volume de Atesto do Lubrificante Acumulado:	4 litros
Volume de Atesto do Lubrificante Recomendado:	8.950 litros (+4.950 litros)
Intervalo de Confiança para volume de atesto recomendado:	[3.866; 14.034]

Tal como nos casos anteriores é difícil analisar a qualidade do modelo com um número reduzido de observações. Neste caso, temos observações dispersas umas das outras, mas pode existir um padrão entre elas. Através das linhas de Cook é possível ver que existe um *outlier* (observação 1), mas mais uma vez é aconselhado uma segunda modelação com mais observações e só depois de confirmada a existência de *outlier* é que se deve retirá-lo. (Vd.Figura 11)

Quanto à normalidade dos resíduos, segundo o teste de *Shapiro-Wilk* apresenta um valor-p = 0.354, o que indica que não se rejeita a hipótese nula, para um nível de significância de 5%, ou seja, os resíduos têm distribuição normal.

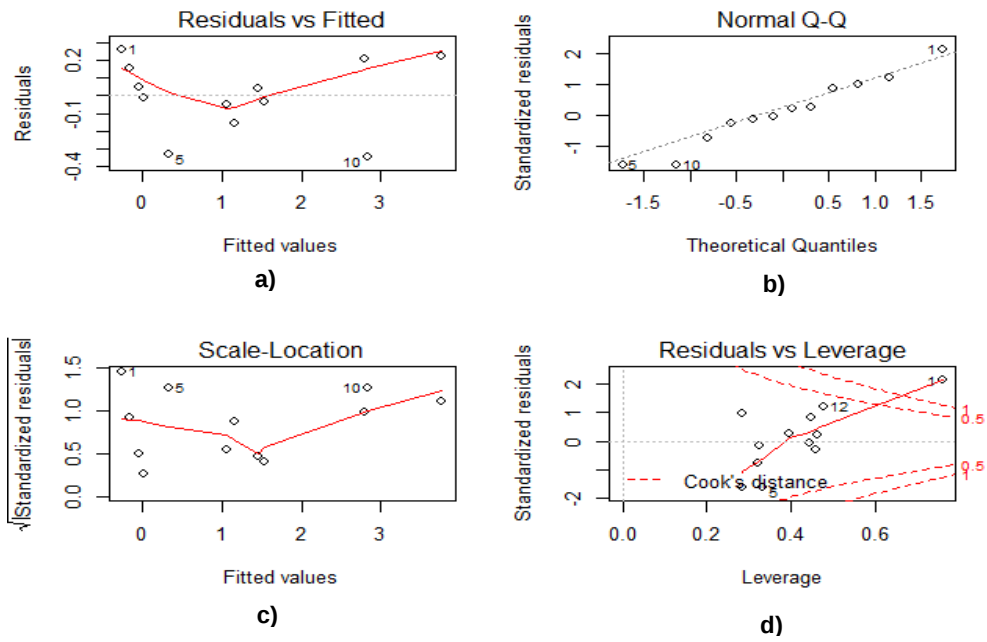


Figura 11-Análise de resíduos modelo de regressão linear múltipla do rolo móvel LOA, Prensa 1:
a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d)
Distância de Cook.

4.2. Prensa 3

- Rolo Fixo LOA

Para este rolo tem-se uma base de dados com 20 observações completas de 09-10-2018 a 14-03-2019 e 22 variáveis. A proporção é ligeiramente menor que uma observação por variável, o que não permite um modelo maduro.

Mesmo assim, foi criado um modelo com um $R^2_{\text{ajust}}=88.0\%$, que retorna as gramas de atesto acumulado, em que os preditores são o Teor de Alumínio, o Teor de Cobre, Teor de Manganês, três indicadores de desgaste mecânico. Aos quais se juntam, Teor de Sódio, indicador de contaminação externa e o desvio padrão do Binário aplicado na prensa (Bin_{dp}), indicador de produção e afinação da prensa. [22]

$$\text{atesto acum} \approx 11344 - 154.697 \cdot \text{Al} + 1575.157 \cdot \text{Cu} + 1438.595 \cdot \text{Mn} - 408.285 \cdot \text{Na} - 517.383 \cdot \text{Bin}_{\text{dp}} \quad (34)$$

Com este modelo é possível fazer uma simulação criando um cenário com valores limite aos preditores, $\text{Al}=100$ ppm, $\text{Cu}=3$ ppm, $\text{Mn}=9.5$ ppm e $\text{Na}=16$ ppm. Contudo sendo o desvio padrão do Binário, uma variável associada ao processo produtivo. As variáveis associadas ao processo produtivo dependem das necessidades e objetivos da própria fábrica, quem faz o diagnóstico do lubrificante em serviço não tem sentido crítico sobre o mesmo. Daí foram gerados três cenários diferentes dos quais resultam diferentes acumulados de atesto. A Tabela 15 apresenta-os, cabe ao cliente verificar o cenário que mais se ajusta à necessidade produtiva que se pretende simular, ou indicar outros valor concreto para este preditor.

Tabela 15-Previsão para modelo do rolo fixo LOA, da prensa 3

Capacidade do Cáster:	15000 gramas
Volume de Atesto do Lubrificante Acumulado:	10410 gramas

<u>Cenários de Desvio Padrão do Binário:</u>	$1^\circ Q=0.6$ MPa	$\text{Md}=1.0$ MPa	$3^\circ Q=2.3$ MPa
Atesto Acumulado (gramas)	7421 gramas (-2989 gr)	7216 gramas (-3194 gr)	6559 gramas (-3851 gr)
Intervalo de confiança	[4807;10036]	[4522;9911]	[3522;9596]

Ao se observar o gráfico de normalidade dos resíduos (Vd.Figura 12) e o teste de *Shapiro-Wilk*, que apresenta um valor- $p = 0.135$, logo não se rejeita a hipótese nula, para um nível de significância de 5%, os resíduos têm distribuição normal.

Reparando nos restantes gráficos, algumas das observações formam um pequeno grupo afastado das restantes observações que apresentam um ligeiro padrão.

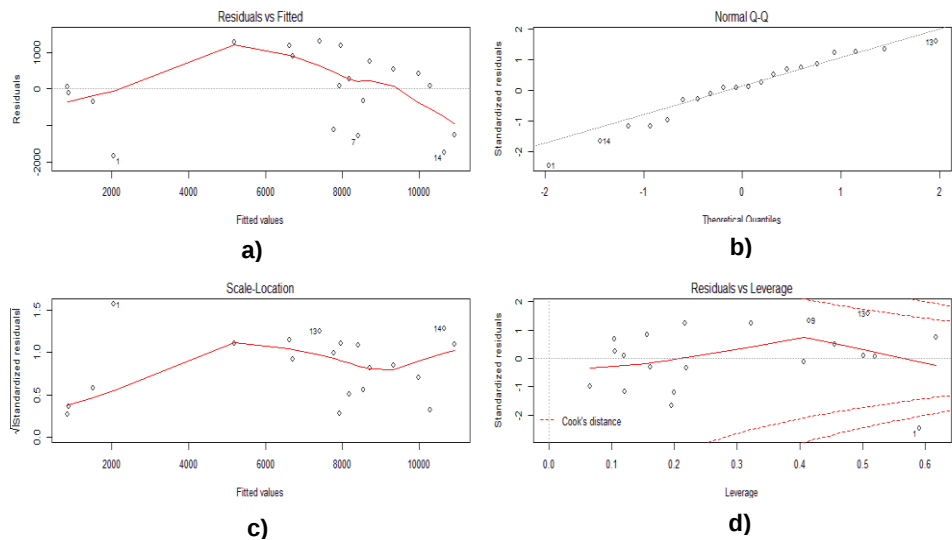


Figura 12-Análise de resíduos modelo de regressão linear múltipla do rolo fixo LOA, Prensa 3: a) Resíduos vs Valores Ajustados; b) Normalidade dos Resíduos; c) Escala dos Resíduos; d) Distância de Cook.

4.3. Discussão de Resultados Caso de Estudo

Neste caso de estudo, rolamentos de uma prensa de pasta de papel, é um exemplo prático como a operação de um equipamento impacta nas suas necessidades de lubrificação. A inclusão das variáveis Binário e Velocidade de Rotação, variáveis oriundas do processo produtivo, aparentemente encaixam-se na base de dados demonstrando importância para o modelo, pois apresentam correlação com a variável resposta.

A prensa de papel, se tiver um maior desvio padrão no seu binário, tipicamente é reflexo de períodos de menor produtividade e/ou paragens. Nesse sentido, as suas necessidades de renovação de lubrificante são menores, e os modelos representam o impacto dessa manutenção condicionada.

Não obstante, o acumulado de atesto de lubrificante real é mais frequentemente superior aos valores recomendados, isso significa que o equipamento não apresenta escassez de lubrificação. Se por outro lado, apontar para a um valor de acumulado inferior ao real, a recomendação é gerar essa reposição extra de lubrificante, para oferecer melhores condições de lubrificidade.

Apesar de alguns dos modelos realizados apresentarem um elevado valor de R^2_{ajust} , não significa que o modelo tenha significância, pois as bases de dados apresentam um reduzido número de observações, face ao número de variáveis, o que não é aconselhado para modelação. Outro facto importante é que os intervalos de confiança são muito alargados, o que significa que se deve promover a captação de mais observações, para que os modelos possam maturar

Também devido ao número reduzido de observações não é possível analisar completamente os gráficos de resíduos, pois as distribuições das observações não são precisas e os valores-p nos testes de normalidade são muito baixos.

5. Conclusão e Sugestões de Trabalho Futuro

Ao concluir o meu estágio sinto que os objetivos propostos foram atingidos, mas é importante salientar que as conclusões deste estágio são limitadas, visto que o tempo despendido não foi o suficiente para que houvesse um número elevado de observações de maneira a que se ajustasse modelos devidamente maturados.

Apesar disso, através do caso de estudo preliminar pude testar diferentes modelos estatísticos por forma a selecionar o mais indicado e cumprir os objetivos definidos.

Ao optar pelo modelo de regressão linear múltipla penso que consegui acabar com a limitação na ação de manutenção dos lubrificantes, assim consegui, para diversos equipamentos, situar no tempo o momento em que o lubrificante deixa de estar em condições para proteger os mesmos.

Após a conclusão do estágio, posso afirmar que os modelos de regressão linear múltipla são os que têm maior potencial para o projeto GALP MODE AIML, pois realizam melhores ajustes e fornecem os *outputs* desejados.

Não obstante, na minha opinião, com o avançar do projeto e por sua vez, o aumento do número de observações, é possível testar modelos de regressão logística, visto que estes preveem a probabilidade de mudança do lubrificante, que se assemelha à variável resposta inicial pretendida para o projeto (LSC).

O projeto GALP MODE AIML terá continuidade e gradualmente geram-se mais observações dos casos de estudo aqui referidos, assim sendo, se tivesse mais tempo, atualizaria os modelos a cada nova observação, enquanto os modelos não se consideram maduros o suficiente, pelo rácio de observações/variáveis. Também seria interessante aprofundar o estudo dos erros de ajuste e previsão, visto que o projeto lida com equipamentos críticos e para além de uma previsão do seu estado deveria ser incorporado o erro de determinação que os resultados numéricos têm.

Este estudo estatístico parece-me o próximo passo na evolução de manutenção proactiva, através da assessoria técnica de lubrificação, a clientes industriais, com equipamentos críticos de qualquer tipo.

Em geral, o meu balanço do estágio é muito positivo, pois constitui uma fase de grande crescimento para a minha formação quer pessoal como profissional, na medida em que foi um momento de ampliação e aprofundamento dos meus conhecimentos teóricos e práticos.

A oportunidade de contactar com o mundo profissional, bem como aprender novos conceitos técnicos sobre o acompanhamento de lubrificantes em serviço

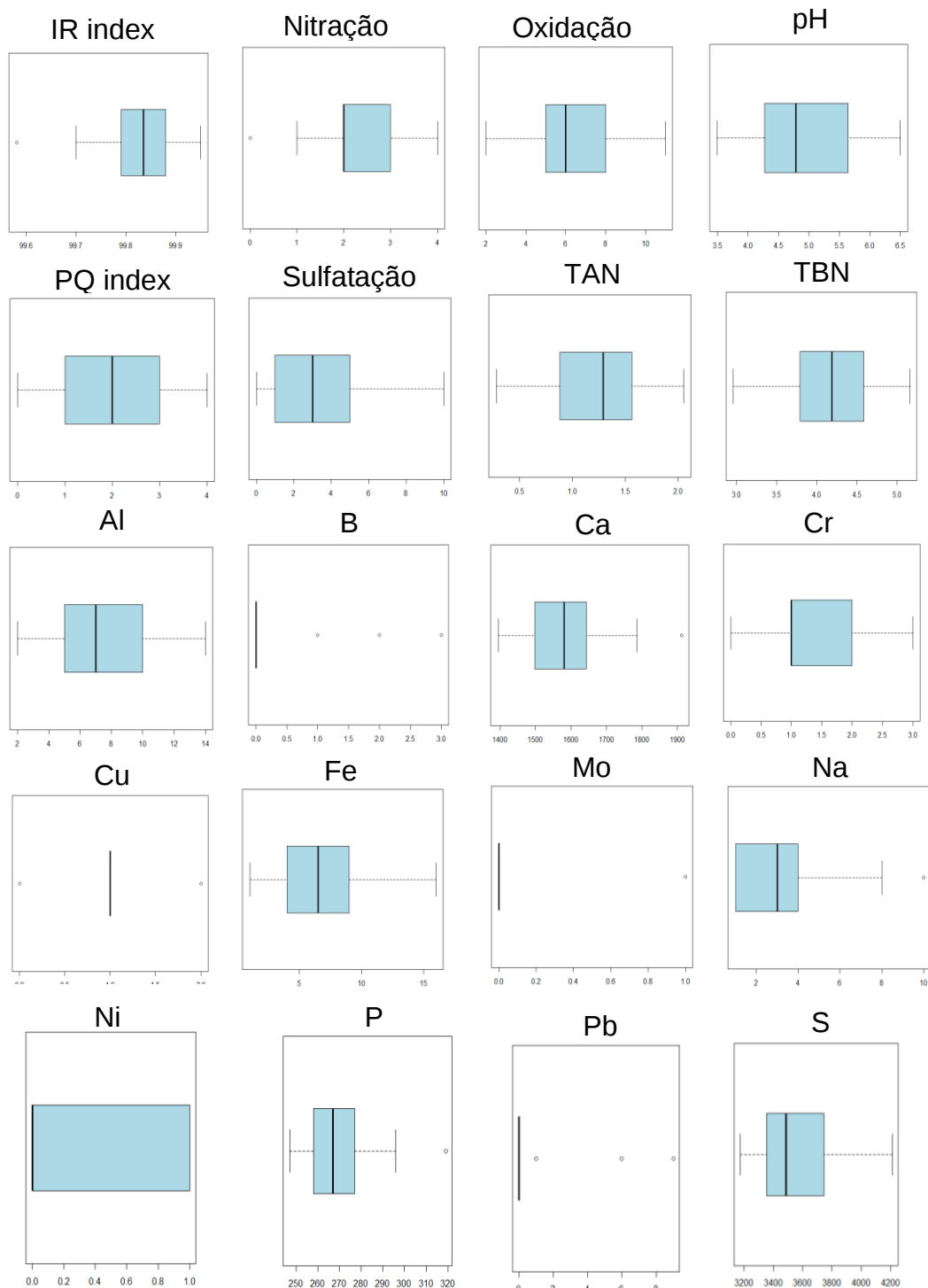
e a utilização de diferentes *softwares* informáticos foi muito importante para mim, pois permitiu-me reconhecer e ultrapassar as minhas limitações, proporcionando-me assim várias vivências para o exercício profissional futuro.

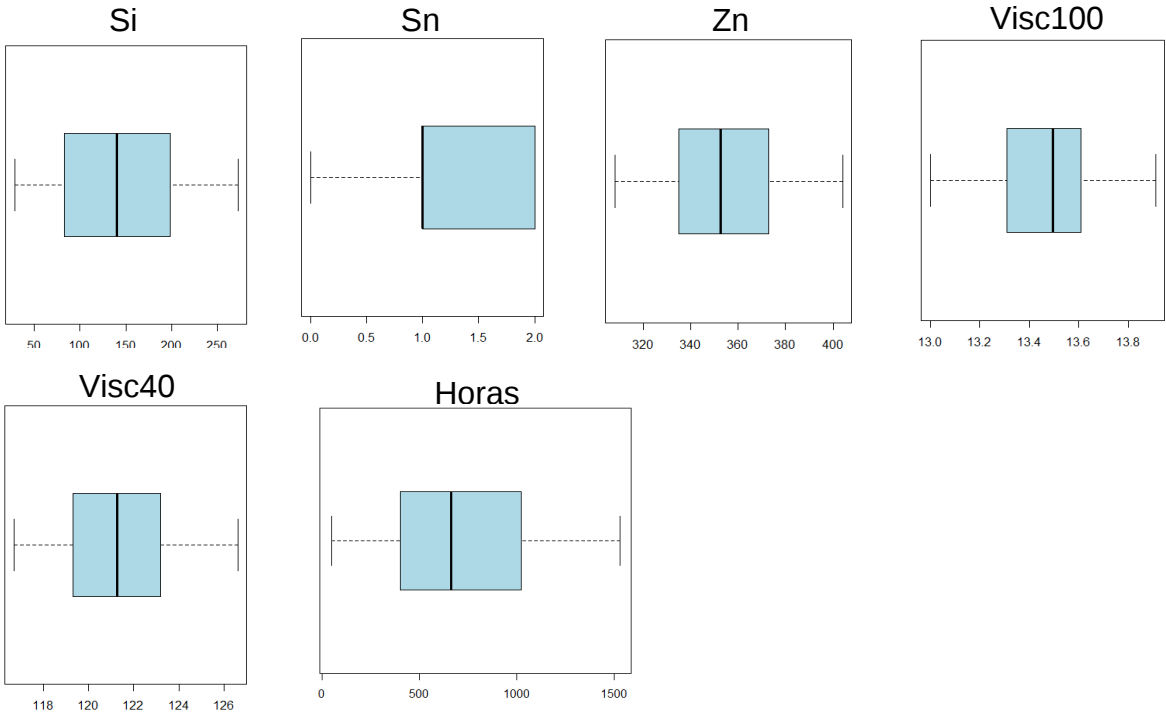
BIBLIOGRAFIA

- [1] P. Ferreira, C. Mota e E. Rocco, Data empiric modelling and principal component analysis in in-service lubricant monitoring for proactive maintance, 2018.
- [2] W. Rich, Lubricant Oxidation Analysis and Control, Maintance Reliability Group, Inc, 2000.
- [3] ASTM D664-11ae1, "Standart Test Method for Acid Number of Petroleum Productys by Pontentiometric Titration," *ASTM International, West Conshohocken, PA*, 2011.
- [4] J. Fitch e D. Troyer, Oil Analysis Basics, Tulsa: Noria Corporation, 2010.
- [5] M. S. Reis, Estatística Para a Melhoria de Processos - A Prespectiva Seis Sigmas, Coimbra: Imprensa da Universidade de coimbra, 2016.
- [6] D. Montgomery e E. & V. G. Peck, Introduction to Linear Regression Analysis, New Jersey: Wiley, 2012.
- [7] I. Jolliffe, Principal Component Analysis, New York: Springer, 2002.
- [8] D. C. Montgomery e G. C. Runger, "Chapter 12 Multiple Linear Regression," in *Applied Statistics and Probability for Engineers, 6th Edition*.
- [9] A. Agresti, Categorical Data Analysis, 3rd Edition, Wiley, 2013.
- [10] A. Dobson e A. Barnett, An Introdution to Generalized Linear Models, 3rd edition, Chapman & Hall, 2008.
- [11] D. Collett, Modelling Survival Data in Medical Research, 2nd Edition, Chapman & Hall, CRC Texts in Statistical Science.
- [12] Lubrication Management, PE-TA.071 Análise Quantitativa de Óleos por Infravermelhos, Bureau Veritas Group Company.
- [13] ASTM D445-17, "Standart Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity)," *ASTM International, West Conshohocken, PA*, 2017.

- [14] ASTM D1744-13, "Standart Test Method for Determination of Water in Liquid Petroleum Products by Karl Fischer Reagent (Withdrawn 2016)," *ASTM International, West Conshohocken, PA*, 2013.
- [15] Galp, Galp Power Gas NGB 40, Lisboa: FPC905, 2016.
- [16] S. Siegel e N. J. Castellan, Nonparametric statistics for the behavioral sciences, New York: McGraw-Hill.
- [17] J. D. Gibbons e S. Chakraborti, Nonparametric Statistical Inference, Boca Ration: CRC Press, 1985.
- [18] D. W. Hosmer e S. Lemeshow, Aplied Logistic Regression, 2nd Edition, Wiley, 2000.
- [19] Z. TZhang, Residuals and regression diagnostics: focusing on logistic regression, Ann Transl Med, 2016.
- [20] Galp, Galp OPS 3200, Lisboa: FPT524, 2011.
- [21] Galp, Galp Gemini Plex 0-1, Lisboa: FPC362, 2018.
- [22] ASTM D5185-18, Standart Test Method for Multielement Determination of Used and Unused Lubricating Oils and Base Oils by Inductively Coupled Plasma Atomic Emission Spectrometry (ICP-AES), West-Conshohocken, PA: ASTM International, 2018.
- [23] J. Zhu, J. M. Yoon, D. He, Y. Qu e E. Bechhoefer, Lubrication Oil Condition Monitoring and Remaining Useful Life Prediction with Particle Filtering, International Journal of Prognostics and Health Management, 2013.

ANEXO A- Boxplots (Caixa de Bigodes) dos dados do Caso de Estudo Preliminar





ANEXO B-Correlações entre Variáveis dos dados do Caso de Estudo Preliminar

	Água	Glicol	IRidx	Nitração	Oxidação	pH	PQidx	Sulfatação	TAN	TBN	Al	B	Ba	Ca	Cr
Água	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Glicol	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
IRidx	NA	NA		0,775	-0,871	0,796	0,054	-0,787	-0,795	0,775	-0,787	-0,070	NA	-0,431	-0,644
Nitração	NA	NA	-0,775		0,937	-0,881	0,136	0,581	0,893	-0,749	0,789	0,126	NA	0,222	0,699
Oxidação	NA	NA	-0,871	0,937		-0,940	0,023	0,721	0,922	-0,781	0,782	0,034	NA	0,305	0,639
pH	NA	NA	0,796	-0,881	-0,940		-0,036	-0,716	-0,897	0,735	-0,712	0,043	NA	-0,352	-0,551
PQidx	NA	NA	0,054	0,136	0,023	-0,036		-0,240	0,014	-0,080	-0,084	0,209	NA	-0,267	0,062
Sulfatação	NA	NA	-0,787	0,581	0,721	-0,716	-0,240		0,615	-0,560	0,582	0,174	NA	0,562	0,501
TAN	NA	NA	-0,795	-0,893	0,922	-0,897	0,014	0,615		-0,763	0,772	0,050	NA	0,202	0,619
TBN	NA	NA	0,775	-0,749	-0,781	0,735	-0,080	-0,560	-0,763		-0,668	0,294	NA	0,196	-0,618
Al	NA	NA	-0,787	0,789	0,782	-0,712	-0,084	0,582	0,772	-0,668		0,003	NA	0,268	0,869
B	NA	NA	-0,070	0,126	0,034	0,043	0,209	-0,174	0,050	-0,294	0,003		NA	-0,321	0,165
Ba	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ca	NA	NA	-0,431	0,222	0,305	-0,352	-0,267	0,562	0,202	-0,196	0,268	-0,321	NA		0,092
Cr	NA	NA	-0,644	0,699	0,639	-0,551	0,062	0,501	0,619	-0,618	0,869	0,165	NA	0,092	
Cu	NA	NA	-0,643	0,727	0,722	-0,655	-0,046	0,492	0,662	-0,560	0,799	-0,082	NA	0,177	0,702
Fe	NA	NA	-0,689	0,714	0,680	-0,569	0,020	0,474	0,664	-0,652	0,864	0,392	NA	0,056	0,863
Mo	NA	NA	-0,189	0,383	0,315	-0,255	0,227	0,092	0,315	-0,229	0,362	0,092	NA	-0,054	0,416
Na	NA	NA	-0,727	0,529	0,664	-0,636	-0,255	0,722	0,572	-0,554	0,664	0,168	NA	0,445	0,468
Ni	NA	NA	-0,449	0,355	0,456	-0,439	0,027	0,492	0,513	-0,378	0,500	0,171	NA	0,128	0,420
P	NA	NA	0,135	-0,104	-0,121	0,067	0,190	0,078	-0,113	0,196	-0,067	0,235	NA	0,262	-0,182
Pb	NA	NA	-0,172	0,066	0,099	-0,091	-0,135	0,218	0,130	-0,083	0,126	0,054	NA	0,091	0,130
S	NA	NA	0,111	-0,139	-0,225	0,182	-0,167	0,107	-0,181	0,107	0,097	-0,033	NA	0,217	0,153
Si	NA	NA	-0,876	0,893	0,932	-0,899	-0,055	0,714	0,899	-0,772	0,923	-0,047	NA	0,379	0,757
Sn	NA	NA	-0,722	0,745	0,718	-0,630	0,086	0,572	0,734	-0,669	0,878	0,036	NA	0,223	0,858
Zn	NA	NA	-0,380	0,275	0,294	-0,351	-0,366	0,529	0,310	-0,276	0,391	-0,221	NA	0,828	0,224
visc100	NA	NA	-0,595	0,591	0,712	-0,781	-0,109	0,721	0,643	-0,460	0,452	-0,352	NA	0,450	0,250
visc40	NA	NA	-0,722	0,677	0,760	-0,770	-0,026	0,612	0,713	-0,553	0,658	-0,292	NA	0,424	0,475
horas	NA	NA	-0,878	0,885	0,948	-0,920	-0,049	0,727	0,906	-0,786	0,863	-0,050	NA	0,378	0,694

	Cu	Fe	Mo	Na	Ni	P	Pb	S	Si	Sn	Zn	visc100	visc40	horas
Água	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Glicol	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
IRidx	-0,643	-0,689	-0,189	-0,727	-0,449	0,135	-0,172	0,111	-0,876	-0,722	-0,380	-0,595	-0,722	-0,878
Nitração	0,727	0,714	0,383	0,529	0,355	-0,104	0,066	-0,139	0,893	0,745	0,275	0,591	0,677	0,885
Oxidação	0,722	0,680	0,315	0,664	0,456	-0,121	0,099	-0,225	0,932	0,718	0,294	0,712	0,760	0,948
pH	-0,655	-0,569	-0,255	-0,636	-0,439	0,067	-0,091	-0,182	-0,899	-0,630	-0,351	-0,781	-0,770	-0,920
PQidx	-0,046	0,020	0,227	0,255	0,027	0,190	0,135	-0,167	-0,055	-0,086	-0,366	-0,109	0,026	-0,049
Sulfatação	0,492	0,474	0,092	0,722	0,492	0,078	0,218	0,107	0,714	0,572	0,529	0,721	0,612	0,727
TAN	0,662	0,664	0,315	0,572	0,513	-0,113	0,130	-0,181	0,899	0,734	0,310	0,643	0,713	0,906
TBN	-0,560	-0,652	-0,229	-0,554	-0,378	0,196	-0,083	0,107	-0,772	-0,669	-0,276	-0,460	-0,553	-0,786
Al	0,799	0,864	0,362	0,664	0,500	-0,067	0,126	0,097	0,923	0,878	0,391	0,452	0,658	0,863
B	-0,082	0,392	0,092	0,168	0,171	0,235	0,054	-0,033	-0,047	0,036	-0,221	-0,352	-0,292	-0,050
Ba	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ca	0,177	0,056	-0,054	0,445	0,128	0,262	0,091	0,217	0,379	0,223	0,828	0,450	0,424	0,378
Cr	0,702	0,863	0,416	0,468	0,420	-0,182	0,130	0,153	0,757	0,858	0,224	0,250	0,475	0,694
Cu		0,669	0,232	0,585	0,453	-0,059	0,067	0,095	0,809	0,712	0,268	0,547	0,738	0,772
Fe	0,669		0,458	0,498	0,373	-0,187	0,146	0,087	0,766	0,749	0,177	0,180	0,391	0,707
Mo	0,232	0,458		-0,031	0,097	-0,295	-0,113	-0,022	0,297	0,354	-0,116	-0,001	0,086	0,232
Na	0,585	0,498	-0,031		0,368	-0,016	0,097	0,100	0,742	0,539	0,562	0,594	0,584	0,746
Ni	0,453	0,373	0,097	0,368		0,083	0,152	0,140	0,508	0,539	0,116	0,396	0,399	0,511
P	-0,059	0,187	0,295	-0,016	0,083		0,124	0,388	-0,093	-0,071	0,313	0,028	-0,067	-0,135
Pb	0,067	0,146	-0,113	0,097	0,152	-0,124		0,065	0,130	0,131	0,122	0,072	0,097	0,176
S	0,095	0,087	-0,022	0,100	0,140	0,388	0,065		-0,032	0,141	0,418	-0,062	-0,078	-0,081
Si	0,809	0,766	0,297	0,742	0,508	-0,093	0,130	-0,032		0,815	0,448	0,686	0,814	0,985
Sn	0,712	0,749	0,354	0,539	0,539	-0,071	0,131	0,141	0,815		0,332	0,396	0,554	0,758
Zn	0,268	0,177	-0,116	0,562	0,116	0,313	0,122	0,418	0,448	0,332		0,431	0,408	0,426
visc100	0,547	0,180	-0,001	0,594	0,396	0,028	0,072	-0,062	0,686	0,396	0,431		0,820	0,729
visc40	0,738	0,391	0,086	0,584	0,399	-0,067	0,097	-0,078	0,814	0,554	0,408	0,820		0,825
horas	0,772	0,707	0,232	0,746	0,511	-0,135	0,176	-0,081	0,985	0,758	0,426	0,729	0,825	

ANEXO C- Gráficos de dispersão entre Variáveis dos dados do Caso de Estudo Preliminar

