08, 09, 10 e 11 de novembro de 2022
ISSN 2177-3866

Técnicas de machine learning como auxílio para a construção e comparação de modelos de previsão do processo de revestimento e liberação de tubos de aço.

LUCAS ALVES DIAS CARDOSO
ESCOLA SUPERIOR DE AGRICULTURA "LUIZ DE QUEIROZ" - USP

FERNANDO FREIRE VASCONCELOS
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

ANDRÉ LEME DA SILVA FLEURY BONINI
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

TÉCNICAS DE *MACHINE LEARNING* COMO AUXÍLIO PARA A CONSTRUÇÃO E COMPARAÇÃO DE MODELOS DE PREVISÃO DO PROCESSO DE REVESTIMENTO E LIBERAÇÃO DE TUBOS DE AÇO.

1 Introdução

Um único fenômeno de estudo pode ser analisado de múltiplas formas e através de perspectivas diversas. Como consequência, valores estimados ou previstos por diferentes técnicas de modelagem podem também ser diferentes. Estas divergências não são necessariamente um problema, uma vez que o intuito dos modelos é representar a realidade de forma simplificada, de modo a se obter a melhor semelhança possível entre os valores observados na realidade e os valores previstos através das modelagens, da forma que melhor atender aos anseios ou resolver as problemáticas, dentro do contexto no qual o estudo é realizado (Silberzahn e Uhlmann, 2015).

Sendo assim, é razoável concluir que os estudos realizados anteriormente acerca de certos fenômenos, podem e devem ser revisitados, aprimorados e atualizados, sob novas perspectivas.

O presente estudo parte inicialmente da observação direta da produção dos tubos de aços e das falhas verificadas em seu processo em uma indústria de médio porte. Destaco que o método observacional é tido como um dos mais modernos, visto que possibilita um maior grau de precisar do problema a ser estudado, especialmente nas ciências sociais (Gil, 2019). Após, procedeu-se com a coleta de dados secundários para montagem do modelo de dados.

Este trabalho foi realizado numa empresa que produz tubos de aço, tendo por objeto a observação da produção de um módulo tubular cujas dimensões são aproximadamente 7 metros de comprimento e 2,1 metros de diâmetro, com paredes que medem cerca de 1,27 centímetros de espessura. Os tubos foram produzidos através de um processo de enrolamento helicoidal. Neste processo de produção, uma bobina de aço é curvada em uma certa angulação, de modo que os lados das bobinas são soldados em outra porção dela mesma, formando assim um objeto cilíndrico com solda helicoidal. A bobina é cortada quando o produto assume o comprimento desejado. Após formado, o tubo passa por uma série de testes de resistência a tensão e pressão. Se aprovado nestes testes, o produto passa por um processo de revestimento polimérico, que protege o aço de oxidação, prolongando sua vida útil. Esta etapa requer um período de espera para secagem e cura, para que só então, depois de uma última avaliação do tubo e do revestimento, o item seja liberado para a entrega ao cliente (Lima, 2008).

Neste processo, foram constatadas altas divergências diárias entre as quantidades de produtos que eram de fato entregues ao cliente e as quantidades que eram planejadas, levando em consideração tão somente a produtividade da fábrica. Visto que o volume de material a ser produzido e entregue periodicamente ao cliente é algo definido contratualmente, o recorrente problema de entregas de quantidade a menor tem também consequências contratuais, na forma de multas.

Sendo assim, para o prosseguimento do projeto, assim como para os futuros contratos de produção da empresa em questão, se faz necessária a aplicação de técnicas que proporcionem um planejamento mais viável, mais próximo do que é realmente possível entregar, reduzindo as chances de se superestimar a produtividade, e consequentemente, evitando futuras divergências e multas contratuais.

O evento estudado foi a etapa final da produção de tubos de aço revestidos com polímeros, que consiste na secagem e cura do revestimento. Esta etapa foi identificada como gargalo do processo de produção dos tubos (ou seja, fase de produção com menor produtividade dentre todas as fases, limitando assim a eficiência de toda a produção). Este processo deve ocorrer sob condições de temperatura e umidade adequados (Lima, 2008). Entretanto, na

empresa em questão, estes processos acontecem em local pouco protegido das possíveis interferências climáticas, que ocasionam atrasos e a necessidade de retrabalho. Alertas de raio, por exemplo, interrompem completamente as atividades do setor estudado. Embora isto ocorra por razões de segurança importantes e reduza as chances de fatalidades, acaba por impedir a movimentação de produtos, a realização de reparos, e até mesmo o próprio carregamento do material a ser transportado para o cliente, resultando assim num significativo impacto negativo no que se refere à quantidade de produtos liberados e entregues.

No momento da realização da coleta de dados, não havia a possibilidade de evitar as influências climáticas (temperatura, umidade, precipitação, descargas elétricas atmosféricas, etc.) no processo estudado, face a natureza da indústria estudada. Por outro lado, se faz necessário compreender, da melhor forma possível, como tais influências dos efeitos climáticos se dão, para que seja possível considerá-los, juntamente com outros fatores, no momento de estimar e planejar a produção e a entrega de produtos.

Portanto, este estudo testou a hipótese de que a liberação de tubos sofreu a influência de fatores relacionados às condições climáticas, assim como da quantidade de funcionários diariamente alocados na realização do processo em questão. Além disso, e principalmente, experimentou a criação de agrupamentos de observações do evento estudado como uma forma de melhor entender os contextos do evento, para que fosse possível proporcionar previsões de produção mais exatas e precisas.

Como contribuição teórica verificou-se que a utilização de técnicas de clusterização antes da montagem dos modelos de previsão melhora bastante sua acurácia, enquanto como contribuição gerencial verificamos a melhora a capacidade preditiva da produção da indústria estudada, inclusive com a possibilidade de economia efetiva de alguns milhares de unidades monetárias em custos decorrentes da mora ou demora no cumprimento dos contratos de fornecimento.

2 Descrição do processo de fabricação dos tubos de aço na empresa estudada

No processo de produção dos tubos de aço uma bobina é curvada em uma certa angulação, de modo que os lados das bobinas são soldados em outra porção dela mesma, formando assim um objeto cilíndrico com solda helicoidal. A bobina é cortada quando o produto assume o comprimento desejado. Após a realização de testes de pressão e resistência, o produto passa por um processo de revestimento polimérico, que protege o aço de oxidação, prolongando sua vida útil, que necessita de um tempo de secagem. (Lima, 2008).

O processo de revestimento, última etapa da produção e extremamente necessário para proteger o aço e aumentar sua durabilidade, requer um período de secagem e cura, no qual o produto deve ser mantido imóvel sob determinadas condições de temperatura e umidade. Tais condições, quando não respeitadas, provocam o aumento do período para que a secagem ocorra. Em casos mais severos, os tubos que não foram mantidos das formas adequadas necessitaram de reparos em seu revestimento, ou seja, que o revestimento fosse reaplicado de forma local, e um novo período de secagem e cura tivesse que ser esperado antes que, finalmente, o produto fosse liberado para a entrega.

O processo de secagem e cura do revestimento dos tubos ocorreu numa área do pátio externo da fábrica, sob uma tenda, denominada como “tenda de liberação”.

Há fatores relacionados às condições às quais os produtos estavam submetidos quando na tenda de liberação, que possivelmente promoveram efeitos na quantidade de produtos liberados.

As especificações técnicas do produto polimérico aplicado como revestimento instruía que sua utilização se desse em superfícies que estivessem a 3º Celsius acima do Ponto de Orvalho, para que sejam evitadas as formações de gotículas de água (Renner Coatings, 2015).

Tais gotículas, quando formadas, ficam aprisionadas entre a superfície de aço e o revestimento polimérico, formando bolhas, prejudicando assim a qualidade do produto.

Como a tenda de liberação fornecia apenas proteção parcial do ambiente, o processo muitas vezes era inviabilizado, ou era realizado mesmo em condições inadequadas, o que frequentemente resulta em defeitos e na necessidade de reparos, que atrasavam ainda mais a liberação.

Os alertas de descargas elétricas atmosféricas (alertas de raio) foram a causa de interrupção total das atividades no local, por motivo de segurança, impossibilitando qualquer manipulação do produto, inclusive no sentido de liberá-lo, bem como a quantidade de funcionários no local influenciaram decisivamente no tempo para finalização dos tubos.

3 Metodologia

Este trabalho consistiu numa de pesquisa explicativa e descritiva, de natureza aplicada, utilizando o método da observação direta e levantamento de dados primários no período de setembro de 2015 até agosto de 2016, no estudo das quantidades de produtos liberados para entrega ao cliente, numa indústria metalúrgica localizada em Pindamonhangaba-SP. Os produtos em questão foram tubos de aço (de dimensões: 7 metros de comprimento, 2,1 metros de diâmetro e 1,27 centímetros de espessura), que passaram por um processo de revestimento em toda a sua extensão, externa e internamente.

O evento estudado foi a etapa final da produção de tubos de aço revestidos com polímeros, que consiste na secagem e cura do revestimento. Esta etapa foi identificada como gargalo do processo de produção dos tubos (ou seja, fase de produção com menor produtividade dentre todas as fases, limitando assim a eficiência de toda a produção)

A quantidade de funcionários alocados na tenda de liberação em cada turno também foi uma variável sobre a qual esperava-se que exercesse uma influência direta na quantidade de tubos liberados, uma vez que que determinava, em grande parte das vezes, o volume de serviço que era realizado no período.

Deste modo, a partir da constatação de aspectos como a situação de exposição ao ambiente no qual o evento estudado acontecia (o que por vezes ia de encontro aos requerimentos técnicos para a aplicação do revestimento), e a variação da quantidade de funcionários, notou-se que era possível se estabelecer relações entre: quantidade de tubos liberados (variável dependente), e as variáveis: Temperatura, Umidade Relativa do Ar; Ponto de Orvalho; Mão de Obra; Incidência de Raios (variáveis explicativas), sendo estas as variáveis coletadas.

Além disso, foram utilizadas técnicas de clusterização para categorizar as observações em diferentes grupos, que representaram contextos latentes nos quais ocorreram os eventos. Tais grupos, após definidos, foram representados por uma variável categórica, onde as observações semelhantes entre si fizeram parte de um mesmo grupo, e os grupos contiveram observações significativamente diferentes das observações presentes nos demais grupos (Ochi et al., 2004).

Deste modo, foi definida como variável resposta do presente estudo a quantidade de tubos liberados pelo setor de reparos no dia como y (número de tubos) e como variáveis explicativas: temperatura ambiente do dia como x_T (ºC); umidade relativa do ar do dia como x_U (%); ponto de orvalho do dia como x_P (ºC); volume de mão-de-obra do dia como x_M (funcionários/ dia); alerta de incidência de raios no dia como x_R (variável qualitativa, definida como 1 em caso da ocorrência de alerta ou 0 no caso da não ocorrência); e grupos de observações como x_C (variável qualitativa, definida a partir das técnicas de clusterização)

Como resultado da coleta dos dados e da criação dos grupos, foi construído o banco de dados que foi utilizado na construção dos modelos supervisionados de *machine learning*, no qual as colunas referem-se às variáveis, e as linhas são referentes às observações (ou dias).

Tabela 1. Banco de dados coletados

Observação/Dia	y_i (tubos)	x_{iT} (ºC)	x_{iU} (%)	x_{iP} (ºC)	x_{iM} (trabalhadores)	x_{iR} (1 ou 0)	x_{iC} (grupo/cluster)
1	y_1	x_{1T}	x_{1U}	x_{1P}	x_{1M}	x_{1R}	x_{1C}
2	y_2	x_{2T}	x_{2U}	x_{2P}	x_{2M}	x_{2R}	x_{2C}
3	y_3	x_{3T}	x_{3U}	x_{3P}	x_{3M}	x_{iR}	x_{iC}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	x_{nT}	x_{nU}	x_{nP}	x_{nM}	x_{nR}	x_{nC}

Fonte: Dados da pesquisa

A clusterização se deu tanto por técnicas hierárquicas, tais como: “Single Linkage” (Vizinho mais próximo), “Complete Linkage” (Vizinho mais distante), “Average Linkage” (Média entre as distâncias) e “Ward‘s Method” (Mínima variância total intra-cluster), mediante um processo progressivo de medir as distâncias ou a variância no método de Ward entre elementos que pertencem ao cluster e os elementos candidatos a fazer parte deste cluster, segundo Rokach e Maimon (2005). Enquanto isso, como método não hierárquico foi utilizado o “k-means” que se dá através da distribuição aleatória de um número k de centroides, a partir dos quais vão sendo associados elementos mais próximos, para formar uma quantidade k de clusters (Bock, HH., 2007). Ressalta-se que a natureza dos dados não permitiu a identificação prévia de um destes métodos que fosse mais adequado que os demais, razão pela qual foram realizados testes das diferentes técnicas de clusterização, com o intuito de definir a que contribuiu para melhores resultados dos modelos de regressão.

Não foi utilizado neste trabalho o método de Clusterização Espacial Baseada em Densidade de Aplicações com Ruído (*Density Based Spatial Clustering of Application with Noise - DBSCAN*), uma vez que esta técnica tem base na densidade estrutural dos dados, e não na redução da variabilidade, de modo que não fez sentido sua aplicação para o estudo em questão (Arlia e Coppola, 2001).

Os modelos foram construídos a partir das seguintes técnicas de Modelos Lineares Generalizados (GLM) de regressão: Linear Múltipla (LM); Logística Múltipla (LgM); Poisson; Binomial Negativa ou Poisson-Gama (NB); Poisson com inflação de Zeros (ZIP) e Poisson-gama com inflação de Zeros (ZINB). O método de regressão linear múltipla, tem como base a relação entre uma variável dependente y e as k variáveis independentes (x_1, x_2, …, x_k), através de uma equação linear, sendo y a quantidade de tubos liberados pela empresa no dia (Montgomery e Runger, 2011). A regressão logística múltipla, por sua vez, é adequada para situações dicotômicas, de respostas binárias, como por exemplo “sim ou não”, “sucesso ou fracasso”. Como esta variável resposta não é numérica, é considerada a relação linear entre as variáveis independents (x_1, x_2, …, x_k) e o logaritmo natural da razão entre as probabilidades de evento e de não-evento, o que é chamado de logito. Desta forma, será estudada a relação entre a probabilidade de liberação de tubo (evento) com a de não-liberação (não-evento), no dia (Montgomery e Runger, 2011). Os modelos Poisson, Poisson-gama e “Zero-Inflated” (inflacionado de zeros) são métodos que levam em consideração dados de contagem (no caso, a contagem de tubos liberados). Entretanto, o modelo Poisson-gamma leva em consideração a superdispersão dos dados (que será testada neste estudo). Já a técnica “Zero-Inflated” pressupõe uma componente logística para explicar as altas incidências do valor “zero” na variável dependente, ou seja, dias nos quais não houve liberação de tubos (Lord et al., 2005).

Os modelos construídos foram estimados pelo método de máxima verossimilhança (“maximum-likelihood estimation [MLE]”), e foram, portanto, comparados pelos seus valores de “Log Likelihood [LogLik]”. Espera-se que, quanto maior o valor do LogLik de um certo modelo, mais este é adequado para a realização de previsões, lembrando que os valores de LogLik se apresentam de forma negativa, de modo que o maior valor é o valor (negativo) que

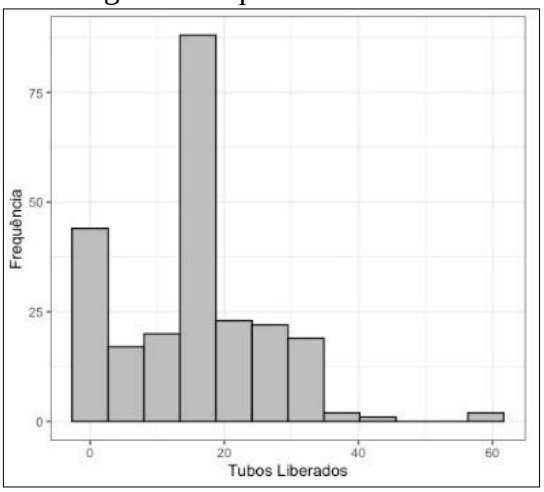
mais se aproxima do zero, ou seja, o menor valor em módulo (Hosmer e Lemeshow, 2000; Long, 1997).

Para realizar o trabalho foi utilizada a linguagem de programação R 3.3.0, a partir da interface do Rstudio IDE.

Resultados e Discussão

Primeiramente, foi realizada uma análise exploratória dos dados, para que fosse possível ter um mínimo entendimento do comportamento do fenômeno estudado, e principalmente para se entender a divergência entre a previsão da empresa e a quantidade real de produtos liberados. Inicialmente, através de um histograma (Figura 1), foi possível analisar as frequências das quantidades de tubos liberados.

Figura 1. Histograma de quantidades de tubos liberados.



Fonte: Resultados originais da pesquisa.

A partir do histograma, pôde-se perceber que, apesar de a maior frequência se dar na faixa dos 16 tubos (foram 88 eventos nesta faixa), a incidência de dias com zero liberações é relativamente alta. Com 44 eventos, dias sem liberação de tubos representaram a segunda maior frequência. Tal fenômeno foi um indicativo para a adequação de modelos de contagem que levam em consideração a inflação de zeros.

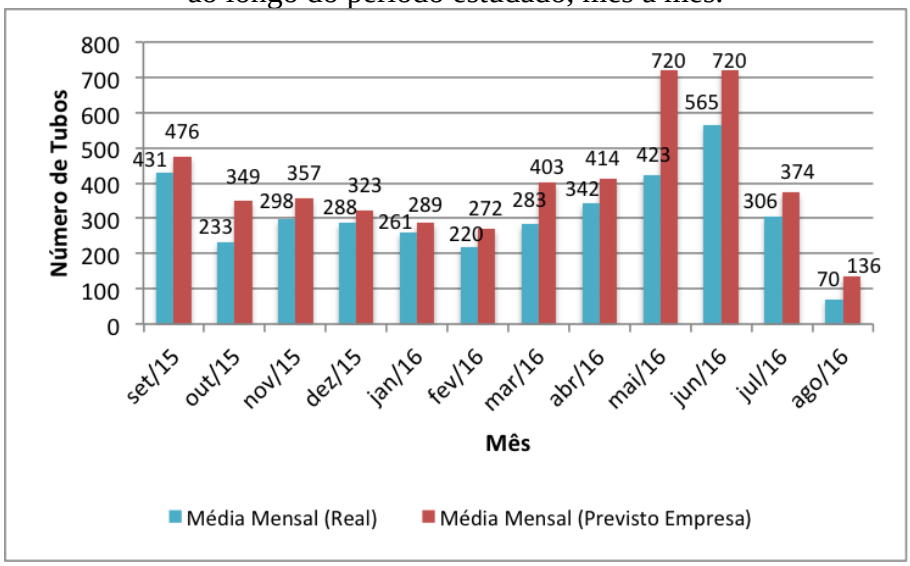
Uma vez que, durante a realização deste estudo, a empresa não previu nenhum evento de zero liberações de tubos, este foi um fator de evidente impacto na divergência entre os valores previstos e os reais, medida que foi aspecto motivador deste estudo. O diagrama abaixo (Figura 2) mostra precisamente a diferença entre o número de produtos real (Y_Real), e o número de produtos previstos pela empresa (Y_Prev_Empresa).

Com base na Figura 2 é possível observar de forma agregada por mês uma clara diferença entre o comportamento do fenômeno real, e a forma como foi previsto pela empresa, o que explica os valore de divergência constatados mês após mês.

Pode-se constatar que houve maior divergência nos meses de maio e junho, sendo estas de 297 e 155 tubos a menos, respectivamente. Estes foram também os meses nos quais a empresa previu uma maior produção, de acordo com os critérios internos e de demanda contratual.

Durante todo o período do estudo, a empresa previu um volume de produção de 4833 tubos, porém a produção real foi de apenas 3720 tubos. Ou seja, com 1113 tubos a menos, a empresa produziu aproximadamente 77% do que era previsto.

Figura 2. Comparação entre quantidade de tubos liberados pela empresa vs. quantidade real, ao longo do período estudado, mês a mês.



Fonte: Resultados originais de pesquisa

Além do simples fato disto representar um déficit próximo de 23% em seu planejamento (ou seja, na capacidade da empresa prever o quanto vai efetivamente produzir e entregar), os impactos dessa divergência também se dão de forma financeira, através da diferença de faturamento, entre o que era esperado e o real. Os cálculos desta diferença foram realizados através de valores estimados no momento da coleta de dados para este estudo, de modo que se possa entender o valor (preço) do produto.

Levando-se em consideração: o volume médio de aço empregado no tubo, de aproximadamente 0,593m³, a densidade média do tubo, 7,8 ton/m³ (Eurokation, 2017), o valor do aço, avaliado em US$ 442,00 por tonelada (Abacusliquid, 2017), e o dólar valendo em média R$ 3,67 (UOL Economia, 2017) à época, o material empregado na fabricação de cada tubo foi estimado em R$ 7.502,40.

Além do material, existe toda uma cadeia que transforma e agrega valor ao produto (energia, maquinário, mão de obra, e etc.). Estes valores, combinados com a própria margem de lucro necessária para a venda, foram os pressupostos para que fosse estipulado um acréscimo de 300% no valor do material bruto, para se obter o valor estimado de venda de cada produto, de aproximadamente R$ 30.000,00.

A partir do valor base calculado, pôde-se estimar uma divergência de faturamento do período. Deixar de produzir 1113 que haviam sido planejados resultou num faturamento menor que o planejado, com uma diferença de mais de 33 milhões de reais. (R$ 33.390.000,00).

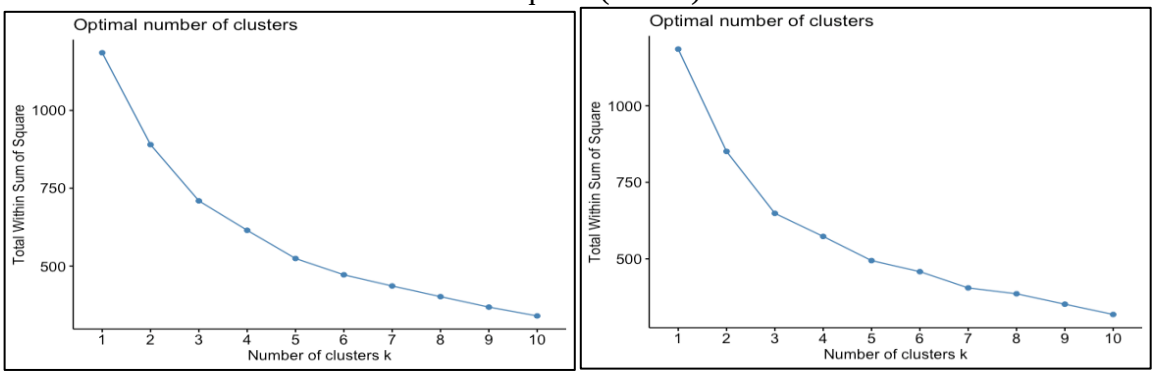
Além deste valor, também incidiram encargos contratuais, uma vez que o cliente precisava também cumprir suas próprias demandas, de modo a depender da entrega conforme havia sido determinada antes mesmo do início do projeto. O não cumprimento, ou seja, a não entrega da quantidade de produtos acordada na periodicidade definida, acarretou multas, previamente definidas em cláusulas contratuais.

Sendo assim, o impacto que a companhia teve por causa de técnicas imprecisas de previsão e planejamento da própria produção se deu, principalmente, pela soma destes dois aspectos: a diferença entre o que se esperava de faturamento e o que realmente foi faturado no período (diferença de cerca de 33 milhões de Reais); e dos valores das multas contratuais aplicadas no período (cujos valores não puderam ser apresentados no presente estudo para que não fossem reveladas informações confidenciais de contrato e de negociação, tanto da própria companhia quanto dos seus clientes).

Para testar a hipótese de que os métodos de clusterização, quando utilizadas em conjunto com as técnicas de Machine Learning podem melhorar suas capacidades preditivas, foram realizadas clusterizações por cinco técnicas diferentes, sendo quatro delas hierárquicas (Single Linkage, Complete Linkage, Average Linkage e Ward's Method), e uma não hierárquica (k-means).

Para definir as quantidades de clusters a serem criados, foram montadas curvas Elbow (Figura 3), sendo a da esquerda para os métodos hierárquicos, e a da direita para o não hierárquico

Figura 3 - Curva Elbow dos modelos de clusterização hierárquicos (esquerda) e do não hierárquico (direita).

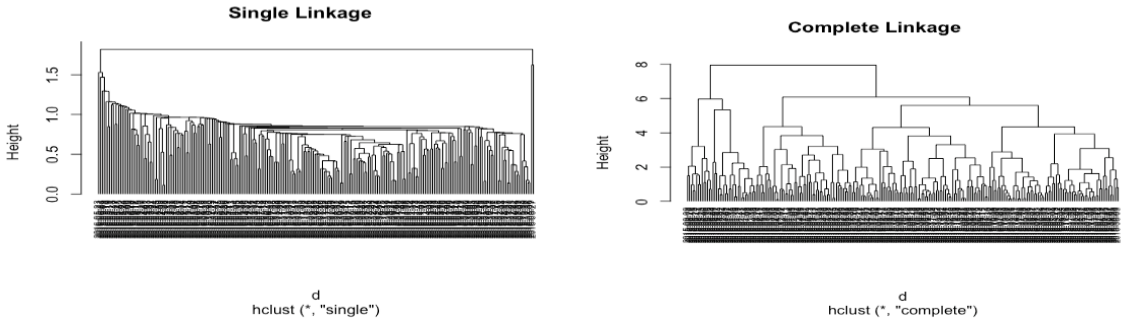


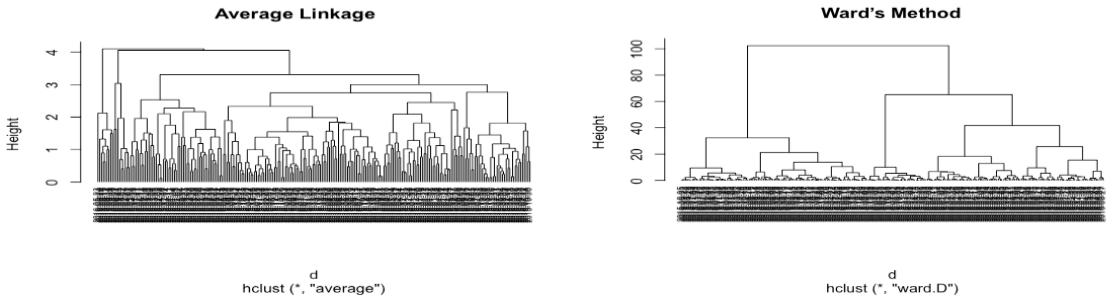
Fonte: Resultados originais da pesquisa.

Nas curvas elbow, a variabilidade dos dados é representada no eixo y, enquanto o número de clusters é mostrada no eixo x. Desse modo, é possível se observar o quanto de variabilidade dos dados é capturada a partir da definição de uma certa quantidade de grupos. Percebe-se que as curvas mostradas têm uma grande variação nas suas variabilidades para o intervalo de dois a cinco grupos, e após este ponto, começa diminuir a sua inclinação e ficar numa posição mais próxima ao horizontal, ou seja, diminui a variação na variabilidade dos dados com a adição de mais do que cinco clusters. Sendo assim, o número de clusters definido será de cinco.

Sendo estabelecido o número de clusters, foram feitos os cinco procedimentos de clusterização. Para observar o comportamento dos dados com relação aos procedimentos de clusterização, os grupos formados a partir das técnicas hierárquicas deram origens aos dendogramas, mostrados a seguir na Figura 4 a seguir:

Figura 4 – Dendrogramas: Métodos de Clusterização Single Linkage, Complete Linkage, Average Linkage e Ward's Method.



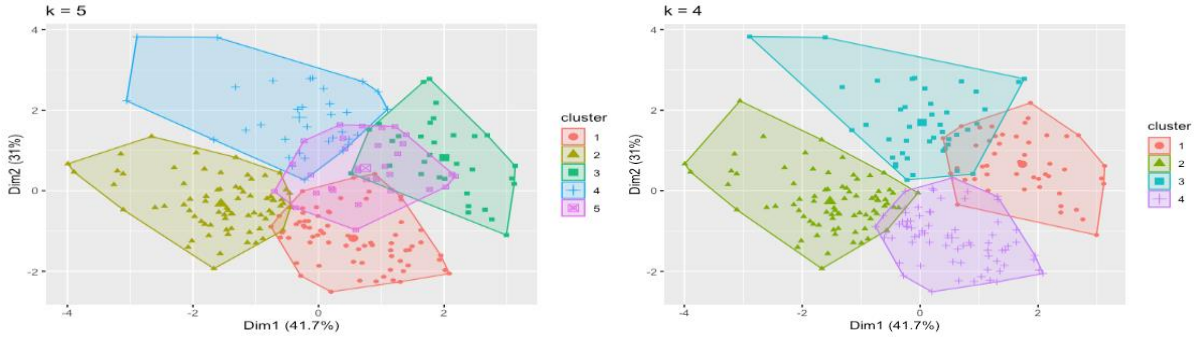


Fonte: Resultados originais da pesquisa.

É possível se observar que as clusterizações feitas pelos métodos Complete Linkage (superior, direita) e Ward's Method (inferior, direita) mostram uma separação em grupos de forma mais clara, enquanto as clusterizações por Single Linkage e Average Linkage (superior, esquerda e inferior, esquerda) parecem não fazer uma separação de grupos que interfiram de forma clara na explicação da variabilidade entre as observações.

Já para visualizar os grupos formados através do método de clusterização não hierárquica k-means, os grupos foram mostrados em gráficos de dispersão, que podem ser observados na Figura 5:

Figura 5. Dispersão – Método de Clusterização K-means com cinco grupos (esquerda) e quatro grupos (direita).



Fonte: Resultados originais da pesquisa.

Notou-se que, na representação bidimensional da clusterização com cinco clusters (esquerda), há um cluster que se sobrepõe a partes de todos os demais. Sendo assim, como teste, a clusterização pelo método k-means foi feita também com quatro grupos (direita). Com apenas quatro clusters, foi possível observar uma maior separação entre os grupos, ainda havendo pequenas sobreposições. É importante lembrar que, nesta visão bidimensional, observa-se a explicação de somente 72,7% da variabilidade dos dados (sendo 41,7% na Dimensão 1, somados a 31% da Dimensão 2). Sendo assim, pode-se supor que a adição de um 5 grupo pode ainda agregar grande valor explicativo da variabilidade dos dados, ainda que isto não seja observável na figura bidimensional. Desse modo, entende-se que a adição do 5 grupo possa influenciar numa maior explicação da variabilidade dos dados, o que possivelmente seria observável através da inclusão de uma Dimensão 3, com potencial para explicar os 27,3% restantes da variabilidade.

Por essa razão, e também para padronizar os procedimentos de clusterização realizados neste estudo, optou-se pela utilização dos 5 grupos, também no método K-means, assim como foi feito nos métodos hierárquicos.

Para seguir com o teste da hipótese que motivou este trabalho, foram realizadas modelagens por 6 técnicas de Machine Learning: Regressão Linear Múltipla, Logística

Múltipla, Poisson (para dados de contagem), Binomial Negativa ou Poisson-Gama (para dados de contagem quando há superdispersão dos dados), Poisson com inflação de zeros e Poisson-gama com inflação de zeros.

Além disso, para cada técnica de Machine Learning, foram realizadas 6 modelagens, sendo uma sem a utilização de clusters, e outras cinco, ou seja, uma modelagem adicional para cada método de clusterização. Em cada modelagem, os grupos (clusters) foram nomeados e utilizados como variáveis categóricas, a fim de que explicassem parte do comportamento dos dados, melhorando o poder preditivo dos modelos.

Como foram utilizadas seis técnicas de modelagem, e para cada técnica, foram construídos seis modelos, ao final do procedimento foi possível a obtenção de 36 modelos de Machine Learning. Os parâmetros de cada modelo podem ser encontrados no Apêndice 1, assim como os seus valores de LogLik. Para uma visualização resumida dos resultados, os modelos foram também agrupados na matriz (Tabela 2) a seguir, onde cada linha representa a sua classificação referente à clusterização (se houve, e qual método foi utilizado), e as colunas representam as diferentes técnicas de regressão:

Tabela 2 - LogLiks dos Modelos de Regressão com e sem Clusterização.

	LM	**LgM**	**Poisson**	**NB**	**ZIP**	**ZINB**
Sem Clusterização	-834,8608	-75,72616	-1093,316	-841,2543	-819,6495	-746,0648
Single Linkage	-816,9487	-75,72616	-1069,209	-841,2543	-793,4	-739,7818
Complete Linkage	-801,6242	-62,37553	-943,8173	-810,807	-752,4474	-713,2631
Average Linkage	-790,819	-73,58934	-999,8132	-829,3198	-748,2831	-719,8972
Ward's Method	-807,084	-55,43368	-904,8311	-793,4221	-799,255	-745,1483
K-Means	-801,0253	-65,37111	-991,8426	-830,3711	-788,5566	-721,3081

Fonte. Resultados originais da Pesquisa.

Numa análise geral, foi possível observar que os valores de LogLik aumentam (ou seja, se aproximam de zero) quando há a utilização de métodos de clusterização, em comparação a quando não há. Como exceção, apenas foram observados dois casos: O modelo de Regressão Logística Múltipla e o Modelo de Regressão Poisson-Gama, ambos com a utilização dos agrupamentos definidos por Single Linkage, apresentaram todas as variáveis criadas a partir dos clusters não significantes para o modelo, o que levou à exclusão delas. Deste modo, os modelos finais, nestes dois casos, foram idênticos aos modelos sem clusterização dos seus respectivos métodos.

Pôde-se constatar que a modelagem de Regressão Logística Múltipla apresenta os melhores valores de LogLik em comparação com os demais métodos, sendo assim os modelos mais adequados para fins preditivos. Entretanto, mesmo sendo bastante preciso, o modelo se limita a uma previsão da probabilidade de evento (liberação de tubos) ou não-evento (não liberação de tubos). Este tipo de previsão, embora útil para algumas tomadas de decisão, não resolve o problema central relacionado à previsão das quantidades de tubos a serem liberados. Dentre os modelos de Regressão Logística Múltipla, o que apresenta melhor LogLik foi o modelo que utilizou o método Ward's de Clusterização, com um valor de -55,43368.

Os Modelos com piores valores de LogLik foram os modelos Poisson para dados de contagem, variando entre -1093,316 e -904,8311, para os quais o melhor valor foi obtido através do modelo que utilizou o método Ward's de Clusterização.

Em geral, os LogLiks dos modelos de Regressão Linear Múltipla, Poisson Gama, e Poisson-Gama com Inflação de Zeros apresentaram valores semelhantes, variando (ao todo) de -841,2543 a -748,2831. Os melhores modelos de cada um destes métodos de regressão foram: Regressão Linear Múltipla com clusterização por Average Linkage (LogLik = -790,819),

Poisson Gama com agrupamentos pelo Método Ward's (LogLik = -793,4221), e Poisson-Gama com Inflação de Zeros com clusterização por Average Linkage (LogLik = -748,2831).

Com relação aos modelos adequados para dados de contagem (Poisson, Poisson-Gama, Poisson com inflação de Zeros e Poisson-Gama com Inflação de Zeros), foi possível perceber que os modelos "Gama" (binomiais negativos) apresentaram maiores valores de LogLik. Isto pôde ser explicado pois tais modelos, sendo eles com e sem inflação de zeros, levam em consideração a superdispersão dos dados, que foi confirmada após a realização do teste Cameron & Trivedi, com t-score de -110,24 e p-valor de aproximadamente 2,2x10-16.

Além disso, ainda sobre os modelos de contagem, notou-se que foram obtidos melhores resultados com os modelos que supõem a inflação de zeros. Uma possível explicação para tal é que estes modelos contam com componentes logísticos, responsáveis por carregar a probabilidade da variável resposta ter valor zero (Fávero & Belfiori, 2017). Como foi observado que os modelos de regressão puramente logísticos são os que apresentaram os melhores resultados dentre todos os demais, é possível concluir então que os modelos que carregam componentes logísticos tenham também uma melhoria colateral nas suas capacidades preditivas.

Finalmente, a técnica de regressão que, ao mesmo tempo, é adequada à característica do problema (previsão da contagem de tubos), e apresentou os melhores valores de LogLik, foi a Regressão Poisson-Gama (Binomial negativa para dados de contagem) com Inflação de Zeros. Isto pode ser explicado pois os modelos referentes a esta técnica somam as vantagens dos modelos de contagem e dos modelos logísticos, sendo capazes de prever as quantidades de tubos a serem liberados e, simultaneamente, calculando a probabilidade de não-liberação (liberação de zero tubos). O modelo sem clusterização apresentou o pior valor de LogLik (-746,0648), enquanto o melhor valor foi observado no modelo com clusterização por Complete Linkage, apresentando um LogLik = -713,2631. Este foi, portanto, o modelo escolhido para melhor representar o fenômeno, e para a realização de previsões.

A equação 1 descreve um modelo Poisson-Gama com Inflação de Zeros do seguinte modo:

$$\lambda_{ZINB} = \left\{1 - \frac{1}{1+e^{-(\delta_0+\delta_1 x_1+\delta_2 x_2+\delta_3 x_3 \dots)}}\right\}.\left\{e^{(\beta_0+\beta_1 x_1+ \beta_2 x_2+ \beta_3 x_3 \dots)}\right\} \qquad (1)$$

Nesta equação, nós temos dois componentes: o logístico e o binomial negativo. Então, no componente logístico, observamos os parâmetros δ, onde δ_0 se refere ao intercepto, e δ_n (n = 1, 2, 3...) se referem aos coeficientes de cada variável que se mostre estatisticamente significante para o modelo.

Os parâmetros foram definidos por meio da Regressão Binomial Negativa com Inflação de Zeros, no software R, e o resultado obtido é descrito na Tabela 3, a seguir:

Tabela 3. Parâmetros do Modelo Final – ZINB com Clusterização por Complete Linkage.

Parâmetro	Valor	Erro Padrão	p-valor
Intercepto	1.368355	0.214860	1.91e-10
T	0.033025	0.007455	9.43e-06
R	-1.797167	0.331397	5.86e-08
MO	0.138253	0.019118	4.77e-13
CL2	-0.545943	0.068291	1.30e-15

Fonte. Resultados originais da Pesquisa.

Para o componente Binomial Negativo, observamos os parâmetros β, onde β_0 se refere ao intercepto, e β_n (n = 1, 2, 3...) se referem aos coeficientes de cada variável que se mostre estatisticamente significante para o modelo (Fávero & Belfiori, 2017).

Tabela 4. Componente do Logito

Parâmetro	Valor	Erro Padrão	p-valor
Intercepto	-0.611402	276.344	0.026934
U	0.07310	0.03259	0.024889
R	2.97563	0.78522	0.000151
MO	-0.41560	0.17953	0.020618
CL2	248.385	0.57064	1.34e-05

Fonte. Resultados originais da Pesquisa.

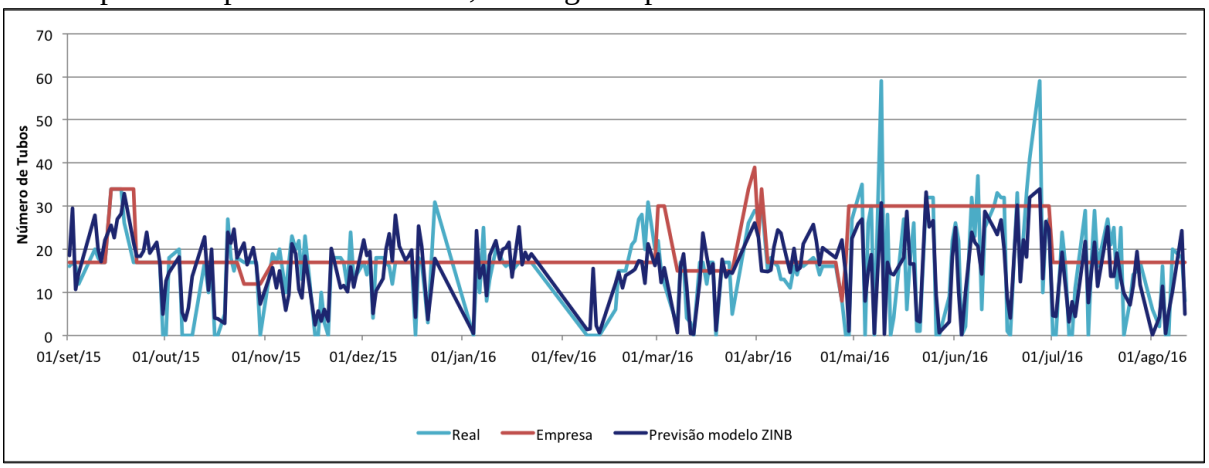
O modelo final é apresentado na equação (2), a seguir:

$$y_{tubos} = \lambda_{ZINB} = \left\{1 - \frac{1}{1 + e^{-(-6{,}11402 + 0{,}0732.x_U + 2{,}97563.x_R - 0{,}41560.x_M + 2{,}48385.x_{CL2})}}\right\}. \left\{e^{(1{,}368355 + 0{,}033025.x_T + -1{,}797167.x_R + 0.138253.x_M - 0.545943.x_{CL2})}\right\} \quad (2)$$

onde λ_{ZINB} é a representação típica dos valores previsto por um modelo binomial negativo com inflação de zeros. Este parâmetro, para o estudo, tem valor igual a y_{tubos} , uma vez que diz respeito à quantidade de tubos liberados prevista pelo modelo. Observou-se que se mantiveram como variáveis explicativas estatisticamente significantes: a temperatura ambiente do dia, como x_T (ºC); alerta de incidência de raios no dia, como x_R (variável qualitativa, definida como 1 em caso da ocorrência de alerta ou 0 no caso da não ocorrência); volume de mão-de-obra do dia, como x_M (funcionários/ dia); e grupos de observações, como x_{CL2} (variável dummy, definida a partir do método Complete Linkage, para o grupo 2), na componente Binomial Negativa. Também foi observada a umidade relativa do ar do dia, como x_U (%); alerta de incidência de raios no dia, como x_R (variável qualitativa, definida como 1 em caso da ocorrência de alerta ou 0 no caso da não ocorrência); volume de mão-de-obra do dia, como x_M (funcionários/ dia); e grupos de observações, como x_{CL2} (variável dummy, definida a partir do método Complete Linkage, para o grupo 2), na componente logística do modelo.

O modelo obteve valores de previsão muito mais precisos, o que pode ser observado na Figura 6, em comparação com a previsão da empresa:

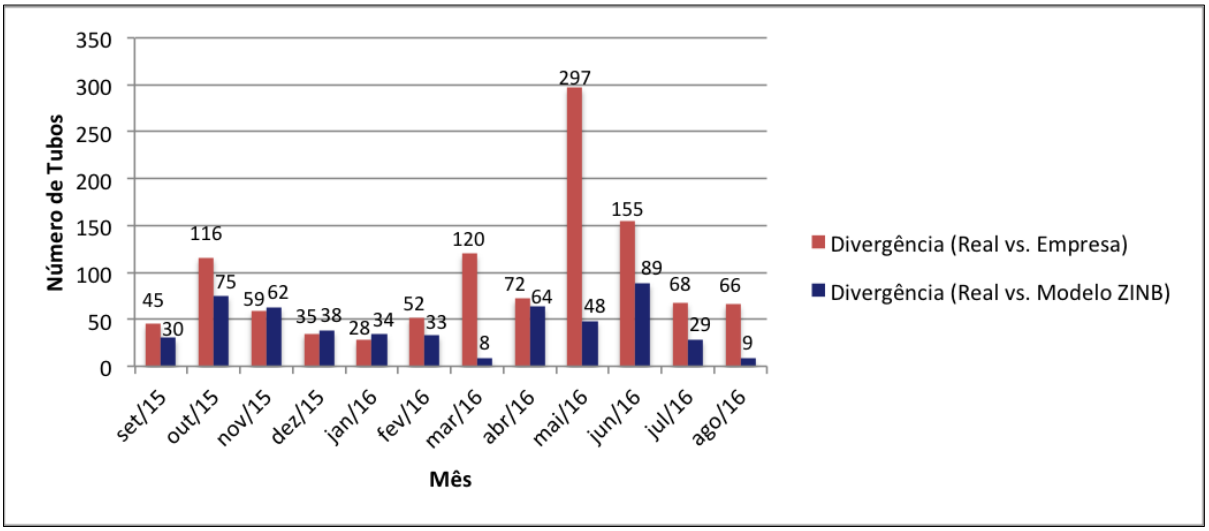
Figura 6. Comparação entre quantidade de tubos liberados pela empresa vs. quantidade real vs. valores previstos pelo Modelo ZINB, ao longo do período estudado.



Fonte: Resultados originais de pesquisa.

Com a aplicação deste modelo, obteve-se uma divergência de 520 tubos ao longo de todo o período de acompanhamento que em comparação com 1113 tubos de divergência previstos pela empresa é bem mais assertivo. Podemos observar o resultado total de divergência por mês na Figura 7, abaixo:

Figura 7. Comparação entre divergências geradas pelos valores previstos pela empresa vs. divergências geradas pelos valores previstos pelo Modelo ZINB, por mês, ao longo do período estudado.



Fonte: Resultados originais de pesquisa.

Em termos financeiros, esta divergência se refere a R$ 15.600.000,00, ou seja, cerca de R$ 17.790.000 a menos que os R$ 33.390.000 de divergência previstos pela própria empresa.

Conclusão

Foi identificada a existência de contextos latentes do fenômeno, que foram traduzidas em grupos de observações por meio de técnicas de clusterização. A utilização destes clusters na construção de modelos de *machine learning* contribuiu para uma melhora da capacidade preditiva dos modelos propostos. A partir do teste e da comparação entre as técnicas estudadas, constatou-se que o modelo mais adequado foi o construído a partir da Regressão Poisson-Gama (Binomial Negativa para dados de contagem) levando em consideração a Inflação de Zeros, com Clusterização pelo método Hierárquico Complete Linkage, pois este apresenta maior precisão e capacidade preditivas, tanto para quantidades quanto para as probabilidades de eventos de não-liberação de tubos.

A contribuição teórica do estudo decorre da afirmação de que a utilização de métodos de clusterização prévios podem melhorar de forma substancialmente os modelos de machine learning. Enquanto isso, como contribuição gerencial, com um método de previsão baseado no modelo proposto, entende-se que possa haver uma significativa melhora na definição ou negociação das quantidades de produtos a serem entregues de forma periódica, ou até mesmo dos prazos contratuais. Além disso, os processos de tomadas de decisão podem se tornar mais eficientes com a possibilidade de alocar mais eficientemente os recursos e o capital humano com base, por exemplo, nas condições climáticas, escalando mais funcionários para atuar no processo de reparo de revestimento em dias com previsão de sol e altas temperaturas, ou ocupando os funcionários em outras atividades em dias cuja previsão é de chuva, tempestade e/ou baixas temperaturas, maximizando assim a utilização efetiva da mão de obra.

Diante os resultados alvissareiros obtidos nas previsões de possíveis economias financeiras, abre-se uma serie de potenciais encaminhamentos em termos organizacionais. A partir da lógica de gestão, vislumbra-se espaços para aprimoramentos nos processos internos, na capacitação do capital humano e na integração de sistemas. Quanto aos processos, o mapeamento do fluxo de atividades e setores envolvidos nos procedimentos produtivos devem ser revistos e atualizados. A respeito da capacitação, se faz necessário treinamentos específicos em áreas afins, no sentido de possibilitar o aproveitamento do profissional em outros setores, para que seja possível atribuir novas funções quando o clima não for favorável. Em relação a integração de sistemas, precisa ser compreendido o momento da empresa em termos de automação e parque tecnológico, com vistas a aprimorar os equipamentos e a infraestrutura para que possa ser aproveitado ao máximo os ativos para proporcionar maior rapidez e qualidade nas entregas dos produtos produzidos na empresa.

Por fim como limitação do estudo indicamos a impossibilidade de realização de testes e coleta de dados em um ambiente controlado, sendo uma sugestão para futuros trabalhos a coleta de dados mediante a utilização de experimentação em ambientes controlados para obtenção dos dados e realização dos testes.

Referências

Abacusliquid (2017). Disponível em: <http://abacusliquid.com/setores-economia/mercado-financeiro/aco-hoje/>. Acesso em: 03 Julho 2017.

Arlia D.; Coppola, M. (2001). Experiments in Parallel Clustering with DBSCAN. In: Euro-Par, Parallel Processing, 2001, Springer, Berlin, Heidelberg, Alemanha. Anais. p. 326-331.

Bock, HH (2007). Clustering Methods: A History of *k*-Means Algorithms. (p. 161-172) In: Brito, P.; Cucumel, G.; Bertrand, P.; de Carvalho, F. Selected. Contributions in Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization. 1ed. Springer, Berlin, Heidelberg, Alemanha.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological), 26(2), 211-243.

Eurokation (2015). Disponível em: <http://www.euroaktion.com.br/Tabela%20de%20Densidade%20dos%20Materiais.>. Acesso em: 03 Julho 2017.

Fávero, L. P., & Belfiore, P. (2017). Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Elsevier Brasil.

Gil, A. (2019). Métodos e técnicas de pesquisa social. rev. atual. São Paulo: Atlas.

Goldstein, H. (2011). Multilevel statistical models. 4ed. John Wiley & Sons, Hoboken, Nova Jersey, Estados Unidos.

Hosmer, D. W.; Lemeshow, S. (2000). Applied logistic regression, 2ed. Wiley, Nova York, Nova York, Estados Unidos.

Lima, V. R. (2008). Fundamentos de caldeiraria e tubulação industrial. Ciência Moderna.

Long, J. S. (1997). Regression models for categorical and limited dependent variables. Sage. Thousand Oaks, California, Estados Unidos.

Lord, D.; Washington, S. P.; Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, *37*(1), 35-46.

Montgomery, D. C.; Runger, G. C. (2011). Applied statistics and probability for engineers, 5ed, John Wiley & Sons, Hoboken, Nova Jersey, Estados Unidos.

Ochi, L. S.; Dias, C. R.; Soares, S. S. F. (2004). Clusterização em mineração de dados. Programa de Pós-Graduação em Computação. Universidade Federal Fluminense. Niterói, Rio de Janeiro, Brasil.

Renner Coatings (2015). Polidura Fundo Anticorrosivo Epóxi. Disponível em: <http://www.rennercoatings.com/uploads/1525114760-78-10-polidura-fundo-anticorrosivo-epoxi-v02.pdf>. Acesso em: 17 nov. 2021.

Rokach, L.; Maimon O. (2005) Clustering Methods. In: Maimon O.; Rokach L. Data Mining and Knowledge Discovery Handbook. Springer. Boston, Massachusetts, Estados Unidos.

Sakia, R.M. (1992). The Box-Cox Transformation Technique: A Review. Journal of the Royal Statistical Society: Series D (The Statistician) 41: 169-178.

Silberzahn, R.; Uhlmann, E. L. (2015). Many hands make tight work. Nature 526: 189-191.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association 58 (301): 236-244.

Apêndice 1

Anexo 1

	Modelo	Cluster	Parâmetros							Clusters				Coeficientes			Parâmetros de Componente Logito (Modelos Zero-Inflated)							
			Intercepto	T	U	PO	R	MO	Theta	Clust. D. 2	Clust. D. 3	Clust. D. 4	Clust. D. 5	R^2	R^2 Ajust.	LogLik	Intercepto	U	R	MO	Clust. D. 2	Clust. D. 3	Clust. D. 4	Clust. D. 5
1	Linear Múltiplo (LM)	n	33,154	*	-0,426	*	-15,613	2,004	n	n	n	n	n	0,424	0,417	-834,8608								
2		Single Linkage	36,34098	*	-0,42383	*	-15,25406	1,42244	n	36,87376	*	*	31,27918	0,5046	0,4939	-816,9487								
3		Complete Linkage	-10,1652	0,5886	*	*	-12,1056	2,5913	n	-11,2956	4,4404	*	*	0,5645	0,5551	-801,6242								
4		Average Linkage	-33,7124	2,6444	*	-1,3877	-14,1126	1,5372	n	*	9,8963	17,4276	38,7308	0,6023	0,5902	-790,819								
5		Ward's Method	-4,9369	1,6116	*	-1,7037	-12,8674	2,4571	n	*	-9,5105	-7,2059	*	0,544	0,5322	-807,084								
6		K-Means	14,193	*	*	*	-13,2911	0,7835	n	-4,6379	*	15,7098	-10,2523	0,5667	0,5573	-801,0253								
7	Logístico Múltiplo (LgM)	n	11,43185	*	-0,13061	*	-3,35427	*	n	n	n	n	n	n	n	-75,72616								
8		Single Linkage	11,43185	*	-0,13061	*	-3,35427	*	n	*	*	*	*	n	n	-75,72616								
9		Complete Linkage	5,92748	*	-0,07133	*	-3,24257	0,42173	n	-2,46244	*	*	*	n	n	-62,37553								
10		Average Linkage	11,95186	*	-0,14063	*	-3,52215	*	n	*	1,09003	*	*	n	n	-73,58934								
11		Ward's Method	20,5661	*	*	*	-2,9247	*	n	*	-21,0345	-18,1839	-18,1382	n	n	-55,43368								
12		K-Means	4,4625	*	*	*	-3,3995	*	n	-2,4404	*	*	-3,9004	n	n	-65,37111								
13	Poisson	n	3,72382	*	-0,02621	*	-2,866968	0,129974	n	n	n	n	n	n	n	-1093,316								
14		Single Linkage	3,950807	*	-0,02666	*	-2,845211	0,097348	n	0,941905	*	*	0,590039	n	n	-1069,209								
15		Complete Linkage	0,248753	0,06065	*	*	-2,253789	0,198491	n	-0,868641	*	*	0,253789	n	n	-943,8173								
16		Average Linkage	-0,60996	0,18101	*	-0,10258	-2,76917	0,114411	n	-0,2247	0,66471	0,78231	0,97634	n	n	-999,8132								
17		Ward's Method	1,632145	0,08279	*	-0,088015	-2,713647	0,13672	n	*	-1,662517	-0,347139	*	n	n	-904,8311								
18		K-Means	2,65069	*	*	*	-2,69875	0,04932	n	-0,30338	0,54462	-0,86157	*	n	n	-991,8426								
19	Poisson-Gama (NB)	n	4,405125	*	-0,03051	*	-2,765028	0,122803	n	n	n	n	n	n	n	-841,2543								
20		Single Linkage	4,405125	*	-0,03051	*	-2,765028	0,122803	n	*	*	*	*	n	n	-841,2543								
21		Complete Linkage	-0,54983	0,0857	*	*	-2,46554	0,22754	n	-0,91073	*	*	0,40183	n	n	-810,807								
22		Average Linkage	1,309067	0,08595	-0,0235	*	-2,662886	0,121084	n	*	0,746396	0,837839	1,053507	n	n	-829,3198								
23		Ward's Method	1,59973	0,08959	*	-0,09061	-2,61362	0,12384	n	*	-1,62081	-0,38159	*	n	n	-793,4221								
24		K-Means	2,97092	*	*	*	-2,61065	*	n	-0,37663	*	0,61445	-0,86484	n	n	-830,3711								
25	Zero Inflated p (ZIP)	n	3,292337	*	-0,01568	*	-1,941624	0,105511	n	n	n	n	n	n	n	-819,6495	-11,46599	0,13106	3,20009	*	n	n	n	n
26		Single Linkage	3,525288	*	-0,01591	*	-1,932825	0,069689	n	0,922859	*	*	0,71286	n	n	-793,4	-11,47222	0,13114	3,19979	*	*	*	*	*
27		Complete Linkage	0,85359	0,04637	*	*	-1,7972933	0,16201	n	-0,510846	*	0,195226	*	n	n	-752,4474	-6,05959	0,07253	3,05212	-0,41542	2,46728	*	*	*
28		Average Linkage	0,63537	0,10512	*	-0,043254	-1,876759	0,065673	n	*	0,424697	0,710579	1,071698	n	n	-748,2831	-12,0278	0,14172	3,31437	*	*	-1,12851	*	*
29		Ward's Method	1,76705	0,0676	*	-0,07233	-1,71587	0,13345	n	*	-0,68168	-0,27946	*	n	n	-799,255	-2,0168	*	2,6046	*	*	*	*	*
30		K-Means	2,42142	*	*	*	-1,97948	0,07116	n	*	*	0,59977	*	n	n	-788,5566	-4,6376	*	3,0611	*	2,6511	*	*	4,0831
31	Zero Inflated p-g (ZINB)	n	3,3816319	*	-0,01632	*	-1,941933	0,098139	2,30652	n	n	n	n	n	n	-746,0648	-11,50247	0,13153	3,13605	*	n	n	n	n
32		Single Linkage	3,546225	*	-0,01628	*	-1,935887	0,07013	2,442077	0,922361	*	*	0,707473	n	n	-739,7818	-11,5048	0,13157	3,14048	*	*	*	*	*
33		Complete Linkage	1,368355	0,03303	*	*	-1,797167	0,138253	2,666562	-0,545943	*	*	*	n	n	-713,2631	-6,11402	0,0731	2,97563	-0,4156	2,48385	*	*	*
34		Average Linkage	0,55829	0,10941	*	-0,04578	-1,88181	0,06785	2,82475	*	0,44036	0,71657	1,0859	n	n	-719,8972	-12,05451	0,1421	3,26027	*	*	-1,13951	*	*
35		Ward's Method	1,75374	0,07043	*	-0,07419	-1,66338	0,13071	2,55689	*	-0,69532	-0,31409	*	n	n	-745,1483	-2,0359	*	2,6048	*	*	*	*	*
36		K-Means	2,97426	*	*	*	-1,98045	*	2,59869	-0,22997	*	0,61111	-0,39176	n	n	-721,3081	-4,723	*	2,916	*	2,757	*	*	4,169
	Modelo	Cluster	Intercepto	T	U	PO	R	MO	Theta	Clust. D. 2	Clust. D. 3	Clust. D. 4	Clust. D. 5	R^2	R^2 Ajust.	LogLik	Intercepto	U	R	MO	Clust. D. 2	Clust. D. 3	Clust. D. 4	Clust. D. 5

*	Parâmetro não-significante para o modelo (p-valor > 0,05)
n	Dado não aplicável
	Inesistência de Parâmetros de Componente Logito (Para modelos que não consideram Inflação de Zeros)
Clust. D. #	Cluster Dummy, variando de 2 a 5 (Cluster 1 tem sua representação imbutida no valor do intercepto)