Revista de Matemática da UFOP
ISSN: 2237-8103

Uma Abordagem Pela Teoria de Valores Extremos: Determinação do Tamanho Amostral Para Inspeção de Equipamentos de Processo Utilizando a Distribuição Generalizada de Pareto e o *Peak-Over-Threshold Method*

Iago Pereira Lemos lemosiago123@gmail.com
Universidade Federal de Uberlândia, Uberlândia, MG, Brazil

Antônio Marcos Gonçalves de Lima amglima@ufu.br
Universidade Federal de Uberlândia, Uberlândia, MG, Brazil

Fabiana Dias Fonseca Martins fabianadf@petrobras.com.br
Petrobras, Rio de Janeiro, RJ, Brazil

Weslley Carlos Dias da Silva weslleycds@petrobras.com.br
Petrobras, Rio de Janeiro, RJ, Brazil

Resumo

Análises de valores extremos estão presentes em diversas áreas atualmente e, são aplicadas, principalmente, na modelagem de eventos extremos e raros. Este trabalho apresenta uma abordagem por meio da teoria de Valores Extremos, pautada no Teorema de Pickands - Balkema - de Haan, utilizando o *Peak-Over-Threshold method* para definição de um conjunto de observações situadas no domínio de atração das distribuições de valores extremos, em conjunto com a Distribuição Generalizada de Pareto, para a modelagem estatística, com o intuito de apresentar um algoritmo que define um tamanho amostral *a posteriori* para equipamentos de processo submetidos à corrosão generalizada. Foi possível, com a implementação do algoritmo, realizar uma clusterização simples, baseada na estrutura dos dados, e promover a reamostragem destes, para um menor tamanho amostral, realizando o ajuste na distribuição e computando o valor de retorno, de forma a extrapolar os dados para regiões hipoteticamente não inspecionadas. Com base no valor de retorno computado e na papulação já conhecida do equipamento de processo em questão, foi realizada a estimativa de um tamanho amostral representativo.

Palavras-chave

Teoria de Valores Extremos, Distribuição Generalizada de Pareto, *Peak-Over-Threshold method*, POT, Tamanho Amostral.

1 Introdução

A análise de dados e aplicação de ferramentas estatísticas poderosas pautadas na teoria de valores extremos estão amplamente presentes na engenharia, atuando, principalmente, na prevenção de eventos extremos [4], e na solução de problemas físicos relacionados à obtenção de dados.

Neste contexto, uma das grandes aplicações das estatísticas relacionadas à análise de valores extremos é no campo de avaliação da integridade física e tomadas de decisão relacionadas a componentes de processos submetidos à corrosão generalizada. A obtenção de dados aplicáveis às estatísticas é realizada por inspeções periódicas do tipo ENDs (ensaios não destrutivos) não convencionais, tais como IRIS (Internal Rotary Inspection System) ou correntes parasitas [9]. Contudo, alguns problemas pontuais são enfrentados no que tange à aplicação de tais metodologias de análise, como dificuldade de acesso à algumas localidades das linhas de processo [11], que podem acontecer por inúmeros motivos ou, ainda atrelado à isso, a impraticabilidade de se inspecionar toda a linha de processo devido ao elevado custo da inspeção [8].

No campo da engenharia de corrosão, normalmente os dados de inspeção são ajustados em distribuições comumente utilizadas, como Normal ou Lognormal, e a profundidade máxima de defeito é estimada com base no limite superior do intervalo de confiança [11]. Ainda, segundo o autor, o emprego das análises de valores extremos permite ao analista de integridade estimar o valor máximo de corrosão de uma forma mais sistemática, considerando que a estimativa do valor máximo é a principal preocupação do analista.

A extrapolação dos dados para áreas não inspecionadas se faz extremamente importante, devido à possibilidade de se estimar o valor da corrosão máxima na região de difícil acesso ou ainda, estimar a corrosão máxima no restante da linha de processo utilizando uma quantidade menor e selecionada de dados, o que implica em menor custo de inspeção. Por meio da aplicação do Peak-

Over-Threshold Method (POT) em conjunto com a Distribuição Generalizada de Pareto (GPD), é possível modelar a cauda de uma distribuição de valores extremos [3] e, por meio da aplicação do conceito de valor de retorno, pode-se estimar a corrosão máxima em regiões não inspecionadas [11]. Contudo, faz-se necessário determinar um tamanho amostral representativo para as amostras de forma a obter uma boa estimativa para a extrapolação. Portanto, a definição de um tamanho amostral representativo reduzido auxilia na redução de custos nas inspeções e, ainda, fornece uma boa estimativa no que tange à extrapolação para as áreas não inspecionadas, exercendo um papel importante no auxílio na tomada de decisão.

Neste contexto, este trabalho apresenta uma abordagem pela teoria de valores extremos aplicada à componentes de processo gerais, cuja inspeção é feita pelo método IRIS, apresentando um algoritmo iterativo, que trabalha com a reamostragem tomando por base clusters, para determinação de um tamanho amostral *a posteriori* representativo para aplicação do método POT em conjunto com a GPD, utilizando a própria extrapolação para verificação do tamanho amostral, neste caso, partindo de equipamentos com todos os tubos inspecionados, isto é, de população conhecida.

2 Conceitos Iniciais

Nesta seção serão apresentados conceitos e definições fundamentais para entendimento do trabalho. Ademais, serão apresentadas algumas justificativas e hipóteses realizadas para o emprego dos métodos a serem descritos.

2.1 A Distribuição Generalizada de Pareto (GPD) e o *Peak-Over-Threshold Method* (POT)

A ideia principal das análises de valores extremos é modelar o risco de eventos extremos e raros, gerando estimativas da frequência destes eventos. Estas análises são baseadas no comportamento assintótico dos extremos observados [4].

O POT é um método estatístico que, atualmente, é uma das mais poderosas ferramentas para análise de probabilidade de eventos extremos [3]. Segundo Tan [11], o método é uma forma natural de se determinar se uma observação é extrema e, de acordo com Bommier [4], a abordagem pelo POT é extremamente utilizada para identificação de eventos extremos como cargas em estruturas, altura de ondas marítimas, velocidades do vento, seguros, etc.

Seja um conjunto de dados observados $x_1, \ldots, x_N$, iid (variáveis aleatórias e identicamente distribuídas), os eventos extremos são identificados definindo um limiar ótimo chamado de u, para o qual tem-se um conjunto de excedências $(x_i : x_i > u)$. Denotando estes excedentes por $x_{(1)}, \ldots, x_{(k)}$, e definindo os excessos acima deste limiar por $y_j = x_{(j)} - u$, para $j = 1, \ldots, k$.

Conforme o Teorema de Pickands – Balkema – De Haan [2] [7], sabe-se que a Distribuição Generalizada de Pareto pode ser utilizada para se modelar uma cauda de uma distribuição de valores extremos [3]. Além disso, Belitsky e Moreira [3] afirmam que, considerando o Resultado de Pickands, o conjunto y_j, sendo valores acima de um limiar alto o suficiente, pode ser ajustado em uma Distribuição Generalizada de Pareto e, se o limiar ótimo, u, pode ser determinado, o conjunto de dados pertence ao domínio de atração das distribuições de valores extremos.

Desta forma, a aplicação do POT permite que seja modelada uma distribuição de valores extremos com base na Distribuição Generalizada de Pareto. A GPD tem sua função de distribuição acumulada dada pela Eq.(1)

$$G_{(\xi,\beta)} = \begin{cases} 1 - \left[1 + \xi \dfrac{(x-u)}{\beta}\right]^{-\frac{1}{\xi}} & se \quad x > u, \\ 0 & se \quad x \leq u \end{cases} \tag{1}$$

onde ξ e β são, respectivamente, os parâmetros de forma e escala, que são estimados por meio da aplicação da função da máxima verossimilhança, dada pela Eq.(2). Alguns outros métodos para a estimativa dos parâmetros já foram propostos (Estimador de Pickands, método do Momento, o *Probability Weighted Moments method* e o método da Máxima Verossimilhança), contudo, o método escolhido é o único que combina eficiência teórica, promove uma base geral para inferências e se estende diretamente aos modelos que incorporam não estacionariedade e dependência covariável [1].

$$l(\xi,\beta) = -n_u log(\beta) - (1+\frac{1}{\xi})\sum_{i=1}^{n_u} log(1+\xi\frac{(x_i-u)}{\beta}) \quad (2)$$

Porém, o estimativa pela função da máxima verossimilhança nem sempre é válida. A função da máxima verossimilhança é válida para $\xi > -1$, mas as propriedades da normalidade assintótica da função são apenas válidas para $\xi > -1/2$. Quando $\xi < -1$, a estimativa não existe [4].

A função de densidade de probabilidade, portanto, é dada pela Eq.(3).

$$g_{(\xi,\beta)} = \frac{dG_{(\xi,\beta)}(x)}{dx} \quad (3)$$

O valor esperado para GPD é dado pela Eq.(4) [4].

$$E(Y) = \frac{\beta}{1-\xi}, \xi < 1 \quad (4)$$

Segundo Tan [11], o valor de retorno associado à um período de retorno $N = 1/p$ é dado pela Eq.(5). Este somado a u computa o valor máximo igualado ou excedido a ser observado em um período de retorno N.

$$y_p = \begin{cases} \frac{\beta}{\xi}(p^{-\xi}-1) & se \quad \xi \neq 0 \\ -\beta(log(p)) & se \quad \xi = 0 \end{cases} \quad (5)$$

A demonstração da Eq.(5) e mais detalhes podem ser encontrados em [4].

No caso em estudo, pode-se determinar a taxa de excedentes em relação à u por área. Conhecendo a área de análise, isto é, a área referente às inspeções, é possível determinar a quantidade de excedentes em uma região específica. Assim, tomando a área de análise como A e a taxa de excedentes como λ, para $\xi \neq 0$, tem-se que:

$$c_{max}(u : \lambda, \xi, \beta) = \frac{\beta}{\xi}(A\lambda^{\xi} - 1) + u \quad (6)$$

onde c_{max} é a estimativa para o valor máximo a ser observado em um período de retorno $N = A\lambda$.

2.2 Definindo o Limiar Ótimo

A escolha do limiar ótimo não é direta e nem provém de uma análise simples. O parâmetro não pode ser demasiado alto, para não haver poucos dados e, por consequência, uma amostra não representativa, mas não pode ser um valor demasiado pequeno, o que enviesa o ajuste na distribuição.

A seguir serão apresentados alguns métodos para definição do limiar ótimo.

2.2.1 A Função da Média Amostral de Excessos (*Mean Residual Life Plot - MRL*)

Davison e Smith [6] propuseram um método gráfico para seleção do limiar ótimo. Este método é baseado na média da GPD. Supondo que a GPD é um modelo válido para excessos acima de um limiar ótimo u_0 gerado por uma série $X_1, \dots, X_n$, pela Eq.(4) tem-se que:

$$E(X - u_0|X > u) = \frac{\beta_{u_0}}{1 - \xi} \tag{7}$$

com $\xi < 1$ e β_{u_0} sendo o parâmetro de escala da GPD para excedentes acima de um limiar ótimo u_0. A propriedade de estabilidade do limiar ótimo significa que se a GPD é um modelo válido para excessos acima de algum limiar ótimo u_0, então é válido para todos os limiares ótimos $u > u_0$.

Desta forma [4]:

$$E(X - u|X > u) = \frac{\beta_u}{1 - \xi} = \frac{\beta_{u_0} + \xi u}{1 - \xi} \tag{8}$$

igualdade esta provinda da relação da GPD com a Distribuição Generalizada de Valores Extremos (GEV). Esta relação nos mostra para que todo $u > u_0$, $E(X - u_0|X - u)$ é uma função linear de u. Além disso, $E(X - u_0|X - u)$ é a média de excessos acima do limiar ótimo u, e pode ser estimada pela média amostral dos excessos acima deste limiar ótimo. Isto define a função da média amostral de excessos (*mean residual life plot, MLR*), dada pelo lócus de pontos:

$$\{(u, \frac{1}{n_u}\sum_{i=1}^{n_u}(X_{(i)} - u)); u < X_{max}\} \tag{9}$$

onde $X_{(1)}, \ldots, X_{(n_u)}$ consiste das n_u observações que excedem u, e X_{max} é a maior observação de X_i. Realizando a obtenção do gráfico referente aos pontos, é necessário procurar pelo valor u_0 onde o gráfico passa a se tornar aproximadamente linear ou deixa de apresentar o comportamento aproximadamente linear.

A interpretação, contudo, é subjetiva e pode ser extremamente desafiadora [5].

2.2.2 Gráfico de Estabilidade dos Parâmetros (*Parameter Stability Plot*)

A ideia é de que se as excedências acima um alto limitar ótimo, u_0, seguem uma GPD com parâmetro de forma ξ e parâmetro de escala β, então, para qualquer limiar ótimo u, sendo $u > u_0$, as excedências vão continuar seguindo a GPD, com parâmetro de forma $\xi_u = \xi$, e parâmetro de escala $\beta_u = \beta_{u_0} + \xi(u - u_0)$ [4].

Seja uma parametrização dada pela Eq.(10), conhecida como modificação do parâmetro de escala.

$$\beta^* = \beta_u - \xi_u u \tag{10}$$

O conjunto de gráficos que estabelecem o *Parameter Stability Plot* é definido pelos lócus de pontos abaixo [4] [11].

$$\{(u, \beta^*); u < X_{max}\}, \{(u, \xi_u); u < X_{max}\} \tag{11}$$

O limiar ótimo deve ser escolhido com base no comportamento aproximadamente constante dos gráficos.

Além disso, em seu extenso trabalho, Coles [5] sugere verificações de gráficos comumente utilizados para verificação do ajuste, como gráficos de probabilidade, gráficos quantis-quantis, gráficos de valor de retorno e comparação gráfica do ajuste com relação ao gráfico da função de distribuição acumulada e o gráfico da função de densidade de probabilidade.

2.2.3 Regras de Ouro (*Rules of Thumb*)

Outro procedimento consiste em escolher uma observação da amostra como um limiar ótimo, escolhendo o k-ésimo X_{n-k+1} da sequência de observações $X_{(1)}, \ldots, X_{(n)}$. Frequentemente é utilizado o nonagésimo quantil. Outras regras podem ser usadas, como a raiz quadrada $k = \sqrt{n}$ ou a regra $k = \dfrac{n^{2/3}}{log(log(n))}$. Tais métodos, que utilizam de fixação de quantis, são inapropriados em um ponto de vista teórico [10].

2.3 Extrapolação das Observações

Conforme definido pela Eq.(6), é possível verificar o valor de retorno máximo que será igualado ou superado em uma região de análise, com base na área de inspeção e na taxa da ocorrência de observação de excedentes.

Seja uma amostra conhecida de tamanho n, provinda de uma área de análise A_n, sabe-se que a área total, isto é, onde se encontra a população da qual a amostra pertence, é dada por A_t. Conhecendo a área de análise e o limiar ótimo provindo da amostra de tamanho n, pode-se determinar a quantidade de observações, n_u, que excedem ao limiar estão presentes na área A_n. Desta forma, tem-se que:

$$\lambda_u = \frac{n_u}{A_n}. \tag{12}$$

Estabelecendo como hipótese que a taxa de excedentes por área na amostra é a mesma para toda a população, pode-se dizer, portanto, que:

$$\lambda_u = \frac{n_{u_{n.inspec}}}{A_t - A_n} \tag{13}$$

onde $n_{u_{n.inspec}}$ é a quantidade de observações, hipotéticas, que iriam exceder o limiar ótimo definido em uma região não inspecionada.

Desta forma, a Eq.(6) pode ser reescrita, utilizando a Eq.(13), resultando na Eq.(14).

$$c_{max}(u : \lambda_u, \xi, \beta) = \frac{\beta}{\xi}([(A_t - A_n)\lambda_u]^{\xi} - 1) + u \tag{14}$$

3 Metodologia

Nesta seção será apresentada a metodologia completa e detalhada de cada etapa do processo para determinação do tamanho amostral com base na aplicação da estatística apresentada anteriormente para extrapolação, desde a definição amostral até o tamanho amostral alcançado, juntamente com uma descrição do algoritmo empregado.

3.1 Primeiro Passo - Definição Amostral

A metodologia desenvolvida é aplicada em equipamentos de processo, cujos dados são obtidos pelo método de ensaio IRIS.

O ensaio IRIS é uma aplicação da técnica de pulso-eco de ultra-som por meio da utilização de um cabeçote especial que contém um cristal piezoelétrico que é excitado por corrente elétrica, gerando um pulso com frequência característica. Este pulso se propaga pelo material do tubo em análise na forma de onda sonora, até a atingir a superfície oposta do tubo, onde aproximadamente, toda energia do pulso é refletida devido à diferença de impedância entre o material ensaiado e o ar. O eco gerado pela reflexão retorna ao cristal, excitando-o e gerando o sinal elétrico. O tempo entre a geração da onda e o retorno é registrado e, conhecendo a velocidade do som no material, calcula-se a distância percorrida e, consequentemente, a espessura remanescente do material é definida [9].

Após um processo de inspeção IRIS, um relatório é gerado contendo uma série de espessuras remanescentes registradas, uma por tubo ensaiado.

Os equipamentos de processo em questão são tubos gerais submetidos à corrosão generalizada devido ao seu *modus operandi*. A origem dos dados é confidencial, portanto, tais observações foram normalizadas para aplicação da metodologia desenvolvida. Desta forma, o intuito é demonstrar o funcionamento do algoritmo.

Considerando o exposto, seja um relatório de inspeção IRIS apresentando o valor da espessura remanescente em N tubos, com população $z = [z_1, \ldots, z_N]$, é realizada uma parametrização nestes dados, de forma a se obter um conjunto de valores de perda de espessura máxima, promovendo uma análise de valores máximos ao invés de valores mínimos. A parametrização é feita segundo a Eq.(15)

$$x_i = e_{nominal} - z_i, i = 1, ..., N. \tag{15}$$

onde e_{nomina} é a espessura nominal dos tubos em análise, desconsiderando a tolerâncias e incertezas de projeto.

3.2 Segundo Passo: Aplicação do *MRL* e do *Parameter Stability Plot*

Com base no conjunto de dados relativos à população dos tubos presentes em um equipamento analisado, são obtidos os gráficos referentes ao *MRL* e ao *Parameter Stability Plot*, cujos pontos são definidos pelas Eqs. (9) e (11), definindo um vetor com valores de u arbitrários, indo de x_{min} até x_{max}.

As regras de ouro não são aplicadas devido às suas ineficiências e teóricas.

Após a obtenção dos gráficos, a análise visual é feita e, com base nesta, se define o limiar ótimo, u, a ser utilizado para realização do ajuste.

Uma vez definido o limiar ótimo, o ajuste é realizados e os parâmetros da população são estimados com um nível de confiança $\alpha = 0.05$. Além disso, com o intuito de se diagnosticar o ajuste, com os parâmetros estimados é possível realizar a obtenção dos gráfico quantil-quantil e de comparação da função de distribuição acumulada teórica e empírica. Desta forma, com base nos gráficos, é possível realizar uma estimativa do quão apropriado está o ajuste, adotando alguns dos métodos sugeridos por Coles [5].

3.3 Terceiro Passo: Aplicação no Algoritmo de Clusterização e Reamostragem

A criação de *clusters* tem como objetivo agrupar os dados de uma amostra de com base em características similares, de forma a encontrar padrões e informações contidas no conjunto de dados. É uma ferramenta extremamente importante atualmente nas aplicações de *Data Mining* (mineração de dados) e, em um sentido mais abrangente, *Data Science* (ciência de dados).

Tomando uma população de N tubos de um equipamento com inspeção completa, isto é, todos os tubos inspecionados, é possível realizar um processo de clusterização simplificado que, com base na estrutura dos dados em análise, agrupa em famílias de valores de corrosão da população em questão, registrando a frequência absoluta e a contribuição de cada família no todo da amostra, em porcentagem. Após a obtenção dos *clusters*, um processo iterativo acontece, de forma a se obter uma amostra menor, com base em um valor pré-determinado de tamanho amostral requerido e um fator chamado de fator de amostragem, ponderando a contribuição de cada família.

Definida a nova amostra e utilizando o limiar ótimo definido para a população, este novo conjunto de dados é ajustado na GPD, utilizando a função da máxima verossimilhança.

É importante ressaltar que, agora, após a definição da nova amostra, esta deve ser considerada uma inspeção a qual não fora inspecionados os tubos do equipamento. Porém, é de conhecimento o restante das observações que tangem à população. Isso serve, portanto, para que se possa comparar a estimativa da corrosão máxima fora da área de inspeção e o valor real desta.

Com os parâmetros estimados e de posse da Eq.(14), estima-se a corrosão máxima fora do conjunto reamostrado. Com a estimativa, computa-se o erro entre corrosão máxima fora do conjunto reamostrado real e a estimada e, se o erro for menor que um erro qualquer estabelecido, o tamanho amostral utilizado é tratado como o novo tamanho amostral *a posteriori*.

O algoritmo para obtenção do novo tamanho amostral é explicado passo a passo abaixo.

Passo 1 – Determina-se, incialmente, o fator de amostragem. Por padrão, o tamanho amostral requerido inicia-se com 30 e este é incrementado a cada finalização de um passo do laço. O fator de amostragem é denotado por $f^{(l)}$ e o tamanho amostral requerido é denotado por $A_{req}^{(j)}$. Definido o fator de amostragem, inicia-se a clusterização. Além disso, determina-se um valor para o erro. Este valor será utilizado para finalizar o laço caso a diferença entre estimativa da corrosão máxima fora da região inspecionada e corrosão máxima real fora da região não inspecionada seja menor ou igual à este erro. Aqui é necessário atenção, à depender da sua quantidade de dados brutos, este erro pode levar à não convergência do método.

Passo 2 – Denotando por k o número de clusters encontrado após a clusterização. Seja X_i, cada uma das famílias de observações de corrosão, com sua frequência absoluta denotada por F_i e sua contribuição percentual dada por P_i, com $i = 1, \ldots, k$. Verifica-se se a porcetagem de contribuição, P_i, de cada família está entre 0 e 5% da contribuição no todo da amostra e, se estiver, computa-se:

$$F_i^{(l)} = F_{+\infty,i}^{(l)} = F_i \times f^{(l)} \tag{16}$$

onde $F_{+\infty,i}^{(l)}$ é a nova frequência arredondada sempre para o inteiro mais próximo na direção de $+\infty$, o que garante para a nova amostra que valores críticos de máximo ou de mínimo estejam sempre presentes, mesmo quando sua nova frequência absoluta seja menor que um.

Caso P_i não esteja na condição citada acima, computa-se:

$$F_i^{(l)} = F_{-\infty,i}^{(l)} = F_i \times f^{(l)} \tag{17}$$

onde $F_{-\infty,i}^{(l)}$ é a nova frequência, arredondada sempre para o inteiro mais próximo na direção de $-\infty$.

Passo 3 – Faz:

$$n^{(l)} = \sum_{i=1}^{k} F_i^{(l)} \tag{18}$$

E verifica-se se $n^{(l)} \leq A_{req}^{(j)}$, de forma a garantir que a nova amostra, $n^{(l)}$, vá ser menor ou igual ao tamanho amostral requerido.

Passo 4-1 – Caso $n^{(l)} \leq A_{req}^{l}$, é realizada uma minoração do fator de amostragem, $f^{(l)}$. Agora $l = l + 1$ e o valor usado para a minoração é denotado por ρ, e é arbitrário. Recomenda-se utilizar um valor na ordem de 10^{-4}.

$f^{(l)}$ é dado pela Eq.(19).

$$f^{(l)} = f^{(l-1)} - \rho \tag{19}$$

$f^{(l)}$ é obtido e processo se realiza novamente, até que a condição seja satisfeita.

Passo 4-2 – Se $n^{(l)} \leq A_{req}^{(j)}$, o ajuste é realizado utilizando o limiar ótimo da população e os parâmetros de forma e escala são estimados. Com base na Eq.(16), estima-se a corrosão máxima na área não inspecionada.

Passo 5-1 – Se a diferença absoluta entre corrosão máxima estimada e a corrosão máxima real fora da área não inspecionada for menor que o erro estabelecido, o laço é finalizado e o tamanho amostral *a posteriori* é definido.

Passo 5-2 – Caso contrário, determina-se um novo tamanho amostral requerido e o laço se realiza novamente, até que a condição no Passo 5.1 seja satisfeita. Agora, $j = j + 1$. Desta forma:

$$A_{req}^{(l)} = A_{req}^{(l-1)} + 1 \tag{20}$$

Para melhor visualização e entendimento do processo, a Fig.1 apresenta um fluxograma do algoritmo.

O valor inicial tomado para $f^{(1)}$ é 0.5, sendo este utilizado em todas as análises, de forma a diminuir a frequência absoluta das famílias de valores de corrosão pela metade. Ademais, ρ foi tido como 10^{-4}.

4 Estudos de Casos e Discussões

Nesta seção serão apresentados os resultados obtidos com a aplicação do método desenvolvido para dois equipamentos de processos genéricos submetidos à corrosão generalizada. Ambos equipamentos são constituídos por feixes de tubos e foram submetidos a uma inspeção a qual todos os tubos presentes foram ensaiados pelo método IRIS.

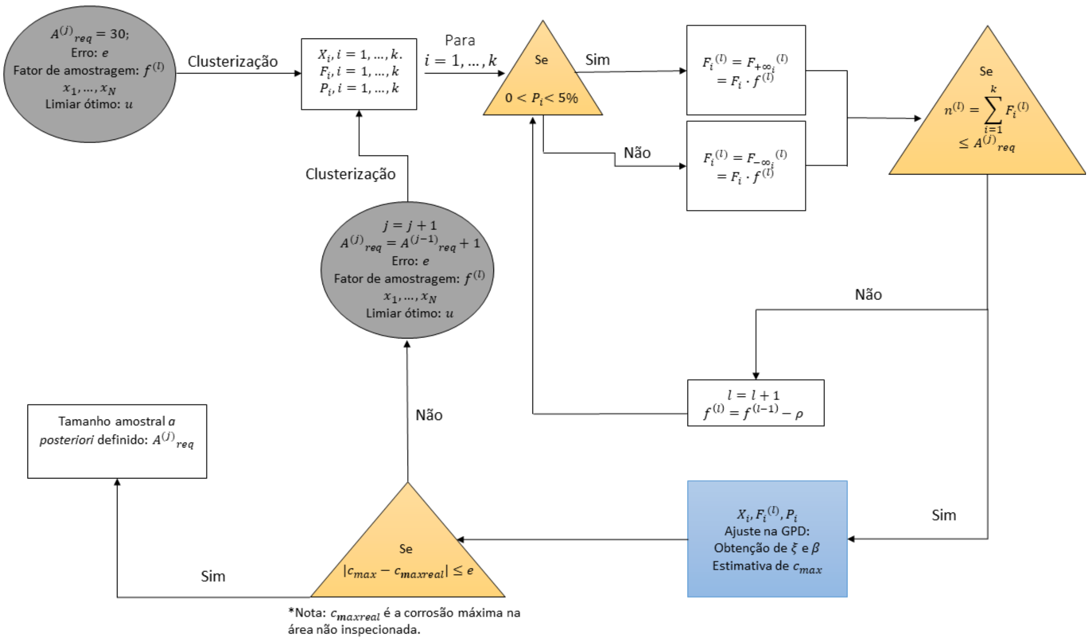


Figure 1: Fluxograma explicitando o algoritmo para definição do tamanho amostral.

4.1 Equipamento de Processo 1

O primeiro equipamento analisado consiste de 329 tubos ensaiados, isto é, população $N = 329$ observações. A Tabela 1 apresenta as características dos tubos relacionados ao equipamento em questão.

Table 1: Características dos tubos analisados referentes ao equipamento de processo 1.

diâmetro (mm)	12.67
comprimento (mm)	4064
$e_{nominal}$ (mm)	1.4
N	329

Conforme o exposto, as observações foram realizadas por meio do ensaio IRIS. Desta forma, aplicando a Eq.(15) é possível obter o novo conjunto, parametrizado, de observações que exprimem a perda de material máxima nos tubos, denotados por $x^{(1)} = [x_1^{(1)}, ..., x_N^{(1)}]$.

Para estimativa dos parâmetros referentes à população, foram obtidos, inicialmente, os gráficos referentes ao *MRL* e ao *Parameter Stability Plot*. A Fig.2 apresenta o gráfico obtido pelo *MRL*. e a Fig.3 apresenta os gráficos obtidos pelo *Parameter Stability Plot*.

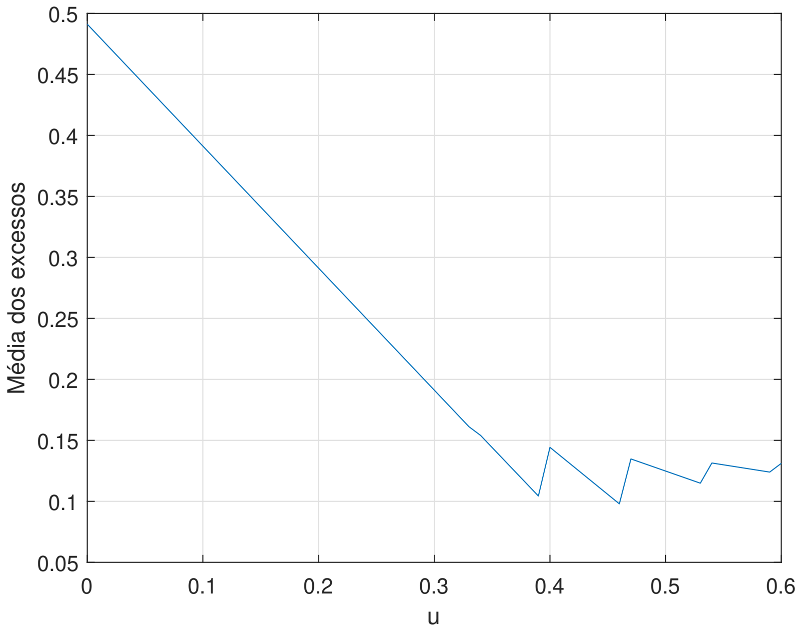


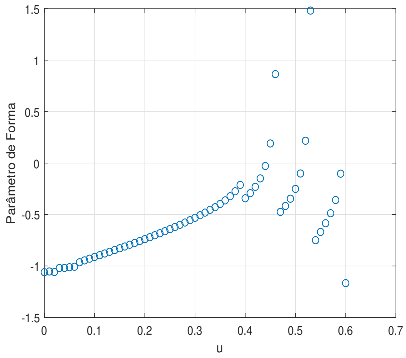
Figure 2: *MRL* referente à população do equipamento 1.

É possível observar que o gráfico apresenta comportamento de serra. Esta fisionomia gráfica é como o gráfico deve ser [3] e está de acordo com o constatado por Tan [11].

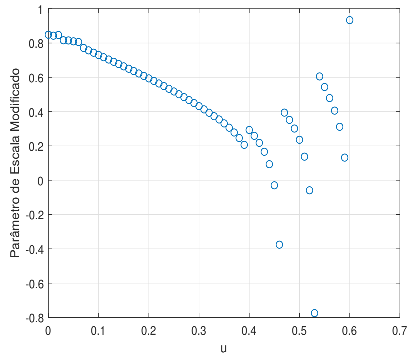
Com base somente no *MRL* foi possível inferir, mesmo que inicialmente, o valor do limiar ótimo, u. O gráfico apresenta o comportamento basicamente linear, decrescente (o que denota um $\xi < 0$), até $u = 0.39$, onde, após este valor, há a quebra no gráfico e a média dos excessos passa a ser aproximadamente constante conforme u cresce.

Esta análise inicial é importante pois, em alguns casos, somente o *MRL* basta para definir o parâmetro u.

A Fig.3 apresenta os gráficos obtidos para o *Parameter Stability Plot* para a população do equipamento em análise.



(a) Estabilidade do parâmetro de forma.



(b) Estabilidade do parâmetro de escala modificado.

Figure 3: *Parameter Stability Plot* referente à população do equipamento 1.

Com base na Fig.3, foi possível observar que o comportamento dos gráficos é aproximadamente constante até $u = 0.39$, isto é, decrescimento e crescimento constantes. Este valor está de acordo com o encontrado no gráfico referente ao *MRL*. Desta forma, foi definido que o limiar ótimo para a população e que será utilizado nas próximas análises é $u = 0.39\ mm$.

Uma vez definido o limiar ótimo, foi realizado o ajuste na GPD, conforme a Eq.(2). O nível de significância utilizado foi $\alpha = 0.05$. Os parâmetros obtidos podem ser verificados na Tabela 2.

Table 2: Valores estimados para a distribuição referente à população de tubos observados no equipamento 1 para $u = 0.39\ mm$.

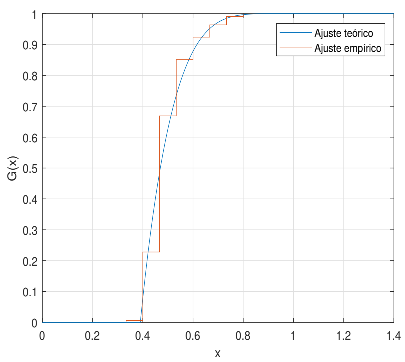
Parâmetros	Valores Estimados	Erro Padrão	Limites de confiança	
			Alto	Baixo
ξ	-0.2121	$\pm$0.0469	-0.1201	-0.3041
β	0.1237	$\pm$0.0089	0.1423	0.1075

Com o intuito de verificar a qualidade do ajuste, foram obtidos o gráfico de comparação da função de probabilidade acumulada e o gráfico quantil-quantil. A Fig.4 apresenta os gráficos construídos.

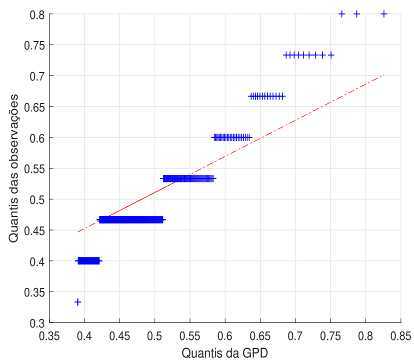
Foi observado que a comparação dada pelo gráfico da função de distribuição acumulada representa, por si só, um bom ajuste, onde a curva do ajuste teórico se adéqua bem ao ajuste empírico.

O gráfico quantil-quantil apresenta uma melhor adequação nas extremidades, enquanto nos quantis centrais o ajuste não é de todo adequado. Contudo, como a análise se pauta em análises de valores extremos e, neste caso, o peso na cauda é maior, o gráfico quantil-quantil obtido é aceitável. Ademais, é possível observar que o comportamento dos pontos não descreve bem uma reta. Isto acontece, simplesmente, devido à resolução das observações prevista pelo ensaio IRIS e, por consequência, tem-se uma quantidade elevada de dados com valores iguais de corrosão, dando a característica de um gráfico quantil-quantil para observações discretas.

É importante ressaltar que a reta em vermelho na Fig.4(b) representa apenas um ajuste para os pontos do gráfico em questão e tem pouco significado para o diagnóstico realizado.



(a) Função de distribuição acumulada estebelecendo uma comparação com o ajuste teórico e o empírico.



(b) Gráfico quantil-quantil da população de tubos em análise.

Figure 4: Gráficos da função de distribuição acumulada e quantil-quantil para diagnóstico do ajuste.

Uma vez verificada a qualidade do ajuste, as observações referentes à população foram aplicadas no algoritmo para definição do tamanho amostral *a posteriori*.

O erro utilizado para verificação do tamanho amostral foi $e = \frac{\alpha}{2} c_{maxreal}$. Ou seja, quando $|c_{max} - c_{maxreal}| \leq \frac{\alpha}{2} c_{maxreal}$, o tamanho amostral *a posteriori* é definido.

Como resultado, a rotina foi finalizada quando $c_{max} = 0.8168\ mm$, onde o $c_{maxreal} = 0.8\ mm$. Os resultados obtidos para o tamanho amostral encontrado estão presentes na Tabela 3.

Table 3: Resultados retornados pelo algoritmo, utilizando $u = 0.39\ mm$.

Parâmetros	Valores Estimados	Tamanho Amostral *a posteriori*
ξ	-0.1957	66 observações
β	0.1260	

Desta forma, de acordo com o algoritmo utilizado, somente com 66 observações seria possível obter uma amostra representativa pautada na extrapolação. Ou seja, este tamanho amostral é apropriado e com ele seria possível modelar a distribuição e realizar uma boa estimativa para a corrosão máxima na região não inspecionada.

4.2 Equipamento de Processo 2

A população referente ao segundo equipamento analisado consiste de 250 tubos ensaiados, portanto, $N = 250$. A Tabela 4 apresenta as características relacionadas aos tubos do equipamento em questão.

Table 4: Características dos tubos analisados referentes ao equipamento de processo 2.

diâmetro (mm)	16.94
comprimento (mm)	3466.7
$e_{nominal}$ (mm)	1.4067
N	250

Como realizado para o primeiro equipamento, os dados obtidos pela inspeção IRIS foram parametrizados confome a Eq.(15). Um novo conjunto de observações foi obtido, denotado por $x^{(2)} = [x_1^{(2)}, ..., x_N^{(2)}]$.

Para definição do limiar ótimo relacionado às observações referentes à população, foram aplicados os métodos do *MRL* e do *Parameter Stability Plot*. A Fig.**??** apresenta o gráfico obtido para o *MRL*.

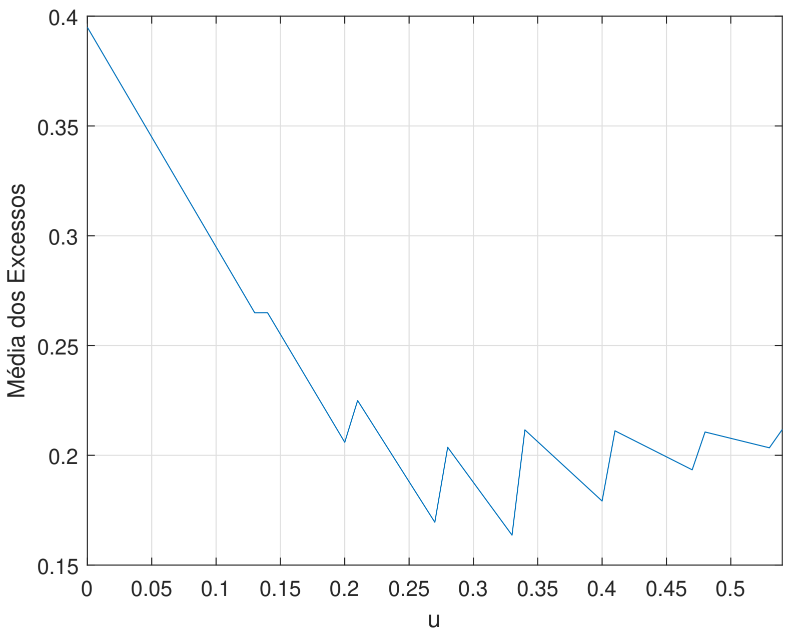
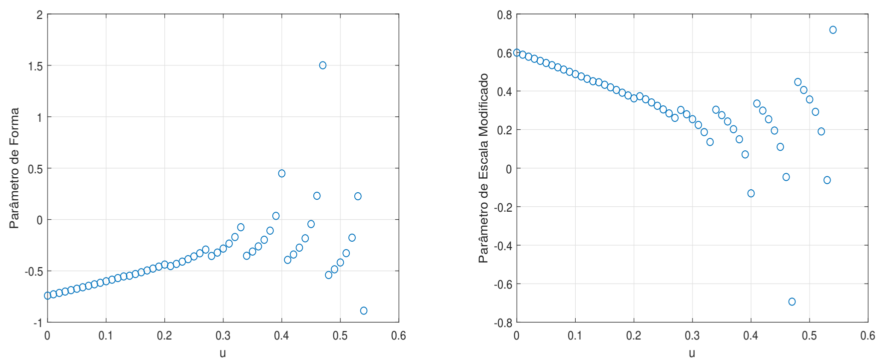


Figure 5: *MRL* referente à população do equipamento 2.

Neste caso, a análise inicial do *MRL* é mais complicada e menos direta com relação ao *MRL* apresentado na Fig.2. Desta forma, faz-se necessário realizar a análise dos gráficos referentes *Parameter Stability Plot* antes de se fazer uma análise inicial.

A Fig.6 apresenta os dois gráficos obtidos para o *Parameter Stability Plot*.



(a) Estabilidade do parâmetro de forma.

(b) Estabilidade do parâmetro de escala modificado.

Figure 6: *Parameter Stability Plot* referente à população do equipamento 2.

Foi possível verificar, pela Fig.6, que o comportamento do gráfico é aproximadamente constante até $u = 0.27$. O decrescimento aproximadamente constante apresentado na 2 vai até $u = 0.27$. Portanto, com base nos dois métodos foi definido $u = 0.27\ mm$.

Com o limar ótimo definido, a função da máxima verossimilhança foi aplicada e os dados ajustados na GPD, utilizando um nível de significância $\alpha = 0.05$. Os parâmetros estimados são apresentados na Tabela 5.

Table 5: Valores estimados para a distribuição referente à população de tubos observados no equipamento 2 para $u = 0.27\ mm$.

Parâmetros	Valores Estimados	Erro Padrão	Limites de confiança	
			Alto	Baixo
ξ	-0.2927	$\pm$0.0473	-0.2000	-0.3854
β	0.1817	$\pm$0.0143	0.2120	0.1557

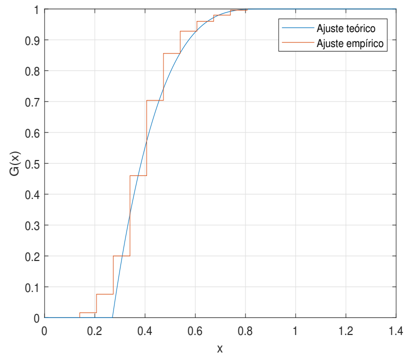
Para diagnóstico do ajuste o gráfico de comparação da função de distribuição acumulada e o gráfico quantil-quantil foram obtidos. A Fig.7 apresenta os gráficos construídos para o diagnóstico.

É possível verificar que a comparação entre o ajuste teórico e o ajuste empírico da função de distribuição acumulada, apresentado na Fig.7(a) se adequam bem, principalmente na extremidade.

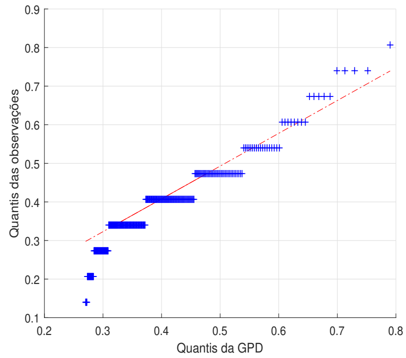
A análise do gráfico quantil-quantil, evidenciado na Fig.7(b), é semelhante à realizada no equipamento 1. O comportamento do gráfico pode ser explicado da mesma forma que gráfico apresentado em Fig.4(b). É mostrado que, principalmente nas extremidades, o gráfico quantil-quantil está adequado. Ressaltando que a reta em vermelho não representa algo muito significativo a análise e esta é somente o ajuste dos pontos apresentados no gráfico.

Após o diagnóstico, as observações foram aplicadas na rotina para determinação do tamanho amostral *a posteriori*. O erro utilizado para verificação do tamanho amostral foi $e = \dfrac{\alpha}{2} c_{maxreal}$, seguindo o mesmo padrão de análise que foi utilizado para o equipamento 1.

Como resultado, a rotina foi finalizada quando $c_{max} = 0.7583\ mm$, sendo $c_{maxreal} = 0.74\ mm$. Os resultados obtidos para o novo tamanho amostral estão



(a) Função de distribuição acumulada estebelecendo uma comparação com o ajuste teórico e o empírico.



(b) Gráfico quantil-quantil da população de tubos em análise.

Figure 7: Gráficos da função de distribuição acumulada e quantil-quantil para diagnóstico do ajuste.

presentes na Tabela 6.

Table 6: Resultados retornados pelo algoritmo, utilizando $u = 0.27\ mm$.

Parâmetros	Valores Estimados	Tamanho Amostral *a posteriori*
ξ	-0.2831	119 observações
β	0.1861	

Portanto, segundo o algoritmo, caso fossem tomadas 119 observações, o modelo estatístico seria válido e uma boa estimativa para a extrapolação seria fornecida.

5 Conclusões

Foi possível, utilizando o método apresentado, realizar uma estimativa para um tamanho amostral *a posteriori* para dois equipamentos de processo compostos por tubos submetidos á corrosão generalizada, pautando a verificação no valor de retorno estimado para uma região não inspecionada, isto é, na extrapolação das observações.

O método empregado se mostrou eficiente. Porém, este funciona apenas caso a população analisada seja conhecida, o que inviabiliza a tratativa para todo e qualquer equipamento. Isto ocorre pois a verificação da extrapolação fica comprometida.

Ademais, a contribuição fornecida pelo método é de fato importante para aplicações de monitoramento de componentes industriais, visto que, com o tamanho amostral retornado, sendo este representativo, além de se ter mais assertividade no que tange às tomadas de decisão quanto ao tamanho amostral a ser obtido na inspeção, a extrapolação para as áreas também é contemplada com uma boa estimativa.

6 Agradecimentos

Agradeço, primeiramente, ao meu orientador Antônio Marcos Gonçalves de Lima por todo tempo tomado e dedicação na ação de me orientar.

Agradeço, também, à empresa Petrobrás, por fomentar o avanço científico e tecnológico, financiamento da pesquisa apresentada e por acreditar no ensino superior brasileiro e fornecer oportunidade de mostrar que as Ifes (Institutos Federais de Ensino Superior) estão aptas a desenvolver pesquisas de ponta.

References

[1] C. Anderson, D. Carter, and Peter Cotton. Wave climate variability and impact on offshore design extremes. 2001.

[2] A. A. Balkema and L. de Haan. Residual life time at great age. *The Annals of Probability*, 2(5):792–804, 1974.

[3] Vladimir Belitsky and Francisco Martins Moreira. *Emprego do método 'Peaks-over-threshold' na estimação de risco; uma exposição abragente, detalhada mas simples*. Apostila do mini-curso da 3-d Brazilian Conference on Statistical Modelling in Insurance and Finance, 2007.

[4] Esther Bommier. Peaks-over-threshold modelling of environmental data. Examensarbete i matematik, Uppsala University, 2014.

[5] Suart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2011.

[6] A. C. Davinson and R. L. Smith. Models for exceedances over high thresholds. *Jounal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442, 1990.

[7] James Pickands III. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131, 1975.

[8] Mohamed Khalifa, Faisal Khan, and Mahmoud Haddara. Bayesian sample size determination for inspection of general corrosion of process components. *Journal of Loss Prevention in the Process Industries*, 25(1):218–223, 2012.

[9] Ricardo Schayer Sabino. Inspeção de feixes tubulares de trocadores de calor. Dissertação de Mestrado, Universidade Federal de Minas Gerias, 2008.

[10] Carl Scarrott and Anna MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT - Statiscal Journal*, 10(1):33–60, 2012.

[11] Hwei-Yang Tan. Analysis of corrosion data for integrity assessments. Ph.D. Thesis, Brunel University, 2017.