UMA APLICAÇÃO DE RANDOM SURVIVAL FORESTS NA AVALIAÇÃO DE DADOS DE FALHA DE BOMBAS CENTRÍFUGAS SUBMERSAS

Ricardo de Melo e Silva Accioly[1]
Professor Adjunto do Instituto de matemática e Estatística da UERJ
raccioly@ime.uerj.br

Rafael de O. Valle dos Santos[2]
PETROBRAS, Rio de Janeiro, Brasil
rvsantos@petrobras.com.br

RESUMO

Neste trabalho será presentada uma aplicação de *random survival forests* na análise de dados de falhas de bombas centrífugas submersas utilizadas na elevação artificial de petróleo. A análise de dados de tempo até um evento é um campo bem desenvolvido na área de estatística, onde o uso de métodos não paramétricos, semi-paramétricos e paramétricos tem sido muito desenvolvido e aplicado. O uso de métodos baseados em árvores, que pode ser enquadrado no contexto do aprendizado estatístico ou de máquina, foi desenvolvido paralelamente, tendo sido usado inicialmente em problemas de regressão e classificação, posteriormente em análise de sobrevivência (*survival trees*). Neste artigo inicialmente é ajustada uma *survival tree*, que permite maior interpretabilidade, mas que geralmente não gera boas previsões, posteriormente são usadas as *random survival trees* buscando aprimorar a acurácia das previsões. Finalmente a acurácia de previsão destas últimas é comparada com os modelos de Cox (semi-paramétrico) e um modelo de tempo de vida acelerado usando o escore de Brier.

1. INTRODUÇÃO

O uso de métodos baseados em árvores, que se enquadra no contexto do aprendizado estatístico ou de máquina, como pode ser visto em [1], é muito usado em problemas de regressão e classificação. A abordagem através de árvores envolve estratificar ou segmentar o espaço de preditores (variáveis explicativas) em uma série de regiões. O conjunto de regras de divisão usados para segmentar o espaço de resultados pode ser resumido em uma árvore, esse tipo de abordagem é muitas vezes denominado de método de árvore de decisão.

O uso de árvores em sua forma original [2], representa uma maneira simples e útil para interpretação dos fatores que influenciam uma determinada variável dependente. No entanto, eles muitas vezes não são competitivos em termos de acurácia de previsão. Para aprimorar seus resultados podem ser aplicados os métodos de *bagging* [3] e *random forests* [4], que geram múltiplas árvores que são combinadas para produzir uma única previsão de consenso (ensemble). Combinar muitas árvores pode, muitas vezes, resultar em melhorias na acurácia da previsão, mas como contraponto há perda na interpretação da árvore.

O método *random forests* (RF) proporciona uma melhoria sobre o *bagging* por meio de um pequeno ajuste que descorrelaciona as árvores. Isso reduz a variância quando fazemos a média das árvores [1]. Como no método *bagging*, construímos uma série de árvores de decisão a partir de amostras *bootstrap* de treinamento. Entretanto, ao construir essas árvores, cada vez que uma divisão é considerada, uma seleção aleatória de m preditores é escolhida como candidatos a dividir o conjunto completo de p preditores. A divisão somente é permitida para apenas um desses m preditores. Uma nova seleção de m preditores é tomada a cada divisão e normalmente se adota para m um valor que é aproximadamente igual à raiz quadrada do número total de preditores.

As primeiras tentativas de aplicação do método de arvores em análise de sobrevivência (*survival trees* - ST) foram apresentadas em [5] e [6]. As principais diferenças entre uma árvore de sobrevivência e a árvore de decisão padrão é que na primeira temos a presença de dados censurados, que é característico neste tipo de análise, ocasionando a necessidade de outros critérios de partição. As regras de partição em árvores de

1 DSc. em Engenharia de Produção UFRJ
2 DSc em Engenharia Elétrica PUC-RIO

sobrevivência, em geral, se baseiam em dois métodos. Medidas de distância em um nó que buscam maximizar a diferença entre observações ou medidas de pureza do nó que buscam agrupar observação semelhante em um único nó. Em [5] foi proposta uma medida de distância baseada no teste de *logrank* e em [6] foi proposta uma medida de pureza baseada nas estimativas de Kaplan-Meier.

As *random survival forests* (RSF) foram propostas por Ishwaran et al. [7] permitindo que as RF pudessem ser aplicadas em dados censurados à direita. A metodologia das RSF segue o mesmo algoritmo que das RF que foi desenvolvido por Breiman [4].

Wang et al. [8] fizeram uma excelente consolidação dos métodos de aprendizado de máquina para análise de dados de sobrevivência. A figura 1 mostra um recorte do que eles apresentaram, selecionando apenas os métodos relacionados ao aprendizado de máquina, com destaque para os relacionados a árvores de sobrevivência.

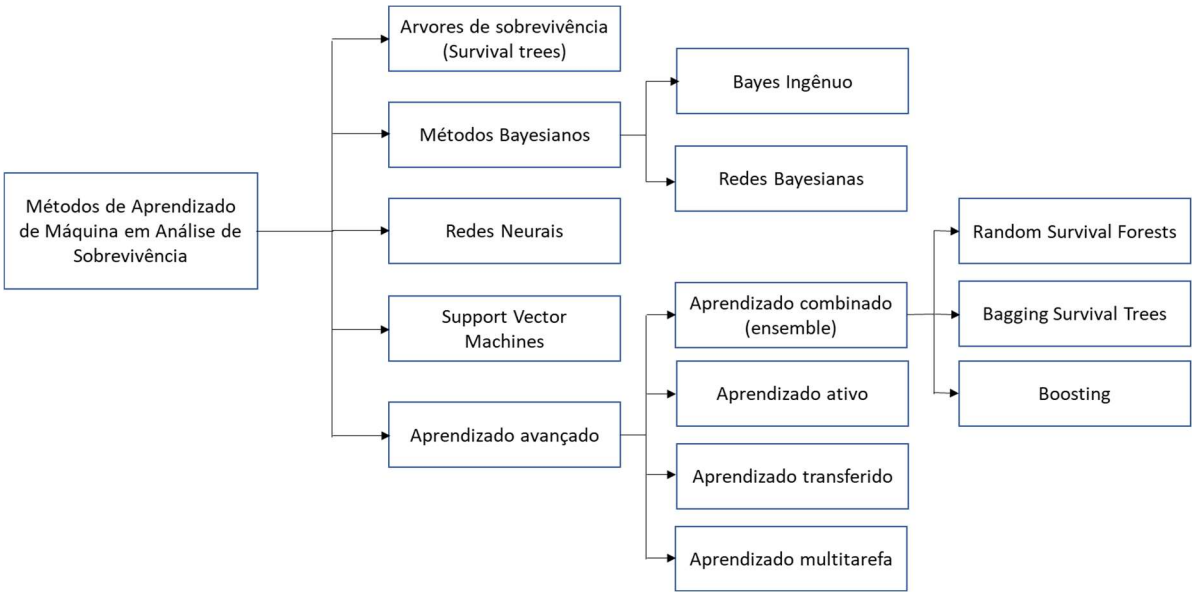

Figura 1 – Métodos de aprendizado de máquina em Análise de Sobrevivência. (Fonte: Wang et al. [8])

Na Seção 2, a seguir, é feita uma breve descrição dos dados utilizados. Na Seção 3 é detalhada a metodologia das ST e RSF. Na Seção 4 são apresentados os resultados da análise dos dados de bombas centrífugas submersas utilizadas na elevação artificial de petróleo e finalmente na Seção 5 é feita uma discussão sobre os resultados obtidos e apontando algumas conclusões com relação ao uso desta metodologia.

2. DESCRIÇÃO DOS DADOS UTILIZADOS

Um reservatório de petróleo pode ter pressão suficiente para levar o hidrocarboneto da sua formação rochosa à superfície sem o uso de qualquer mecanismo de elevação. Esses poços são conhecidos como poços surgentes. Quando a pressão do reservatório é baixa, os fluidos oriundos do poço não têm capacidade de surgir na superfície, sendo então necessário o uso de um método de elevação artificial.

Um dos métodos de elevação utilizados são as bombas centrífugas submersas (BCS), que são o escopo desta análise de falhas deste trabalho. Na figura 2 é apresentado uma representação esquemática de uma BCS.

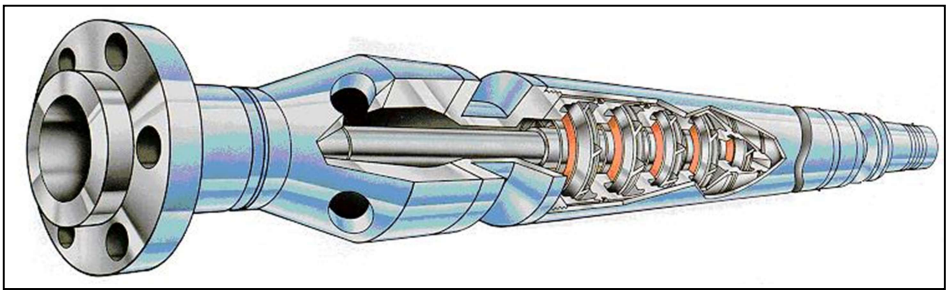

Figura 2 – Esquema de uma bomba centrífuga submersa. (Fonte: Petrobras)

Os dados de tempo de vida das de bombas centrífugas submersas (BCS) analisados são relativos ao período de 1989 a 1994 e se referem às instalações de BCS no Polo Nordeste da Bacia de Campos, situado no Estado do Rio de Janeiro. A amostra contém 338 observações de tempos de vida (210 falhas e 128 censuras) e as respectivas variáveis explicativas da sua duração. Estes dados foram utilizados anteriormente em [9].

Tabela 1 – Descrição das possíveis variáveis explicativas

Variável	**Descrição**	**Unidade**
Fab	Fabricante da bomba e motor	0 e 1
Hp	Potência do motor	Hp
Sep	Existência ou não de separador de gás	0 e 1
Ang	Ângulo máximo de desvio	Graus
Temp	Temperatura média do reservatório	°C
Form	Reservatório produtor na época da instalação	0 e 1
Rgo	Razão gás-óleo média no ano da instalação	m3/m3
Bsw	Porcentagem média de água de formação no ano da instalação	%

3. METODOLOGIA

Para demonstrar o uso de *survival trees* foi escolhida a metodologia baseada na inferência condicional [10], que busca evitar o viés na seleção das variáveis a serem particionadas. Simplificadamente o algoritmo funciona da seguinte forma:

1. Teste a hipótese nula global de independência entre qualquer uma das variáveis explicativas (X) e a variável dependente (Y). Pare se essa hipótese não puder ser rejeitada. Caso contrário, selecione a variável explicativa (X) com maior associação à Y. Esta associação é medida por um teste de classificação linear baseado em permutações [10]. Utilizando-se a distribuição da estatística de classificação resultante, os p-valores são avaliados e a variável explicativa com o menor p-valor é a que tem a associação mais forte com a variável resposta;
2. Implemente uma divisão binária na variável de explicativa selecionada;
3. Repetir recursivamente os passos (1) e (2).

Maiores detalhes sobre a metodologia de árvores de inferencial condicional podem ser encontrados em Hothorn et al. [10].

A metodologia das RSF desenvolvida por Ishwaran et al. [7] segue os princípios gerais do que foi proposto por Breiman [4], mas com adaptações devido a censura à direita:

1. Gerar *B* amostras *bootstrap*;
2. Crescer uma árvore para cada amostra *bootstrap* $b = 1, \cdots, B$:
   a. Em cada nó da árvore selecione um subconjunto $m\ (m < p)$ das p variáveis explicativas;
   b. Entre todas as partições binárias das variáveis explicativas selecionadas em (a), ache a melhor partição entre dois subconjuntos (nós filhos) através de um critério de partição adequado a dados censurados à direita (ex.: *logrank* [11]);
   c. Repetir (a) e (b) recursivamente em cada nó filho até que um critério de parada seja encontrado.
3. Agregar toda informação obtida dos nós terminais das *B* arvores de sobrevivência para obter uma previsão de consenso (ensemble). A combinação é calculada através da média das previsões.

As RSF geram dois tipos de previsões, curvas de sobrevivência e a função de risco acumulada (FRA). Em cada nó terminal é gerada uma previsão da curva de sobrevivência e da função de risco acumulada. Seja *j* um nó terminal da árvore em que,

$$t_{1,j} < t_{2,j} < \cdots < t_{k(j),j}$$

são mortes distintas em *j*. Sejam $d^*_{i,j}$ e $R^*_{i,j}$ iguais ao número de mortes e de indivíduos sob risco no tempo $t_{i,j}$, sendo que aqui o asterisco em *d* e *R* representa uma amostra *bootstrap*. A FRA e a curva de sobrevivência para *j* são estimadas usando as estimativas de *bootstrap* dos estimadores de Nelson-Aalen e de Kaplan-Meier [11], que neste caso representam o que se denomina estimativas "in-bag" (*IB*).

$$H_j^{IB}(t) = \sum_{t_{i,j} \le t} \frac{d^*_{i,j}}{R^*_{i,j}} \qquad (1)$$

$$S_j^{IB}(t) = \prod_{t_{i,j} \le t} \left(1 - \frac{d^*_{i,j}}{R^*_{i,j}}\right) \quad (2)$$

Para se estimar $H(t|X)$ e $S(t|X)$ para uma variável explicativa *X*, nós seguimos as partições da árvore e localizaremos um nó terminal que contenha *X*. Este nó será único devido à natureza binária das árvores. A FRA e a curva de sobrevivência para *X* são as amostras de *bootstrap* de Nelson-Aalen e de Kaplan-Meier no nó terminal correspondente a *X*.

$$H^{IB}(t|X) = H_j^{IB}(t),\ se\ X \in j \quad (3)$$

$$S^{IB}(t|X) = S_j^{IB}(t), se\ X \in j \quad (4)$$

As amostras *bootstrap* usam em média 2/3 da amostra original ("in-bag") deixando cerca de 1/3 de fora da amostra ("out-of-bag" ou *OOB*) [1]. Sabendo disso é possível se obter estimativas *OOB* dos estimadores e Nelson-Aalen e de Kaplan-Meier. Estas estimativas nos permitem ter uma prévia do erro de previsão de uma amostra de teste.

$$H^{OOB}(t|X_{OOB}) = H_j^{IB}(t), se\ X \in OOB \quad (5)$$

$$S^{OOB}(t|X_{OOB}) = S_j^{IB}(t), se\ X \in OOB \quad (6)$$

As estimativas de consenso da FRA e da curva de sobrevivência são obtidas fazendo-se a média das estimativas obtidas nos nós terminais. Estas estimativas *IB* são dadas por,

$$\bar{H}^{IB}(t|X) = \frac{1}{B}\sum_{b=1}^{B} H_b^{IB}(t|X) \quad (7)$$

$$\bar{S}^{IB}(t|X) = \frac{1}{B}\sum_{b=1}^{B} S_b^{IB}(t|X) \quad (8)$$

O mesmo critério é utilizado para se obter as estimativas de consenso das amostras *OOB*. Considerando a contabilização dos casos em que $X \in OOB$ em N_{OOB}, temos que,

$$\bar{H}^{OOB}(t|X_{OOB}) = \frac{1}{N_{OOB}}\sum_{b\in N_{OOB}} H_b^{IB}(t|X_{OOB}) \quad (9),$$

$$\bar{S}^{OOB}(t|X_{OOB}) = \frac{1}{N_{OOB}}\sum_{b\in N_{OOB}} S_b^{IB}(t|X_{OOB}) \quad (10)$$

Observar que as estimativas *OOB* só servem para estimar o erro de previsão do modelo. As estimativas *IB*, no entanto, podem ser usadas para previsão da FRA e da curva de sobrevivência.

Os erros de previsão são obtidos através do índice de concordância de Harrell (C) [7]. Este índice é muito usado em modelos de sobrevivência. Ele pode variar de 0 a 1, onde 1 significa desempenho perfeito e 0 significa pior desempenho possível. Se um modelo não levasse em conta qualquer informação dos dados, ou seja, fosse feita uma previsão aleatória, então o índice C correspondente seria em torno de 0,5. No caso das RSF este índice é calculado a partir da mortalidade. Seja $t_1 < \cdots < t_m$ o conjunto de tempos (únicos) em que ocorrem eventos no conjunto de aprendizado (*IB*). A mortalidade de consenso para uma variável explicativa *X* é definida por,

$$\bar{M}^{IB}(X) = \sum_{j=1}^{m} \bar{H}^{IB}\left(t_j|X\right) \quad (11)$$

A estimativa acima nos dá o número de mortes esperadas se todos os casos fossem similares a *X*. Para o cálculo do índice de concordância de Harrell (índice C) usamos a mortalidade de consenso do conjunto *OOB* que é definida por [7,11],

$$\bar{M}^{OOB}(X_{OOB}) = \sum_{j=1}^{m} \bar{H}^{OOB}\left(t_j|X_{OOB}\right) \quad (12)$$

A partir do valor acima é calculado o índice C conforme pode ser visto em [7]. A taxa de erro (TE) é $TE = 1 - C$ e apresenta valores entre $0 \leq TE \leq 1$. Quanto mais próximo de zero melhor, sendo que um valor igual a 0,5 corresponde a um modelo semelhante a um chute aleatório.

4. APLICAÇÃO

Para a aplicação das *survival trees* foram geradas duas amostras aleatórias de treino, através de sementes diferentes, contendo 90% dos dados originais. O objetivo ao se criar duas amostras de treino foi para destacar o principal problema das *survival trees* que é sua instabilidade em função de mudança no conjunto de dados de treino. Para obtenção das arvores foi utilizada a biblioteca do R partykit [12].

A figura 3 apresenta a árvore obtida para o 1º conjunto de treino e a Figura 4 apresenta a árvore para o 2º conjunto de treino. É fácil observar a diferença entre os dois resultados. No 1º caso a **rgo** é vista como uma variável importante na partição da árvore, enquanto na 2ª ela desaparece. Esta instabilidade afeta os resultados dos nós terminais onde vemos que na 1ª árvore existem quatro nós terminais com as respectivas curvas de sobrevivência e na 2ª árvore apenas três nós terminais. Apesar da clareza da apresentação dos resultados, este ponto fraco limita a utilidade das *survival trees*.

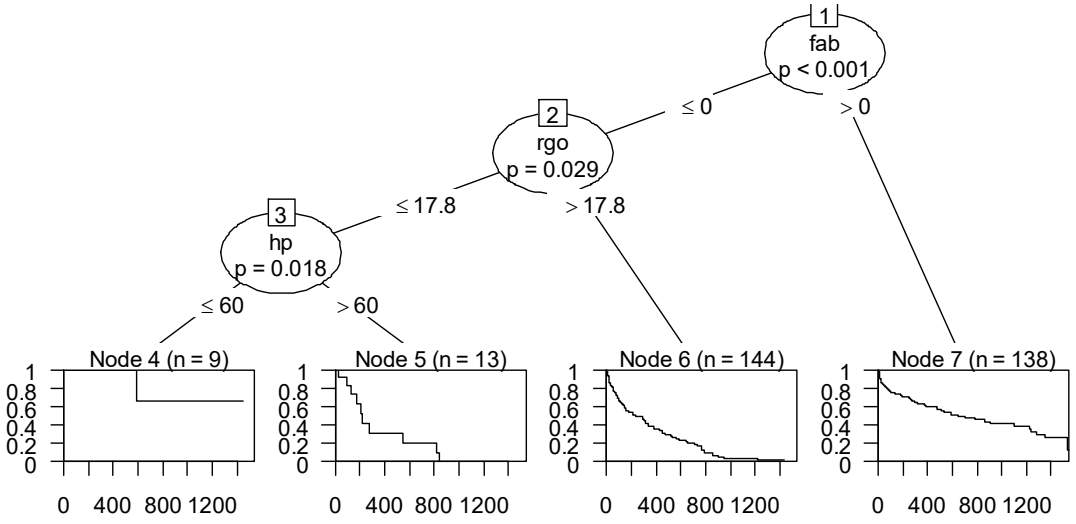

Figura 3 – *Survival tree* para o 1º conjunto de treino. (Fonte: Autor)

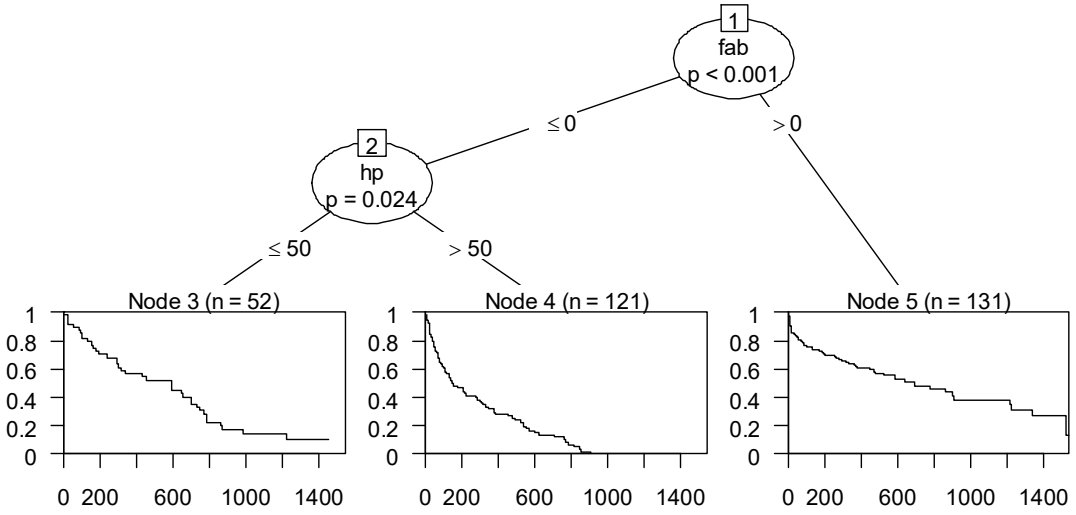

Figura 4 – *Survival tree* para o 2º conjunto de treino. (Fonte: Autor)

A metodologia das RSF não tem este problema, pois nela as previsões são baseadas no consenso (ensemble) de um conjunto de amostras *bootstrap*, mas é perdida a clareza da visualização do resultado da *survival tree*, ou seja, a árvore com suas partições e grupos de dados.

Para desenvolver a aplicação foi utilizada a biblioteca do R randomForestSRC [13]. Alguns parâmetros necessitam ser definidos antes da obtenção da RSF:

1. Vamos gerar RSF para B =100 e B =600 amostras *bootstrap*;
2. Como temos p = 8 variáveis explicativas, vamos definir com o número máximo de variáveis a serem escolhidas aleatoriamente a cada etapa como sendo $m = \sqrt{8} \cong 3$, que é o valor recomendado pela literatura;
3. O nó terminal deverá ter pelos menos 10 casos únicos, que foi o valor sugerido após otimização deste parâmetro.

Na figura 5 são apresentados os resultados obtidos considerando $B = 100$, através dele vemos que a taxa de erro desce até 20 árvores e depois começa a subir e posteriormente temos nova queda. Em termos de importância das variáveis explicativas analisadas vemos que **hp** e **fab** são bem mais relevantes que as demais. A medida de importância da variável é obtida de forma não paramétrica usando a medida de Breiman-Cutler [7, 11] que costuma ser denominada de importância de permutação.

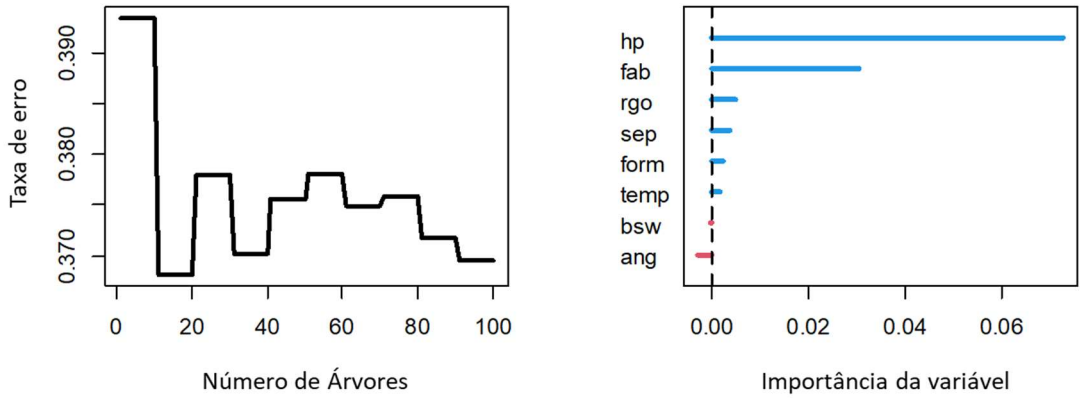

Figura 5 – (a) Taxa de Erro por n° de árvores e (b)Importância das variáveis (Fonte: Autor)

Na figura 6 são apresentados os resultados obtidos considerando $B = 600$, através dele vemos que a taxa de erro desce no início até 350 árvores e depois flutua até chegarmos a 600 árvores. Esta quantidade de amostras apresentou maior estabilidade na taxa de erro, portanto será usada nas demais análises. Observar que esta taxa de erro é calculada com base nas amostras *OOB* e não com um conjunto de teste. Em termos de importância das variáveis explicativas foi obtido resultado semelhante ao anterior em que **hp** e **fab** se destacam. O resultado de importância de variável está compatível com o encontrado em [9] no qual foram usados métodos semi-paramétricos e paramétricos.

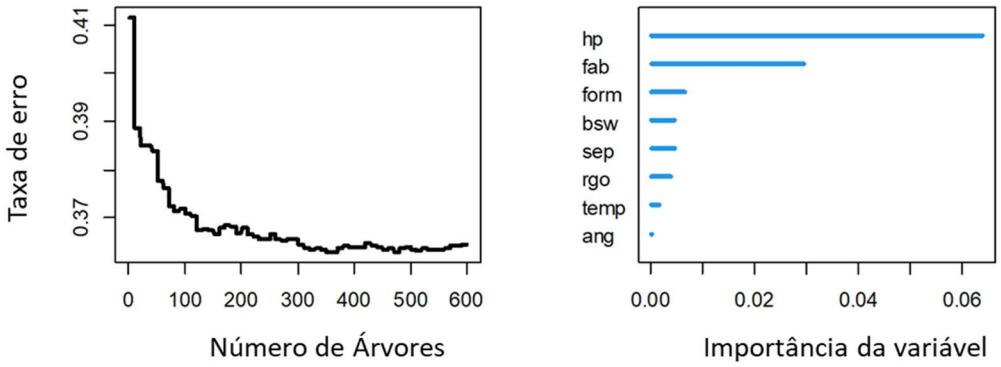

Figura 6 – (a) Taxa de Erro por n° de árvores e (b)Importância das variáveis (Fonte: Autor)

A partir dos resultados da RSF é possível se obter curvas de sobrevivência baseados em valores fixos das variáveis explicativas. As variáveis mais relevantes são a potência do motor (**hp**) e o fabricante (**fab**). Inicialmente foi selecionado o **hp** para variar, mantendo-se o fabricante igual a 0 (categoria) e o separador em 1 (presente), as variáveis contínuas ficam fixadas nos seus valores medianos. A figura 7 mostra as duas curvas de sobrevivência para os valores de potência do motor de 60 HP e 105 HP. Fica evidente que as BCS com potências mais elevadas têm maior probabilidade de falhas e consequentemente menores chances de sobrevivência.

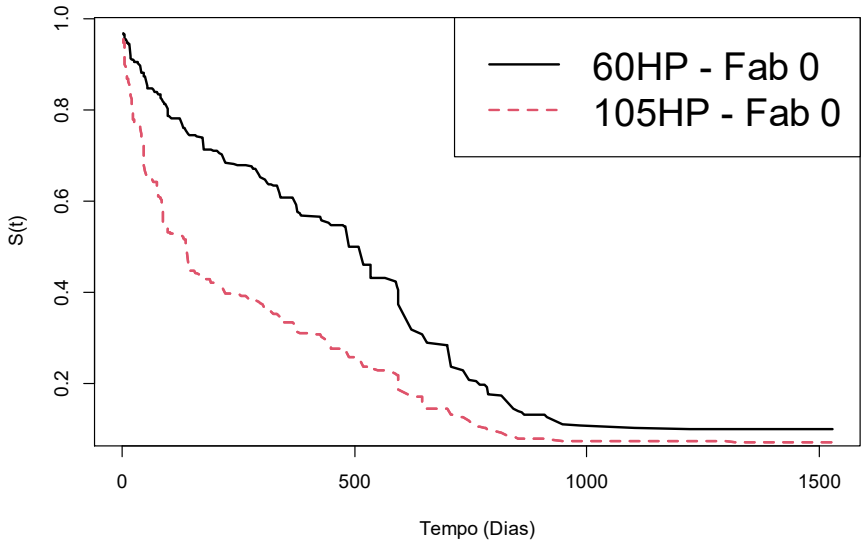

Figura 7 – Curva de Sobrevivência para BCS com potência de 60 HP e 105 HP. (Fonte: Autor)

Agora serão apresentadas curvas de sobrevivência mantendo o **hp** constante em 90 e variando o fabricante (**fab** = 0 ou 1), com as demais variáveis mantidas no mesmo valor anterior. A figura 8 mostra as duas curvas de sobrevivência para os dois fabricantes na potência do motor de 90 HP. A partir de 500 dias as duas curvas de sobrevivência se separam com a do fabricante 1 indicando maiores probabilidades de sobrevivência.

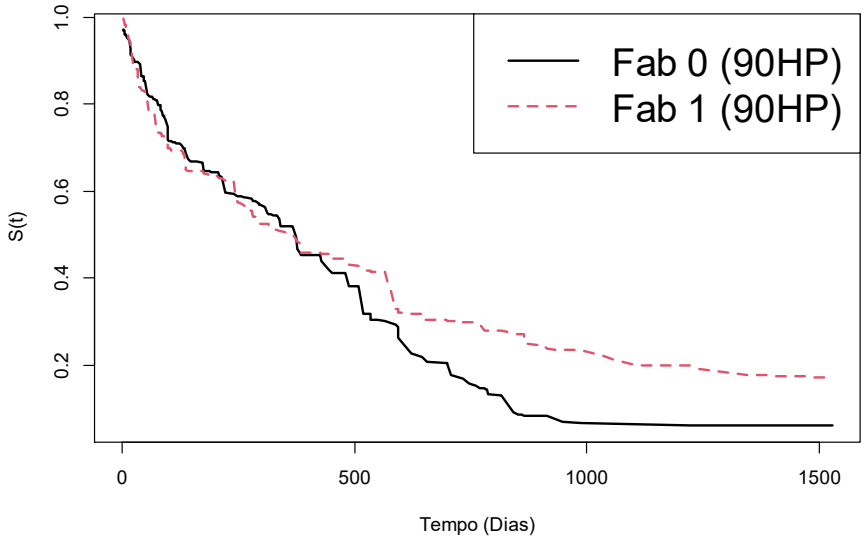

Figura 8 – Curva de Sobrevivência para BCS dos dois fabricantes analisados. (Fonte: Autor)

Quando estamos lidando com modelos paramétricos podemos avaliar qual seria o melhor modelo através do teste da razão de verossimilhança [14], a partir de um modelo que englobe todos os demais. As RSF são modelos não paramétricos, portanto é necessário o uso de alguma outra medida para avaliar o modelo e principalmente possibilitar a comparação com alternativas de modelagem. O Brier score é uma das medidas usadas na avaliação de modelos de sobrevivência, podendo ser usada na geração de curvas de erro de predição para diferentes tipos de modelos [15]. Para informações detalhadas do processo de cálculo consultar [15, 16].

A biblioteca do R pec [17] nos permite obter estar curvas de erro de predição, que no caso foram utilizadas para a RSF, para um modelo semi-paramétrico de Cox [18] e um modelo paramétrico de tempo de vida acelerado com base na distribuição de Weibull [18], utilizando as mesmas variáveis explicativas e conjunto de dados. Estes dois últimos foram ajustados através da biblioteca do R rms [19]. Na figura 9 são apresentados os resultados obtidos do Brier score ao longo do tempo, quanto menor o valor do escore melhor o modelo, indicando que a RSF apresentou resultados muito próximos do modelo de Cox e do modelo de tempo de vida acelerado de Weibull. A RSF apresentou valores ligeiramente maiores em alguns intervalos de tempo. O modelo de referência é baseado na estimativa de Kaplan-Meier (não paramétrica). Chama a atenção que, apesar de são um modelo não paramétrico, a RSF apresentou um desempenho muito mais próximo do modelo paramétrico e semi-paramétrico do que o modelo de referência. Este comportamento está associado a melhor utilização das variáveis explicativas.

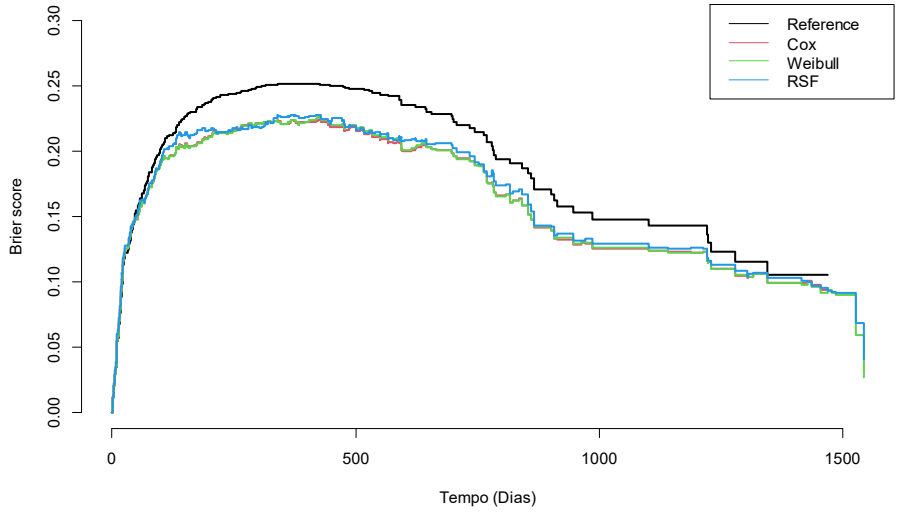

Figura 9 – Curvas de erro de predição para as RSF, modelo de Cox e Weibull. (Fonte: Autor)

5. CONCLUSÕES

Neste artigo avaliou-se a utilização de um modelo de *random survival forests* (RSF) na análise de dados de falhas de bombas centrífugas submersas utilizadas na elevação artificial de petróleo. As RSF são muito usadas na área médica e nas ciências biológicas, com menor presença na área de engenharia. A utilização das RSF tem como principal apelo sua característica não paramétrica que evita suposições teóricas que são comuns em modelos paramétricos (MP) ou semi-paramétricos (MSP), tais como, a suposição de taxas proporcionais. As RSF criam suas árvores com variáveis contínuas ou discretas, como ocorre nos outros modelos.

Nos modelos MP e SMP podemos avaliar a significância estatística das variáveis explicativas, o que não é possível nas RSF. As RSF geram informação sobre a importância das variáveis, que neste artigo, apresentou resultados semelhantes ao obtido em [9], mas sem o mesmo poder de avaliação (ou seleção) de variáveis.

Para o conjunto de dados analisado o erro de predição da RSF, através do escore de Brier [15], não teve desempenho melhor que o modelo de Cox e o modelo de tempo de vida acelerado de Weibull. Cabe-se destacar que este artigo foi uma aplicação inicial das RSF neste conjunto de dados, onde as opções de otimização dos parâmetros das RSF foram limitadas, o que dá margem para novos estudos e simulações visando aprimorar esta primeira aplicação, ou seja, buscar uma combinação de parâmetros que leve a melhores resultados.

A experiência deste trabalho sugere que as RSF sejam incorporadas ao conjunto de modelos potenciais para análise de dados de falha, de forma a se adicionar na análise uma abordagem não paramétrica que permita a incorporação de variáveis explicativas.

6. REFERENCIAS:

[1] JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R., An Introduction to Statistical Learning, with applications in R, Springer, (2013)

[2] BREIMAN, LEO, FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J., Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, (1984)

[3] BREIMAN, L., Bagging predictors, Machine Learning, 24(2), 123-140, (1996)

[4] BREIMAN, L., Random Forests, Machine Learning, 45(1), 5-32, (2001)

[5] CIAMPI, A., THIFFAULT, J., NAKACHE, J. P., ASSELAIN, B., Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. Computational statistics & data analysis 4, 3 (1986), 185–204, (1986)

[6] GORDON, L., OLSHEN, R. A., Tree-structured survival analysis. Cancer treatment reports 69, 10, 1065–1069, (1985)

[7] ISHWARAN, H., KOGALUR, U.B., BLACKSTONE, E.H., LAUER, M.S., Random survival forests, The Annals of Applied Statistics, 2(3), 841-860, (2008)

[8] WANG, P., LI, Y., REDDY, C.K., Machine Learning for Survival Analysis: A Survey, ACM Computing Surveys, Vol. 51, Issue 6, Article No.: 110, pp 1–36, (2019)

[9] ACCIOLY, R. M. S., Análise da duração do tempo de vida de bombas centrífugas submersas. Dissertação de MSc., COPPE/UFRJ, RJ, RJ, Brasil, (1995)

[10] HOTHORN, T., HORNIK, K., ZEILEIS, A., Unbiased recursive partitioning: A conditional inference framework. Journal of Computational Graphical Statistics.V.15, N.3, pp 651–74, (2006)

[11] ISHWARAN, H., LU, M., Random Survival Forests, Wiley StatsRef: Statistics Reference Online, 2014–2019 John Wiley & Sons, Ltd. DOI: 10.1002/9781118445112.stat08188, (2019)

[12] HOTHORN, T., SEIBOLD, H., ZEILEIS, A., partykit: A Toolkit for Recursive Partytioning. R package version 1.2-15, URL https://cran.r-project.org/package=partykit, (2021)

[13] ISHWARAN, H., KOGALUR, U.B., randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 2.11.0, URL https://cran.r-project.org/package=randomForestSRC , (2021)

[14] BUSE, A., “The likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note”, The American Statistician, 36, 153-157 , (1982).

[15] MOGENSEN, U.B., ISHWARAN, H., GERDS, T. A., Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. Journal of Statistical Software, 50(11), 1-23. URL https://www.jstatsoft.org/v50/i11. (2012)

[16] GERDS, T.A., SCHUMACHER, M., Consistent estimation of the expected brier score in general survival models with right-censored event times. Biom. J., 48 (6), 1029–1040, (2006)

[17] GERDS, T.A., pec: Prediction Error Curves for Risk Prediction Models in Survival Analysis. R package version 2020.11.17, URL https://cran.r-project.org/package=pec, (2020)

[18] HOSMER, D. W., LEMESHOW, S., MAY, S., Applied Survival Analysis: Regression Modeling of Time ‐ to ‐ Event Data, Second Edition, John Wiley & Sons, (2008)

[19] HARREL, F.E., rms: Regression Modeling Strategies. R package version 6.2-0, URL http://cran.r-project.org/package=rms, (2021)