UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO

FACULDADE DE CIÊNCIA E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

MODELOS DE MACHINE LEARNING EM ANÁLISE DE SOBREVIVÊNCIA NA PRESENÇA DE FRAÇÃO DE CURA

JOÃO DEBASTIANI NETO

PRESIDENTE PRUDENTE

2026

João Debastiani Neto

Modelos de Machine Learning em análise de sobrevivência na presença de fração de cura

Relatório Final Científico do Programa de Pós-Doutorado submetido ao Departamento de Estatística, Faculdade de Ciência e Tecnologia, Universidade Estadual Paulista

Supervisor: Prof. Dr. Ricardo Puziol de Oliveira

Presidente Prudente

2026

SUMÁRIO

1 Atividades de Desenvolvimento do Projeto Realizadas **1**

1.1 Identificação . . . 1

1.2 Artigos Científicos Produzidos na Temática do Projeto . . . 1

1.3 Artigos Científicos em Produção na Temática do Projeto . . . 1

1.4 Artigos Científicos Produzidos em Outra Temática (Com Interseção com Temática do Projeto) 2

1.5 Apresentação de Trabalhos em Eventos Científicos . . . 2

1.6 Participação em Eventos como Avaliador de Trabalhos . . . 3

1.7 Participação em Grupos de Pesquisa . . . 3

1.8 Aulas Ministradas . . . 3

2 Descrição e Desenvolvimento do Projeto **4**

2.1 Introdução . . . 4

2.2 Objetivos do Projeto . . . 5

2.2.1 Objetivos Específicos . . . 5

2.3 Materiais e Métodos . . . 5

2.3.1 Estrutura dos Dados de Sobrevivência . . . 5

2.3.2 Estratégia de Particionamento de Dados . . . 6

2.3.3 Modelos com Fração de Cura . . . 7

2.3.4 Modelos de Transformação de Risco . . . 9

2.4 Cronograma Desenvolvido . . . 11

3 Resultados e Discussão **14**

3.1 Artigo 1 . . . 14

3.1.1 Dados do estudo . . . 14

3.1.2 Modelos de Transformação de Risco . . . 14

3.1.3 Incorporando a Taxa de Cura . . . 15

3.1.4 Resultados e Discussão . . . 16

3.2 Artigo 2 . . . 21

3.2.1 Dados do estudo . . . 21

3.2.2 Fundamentação Estatística . . . 22

3.2.3 Resultados e Discussão . . . 23

3.3 Artigo 3 . . . 29

3.3.1 Dados do Estudo . . . 29

3.3.2 Modelo de Transformação Monótona Adaptativa aos Dados 31
3.3.3 Resultados 32

4 Considerações Finais **35**

Apêndice A: Atividades Docente **42**

Apêndice B: Outros Artigos Desenvolvidos **43**
4.1 Artigo 1 43
4.2 Artigo 2 44
4.3 Artigo 3 45
4.4 Artigo 4 46
4.5 Artigo 5 47
4.6 Artigo 6 48

Comprovantes e Declarações **49**

CAPÍTULO 1

ATIVIDADES DE DESENVOLVIMENTO DO PROJETO REALIZADAS

1.1 IDENTIFICAÇÃO

Título do Projeto: Modelos de *Machine Learning* em análise de sobrevivência na presença de fração de cura

Candidato: João Debastiani Neto

Supervisor: Ricardo Puziol de Oliveira

Período: 27 de fevereiro de 2025 à 27 de fevereiro de 2026

Período de Trabalho: Compromisso de 30 horas semanais.

Total de horas dedicadas ao projeto: 1560 horas.

1.2 ARTIGOS CIENTÍFICOS PRODUZIDOS NA TEMÁTICA DO PROJETO

Artigo 1: Extending the Cox Proportional Hazards Model with a Bayesian Semiparametric Cumulative Hazard Transformation Mixture Cure Model for Long-Term Survival Estimation.

Status: Submetido para *Statistical Methods in Medical Research*.

Artigo 2: Random Survival Forests for Survival Prediction in Heart Failure: External Validation and Predictor Importance.

Status: Submetido para *Statistical Methods for Medical Research*.

1.3 ARTIGOS CIENTÍFICOS EM PRODUÇÃO NA TEMÁTICA DO PROJETO

Artigo 1: A Penalized Bayesian Semiparametric Cumulative Hazard Transformation Model for Long-Term Survival.

Observação: O artigo consolida os resultados finais obtidos no âmbito do projeto, os quais são apresentados e descritos de forma detalhada neste relatório, assegurando coerência entre a metodologia adotada, as evidências empíricas e as conclusões alcançadas. Ressalta-se, contudo, que o manuscrito ainda se encontra em fase de finalização, passando por ajustes estruturais e refinamentos textuais antes de sua submissão.

1.4 Artigos Científicos Produzidos em Outra Temática (Com Interseção com Temática do Projeto)

Artigo 1: Nonlinear Semiparametric Modeling of Lifetime Data Using Polynomial Approximations for Hazard Functions.

Status: O artigo consolida os resultados finais obtidos no âmbito do projeto. Ressalta-se, contudo, que o manuscrito ainda se encontra em fase de finalização, passando por ajustes estruturais e refinamentos textuais antes de sua submissão.

Artigo 2: A Hierarchical Bayesian Lagged-Effects Regression Model for Analyzing Case-Fatality Rates (CFR).
Status: Submetido para *Biostatistics & Epidemiology*.

Artigo 3: Modeling Dependent (Informative) Censoring in Survival Data: A Bayesian Comparison Through Frailty and Marshall-Olkin Bivariate Models.
Status: Submetido para *Journal of Applied Statistics.*

Artigo 4: A Decision Tree–Based Framework for the Classification of Peckoltia Species Using Morphometric Measurements.
Status: Submetido para *Journal of Fish Biology.*

Artigo 5: Modeling Incidence, Mortality Rates, and Patient Survival in Yellow Fever Cases in Brazil.
Status: Submetido para *Revista Colombiana de Estadística.*

Artigo 6: Analysis of the Ammonia Nitrogen Dynamics in Washington State Rivers: An Approach using Tobit Model.
Status: Submetido para *Annals of Data Science.*

1.5 Apresentação de Trabalhos em Eventos Científicos

Evento 1: *Semiparametric Transformation Models with Cure Fraction: Extensions of the Cox Model under a Bayesian Approach*, apresentado no XIV ERMAC- Encontro Regional de Matemática Aplicada e Computacional, UEM, Maringá, PR.

Evento 2: *Extended Semiparametric Cox Model: A Cure Fraction Approach for Long-Term Survival*, apresentado na 69ª RBras e 21° SEAGRO, Vitória, ES.

Evento 3: *Modelagem Semiparamétrica de Dados de Confiabilidade Usando Aproximações Polinomiais de Funções de Risco*, apresentado na 69ª RBras e 21° SEAGRO, Vitória, ES.

Evento 4: *A New Semiparametric Regression Framework with Lagged Effects for Analyzing Epidemiology Non-Linear Data*, apresentado na XIX Escola de Modelos de Regressão (XIX EMR), João Pessoa, PB.

1.6 Participação em Eventos como Avaliador de Trabalhos

Evento 1: *31ª Edição do Prêmio Rocha Lima (PRL)*, promovido pelo Departamento Científico da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, USP, Ribeirão Preto, SP.

Evento 2: *33º SIICUSP - Simpósio Internacional de Iniciação Científica e Tecnológica da Universidade de São Paulo*, promovido pela Universidade de São Paulo, USP.

Evento 3: *XXXVII Congresso de Iniciação Científica da Unesp - FCT/Presidente Prudente*, promovido pela Universidade Estadual Paulista (Unesp) - Faculdade de Ciências e Tecnologia, Presidente Prudente.

1.7 Participação em Grupos de Pesquisa

Grupo 1: Vice-Líder do Grupo de Pesquisa em Estatística Computacional e Aprendizado de Máquina - link para acesso ao grupo: dgp.cnpq.br/dgp/espelhogrupo/2635888148428276.

Grupo 2: Participante do Grupo de Pesquisa em Estatística e Inferência Bayesiana - link para acesso ao grupo: dgp.cnpq.br/dgp/espelhogrupo/9579426422776268.

1.8 Aulas Ministradas

Fundamentos da Matemática: Ministrou 20 aulas da disciplina de Fundamentos da Matemática para o curso de Bacharelado em Estatística da Faculdade de Ciências e Tecnologia da UNESP (FCT/UNESP) – Campus Presidente Prudente.

CAPÍTULO 2

DESCRIÇÃO E DESENVOLVIMENTO DO PROJETO

2.1 INTRODUÇÃO

Em estudos médicos, especialmente em oncologia, é frequente a presença de uma proporção de indivíduos que não experimentam o evento de interesse ao longo do acompanhamento, sendo interpretados como curados ou sobreviventes de longo prazo. Modelos clássicos de sobrevivência, como o modelo de riscos proporcionais de Cox (Cox, 1972), não fornecem estimativa direta da fração de cura, o que motivou o desenvolvimento de modelos específicos. A literatura distingue duas principais abordagens: modelos de cura com mistura, amplamente utilizados na análise de dados com heterogeneidade populacional (De Angelis et al., 1999; Tsodikov et al., 2003; Lambert et al., 2006), e modelos sem mistura (Achcar et al., 2012; Vahidpour, 2016). Trabalhos como Farewell (1982) introduziram a modelagem da fração de cura em dados oncológicos, sendo posteriormente estendidos para incluir covariáveis clínicas e demográficas (De Angelis et al., 1999; Price and Manatunga, 2001). Estudos subsequentes incorporaram estruturas paramétricas e semiparamétricas para tratar a heterogeneidade e estimar a fração curada (Lambert et al., 2006; Othus et al., 2012), enquanto abordagens Bayesianas passaram a ser consideradas na estimação de modelos com fração de cura (Fernandes, 2014). Extensões para dados bivariados foram propostas por Wienke et al. (2003, 2006), incluindo componentes de fragilidade e dependência entre tempos até o evento. Paralelamente, modelos defectivos (Gompertz, 1825) e propostas semiparamétricas não lineares (Chen et al., 2002; Zeng and Lin, 2007) buscaram alternativas que incorporam a fração de cura mantendo estrutura semelhante à do modelo de Cox, porém com maior flexibilidade na especificação da função de risco basal.

Neste projeto, a modelagem da fração de cura é desenvolvida a partir de modelos semiparamétricos estruturados sob uma perspectiva preditiva, integrando conceitos de aprendizado de máquina, como penalizações para controle de complexidade e regularização funcional, particularmente na modelagem da função de risco basal ou de transformações do risco. A avaliação preditiva é conduzida por meio de estratégia de hold-out estratificado para dados de sobrevivência, preservando a distribuição marginal do evento; 70% das observações em cada estrato (evento e censura) compõem o conjunto de treinamento, e 30% formam o conjunto de teste, sendo o ajuste e a seleção de hiperparâmetros realizados exclusivamente nos dados de treinamento. A inferência é conduzida sob abordagem Bayesiana, com especificação de distribuições a priori compatíveis com as penalizações adotadas e estimação via algoritmos de Cadeias de Markov por Monte Carlo (MCMC), permitindo obter distribuições posteriores dos parâmetros e quantificar incerteza de forma integrada ao processo de regularização.

2.2 Objetivos do Projeto

Este projeto se concentrou no objetivo propor novos modelos semiparamétricos com fração de cura para descrever a proporção de indivíduos considerados sobreviventes de longo prazo em estudos clínicos, adotando uma abordagem preditiva fundamentada em técnicas de aprendizado de máquina e incorporando funções de penalização e estratégias de regularização próprias do *machine learning*, integradas à inferência Bayesiana. A proposta busca reduzir a dependência de pressupostos paramétricos restritivos, permitindo modelagem flexível da função de risco e da fração de cura, com estimação realizada por meio de algoritmos de Cadeias de Markov via Monte Carlo (MCMC), investigando como a integração entre aprendizado de máquina e inferência Bayesiana pode contribuir para a análise de dados clínicos.

2.2.1 Objetivos Específicos

1. Desenvolver modelos semiparamétricos com fração de cura incorporando funções de penalização inspiradas em técnicas de aprendizado de máquina.

2. Estruturar componentes não paramétricos do modelo (como função de risco basal ou transformação do risco) utilizando mecanismos de regularização que permitam controle de suavização e complexidade.

3. Formular a inferência sob perspectiva Bayesiana, especificando distribuições a priori compatíveis com as penalizações adotadas, e implementar algoritmos de estimação baseados em Cadeias de Markov via Monte Carlo (MCMC) para obtenção das distribuições posteriores dos parâmetros.

4. Avaliar propriedades inferenciais e desempenho preditivo dos modelos propostos, e aplicar a metodologia desenvolvida a dados clínicos reais com presença de fração de cura, analisando impacto na estimação da sobrevivência de longo prazo.

2.3 Materiais e Métodos

2.3.1 Estrutura dos Dados de Sobrevivência

Considere uma população de interesse em que cada unidade está associada a um tempo até a ocorrência de um evento de interesse (ou tempo de falha), representado por uma variável aleatória não-negativa T. Em estudos aplicados, esse tempo nem sempre é observado integralmente, pois o acompanhamento pode ser interrompido antes da ocorrência do evento, caracterizando a presença de *censura*. Nesse caso, para cada unidade amostral $i = 1, \ldots, n$, o acompanhamento se estende até um tempo aleatório C_i, associado ao encerramento da observação, de modo que o tempo efetivamente observado é definido por:

$$t_i = \min(T_i, C_i), \qquad \delta_i = \mathbb{I}(T_i \leqslant C_i), \tag{2.1}$$

em que t_i corresponde ao tempo efetivamente observado e δ_i é uma variável indicadora binária que registra se o evento de interesse ocorreu durante o período de acompanhamento. Assim, a informação disponível para cada unidade pode ser resumida pelo par (t_i, δ_i), o qual sintetiza toda a informação observável sobre o tempo até o evento. Além disso, é possível associar a cada unidade um vetor de covariáveis:

$$\mathbf{x}_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_p})^\top \in \mathcal{X} \subseteq \mathbb{R}^p, \tag{2.2}$$

que pode incluir características demográficas, clínicas, operacionais ou comportamentais, mensuradas em um instante inicial ou agregadas ao longo do período de acompanhamento, definindo o espaço de entrada sobre o qual os métodos de aprendizado buscam identificar padrões associados ao comportamento do tempo até evento. Dessa forma, o conjunto de dados observados pode ser representado de forma compacta como:

$$\mathcal{D} = \{(t_i, \delta_i, \mathbf{x}_i) : i = 1, \ldots, n\}. \tag{2.3}$$

Essa estrutura, em particular, distingue a análise de sobrevivência de problemas supervisionados clássicos, pois a variável resposta não é observada diretamente como o tempo T_i, mas apenas por meio do par (t_i, δ_i), determinado pelo mecanismo de censura. Consequentemente, a inferência é realizada com base em informação incompleta, considerando-se que as observações $(t_i, \delta_i, \mathbf{x}_i)$ são independentes e identicamente distribuídas.

2.3.2 Estratégia de Particionamento de Dados

O particionamento de dados em análise de sobrevivência deve preservar simultaneamente a estrutura de censura e a distribuição temporal dos eventos. Diversas estratégias de validação podem ser empregadas para esse fim, tais como *hold-out*, validação cruzada (*k-fold cross-validation*), validação cruzada repetida, *leave-one-out*, métodos baseados em bootstrap (incluindo as variações .632 e .632+), esquemas dependentes do tempo (*time-dependent splitting*) e validação externa em coortes independentes. Cada abordagem apresenta vantagens específicas em termos de viés, variância e eficiência amostral. Todavia, para fins de clareza metodológica e ilustração do procedimento de particionamento sob censura, neste trabalho foi considerado apenas o *método hold-out*, devidamente adaptado para o contexto de dados de sobrevivência, no qual, para cada unidade $i = 1, \ldots, n$, observa-se $(t_i, \delta_i, \mathbf{x}_i)$, em que $t_i = \min(T_i, C_i)$ é o tempo observado, $\delta_i = \mathbb{I}(T_i \leqslant C_i)$ é o indicador de ocorrência do evento e $\mathbf{x}_i \in \mathbb{R}^p$ representa o vetor de covariáveis, sendo a variável resposta T_i parcialmente observada por (t_i, δ_i).

Nesse cenário, a aplicação direta do método *hold-out* clássico, isto é, o sorteio independente e equiprovável das unidades, apresenta uma limitação importante: a proporção de eventos $\delta_i = 1$ no conjunto de teste pode diferir da proporção observada na amostra original, especialmente sob elevada taxa de censura. A consequência imediata é a perda de representatividade estrutural: o conjunto de teste pode conter número insuficiente de eventos para a avaliação do desempenho preditivo com precisão. Para contornar essa dificuldade, adota-se a

amostragem estratificada pelo indicador de censura δ_i. Neste caso, sejam:

$$\mathcal{I}_1 = \{i : \delta_i = 1\} \quad \text{(unidades com evento observado)}, \tag{2.4}$$

$$\mathcal{I}_0 = \{i : \delta_i = 0\} \quad \text{(unidades censuradas)}, \tag{2.5}$$

com cardinalidades $n_1 = |\mathcal{I}_1|$ e $n_0 = |\mathcal{I}_0|$, tais que $n_1 + n_0 = n$. Define-se, para cada estrato, uma variável indicadora de alocação independente:

$$Z_i \mid \delta_i = 1 \sim \text{Bernoulli}(p), \quad i \in \mathcal{I}_1, \tag{2.6}$$

$$Z_i \mid \delta_i = 0 \sim \text{Bernoulli}(p), \quad i \in \mathcal{I}_0, \tag{2.7}$$

com sorteios independentes entre unidades e entre estratos. Os conjuntos de treinamento e teste são então definidos por:

$$\mathcal{D}_{\text{train}} = \{(t_i, \delta_i, \mathbf{x}_i) : Z_i = 1\}, \tag{2.8}$$

$$\mathcal{D}_{\text{test}} = \{(t_i, \delta_i, \mathbf{x}_i) : Z_i = 0\}. \tag{2.9}$$

Por construção, a proporção amostral de eventos em cada subconjunto é preservada em valor esperado. De fato, condicional às cardinalidades n_1 e n_0, o número esperado de eventos no conjunto de treinamento é $p \cdot n_1$, e a proporção esperada de eventos é $(p \cdot n_1)/(p \cdot n_1 + p \cdot n_0) = n_1/n$, que coincide com a proporção global. O mesmo raciocínio aplica-se ao conjunto de teste. A estratificação por δ_i assegura, portanto, que ambos os subconjuntos preservem, em valor esperado, a taxa de incidência do evento observada na amostra original.

2.3.3 Modelos com Fração de Cura

Seja T uma variável aleatória não negativa representando o tempo até a ocorrência do evento de interesse. Na formulação de mistura (Vahidpour, 2016), assume-se que a população é composta por dois subgrupos latentes: indivíduos suscetíveis ao evento e indivíduos curados (ou não suscetíveis). Introduz-se, para tanto, uma variável indicadora latente $Y \in \{0, 1\}$ tal que $Y = 1$ denota indivíduo suscetível e $Y = 0$ indivíduo curado. Para os indivíduos curados, admite-se $\mathbb{P}(T = \infty \mid Y = 0) = 1$, enquanto, para os suscetíveis, T possui distribuição própria com função de sobrevivência condicional $S_u(t \mid \mathbf{x})$, onde $\mathbf{x} \in \mathbb{R}^p$ é um vetor de covariáveis associado ao mecanismo de latência. Seja, agora, $\mathbf{z} \in \mathbb{R}^q$ um vetor de covariáveis associado ao mecanismo de incidência, que governa a probabilidade de suscetibilidade. Define-se $\pi(\mathbf{z}) = \mathbb{P}(Y = 0 \mid \mathbf{z})$ e $1 - \pi(\mathbf{z}) = \mathbb{P}(Y = 1 \mid \mathbf{z})$. Pela lei da probabilidade total, a função de sobrevivência populacional condicional a $(\mathbf{x}, \mathbf{z})$ é dada por:

$$\begin{aligned} S(t \mid \mathbf{x}, \mathbf{z}) &= \mathbb{P}(T > t \mid \mathbf{x}, \mathbf{z}) \\ &= \mathbb{P}(T > t \mid Y = 0, \mathbf{x}, \mathbf{z})\mathbb{P}(Y = 0 \mid \mathbf{z}) + \mathbb{P}(T > t \mid Y = 1, \mathbf{x}, \mathbf{z})\mathbb{P}(Y = 1 \mid \mathbf{z}) \\ &= \pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]\, S_u(t \mid \mathbf{x}). \end{aligned} \tag{2.10}$$

Como $S_u(t \mid \mathbf{x}) \to 0$ quando $t \to \infty$, segue que:

$$\lim_{t\to\infty} S(t \mid \mathbf{x}, \mathbf{z}) = \pi(\mathbf{z}), \tag{2.11}$$

de modo que $\pi(\mathbf{z})$ representa a fração de cura condicional ao perfil $\mathbf{z}$. A função de distribuição populacional é $F(t \mid \mathbf{x}, \mathbf{z}) = 1 - S(t \mid \mathbf{x}, \mathbf{z})$, e a densidade populacional, para $t > 0$, assume a forma:

$$f(t \mid \mathbf{x}, \mathbf{z}) = [1 - \pi(\mathbf{z})]\, f_u(t \mid \mathbf{x}), \tag{2.12}$$

onde $f_u(t \mid \mathbf{x}) = -\partial/\partial t\, S_u(t \mid \mathbf{x})$ é a densidade do tempo até evento entre suscetíveis. A função de risco populacional $h(t \mid \mathbf{x}, \mathbf{z})$ pode ser escrita como:

$$h(t \mid \mathbf{x}, \mathbf{z}) = \frac{[1 - \pi(\mathbf{z})]\, f_u(t \mid \mathbf{x})}{\pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]\, S_u(t \mid \mathbf{x})}, \tag{2.13}$$

ou, de forma equivalente, em termos do risco entre suscetíveis $h_u(t \mid \mathbf{x})$:

$$h(t \mid \mathbf{x}, \mathbf{z}) = \frac{[1 - \pi(\mathbf{z})]\, h_u(t \mid \mathbf{x}) S_u(t \mid \mathbf{x})}{\pi(\mathbf{z}) + [1 - \pi(\mathbf{z})]\, S_u(t \mid \mathbf{x})}. \tag{2.14}$$

Para completar a especificação do modelo, a componente de incidência é usualmente modelada por regressão logística:

$$\pi(\mathbf{z}) = \frac{\exp(\mathbf{z}^\top \boldsymbol{\gamma})}{1 + \exp(\mathbf{z}^\top \boldsymbol{\gamma})}, \tag{2.15}$$

com $\boldsymbol{\gamma} \in \mathbb{R}^q$, enquanto a componente de latência pode assumir especificações paramétricas ou semiparamétricas. Agora, considere uma amostra independente $\mathcal{D} = \{(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^n$, em que $t_i = \min(T_i, C_i)$ representa o tempo observado e $\delta_i = \mathbb{I}(T_i \leqslant C_i)$ indica a ocorrência do evento. Sob censura à direita, a verossimilhança assume a forma:

$$L(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \{[1 - \pi(\mathbf{z}_i; \boldsymbol{\gamma})]\, f_u(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})\}^{\delta_i} \tag{2.16}$$

$$\times \{\pi(\mathbf{z}_i; \boldsymbol{\gamma}) + [1 - \pi(\mathbf{z}_i; \boldsymbol{\gamma})]\, S_u(t_i \mid \mathbf{x}_i; \boldsymbol{\theta})\}^{1-\delta_i}. \tag{2.17}$$

Os estimadores de $(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\theta}})$ são obtidos por maximização de $L(\boldsymbol{\gamma}, \boldsymbol{\theta})$, ou, alternativamente, por inferência Bayesiana, na qual $\boldsymbol{\gamma}$ e $\boldsymbol{\theta}$ são tratados como parâmetros aleatórios e estimados a partir da distribuição a posteriori. Nesse caso, podem-se incorporar estruturas de regularização e penalizações de aprendizado de máquina por meio de distribuições a priori equivalentes a penalidades, bem como empregar procedimentos de amostragem MCMC ou aproximações variacionais em um enfoque de aprendizado de máquina Bayesiano.

2.3.4 Modelos de Transformação de Risco

Os modelos de transformação de risco constituem uma classe de modelos semiparamétricos na qual a relação entre o desfecho de sobrevivência e as covariáveis é decomposta em um componente paramétrico e um componente não paramétrico. Diferentemente do modelo de riscos proporcionais de Cox, essa classe é caracterizada pela modelagem da função de risco acumulada como uma transformação de um risco *baseline* não especificado, combinada com os efeitos das covariáveis, permitindo maior flexibilidade além da suposição de riscos proporcionais (Achcar and Barili, 2024). Para a construção dessa classe de modelos, considere T o tempo até o evento e $\mathbf{Z}$ um vetor de covariáveis; então a função de risco cumulativa condicional é expressa como:

$$\Lambda(t;\mathbf{z}) = G\left(\int_0^t e^{\boldsymbol{\beta}^\top \mathbf{z}} h_0(u)\, du\right), \tag{2.18}$$

onde $G(\cdot)$ é uma função de transformação monotônica conhecida que incorpora os efeitos das covariáveis, $\Lambda_0(t) = \int_0^t h_0(u)\, du$ representa a função de risco acumulada *baseline* (não-especificada), e $\boldsymbol{\beta}$ é um vetor de parâmetros de regressão. Reorganizando os termos, tem-se que:

$$\Lambda(t;\mathbf{z}) = G\left(e^{\boldsymbol{\beta}^\top \mathbf{z}} \Lambda_0(t)\right). \tag{2.19}$$

Essa formulação, em particular, permite modelar variações específicas de cada indivíduo no risco de sobrevivência, superando as limitações do modelo de riscos proporcionais de Cox, que assume uma estrutura multiplicativa específica para os efeitos das covariáveis. Além disso, dependendo da escolha da função de transformação $G(\cdot)$, diferentes modelos podem ser construídos, cada um capturando distintos aspectos da dinâmica subjacente de sobrevivência. Alguns exemplos incluem:

(a) Modelos de Riscos Proporcionais de Cox: Se $G(x) = x$, então a função de risco acumulada expressa em (2.19) se transforma em:

$$\Lambda(t;\mathbf{z}) = e^{\boldsymbol{\beta}^\top \mathbf{z}} \Lambda_0(t), \tag{2.20}$$

em que $\Lambda_0(t) = \int_0^t h_0(u)du$é a função de risco cumulativo de referência, isto é, a *baseline* (com $h_0(u)$ desconhecida). Essa formulação, em particular, corresponde ao *modelo de riscos proporcionais de Cox*, com a função de risco definida por:

$$h(t;\mathbf{z}) = e^{\boldsymbol{\beta}^\top \mathbf{z}} h_0(t). \tag{2.21}$$

Neste caso, para dois indivíduos i e j com covariáveis $\mathbf{z}_i$ e $\mathbf{z}_j$, a *razão de riscos* é dada por:

$$\frac{h(t;\mathbf{z}_i)}{h(t;\mathbf{z}_j)} = \frac{e^{\boldsymbol{\beta}^\top \mathbf{z}_i} h_0(t)}{e^{\boldsymbol{\beta}^\top \mathbf{z}_j} h_0(t)} = e^{\boldsymbol{\beta}^\top (\mathbf{z}_i - \mathbf{z}_j)}. \tag{2.22}$$

(b) Modelo de Razão de Chances Proporcional: Se $G(x) = \ln(1+x)$, então a função de risco acumulada expressa em (2.19) se transforma em:

$$\Lambda(t;\mathbf{z}) = \ln\left(1 + e^{\boldsymbol{\beta}^\top \mathbf{z}}\Lambda_0(t)\right), \tag{2.23}$$

e a função de sobrevivência $S(t;\mathbf{z})$ em:

$$S(t;\mathbf{z}) = \exp\left(-\Lambda(t;\mathbf{z})\right) = \frac{1}{1 + e^{\boldsymbol{\beta}^\top \mathbf{z}}\Lambda_0(t)}. \tag{2.24}$$

Neste caso, a *razão de chances* (ou *odds ratio*) entre dois indivíduos i and j com covariáveis $\mathbf{z}_i$ e $\mathbf{z}_j$ é dada por:

$$\frac{\mathrm{OR}_i}{\mathrm{OR}_j} = \frac{e^{\boldsymbol{\beta}^\top \mathbf{z}_j}\Lambda_0(t)}{e^{\boldsymbol{\beta}^\top \mathbf{z}_i}\Lambda_0(t)} = e^{\boldsymbol{\beta}^\top(\mathbf{z}_j - \mathbf{z}_i)}. \tag{2.25}$$

(c) Modelo de Transformação Exponencial: Se $G(x) = e^x$, então a função de risco acumulada expressa em (2.19) se transforma em:

$$\Lambda(t;\mathbf{z}) = e^{e^{\boldsymbol{\beta}^\top \mathbf{z}}\Lambda_0(t)}. \tag{2.26}$$

Essa transformação leva a um *modelo de transformação exponencial*, no qual o risco acumulado aumenta exponencialmente em função do tempo e das covariáveis. Para este modelo, a função de risco resultante é dada por:

$$h(t;\mathbf{z}) = e^{e^{\boldsymbol{\beta}^\top \mathbf{z}}\Lambda_0(t)} \cdot e^{\boldsymbol{\beta}^\top \mathbf{z}} h_0(t), \tag{2.27}$$

em que $h_0(t)$ é o risco de referência *baseline*, e as covariáveis são transformadas exponencialmente.

(d) Modelo de Transformação Adaptativa: transformação adaptativa aos dados construída por meio de um gerador spline monotônico. Especificamente, definimos a função de transformação como

$$G(x;\boldsymbol{\theta}) = \int_0^x \exp\left\{\sum_{k=1}^{K} \theta_k B_k\big(\log(1+u)\big)\right\} du, \qquad x \geqslant 0, \tag{2.28}$$

em que $\{B_k(\cdot)\}_{k=1}^K$ denota um conjunto de funções base spline e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ é um vetor de parâmetros de transformação. Por construção, a derivada de G é dada por

$$G'(x;\boldsymbol{\theta}) = \exp\left\{\sum_{k=1}^{K} \theta_k B_k\big(\log(1+x)\big)\right\} > 0, \qquad x \geqslant 0, \tag{2.29}$$

o que garante a monotonicidade estrita de $G(\cdot;\boldsymbol{\theta})$. Consequentemente, o risco cumulativo $\Lambda(t \mid \mathbf{Z}) = G\left(e^{\boldsymbol{\beta}^\top \mathbf{Z}}\Lambda_0(t);\boldsymbol{\theta}\right)$ permanece não decrescente em t, assegurando a validade do modelo de sobrevivência.

2.4 CRONOGRAMA DESENVOLVIDO

O plano de execução do projeto foi estruturado para um período de 12 meses. As etapas da pesquisa foram concluídas adequadamente, e o trabalho manteve o alinhamento com os objetivos inicialmente estabelecidos. A Tabela 2.1 apresenta o cronograma considerado com as etapas realizadas, em que, com dedicação de 30 horas semanais, foram totalizadas 1560 horas de dedicação ao projeto.

Tabela 2.1: Cronograma do projeto de pesquisa proposto (12 meses).

Atividades	**Meses**					
	01–02	03–04	05–06	07–08	09–10	11–12
Revisão de literatura	X	X	X	X	X	X
Coleta de dados	X	X	X	X		
Desenvolvimento dos modelos		X	X	X	X	
Análise dos resultados		X	X	X	X	
Redação e discussão final			X	X	X	X
Iteração com corpo discente e docente		X	X	X		
Disseminação de resultados		X	X	X	X	X

Detalhamento

Revisão de literatura

Nesta etapa, foram conduzidas buscas sistemáticas e análise crítica de artigos científicos, livros e publicações especializadas em análise de sobrevivência, com ênfase em modelos com fração de cura e extensões semiparamétricas. O objetivo foi consolidar o embasamento teórico do projeto, identificando avanços recentes em modelagem de risco, formulações de mistura, estruturas de transformação e estratégias de regularização. Também foram revisados trabalhos que incorporam técnicas de *machine learning* ao contexto de dados de sobrevivência, especialmente aqueles que tratam da integração entre métodos estatísticos tradicionais e abordagens baseadas em aprendizado adaptativo. A literatura específica sobre aplicações clínicas de modelos com fração de cura foi examinada, permitindo contextualizar as contribuições metodológicas no âmbito de estudos médicos e de saúde pública.

Coleta de dados

Nesta etapa, realizou-se a obtenção e preparação de bases de dados de sobrevivência provenientes de repositórios públicos e estudos clínicos que disponibilizaram informações adequadas ao contexto de fração de cura. A seleção considerou a qualidade dos registros, a presença de variáveis clínicas relevantes e a existência de censura à direita. Os dados passaram por procedimentos de limpeza, padronização e organização estrutural, incluindo tratamento de valores ausentes, codificação de covariáveis e identificação adequada dos tempos de evento e indicadores de censura. Adicionalmente, foi implementada uma estratégia de validação do tipo *hold-out*

adaptada para dados de sobrevivência, na qual a base foi particionada em subconjuntos de treinamento e teste, preservando a proporção de eventos e censuras. Essa separação possibilitou avaliar o desempenho preditivo dos modelos em dados não utilizados no ajuste, por meio de métricas apropriadas para sobrevivência. Por fim, os dados foram estruturados de modo a permitir a aplicação tanto de modelos semiparamétricos com componentes penalizados quanto de abordagens complementares baseadas em *machine learning*, assegurando consistência metodológica, comparabilidade entre métodos e reprodutibilidade das análises realizadas.

Desenvolvimento dos modelos

O foco desta etapa consistiu na implementação de modelos semiparamétricos não lineares ajustados ao contexto de fração de cura. Foram exploradas abordagens que permitissem maior flexibilidade na modelagem da função de risco, como modelos de transformação semiparamétricos e técnicas baseadas em aproximações polinomiais e representações spline para a função de risco e para a transformação do risco cumulativo, visando descrever de forma mais precisa a heterogeneidade presente nos dados clínicos e possíveis desvios da suposição de riscos proporcionais. Adicionalmente, esta etapa contemplou a integração de técnicas de *machine learning* para dados de sobrevivência ao contexto semiparamétrico. Foram investigadas estratégias híbridas em que componentes estruturais do modelo — como a função de transformação ou a função de risco basal — são aproximadas por bases flexíveis (splines monotônicas, expansões em bases polinomiais ou representações baseadas em árvores), combinadas com regularização Bayesiana. Penalizações de priori, tais como estruturas do tipo passeio aleatório de segunda ordem sobre incrementos logarítmicos do risco basal ou sobre coeficientes de spline,

$$\theta_k - 2\theta_{k-1} + \theta_{k-2} \sim \mathcal{N}(0, \tau^{-1}), \tag{2.30}$$

foram utilizadas para controlar suavidade e evitar sobreajuste, mantendo interpretabilidade e estabilidade numérica. Também foram avaliadas conexões com métodos baseados em árvores e florestas para sobrevivência, explorando sua incorporação como mecanismos de aprendizado da estrutura funcional dentro de modelos com fração de cura, seja como ferramentas exploratórias para detecção de não linearidades e interações, seja como componentes integráveis ao modelo Bayesiano hierárquico. Essa integração possibilitou combinar a interpretabilidade dos modelos semiparamétricos com a capacidade adaptativa de métodos de aprendizado de máquina, ampliando o poder preditivo e a robustez inferencial em cenários clínicos complexos.

Análise dos resultados

A partir da implementação, os resultados obtidos foram analisados com o objetivo de interpretar a eficácia dos modelos desenvolvidos na previsão da fração de cura e dos tempos de sobrevivência, estabelecendo comparações sistemáticas com abordagens tradicionais, como modelos de riscos proporcionais e formulações paramétricas clássicas. A avaliação contemplou tanto aspectos inferenciais — como precisão dos estimadores,

estabilidade numérica e coerência dos intervalos de credibilidade, quanto aspectos preditivos, incluindo medidas de desempenho fora da amostra, capacidade de discriminação e calibração. No contexto dos modelos com fração de cura, foi examinada a habilidade das abordagens propostas em separar adequadamente os mecanismos de incidência (probabilidade de cura) e de latência (tempo até o evento entre indivíduos suscetíveis), verificando se a modelagem mais flexível da função de risco resultava em melhor ajuste aos dados observados. Também foram analisados padrões de não linearidade e possíveis interações entre covariáveis, especialmente quando incorporadas por meio de estruturas mais adaptativas inspiradas em técnicas de *machine learning*. A comparação com métodos tradicionais permitiu identificar ganhos em termos de flexibilidade estrutural e capacidade de capturar heterogeneidade não observada. Em particular, modelos semiparamétricos penalizados e suas extensões com componentes de aprendizado de máquina mostraram-se mais sensíveis à detecção de padrões complexos nos dados clínicos, sem sacrificar interpretabilidade. Foram ainda conduzidas análises de sensibilidade relacionadas às especificações de priori e aos mecanismos de regularização, avaliando o impacto dessas escolhas na estabilidade das estimativas e no desempenho preditivo.

Redação e discussão final

Nas etapas finais, o pesquisador concentrou-se na sistematização dos resultados e na redação dos produtos científicos decorrentes do projeto. Foram elaborados manuscritos descrevendo de forma estruturada os fundamentos metodológicos, as estratégias de modelagem adotadas, incluindo extensões semiparamétricas com fração de cura e componentes de *machine learning*, bem como as evidências empíricas obtidas. A discussão dos resultados enfatizou as contribuições metodológicas dos modelos desenvolvidos, especialmente quanto à maior flexibilidade na modelagem do risco, à separação entre incidência e latência e ao impacto das penalizações de priori na estabilidade inferencial. Também foram destacadas as implicações práticas para a análise de dados clínicos e para a tomada de decisão em contextos de saúde pública. Os artigos produzidos foram ou serão submetidos a periódicos científicos da área, consolidando a contribuição do estudo tanto sob a perspectiva metodológica quanto aplicada.

Interação com corpo docente e discente

Foram realizados encontros regulares para discussão aprofundada de métodos estatísticos aplicados a dados clínicos, com ênfase em modelos de sobrevivência, fração de cura e extensões que incorporam regularização e componentes de *machine learning*. Essas reuniões possibilitaram a análise crítica de artigos recentes, a discussão de estratégias metodológicas e o aprimoramento das abordagens implementadas no projeto. Houve também participação em disciplinas de graduação, com contribuição na elaboração de materiais didáticos, desenvolvimento de exemplos aplicados e apoio à orientação de trabalhos. As atividades favoreceram a integração entre pesquisa e ensino, fortalecendo o diálogo acadêmico e promovendo um ambiente de formação que estimulou a troca de conhecimento, o pensamento crítico e a consolidação de competências metodológicas entre docentes e discentes.

CAPÍTULO 3

RESULTADOS E DISCUSSÃO

3.1 ARTIGO 1

Título: Extending the Cox Proportional Hazards Model with a Bayesian Semiparametric Cumulative Hazard Transformation Mixture Cure Model for Long-Term Survival Estimation.

Status: Submetido para *Statistical Methods in Medical Research*.

3.1.1 DADOS DO ESTUDO

Para este estudo, consideramos um conjunto de dados de uma coorte retrospectiva compreendendo 272 casos de carcinoma de mama invasivo primário, coletados como dados secundários a partir de van de Vijver et al. (2002), provenientes do Instituto do Câncer da Holanda (NKI). A coorte inclui pacientes do sexo feminino diagnosticadas entre 1984 e 1995, com os seguintes critérios de inclusão: diâmetro do tumor inferior a 5 cm no exame anatomopatológico (pT1 ou pT2), linfonodos axilares apicais negativos, idade no diagnóstico de 52 anos ou menos, e sem histórico prévio de câncer, exceto câncer de pele não melanoma. Todas as pacientes foram submetidas a mastectomia radical modificada ou cirurgia conservadora da mama, com dissecção de linfonodos axilares, seguida de radioterapia adjuvante quando clinicamente indicada. Apesar de o conjunto de dados ter origem no período de 1984 a 1995, ele permanece altamente significativo devido à sua coleta abrangente e bem documentada de características clínicas e tumorais.

3.1.2 MODELOS DE TRANSFORMAÇÃO DE RISCO

Os modelos de transformação semiparamétricos formam uma classe de modelos estatísticos em que a relação entre o desfecho de sobrevida e as covariáveis é expressa por meio de um componente paramétrico e um componente não paramétrico. Na análise de sobrevida, essa classe é caracterizada pela modelagem da função de risco cumulativo como uma transformação de um risco basal não especificado combinado com os efeitos das covariáveis, permitindo maior flexibilidade além da suposição de riscos proporcionais. Formalmente, seja T o tempo de falha e $\mathbf{Z}$ um vetor de covariáveis; então a função de risco cumulativo condicional é expressa como:

$$\Lambda(t;\mathbf{z}) = G\left(\int_0^t e^{\boldsymbol{\beta}^\top \mathbf{z}} h_0(u)\,du\right), \tag{3.1}$$

onde $G(\cdot)$ é uma função de transformação monótona conhecida que incorpora os efeitos das covariáveis, $\Lambda_0(t) = \int_0^t h_0(u)\,du$ representa a função de risco cumulativo basal não especificada, e $\boldsymbol{\beta}$ é um vetor de parâmetros de regressão. Assim, a função de risco cumulativo condicional pode ser simplificada para $\Lambda(t;\mathbf{z}) = G\left(e^{\boldsymbol{\beta}^\top \mathbf{z}}\Lambda_0(t)\right)$.

Agora, para a construção da função de verossimilhança (também chamada de *função custo* no contexto de *machine learning*), considerea t_i o tempo observado até o evento ou censura para o i-ésimo indivíduo, e seja $\mathbf{Z}_i$ o vetor de covariáveis de dimensão d associado. Defina δ_i como o indicador de censura, onde $\delta_i = 1$ se o evento for observado, e $\delta_i = 0$ se a observação for censurada à direita. Assim, no modelo de transformação de risco cumulativo, a função de sobrevida condicional é dada por:

$$S(t;\mathbf{Z}) = \exp\left\{-G\left(e^{\boldsymbol{\beta}^\top \mathbf{Z}}\Lambda_0(t)\right)\right\}, \tag{3.2}$$

onde $G(\cdot)$ é a função de transformação que caracteriza a relação entre o risco cumulativo e os efeitos das covariáveis, e $\Lambda_0(t)$ é a função de risco cumulativo basal. Diferenciando a função de sobrevida em relação ao tempo, obtém-se a função de densidade condicional:

$$f(t;\mathbf{Z}) = h(t;\mathbf{Z})S(t;\mathbf{Z}), \tag{3.3}$$

onde a função de risco condicional é dada, de acordo com a regra da cadeia, por:

$$h(t;\mathbf{Z}) = G'\left(e^{\boldsymbol{\beta}^\top \mathbf{Z}}\,\Lambda_0(t)\right)\cdot e^{\boldsymbol{\beta}^\top \mathbf{Z}}\,h_0(t), \tag{3.4}$$

onde $G'(\cdot)$ denota a derivada de $G(\cdot)$ em relação ao seu argumento, e $h_0(t) = \Lambda_0^{'}(t)$ é a função de risco basal. Portanto, a função de verossimilhança para uma amostra de tamanho n é dada por:

$$L(\boldsymbol{\beta}, \Lambda_0) = \prod_{i=1}^{n}\left[G'\left(e^{\boldsymbol{\beta}^\top \mathbf{Z}_i}\,\Lambda_0(t_i)\right)\cdot e^{\boldsymbol{\beta}^\top \mathbf{Z}_i}\,h_0(t_i)\right]^{\delta_i}\exp\left\{-G\left(e^{\boldsymbol{\beta}^\top \mathbf{Z}_i}\,\Lambda_0(t_i)\right)\right\}. \tag{3.5}$$

3.1.3 Incorporando a Taxa de Cura

O *modelo de mistura com taxa de cura*, também conhecido como *modelo padrão de taxa de cura*, tem visto ampla aplicação na análise de sobrevida onde taxas de cura são consideradas (De Angelis et al., 1999; Tsodikov et al., 2003; Lambert et al., 2007), conforme enfatizado por Vahidpour (2016). Este modelo assume que a população é composta por dois subgrupos: (1) indivíduos que são suscetíveis ao evento e (2) indivíduos que estão "curados"ou são "não suscetíveis"e nunca experimentarão o evento. Assim, seguindo Maller and Zhou (1996), a função de sobrevida para um indivíduo baseada em um vetor de covariáveis $\mathbf{Z}$ é definida como:

$$S(t;\mathbf{Z}) = \rho + (1-\rho)S_0(t;\mathbf{Z}), \tag{3.6}$$

onde $0 < \rho < 1$ é o parâmetro de *taxa de cura*, representando a proporção de indivíduos que estão curados ou imunes ao evento; e $S_0(t;\mathbf{Z})$ é a função de sobrevida para o *grupo suscetível*, representando os indivíduos que estão em risco de experimentar o evento. Em nossa estrutura proposta, assumimos que a função de sobrevida

para o grupo suscetível segue um modelo de transformação de risco cumulativo, a saber:

$$S_0(t;\mathbf{Z}) = \exp\left\{-G\left(e^{\boldsymbol{\beta}^\top \mathbf{Z}}\Lambda_0(t)\right)\right\}, \tag{3.7}$$

onde $G(\cdot)$ é uma função de transformação estritamente crescente que define como as covariáveis influenciam a distribuição de sobrevida, e $\Lambda_0(t)$ denota a função de risco cumulativo basal. Além disso, a diferenciação do logaritmo da função de sobrevida global $S(t;\mathbf{Z})$ fornece a função de risco:

$$h(t;\mathbf{Z}) = -\frac{d}{dt}\log S(t;\mathbf{Z}). \tag{3.8}$$

Aplicando a regra da cadeia, a forma explícita da função de risco torna-se:

$$h(t;\mathbf{Z}) = \frac{(1-\rho)e^{\boldsymbol{\beta}^\top \mathbf{Z}}h_0(t)S_0(t;\mathbf{Z})}{[\rho + (1-\rho)S_0(t;\mathbf{Z})]\left[\frac{d}{dt}G\left(e^{\boldsymbol{\beta}^\top \mathbf{Z}}\Lambda_0(t)\right)\right]^{-1}}. \tag{3.9}$$

Assim, para uma amostra de n indivíduos, a função de verossimilhança é dada por:

$$L(\boldsymbol{\beta},\, \rho,\, \Lambda_0) = \prod_{i=1}^{n}\left[(1-\rho)\, h_0(t_i;\mathbf{Z}_i)S_0(t_i;\mathbf{Z}_i)\right]^{\delta_i}\left[\rho + (1-\rho)S_0(t_i;\mathbf{Z}_i)\right]^{1-\delta_i}. \tag{3.10}$$

3.1.4 Resultados e Discussão

Para a análise inicial dos dados, considerou-se o modelo PH de Cox ao conjunto de dados de câncer de mama do NKI para estimar os efeitos dos seguintes fatores de risco nos desfechos de sobrevida:

- Z_1 (idade): Idade no diagnóstico (anos);
- Z_2 (quimio): Recebeu quimioterapia (Sim = 1, Não = 0);
- Z_3 (hormonal): Recebeu terapia hormonal (Sim = 1, Não = 0);
- Z_4 (amputação): Tipo de cirurgia (Amputação = 1, Conservadora = 0);
- Z_5 (tipo hist): Tipo histológico do tumor (Ductal = 1, Outros = 0);
- Z_6 (diam): Diâmetro do tumor (cm);
- Z_7 (linfonodos pos): Número de linfonodos positivos;
- Z_8 (grau): Grau do tumor (1 = Alto Grau, Baixo/Intermediário Grau = 0).

Os resultados do modelo de riscos proporcionais (PH) de Cox estão resumidos na Tabela 3.1, que apresenta coeficientes estimados, razões de riscos (HR), erros padrão, valores z e valores p para cada covariável, com o nível zero como a *linha de base* (*baseline*) para variáveis binárias. Entre as covariáveis, apenas o

grau do tumor é um preditor estatisticamente significativo de sobrevida ($p < 0.0001$). Sua razão de riscos ($\exp(\hat{\beta}) = 2.9909$) indica um aumento de 199.09% no risco de experimentar o evento, destacando o forte impacto da agressividade do tumor no prognóstico. Outras covariáveis, incluindo idade, quimioterapia, terapia hormonal, amputação, tipo histológico, diâmetro do tumor e linfonodos positivos, não foram estatisticamente significativas.

Tabela 3.1: Estimativas para o modelo PH de Cox para os dados de câncer de mama do NKI.

Parâmetro	**Estimativa**	**HR**	**Erro Padrão**	**Valor z**	**Valor p**
β_1 (idade)	-0.0340	0.9666	0.0203	-1.6690	0.0950
β_2 (quimio)	-0.4106	0.6663	0.2979	-1.3780	0.1681
β_3 (hormonal)	-0.2635	0.7683	0.4420	-0.5960	0.5510
β_4 (amputação)	-0.0106	0.9895	0.2504	0.0420	0.9663
β_5 (tipo hist)	0.6179	1.8550	0.4956	1.2470	0.2125
β_6 (diam)	0.0198	1.0200	0.0127	1.5570	0.1196
β_7 (linfonodos pos)	0.0917	1.0960	0.0541	1.6930	0.0905
β_8 (grau)	1.0956	2.9909	0.2650	4.1340	< 0.0001

A suposição de riscos proporcionais foi avaliada usando resíduos de Schoenfeld (Schoenfeld, 1982; Grambsch and Therneau, 1994), mostrados nas Figuras 3.1 e 3.2. A suposição parece ser violada para o grau do tumor (valor p < 0,05), enquanto é satisfeita para as covariáveis restantes. Dado que o grau do tumor é o preditor mais significativo neste estudo, essa violação sugere que o modelo PH de Cox pode não satisfazer totalmente suas suposições, destacando a necessidade de abordagens de modelagem alternativas que possam contabilizar efeitos não proporcionais e sobreviventes de longo prazo.

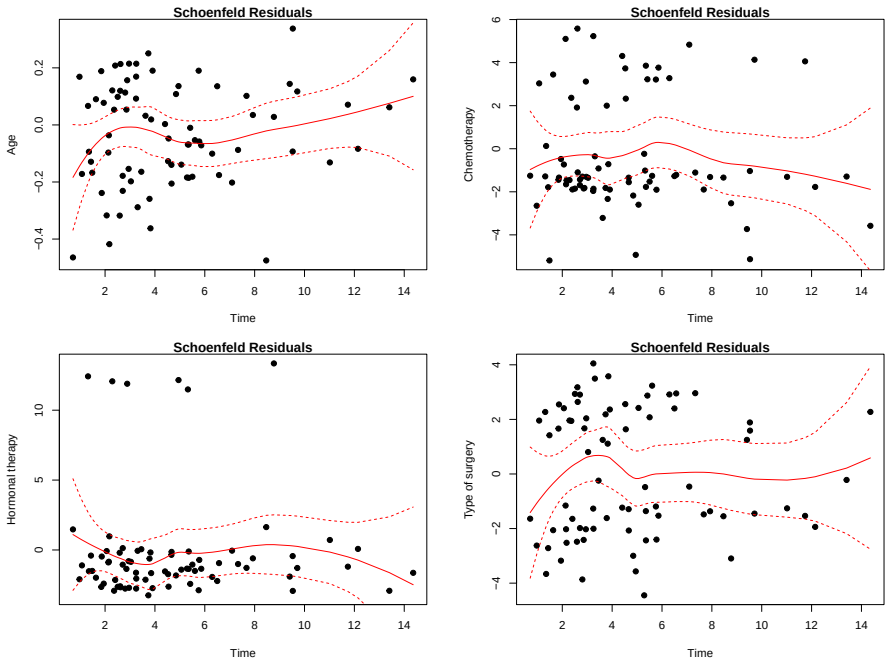


Figura 3.1: Resíduos de Schoenfeld para os fatores de risco empregados na análise - Idade, Quimioterapia recebida, Terapia hormonal recebida e Tipo de cirurgia.

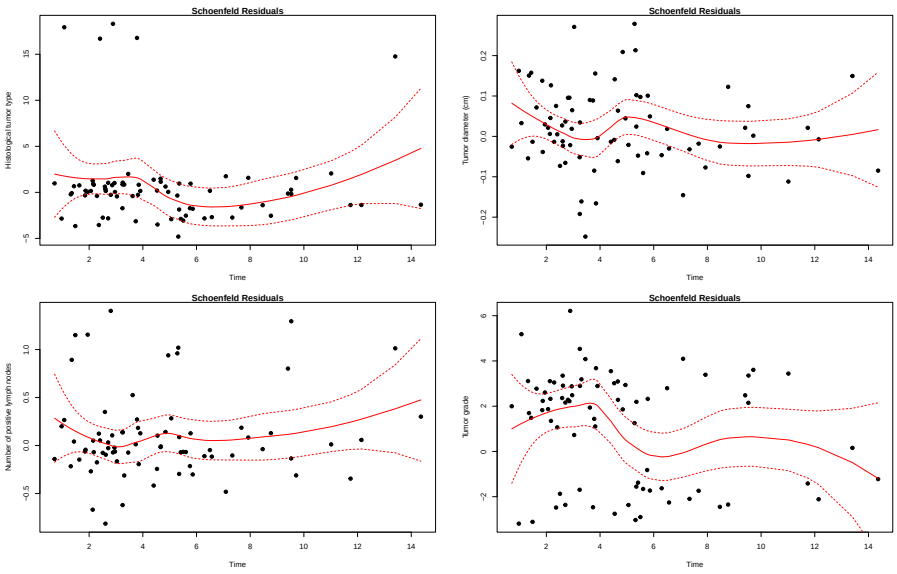


Figura 3.2: Resíduos de Schoenfeld para os fatores de risco empregados na análise - Tipo histológico do tumor, Diâmetro do tumor, Número de linfonodos positivos e Grau do tumor.

Para abordar essas limitações, foi desenvolvida uma extensão do modelo de riscos proporcionais de Cox, levando a um *modelo de Cox com taxa de cura* semiparamétrico formulado sob a estrutura de transformação de risco cumulativo com a função de transformação $G(x) = x$. Dentro deste modelo, incorporamos informações *a priori* em uma estrutura totalmente Bayesiana, atribuindo distribuições *a priori* Gama aos coeficientes de regressão reparametrizados. Especificamente, definimos os parâmetros transformados como $\theta_j = \exp(\beta_j), \quad j = 1, 2, \ldots, k$, e especificamos *a prioris* Gama independentes da forma

$$\theta_j \sim \text{Gamma}(a_j, 1), \tag{3.11}$$

onde os parâmetros de forma a_j são derivados de estimativas obtidas sob o modelo padrão de riscos proporcionais de Cox, fornecendo assim *a prioris* informativas para os efeitos de regressão. Para o parâmetro de taxa de cura ρ, assume-se uma *a priori* Beta(1,1), representando uma distribuição uniforme não informativa sobre $(0, 1)$. Além disso, aos incrementos da função de risco cumulativo basal são atribuídas *a prioris* Gama aproximadamente não informativas da forma $\text{Gamma}(0.001, 0.001)$, refletindo conhecimento *a priori* mínimo e permitindo que os dados informem primariamente a estimativa. Adicionalmente, foram incorporadas penalizações às distribuições *a priori*, especialmente sobre os incrementos da função de risco basal, com o objetivo de controlar suavidade e reduzir risco de sobreajuste, reforçando a estabilidade inferencial em cenários com amostras moderadas ou alta censura. Para avaliação preditiva, foi adotada uma estratégia de validação do tipo *hold-out* em dados de sobrevivência, permitindo mensurar o desempenho fora da amostra e comparar o modelo proposto a abordagens alternativas sob critérios de discriminação e calibração.

Todos os cálculos Bayesianos para o modelo proposto foram realizados usando o pacote `R2jags` (Su and Yajima, 2024) no ambiente de software estatístico `R` (R Core Team, 2024), versão 4.5.1. Para obter as distribuições marginais *a posteriori* e os resumos correspondentes, empregamos o algoritmo Metropolis-within-

Gibbs (MwG) descrito anteriormente. Especificamente, três cadeias MCMC independentes foram executadas, cada uma com 100000 iterações, e os 5% iniciais de cada cadeia foram descartados como *burn-in*. A amostra *a posteriori* final consistiu em 1000 extrações, obtidas retendo cada 95ª iteração das amostras pós-*burn-in*. Os resumos *a posteriori* obtidos são apresentados na Tabela 3.2.

Tabela 3.2: Resumos *a posteriori* para o modelo semiparamétrico de mistura com taxa de cura proposto baseado em $G(x) = x$ para os dados de câncer de mama do NKI.

				Intervalo de Credibilidade 95%	
Parâmetro	**Média**	**HR**	**Desvio Padrão**	**Limite Inferior**	**Limite Superior**
β_1 (idade)	-0.0231	0.9772	0.0157	-0.0500	0.0106
β_2 (quimio)	-0.0357	0.9649	0.3901	-0.9693	0.6212
β_3 (hormonal)	-1.0865	0.3374	0.5667	-2.0880	0.0379
β_4 (amputação)	0.1438	1.1547	0.3495	-0.5440	0.6934
β_5 (tipo hist)	0.1588	1.1721	0.3243	-0.4642	0.6984
β_6 (diâmetro)	0.0160	1.0161	0.0242	-0.0352	0.0573
β_7 (linfonodos pos)	0.1929	1.2128	0.0945	0.0351	0.3782
β_8 (grau)	1.0192	2.7710	0.1979	0.6605	1.4105
ρ (taxa de cura)	0.5550	—	0.0498	0.4679	0.6459

A partir da Tabela 3.2, os resultados do *modelo de Cox com taxa de cura* proposto indicam que o *grau do tumor* e o *número de linfonodos positivos* são preditores significativos (os intervalos de credibilidade de 95% para os parâmetros de regressão não incluem o valor zero) de sobrevida a longo prazo. Em particular, um grau de tumor mais alto (β_8) tem uma razão de riscos de 2.77, sugerindo que pacientes com tumores de grau mais alto enfrentam um aumento de quase três vezes no risco cumulativo em comparação com aquelas com graus mais baixos. Da mesma forma, cada linfonodo positivo adicional (β_7) aumenta o risco cumulativo em cerca de 21% (HR = 1.21). Esses achados ressaltam o forte impacto da agressividade do tumor e do envolvimento nodal nos desfechos a longo prazo. A comparação desses achados com o modelo padrão de riscos proporcionais de Cox revela algumas diferenças importantes. As outras covariáveis, incluindo idade, quimioterapia, terapia hormonal, tipo de cirurgia, tipo histológico e diâmetro do tumor, exibem razões de risco próximas à unidade com intervalos de credibilidade englobando o valor nulo, indicando nenhum efeito significativo no risco cumulativo dentro da estrutura do modelo de Cox com taxa de cura.

A taxa de cura estimada sob o modelo de Cox com taxa de cura é de 55.5% (intervalo de credibilidade de 95%: 46.8%–64.6%), indicando que aproximadamente 55.5% das pacientes devem ser sobreviventes de longo prazo que podem nunca experimentar o evento. Essa estimativa é ligeiramente inferior, mas amplamente consistente com a fração de cura de 60.4% obtida do estimador não paramétrico de KM, corroborando a capacidade do modelo de capturar a proporção curada enquanto incorpora os efeitos das covariáveis. Por fim, a Figura 3.3 ilustra o envelope simulado para os resíduos de Cox-Snell do modelo de Cox com taxa de cura. Os resíduos observados seguem a curva mediana simulada de perto e situam-se dentro dos limites do envelope, indicando um bom ajuste do modelo e nenhuma evidência de desvios importantes das suposições subjacentes.

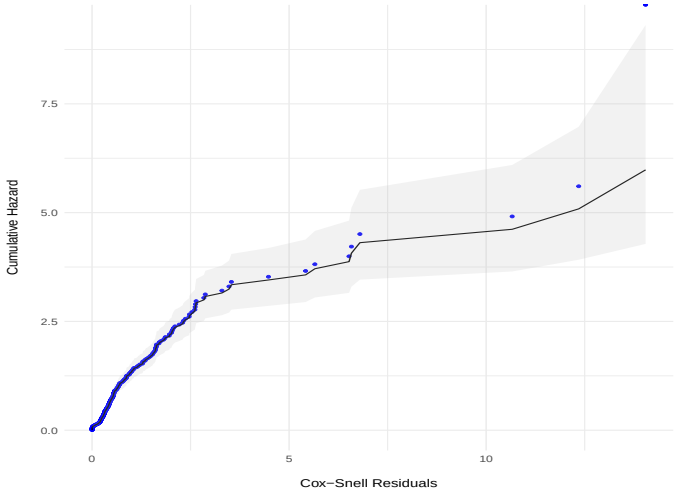


Figura 3.3: Envelope simulado para os resíduos de Cox-Snell.

O grau do tumor, identificado como um fator de risco significativo em ambos os modelos, teve uma relação proporcional clara com o risco de mortalidade. Cânceres de grau mais alto, que demonstram um comportamento celular mais agressivo, têm taxas de sobrevida mais baixas porque se disseminam precocemente e são menos suscetíveis à terapia tradicional (da Costa et al., 2024). Em contraste, o número de linfonodos positivos foi um preditor significativo apenas dentro do modelo de Cox com taxa de cura, com uma razão de riscos de 1.21 (intervalo de credibilidade de 95%: 1.04–1.46), indicando que cada linfonodo positivo adicional aumenta o risco cumulativo em aproximadamente 21%. Embora o tamanho do tumor não tenha sido incluído no modelo atual, está bem documentado que tumores maiores estão frequentemente associados a um maior envolvimento de linfonodos e metástases à distância, fatores conhecidos por impactar negativamente o prognóstico (Min et al., 2021). Essa associação biológica é ainda mais corroborada pela presença de características tumorais como aneuploidia, microambientes hipóxicos e angiogênese desregulada em tumores volumosos, as quais contribuem para a progressão do câncer e piores desfechos para as pacientes (Xu et al., 2016; Tang et al., 2021). Embora não seja estatisticamente significativa ao nível convencional de 5%, a terapia hormonal demonstrou uma razão de riscos de aproximadamente 0.34 (intervalo de credibilidade de 95%: 0.12–1.04), sugerindo uma tendência a um efeito protetor. Essa observação é consistente com os benefícios terapêuticos bem estabelecidos dos tratamentos hormonais no câncer de mama positivo para receptor hormonal, que funcionam primariamente inibindo a proliferação tumoral e reduzindo o risco de recorrência da doença (Murphy et al., 2012).

3.2 ARTIGO 2

Título: Random Survival Forests for Survival Prediction in Heart Failure: External Validation and Predictor Importance.

Status: Submetido para *Statistical Methods for Medical Research*.

3.2.1 DADOS DO ESTUDO

Para este estudo, foi considerada uma coorte previamente descrita por Ahmad et al. (2017). O conjunto de dados é composto por pacientes com insuficiência cardíaca admitidos no Institute of Cardiology and Allied Hospital, em Faisalabad, Paquistão, entre abril e dezembro de 2015. Todos os indivíduos tinham 40 anos ou mais, apresentavam disfunção sistólica do ventrículo esquerdo e foram classificados como classe funcional III ou IV da New York Heart Association (NYHA). A coorte inclui 299 pacientes e 13 variáveis demográficas, clínicas e laboratoriais rotineiramente coletadas no manejo da insuficiência cardíaca. As variáveis contínuas incluem idade (anos), sódio sérico (mEq/L), creatinina sérica (mg/dL), contagem de plaquetas e creatina fosfoquinase (CPK; mcg/L), além da fração de ejeção do ventrículo esquerdo (percentual). Indicadores binários codificam sexo, tabagismo, histórico de diabetes, hipertensão arterial e anemia.

Do ponto de vista da modelagem, este conjunto de dados apresenta diversas características relevantes para a análise de sobrevivência. Primeiro, a coorte representa uma população com disfunção sistólica avançada, implicando uma distribuição de risco relativamente concentrada e heterogeneidade limitada na classe funcional NYHA. Segundo, os preditores incluem tanto medidas específicas de órgãos (fração de ejeção, creatinina) quanto indicadores sistêmicos (idade, sódio, anemia), permitindo a avaliação de eixos prognósticos complementares. Terceiro, o tamanho amostral moderado em relação ao número de preditores motiva o uso de métodos regularizados e baseados em ensembles, capazes de capturar relações não lineares sem impor pressupostos paramétricos fortes. Um resumo das características basais e da distribuição do desfecho é apresentado na Tabela 3.3.

Tabela 3.3: Características basais estratificadas pelo estado vital ao final do acompanhamento. As variáveis contínuas são apresentadas como médias por grupo; as variáveis categóricas são apresentadas como frequências (percentuais).

Variáveis Contínuas		
Variável	**Óbito** (n = 96)	**Censurado** (n = 203)
Creatinina sérica (mg/dL)	1.83	1.18
Sódio sérico (mEq/L)	135.39	137.22
Creatina fosfoquinase (CPK, mcg/L)	670	540
Idade (anos)	65.21	58.76
Contagem de plaquetas	256.381	266.657
Fração de ejeção (%)	33.46	40.27
Variáveis Categóricas		
Variável	**Óbito** (n = 96)	**Censurado** (n = 203)
Sexo masculino	62 (64%)	132 (65%)
Fumante atual	30 (31%)	66 (32%)
Diabetes mellitus	40 (42%)	85 (42%)
Hipertensão arterial	40 (42%)	66 (32%)
Anemia	54 (56%)	83 (40%)

3.2.2 Fundamentação Estatística

Sejam $(T_i, \delta_i, \mathbf{X}_i)$, $i = 1, \ldots, 299$, os dados observados, onde T_i representa o tempo de acompanhamento, $\delta_i \in 0, 1$ indica óbito e $\mathbf{X}_i$ é o vetor de covariáveis demográficas, clínicas e laboratoriais (idade, fração de ejeção, creatinina sérica, sódio, anemia, status de diabetes, contagem de plaquetas, creatinofosfoquinase, sexo, status de tabagismo e status de pressão arterial). Para avaliar o desempenho preditivo fora da amostra, os dados foram particionados usando um desenho hold-out estratificado que preservou a distribuição marginal de eventos. Setenta por cento dos indivíduos dentro de cada estrato de desfecho (óbito versus censurado) foram alocados aleatoriamente para o conjunto de treinamento, e os 30% restantes formaram um conjunto de teste independente. Todo o desenvolvimento do modelo, incluindo o ajuste de hiperparâmetros, foi conduzido exclusivamente na amostra de treinamento para evitar o vazamento de informações.

A previsão de sobrevida foi realizada usando um RSF (Ishwaran et al., 2008), um estimador de ensemble não paramétrico que estende as florestas aleatórias de Breiman para dados censurados à direita. O RSF constrói um ensemble $\{\mathcal{T}_b\}_{b=1}^{B}$ de árvores de sobrevida cultivadas em subamostras extraídas sem reposição do conjunto de treinamento. Cada árvore é obtida via particionamento binário recursivo. Em cada nó interno $\mathcal{N}$, um subconjunto aleatório de m_{try} preditores candidatos é selecionado, ae a divisão ideal é determinada maximizando a estatística de divisão log-rank. Para uma divisão candidata separando $\mathcal{N}$ em nós filhos esquerdo e direito $\mathcal{N}_L$ e $\mathcal{N}_R$, a estatística log-rank assume a forma:

$$Z = \frac{\sum_j \left(d_{Lj} - \frac{Y_{Lj}}{Y_j} d_j \right)}{\sqrt{\sum_j \frac{Y_{Lj}}{Y_j} \left(1 - \frac{Y_{Lj}}{Y_j} \right) \frac{Y_j - d_j}{Y_j - 1} d_j}}, \tag{3.12}$$

onde d_{Lj} e Y_{Lj} denotam o número de eventos e sujeitos em risco no nó filho esquerdo no tempo t_j, e d_j, Y_j denotam os totais correspondentes no nó pai. A divisão que maximiza $|Z|$ é selecionada, visando assim a máxima separação entre as distribuições de sobrevida dos nós filhos. Dentro de cada nó terminal $\mathcal{N}$, a função de risco cumulativo é estimada usando o estimador de Nelson-Aalen:

$$\hat{H}_{\mathcal{N}}(t) = \sum_{t_j \leqslant t} \frac{d_j}{Y_j}, \qquad \hat{S}_{\mathcal{N}}(t) = \exp\{-\hat{H}_{\mathcal{N}}(t)\}. \tag{3.13}$$

Para um indivíduo com covariáveis $\mathbf{x}$, seja $\mathcal{N}_b(\mathbf{x})$ o nó terminal na árvore b o qual o indivíduo é atribuído. A estimativa de sobrevida do ensemble é definida como:

$$\hat{S}(t \mid \mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{S}_{\mathcal{N}_b(\mathbf{x})}(t), \tag{3.14}$$

o que reduz a variância via agregação. Os hiperparâmetros que controlam a complexidade do modelo incluíram o número de preditores candidatos considerados a cada divisão e o tamanho mínimo do nó terminal. A seleção foi conduzida usando a estimativa de erro out-of-bag (OOB), que explora o mecanismo de subamostragem para fornecer uma estimativa interna aproximadamente não viesada do erro de predição, sem a necessidade de uma

reamostragem separada. O desempenho preditivo foi avaliado na amostra de validação. A discriminação foi quantificada usando o índice de concordância de Harrell (Harrell et al., 1982),

$$C = \Pr\left(\hat{r}_i > \hat{r}_j \mid Y_i < Y_j\right), \tag{3.15}$$

onde $\hat{r}_i$ denota um escore de risco escalar derivado da incidência cumulativa predita em um horizonte de tempo pré-especificado. O erro de predição dependente do tempo foi avaliado usando o escore de Brier ponderado pela probabilidade inversa de censura (IPCW) (Prince et al., 2025):

$$\mathrm{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{I}(T_i \le t, , \delta_i = 1)}{\hat{G}(T_i)}\left(0 - \hat{S}(t \mid X_i)\right)^2 + \frac{\mathbb{I}(T_i > t)}{\hat{G}(t)}\left(1 - \hat{S}(t \mid X_i)\right)^2\right], \tag{3.16}$$

onde $\hat{G}(t)$ denota o estimador de Kaplan-Meier da distribuição de censura. A calibração foi avaliada comparando as curvas de sobrevida marginais do ensemble com o estimador não paramétrico de Kaplan-Meier. As propriedades de estratificação de risco foram examinadas agrupando os indivíduos de acordo com os quantis da incidência cumulativa predita e avaliando a separação das curvas de sobrevida. Por fim, a importância das variáveis baseada em permutação foi calculada avaliando o aumento no erro de predição induzido pela perturbação aleatória de cada covariável.

3.2.3 Resultados e Discussão

Para a modelagem preditiva, a coorte completa (n = 299) foi particionada usando um esquema hold-out estratificado para preservar a distribuição marginal do indicador de evento. O conjunto de treinamento resultante compreendeu 211 pacientes, incluindo 68 óbitos observados, enquanto o conjunto de teste conteve 88 pacientes. Um modelo RSF foi ajustado aos dados de treinamento usando o tempo de sobrevida como desfecho e as seguintes covariáveis: idade, fração de ejeção, creatinina sérica, sódio, anemia, status de diabetes, contagem de plaquetas, creatinofosfoquinase, sexo, status de tabagismo e status de pressão arterial. A configuração ideal para o modelo proposto foi $m_{\text{try}} = 9$, e $n_{\min} = 20$.

O ensemble final compreendeu 5000 árvores de sobrevida para garantir a estabilidade de Monte Carlo das predições do ensemble. A divisão dos nós baseou-se na estatística log-rank, visando diretamente a heterogeneidade nas distribuições de sobrevida entre os nós filhos. Em cada nó, foram avaliados 20 pontos de divisão candidatos selecionados aleatoriamente por variável. As árvores foram cultivadas usando subamostragem sem reposição, com um tamanho médio de amostra in-bag de 133 observações por árvore. Sob a restrição de nó terminal selecionada, as árvores exibiram uma média de 10.5 nós terminais, indicando profundidade moderada e complexidade controlada do modelo. O escore de probabilidade ranqueada contínua (CRPS, do inglês continuous ranked probability score) foi de 33.95 (CRPS padronizado = 0.141) e o erro de predição OOB foi de 0.286. Essas métricas indicam um desempenho preditivo estável nos dados de treinamento e sugerem que o RSF capturou efeitos não lineares de covariáveis e estruturas de interação relevantes.

A Figura 3.4 mostra a comparação entre a função de sobrevida média predita pelo RSF e o estimador de Kaplan-Meier calculado a partir do conjunto de teste independente (n = 88). Essa comparação, em particular, avalia o desempenho do modelo sob validação externa, uma vez que os dados de teste não foram utilizados durante o ajuste do modelo ou na otimização de hiperparâmetros.

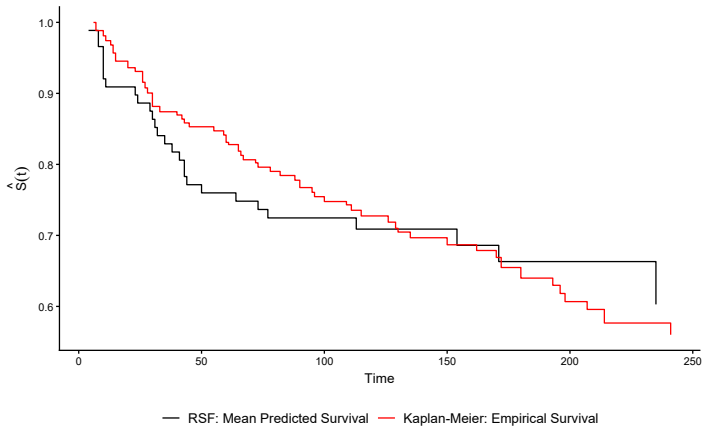


Figura 3.4: Comparação entre a função de sobrevida média predita pelo RSF (linha preta) e o estimador de Kaplan-Meier calculado a partir da amostra de teste independente (linha vermelha).

Ao longo do período de acompanhamento observado, a curva de sobrevida predita segue o estimador empírico de Kaplan-Meier com comportamento temporal semelhante. Nas fases inicial e intermediária do acompanhamento, onde a densidade de eventos é maior, as duas curvas exibem padrões de declínio comparáveis. A ausência de separação vertical sistemática sugere não haver superestimativa ou subestimativa global consistente das probabilidades de sobrevida em nível populacional. Em tempos de acompanhamento mais longos, observa-se alguma divergência. A estimativa baseada em ensemble apresenta uma trajetória mais suave devido à agregação entre as árvores, enquanto a função de Kaplan-Meier muda apenas nos tempos de eventos observados e torna-se cada vez mais variável à medida que o número em risco diminui. Esse comportamento reflete a redução no tamanho efetivo da amostra e a maior variância do estimador não paramétrico na região da cauda da distribuição de sobrevida.

A Figura 3.5 apresenta a comparação entre a função de incidência cumulativa média predita pelo RSF e a incidência cumulativa empírica derivada do estimador de Kaplan-Meier na amostra de teste independente. A incidência cumulativa foi calculada como $1 - \hat{S}(t)$, representando assim a probabilidade estimada de óbito ao longo do tempo. Ao longo do período de acompanhamento, a função de incidência cumulativa baseada no ensemble segue a trajetória geral do estimador empírico. Na fase inicial, a curva baseada no modelo aumenta mais rapidamente do que a curva de Kaplan-Meier, indicando uma maior probabilidade de evento a curto prazo predita em relação à incidência empírica observada. Durante os tempos de acompanhamento intermediários, as curvas exibem uma concordância mais estreita, com padrões de crescimento incremental semelhantes. Nas fases mais tardias do acompanhamento, a divergência entre as curvas torna-se mais aparente. A incidência cumulativa empírica aumenta em degraus maiores, refletindo os tempos de eventos individuais e o número reduzido em

risco, enquanto a estimativa baseada no ensemble evolui mais suavemente devido à agregação entre as árvores. Essas diferenças são consistentes com o aumento da variabilidade do estimador não paramétrico na cauda da distribuição de sobrevida e com a suavização inerente às predições do ensemble.

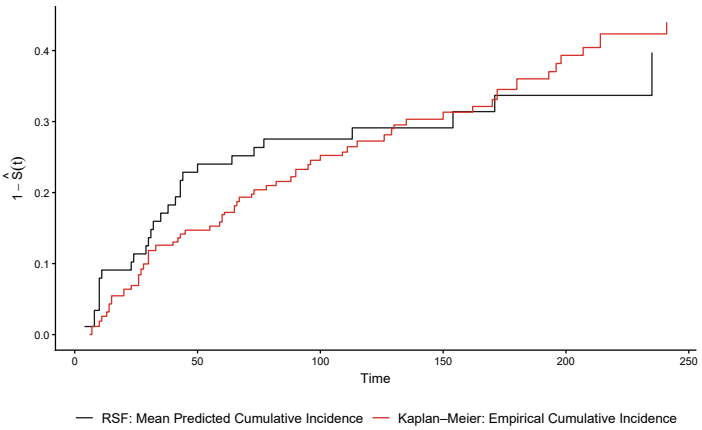


Figura 3.5: Comparação entre a incidência cumulativa média predita pelo RSF (linha preta) e a incidência cumulativa empírica estimada pelo método de Kaplan-Meier (linha vermelha) na amostra de teste independente.

A Figura 3.6 apresenta as curvas de sobrevida de Kaplan-Meier na amostra de teste independente estratificadas de acordo com os grupos de risco preditos derivados do RSF. Os escores de risco individuais foram definidos como a incidência cumulativa predita $\hat{F}(t^*) = 1 - \hat{S}(t^*)$ avaliada em um horizonte de tempo pré-especificado t^*, correspondente ao percentil 60 dos tempos de acompanhamento observados no conjunto de teste. Os pacientes foram categorizados em três grupos (baixo, intermediário e alto risco) de acordo com os tercis da distribuição de risco predita. As curvas de sobrevida resultantes ilustram uma separação clara entre os três estratos. Indivíduos classificados como de baixo risco exibem probabilidades de sobrevida consistentemente mais altas ao longo do acompanhamento, com declínio limitado ao longo do tempo. O grupo de risco intermediário apresenta uma redução moderada da sobrevida, enquanto o grupo de alto risco mostra uma diminuição acelerada na probabilidade de sobrevida, particularmente nas fases inicial e intermediária do acompanhamento. A ordenação das curvas permanece preservada ao longo do tempo, indicando uma estratificação de risco monotônica.

O índice de concordância de Harrell foi calculado na amostra de teste independente usando a incidência cumulativa predita avaliada no horizonte de tempo pré-especificado t^* (percentil 60 do acompanhamento). O índice de concordância estimado foi $C = 0.236$. Uma vez que a concordância quantifica a probabilidade de que, para um par comparável selecionado aleatoriamente, o sujeito com o maior risco predito experimente o evento mais cedo, esse valor indica uma discriminação limitada sob a definição de risco de horizonte fixo adotada nesta análise. Como o escore de risco foi definido em um horizonte de tempo fixo em vez de um índice prognóstico totalmente dependente do tempo, a concordância relatada reflete a discriminação em relação a essa medida de resumo específica. Além disso, a precisão da predição foi avaliada posteriormente por meio de escores de Brier dependentes do tempo calculados em quantis da distribuição de acompanhamento do teste. O escore de Brier variou de 0.144 em tempos de avaliação mais precoces a 0.201 em pontos de tempo mais tardios. O aumento

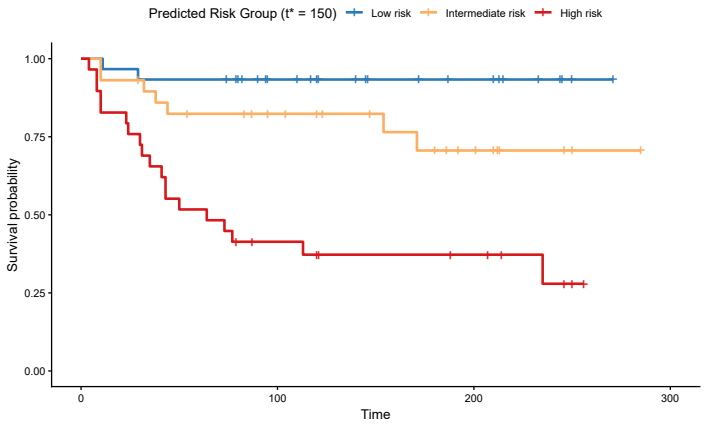


Figura 3.6: Curvas de sobrevida de Kaplan-Meier na amostra de teste independente estratificadas por grupos de risco preditos. Os grupos de risco (baixo, intermediário, alto) foram definidos de acordo com os tercis da incidência cumulativa predita avaliada no horizonte de tempo pré-especificado t^* (percentil 60 do acompanhamento).

gradual nos valores de Brier ao longo do tempo é consistente com a crescente incerteza nas predições de longo prazo à medida que o número em risco diminui e a censura se acumula.

A trajetória do erro de predição out-of-bag (OOB) em função do número de árvores é mostrada na Figura 3.7. O erro se estabiliza após a fase de crescimento inicial do ensemble, com as flutuações diminuindo à medida que o número de árvores aumenta. Além de aproximadamente 1000 árvores, o erro OOB exibe variação mínima, indicando a convergência das estimativas de risco do ensemble. O nível final de erro OOB é consistente com o valor relatado na fase de treinamento e corrobora a estabilidade numérica da especificação de 5000 árvores.

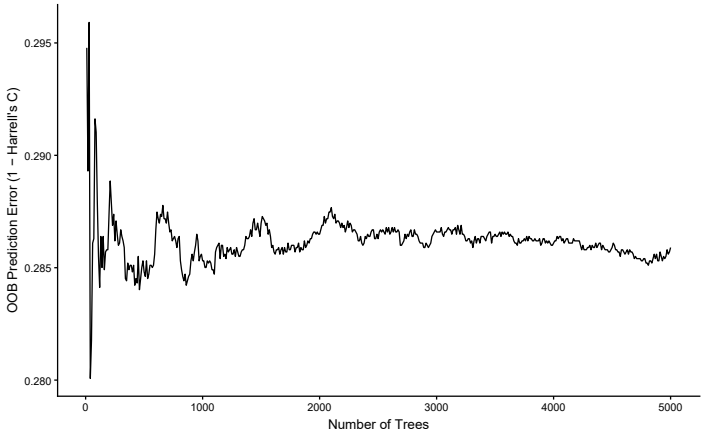


Figura 3.7: Erro de predição *out-of-bag* (1 - C de Harrell) em função do número de árvores no RSF.

A Figura 3.8 mostra o escore de Brier dependente do tempo avaliado na amostra de teste independente para o RSF e para um modelo de referência basal sem covariáveis. Ao longo dos tempos de acompanhamento, o RSF exibe consistentemente escores de Brier mais baixos do que o modelo basal, indicando um melhor desempenho preditivo em relação a um modelo que não incorpora informações de covariáveis. A magnitude da diferença é mais pronunciada em pontos de tempo mais precoces e permanece presente durante todo o acompanhamento. À medida que o tempo avança, os escores de Brier para ambos os modelos aumentam, o que

é consistente com o acúmulo de censura e a diminuição do número em risco, levando a uma maior incerteza na predição. Os intervalos de confiança aumentam em pontos de tempo mais tardios para o RSF, refletindo a redução do tamanho efetivo da amostra e o aumento da variabilidade na estimativa de risco a longo prazo. No entanto, o modelo mantém um erro de predição menor do que a especificação basal na maior parte do horizonte de tempo considerado, o que é uma indicação de um bom ajuste.

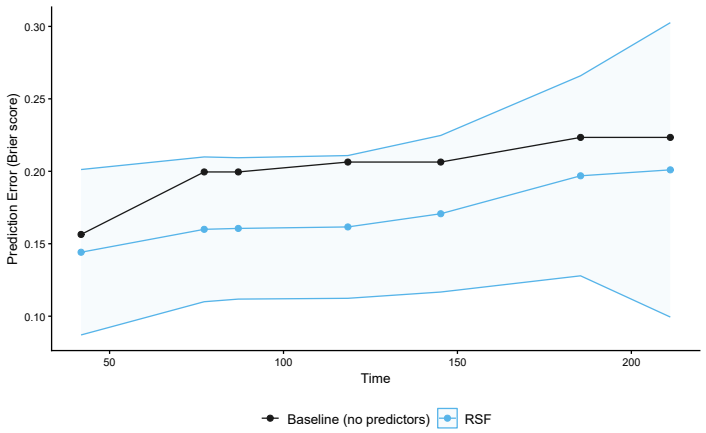


Figura 3.8: Escore de Brier dependente do tempo na amostra de teste independente comparando o RSF (azul) com um modelo basal sem preditores (preto). As regiões sombreadas representam os intervalos de confiança para as estimativas do RSF. Escores de Brier mais baixos indicam melhor precisão de predição.

Além das métricas de desempenho global, o exame das contribuições dos preditores fornece insights sobre a estrutura clínica capturada pelo modelo. A Figura 3.9 apresenta as medidas de importância das variáveis baseadas em permutação derivadas do RSF ajustado. A Fração de Ejeção (FE) foi identificada como o preditor mais influente, seguida pela Creatinina Sérica e pela Idade. Do ponto de vista clínico e fisiopatológico, essa classificação é coerente com a estrutura estabelecida da progressão da insuficiência cardíaca. Em termos de modelagem de risco, essas variáveis representam (i) a gravidade da disfunção sistólica ventricular (FE), (ii) a disfunção renal e a interação cardiorrenal (Creatinina) e (iii) a vulnerabilidade biológica global e a carga de multimorbidade (Idade).

A fração de ejeção do ventrículo esquerdo (FE) tem um papel fundamental na fisiopatologia, classificação e avaliação prognóstica da insuficiência cardíaca (Yancy et al., 2013; Ponikowski et al., 2016). Como uma medida integrativa do desempenho sistólico do ventrículo esquerdo, a FE reflete a consequência hemodinâmica da disfunção miocárdica, incluindo débito cardíaco reduzido, pressões de enchimento elevadas e ativação de sistemas neuro-hormonais que impulsionam o remodelamento ventricular progressivo. Na insuficiência cardíaca sistólica avançada, uma FE mais baixa está consistentemente associada a um risco aumentado de mortalidade em registros observacionais e ensaios clínicos randomizados. Sua liderança na importância de permutação no presente modelo indica que, dentro desta coorte NYHA III–IV, a heterogeneidade na sobrevida é estruturada principalmente em torno da gravidade da disfunção sistólica. Esse achado é concordante com modelos prognósticos estabelecidos, como o Seattle Heart Failure Model e o escore MAGGIC, os quais mantêm

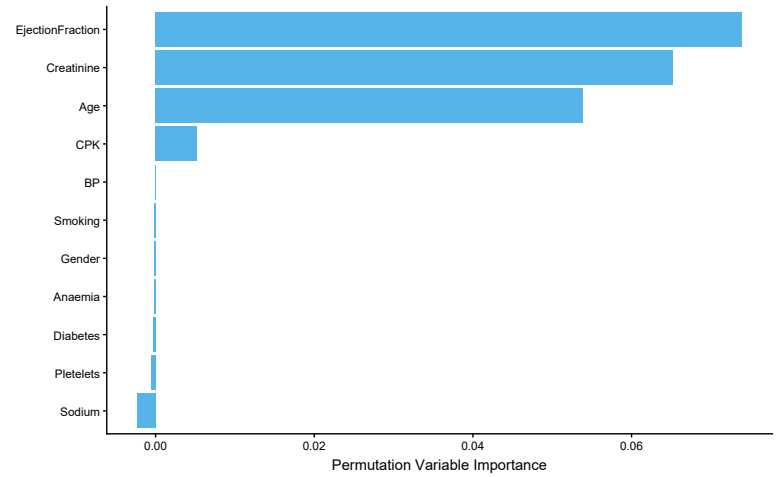


Figura 3.9: Importância das variáveis baseada em permutação a partir do RSF. Os valores de importância representam o aumento no erro de predição após a permutação aleatória de cada covariável.

a FE como um determinante central do risco de mortalidade (Levy et al., 2006; Pocock et al., 2013).

A função renal, representada aqui pela creatinina sérica, surge como a segunda principal dimensão da estratificação prognóstica. Na insuficiência cardíaca avançada, o comprometimento renal reflete mais do que uma doença renal isolada; representa a expressão clínica do acoplamento cardiorrenal, englobando perfusão renal reduzida, pressões venosas elevadas, ativação neuro-hormonal e resposta natriurética prejudicada (Hillege et al., 2000; Smith et al., 2006). Inúmeros estudos de coorte demonstram que níveis crescentes de creatinina estão independentemente associados a uma maior mortalidade, mesmo após o ajuste para a função cardíaca (Hillege et al., 2000). A idade, por outro lado, captura a vulnerabilidade biológica sistêmica em vez de falência órgão-específica. Ela integra fragilidade, carga cumulativa de comorbidades, rigidez vascular e reserva adaptativa diminuída (Roger, 2013). Grandes análises prognósticas agrupadas confirmam que a idade mantém seu valor preditivo independente além dos parâmetros hemodinâmicos e laboratoriais (Pocock et al., 2013).

Em contraste, o sódio sérico, a anemia, o status de diabetes e a contagem de plaquetas exibiram importância de permutação limitada na presente análise. Embora a hiponatremia tenha sido consistentemente associada à ativação neuro-hormonal, ao desequilíbrio hídrico e a desfechos adversos na insuficiência cardíaca crônica e descompensada (Klein et al., 2005), e a anemia tenha sido associada à capacidade funcional reduzida e ao aumento do risco de mortalidade (Groenveld et al., 2008), seu valor prognóstico incremental pode diminuir uma vez que marcadores dominantes de gravidade da doença sejam incorporados ao modelo. Na insuficiência cardíaca sistólica avançada, a FE e a creatinina sérica resumem eixos fisiopatológicos centrais, nomeadamente a falência de bomba ventricular e a interação cardiorrenal, que podem mediar uma fração substancial do risco a jusante. Preditores parcialmente correlacionados com esses mecanismos podem, portanto, exibir contribuição adicional limitada em configurações multivariáveis, apesar de associações independentes em análises não ajustadas. Essa estrutura de importância de variáveis é consistente com o estudo original de Ahmad et al. (2017), que aplicou um modelo de riscos proporcionais de Cox ao mesmo conjunto de dados e identificou a fração de ejeção reduzida e a creatinina sérica elevada como os principais preditores independentes de mortalidade.

3.3 ARTIGO 3

Título: A Penalized Bayesian Semiparametric Cumulative Hazard Transformation Model for Long-Term Survival.

Observação: O artigo consolida os resultados finais obtidos no âmbito do projeto, os quais são apresentados e descritos de forma detalhada neste relatório, assegurando coerência entre a metodologia adotada, as evidências empíricas e as conclusões alcançadas. Ressalta-se, contudo, que o manuscrito ainda se encontra em fase de finalização, passando por ajustes estruturais e refinamentos textuais antes de sua submissão.

3.3.1 DADOS DO ESTUDO

Para a aplicação da metodologia proposta, consideramos a coorte retrospectiva analisada por Puchner et al. (2017), publicamente disponível no repositório PLOS ONE (DOI: 10.1371/journal.pone.0172203). O conjunto de dados inclui pacientes diagnosticados com sarcoma pélvico maligno primário que foram submetidos a tratamento cirúrgico em um centro terciário de oncologia ortopédica entre 1980 e 2012. Todos os casos foram confirmados histologicamente, e os pacientes foram acompanhados longitudinalmente para desfechos de sobrevida após a cirurgia. A coorte compreende 147 pacientes e inclui covariáveis demográficas, patológicas e relacionadas ao tratamento coletadas no momento basal (*baseline*). As variáveis demográficas incluem idade no diagnóstico e sexo. As características do tumor englobam subtipo histológico, grau do tumor e volume do tumor. As variáveis cirúrgicas incluem o estado da margem de ressecção e o tipo de reconstrução. Indicadores adicionais de tratamento codificam a administração de quimioterapia e radioterapia. O desfecho primário é a sobrevida global, definida como o tempo da intervenção cirúrgica até a morte ou último acompanhamento, com censura à direita para indivíduos vivos no final da observação. Um resumo das características basais e da distribuição dos desfechos é apresentado na Tabela 3.4.

Do ponto de vista da modelagem, este conjunto de dados apresenta características estruturais que motivam o uso da estrutura de transformação semiparamétrica Bayesiana proposta com *a prioris* penalizadas. A população do estudo exibe heterogeneidade na biologia tumoral e na apresentação clínica, uma vez que os sarcomas pélvicos incluem múltiplos subtipos histológicos com dinâmicas de crescimento e padrões de sobrevida distintos. Tal variabilidade pode levar a estruturas de risco que não são adequadamente representadas por uma formulação estritamente proporcional, apoiando assim o uso de uma especificação flexível baseada em transformação. As covariáveis cobrem domínios prognósticos complementares, incluindo características demográficas, grau e volume do tumor, status da margem cirúrgica e indicadores de tratamento adjuvante. Esses fatores podem influenciar a sobrevida por meio de mecanismos não lineares ou não multiplicativos. O componente de transformação do modelo permite a modificação adaptativa aos dados da estrutura de risco cumulativo, enquanto a representação semiparamétrica em degraus (*stepwise*) do risco basal evita suposições paramétricas restritivas. Além disso, o tamanho moderado da amostra em relação ao número de preditores requer regularização para garantir uma estimação estável. A incorporação de *a prioris* penalizadas, como estruturas de

passeio aleatório (*random-walk*) de segunda ordem nos incrementos basais e coeficientes de transformação, induz suavização e encolhimento (*shrinkage*), reduzindo o sobreajuste (*overfitting*) enquanto preserva a flexibilidade do modelo proposto. Dentro desta estrutura Bayesiana unificada, os efeitos de regressão, os parâmetros de transformação e os componentes de risco basal são estimados conjuntamente, permitindo uma quantificação coerente da incerteza.

Tabela 3.4: Características basais e dados operatórios da coorte retrospectiva analisada por Puchner et al. (2017).

Variável	**Valor**
Pacientes (número; %)	147
Sexo	
Masculino	68 (46%)
Feminino	79 (54%)
Histologia	
Condrossarcoma	54 (37%)
Sarcoma de Ewing/PNET	37 (25%)
Osteossarcoma	32 (22%)
Leiomiossarcoma	4 (3%)
Sarcoma–Não especificado de outra forma	4 (3%)
Hemangiopericitoma	3 (2%)
Outros	13 (9%)
Grau	
G3	101 (69%)
G2	38 (26%)
G1	8 (5%)
Idade no momento da cirurgia (anos; DP)	38 ± 20
Tamanho (cm^3; DP)	1023 ± 1848
Localização	
Íleo	110 (75%)
Ísquio	9 (6%)
Púbis	28 (19%)
Envolvimento periacetabular	67 (46%)
Tipo de cirurgia	
Ressecção sem reconstrução	46 (31%)
Reconstrução endoprotética	47 (32%)
Reconstrução biológica	21 (14%)
Hemipelvectomia interna com transposição	14 (10%)
Hemipelvectomia externa	19 (13%)
Tipo de ressecção	
Tipo I	27 (18%)
Tipo III	14 (10%)
Tipo I/II	19 (13%)
Tipo I/IV	10 (7%)
Tipo II/III	5 (3%)
Tipo II/III	25 (17%)
Tipo I/II/III	33 (22%)
Tipo I/II/III/IV	14 (10%)

Em nossa análise, o tempo de sobrevida global (SG) foi definido como o tempo desde a cirurgia até a morte ou último acompanhamento, com censura à direita para os indivíduos vivos no final da observação. O indicador de evento foi codificado como $\delta = 1$ para morte e $\delta = 0$ para censura. O conjunto de covariáveis foi construído para refletir variáveis demográficas, patológicas e relacionadas ao tratamento disponíveis na coorte. Preditores binários incluíram sexo, status de reconstrução, recebimento de radioterapia e recebimento de quimioterapia. Preditores contínuos compreenderam uma medida de alteração operatória e volume do tumor; ambos foram padronizados antes da análise para melhorar a estabilidade numérica e a calibração da *a priori*. Variáveis categóricas representando o grau do tumor (três níveis) e o tipo cirúrgico (cinco categorias) foram incorporadas usando expansões de indicadores, produzindo uma matriz de delineamento de dimensão p usada em ambos os componentes de incidência e latência do modelo.

3.3.2 Modelo de Transformação Monótona Adaptativa aos Dados

A função de transformação $G(\cdot;\theta)$ permite relações estruturais flexíveis entre as covariáveis e o risco cumulativo. Escolhas clássicas, como riscos proporcionais, chances proporcionais ou transformações do tipo logarítmica, correspondem a formas paramétricas fixas. Embora convenientes, tais especificações podem ser restritivas quando a relação de risco subjacente desvia dessas estruturas canônicas. Para aumentar a flexibilidade enquanto se preserva a monotonicidade e a interpretabilidade, introduzimos uma transformação adaptativa aos dados construída por meio de um gerador de *spline* monótono. Especificamente, definimos a função de transformação como,

$$G(x;\boldsymbol{\theta}) = \int_0^x \exp\left\{\sum_{k=1}^{K} \theta_k B_k\big(\log(1+u)\big)\right\} du, \qquad x \geqslant 0, \tag{3.17}$$

onde $\{B_k(\cdot)\}_{k=1}^{K}$ denota um conjunto de funções de base *spline* e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ é um vetor de parâmetros de transformação. Por construção, a derivada de G é,

$$G'(x;\boldsymbol{\theta}) = \exp\left\{\sum_{k=1}^{K} \theta_k B_k\big(\log(1+x)\big)\right\} > 0, \qquad x \geqslant 0, \tag{3.18}$$

o que garante a monotonicidade estrita de $G(\cdot;\boldsymbol{\theta})$. Consequentemente, o risco cumulativo $\Lambda(t \mid \mathbf{Z}) = G\left(e^{\beta^\top \mathbf{Z}}\Lambda_0(t);\boldsymbol{\theta}\right)$, permanece não decrescente em t, garantindo a validade do modelo de sobrevida. O uso de $\log(1+u)$ dentro da base *spline* estabiliza o comportamento próximo de zero e reduz a sensibilidade numérica para valores grandes de u. Além disso, os modelos de transformação clássicos surgem como casos especiais de (3.17). Por exemplo, valores constantes de θ_k produzem transformações lineares correspondentes aos riscos proporcionais, enquanto escolhas estruturadas específicas aproximam-se da curvatura do tipo chances proporcionais. Assim, a formulação proposta estende os modelos clássicos enquanto permite que os dados determinem a distorção estrutural apropriada do risco basal. Para evitar o sobreajuste (*overfitting*) e garantir uma variação estrutural suave, impomos uma penalização de passeio aleatório (*random-walk*) de segunda ordem aos

coeficientes *spline*. Sendo $\boldsymbol{\theta}$ o vetor de parâmetros de transformação, assumimos,

$$p(\boldsymbol{\theta} \mid \tau_\theta) \propto \exp\left\{-\frac{\tau_\theta}{2}\sum_{k=3}^{K}(\theta_k - 2\theta_{k-1} + \theta_{k-2})^2\right\}, \tag{3.19}$$

onde τ_θ é um parâmetro de precisão de suavização. Uma hiper-priori Gama é atribuída a τ_θ, ou seja, $\tau_\theta \sim \text{Gamma}(a_\theta, b_\theta)$. Sob esta especificação, a função de sobrevida dos suscetíveis é definida por,

$$S_0(t \mid \mathbf{Z}) = \exp\left\{-G\left(e^{\boldsymbol{\beta}^\top \mathbf{Z}}\Lambda_0(t); \boldsymbol{\theta}\right)\right\}. \tag{3.20}$$

Todos os componentes são estimados conjuntamente dentro da estrutura bayesiana descrita anteriormente, com $\boldsymbol{\theta}$ atualizado como um bloco de parâmetros adicional no esquema MCMC.

3.3.3 Resultados

Para a linha de base semiparamétrica, o suporte foi definido pelos tempos de evento distintos ordenados $\{t_{(j)}\}_{j=1}^{J}$ obtidos a partir das mortes observadas, com J igual ao número de tempos de evento únicos. Os incrementos basais foram inicializados em $\gamma_{0j} = 0.02$ e receberam *a prioris* gama fracamente informativas com hiperparâmetros $a_0 = b_0 = 0.01$. O componente de transformação foi representado usando um gerador de *spline* monótono com $K = 4$ funções de base cúbica avaliadas em $\log(1 + u)$, e nós de contorno fixados em $(0, \log\{1 + 2\max(\text{SG})\})$ para estabilizar a extrapolação na escala transformada. Aos coeficientes de regressão no bloco de latência foram atribuídas *a prioris* independentes $\mathcal{N}(0, 1)$. Para o bloco de incidência, o intercepto foi centralizado em uma probabilidade de cura basal de ≈ 0.38, correspondendo ao nível empírico de sobrevida a longo prazo sugerido pelo estimador de Kaplan-Meier, por meio de uma *a priori* normal na escala logit, enquanto os coeficientes restantes receberam *a prioris* independentes $\mathcal{N}(0, 1)$. *A prioris* penalizadas foram impostas via penalidades de passeio aleatório de segunda ordem nos coeficientes de transformação e nos incrementos log-basais $\psi_j = \log(\gamma_{0j})$, com precisões de suavização τ_θ e τ_ψ recebendo hiper-prioris gama $\text{Gamma}(1, 1)$.

A inferência *a posteriori* foi conduzida usando um amostrador Metropolis-within-Gibbs. Os blocos de parâmetros $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\theta})$ foram atualizados por propostas de passeio aleatório gaussiano, e os incrementos basais foram atualizados na escala logarítmica através de uma proposta de passeio aleatório para $\boldsymbol{\psi}$. Os desvios padrão das propostas foram fixados em 0.05. Atualizações condicionais para os parâmetros de suavização (τ_ψ, τ_θ) foram realizadas por amostragem de Gibbs a partir de suas distribuições condicionais completas gama. A cadeia de Markov foi executada por 110000 iterações, com as primeiras 10000 iterações descartadas como *burn-in*, resultando em 100000 amostras *a posteriori* retidas. A convergência foi avaliada por meio de gráficos de traço (*trace plots*) e estabilidade dos resumos *a posteriori*. A inferência baseou-se nas médias *a posteriori* e nos intervalos de credibilidade de 95%, e a adequação do modelo foi avaliada através da comparação da curva de sobrevida marginal *a posteriori* com o estimador de Kaplan-Meier e através de diagnósticos baseados em resíduos. Os resumos *a posteriori* dos parâmetros de regressão, transformação e suavização são reportados na Tabela 3.5.

Tabela 3.5: Resumos *a posteriori* para o modelo de cura de transformação semiparamétrica bayesiana proposto. Médias *a posteriori* e intervalos de credibilidade de 95% são reportados.

				Intervalo de Credibilidade 95%	
Bloco	**Parâmetro**	**Média**	**DP**	**2.5%**	**97.5%**
Componente de latência (β)					
	$\beta_{\text{Masculino}}$	0.0027	0.0110	-0.0152	0.0290
	$\beta_{\text{Reconstrução}}$	-0.0044	0.0081	-0.0222	0.0118
	β_{Radio}	-0.0032	0.0084	-0.0163	0.0187
	β_{Quimio}	0.0011	0.0107	-0.0185	0.0201
	β_{AltOper_z}	-0.0011	0.0084	-0.0131	0.0182
	β_{Volume_z}	0.0026	0.0087	-0.0141	0.0137
	$\beta_{\text{Grau G2}}$	0.0006	0.0088	-0.0138	0.0162
	$\beta_{\text{Grau G3}}$	-0.0017	0.0092	-0.0171	0.0156
	$\beta_{\text{Cirurgia 2}}$	0.0006	0.0111	-0.0205	0.0211
	$\beta_{\text{Cirurgia 3}}$	-0.0001	0.0081	-0.0126	0.0160
	$\beta_{\text{Cirurgia 4}}$	-0.0019	0.0117	-0.0255	0.0171
	$\beta_{\text{Cirurgia 5}}$	-0.0007	0.0094	-0.0204	0.0152
Componente de incidência (γ)					
	γ_0 (Intercepto)	-0.0584	0.0113	-0.0804	-0.0375
	$\gamma_{\text{Masculino}}$	0.0034	0.0093	-0.0220	0.0200
	$\gamma_{\text{Reconstrução}}$	-0.0047	0.0097	-0.0202	0.0211
	γ_{Radio}	0.0017	0.0107	-0.0167	0.0155
	γ_{Quimio}	-0.0060	0.0088	-0.0208	0.0103
	$\gamma_{\text{AltOper}_z}$	-0.0054	0.0107	-0.0219	0.0174
	γ_{Volume_z}	-0.0019	0.0111	-0.0215	0.0183
	$\gamma_{\text{Grau G2}}$	0.0008	0.0066	-0.0103	0.0131
	$\gamma_{\text{Grau G3}}$	0.0018	0.0107	-0.0134	0.0179
	$\gamma_{\text{Cirurgia 2}}$	-0.0029	0.0085	-0.0164	0.0127
	$\gamma_{\text{Cirurgia 3}}$	0.0026	0.0091	-0.0161	0.0216
	$\gamma_{\text{Cirurgia 4}}$	-0.0050	0.0071	-0.0161	0.0157
	$\gamma_{\text{Cirurgia 5}}$	-0.0046	0.0086	-0.0262	0.0124
Componente de transformação (θ)					
	θ_1	-0.1940	0.1344	-0.4236	0.0655
	θ_2	0.1840	0.0378	0.1123	0.2513
	θ_3	0.1781	0.0884	0.0154	0.3205
	θ_4	0.1671	0.1463	-0.0971	0.3256

Os resultados indicam que, na componente de latência, os coeficientes $\beta_{\text{Masculino}}$, $\beta_{\text{Reconstrução}}$, β_{Radio}, β_{Quimio}, β_{AltOper_z}, β_{Volume_z}, $\beta_{\text{Grau G2}}$, $\beta_{\text{Grau G3}}$ e os parâmetros associados às categorias de cirurgia apresentam médias muito próximas de zero e intervalos de credibilidade que incluem o valor nulo. Isso sugere ausência de evidência posterior consistente de efeito dessas covariáveis sobre o tempo até o evento entre os indivíduos suscetíveis. Na componente de incidência, o intercepto γ_0 apresenta intervalo de credibilidade abaixo de zero,

indicando uma probabilidade basal de cura inferior a 0,5 no grupo de referência. As demais covariáveis, como $\gamma_{\text{Masculino}}$, $\gamma_{\text{Reconstrução}}$, γ_{Radio}, γ_{Quimio}, $\gamma_{\text{AltOper}_z}$, γ_{Volume_z} e os efeitos de grau e cirurgia, possuem intervalos que incluem zero, sugerindo que não há evidência forte de associação com a probabilidade de cura após ajuste conjunto. No componente de transformação, observa-se que θ_2 e θ_3 apresentam intervalos de credibilidade totalmente positivos, indicando evidência de estrutura não proporcional no risco cumulativo. Já θ_1 e θ_4 possuem intervalos que incluem zero, sugerindo contribuição menos pronunciada desses termos.

É importante destacar que os resultados aqui apresentados devem ser interpretados como resultados iniciais do processo de modelagem. Embora forneçam uma primeira evidência sobre o comportamento das covariáveis nas componentes de incidência, latência e transformação, análises adicionais serão conduzidas com o objetivo de aprofundar a investigação inferencial. Em etapas subsequentes, serão exploradas diferentes especificações para os hiperparâmetros das distribuições *a priori*, avaliando a sensibilidade das estimativas a escolhas alternativas de informação prévia. Também serão consideradas outras famílias de prioris e estratégias de penalização, especialmente sobre os componentes de transformação e sobre possíveis estruturas de suavização, com vistas a aprimorar a estabilidade, flexibilidade e capacidade preditiva do modelo. Essas extensões permitirão uma avaliação mais abrangente da robustez dos resultados e da adequação da especificação Bayesiana adotada.

CAPÍTULO 4

CONSIDERAÇÕES FINAIS

O projeto de pós-doutorado teve como objetivo o desenvolvimento, implementação e avaliação de abordagens modernas para análise de sobrevivência em diferentes cenários clínicos, com ênfase em (i) modelos semiparamétricos flexíveis capazes de acomodar violações da suposição de riscos proporcionais e (ii) métodos de *machine learning* voltados à predição de desfechos em presença de censura. Ao longo do período, foram conduzidas atividades de pesquisa metodológica e aplicada que resultaram na elaboração de três frentes principais materializadas em manuscritos científicos, integrando contribuições estatísticas, computacionais e de interpretação clínica.

No primeiro eixo, o projeto avançou na literatura de modelos com fração de cura ao propor e estudar uma extensão do modelo de Cox via uma formulação Bayesiana semiparamétrica com mistura e transformação do risco cumulativo, aplicada a uma coorte clássica de câncer de mama do NKI. A análise mostrou que a violação da suposição de riscos proporcionais em covariáveis clinicamente relevantes motiva o uso de estruturas mais gerais. A modelagem proposta permitiu quantificar a taxa de cura e avaliar simultaneamente efeitos de covariáveis sobre o risco, produzindo inferência coerente e interpretação alinhada ao contexto de sobreviventes de longo prazo. Além disso, foi adotada uma estratégia de validação do tipo *hold-out* adaptada para dados de sobrevivência, com particionamento em conjuntos de treinamento e teste, possibilitando avaliar o desempenho preditivo fora da amostra e comparar o modelo proposto a abordagens tradicionais sob critérios de discriminação e calibração. Esta etapa consolidou a capacidade do aluno em formular modelos, construir a verossimilhança e implementar inferência Bayesiana com MCMC em um ambiente reprodutível.

No segundo eixo, o projeto incorporou métodos de *machine learning* para predição de sobrevivência, com destaque para Florestas Aleatórias de Sobrevivência (RSF) aplicadas a uma coorte de insuficiência cardíaca. Foi implementado um desenho de validação *hold-out* estratificado e conduzida a seleção de hiperparâmetros via erro *out-of-bag*, com avaliação preditiva fora da amostra por medidas dependentes do tempo e inspeção de calibração. Além da predição, a análise de importância de variáveis forneceu evidências coerentes com a fisiopatologia da doença, destacando preditores dominantes como fração de ejeção, creatinina e idade. Essa frente reforçou a contribuição do projeto ao integrar aprendizagem estatística e avaliação preditiva em um desenho compatível com censura e incerteza longitudinal.

No terceiro eixo, foi desenvolvido um modelo Bayesiano semiparamétrico de transformação monótona adaptativa aos dados, com priors penalizadas para controle de suavidade e redução de sobreajuste, aplicado a uma coorte de sarcoma pélvico. A formulação combinou uma transformação por *spline* monótono e penalizações do tipo passeio aleatório de segunda ordem tanto para os coeficientes de transformação quanto para incrementos

do risco basal. Os resultados iniciais indicaram que, embora os efeitos lineares nas componentes de incidência e latência tenham apresentado magnitudes pequenas sob a especificação atual, a estrutura de transformação capturou evidência de não proporcionalidade (especialmente via θ_2 e θ_3), sustentando a pertinência da metodologia proposta para representar heterogeneidade estrutural em dados clínicos reais. Ressalta-se que esta etapa permanece em aprofundamento, com análise de sensibilidade planejada para diferentes hiperparâmetros, alternativas de prioris e esquemas de penalização, de modo a avaliar robustez inferencial e potenciais ganhos preditivos.

De forma integrada, o projeto consolidou uma agenda de pesquisa que conecta inferência Bayesiana semiparamétrica, modelagem com fração de cura, transformação do risco cumulativo e métodos de *machine learning* para sobrevivência, com avaliação preditiva baseada em validação externa e métricas apropriadas para censura. Além das contribuições científicas, o aluno demonstrou domínio técnico na implementação computacional (incluindo rotinas MCMC e fluxos de validação), capacidade de interpretação clínica dos resultados e maturidade acadêmica na redação de manuscritos e relatórios. Em síntese, o projeto atingiu seus objetivos ao produzir um conjunto coerente de resultados metodológicos e aplicados, estabelecendo bases sólidas para continuidade das pesquisas em modelagem de sobrevivência flexível, regularização Bayesiana e aprendizagem estatística em saúde.

REFERÊNCIAS BIBLIOGRÁFICAS

Achcar, J. A. and Barili, E. (2024). Semiparametric transformation model in presence of cure fraction: a hierarchical bayesian approach assuming the unknown hazards as latent factors. *Statistical Methods & Applications*, 33(2):357–380.

Achcar, J. A., Coelho-Barros, E. A., and Mazucheli, J. (2012). Cure fraction models using mixture and non-mixture models. *Tatra Mountains Mathematical Publications*, 51(1):1–9.

Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., and Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. *PloS one*, 12(7):e0181001.

Armbruster, J. W. and Lujan, N. K. (2016). Data from: A new species of Peckoltia from the Upper Orinoco (Siluriformes, Loricariidae). Dryad Digital Repository.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89:659–668.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794. ACM.

Christin, S., Hervet, É., and Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

da Costa, R. E. A. R., da Silva Rodrigues, S. V., Mendes, A. C. V., Costa, R. R. E., de Sá, M. C., Vieira, S. C., de Sá, C. E. C., and de Vasconcelos Valença, R. J. (2024). Sobrevida e fatores prognósticos em pacientes com câncer de mama: Estudo de coorte. *Revista Brasileira de Cancerologia*, 70(4).

Dauda, K. A. (2022). Optimal tuning of random survival forest hyperparameter with an application to liver disease. *Malaysian Journal of Medical Sciences*, 29(6):67–76.

De Angelis, R., Capocaccia, R., Hakulinen, T., Soderman, B., and Verdecchia, A. (1999). Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in medicine*, 18(4):441–454.

Dietrich, S. et al. (2016). Random survival forest in practice: A method for modelling complex metabolomics data in time to event analysis. *International Journal of Epidemiology*, 45(5):1406–1420.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046.

Fernandes, L. M. (2014). *Inferência Bayesiana em modelos discretos com fração de cura*. PhD thesis.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society B*, (115):513–583.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526.

Groenveld, H. F., Januzzi, J. L., Damman, K., van Wijngaarden, J., Hillege, H. L., and van Veldhuisen, D. J. (2008). Anemia and mortality in heart failure patients: A systematic review and meta-analysis. *Journal of the American College of Cardiology*, 52(10):818–827.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546.

Hillege, H. L., Girbes, A. R. J., de Kam, P. J., Boomsma, F., de Zeeuw, D., Charlesworth, A., Hampton, J. R., and van Veldhuisen, D. J. (2000). Renal function, neurohormonal activation, and survival in patients with chronic heart failure. *Circulation*, 102(2):203–210.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

Klein, L., O'Connor, C. M., Leimberger, J. D., Gattis-Stough, W., Pina, I. L., Felker, G. M., Adams, K. F., and Califf, R. M. (2005). Lower serum sodium is associated with increased short-term mortality in hospitalized patients with worsening heart failure. *Circulation*, 111(19):2454–2460.

Lambert, P. C., Thompson, J. R., Weston, C. L., and Dickman, P. W. (2006). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8(3):576–594.

Lambert, P. C., Thompson, J. R., Weston, C. L., and Dickman, P. W. (2007). Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8(3):576–594.

Levy, W. C., Mozaffarian, D., Linker, D. T., Sutradhar, S. C., Anker, S. D., Cropp, A. B., Anand, I. S., Maggioni, A. P., Burton, P., Sullivan, M. D., Pitt, B., Poole-Wilson, P. A., Mann, D. L., and Packer, M. (2006). The seattle heart failure model: Prediction of survival in heart failure. *Circulation*, 113(11):1424–1433.

Maller, R. A. and Zhou, X. (1996). Survival analysis with long-term survivors.

Min, S. K., Lee, S. K., Woo, J., Jung, S. M., Ryu, J. M., Yu, J., Lee, J. E., Kim, S. W., Chae, B. J., and Nam, S. J. (2021). Relation between tumor size and lymph node metastasis according to subtypes of breast cancer. *Journal of Breast Cancer*, 24(1):75.

Murphy, C., Bartholomew, L., Carpentier, M., Bluethmann, S., and Vernon, S. (2012). Adherence to adjuvant hormonal therapy among breast cancer survivors in clinical practice: a systematic review. *Breast Cancer Research and Treatment*, 134(2):459–478.

Othus, M., Barlogie, B., LeBlanc, M. L., and Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, 18(14):3731–3736.

Pocock, S. J., Ariti, C. A., McMurray, J. J. V., Maggioni, A., Köber, L., Squire, I. B., Swedberg, K., Dobson, J., Poppe, K., Whalley, G. A., and Doughty, R. N. (2013). Predicting survival in heart failure: A risk score based on 39 372 patients from 30 studies. *European Heart Journal*, 34(19):1404–1413.

Ponikowski, P., Voors, A. A., Anker, S. D., Bueno, H., Cleland, J. G. F., Coats, A. J. S., Falk, V., González-Juanatey, J. R., Harjola, V.-P., Jankowska, E. A., Jessup, M., Linde, C., Nihoyannopoulos, P., Parissis, J. T., Pieske, B., Riley, J. P., Rosano, G. M. C., Ruilope, L. M., Ruschitzka, F., Rutten, F. H., and van der Meer, P. (2016). 2016 esc guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal*, 37(27):2129–2200.

Price, D. L. and Manatunga, A. K. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine*, 20(9–10):1515–1527.

Prince, T., Bommert, A., Rahnenführer, J., and Schmid, M. (2025). On the estimation of inverse-probability-of-censoring weights for the evaluation of survival prediction error. *PloS one*, 20(1):e0318349.

Puchner, S. E., Funovics, P. T., Böhler, C., Kaider, A., Stihsen, C., Hobusch, G. M., Panotopoulos, J., and Windhager, R. (2017). Oncological and surgical outcome after treatment of pelvic sarcomas. *PloS one*, 12(2):e0172203.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roger, V. L. (2013). Epidemiology of heart failure. *Circulation Research*, 113(6):646–659.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69:239–241.

Smith, G. L., Lichtman, J. H., Bracken, M. B., Shlipak, M. G., Phillips, C. O., DiCapua, P., and Krumholz, H. M. (2006). Renal impairment and outcomes in heart failure: Systematic review and meta-analysis. *Journal of the American College of Cardiology*, 47(10):1987–1996.

Sobreiro, P., Alonso, J. G., Martinho, D., and Berrocal, J. (2022). Hybrid random forest survival model to predict customer membership dropout. *Electronics*, 11:3328.

Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, 2 edition.

Su, Y.-S. and Yajima, M. (2024). *R2jags: Using R to Run 'JAGS'*. R package version 0.8-9.

Tang, M., Bolderson, E., O'Byrne, K. J., and Richard, D. J. (2021). Tumor hypoxia drives genomic instability. *Frontiers in cell and developmental biology*, 9:626229.

Tsodikov, A. D., Ibrahim, J. G., and Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98(464):1063–1078.

Vahidpour, M. (2016). *Cure rate models*. Ecole Polytechnique, Montreal (Canada).

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, 44(12):1731–1742.

Valletta, J. J., Torney, C., Kings, M., Thornton, A., and Madden, J. (2017). Applications of machine learning in animal behaviour studies. *Animal Behaviour*, 124:203–220.

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009.

Wienke, A., Lichtenstein, P., and Yashin, A. I. (2003). A bivariate frailty model with a cure fraction for modeling familial correlations in diseases. *Biometrics*, 59(4):1178–1183.

Wienke, A., Locatelli, I., and Yashin, A. I. (2006). The modelling of a cure fraction in bivariate time-to-event data. *Austrian Journal of Statistics*, 35(1):67–76.

Xu, J., Huang, L., and Li, J. (2016). Dna aneuploidy and breast cancer: a meta-analysis of 141,163 cases. oncotarget. 2016; 7: 60218–29.

Yancy, C. W., Jessup, M., Bozkurt, B., Butler, J., Casey, D. E., Drazner, M. H., Fonarow, G. C., Geraci, S. A., Horwich, T., Januzzi, J. L., Johnson, M. R., Kasper, E. K., Levy, W. C., Masoudi, F. A., McBride, P. E., McMurray, J. J. V., Mitchell, J. E., Peterson, P. N., Riegel, B., Sam, F., Stevenson, L. W., Tang, W. H. W., Tsai, E. J., and Wilkoff, B. L. (2013). 2013 accf/aha guideline for the management of heart failure. *Journal of the American College of Cardiology*, 62(16):e147–e239.

Zeng, D. and Lin, D. (2007). Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association*, 102(477):167–180.

Zhang, X., Liu, G., and Peng, X. (2023). A random forest model for post-treatment survival prediction in patients with non-squamous cell carcinoma of the head and neck. *Journal of Clinical Medicine*, 12:5015.

APÊNDICE A: ATIVIDADES DOCENTE

Durante o período do pós-doutorado, o pesquisador também exerceu atividades de ensino no curso de Bacharelado em Estatística da Faculdade de Ciências e Tecnologia da UNESP (FCT/UNESP) – Campus Presidente Prudente. Foram ministradas 20 aulas na disciplina *Fundamentos da Matemática*, contribuindo diretamente para a formação básica dos discentes em conteúdos matemáticos essenciais à graduação em Estatística. A disciplina abrange tópicos fundamentais, incluindo propriedades numéricas e manipulação algébrica, funções reais (lineares, afins, quadráticas e polinomiais), equações e inequações, funções exponenciais e logarítmicas, aplicações e modelagem exponencial, além de conteúdos de trigonometria e progressões.

No âmbito das aulas ministradas, o pesquisador atuou especificamente nos conteúdos de Trigonometria e Progressões. Em Trigonometria, foram abordados desde os conceitos elementares — arcos, ângulos, unidades de medida e ciclo trigonométrico — até o estudo completo das funções trigonométricas (seno, cosseno, tangente, cotangente, secante e cossecante), incluindo relações fundamentais, identidades trigonométricas e transformações. Foram desenvolvidas aplicações em triângulos retângulos e exercícios voltados à consolidação das propriedades algébricas e geométricas das funções circulares.

No tópico de Progressões Aritméticas (PA) e Progressões Geométricas (PG), foram trabalhadas definições formais, dedução e interpretação das fórmulas do termo geral, soma de termos finitos e, no caso das PGs, soma infinita sob condição de convergência. Problemas aplicados foram discutidos para conectar os conceitos formais à resolução de situações práticas e ao raciocínio quantitativo.

APÊNDICE B: OUTROS ARTIGOS DESENVOLVIDOS

4.1 ARTIGO 1

Título: Nonlinear Semiparametric Modeling of Lifetime Data Using Polynomial Approximations for Hazard Functions.

Resumo: This work introduces a novel semiparametric, polynomial-based model designed to effectively capture complex hazard patterns, including bathtub-shaped behaviors and other nonstandard variations, through polynomial approximations of the hazard function. Parameter estimation is performed via a least-squares approach, while LASSO regularization is employed to select the most relevant polynomial terms and mitigate overfitting. The proposed model is applied to real-world engineering datasets, and the Kaplan–Meier estimator is used to obtain the empirical cumulative hazard functions in a nonparametric manner. The empirical results highlight the adaptability and robustness of the proposed methodology, which successfully captures the three canonical phases in reliability analysis—early failure, useful life, and wear-out—while accurately modeling complex hazard dynamics.

Status: O artigo consolida os resultados finais obtidos no âmbito do projeto. Ressalta-se, contudo, que o manuscrito ainda se encontra em fase de finalização, passando por ajustes estruturais e refinamentos textuais antes de sua submissão.

4.2 ARTIGO 2

Título: A Hierarchical Bayesian Lagged-Effects Regression Model for Analyzing Case-Fatality Rates (CFR).

Resumo: This study reports the Case-Fatality Rate (CFR) for COVID-19 in six severely impacted countries—Italy, the UK, Brazil, the United States, Germany, and Mexico—in the year 2020. The statistical analysis focused on the death-to-confirmed-case ratio in each country from the onset of the pandemic to September 10, 2020. Given that the CFR variability changed over time due to factors such as testing rates, healthcare capacity, public health interventions, and so on, it wasdopted a hierarchical Bayesian linear regression model with lagged effects assuming normal errors with non-constant variances (heteroscedasticity). Standard time series models, such as ARIMA and Moving Average (MA), were also taken into consideration for comparison purposes. The obtained results showed that the CFR initially had peaks at the beginning of the pandemic and then gradually decreased in each country, with great accurate CFR projections provided by the proposed model.

Submetido no Periódico: Biostatistics & Epidemiology.

4.3 ARTIGO 3

Título: Modeling Dependent (Informative) Censoring in Survival Data: A Bayesian Comparison Through Frailty and Marshall-Olkin Bivariate Models.

Resumo: In medical research, survival times and dropout events often show a significant dependence structure. This dependency is evident in terminally ill patients, whose deteriorating health frequently leads to early withdrawal from clinical trials. Consequently, these dropouts introduce incomplete survival records, and the assumption that censoring is independent of survival outcomes fails, producing biased and inaccurate survival estimates. This work provides a comprehensive Bayesian framework for informative censoring that employs two complementary approaches: (1) parametric bivariate lifetime distributions derived from Marshall-Olkin methods and (2) frailty models. Using the Metropolis-within-Gibbs algorithm, we carried out posterior inference within a hierarchical Bayesian framework employing weakly informative priors. The methodology was applied to the German Breast Cancer Study dataset, which has clinically informative censoring. Our findings show that the Marshall-Olkin models offer more accurate survival estimates at lower computing costs than frailty models. Furthermore, under the informative censoring framework, important clinical indicators such as the use of hormones, tumor grade and nodes were significantly associated to the outcome through the proposed models. In summary, these results emphasize the importance of explicitly modeling dependence structures in survival analysis, which has significant impacts for clinical trials in contexts where informative censoring is a common feature.

Submetido no Periódico: Journal of Applied Statistics.

4.4 ARTIGO 4

Título: A Decision Tree–Based Framework for the Classification of *Peckoltia* Species Using Morphometric Measurements.

Resumo: Accurate species identification plays an important role in ichthyological research based on morphometric data, particularly in taxonomically challenging freshwater genera such as *Peckoltia*. This study evaluates supervised machine learning approaches for species-level discrimination using standardized linear morphometric measurements derived from 292 specimens representing six *Peckoltia* species, each described by 34 traits scaled by standard length. The dataset was divided into independent training (70%) and testing (30%) subsets, and tree-based ensemble methods were implemented within a supervised classification framework. Hyperparameters were optimized via repeated cross-validation, and predictive performance was assessed on the held-out test set using overall accuracy and macro-averaged F_1 score, along with class-wise metrics. Both classifiers showed consistent global performance, although classification effectiveness varied among species, mainly due to differences in sensitivity. Most classification errors occurred between species that share similar morphometric profiles, indicating that misassignments were associated with overlapping measurements rather than random confusion across classes. Permutation-based variable importance analysis identified cranial and caudal measurements, particularly barbel length and caudal peduncle depth, as the primary contributors to discrimination, whereas several trunk-related traits showed limited influence.

Submetido no Periódico: Journal of Fish Biology.

4.5 ARTIGO 5

Título: Modeling Incidence, Mortality Rates, and Patient Survival in Yellow Fever Cases in Brazil (1995 – 2023).

Resumo: This paper reports a broad study of the incidence, mortality rates, and patient survival outcomes of yellow fever in Brazil from 1995 to 2023, utilizing epidemic-related data obtained from the Brazilian Ministry of Health and the General Coordination of Arbovirus Surveillance (CGARB). For the analysis, we employed a range of statistical approaches, including Cox proportional hazards (PH) models, cure rate models, and logistic regression models, to identify factors affecting the disease's behavior. These factors include municipality-level case counts, the population density, the Municipal Human Development Index (MHDI), death rates, age, gender, patient survival outcomes, among others. The findings suggested that yellow fever incidence, mortality, and survival outcomes are highly influenced by population density, MHDI, age, and gender. Specifically, older age and male gender were associated to higher mortality risks and shorter survival times, whereas higher population density and lower MHDI were associated to higher case counts. These results, in particular, emphasize the need for focused efforts in high-risk areas and among vulnerable populations, such as enhancing vaccination campaigns.

Submetido no Periódico: Revista Colombiana de Estadística.

4.6 ARTIGO 6

Título: Analysis of the Ammonia Nitrogen Dynamics in Washington State Rivers: An Approach using Tobit. Model

Resumo: This study introduces a new parametric mixture model, named Tobit-Weibull cure rate model, based on Tobit model, to analyze environmental data under a left-censoring schema. A hierarchical Bayesian approach was adopted to estimate the model's parameters, assuming non-informative prior distributions. An extensive simulation study was carried out to evaluate the performance of the Bayesian estimators assuming different censoring scenarios. The proposed model usefulness is archived through the analysis of ammonia nitrogen concentrations in rivers across Washington State, using public data collected between 2011 and 2021 by the Washington Department of Ecology's monitoring program. The dataset includes measurements of water quality parameters such as pH, bacteria, phosphorus, ammonia nitrogen, dissolved oxygen, temperature, conductivity and so on. The results showed that the proposed model captures the variability in ammonia nitrogen concentrations, and identifies the water quality parameters that affects those concentrations, establishing a robust framework for environmental risk assessment.

Submetido no Periódico: Annals of Data Science.

COMPROVANTES E DECLARAÇÕES

ARTIGOS NA TEMÁTICA DO PROJETO

Statistical Methods in Medical Research

Extending the Cox Proportional Hazards Model with a Bayesian Semiparametric Cumulative Hazard Transformation Mixture Cure Model for Long-Term Survival Estimation

Journal:	*Statistical Methods in Medical Research*
Manuscript ID	SMM-25-0748
Manuscript Type:	Original Research Article
Keywords:	Bayesian inference, breast cancer survival, long-term survival analysis, cox proportional hazards model, semiparametric cumulative hazard transformation models
Abstract:	In this paper, we propose an extension of the standard Cox proportional hazards model through a semiparametric mixture cure framework based on cumulative hazard transformation models, aiming to accommodate long-term survival outcomes and to estimate the plateau observed in survival curves. Our approach effectively distinguishes between individuals who remain at risk and those who achieve long-term remission or cure, thereby addressing a fundamental limitation of the classical Cox model, which presumes that all individuals remain perpetually at risk of the event. Specifically, within our modeling framework, the unknown cumulative hazard function is treated as a latent component and is estimated under a Bayesian paradigm, incorporating prior information derived from standard Cox model estimates. To illustrate the usefulness of the proposed methodology, we employed it to a retrospective cohort comprising 295 cases of primary invasive breast carcinoma from the Netherlands Cancer Institute (NKI). The findings indicate that the proposed semiparametric mixture cure model offers a more accurate representation of survival probabilities by appropriately capturing the heterogeneity between cured individuals and those still at risk.

Submission Confirmation

Print

Thank you for your submission

Submitted to
Statistical Methods in Medical Research

Manuscript ID
SMM-26-0131

Title
Random Survival Forests for Survival Prediction in Heart Failure: External Validation and Predictor Importance

Authors
Rodrigues, Naiara
Puziol de Oliveira, Ricardo
Neto, João
Achcar, Jorge

Date Submitted
26-Feb-2026

Author Dashboard

OUTROS ARTIGOS COM INTERSEÇÃO NA TEMÁTICA DO PROJETO

Gmail Ricardo Puziol <rpuziol.oliveira@gmail.com>

Submission received for Biostatistics & Epidemiology (Submission ID: 250324579)
1 mensagem

TBEP-peerreview@journals.tandf.co.uk <TBEP-peerreview@journals.tandf.co.uk> 17 de outubro de 2025 às 17:43
Para: rpuziol.oliveira@gmail.com

Taylor & Francis
Taylor & Francis Group

Dear Ricardo Oliveira,

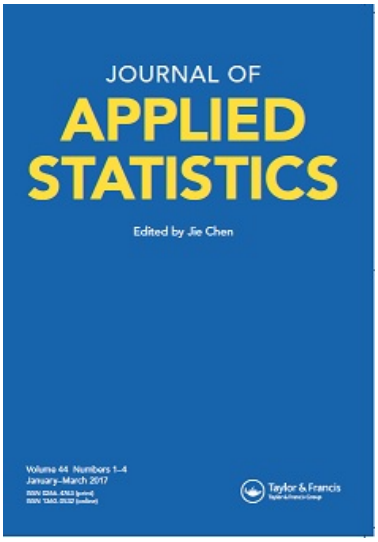
Thank you for your submission.

Submission ID	**250324579**
Manuscript Title	**The Use of a Hierarchical Bayesian Lagged Effects Regression model for COVID-19 Case-Fatality Rates (CFR)**
Journal	**Biostatistics & Epidemiology**

If you made the submission, you can check its progress and make any requested revisions on the Author Portal

Thank you for submitting your work to our journal.
If you have any queries, please get in touch with TBEP-peerreview@journals.tandf.co.uk.

Kind Regards,
Biostatistics & Epidemiology Editorial Office



Modeling Dependent (Informative) Censoring in Survival Data: A Bayesian Comparison Through Frailty and Marshall-Olkin Bivariate Models

Journal:	*Journal of Applied Statistics*
Manuscript ID	CJAS-2025-1061
Manuscript Type:	Research Article
Date Submitted by the Author:	13-Nov-2025
Complete List of Authors:	Oliveira, Ricardo; Universidade Estadual Paulista "Júlio de Mesquita Filho" Brazil, Statistics Achcar, Jorge; University of São Paulo Brazil, Social Medicine Debastiani Neto, João; Universidade Estadual de Maringa, Ciencias Martinez, Edson; Universidade de Sao Paulo, Social Medicine de Oliveira Peres, Marcos; Universidade Estadual de Maringa, Statistics Aguilar, Guilherme Aparecido Santos; Universidade Estadual Paulista "Júlio de Mesquista Filho" Brazil, Statistics Fernandes, Maria Graziela da Silva; Universidade Estadual de Londrina, PGMAC/UEL
Keywords:	Informative censoring, Bayesian survival analysis, frailty models, Marshall-Olkin method, Markov Chains Monte Carlo methods, oncology studies
Maths:	62H10, 62N02

JOURNAL OF FISH BIOLOGY fsbi

Regular Article

A Decision Tree–Based Framework for the Classification of Peckoltia Species Using Morphometric Measurements

Submission ID 5cd994e7-1b2c-4251-a945-fbda2a71340a

Submission Version Initial Submission

PDF Generation 02 Mar 2026 07:43:47 EST by Atypon ReX

Authors

Dr. Ricardo Puziol de Oliveira

Affiliations

- Department of Statistics, Faculty of Science and Technology, State University of São Paulo, Brazil

Dr. João Debastiani Neto
Corresponding Author
Submitting Author

ORCiD
https://orcid.org/0000-0003-4402-1682

Affiliations

- Department of Statistics, Faculty of Science and Technology, State University of São Paulo, Brazil
- Department of Science, Maringá State University, Brazil


Revista Colombiana de Estadística
Submissions
My Queue 1
Archived 1
Help
My Assigned
Search
Filters
New Submission
120297 Alberto Achcar et al.
Modeling Incidence, Mortality Rates, and Patient Survival in Yellow Fever Cases in Brazil (1995 – 2023)
0/3
Review
View


Revista Colombiana de Estadística
← Back to Submissions
120297 / Alberto Achcar et al. / Modeling Incidence, Mortality Rates, and Patient Survival in Yellow Fever Cases in Brazil (1995 – 2023)
Library
Workflow
Publication
Submission
Review
Copyediting
Production
Submission Files
Search
726658 Yellow_Fever_Article.pdf
May 9, 2025
Article Text
Download All Files

Annals of Data Science

Analysis of the Ammonia Nitrogen Dynamics in Washington State Rivers: An Approach using Tobit Model

--Manuscript Draft--

Manuscript Number:	AODS-D-25-00750
Full Title:	Analysis of the Ammonia Nitrogen Dynamics in Washington State Rivers: An Approach using Tobit Model
Article Type:	Original Article
Funding Information:	
Abstract:	This study introduces a new parametric mixture model, named Tobit-Weibull cure rate model, based on Tobit model, to analyze environmental data under a left-censoring schema. A hierarchical Bayesian approach was adopted to estimate the model's parameters, assuming non-informative prior distributions. An extensive simulation study was carried out to evaluate the performance of the Bayesian estimators assuming different censoring scenarios. The proposed model usefulness is archived through the analysis of ammonia nitrogen concentrations in rivers across Washington State, using public data collected between 2011 and 2021 by the Washington Department of Ecology's monitoring program. The dataset includes measurements of water quality parameters such as pH, bacteria, phosphorus, ammonia nitrogen, dissolved oxygen, temperature, conductivity and so on. The results showed that the proposed model captures the variability in ammonia nitrogen concentrations, and identifies the water quality parameters that affects those concentrations, establishing a robust framework for environmental risk assessment.
Corresponding Author:	Ricardo Puziol de Oliveira Universidade Estadual Paulista: Universidade Estadual Paulista Julio de Mesquita Filho BRAZIL
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universidade Estadual Paulista: Universidade Estadual Paulista Julio de Mesquita Filho
Corresponding Author's Secondary Institution:	
First Author:	Danielle Peralta
First Author Secondary Information:	
Order of Authors:	Danielle Peralta Ricardo Puziol de Oliveira Jorge Alberto Achcar João Debastiani Neto Josmar Mazuchelli
Order of Authors Secondary Information:	
Author Comments:	Cover Letter Dear Editor-in-Chief, On behalf of my co-authors, I am pleased to submit our manuscript entitled "Analysis of the Ammonia Nitrogen Dynamics in Washington State Rivers: An Approach Using Tobit Model" for consideration for publication in Annals of Data Science. In this study, we propose a new parametric mixture model, named the Tobit–Weibull cure rate model, designed to analyze environmental data under left-censoring

XIV ERMAC
Regional 11

Certificado de Participação em Evento

Declaramos, para os devidos fins, que **João Debastiani Neto**, apresentou o trabalho intitulado **Semiparametric Transformation Models with Cure Fraction: Extensions of the Cox Model under a Bayesian Approach**, na forma de pôster, no evento *XIV ERMAC- Encontro Regional de Matemática Aplicada e Computacional*, que foi realizado de 05 a 07 de maio de 2025, na Universidade Estadual de Maringá.

Maringá, 11 de maio de 2025.

Emerson Vitor Castelani
Coordenador do evento

69ª Reunião Anual da RBras
21° SEAGRO

Estatística e Ciência de Dados
a serviço do Desenvolvimento Sustentável

Vitória - ES

4 a 8 de agosto de 2025

CERTIFICADO DE APRESENTAÇÃO

Certificamos que o trabalho **Extended Semiparametric Cox Model: A Cure Fraction Approach for Long-Term Survival**, de autoria de **João Debastiani Neto, Jorge Alberto Achcar, Ricardo Puziol de Oliveira, Maria Graziela da Silva Fernandes**, foi apresentado na categoria pôster na **69ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria (RBras)** e no **21° Simpósio de Estatística Aplicada à Experimentação Agronômica (SEAGRO)**, realizados em Vitória/ES, no período de 4 a 8 de agosto de 2025.

Vitória/ES, 08 de agosto de 2025

Código de autenticação:LASbct6FUo6ViCRhJ37tAWzFeFMdNh

Diogo Rossoni
Presidente da RBras

Agatha Rodrigues
Coordenadora da Comissão Organizadora Local

69ª Reunião Anual da RBras
21° SEAGRO

Estatística e Ciência de Dados
a serviço do Desenvolvimento Sustentável

Vitória - ES

4 a 8 de agosto de 2025

CERTIFICADO DE APRESENTAÇÃO

Certificamos que o trabalho **Modelagem Semiparamétrica de Dados de Confiabilidade Usando Aproximações Polinomiais de Funções de Risco**, de autoria de **Ricardo Puziol de Oliveira, Jorge Alberto Achcar, João Debastiani Neto, Maria Graziela da Silva Fernandes**, foi apresentado na categoria pôster na **69ª Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria (RBras)** e no **21° Simpósio de Estatística Aplicada à Experimentação Agronômica (SEAGRO)**, realizados em Vitória/ES, no período de 4 a 8 de agosto de 2025.

Vitória/ES, 08 de agosto de 2025

Código de autenticação:800HVQb4uosEoC19YrcCj5xa3RpfwO

Diogo Rossoni
Presidente da RBras

Agatha Rodrigues
Coordenadora da Comissão Organizadora Local


XIX EMR
Escola de Modelos de Regressão
XIX ESCOLA DE MODELOS DE REGRESSÃO
Certificamos que o trabalho intitulado " A New Semiparametric Regression Framework with Lagged Effects for Analyzing Epidemiology Non-Linear Data" de autoria de "Ricardo Puziol de Oliveira, Maria Luisa Scherer Belini, João Debastiani Neto, Jorge Alberto Achcar" foi apresentado na modalidade pôster na XIX Escola de Modelos de Regressão (XIX EMR), ocorrida em João Pessoa (PB), no período de 29 a 31 de Outubro de 2025.
Código de autenticação:bQAjDdti4ZSYADiaYHxvuQLkN9SJQA
João Pessoa, 31 de Outubro de 2025
Hemílio Fernandes Campos Coelho
Presidente da Comissão Organizadora
Flávio Ziegelmann
Presidente da Associação Brasileira de Estatística
UFPB
ABE
UEPB

AVALIADOR DE PESQUISAS EM EVENTOS CIENTÍFICOS

Universidade de São Paulo
Brasil

33º SIICUSP - Simpósio Internacional de Iniciação Científica e Tecnológica da Universidade de São Paulo

Declaração de Participação

Declaro para os devidos fins que **João Debastiani Neto** participou como Avaliador(a) do **33º** Simpósio Internacional de Iniciação Científica e Tecnológica da USP - SIICUSP

Prof. Dr. Paulo Alberto Nussenzveig
Pró Reitor de Pesquisa e Inovação
Universidade de São Paulo

Documento emitido às **10:09:07** horas do dia **15/10/2025** (hora e data de Brasilia).

Código de controle: **NRXA-NVK7-VUNB-DL9M**

A autenticidade deste documento pode ser verificada na página da Universidade de São Paulo

http://uspdigital.usp.br/webdoc


31ª edição
XXXI PRÊMIO ROCHA LIMA
Certificamos que João Debastiani Neto participou como avaliador de trabalhos científicos, apresentados na modalidade "Submissão de Resumos", na 1ª fase da 31ª Edição do Prêmio Rocha Lima (PRL), realizada no período de 1 a 22 de setembro de 2025, pelo Departamento Científico.
O Departamento Científico é vinculado ao Centro Acadêmico Rocha Lima (CARL), entidade representativa dos estudantes de medicina da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo.
Ribeirão Preto, 10 de novembro de 2025.
Pedro Mussi Perroni
Diretor do Departamento Científico
Daniel Freitas dos Santos
Presidente do Centro Acadêmico Rocha Lima
Isabella Pacheco Ferreira
Secretária do Centro Acadêmico Rocha Lima
Este certificado está registrado no Centro Acadêmico "Rocha Lima" como: COMA 2502297, Folha 39, Livro 05.

XXXVII CIC Unesp
Congresso de Iniciação Científica e Tecnológica da Unesp
"Ciência e liberdade de pensamento em uma época de extremos"

CERTIFICADO

Certificamos que **João Debastiani Neto**, participou na qualidade de **AVALIADOR**, do evento **XXXVII Congresso de Iniciação Científica da Unesp - FCT/Presidente Prudente**, avaliando nas Modalidade(s) **Ciências Exatas e da Terra** e Área(s) Temática(s) **Ciências Exatas e da Terra - Probabilidade e Estatística**.

Presidente Prudente, 07 de outubro de 2025.

Edson Cocchieri Botelho
Pró-Reitor de Pesquisa
Presidente do CIC

Reginaldo Barboza da Silva
Coordenador de Iniciação Científica
Vice-Presidente do CIC

unesp.br/xxxviiicic

Grupo de pesquisa

Grupo de Pesquisa em Estatística Computacional e Aprendizado de Máquina

Endereço para acessar este espelho: dgp.cnpq.br/dgp/espelhogrupo/2635888148428276

Identificação

Situação do grupo:	Certificado
Ano de formação:	2025
Data da Situação:	04/08/2025 09:46
Data do último envio:	16/09/2025 10:07
Líder(es) do grupo:	Ricardo Puziol de Oliveira
	João Debastiani Neto
Área predominante:	Ciências Exatas e da Terra; Probabilidade e Estatística
Instituição do grupo:	Universidade Estadual Paulista Júlio de Mesquita Filho - UNESP
Unidade:	Faculdade de Ciências e Tecnologia de Presidente Prudente

Endereço / Contato

Endereço

Logradouro:	Rua Roberto Simonsen 305 Universidade Estadual Paulis
Número:	305
Complemento:	
Bairro:	Vila Santa Helena
UF:	SP
Localidade:	Presidente Prudente

Grupo de pesquisa

GRUPO DE PESQUISA EM ESTATÍSTICA E INFERÊNCIA BAYESIANA

Endereço para acessar este espelho: dgp.cnpq.br/dgp/espelhogrupo/9579426422776268

Identificação

Situação do grupo: Certificado

Ano de formação: 2011

Data da Situação: 09/09/2013 15:53

Data do último envio: 18/09/2025 20:43

Líder(es) do grupo: Fernando Antonio Moala

Guilherme Aparecido Santos Aguilar

Área predominante: Ciências Exatas e da Terra; Probabilidade e Estatística

Instituição do grupo: Universidade Estadual Paulista Júlio de Mesquita Filho - UNESP

Unidade: Campus Presidente Prudente

Endereço / Contato

Endereço

Logradouro: Universidade Estadual Paulista - Campus Presidente Pr

Número: 305

Complemento:

Bairro: Vila Santa Helena

UF: SP

Localidade: Presidente Prudente

Linhas de pesquisa

Nome da linha de pesquisa	Quantidade de Estudantes	Quantidade de Pesquisadores
Processos Não-Gaussianos	0	3
Análise de Regressão e Modelos Lineares Generalizados	0	4
Análise de Sobrevivência e Confiabilidade	2	8
Aprendizado de máquina	0	1
Elicitação	1	1
Inferência Bayesiana Não-Paramétrica	0	1
Inferência Bayesina	2	6
Modelagem Estatística em Hidrologia, Climatologia e Ambiental	1	3

Recursos humanos

Pesquisadores	Titulação máxima	Data inclusão
Carlos Aparecido dos Santos	Doutorado	18/10/2016
Clovis Augusto Niiyama	Mestrado	Anterior a abril de 2014
Fernando Antonio Moala	Doutorado	Anterior a abril de 2014
Guilherme Aparecido Santos Aguilar	Doutorado	23/10/2017
João Debastiani Neto	Doutorado	11/06/2025
Jorge Alberto Achcar	Doutorado	Anterior a abril de 2014
Marcelo Hartmann	Doutorado	11/12/2019
Pedro Luiz Ramos	Doutorado	Anterior a abril de 2014
Ricardo Puziol de Oliveira	Doutorado	09/12/2021
Rick Anderson Freire Mangueira	Doutorado	18/09/2025
Sérgio Minoru Oikawa	Doutorado	Anterior a abril de 2014

PLANO DE ATIVIDADES DE ENSINO

Disciplina: Fundamentos de Matemática
Curso: Estatística
Carga horária: 60 horas
Professores Responsáveis: José Gilberto Spasiani Rinaldi e Ricardo Puziol de Oliveira
Pós-Doutorando: João Debastiani Neto
Período Letivo: 1º semestre de 2025.
Modalidade: Presencial com atividades semipresenciais (até 20% da carga horária)
Carga horária a ser ministrada pelo pós-doutorando: 20 horas.

OBJETIVOS

- Apresentar domínio dos conteúdos matemáticos presentes na educação básica, de modo a prosseguir no aprendizado de novos conceitos;
- Comunicar-se matematicamente e expressar-se com clareza, precisão e objetividade;
- Compreender matemática, para estabelecer relações com outras áreas do conhecimento e utilizar os conhecimentos na compreensão do mundo que o cerca;
- Integrar os diversos conteúdos e utilizá-los na resolução de problemas.

CONTEÚDO PROGRAMÁTICO

- Operações fundamentais com números: propriedades numéricas, expressões matemáticas e simplificação de expressões.
- Funções de uma variável real: linear, afim, quadrática, polinomial. Equações e Inequações.
- Funções exponenciais e logarítmicas. Potências e raízes. Funções exponenciais. A função exponencial de base e. Equações e inequações exponenciais. Aplicações das funções exponenciais. Modelos exponenciais. Logaritmos: origem, conceito, nomenclatura, propriedades. Funções logarítmicas. Equações e inequações logarítmicas. Logaritmo decimal. Logaritmo natural. Aplicações dos logaritmos.
- Funções trigonométricas. Trigonometria no triângulo retângulo. Conceitos trigonométricos básicos: arcos e ângulos, unidades, ciclo trigonométrico, arcos congruentes, quadrantes. Funções circulares: seno, cosseno, tangente, cotangente, secante e cossecante. Relações trigonométricas. Transformações trigonométricas. Aplicações.
- Progressões aritméticas e geométricas. Sequências. Progressões aritméticas: definição, fórmula para o termo geral, soma de uma PA finita. Progressões geométricas: definição, fórmula para o termo geral, soma dos n primeiros termos de uma PG finita; soma dos termos de uma PG infinita. Problemas envolvendo PA e PG

METODOLOGIA

- Aulas expositivas.
- Resolução de situações-problemas.
- Horários de atendimento extra-classe.
- Poderão ser programadas atividades na modalidade semipresencial, até o limite máximo de 20% da carga horária, conforme prevê a Portaria 4059 do MEC, de 10/12/2024. As atividades semipresenciais poderão ser realizadas por meio de trabalhos práticos e estudos dirigidos.

CRITÉRIOS DE AVALIAÇÃO DE APRENDIZAGEM

- *Resolução Unesp nº 106/2012, alterada pelas Resoluções nº 23/2013 e 75/2016 (notadamente quanto à recuperação)*
- Média ponderada das avaliações (provas, listas de exercícios e trabalhos). Será considerado aprovado o aluno que atingir média do período regular maior ou igual a 5,0 e cumprir com a frequência mínima exigida.

RECURSOS DIDÁTICOS

- Quadro branco, projetor multimídia.
- Softwares: GeoGebra e Microsoft Excel.
- Plataforma de apoio (Google Classroom).
- Calculadoras científicas, material de apoio visual.

OBSERVAÇÕES

- O cronograma poderá sofrer ajustes conforme o desenvolvimento da turma e calendário acadêmico.

Documento assinado digitalmente
gov.br
JOAO DEBASTIANI NETO
Data: 02/03/2026 16:23:47-0300
Verifique em https://validar.iti.gov.br

João Debastiani Neto
Pós-Doutorando

Documento assinado digitalmente
gov.br
RICARDO PUZIOL DE OLIVEIRA
Data: 02/03/2026 15:56:43-0300
Verifique em https://validar.iti.gov.br

Ricardo Puziol de Oliveira
Professor Supervisor

Documento assinado digitalmente
gov.br
RICARDO PUZIOL DE OLIVEIRA
Data: 02/03/2026 15:54:57-0300
Verifique em https://validar.iti.gov.br

Ricardo Puziol de Oliveira
Docente da disciplina de
Fundamentos da Matemática

Documento assinado digitalmente
gov.br
JOSE GILBERTO SPASIANI RINALDI
Data: 02/03/2026 15:46:11-0300
Verifique em https://validar.iti.gov.br

José Gilberto Spasiani Rinaldi
Docente da disciplina de
Fundamentos da Matemática

Documento assinado digitalmente
gov.br
EDILSON FERREIRA FLORES
Data: 02/03/2026 11:09:23-0300
Verifique em https://validar.iti.gov.br

Edilson Ferreira Flores
Chefe do Departamento de Estatística