

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS

Nathan Silva Bossi

**Análise de Sobrevivência aplicada ao estudo de ocorrências de sinistros no
contexto da Proteção Veicular**

Belo Horizonte

2023

Nathan Silva Bossi

Análise de Sobrevivência aplicada ao estudo de ocorrências de sinistros no contexto da Proteção Veicular

Monografia apresentada por Nathan Silva Bossi à UFMG, como um dos requisitos para conclusão do Curso de Bacharel em Ciências Atuariais.

Orientador: Prof. Dr. Enrico Antônio Colosimo

Belo Horizonte

2023

Resumo

No mercado de Seguradoras e Associações de Proteção Veicular, os eventos de Perda Total representam grande impacto financeiro para uma entidade de gestão de risco. O presente trabalho pretende desenvolver algumas metodologias que sirvam a propósitos atuariais de mensurar e gerir o risco deste tipo de evento. Para tanto, foram usadas técnicas de Análise de Sobrevivência, frequentemente usadas em estudos estatísticos relacionados ao tempo até ocorrência de eventos diversos: desde a durabilidade de componentes eletrônicos até a longevidade de empresas. No estudo da distribuição da variável aleatória “tempo até ocorrência de Perda Total”, foram construídos indicadores de frequência, formulados métodos de cálculo de Provisão usando a metodologia de Seguros de Vida e mensurados os efeitos de algumas variáveis de interesse. A tentativa de construção de uma Esperança de Vida Condicional não obteve êxito, por não considerar processos comuns a veículos automotivos, que os tornam inutilizáveis em um período de tempo já reconhecido e documentado.

Sumário

INTRODUÇÃO.....	5
1.1 Apresentação.....	5
1.2 Objetivos.....	6
1.2.1 Objetivo geral.....	6
1.2.2 Objetivos específicos.....	6
1.3 Definição de conceitos.....	6
1.3.1 Associações de Proteção Veicular.....	6
1.3.2 Sinistro, Falha e Perda Total.....	7
1.3.3 Análise de Sobrevivência.....	7
1.3.4 Censura.....	7
1.3.4 Tempo de Contrato.....	7
1.4 Base de Dados.....	8
METODOLOGIA ESTATÍSTICA E ATUARIAL.....	8
RESULTADOS.....	15
3.1 Análise Descritiva.....	15
3.1.1 Tipo Veículo.....	16
3.1.2 Ano Fabricação (Idade Veículo).....	17
3.1.3 Cor do Veículo.....	18
3.1.4 Valor Fipe.....	19
3.1.5 Estado/ macrorregião.....	20
3.1.6 Histórico de Perdas Parciais.....	22
3.1.7 Estado Civil.....	22
3.1.8 Idade do Associado.....	23
3.1.9 Sexo do Associado.....	24
3.1.10 Melhorias para a base de dados.....	25
3.2 Modelagem Não Paramétrica.....	26
3.3 Modelagem Paramétrica.....	30
3.3.1 Escolhendo o modelo.....	30
3.3.2 Tempo estimado para probabilidades de Perda Total.....	31

3.3.3 Valor Esperado de custos com Perda Total.....	32
3.3.4 Construção do modelo completo.....	33
3.3.5 Analisando impacto da natureza do evento.....	36
CONSIDERAÇÕES FINAIS.....	39
REFERÊNCIAS.....	40
APÊNDICE A – CURVAS KAPLAN-MEIER.....	41
APÊNDICE B – ESPERANÇA CONDICIONAL.....	43
APÊNDICE C – TEMPO ESTIMADO PARA PROBABILIDADES DE PERDA TOTAL COM COVARIÁVEIS.....	46

1. INTRODUÇÃO

1.1 Apresentação

A mensuração do risco em eventos aleatórios sempre foi um desafio na história humana. Com o tempo, foram sendo desenvolvidos métodos cada vez mais sofisticados para expressar incertezas e sistematizar “crenças” acerca de um evento aleatório. A “lei de evidências” foi a primeira sistematização de que se tem conhecimento, desenvolvida na Idade Média, como uma classificação dos graus de provas para lidar com as incertezas de evidências em tribunais. Já na Renascença, há registros de apostas que foram discutidas em termos rudimentares de probabilidade. Neste mesmo período histórico, mais especificamente no século XIV, os primeiros contratos de seguro foram criados com base em um percentual do valor de um bem e em uma estimativa intuitiva dos riscos de dano ao mesmo. Mas não havia nenhuma teoria sobre a forma de calcular probabilidades de sinistro e prêmios (*Franklin, James, 2001, “The Science of Conjecture: Evidence and Probability Before Pascal”*). Apesar de evidências históricas indicarem o uso de técnicas de dissolução do risco e mutualismo desde a antiga civilização Babilônica (Vaughan, E. J., 1997, *Risk Management*), foi apenas no século XVIII que se desenvolveram métodos matemáticos para a estimativa do risco em seguros, com a criação de indicadores como a Taxa de Mortalidade em seguros de vida (*“Today and History: The History of Equitable Life”, 2009, Retrieved*). Neste mesmo século, desenvolveram-se trabalhos de matemáticos como o Jacob Bernoulli, *Ars Conjectandi* (póstuma, 1713), que foram de grande importância para o desenvolvimento de ferramentas no cálculo de probabilidade e que, por sua vez, contribuiu com o aprimoramento do ramo de seguros. Hoje, as seguradoras de veículos automotivos utilizam diversas técnicas dos campos de estatística e de ciência de dados para um melhor gerenciamento de risco. Este trabalho busca contribuir com este desenvolvimento metodológico, propondo uma aplicação das técnicas de Análise de Sobrevivência na mensuração do risco de Perda Total para veículos automotivos. A Perda Total é um evento que representa grande impacto financeiro para entidades de gestão de risco e, portanto, deve ter uma análise própria. Pretende-se que a metodologia desenvolvida possa ser útil a seguradoras e demais entidades de gestão de risco do ramo automotivo.

1.2 Objetivos

1.2.1 Objetivo geral

Usar técnicas de análise de sobrevivência na criação de modelos preditivos para a variável tempo de vida de um automóvel. Tempo este, cujo ponto de partida é o momento do contrato, que perdura até a ocorrência da perda total.

1.2.2 Objetivos específicos

Encontrar um modelo probabilístico que explique a relação entre as variáveis explicativas presentes no banco como tipo de veículo, modelo e ano de fabricação com a variável preditora tempo médio até um sinistro de perda total.

Criar tabelas com estimativas de esperança de vida condicionadas ao tempo de contrato, inspirando-se nas tabelas atuariais usadas na Previdência, para fins de gestão de risco.

Criar uma tabela com probabilidades acumuladas de Perda Total associadas a tempos de sobrevivência, para fins de gestão de risco.

Comparar a metodologia de cálculo de provisões para o Ramo Não-Vida de seguros com a metodologia usada em seguros do Ramo Vida, usando o resultado obtido de modelos paramétricos com a distribuição selecionada do tempo de sobrevivência.

1.3 Definição de conceitos

1.3.1 Associações de Proteção Veicular

Associações de Proteção Veicular (APVs) são entidades de gestão de risco que, bem como seguradoras, se orientam pelos princípios de mutualismo e dissolução do risco. Tais entidades, diferentemente de empresas de seguros, não possuem fins lucrativos, distribuindo as despesas com sinistros (eventos indesejados contemplados por um contrato) entre os associados que fazem parte da entidade. Qualquer receita gerada é destinada ao pagamento de despesas administrativas ou operacionais, reduzindo desta forma o valor pago pelos segurados (que, neste contexto, são chamados associados). Uma gestão eficiente do risco, tal qual em seguradoras, se utiliza de ferramentas estatísticas que são constantemente desenvolvidas e aprimoradas.

1.3.2 Sinistro, Falha e Perda Total

No contexto das APVs e seguradoras de veículos automotivos, sinistro é todo evento indesejado que ocorre com um veículo e cujo custo associado é coberto integral ou parcialmente pela entidade de gestão de risco. Tal evento pode ser classificado como sendo de Perda Parcial, em que o dano pode ser reparado, podendo acontecer mais de uma vez ao veículo, ou Perda Total (PT), que torna o veículo inutilizável. O critério para classificar um evento como sendo de Perda Total é que o custo associado a este seja superior a 75% do valor do veículo (mais especificamente, seu valor FIPE). Este será o evento de interesse, que, no contexto da Análise de Sobrevivência, é denominado “falha”.

1.3.3 Análise de Sobrevivência

A Análise de Sobrevivência é uma área da Estatística que estima ou caracteriza a distribuição do tempo até a ocorrência de um evento de interesse (falha), com base em uma amostra formada por indivíduos que podem ou não passar por este evento. Tradicionalmente utilizada em estudos clínicos da Medicina, na avaliação da eficácia de tratamentos e medicamentos, a Análise de Sobrevivência tem sido cada vez mais aplicada a outros contextos. Uma aplicação atual tem sido na Engenharia, mais especificamente na avaliação da durabilidade dos componentes de equipamentos.

1.3.4 Censura

No contexto da Análise de Sobrevivência, Censura é a interrupção do tempo até a ocorrência de uma falha (no caso deste trabalho, a Perda Total de um veículo seria esta falha). A interrupção que se deve à limitação do período de análise do banco de dados é chamada de censura administrativa. Este é o tipo mais comum, também chamado de censura à direita. A interrupção também pode se dever a um evento não considerado como falha e que encerra o tempo de falha (neste caso, seria a interrupção do contrato do veículo com a APV). Se o veículo não passa pelo evento de falha, ele necessariamente passa por censura.

1.3.4 Tempo de Contrato

No contexto deste trabalho, tempo de contrato será a denominação para o tempo que um veículo já se encontrava na base, em um referencial escolhido na linha do tempo. Mais especificamente, é o tempo vivido sem passar pelo evento de falha de interesse, marcado entre a data de contrato e um momento escolhido como referencial. Também será referido como

“tempo vivido antecipadamente”. Este valor é determinístico para cada veículo, uma vez que já ocorreu, ao contrário da variável aleatória tempo até uma Perda Total.

1.4 Base de Dados

São usados dados primários de uma APV obtidos pela empresa de consultoria Brasil Atuarial, sob condição de sigilo de informações sobre a identidade da APV e de seus associados. A unidade de análise é o veículo de um associado e, sendo uma análise longitudinal, este trabalho irá tratar do conjunto de veículos que começaram seu contrato entre 01 de janeiro de 2018 e 31 de dezembro de 2022. Tal evento é o que demarca a linha de base. Estes veículos podem ter passado pelo evento de falha de interesse, no caso, a Perda Total, ou por censura, caso tenham cancelado o contrato ou passado pelo evento após o final do período de análise. As linhas de vida são os tempos de exposição (em dias) ao risco de sofrer o evento de falha, para cada veículo. Não são consideradas relevantes, portanto, a cronologia nem a ordem das entradas dos veículos na base de associados da APV.

O banco consiste em 23.533 tempos de exposição, que variam de 1 a 1471 dias, dentre os quais 334 se encerraram com um evento de falha (Perda Total). As potenciais variáveis explicativas que se encontram no banco são: Cor do veículo, Tipo de Veículo, Valor Fipe, Estado (agrupado por Macrorregião), número de Perdas Parciais prévias, Ano de Fabricação, Sexo do proprietário/ associado, Estado Civil e a Idade do mesmo. Ainda há uma variável que identifica a natureza do evento (furto, colisão, evento da natureza, capotamento, entre outros).

2. METODOLOGIA ESTATÍSTICA E ATUARIAL

A princípio é feita uma análise descritiva do banco de dados disponível, para visualização e uma melhor orientação das decisões a serem tomadas na criação dos modelos de Análise de Sobrevivência.

Após a análise descritiva dos dados, foi feita uma modelagem não paramétrica Kaplan-Meier a fim de determinar o comportamento da variável “tempo até Perda Total”, considerando os indivíduos que foram censurados em algum momento. Kaplan-Meier é um método que consiste na estimação das probabilidades de sobrevivência $S(t) = P(T > t)$, a partir da divisão do tempo de exposição em intervalos e do cálculo, em cada intervalo, da proporção de expostos que não passaram por falhas (sobreviventes). O estimador é dado pelo produtório

das proporções de sobreviventes em cada intervalo. Sendo não viciado, apresenta a seguinte forma:

$$\hat{S}(t) = \prod_{j: t_j < t} (n_j - d_j) / n_j$$

j = índice identificador de ordem do intervalo

t_j = tempo de uma única falha ou de um conjunto de falhas simultâneas (delimitador dos intervalos)

n_j = número de indivíduos expostos ao risco de falha no tempo t_j . São todos que não passaram por falhas ou censuras até o momento imediatamente anterior a t_j .

d_j = número de falhas em t_j .

A modelagem Kaplan-Meier também foi usada para determinar a significância das covariáveis, usando o teste Log-Rank. O teste consiste em avaliar a significância da diferença entre as curvas de sobrevivência para diferentes categorias de uma mesma potencial covariável. Se diferentes categorias produzem curvas de sobrevivência suficientemente distantes entre si, a variável é significativa para explicar o tempo até o evento de falha. Variáveis numéricas precisam ser categorizadas. A estatística de teste possui distribuição Qui-Quadrado com $r-1$ graus de liberdade, sendo r o número de grupos comparados. Tal estatística possui a seguinte forma:

$$T = \mathbf{v}' \cdot \mathbf{V}^{-1} \cdot \mathbf{v}$$

\mathbf{v} = vetor com números de falhas observados em cada intervalo de tempo, subtraídos pelo valor esperado para esta mesma contagem

\mathbf{V} = Matriz de variâncias e covariâncias entre as contagens de falhas para cada intervalo de tempo.

Em seguida foram testadas 3 distribuições probabilísticas comumente usadas em Análise de Sobrevivência (Exponencial, Weibull e Log-Normal), a fim de encontrar a que melhor se adéqua ao comportamento da variável “tempo até Perda Total”. A validação das distribuições foi feita comparando as probabilidades de sobrevivência estimadas de forma paramétrica com as estimadas pelo ajuste Kaplan-Meier. Também se verificaram pressupostos de linearidade, os valores de AIC/BIC e foi realizado um teste estatístico que compara a distribuição ajustada com a distribuição Gama Generalizada. Esta distribuição, que é um caso geral das três comparadas (Exponencial, Weibull e Log-Normal), costuma se ajustar melhor aos tempos de sobrevivência, mas não foi usada devido a natureza pouco interpretável de seus parâmetros. A estatística de teste tem distribuição Qui-Quadrado com $g-k$ graus de liberdade,

sendo g o número de parâmetros da Gama Generalizada e k o número de parâmetros da distribuição testada. A estatística de teste é dada pela seguinte função:

$$TRV = 2 * | l_{maior}(\beta|x) - l_{menor}(\beta|x) |$$

l_{maior} = log-verossimilhança do modelo com mais parâmetros de distribuição (Gama Generalizada)

l_{menor} = log-verossimilhança do modelo com menos parâmetros de distribuição

Uma vez definida a melhor distribuição probabilística, é feito o ajuste do modelo paramétrico de Análise de Sobrevivência. Modelos de Vida Acelerado (paramétricos) são do tipo que assumem uma distribuição de probabilidade para a variável “tempo até a falha”. Permitem estimar coeficientes que quantificam a relação entre a variável resposta e uma variável explicativa. Estes coeficientes são interpretados como o quanto uma variável explicativa acelera ou desacelera o tempo de vida. O método utilizado para a escolha das covariáveis a serem usadas no modelo é derivado da proposta de Collett (2003a), tendo sido usado no livro “Análise de Sobrevivência Aplicada” (COLOSIMO, 2006). O uso do mesmo é uma alternativa aos métodos automatizados de forward, backward e stepwise, que têm como limitação não oferecer ao pesquisador todos os possíveis conjuntos igualmente bons de combinações de covariáveis, já que apenas uma combinação é escolhida de forma automática pelo software. O modelo apresenta a seguinte forma:

$$\mu(x) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = x' \beta$$

$$Y = \log(T) = \mu(x) + \sigma v$$

$$v = \log(\epsilon); \epsilon \sim N(0, 1)$$

$$L(\theta) = \prod_{i=1}^n [f(y_i|x_i)]^{\delta_i} [S(y_i|x_i)]^{(1-\delta_i)}$$

A partir do modelo paramétrico, foram calculadas esperanças condicionadas ao tempo de contrato em anos. A ideia por trás do conceito é emular as Esperanças de Vida usadas nas tabelas atuariais para cálculos previdenciários e de pensão. Mas ao invés de condicionar o tempo de vida futuro ao tempo de vida passado (a idade do veículo), o mesmo foi

condicionado ao tempo de contrato. Esta escolha se deveu ao desconhecimento de riscos de falha anteriores ao momento do contrato, por limitações da base de dados, e à baixa significância demonstrada pela variável “Ano Fabricação”. A metodologia e resultados se encontram no Apêndice B.

Também foram construídas tabelas de probabilidades acumuladas de falha para um conjunto de associados que iniciam o contrato na mesma data. A interpretação frequentista da tabela a ser usada na gestão de risco da entidade será: a estimativa do percentual de ocorrências de Perda Total para o conjunto de associados que permanecerem na base da associação por um tempo t ou uma fração de t , em anos. As probabilidades, sob suposição da distribuição Log-Normal, constatada como a mais adequada, foram calculadas usando a função de percentis apresentada a seguir:

$$P(T \leq t_p) = F(t_p) = p$$

$$t_p = e^{z_p \cdot \sigma + \mu} = \exp(z_p \cdot \sigma + \mu)$$

A partir da distribuição Log-Normal dos tempos de falha, foi calculado o Valor Presente Atuarial (estimação do Valor Esperado) das despesas/ custos com os eventos de Perda Total. Esta metodologia é usada no Ramo-Vida de seguros para calcular, por exemplo, o valor presente esperado das obrigações futuras com o pagamento de indenizações dos Seguros de Vida. Para fins de comparação, também será usada a metodologia mais comum de Seguros do Ramo Não-Vida para o cálculo do Valor Esperado da Perda Agregada. Além do Valor Esperado, também foi calculada a Variância por estes dois métodos distintos. Para que os resultados fossem comparáveis, foram adotadas as seguintes premissas a ambos:

- Os valores devem ser estimados para uma coorte de veículos que experienciam o risco de sofrer o evento de Perda Total ao longo do ano de 2023.
- Os potenciais valores de despesas serão iguais aos Valores Fipe dos veículos que iniciaram seu contrato somente em 2022.
- Taxa de juros para trazer despesas a valor presente= 0. (Uma vez que não há forma determinada de aplicar uma taxa de juros pelo método do Ramo Não-Vida)
- Tempo de contrato/ tempo vivido antecipadamente pelos veículos, a princípio, não afetará as probabilidades de Perda Total. Pelo método do Ramo Não-Vida de seguros, estes tempos já não são considerados.

- Probabilidade de ocorrência do evento de Perda Total baseada na experiência de 5 anos da base (2018-2022).

Usando a mesma metodologia dos Seguros de Vida, tem-se que a esperança individual da despesa com cada veículo, ao longo de 1 ano e em tempo contínuo, é dada por:

$$VP \cdot \int_0^1 e^{-\delta t} \cdot {}_t p_x \cdot \lambda(x+t) dt$$

δ = taxa contínua de juros

${}_t p_x$ = probabilidade de sobrevivência por t anos, dado que se viveram x anos

$\lambda(x+t)$ = taxa de falha (ou força de mortalidade) aos $x+t$ anos

VP = Valor Protegido (no caso, o Valor Fipe de veículos cujo contrato se fez em 2022)

Considerando a taxa de juros igual a zero e o total de anos vividos antecipadamente também igual a zero, a esperança pode ser reescrita como:

$$VP \cdot \int_0^1 S(t) \cdot \lambda(t) dt = VP \cdot \int_0^1 f(t) dt$$

$f(t)$ = função densidade de probabilidade do evento de Perda Total (densidade da Log-Normal) no tempo t

Já a variância, esta possui a seguinte fórmula:

$$VP^2 \cdot \left[\int_0^1 e^{-2\delta t} \cdot {}_t p_x \cdot \lambda(x+t) dt - \left(\int_0^1 e^{-\delta t} \cdot {}_t p_x \cdot \lambda(x+t) dt \right)^2 \right]$$

Considerando as mesmas premissas, a fórmula pode ser reescrita como:

$$VP^2 \cdot \left[\int_0^1 f(t) dt - \left(\int_0^1 f(t) dt \right)^2 \right]$$

Como os veículos são independentes entre si, tanto a esperança quanto a variância da despesa total são dadas pela soma das estimativas individuais.

Para considerar os tempos vividos antecipadamente, mais especificamente, os tempos entre a data de contrato em 2022 e o dia 1 de janeiro de 2023 (demarcado como o início da exposição ao risco de Perda Total), a estimativa de esperança seria dada por:

$$VP \cdot \int_0^1 e^{-\delta t} \cdot {}_t p_x \cdot \lambda(x+t) dt$$

$$VP \cdot \int_0^1 \frac{S(t+x)}{S(x)} \cdot \lambda(t+x) dt = VP \cdot \int_0^1 \frac{f(t+x)}{S(x)} dt \quad \begin{array}{l} x = \text{tempo} \\ \text{vivido} \end{array}$$

antecipadamente

$S(x+t)$ = Sobrevivência ao evento de Perda Total até $x+t$ anos

$S(x)$ = Sobrevivência ao evento de Perda Total até x anos

$f(x+t)$ = função densidade de probabilidade do evento de Perda Total (densidade da Log-Normal) em $x+t$ anos

Já a variância, seria estimada da seguinte forma:

$$VP^2 \cdot \left[\int_0^1 e^{-2\delta t} \cdot {}_t p_x \cdot \lambda(x+t) dt - \left(\int_0^1 e^{-\delta t} \cdot {}_t p_x \cdot \lambda(x+t) dt \right)^2 \right]$$

$$VP^2 \cdot \left[\int_0^1 \frac{f(t+x)}{S(x)} dt - \left(\int_0^1 \frac{f(t+x)}{S(x)} dt \right)^2 \right]$$

* A metodologia de seguros de vida é baseada no livro “Actuarial Mathematics” (BOWERS, Newton et al., 1997).

Usando a metodologia de seguros do ramo Não-Vida, o Valor da Perda Agregada é descrita como uma distribuição Poisson Composta. Uma Poisson(λ) para o número de sinistros ocorridos em 1 ano e uma distribuição secundária Exponencial(α) para o valor das despesas. Assim sendo, a esperança da distribuição Composta é calculada como:

$$E(N) \cdot E(X) = \lambda \cdot \alpha$$

N = variável aleatória representando número de Perdas Totais no ano

X = variável aleatória representando valor da Perda Total

λ = parâmetro da Poisson

α = parâmetro da Exponencial

A variância da distribuição composta, por sua vez, possui a forma:

$$E(N) \cdot Var(X) + Var(N) \cdot E^2(X) = \lambda \cdot (Var(X) + E^2(X))$$

Que pode ser reescrita como:

$$\lambda \cdot (E(X^2)) = \lambda \cdot 2/\alpha^2$$

λ = parâmetro da Poisson

α = parâmetro da Exponencial

O parâmetro da Poisson deve ser estimado pela média amostral do número de sinistros em 1 ano. Para impedir a influência do crescimento da base de veículos na estimação, foi feita uma média amostral das frequências relativas dos anos, que por sua vez foi multiplicada pelo número de veículos no início de 2022. O parâmetro da Exponencial, por sua vez, foi estimado como o inverso da média amostral dos Valores Fipe de veículos cujo contrato se realizou em 2022.

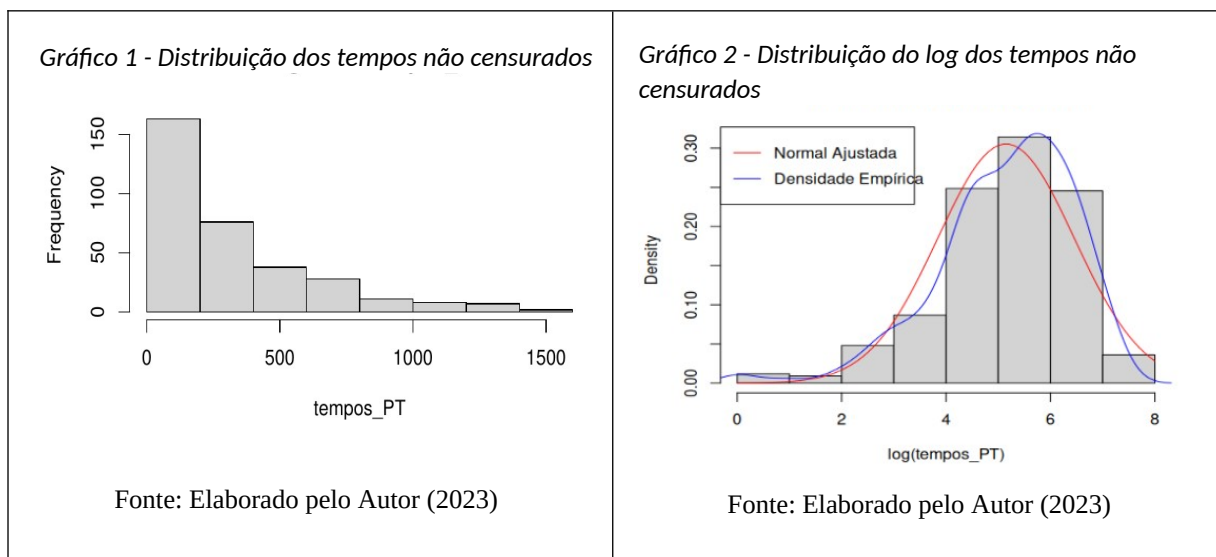
* A metodologia usada foi baseada no livro “NONLIFE ACTUARIAL MODELS” (Y.-K. TSE, 2009).

3. RESULTADOS

3.1 Análise Descritiva

Uma vez consolidada a base de dados, procura-se, através da análise descritiva, constatar suas limitações de informação, bem como características gerais das variáveis que serão importantes para a construção de modelos. Mais especificamente, características da variável resposta “tempo até Perda Total” e das potenciais variáveis explicativas.

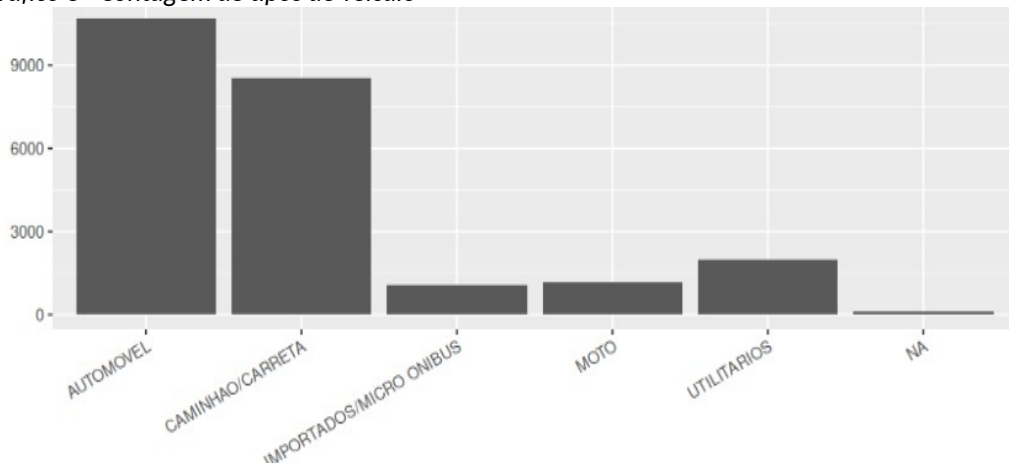
No Gráfico 1 é demonstrado que a distribuição dos tempos não censurados até a Perda Total é assimétrica. Mas quando fazemos uma transformação com a função $\log()$ (Gráfico 2), a distribuição passa a ter uma forma senoidal. Os tempos não censurados representam apenas 1,42% das linhas de vida (ou seja, 98,58% dos veículos da base passaram por censura).



3.1.1 Tipo Veículo

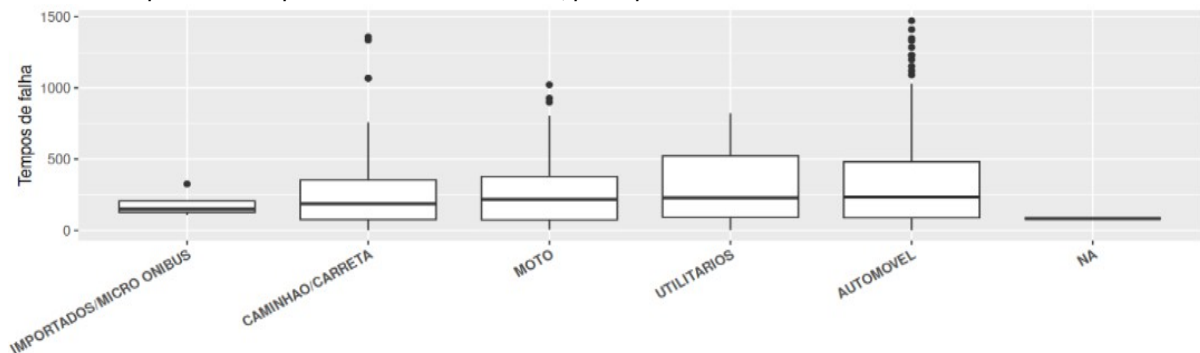
Esta potencial covariável demonstrou, no Gráfico 3, maior presença das categorias Automóvel e Caminhão/Carreta. Já no Gráfico 4, as que apresentaram maior tempo mediano até a falha (e portanto menor risco de PT) foram Automóvel, Utilitários e Moto. Vale destacar sobre o Gráfico 4 que a análise, ao desconsiderar os tempos censurados, perde informação relevante para determinar os riscos de cada categoria. Por tanto, este gráfico presta somente a uma análise superficial de riscos, que deve ser tomada com cautela.

Gráfico 3 - Contagem de tipos de veículo



Fonte: Elaborado pelo Autor (2023)

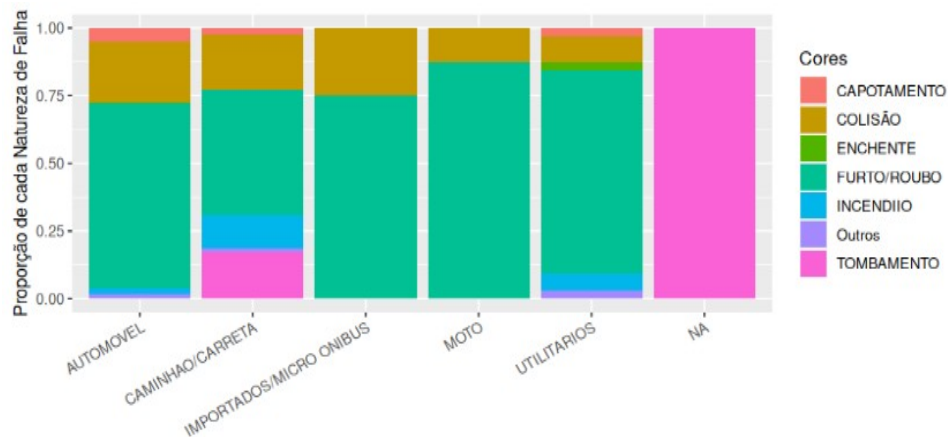
Gráfico 4 - Boxplot dos tempos não censurados de PT, por Tipo Veículo



Fonte: Elaborado pelo Autor (2023)

Observa-se no Gráfico 5 que a natureza dos eventos de falha mudam, a depender do Tipo de Veículo considerado. Incêndios e Tombamentos ocorrem com mais frequência para Caminhões/ Carretas. E Motos só passam por eventos do tipo Colisão e Furto.

Gráfico 5 - Proporção da natureza de eventos PT, por Tipo veículo

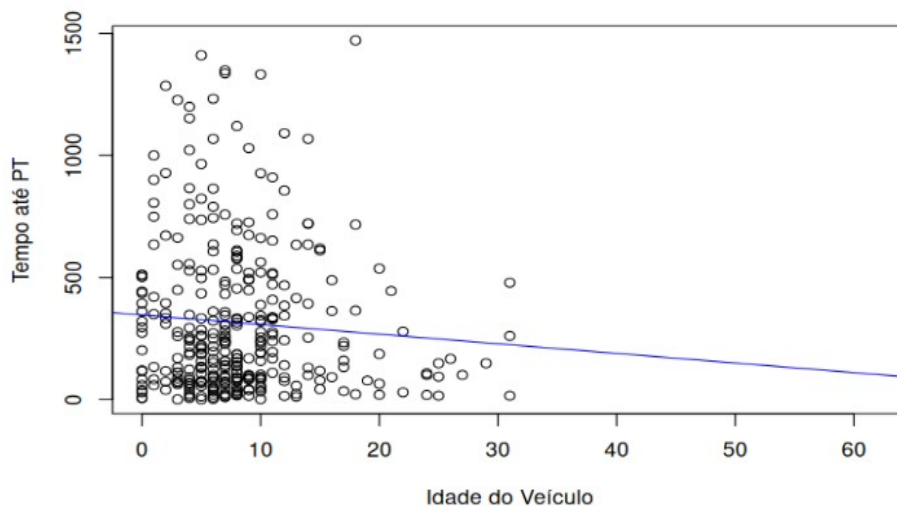


Fonte: Elaborado pelo Autor (2023)

3.1.2 Ano Fabricação (Idade Veículo)

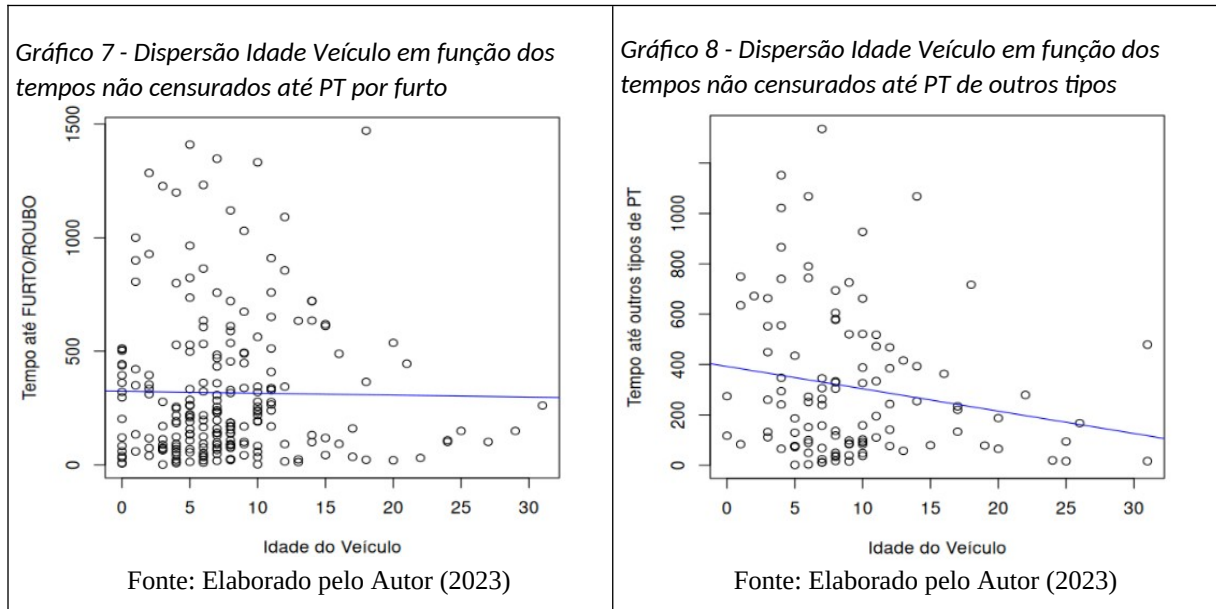
Foi constatado que seria melhor a criação de uma outra variável correlacionada, para a análise desta característica dos veículos: a idade. Mais especificamente, foi considerada a idade do veículo no momento do contrato. Desta forma, a interpretação do efeito da potencial covariável se torna mais direta. A variável criada apresenta média de 9,57 anos e desvio padrão de 7,12. Além disso, possui valores de máximo e mínimo, respectivamente, iguais a 0 e 62 e uma mediana de 9 anos. Para o teste Logrank da variável foi feita uma categorização usando os quantis. Segundo a análise do Gráfico 6, desconsiderando censuras, quanto mais velho é um veículo, menor tende a ser o seu tempo até a falha. Vale observar que há 81 veículos com dados faltantes, tendo 2 deles passado pelo evento de falha.

Gráfico 6 - Dispersão Idade Veículo em função dos tempos não censurados até PT



Fonte: Elaborado pelo Autor (2023)

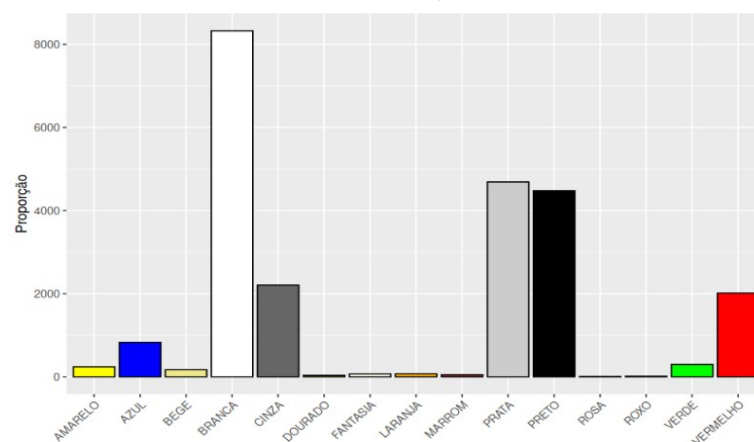
A variável Idade Veículo demonstrou diferença de influência sobre a variável de interesse, a depender da Natureza do Evento, como mostram os Gráficos 7 e 8. Estes, bem como o Gráfico 6, devem ser analisados com cautela, uma vez que perdem informações relevantes de dados censurados.



3.1.3 Cor do Veículo

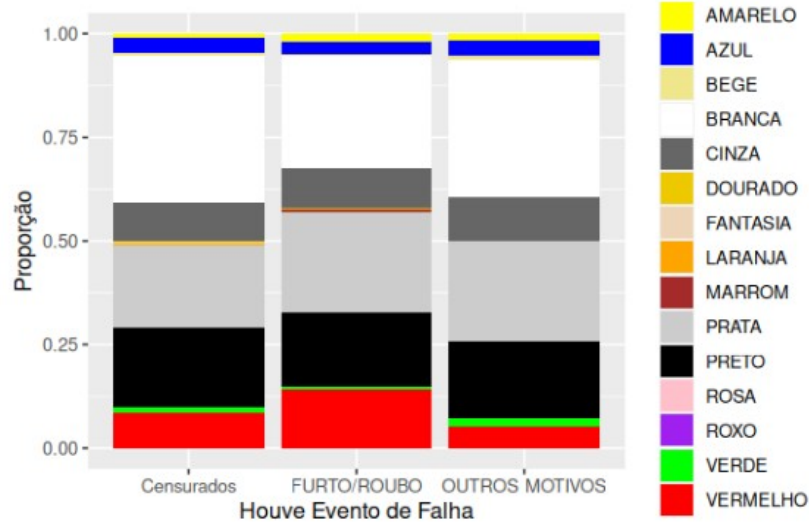
A distribuição da variável “Cor” demonstra, no Gráfico 9, predominância de veículos brancos, pratas e pretos. Possui 26 dados faltantes. No Gráfico não se constatou diferença significativa entre indivíduos censurados e indivíduos que sofreram o evento de falha. Para fins de comparação, foi feita no Gráfico 10 uma estratificação das ocorrências de falha entre aqueles que passaram por Furto/Roubo e aqueles que passaram por outras causas.

Gráfico 9 - Distribuição da variável categórica Cor



Fonte: Elaborado pelo Autor (2023)

Gráfico 10 - Distribuição da variável Cor por natureza do evento

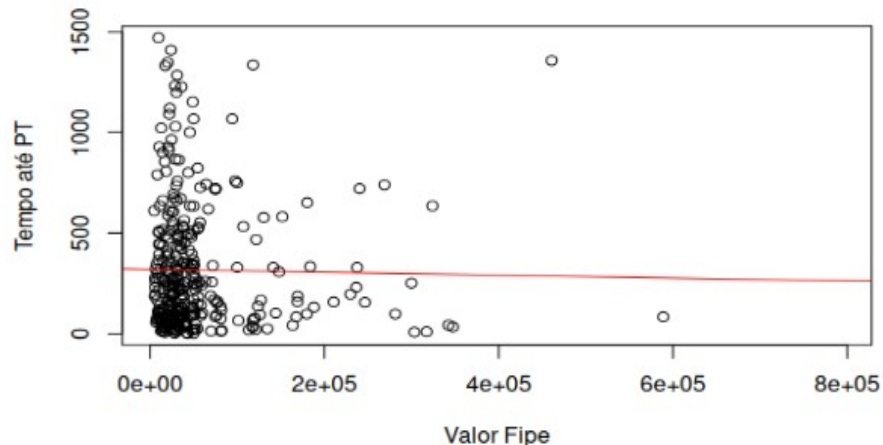


Fonte: Elaborado pelo Autor (2023)

3.1.4 Valor Fipe

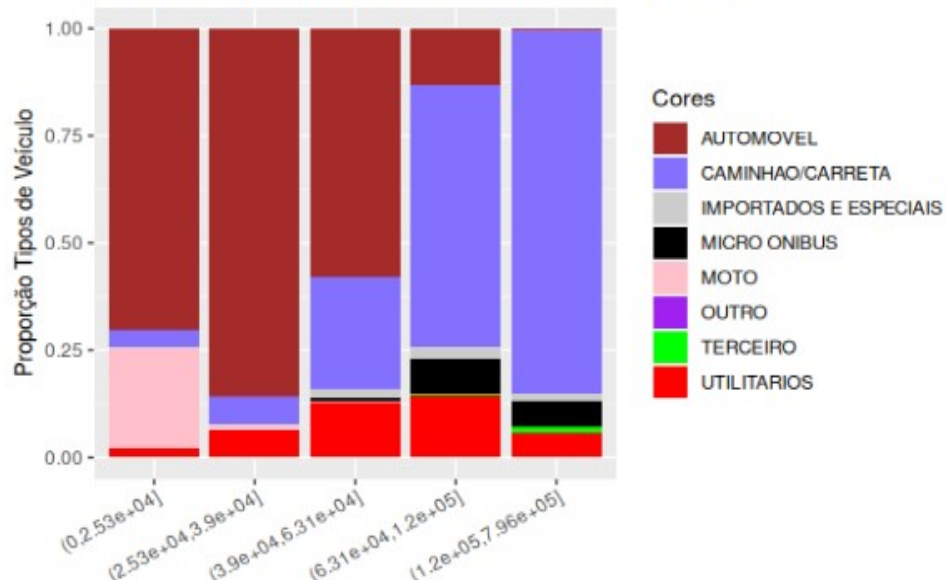
Tal variável numérica possui média R\$ 84.425,27 e desvio padrão de R\$ 97.668,30. Os valores de mínimo e máximo foram, respectivamente, R\$ 100,00 e R\$ 795.996,00. Já a Mediana, R\$ 84.328,00. Há 27 dados faltantes. Esta potencial covariável possui, provavelmente, uma forte associação com a variável Tipo Veículo, como mostra o Gráfico 12, de barras, em que a variável numérica foi categorizada de forma que as categorias possuíssem quantidades semelhantes de veículos (foram delimitadas por quantis). Além disso, fazendo uma análise sem considerar censuras no Gráfico 11 (necessária mesma cautela de gráficos anteriores), a variável apresentou uma correlação fraca e negativa com o tempo até o evento de falha.

Gráfico 11 - Dispersão da variável Valor Fipe em função dos tempos não censurados até PT



Fonte: Elaborado pelo Autor (2023)

Gráfico 12 - Proporção da variável Tipo Veículo por categoria de Valor Fipe

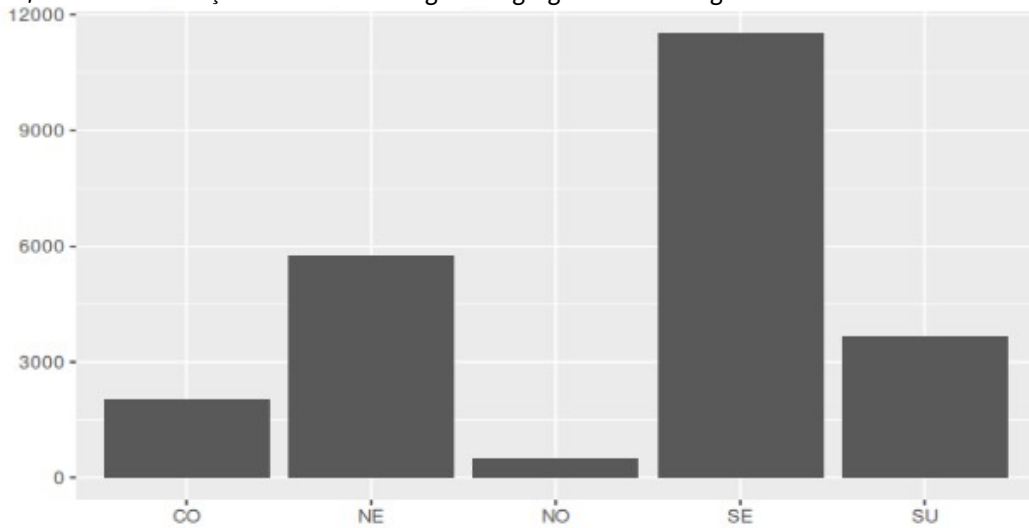


Fonte: Elaborado pelo Autor (2023)

3.1.5 Estado/ macrorregião

A variável Estado foi recategorizada nas 5 macrorregiões do Brasil, a fim de diminuir o número de parâmetros (β 's) a serem estimados no modelo paramétrico. A distribuição desta variável agregada pode ser vista no gráfico 13.

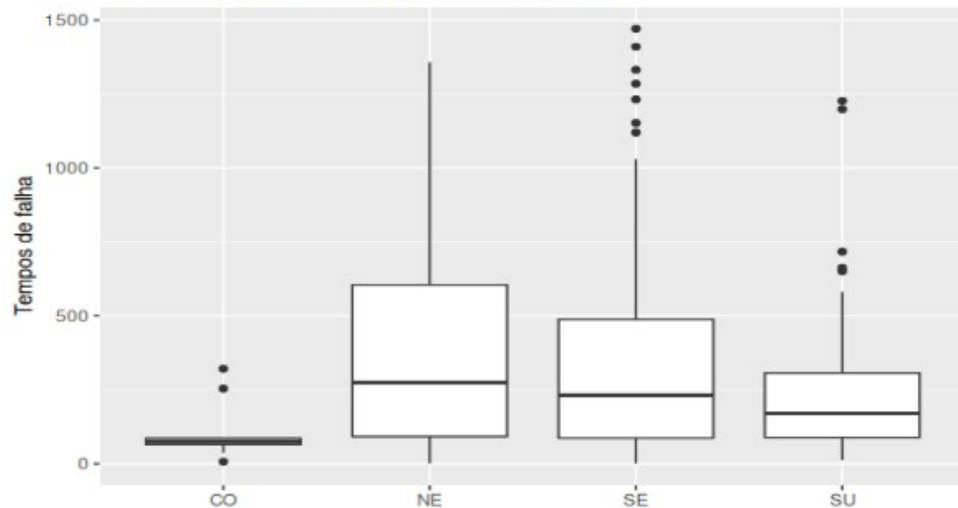
Gráfico 13 - Distribuição da variável categórica agregada Macrorregião



Fonte: Elaborado pelo Autor (2023)

A região Norte não apresentou eventos. Enquanto o Nordeste apresenta maior tempo mediano de falha, quando não consideramos as censuras (necessária cautela devido perda de informação). Ou seja, demonstra indícios de ser a região com o menor risco de falha.

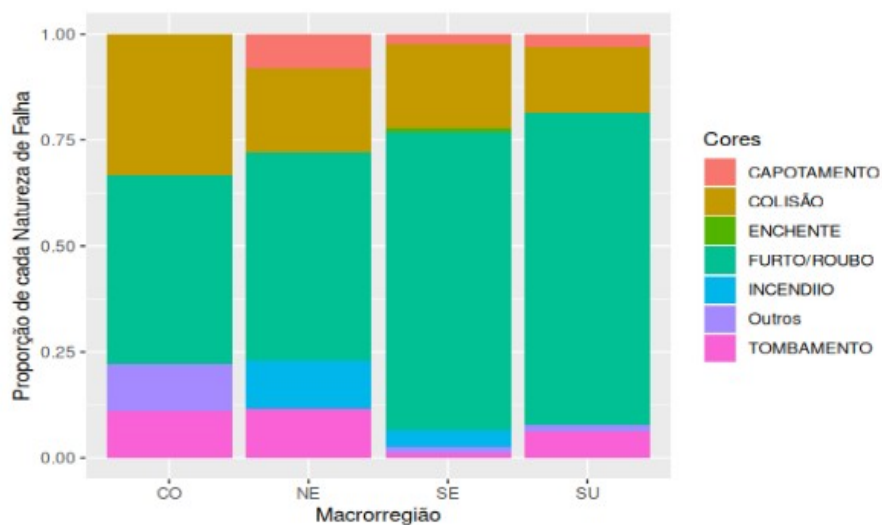
Gráfico 14 - Boxplot dos tempos não censurados de PT, por Macrorregião



Fonte: Elaborado pelo Autor (2023)

Fazendo uma análise da natureza dos eventos de falha, vemos no Gráfico 15 que o Nordeste possui muitos tombamentos e incêndios, que são eventos tipicamente relacionados a Caminhões e Carretas. E esta categoria de tipo de veículo normalmente apresenta menos eventos de falha, quando comparada a outras como Moto e Automóvel (Gráfico 4).

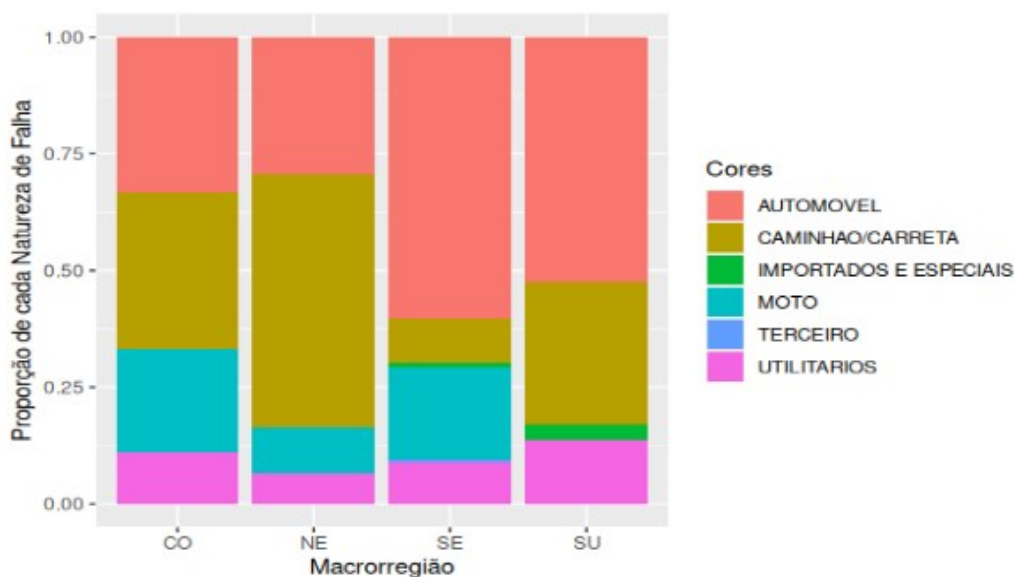
Gráfico 15 - Proporção da natureza dos eventos PT por Macrorregião



Fonte: Elaborado pelo Autor (2023)

No Gráfico 16, torna-se mais evidente a associação entre Macrorregião e Tipo de Veículo. Ou seja, a configuração de tipos de veículo na base muda em cada macrorregião:

Gráfico 16 - Proporção da variável Tipo Veículo por Macrorregião



Fonte: Elaborado pelo Autor (2023)

3.1.6 Histórico de Perdas Parciais

A variável “número de Perdas Parciais prévias” se mostrou pouco vantajosa. Os únicos 16 veículos da base que passaram por 2 eventos de perda parcial, bem como o único que passou por 3 eventos do tipo, não chegaram a passar por Perda Total. Ainda que fosse feita uma categorização do tipo binária para esta variável (1 caso tenha ocorrido perda parcial no histórico do veículo e 0 caso contrário), as duas categorias seriam pouco comparáveis, tendo apenas 6 veículos na primeira e 328 na segunda. Por este motivo, é inconclusiva a hipótese de influência da variável “número de Perdas Parciais prévias” sobre a variável de interesse, tempo até uma Perda Total.

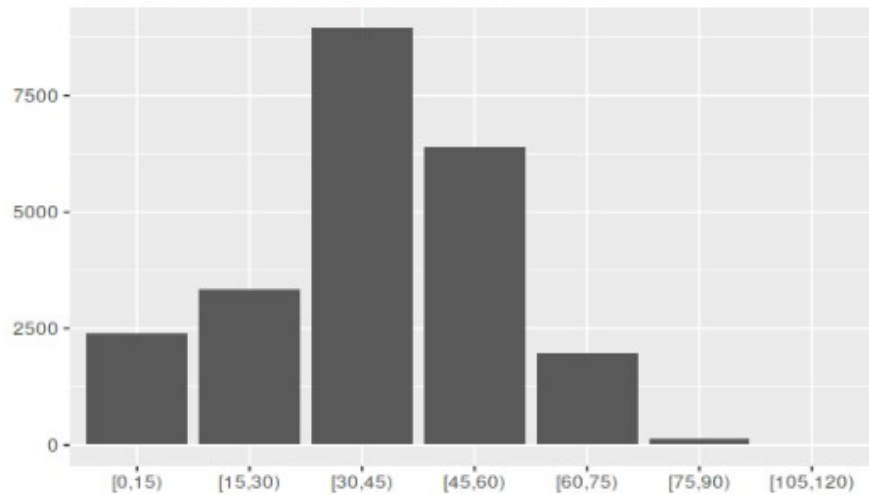
3.1.7 Estado Civil

O número de dados faltantes para esta variável impossibilita o uso da mesma. Os mesmos constituem-se como uma categoria que representa 99,987% da base, enquanto o restante consta como sendo da categoria “casados”. Ou seja, também não há uma categoria para solteiros.

3.1.8 Idade do Associado

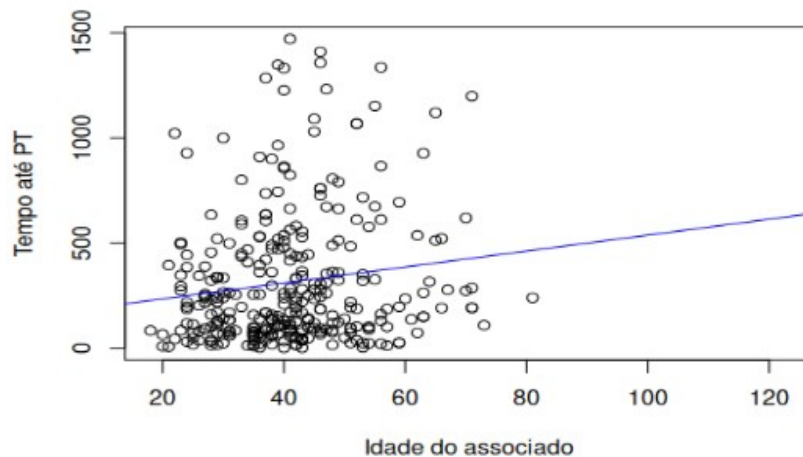
Vale destacar que esta variável, obtida pela diferença entre a data de contrato e a data de nascimento, possui, devido a um mau preenchimento dos dados referentes à segunda, muitos valores duvidosos. Há várias idades inferiores a 18 anos (cerca de 2.993 associados) e outras superiores a 100, além de outros 369 dados faltantes. A média é de 31,8 anos e o desvio padrão, aproximadamente 16 anos. A distribuição da variável categorizada em faixas etárias de 15 anos se observa no Gráfico 17. Apesar dos problemas, quando desconsideramos censuras, a variável demonstra correlação significativamente positiva com o tempo até a falha (ainda que desconsiderássemos as idades discrepantes inferiores a 18 anos – Gráfico 18). Ou seja, quanto mais velho é o associado/ proprietário do veículo, maior tende a ser o tempo até uma falha.

Gráfico 17 - Distribuição da variável Idade Associado categorizada em faixas etárias



Fonte: Elaborado pelo Autor (2023)

Gráfico 18 - Dispersão da variável Idade Associado em função dos tempos não censurados até PT

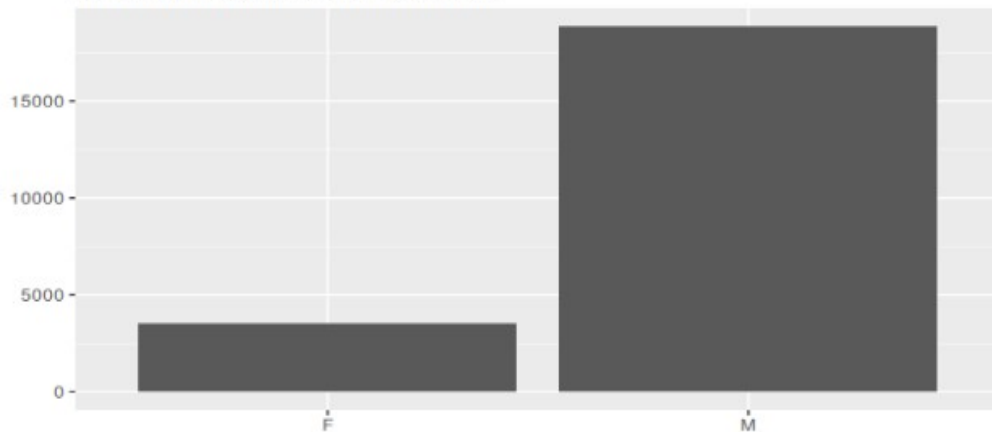


Fonte: Elaborado pelo Autor (2023)

3.1.9 Sexo do Associado

Esta variável possui 1094 dados faltantes (cerca de 4,6% da base). Aproximadamente 80,3% da base é de veículos cujos proprietários são do sexo masculino. Devido a esta diferença verificada no Gráfico 19, o sexo masculino acaba sendo, por isto, preponderante para todos os Tipos de Veículo e todas as Macrorregiões.

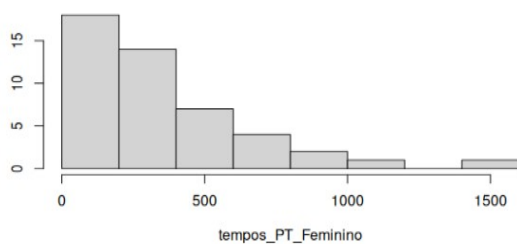
Gráfico 19 - Distribuição da variável Sexo



Fonte: Elaborado pelo Autor (2023)

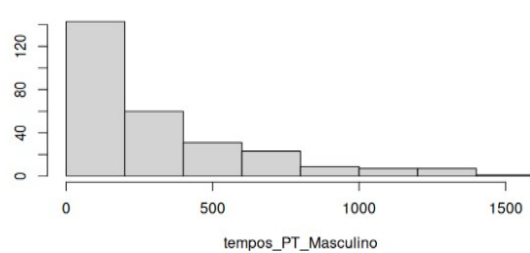
Devido a disparidade, torna-se difícil comparar as duas categorias. Uma possível forma é através de histogramas contendo as distribuições dos tempos não censurados de falha, para cada sexo. Percebe-se que a cauda direita é mais pesada no caso das mulheres, em comparação com os homens. Isto pode ser uma indicação de que o tempo até a falha para a primeira categoria tende a ser maior em relação à segunda.

Gráfico 21 - Distribuição dos tempos não censurados até PT, para mulheres



Fonte: Elaborado pelo Autor (2023)

Gráfico 20 - Distribuição dos tempos não censurados até PT, para homens



Fonte: Elaborado pelo Autor (2023)

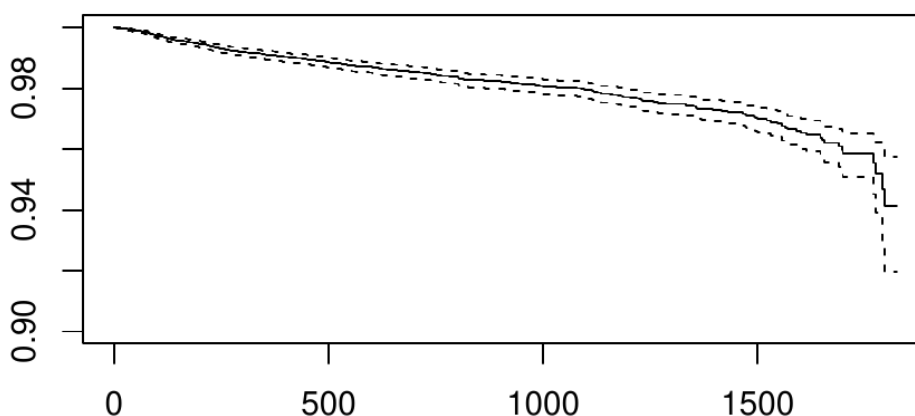
3.1.10 Melhorias para a base de dados

A importância do correto preenchimento das bases de dados é vista na análise descritiva. Um número excessivo de dados faltantes pode impossibilitar o uso de uma potencial covariável, como foi o caso do “Estado Civil”. Também podem diminuir a significância de algumas covariáveis, por estas apresentarem, por exemplo, categorias com dados escassos. É o que provavelmente ocorreu com a variável Sexo, como veremos adiante. Além dos dados faltantes, há indícios de preenchimento incorreto de algumas informações, que podem criar interpretações enganosas sobre a relação entre a variável resposta e uma potencial covariável (pode ser o caso da “Idade do Associado”). Alerta-se sobre a importância de se atentar para a qualidade dos dados, afinal, são a partir destes que se constrói um modelo estatístico. Tanto a precisão quanto o viés do modelo são, em parte, dependentes da fidelidade da base à realidade que se encontra. Diminuir o número de categorias para as variáveis categóricas é uma das poucas formas de contornar (parcialmente) este problema.

3.2 Modelagem Não Paramétrica

Foram utilizadas técnicas de modelagem não paramétrica para a visualização do comportamento da variável “tempo até Perda Total”, quando consideramos censuras. Afinal, dados censurados carregam informação sobre o risco de falha, ainda que não tenham passado pela mesma. Mais especificamente, busca-se compreender o comportamento de uma função da variável de interesse: a função de sobrevivência, $S(t)$, ilustrada no Gráfico 22.

Gráfico 22 - Função de Sobrevivência ($S(t)$) estimada por Kaplan-Meier

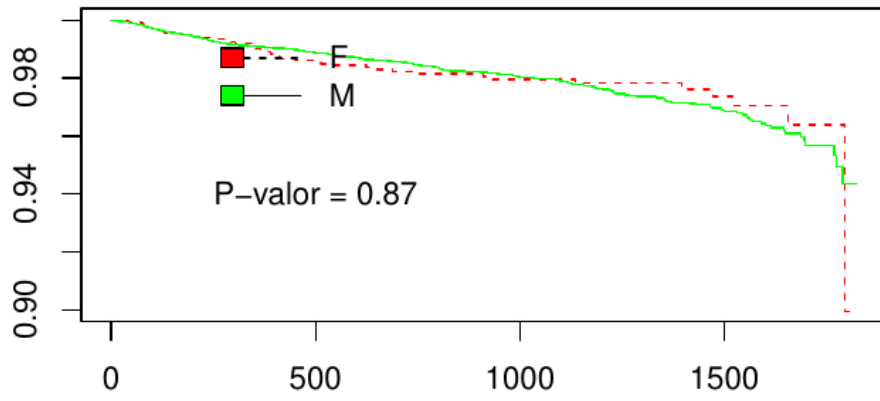


Fonte: Elaborado pelo Autor (2023)

Devidos às censuras, não é possível ver a curva de sobrevivência completa com os dados disponíveis, sendo esta restrita aos tempos em que as probabilidades de sobrevivência são superiores a 90%. Para tanto, seria necessário pressupor uma distribuição de probabilidade para os tempos.

Ainda por técnicas não paramétricas, testamos a significância das potenciais covariáveis pelo teste Log-rank. Nos Gráficos de número 23 a 29, são escolhidas duas categorias de cada potencial covariável a fim de demonstrar a significância da mesma. No caso de variáveis numéricas, foi necessária uma categorização, feita com base em percentis amostrais múltiplos de 10% (todas as categorias criadas para estas variáveis são citadas no Apêndice A). Os critérios para a seleção das duas categorias comparadas são: uma diferença perceptível, que evidencie a significância da variável, e um número de eventos comparável. No Apêndice A são apresentados gráficos envolvendo todas as categorias, com o p-valor do teste Log-Rank.

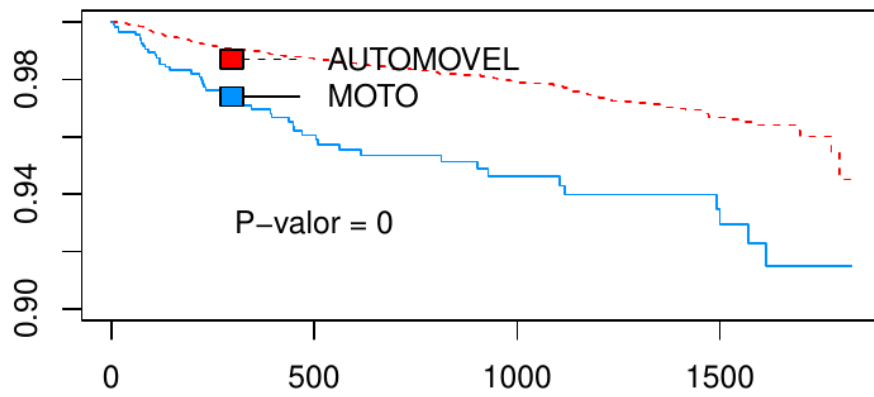
Gráfico 23 - Sobrevivência Kaplan-Meier $\hat{S}(t)$ para variável categórica Sexo



Tempo t até Perda Total (dias)

Fonte: Elaborado pelo Autor (2023)

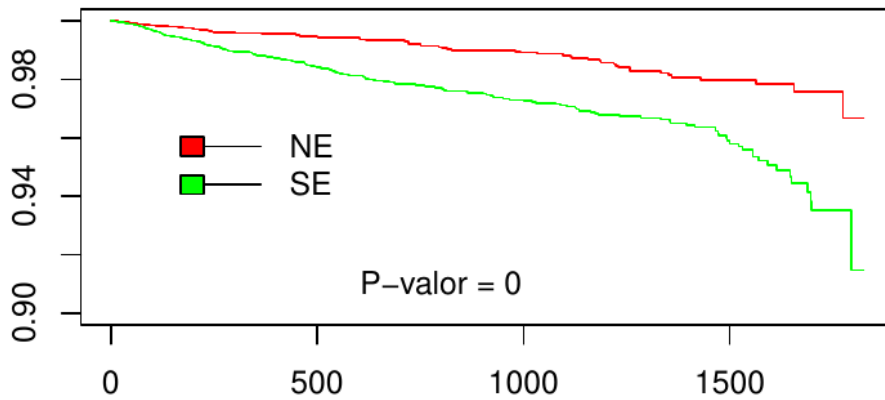
Gráfico 24 - Sobrevivência Kaplan-Meier $\hat{S}(t)$ para variável categórica Tipo Veículo



Tempo t até Perda Total (dias)

Fonte: Elaborado pelo Autor (2023)

Gráfico 25 - Sobrevivência Kaplan-Meier $\hat{S}(t)$ para variável categórica Macrorregião



Tempo t até Perda Total (dias)

Fonte: Elaborado pelo Autor (2023)

Fonte: Elaborado pelo Autor (2023)

Gráfico 26 - Sobrevivência Kaplan-Meier $\hat{S}(t)$ para variável categórica Cor

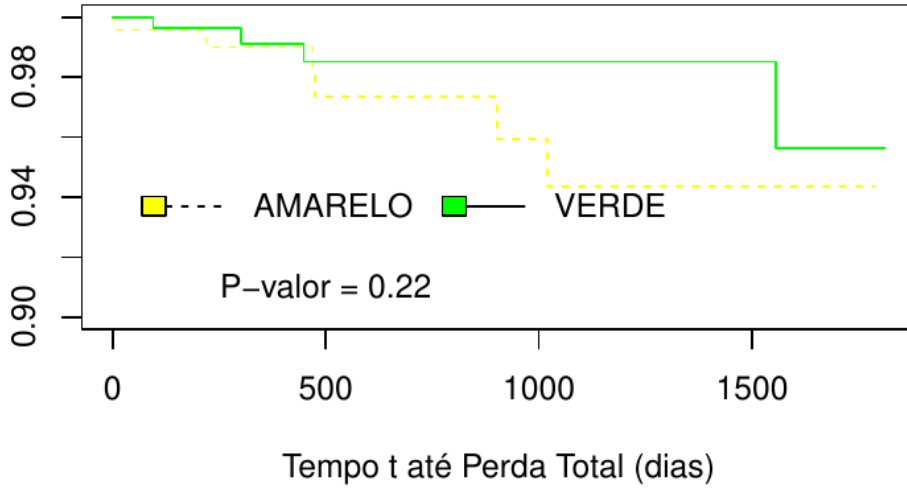
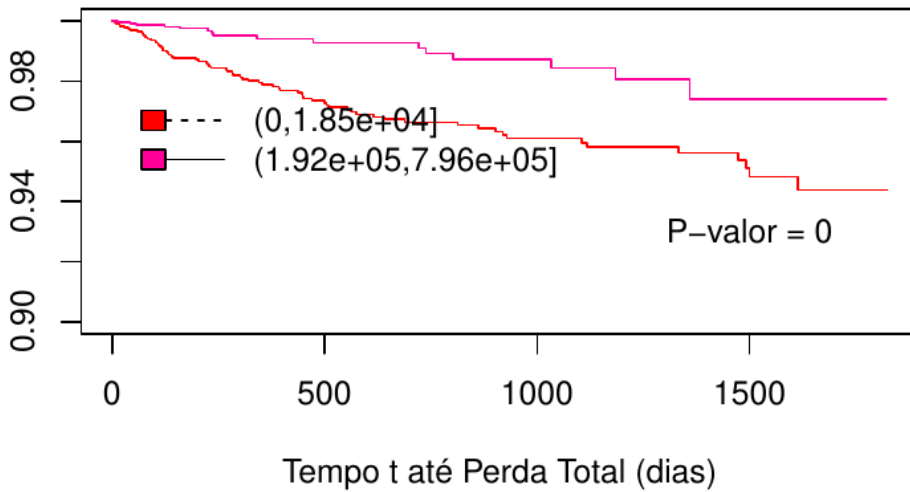
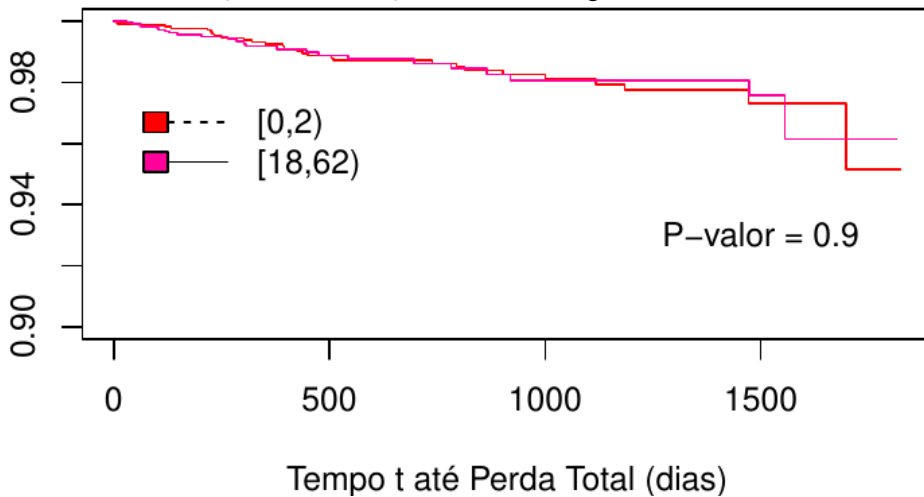


Gráfico 27 - Sobrevivência Kaplan-Meier $\hat{S}(t)$ para variável categorizada Valor Fipe

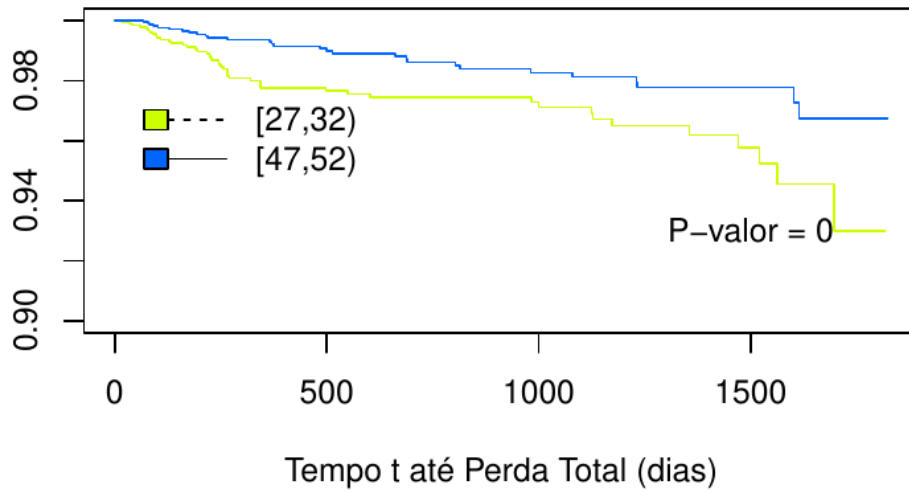


Fonte: Elaborado pelo Autor (2023)

Gráfico 28 - Sobrevivência Kaplan-Meier $\hat{S}(t)$ para variável categorizada Idade Veículo



Fonte: Elaborado pelo Autor (2023)

Gráfico 29 - Sobrevivência Kaplan-Meier $\hat{S}(t)$ para variável categorizada Idade Associado

Fonte: Elaborado pelo Autor (2023)

Os resultados dos testes Log-Rank, realizados para cada potencial covariável no Apêndice A, seguem na Tabela 1.

Tabela 1 - Testes Log-Rank das potenciais variáveis explicativas do banco

Variável Preditora	Valor-p Log-Rank	g.l
Sexo	0.865	1
Tipo Veículo	<0.001	4
Macro-região	<0.001	3
Cor	0.644	14
Valor FIPE Veiculo	<0.001	9
Ano Fab. (Idade Veiculo)	0.159	9
Idade Associado	<0.001	9

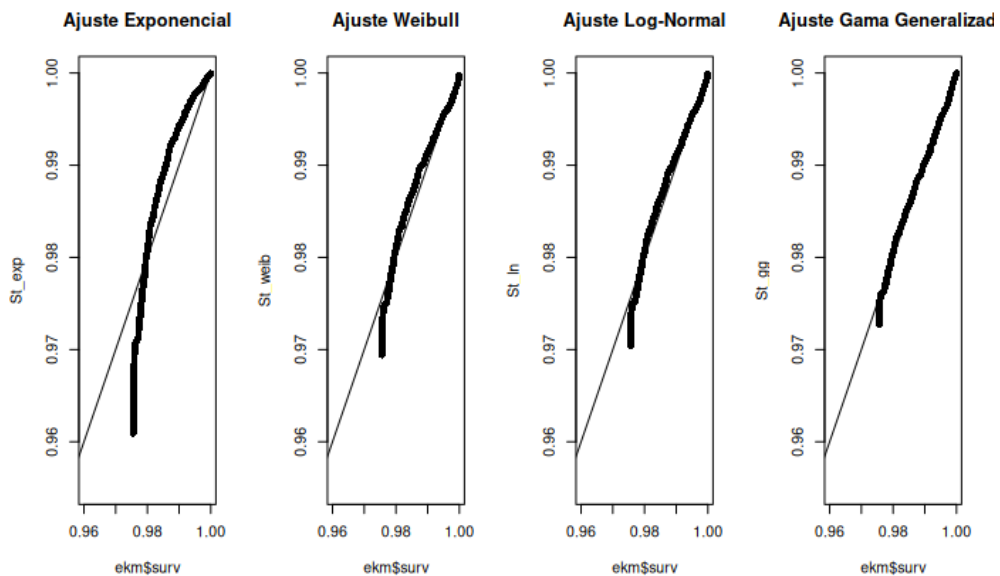
Fonte: Elaborado pelo Autor (2023)

3.3 Modelagem Paramétrica

3.3.1 Escolhendo o modelo

Comparando, no Gráfico 30, os ajustes dos modelos Exponencial, Weibull e Log-Normal às estimativas do Kaplan-Meier, o ajuste Log- Normal demonstrou ser o melhor, depois da Gama Generalizada. Como dito na Metodologia, não usamos esta última devido à dificuldade de interpretação dos resultados, dada pela natureza dos parâmetros da distribuição.

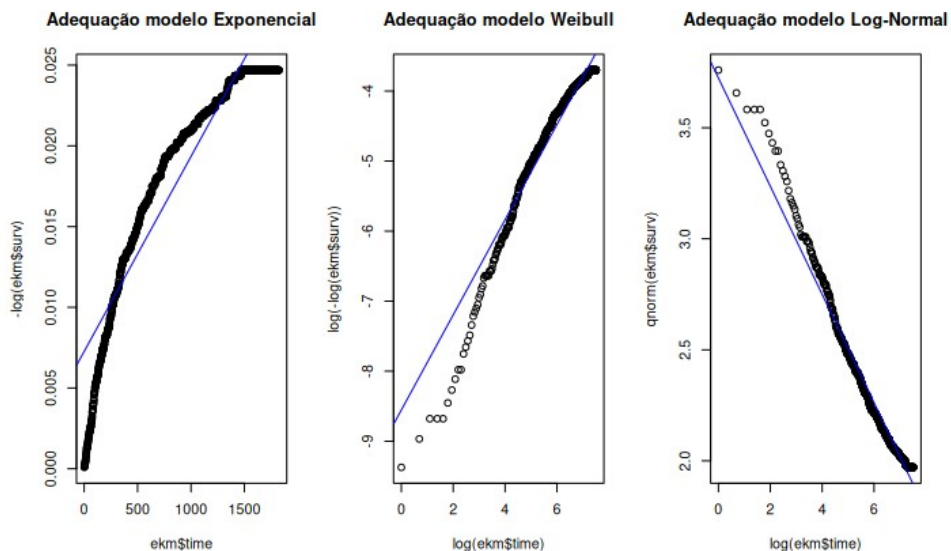
Gráfico 30 - Ajuste dos modelos de probabilidade às estimativas Kaplan-Meier



Fonte: Elaborado pelo Autor (2023)

Verifica-se, no Gráfico 31, se existe linearidade entre a variável tempo (ou uma modificação desta) e a transformação da probabilidade de sobrevivência adequada a cada distribuição. O modelo Log- Normal demonstra novamente ser o melhor, por apresentar menos desvios da reta que indica a relação linear esperada.

Gráfico 31 - Verificação da linearidade de cada ajuste



Fonte: Elaborado pelo Autor (2023)

Na Tabela 2 se encontram os testes de comparação dos ajustes com a Gama-Generalizada. O teste rejeita todas as distribuições a um nível de 5% (apesar de o p-valor da Log-Normal ter sido o maior), o que é uma ocorrência comum para grandes bases de dados.

Tabela 2 - Teste de adequação em relação a Gama-Generalizada

Modelo	$\log(L(\theta))$	TRV	valor-p
Gama Generalizada	-3858.0231		
Exponencial	-3918.0773	120.1085	<0.0001
Weibull	-3875.1423	34.2384	<0.0001
Log-Normal	-3865.2732	14.5002	0.0001

Fonte: Elaborado pelo Autor (2023)

Porém, na Tabela 3, os valores de AIC e BIC para cada modelo reforçam a constatação de que o modelo Log- Normal demonstra ser o melhor, em relação a Gama-Generalizada, por apresentar maior ganho de verossimilhança e, portanto, menores AIC e BIC.

Tabela 3 - AIC e BIC dos modelos paramétricos sem covariáveis

Modelo	AIC	BIC
Gama Generalizada	7718.05	7726.11
Exponencial	7838.15	7846.22
Weibull	7752.28	7760.35
Log-Normal	7732.55	7740.61

Fonte: Elaborado pelo Autor (2023)

3.3.2 Tempo estimado para probabilidades de Perda Total

Assumindo, agora, distribuição Log-Normal, foi construída a Tabela 4 com Probabilidades acumuladas de Perda Total, junto a seus respectivos quantis representados em meses. Para tanto, o modelo construído não considerou covariáveis.

Tabela 4 - Tempos estimados para Perda Total de p% dos veículos

Percentual de Falha	Tempo em meses até percentual Falha	IC 95%
0.1%	0.50	[0.24 ; 0.76]
0.2%	1.17	[0.54 ; 1.8]
0.3%	1.96	[0.88 ; 3.05]
0.4%	2.88	[1.26 ; 4.49]
0.5%	3.90	[1.69 ; 6.1]
0.6%	5.02	[2.15 ; 7.89]
0.7%	6.25	[2.65 ; 9.85]
0.8%	7.58	[3.18 ; 11.98]
0.9%	9.00	[3.74 ; 14.26]
1%	10.53	[4.34 ; 16.71]

Fonte: Elaborado pelo Autor (2023)

3.3.3 Valor Esperado de custos com Perda Total

Usando o mesmo modelo de distribuição Log-Normal para a modelagem do tempo até o evento de falha, foram feitas estimativas de esperança e variância das despesas com sinistros de Perda Total, usando a metodologia de seguros do ramo Vida. Comparativamente, usou-se uma metodologia típica do ramo Não-Vida. As estimativas de Esperança e Variância foram semelhantes, como mostra a Tabela 5.

Tabela 5 - Custos com Perda Total para 2023: comparação de metodologias

Método Usado	Valor Esperado dos Custos	Variância dos Custos
Seguro Vida	10112149	2.469068e+12
Seguro Não-Vida	14421491	3.190650e+12

Fonte: Elaborado pelo Autor (2023)

O valor esperado dos custos com sinistros de Perda Total, para a entidade de risco, ao longo de 2023, foi estimado em R\$ 10.112.149,00, segundo a metodologia de seguros de vida. E o valor de R\$ 14.421.491,00 foi a estimativa obtida usando a metodologia do Ramo-Não-Vida de seguros. O primeiro método resultou em uma redução no valor da esperança, bem como da variância, ainda que o segundo método, vale destacar, seja sensivelmente afetado por pequenas variações na estimação do parâmetro da Poisson (estimado como a média do número de eventos de Perda Total no ano). No entanto, a vantagem da metodologia de seguros do Ramo-Vida é que esta pode considerar o tempo vivido previamente por cada veículo na base como condição para as probabilidades de falha. Se isto for feito, o valor a ser provisionado pela entidade de risco pode ser mais fidedigno à realidade da base. Com este intuito foi construída a Tabela 6. Ao considerar o tempo vivido antecipadamente pelos veículos, o valor a ser provisionado se demonstrou menor (tanto a esperança quanto a variância), o que pode ser positivo para a entidade de gestão de risco, permitindo que esta realoque parte de seus recursos que estariam constituindo um passivo.

Tabela 6 - Custos com Perda Total para 2023: avaliando o efeito do tempo previamente vivido

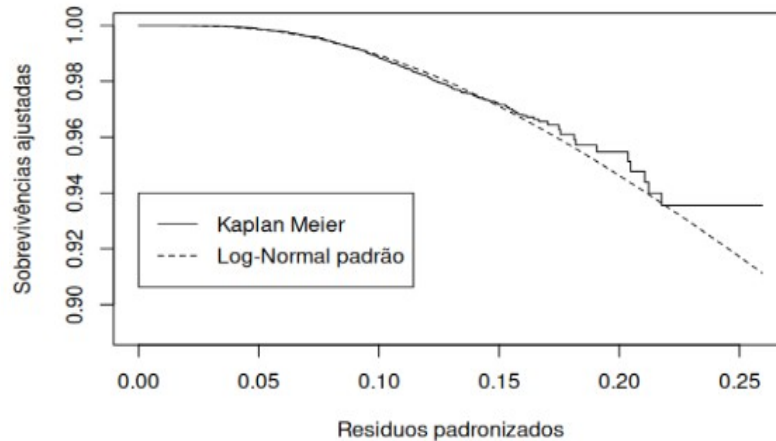
Método Usado	Valor Esperado dos Custos	Variância dos Custos
Seguro Vida não considerando tempo vivido	10112149	2.469068e+12
Seguro Vida considerando tempo vivido	7360870	5.010678e+07

Fonte: Elaborado pelo Autor (2023)

3.3.4 Construção do modelo completo

O modelo final obteve resíduos padronizados que mostraram bom ajuste do modelo paramétrico Log-Normal em relação as estimativas não paramétricas Kaplan-Meier, o que é ilustrado no Gráfico 32.

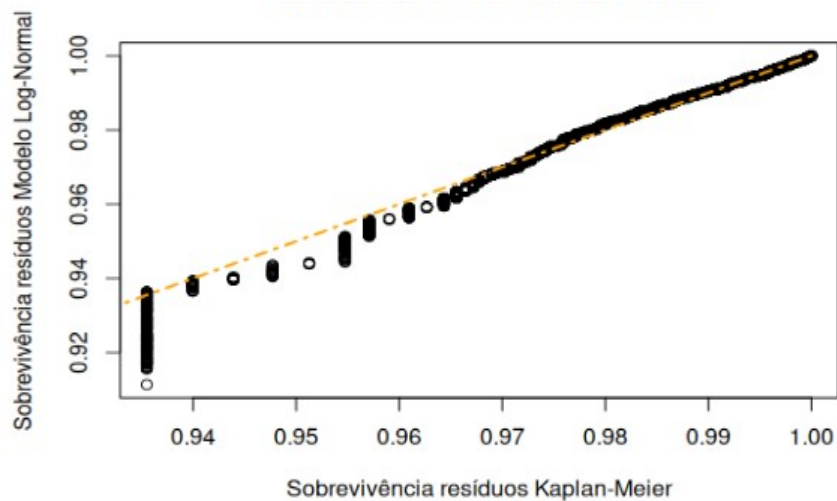
Gráfico 32 - Adequação resíduos do modelo Log-Normal



Fonte: Elaborado pelo Autor (2023)

Os ajustes das Probabilidades de Sobrevivência dos resíduos dos dois modelos, o paramétrico e o não paramétrico, parecem coincidir no Gráfico 33, especialmente em probabilidades de sobrevivência maiores, as quais se encontram em menores tempos de vida. Naturalmente, quanto mais ao futuro é feita a projeção, maior é a incerteza.

Gráfico 33 - Probabilidades de sobrevivência dos resíduos Log-Normal e Kaplan-Meier



Fonte: Elaborado pelo Autor (2023)

Segue o resultado do modelo paramétrico completo na Tabela 7, o qual manteve em sua construção apenas as variáveis macrorregião (transformação da variável Estado) e Tipo Veículo como significativas.

Tabela 7 - Modelo Log-Normal com covariáveis Tipo Veículo e Macrorregião

	Estimativa	EP	Estatística de Teste	Valor-p
(Intercepto)	15.37	0.45	34.07	0.00
CAMINHAO/CARRETA	0.49	0.22	2.22	0.03
IMPORTADOS/MICRO ONIBUS	2.05	0.69	2.98	0.00
MOTO	-1.57	0.30	-5.27	0.00
UTILITARIOS	-0.18	0.31	-0.58	0.56
Região NO/CO	1.13	0.48	2.36	0.02
Região SE	-0.94	0.24	-3.99	0.00
Região SU	-1.06	0.28	-3.81	0.00
Log(sigma)	1.36	0.04	30.24	0.00

Fonte: Elaborado pelo Autor (2023)

Forma do modelo:

$$\mu = 15,37 + 0,49 * X1 + 2,05 * X2 - 1,57 * X3 - 0,18 * X4 + 1,13 * X5 - 0,94 * X6 - 1,06 * X7$$

$$\sigma = \exp(1,358547) \approx 3.89$$

$$S(t) = \phi\left(\frac{-\log(t) + \mu}{\sigma}\right)$$

X1= 1 caso o veículo pertença à categoria Caminhão/Carreta e 0 casos contrário

X2= 1 caso o veículo pertença à categoria Importados/Micro Onibus e 0 casos contrário

X3= 1 caso o veículo pertença à categoria Moto e 0 casos contrário

X4= 1 caso o veículo pertença à categoria Utilitários e 0 casos contrário

X5= 1 caso o veículo pertença à categoria NO/CO (Norte/Centro-Oeste) e 0 casos contrário

X6= 1 caso o veículo pertença à categoria SE (Sudeste) e 0 casos contrário

X7= 1 caso o veículo pertença à categoria SU (Sul) e 0 casos contrário

Destaca-se que, no caso das duas variáveis mantidas, foram necessárias recategorizações, devido a existência de uma categoria pouco representativa na base que não contém eventos. Isto impossibilita a estimação de um Beta que represente o efeito desta categoria no tempo de falha. Para a variável Tipo Veículo, a categoria sem eventos “MICRO

ONIBUS” foi agregada com outra categoria pouco representativa na base, que ao menos contém eventos: “IMPORTADOS E ESPECIAIS”. Já no caso da variável Macrorregião (formada com a agregação de Estados), a categoria sem eventos “NO” (região Norte) passou pelo mesmo processo, sendo agregada com uma categoria de pouca representação na base, mas que contém eventos: “CO” (região Centro-Oeste).

Ao longo do processo, a variável Cor foi definida pouco relevante para explicar a variabilidade do tempo até uma falha. Valor Fipe, que apresenta forte correlação com a variável Tipo Veículo, não explica tão bem a variabilidade da resposta quanto a segunda. As variáveis Sexo e Idade Veículo (sendo esta segunda a transformação da variável Ano Fabricação), que já tinham sido rejeitadas no teste não paramétrico Log-Rank, mostraram pouca significância dos respectivos coeficientes Beta estimados. Por fim, a variável Idade Associado apresentou a mesma inconsistência das duas variáveis supracitadas, por mais que não tenha sido rejeitada no teste Log-Rank.

No resultado do modelo, a categoria base é de Automóvel no Nordeste. Os betas são interpretados usando a seguinte relação:

$$\begin{aligned} \text{tempo}(x=1, \beta) / \text{tempo}(x=0, \beta) &= \exp(\beta) \\ \text{tempo}(x=1, \beta) &= \exp(\beta) * \text{tempo}(x=0, \beta) \end{aligned}$$

O que consiste em dizer, por exemplo, que fazer parte da categoria CAMINHÃO/CARRETA faz com que o tempo até uma falha seja aumentado/ desacelerado em $\exp(0.4907156)$ ou 1.6 vezes, em relação a categoria base (aumento de 60%). De forma semelhante, fazer parte da categoria SE (Sudeste) diminui/ acelera o tempo até a falha em $\exp(-0.9421209)$ ou 0.4 vezes, em relação a categoria base (redução de 60%).

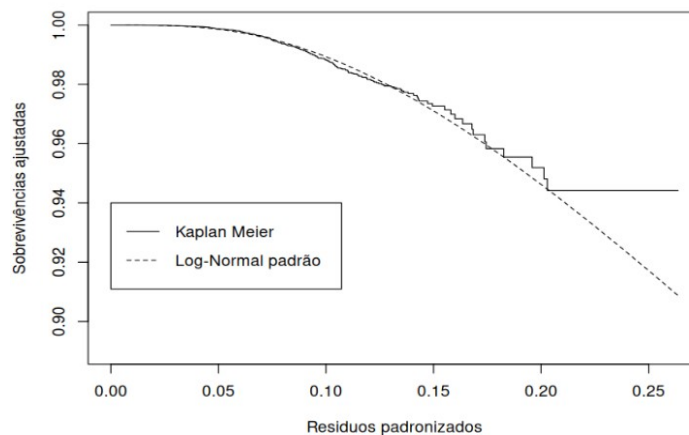
A partir do modelo paramétrico com as duas covariáveis, podem ser construídas, para diferentes configurações de categorias, tabelas de tempos estimados para probabilidades acumuladas de Perda Total, como a construída no tópico 3.3.2. Algumas tabelas de tempos estimados foram construídas e expostas no Apêndice C.

3.3.5 Analisando impacto da natureza do evento

Devido a indícios de que as covariáveis do modelo influenciam o tempo de falha de forma diferente, a depender da natureza dos evento de Perda Total analisados, foram construídos modelos a fim de evidenciar este impacto. Em Análise de Sobrevivência, este tipo de estudo é uma avaliação de Riscos Competitivos. O que será feito a seguir é uma análise de causa específica.

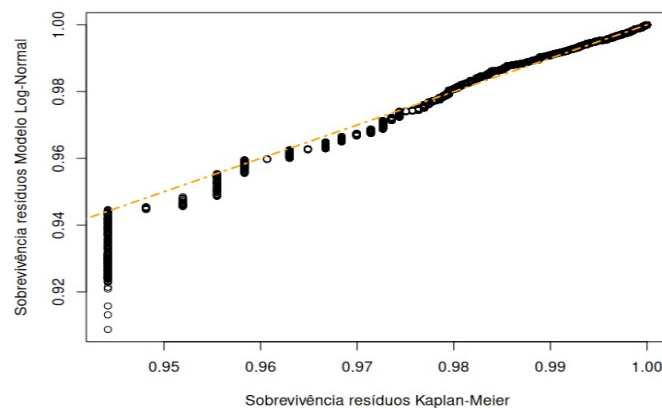
Em um processo análogo ao utilizado para seleção de variáveis do modelo paramétrico apresentado no tópico anterior, foi construído um modelo censurando os eventos de falha cuja natureza difira de Furto/Roubo. A ideia é que o único evento de falha possível seja Furto/Roubo e que os veículos censurados possam passar apenas por este tipo de evento. Vale ressaltar que eventos desta natureza representam aproximadamente 2/3 do total de falhas.

Gráfico 34 - Adequação resíduos do modelo Log-Normal para Furto/Roubo



Fonte: Elaborado pelo Autor (2023)

Gráfico 35 - Probabilidades de sobrevivência dos resíduos Log-Normal e Kaplan-Meier, para Furto/Roubo



Fonte: Elaborado pelo Autor (2023)

Percebe-se que o ajuste foi semelhante ao do tópic anterior. Porém apresentou uma diferença maior dos resíduos nos Gráficos 34 e 35 devido às censuras.

Tabela 8 - Modelo Log-Normal para Furto/Roubo com covariáveis Tipo Veículo e Macrorregião

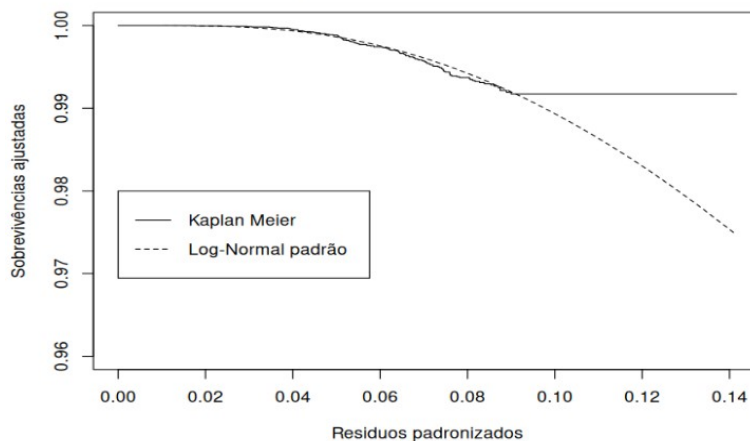
	Estimativa	EP	Estatística de Teste	Valor-p
(Intercepto)	16.50	0.61	26.94	0.00
CAMINHAO/CARRETA	0.95	0.29	3.27	0.00
IMPORTADOS/MICRO ONIBUS	1.92	0.79	2.44	0.01
MOTO	-2.04	0.33	-6.14	0.00
UTILITARIOS	-0.34	0.35	-0.97	0.33
Região NO/CO	1.24	0.68	1.82	0.07
Região SE	-1.29	0.31	-4.21	0.00
Região SU	-1.67	0.35	-4.71	0.00
Log(sigma)	1.38	0.06	25.05	0.00

Fonte: Elaborado pelo Autor (2023)

Para o modelo construído na Tabela 8, acabaram sendo selecionadas as mesmas covariáveis, tendo um impacto diferente de algumas categorias. Na variável Tipo Veículo, a categoria “Moto” demonstrou um impacto muito maior para acelerar o tempo de falha, o que indica ser mais afetada por roubos. O mesmo ocorre com as categorias SE (Sudeste) e SU (Sul) da variável macrorregião.

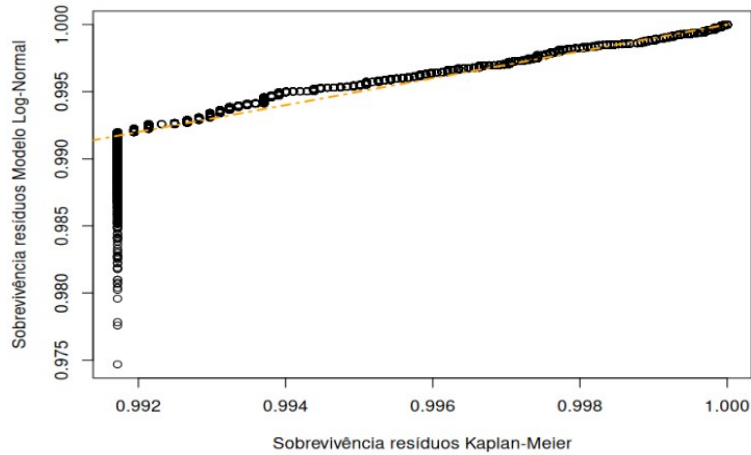
Em seguida, foi construído um modelo censurando os veículos que passaram por um evento de falha cuja natureza seja Furto/Roubo. A ideia é que os eventos de falha pelos quais um veículo possa passar sejam: Capotamento, Colisão, Enchente, Tombamento, Incêndio e outros diferentes de Furto/Roubo. Vale ressaltar que eventos desta natureza representam em torno de 1/3 do total de falhas.

Gráfico 36 - Adequação resíduos do modelo Log-Normal para outros motivos de PT



Fonte: Elaborado pelo Autor (2023)

Gráfico 37 - Probabilidades de sobrevivência dos resíduos Log-Normal e Kaplan-Meier, para outros motivos de PT



Fonte: Elaborado pelo Autor (2023)

Percebe-se o efeito de censuras ainda mais intenso nos Gráficos 36 e 37. Afinal, esta parcela de Perdas Totais é menos representativa do total.

Tabela 9 - Modelo Log-Normal para outras causas de PT com covariáveis Tipo Veículo e Idade Veículo

	Estimativa	EP	Estatística de Teste	Valor-p
(Intercepto)	18.20	0.95	19.08	0.00
CAMINHAO/CARRETA	0.17	0.34	0.52	0.61
IMPORTADOS/MICRO ONIBUS	2.53	1.40	1.80	0.07
MOTO	-0.07	0.69	-0.10	0.92
UTILITARIOS	0.35	0.59	0.59	0.56
Idade Veículo	-0.04	0.02	-1.89	0.06
Log(sigma)	1.50	0.08	19.10	0.00

Fonte: Elaborado pelo Autor (2023)

Mostraram-se significativas apenas as variáveis “Tipo Veículo” e “Idade Veículo”, tendo sido, a segunda, rejeitada nos demais modelos. Além disso, a categoria “Utilitários” da variável Tipo Veículo demonstrou efeito contrário no tempo de falha (desacelerando-o), ainda que o respectivo coeficiente Beta tenha se mostrado pouco significativo.

4. CONSIDERAÇÕES FINAIS

Os modelos de Análise de Sobrevivência mostraram-se adequados à análise de eventos de Perda Total da base de veículos, apresentando-se como uma alternativa para estimar Provisões Atuariais, quando combinada à metodologia de Seguros de Vida. Além disso, possibilita a construção de outros indicadores de gestão de risco, permitindo análises especializadas no perfil do veículo ou de seu dono. Os métodos usados se mostraram ineficientes para estimar a Esperança de Vida Condicional. Na mensuração do efeito de outras variáveis sobre o tempo até uma Perda Total, Tipo Veículo e Macrorregião se mostraram as mais relevantes. Porém, ao desconsiderar Perdas Totais do tipo Furto/Roubo, a variável Idade Veículo passou a ser significativa, o que não ocorreu para Macrorregião.

5. REFERÊNCIAS

- COLOSIMO, Enrico & GIOLO, Suely. *Análise de Sobrevivência Aplicada*. 1.ed. São Paulo, SP: Edgard Blücher, 2006.
- COLOSIMO, Enrico & FREITAS, Marta. *Confiabilidade: Análise de Tempo de Falha e Testes de Vida Acelerados*. Belo Horizonte, MG: Fundação Christiano Ottoni, 1997.
- BOWERS, Newton et al., *Actuarial Mathematics*. The Society of Actuaries, 1997.
- TSE, YIU-KUEN, *Nonlife Actuarial Models: Theory, Methods and Evaluation*. New York (USA) Cambridge University Press, 2009
- CÉSPEDES, Carlos; FOCHEZATTO, Adelar; VELOSO, Leandro. **ANÁLISE DE SOBREVIVÊNCIA DE EMPRESAS: UM ESTUDO LONGITUDINAL DA COORTE DE 2007 NO RIO GRANDE DO SUL**. 2020. v. 35, n. 76, p. 557-579. Artigo –Geosul, Florianópolis, 2020.
- LEOW, Mindy & CROOK, Jonathan. **The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis**. 2014. 8 f. Artigo, Credit Research Centre, University of Edinburgh Business School, Scotland (UK), 2014.
- BERAN, Jan & DJAÏDJA, Abdel-Yazid. **Credit risk modeling based on survival analysis with immunes**. 2006. 26 f. Artigo, Department of Mathematics and Statistics, University of Konstanz, Germany, 2006.
- Gráfico de taxas de falha “Curva da banheira e ciclo de vida de equipamentos”: SELLITTO, Miguel. *Formulação estratégica da manutenção industrial com base na confiabilidade dos equipamentos*. 2005, p.47 Artigo, Revista Produção, Universidade do Vale do Rio dos Sinos – Unisinos, Brasil, 2005.

APÊNDICE A – CURVAS KAPLAN-MEIER

A seguir, as curvas de Kaplan-Meier para cada covariável:

Gráfico 38 - Kaplan-Meier e Log-Rank para todas categorias da variável Sexo

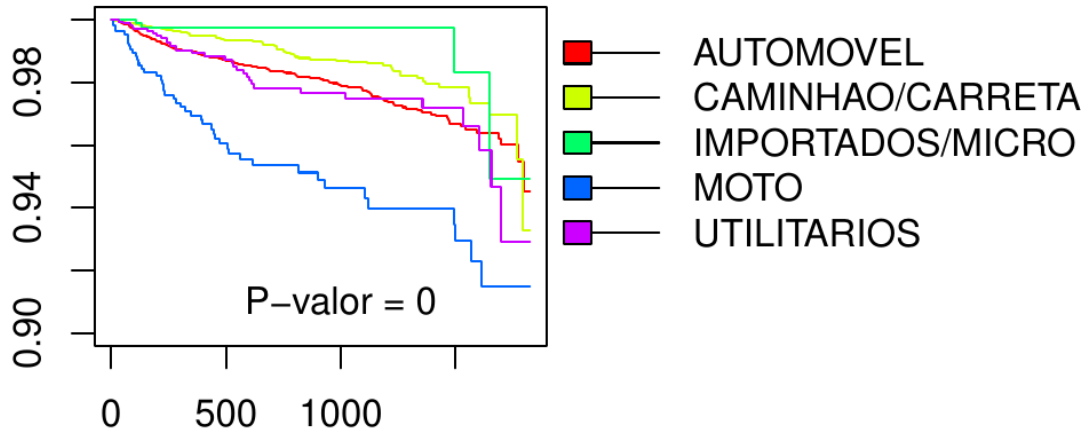


Gráfico 39 - Kaplan-Meier e Log-Rank para todas categorias da variável Macrorregião

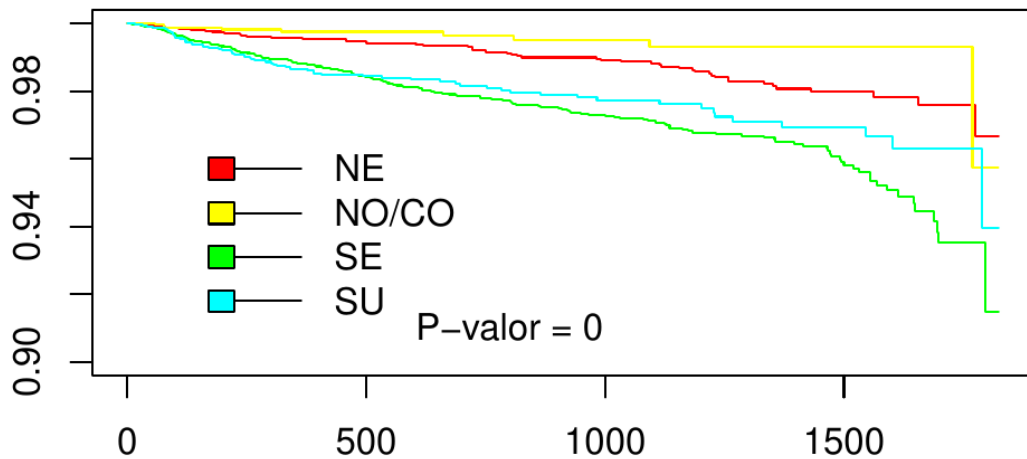


Gráfico 40 - Kaplan-Meier e Log-Rank para todas categorias da variável Cor

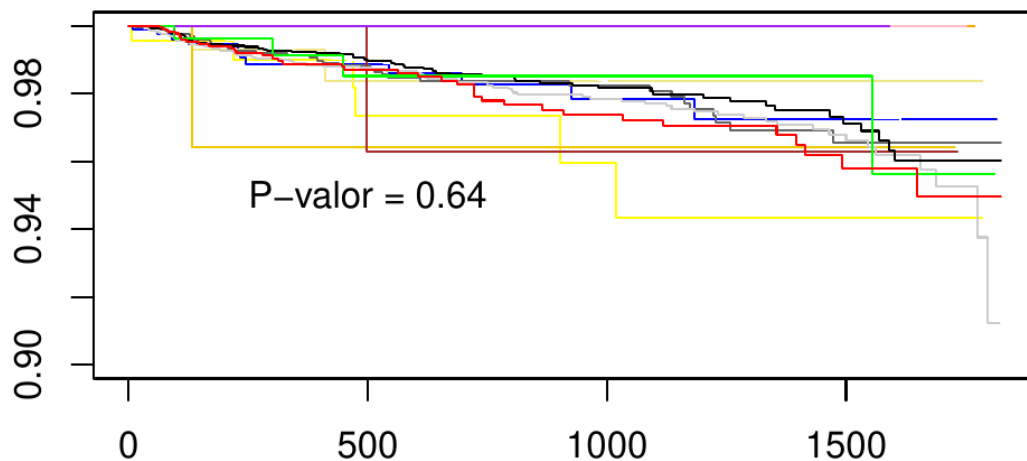


Gráfico 41 - Kaplan-Meier e Log-Rank para todas categorias da variável Valor Fipe

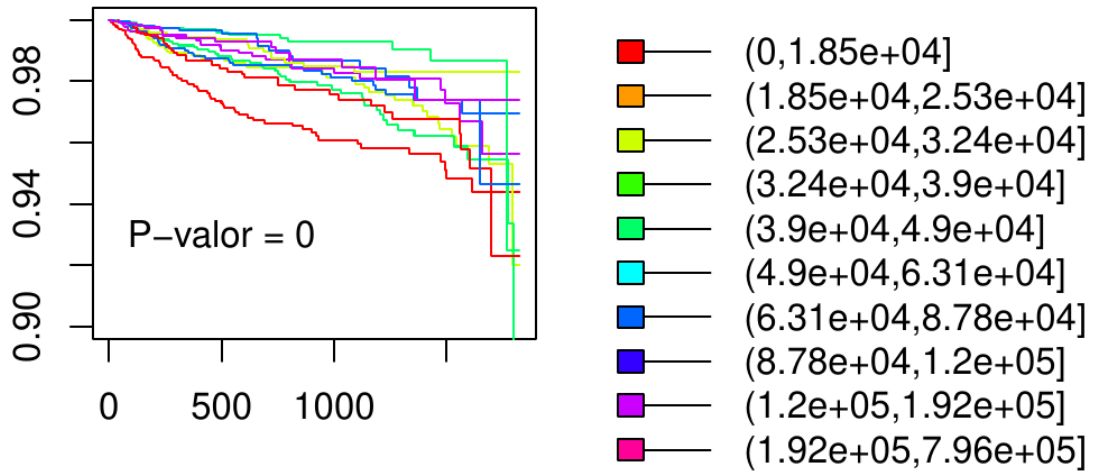


Gráfico 42 - Kaplan-Meier e Log-Rank para todas categorias da variável Idade Veículo

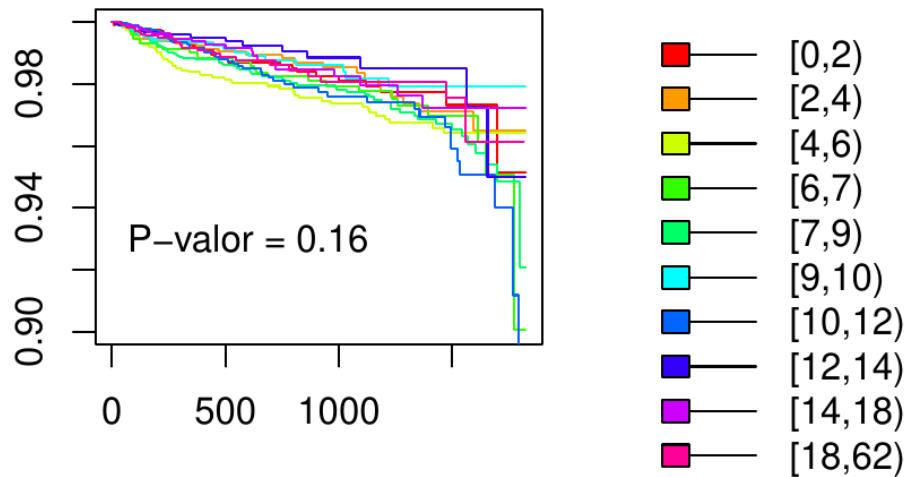
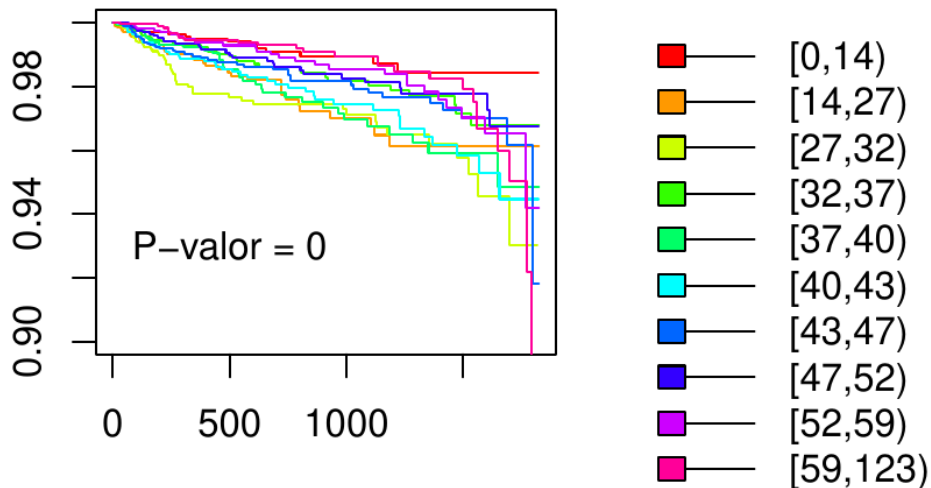


Gráfico 43 - Kaplan-Meier e Log-Rank para todas categorias da variável Idade Associado



APÊNDICE B – ESPERANÇA CONDICIONAL

A partir do modelo paramétrico, foram calculadas esperanças condicionadas ao tempo de contrato em anos. A ideia por trás do conceito é emular as Esperanças de Vida usadas nas tabelas atuariais para cálculos previdenciários e de pensão. Mas ao invés de condicionar o tempo de vida futuro à idade real do veículo, o mesmo foi condicionado ao tempo de contrato. Esta escolha se deveu ao desconhecimento de riscos de falha anteriores ao momento do contrato, por limitações da base de dados, e à baixa significância demonstrada pela variável “Ano Fabricação”, quando desconsideramos a natureza do evento de falha. Vale ressaltar que as tabelas foram construídas desconsiderando três fatores importantes:

- Os veículos automotivos, hoje, são fabricados visando a uma duração média de 10 anos. Podemos considerar que a taxa de falha deveria possuir, portanto, um comportamento de crescimento acelerado para idades próximas a 10 anos. Mas as distribuições de probabilidade consideradas para o estudo não possuem tal propriedade.
- Foram considerados apenas riscos de falha relacionados à Perda Total por motivos de colisão, capotamento, incêndio, fenômenos da natureza e furto. Seria como calcular a expectativa de vida de uma pessoa considerando apenas mortes por doenças infecciosas e fatores externos (acidentes, violência), quando existem outras causas que impactam pessoas mais velhas e demarcam um limite para a longevidade. Isto provocou uma superestimação da medida.
- A distribuição dos tempos de falha é necessariamente assimétrica. Devidos a esta natureza, as estimativas de esperança serão sempre uma medida questionável de centralidade. Por este motivo, costuma-se usar a mediana no campo da Análise de Sobrevivência.

As esperanças condicionais foram calculadas assumindo distribuição Log-Normal, constatada como a mais adequada, usando a função desenvolvida a seguir:

$$E(T|X) = \int_0^{\infty} tpx \cdot dt = \int_0^{\infty} \frac{S(x+t)}{S(x)} \cdot dt$$

$$S(x) = \phi \left(\frac{-\log(x) + \mu}{\sigma} \right); S(x+t) = \phi \left(\frac{-\log(x+t) + \mu}{\sigma} \right)$$

O cálculo da Esperança de Vida condicional (no caso, condicionada ao tempo de contrato em anos) utilizando técnicas de modelagem paramétrica e assumindo distribuição Log-Normal para os tempos até a ocorrência de Perda Total obteve os resultados que se encontram na Tabela 10.

Tabela 10 - Esperança de vida de um veículo em anos condicionada ao tempo contrato

Tempos de contrato em anos (x)	Esperança de Vida em anos ($E(T x)$)
0	25632084
2	26079911
4	26318091
6	26503092
8	26659140
10	26796212

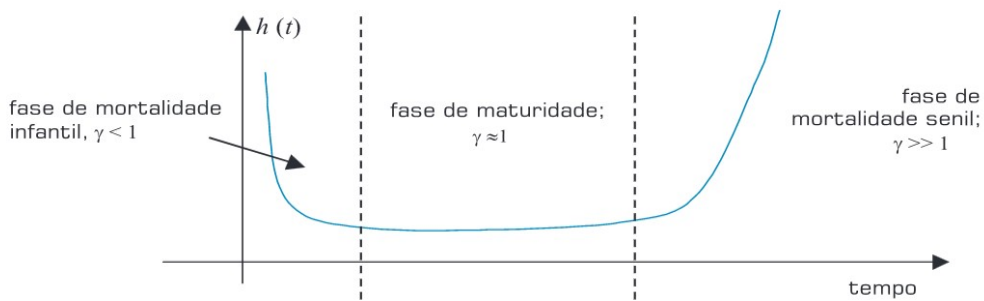
Fonte: Elaborado pelo Autor (2023)

O resultado obtido para as esperanças condicionais é de valores inflados e que aumentam quanto maior é o tempo de contrato vivido antecipadamente pelo veículo, o que se pode dizer contraintuitivo. O aumento se deve ao comportamento da distribuição Log-Normal, que possui taxa de falha decrescente no tempo, ou seja, quanto maior o tempo previamente vivido, menor é o risco de Perda Total e, por isto, maior é a esperança do tempo futuro. As outras duas distribuições consideradas no trabalho também teriam suas limitações. Para a distribuição Exponencial, a esperança do tempo até a Perda Total é independente do tempo previamente vivido, uma vez que sua taxa de falha é constante. Logo, a Esperança também seria constante (e igual a 125,32 anos, segundo o ajuste feito). Já a distribuição Weibull demonstra valores decrescentes em relação ao tempo de contrato apenas quando seu parâmetro gama é maior que 1. Quando é igual a um, possui o mesmo comportamento da Exponencial. Já quando o parâmetro gama é menor que um (o que se mostrou no ajuste da Weibull para os dados da APV), os valores de esperança também crescem em relação ao tempo de contrato.

Como já foi observado, as ferramentas usadas são insuficientes para a estimação da esperança condicional. Além de não serem consideradas todas as circunstâncias de depreciação que levam a uma Perda Total, podemos supor que a incidência da taxa de falha da Log-Normal não parece corresponder à realidade, em tempos superiores a 5 anos (o limite da base de dados). Talvez a distribuição Log-Normal só tenha demonstrado um bom ajuste pela ausência de dados longitudinais com tempos suficientemente grandes. Se considerarmos um veículo como um conjunto complexo de componentes que podem, a partir de um determinado

momento, começar a falhar simultaneamente devido ao desgaste, poderíamos considerar que o desenvolvimento da taxa de falha no tempo é semelhante à taxa de falha de seres humanos (também chamada força de mortalidade). O ritmo de queda da taxa de falha da Log-Normal (Gráfico 45) demonstrou descrever adequadamente o tempo até Perdas Totais para veículos em momentos iniciais (até 5 anos, tempo máximo encontrado na base de dados). Mas é razoável supor que, a partir de um dado momento, a taxa de falha volte a crescer. A curva de banheira para taxas de falha tem sido usada no estudo da durabilidade de alguns equipamentos e componentes elétricos, como é apresentado no Gráfico 44.

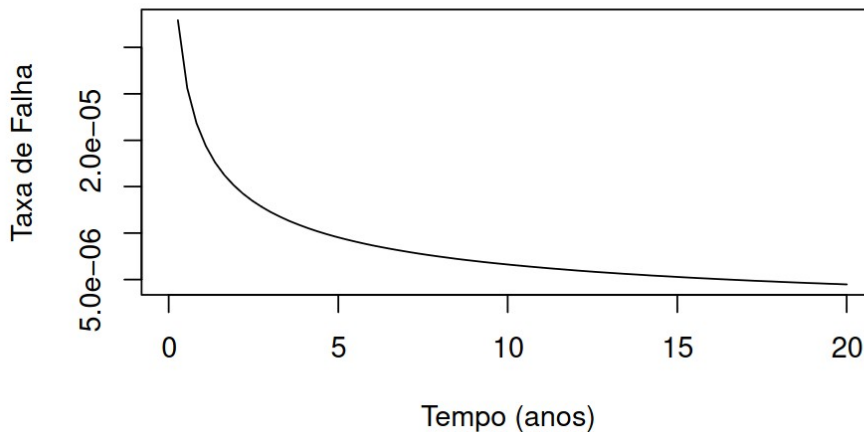
Gráfico 44 - Curva da banheira e ciclo de vida de equipamentos



Fonte: Elaborado por Sellitto, Miguel, 2005

A taxa de falha da Log-Normal, por outro lado possui um desenvolvimento que provavelmente subestima as falhas em idades avançadas, como mostra o Gráfico 45.

Gráfico 45 - Função taxa de falha para o ajuste Log-Normal



Fonte: Elaborado pelo Autor (2023)

O indicador obtido não possui utilidade prática para a gestão de risco, mas espera-se que esta tentativa possa ser reavaliada por outros trabalhos, com o uso de outros pressupostos.

APÊNDICE C – TEMPO ESTIMADO PARA PROBABILIDADES DE PERDA TOTAL COM COVARIÁVEIS

Fazendo estimativas dos tempos até a Perda Total de p% de veículos, considerando as covariáveis do modelo paramétrico, obteve-se os seguintes resultados, para a combinação das categorias “Nordeste”, “Sul”, “Automóvel” e “Moto”:

Tabela 11- Automóveis do Nordeste: Tempos estimados para Perda Total de p% dos veículos

Percentual de Falha	Tempo em meses
0.1%	0.95
0.2%	2.17
0.3%	3.60
0.4%	5.23
0.5%	7.03
0.6%	9.01
0.7%	11.16
0.8%	13.47
0.9%	15.94
1%	18.57

Fonte: Elaborado pelo Autor (2023)

Tabela 12- Motos do Nordeste: Tempos estimados para Perda Total de p% dos veículos

Percentual de Falha	Tempo em meses
0.1%	0.20
0.2%	0.45
0.3%	0.75
0.4%	1.09
0.5%	1.47
0.6%	1.88
0.7%	2.33
0.8%	2.81
0.9%	3.32
1%	3.87

Fonte: Elaborado pelo Autor (2023)

Tabela 14- Automóveis do Sul: Tempos estimados para Perda Total de p% dos veículos

Percentual de Falha	Tempo em meses
0.1%	0.33
0.2%	0.75
0.3%	1.25
0.4%	1.81
0.5%	2.43
0.6%	3.11
0.7%	3.86
0.8%	4.65
0.9%	5.51
1%	6.42

Fonte: Elaborado pelo Autor (2023)

Tabela 13- Motos do Sul: Tempos estimados para Perda Total de p% dos veículos

Percentual de Falha	Tempo em meses
0.1%	0.07
0.2%	0.16
0.3%	0.26
0.4%	0.38
0.5%	0.51
0.6%	0.65
0.7%	0.80
0.8%	0.97
0.9%	1.15
1%	1.34

Fonte: Elaborado pelo Autor (2023)