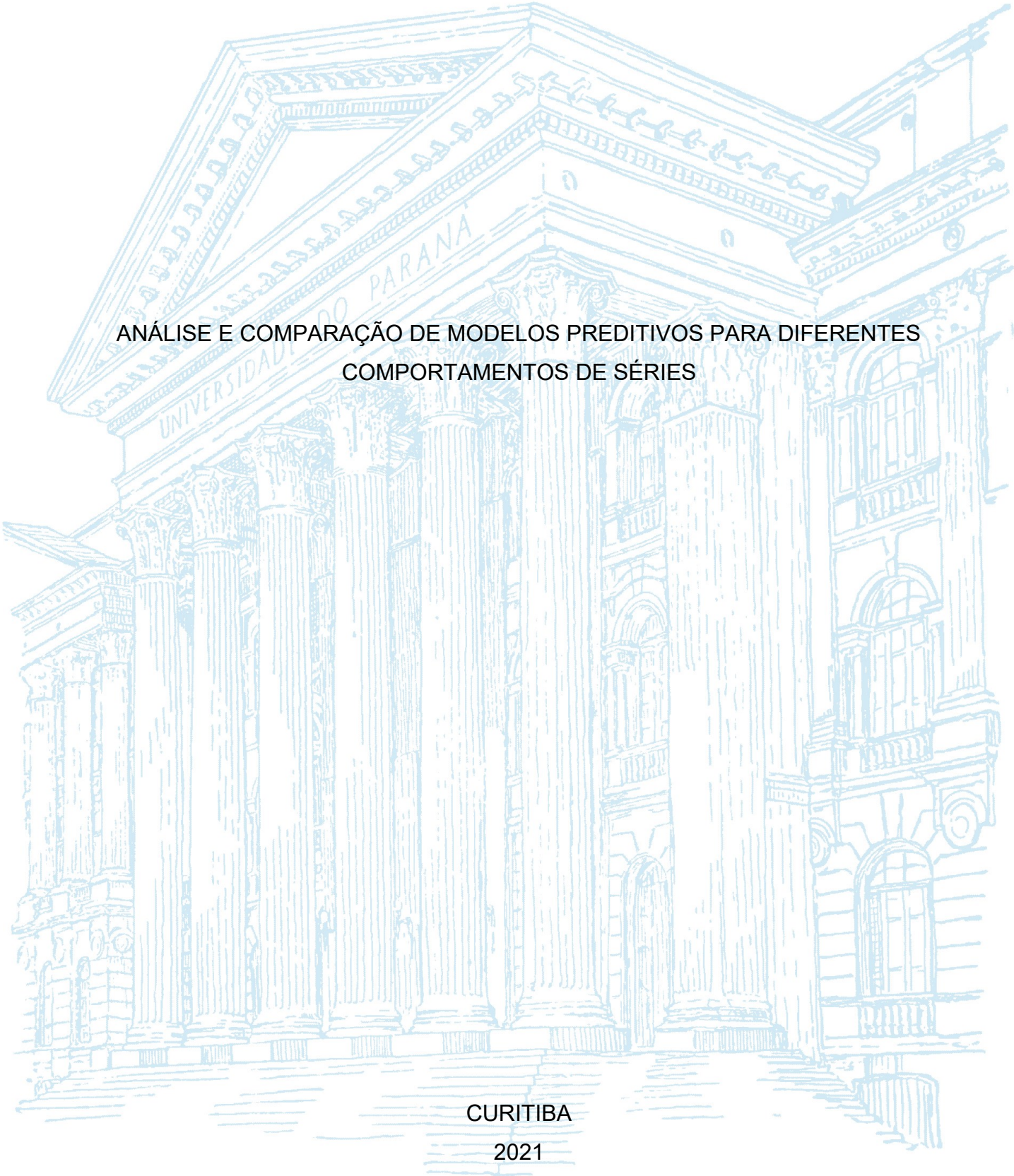
UNIVERSIDADE FEDERAL DO PARANÁ

PIETRO ROEHRIG MEGGETTO



ANÁLISE E COMPARAÇÃO DE MODELOS PREDITIVOS PARA DIFERENTES COMPORTAMENTOS DE SÉRIES

CURITIBA

2021

PIETRO ROEHRIG MEGGETTO

ANÁLISE E COMPARAÇÃO DE MODELOS PREDITIVOS PARA DIFERENTES COMPORTAMENTOS DE SÉRIES

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Produção da Universidade Federal do Paraná como requisito à obtenção do título de obtenção do grau de Engenheiro de Produção

Orientadora: Profa. Dra. Mariana Kleina

CURITIBA

2021

RESUMO

Dentre os inúmeros métodos preditivos existentes decidir qual o melhor não é uma tarefa fácil e, com as demandas do mercado cada vez mais rápidas e exigentes, essa escolha deve ser tomada o mais rápido e assertivamente possível. Nesse estudo foi realizada a comparação entre três diferentes modelos preditivos (ARIMA, GLM e Redes Neurais) com o intuito de avaliar seu desempenho com diferentes séries de dados. Foram considerados, para a avaliação dos resultados, a estacionariedade e a volatilidade das séries de dados utilizadas, duas medidas de erro, o volume de dados utilizados para o treinamento dos modelos e a quantidade de variáveis regressoras utilizadas para a previsão. Os resultados mostram uma maior consistência nos resultados obtidos com o método de Redes Neurais, não apresentando grandes erros em nenhuma das séries avaliadas nesse estudo. Constatou-se também que, quando existe uma grande quantidade de informação os modelos gerados tem um erro maior que quando treinados utilizando apenas dados mais recentes. Por fim, concluiu-se que, dentro das restrições desse estudo, Redes Neurais teve um desempenho superior aos demais métodos, mas que é necessário realizar outros estudos para explorar outros métodos e parâmetros.

Palavras-chave: Modelos Preditivos. ARIMA. GLM. Redes Neurais.

ABSTRACT

Among the innumerable existing predictive methods determine which one is the most appropriate is not an easy task and, with the market demand turning faster and exigent, that choice should be made fastest and assertive as possible. This paper intends to compare three different forecast methods (ARIMA, GLM and Neural Network) in order to appraise their performance with different data series. For the results analysis this paper considerate the stationarity and volatility of the data series, two different error measures, the amount of data considered for the training and the number of explanatory variables in the models. The results shows that the Neural Network method tends to be more stable than the other two, due to the absence of large errors in the prediction of the analyzed series. In addition, it was found out that the big amount of information given to train the models provoke bigger errors in the prediction than the models trained using latest information. Finally, considering the restrains of this paper, the Neural Network models had a superior performance than the other methods, but further studies are necessary in behalf of the countless number of prediction methods and parameters existent.

Keywords: Forecasting Methods. ARIMA. GLM. Neural Network.

SUMÁRIO

1. INTRODUÇÃO

Planejamento e organização são palavras-chave para o sucesso de muitas empresas, já que, com elas pode-se traçar metas mais claras, encontrar possíveis pontos de melhoria e falhas no processo produtivo com mais facilidade, aumentando, assim, sua competitividade no mercado. Segundo Dumas em tempos de crise, como o enfrentado durante a pandemia do COVID 19, o planejamento estratégico pode fazer com que as organizações se mantenham pro ativas e evitar tomada de decisão influenciadas pelo cenário atual. Por isso é importante manter um planejamento que permita a análise de diversos cenários críticos, permitindo a criação de planos de ação caso eles se concretizem. (DUMAS, 2020)

Para um bom planejamento é necessário ter uma boa fonte de informação, possibilitando uma tomada de decisão mais rápida e acurada, o que, com o aumento do uso de ferramentas da indústria 4.0 - como o Big Data, Inteligência Artificial e Internet das Coisas -, passou a ser um dos grandes desafios dentro de uma empresa visto que quanto mais rápida e precisa for a obtenção dos dados, mais competitiva a empresa se tornará.

Assim, a busca por novos métodos preditivos, com menos limitações é constante, aumentando cada vez mais as opções disponíveis para realizar uma previsão. Esse grande número de alternativas para realizar uma predição e a grande diversidade de conjuntos de dados podem dificultar a escolha do método mais adequado.

Segundo dados apresentados no Boletim IRB + Mercado, iniciativa do IRB Brasil RE, o mercado segurador apresentou, em fevereiro de 2021, faturamento 13,1% maior que o mesmo período de 2020. Com esse crescimento é necessário que as seguradoras se atentem cada vez mais aos riscos a que estão expostas e fortaleçam seu planejamento para conseguir atender a demanda do mercado em ascensão. Para seguradoras é de suma importância que se tenha um modelo preditivo bem consolidado, já que é um setor exposto a diversos riscos, e quanto mais bem construídas e consolidadas forem as previsões de cenários realizadas, mais fácil será a tomada de decisão e o planejamento de risco. (IRB BRASIL)

1.1. Objetivos do Trabalho

Nessa seção serão apresentados os objetivos do estudo.

1.1. Objetivo Geral

O presente estudo tem como objetivo comparar diferentes métodos preditivos para comportamentos diferentes de séries de dados do mercado de seguros e utilizando variáveis macroeconômicas a fim de avaliar qual método é mais adequado para cada situação.

1.2. Objetivos Específicos

Durante a análise avaliar fatores como:

- Volatilidade da série;
- Volume de dados utilizados para realizar o treinamento;
- Estacionariedade da série;
- Quantidade de variáveis regressoras;
- Quais os fatores limitantes de cada modelo;
- As dificuldades para ajuste.

1.2. Limitações

Para o presente estudo existem alguns pontos que devem ser levados em consideração na análise dos resultados obtidos:

- As bases de dados utilizadas são públicas, ou seja, existe uma limitação de informação que pode acabar dificultando o ajuste de alguns dos modelos testados.
- Por conta da grande diversidade de métodos preditivos existentes não serão avaliadas todas as possibilidades existentes.
- Além da grande diversidade de métodos, existe, para cada um deles, uma grande quantidade de parâmetros a serem testados, assim apenas alguns serão avaliados.

2. REVISÃO DE LITERATURA

Durante esse capítulo serão abordados os conceitos de cada um dos modelos preditivos que serão testados nesse estudo, bem como os testes que devem ser realizados para a validação de cada um deles. Além disso serão retratados os critérios a serem utilizados para a comparação dos resultados dos modelos gerados.

2.1. Modelos Preditivos

Nessa seção serão apresentados os modelos preditivos que serão avaliados nesse trabalho: Regressão Linear Múltipla, Modelos Lineares Generalizados, ARIMA e Redes Neurais Artificiais.

2.1.1. Regressão Linear Múltipla

O método consiste em encontrar os valores de β que minimizem a soma do quadrado dos resíduos da equação 1.

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_3\beta_3 + \ \dots + x_{ik}\beta_k + \ \epsilon_i \qquad (1)$$

Onde:

x_i são as variáveis regressoras de cada uma das i observações, com $i = 1, \dots, n$;

y_i é o resultado do modelo preditivo para cada uma das i observações, com $i = 1, \dots, n$;

β_k são os coeficientes parciais de cada uma das k variáveis regressoras;

ϵ_i são os resíduos de cada uma das i observações, com $i = 1, \dots, n$.

Para isso é feita uma estimação dos β utilizando o método dos Mínimos Quadrados e, para facilitar os cálculos, será utilizada a forma matricial da equação 1, que é dado pela equação 2. (BARBOSA, 2010; ESTATCAMP, 2021)

$$\underline{y} = \boldsymbol{X}\underline{\beta} + \ \underline{\epsilon} \qquad (2)$$

Onde:

$\boldsymbol{X}$ é a matriz com as variáveis regressoras;

$\underline{y}$ é o vetor resultado;

$\underline{\beta}$ é o vetor dos coeficientes parciais de regressão;

$\underline{\epsilon}$ é o vetor de resíduos.

Para obter-se a estimação de β resolve-se a equação 3:

$$\underline{\hat{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'\underline{y} \qquad (3)$$

2.1.1.1. Testes a serem realizados

Depois de se ter os β não é possível afirmar que o modelo gerado é aplicável, visto que o resultado da regressão pode não ser válido por não atender algum desses critérios:

- Existir alguma variável regressora que contribui significativamente para o modelo;
- Os resíduos seguem uma distribuição normal com média 0;
- Há variância constante entre os resíduos;
- Não existe colinearidade entre as variáveis regressoras.

Tendo isso em vista é necessário fazer uma série de testes para validar os resultados encontrados. (ESTATCAMP, 2021)

2.1.2. Modelos Lineares Generalizados

Uma extensão ao modelo de regressão linear tradicional são os Modelos Lineares Generalizados (GLM), introduzidos por Nelder e Wedderburn (1972), e tem como diferencial a possibilidade de utilizar funções de ligação que permitam que a variável resposta (y) assuma uma distribuição de erro diferente da normal.
Esse modelo é definido por três componentes: o componente aleatório, o componente sistemático e a função de ligação. (PAULA, 2013; TURKMAN, 2000; CORDEIRO & DEMÉTRIO, 2013; NELDER & WEDDERBURN, 1972; ROCHA, 2015)

2.1.2.1. Componente aleatório

Conjunto de variáveis aleatórias e independentes (y_i) que tenham distribuição pertencente à família exponencial de dispersão. Ou seja, assume-se que a função de probabilidades de y_i pode ser escrita como apresentado na equação 4. (PAULA, 2013)

$$f(y_i, \theta_i, \Phi) = exp\left\{\frac{y_i * \theta_i - b(\theta_i)}{a(\Phi)} + c(y_i, \Phi)\right\} \quad (4)$$

Onde:
y_i é o resultado de cada uma das i observações, com $i = 1, \dots, n$;
θ_i é o parâmetro canônico de cada uma das i observações, com $i = 1, \dots, n$;
Φ é o parâmetro de dispersão da distribuição.

2.1.2.2. Componente sistemático

É a uma combinação linear das variáveis explicativas, tem como objetivo encontrar valores de β que minimizem o erro na equação: (PAULA, 2013)

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_3\beta_3 + \ \dots + x_{ik}\beta_k + \ \epsilon_i \quad (5)$$

Onde:

x_i são as variáveis regressoras de cada uma das i observações, com $i = 1, \dots, n$;

y_i é o resultado de cada uma das i observações, com $i = 1, \dots, n$;

β_k são os coeficientes parciais de cada uma das k variáveis regressoras;

ϵ_i são os resíduos de cada uma das i observações, com $i = 1, \dots, n$.

2.1.2.2.1. Funções de Ligação

São funções que permitem a associação do componente aleatório ao componente sistemático, essa função deve ser real, monótona e diferenciável. A TABELA 1 apresenta as funções de ligação que serão testadas neste estudo considerando as características da base a ser estudada. (TURKMAN, 2000; CORDEIRO & DEMÉTRIO, 2013; NELDER & WEDDERBURN, 1972)

TABELA 1 – FUNÇÕES DE LIGAÇÃO

Ligação	Função de ligação	Função inversa	Tipo de ligação
Identidade	$\underline{\mu}$	$\underline{\eta}$	Normal
Inversa	$\underline{\mu}^{-1}$	$\underline{\eta}^{-1}$	Gama
Inversa quadrada	$\underline{\mu}^{-2}$	$\underline{\eta}^{-2}$	Normal inversa

FONTE: O Autor (2021)

Onde

$\underline{\eta}$ é o vetor das variáveis resposta;

$\underline{\mu}$ é o vetor de valores estimados pelo modelo.

2.1.3. ARIMA

O modelo autorregressivo integrado de médias móveis (ARIMA), consiste na integração de três partes: a parte autorregressiva (AR), a parte integrada (I) e, por fim, a parte de médias móveis (MA). (HYNDMAN, 2021)

2.1.3.1. Parte Integrada (I)

Com o intuito de manter a estacionariedade da série é realizada a diferenciação da série, que consiste na substituição da variável resposta pela variação entre duas observações consecutivas. A diferenciação de um período pode ser escrita como exposto na equação 6. *(*KUMAR *et al., 2020)*

$$y'_t = y_t - y_{t-1} \quad (6)$$

Assim, uma diferenciação de ordem d pode ser escrita com notação *backward shift*, como apresentado na equação 7.

$$y_t^d = (1 - B)^d y_t \quad (7)$$

Onde

y_t é a variável resposta no período t;

B é o operador *backward shift*, ($By_t = y_{t-1}$*);*

d é o número de diferenciações a serem feitas.

2.1.3.2. Autorregressão (AR)

O componente de autorregressão consiste na previsão da variável resposta por meio da combinação linear dos valores históricos, ou seja, utiliza o histórico (de p períodos anteriores) da variável resposta para prever seus valores no futuro. Assim um modelo autorregressivo de ordem p pode ser escrito como na equação 8. (HYNDMAN, 2021)

$$y_t = c + \Phi_1 * y_{t-1} + \cdots + \Phi_j * y_{t-j} + \varepsilon_t \quad (8)$$

Onde:

c é uma constante;

y_t é a variável resposta no período t;

y_{t-j} é a variável resposta no período anterior j, com $j = 1, \ldots, p$;

Φ_j é o peso dado a cada período anterior j, com $j = 1, \ldots, p$;

ε_t é o ruído branco no período t;

2.1.3.3. Médias Móveis (MA)

O componente de médias móveis consiste na previsão da variável resposta por meio da combinação linear dos ruídos brancos (ε_t) históricos, ou seja, utiliza o histórico (de q períodos anteriores) do ruído branco para prever seus valores no

futuro. Assim um modelo de médias móveis ordem q pode ser escrito como na equação 9. (HYNDMAN, 2021)

$$y_t = c + \varepsilon_t + \theta_1 * \varepsilon_{t-1} + \cdots + \theta_k * \varepsilon_{t-k} \quad (9)$$

Onde:

c é uma constante;

y_t é a variável resposta no período t;

ε_t é o ruído branco no período t;

ε_{t-k} é o ruído branco para cada período anterior k, com $k = 1, \dots, q$.

θ_k é o peso dado a cada período anterior k, com $k = 1, \dots, q$;

2.1.3.4. Modelo Completo

Assim, ao agrupar as três componentes do modelo tem-se o modelo $ARIMA(p, d, q)$, que pode ser escrito como apresentado na equação 10. *(PINHO, 2019)*

$$\phi(B)(1 - B)^d y_t = c + \theta_t(B)\varepsilon_t \quad (10)$$

Onde:

c é uma constante;

B é o operador *backward shift*, ($By_t = y_{t-1}$);

y_t é a variável resposta no período t;

Φ_t é o peso dado a cada período t, com $t = 1, \dots, p$;

θ_t é o peso dado a cada período *t* , com $t = 1, \dots, q$;

ε_t é o ruído branco no período t, com $t = 1, \dots, p$.

2.1.4. Redes Neurais Artificiais

O modelo de Redes Neurais Artificiais consiste em um modelo inspirado nos neurônios humanos, e possui três elementos principais: a camada de entrada, a camada escondida e a camada de saída.

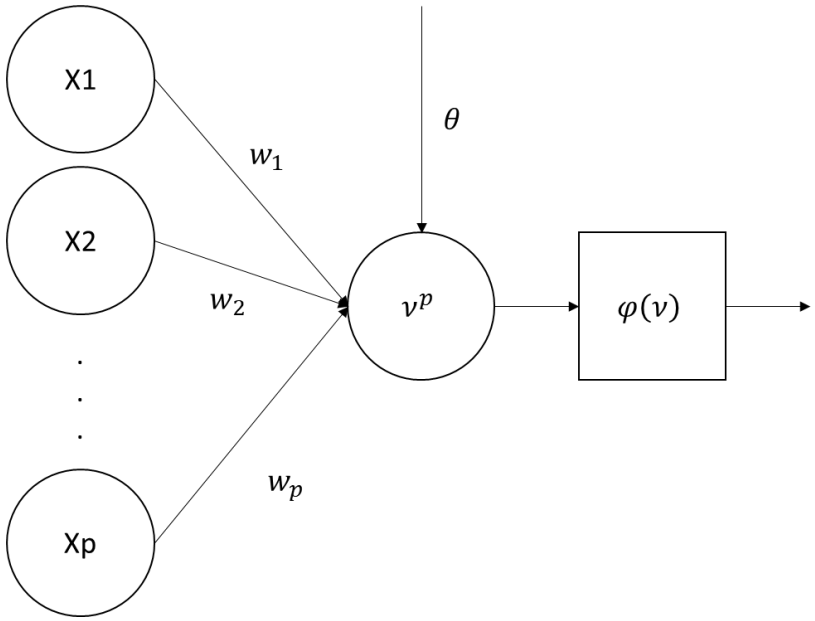
Assim como os neurônios no sistema nervoso, que são conectados uns aos outros para transmitir informações, essas camadas são conectadas por pesos e *bias* que tem a função de transformar as informações recebidas em uma resposta.

Existem diversas estruturas de redes neurais, para este estudo será utilizado o algoritmo Perceptron, que é um dos algoritmos mais simples de redes neurais e comumente utilizado na literatura. (POZZOLO, 2010)

2.1.4.1. Perceptron

Esse método consiste no aprendizado supervisionado do modelo, e utiliza o método *backpropagation* – dividido em duas etapas, a fase *forward* e a fase *backward* - para o seu treinamento. Possui três principais elementos: os pesos sinápticos, os *bias* e a função de ativação, além disso possui duas camadas (a de entrada e a de saída). Durante o processo de treinamento os valores dos pesos são atualizados a cada iteração com base no erro apresentado entre o valor desejado e o valor calculado, até que algum dos critérios de parada seja atendido – erro mínimo ou número máximo de iterações. Um exemplo da estrutura de uma rede Perceptron é apresentado na FIGURA 1. (SIQUEIRA, 2014)

FIGURA 1 - EXEMPLO REDE PERCEPTRON



FONTE: O Autor (2021)

2.1.4.1.1. Fase forward

Durante essa etapa são calculados os valores de saída da rede, ou seja, é realizada uma estimação de resultado utilizando os valores dos *bias* e pesos daquela iteração, assim como exposto na equação 11.

$$v^p = \sum w_j \, x_j^p + \theta \tag{11}$$

Onde:

v^p é o valor de saída calculado para a observação p;

w_j é o peso dado a variável explicativa j;

x_j^p é valor da variável explicativa j para a observação p;

θ é o *bias*.

2.1.4.1.2. Função de ativação

Utilizada para evitar acréscimos progressivos, limita a saída do neurônio a um determinado intervalo e introduz a não-linearidade no modelo. As mais utilizadas são as apresentadas na TABELA 2. *(*PINHO, 2019*;* SAMPAIO et al., 2019)

TABELA 2 – FUNÇÕES DE ATIVAÇÃO

Função	**Equação**
Sigmoide Logística	$\varphi(v) = \frac{1}{1+e^{-v}}$
Tangente Hiperbólica	$\varphi(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}}$
Gaussiana	$\varphi(v) = e^{\frac{-(v-\mu)^2}{2\sigma^2}}$
Linear	$\varphi(v) = \alpha v, \alpha \epsilon \mathbb{R}$

FONTE: O Autor (2021)

Onde:

φ é a função de ativação;

v é o valor de saída calculado;

μ é o centro da função de ativação;

σ é o desvio padrão de μ em relação a v.

2.1.4.1.3. Fase backward

É nessa etapa que os valores dos pesos sinápticos e *bias* vão sendo atualizados com base no erro resultante da fase *forward*. As equações 12, 13 e 14 apresentam os cálculos como é feita essa atualização. (SIQUEIRA, 2014)

$$\delta^p = d^p - \varphi(v^p) \tag{12}$$

$$\overline{w_j} = w_j + \gamma x_j^p \delta^p \tag{13}$$

$$\bar{\theta} = \theta + \gamma \delta^p \tag{14}$$

Onde:

δ^p é o erro calculado para a observação p;

d^p é o valor real da observação p;

v^p é o valor de saída calculado para a observação p;

φ é a função de ativação;

w_j é o peso da variável explicativa j;

$\overline{w_j}$ é o peso atualizado da variável explicativa j;

x_j^p é valor da variável explicativa j para a observação p;

θ é o valor do *bias*;

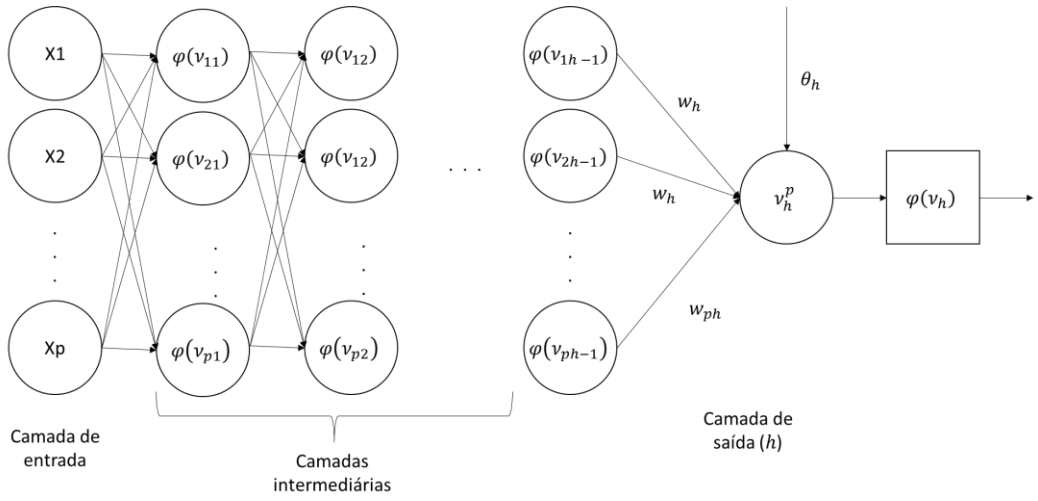
$\bar{\theta}$ é o valor atualizado do *bias;*

γ é a taxa de aprendizagem do modelo.

2.1.4.2. Multilayer Perceptron

Assim como o Perceptron, o Multilayer Perceptron possui três principais elementos: os pesos sinápticos, os *bias* e a função de ativação, mas apresenta como diferença a existência de camadas ocultas além das camadas de entrada e de saída, a estrutura desse modelo de rede é apresentado na FIGURA 2. O processo de treinamento continua o mesmo, os valores dos pesos são atualizados a cada iteração com base no erro apresentado entre o valor desejado e o valor calculado, até que algum dos critérios de parada seja atendido – erro mínimo ou número máximo de iterações, porém a complexidade dos cálculos aumenta por conta da inserção de novas camadas. (SIQUEIRA, 2014)

FIGURA 2 – EXEMPLO MULTILAYER PERCEPTRON



Fonte (O Autor, 2021)

2.1.4.2.1. Fase forward

Durante essa etapa são calculados os valores de saída da rede, ou seja, é realizado uma estimação de resultado utilizando os valores dos *bias* e pesos daquela iteração, como apresentado na equação 15. (SIQUEIRA, 2014)

$$v_i^p = \sum w_{ij}^p \, x_j^p + \theta_i \tag{15}$$

Onde:

v_i^p é o valor de saída calculado para a observação p na camada i, com $i = 1, \dots, h$;

w_{ij}^p é o peso dado a variável explicativa j na camada i para a observação p, com $i = 1, \dots, h$;

x_j^p é valor da variável explicativa j para a observação p;

θ_i é o *bias* da camada i, com $i = 1, \dots, h$;

h é o índice da camada de saída.

2.1.4.2.2. Fase Backward

É nessa etapa que os valores dos pesos sinápticos e *bias* vão sendo atualizados com base no erro resultante da fase *forward*. Primeiramente são atualizados os pesos e *bias* da última camada, a camada de saída. Para a última camada o erro resultante da fase *forward* é calculado pela equação 16, já para as demais camadas calcula-se o erro pela equação 17 e a atualização dos pesos e *bias* são dadas pelas equações 18, 19, 20 e 21.

(SIQUEIRA, 2014)

$$\delta_h^p = (d^p - \varphi(v_h^p)) \, \varphi(v_h^p) \, (1 - \varphi(v_h^p)) \tag{16}$$

$$\delta_i^p = \varphi(v_i^p) \left(1 - \varphi(v_h^p)\right) \delta_{i+1}^p w_{ij}^p \tag{17}$$

$$\Delta w_{hi}^p(k) = \gamma \varphi(v_{i-1}^p) \, \delta_i^p + \alpha \Delta w_{hi}^p(k-1) \tag{18}$$

$$\Delta \theta_i^p(k) = \gamma \delta_i^p + \alpha \Delta \theta_i^p(k-1) \tag{19}$$

$$\overline{w_{ij}} = w_{hi}^p + \Delta w_{hi}^p(k) \tag{20}$$

$$\overline{\theta_i} = \theta_i^p + \Delta \theta_i^p(k) \tag{21}$$

Onde:

δ_h^p é o erro calculado para a observação p na camada de saída h;

δ_i^p é o erro calculado para a observação p na camada i, com $i = h - 1, \dots, 1$;

d^p é o valor real da observação p;

v_h^p é o valor de saída calculado para a observação p na camada h;

v_i^p é o valor de saída calculado para a observação p na camada i, com $i = h, \dots, 1$;

φ é a função de ativação;

w_{ij} é o peso do neurônio j na camada i;

$\overline{w_{ij}}$ é o peso atualizado da variável explicativa j na camada i;

θ_i é o valor do *bias* na camada i;

$\overline{\theta_i}$ é o valor atualizado do *bias* na camada i;

γ é a taxa de aprendizagem do modelo;

α é o fator de momento do modelo.

2.2. Testes de validação

Nessa seção serão expostos os testes estatísticos utilizados para a validação das premissas dos modelos de regressão utilizados.

2.2.1. Teste de Normalidade dos resíduos (Shapiro-Wilk)

Para que se tenha um resultado confiável da regressão é necessário que os resíduos se ajustem a uma distribuição normal. Para avaliar a normalidade dos resíduos será utilizado o teste de Shapiro-Wilk.

Para realizar o teste calcula-se a estatística W (dada pela equação 22) e compara-se esse valor ao valor tabelado $w_{n;\alpha}$, se o valor calculado for menor que o tabelado, rejeita-se a hipótese de normalidade ao nível α de significância. (MOHD & YAP, 2011)

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \quad (22)$$

Onde:

a_i são os coeficientes tabelados, com $i = 1, \dots, n$;

y_i é a variável i observada, com $i = 1, \dots, n$;

$\overline{y}$ é a média das i variáveis y observadas.

2.2.2. Teste de homocedasticidade (Goldfeld-Quandt)

Para verificar se os resíduos são homocedásticos, ou seja, as suas variâncias são constantes realiza-se o teste de Goldfeld-Quandt.

Para realizar esse teste é necessário ordenar as variáveis regressoras de acordo a variável que se acredita ser responsável pela heterocedasticidade do modelo e, em seguida, dividi-las em três grupos, sendo que o segundo grupo deve conter 20% do total de observações. (ESTATCAMP, 2021)

Após a divisão, descarta-se a segunda parcela da seleção e calcula-se a estatística de teste (dado pela equação 23). (ESTATCAMP, 2021)

$$FCG = \frac{\frac{SQE^b}{(n_3-(p-1))}}{\frac{SQE^a}{(n_1-(p-1))}} \quad (23)$$

Onde

n_1 é o número de observações na primeira parcela de divisão;

n_3 é o número de observações na terceira parcela de divisão;

SQE^a é a soma de quadrados dos resíduos da regressão da primeira parcela de divisão;

SQE^b é a soma de quadrados dos resíduos da regressão da terceira parcela de divisão;

p é o número de variáveis regressoras.

Depois de calculada a estatística teste, ela é comparada com um F tabelado, caso FCG seja maior que o tabelado não se rejeita a homocedasticidade.

2.2.3. Teste de independência (Durbin-Watson)

Visando avaliar se os resíduos são independentes realiza-se o teste de Durbin-Watson, que consiste na comparação da estatística do teste de Durbin-Watson (dado pela equação 24) com os valores críticos tabelados dL e dU. (ESTATCAMP, 2021)

$$dw = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n}(\epsilon_i)^2} \quad (24)$$

Onde:

ϵ_i é o resíduo da observação i, com $i = 1, \dots, n$;

n é o número de observações.

Após o cálculo de dw compara-se a dL e dU:

- se $0 \leq dw < dL$ então a dependência é rejeitada;
- se $dL \leq dw < dU$ então o teste é inconclusivo;
- se $dU \leq dw < 4 - dU$ então a dependência não é rejeitada;
- se $4 - dU \leq dw < 4 - dL$ então o teste é inconclusivo;
- se $4 - dL \leq dw < 4$ então a dependência é rejeitada

2.2.4. Teste de estacionariedade (Kwiatkowski–Phillips–Schmidt–Shin)

Para avaliar a estacionariedade das séries históricas que serão testadas no presente estudo será utilizado o teste de Kwiatkowski–Phillips–Schmidt–Shin (KPSS). Esse método consiste em avaliar se existe uma raiz unitária na série, para isso supõe-se as equações 25 e 26. (NELIZE, 2018)

$$y_t = \xi D_t + r_t + \varepsilon_t \quad (25)$$

$$r_t = r_{t-1} + u_t \quad (26)$$

Onde:

y_t é a variável resposta no período t;

D_t é o coeficiente que define a raiz unitária;

r_t é um passeio aleatório;

u_t é uma Distribuição Normal Identicamente distribuida;

ε_t são os resíduos da regressão.

O teste de hipótese é baseado na estatística LM como apresentado na equação 27, caso esse valor seja menor que o valor crítico tabelado, rejeita-se a hipótese nula de que a série é estacionária.

$$LM = \frac{1}{N^2} * \left(\sum_{t=1}^{N} \frac{S_t^2}{\sigma^2}\right) \quad (27)$$

Onde:

LM é o valor da estatística de LM;

N é o número de observações;

S_t é o somatório dos erros, com $t = 1, \dots N$;

σ é a estimativa do erro da variância dessa regressão.

2.3. Métodos Comparativos

Nessa seção será apresentado quais as métricas que serão utilizadas para comparação dos diferentes modelos preditivos avaliados neste estudo.

2.3.1. Mean Squared Error (MSE)

Uma das métricas mais utilizadas para comparação de modelos preditivos é o MSE, esse método consiste em calcular o somatório do quadrado da variação entre o valor real e o valor predito sobre o total de observações, como apresentado na equação 28. Para essa métrica quanto menor for o resultado, melhor o modelo. (BAZIEWICZ, 2019; SAMPAIO et al., 2019)

$$MSE(\hat{y}) = \frac{1}{n} * \sum_1^n (y_i - \hat{y}_i)^2 \quad (28)$$

Onde:

γ_i é o valor real da observação i, com $i = 1, \dots, n$

$\hat{\gamma}_i$ é o valor estimado pelo modelo para a observação i, com $i = 1, \dots, n$

n é o número de observações.

2.3.2. Mean Absolute Percentage Error (MAPE)

Semelhante ao método MSE, o MAPE é uma métrica normalizada e consiste em calcular o somatório do módulo da variação entre o valor real e o valor predito sobre o módulo da variação entre o valor real e a média dos valores reais, sobre o total de observações, como apresentado na equação 29. (BAZIEWICZ, 2019; SAMPAIO et al., 2019)

$$MAPE(\hat{y}) = \frac{1}{n} * \sum_1^n \frac{|y_i - \hat{y}_i|}{|y_i - \bar{y}|} \quad (29)$$

Onde:

γ_i é o valor real da observação i, com $i = 1, \dots, n$;

$\hat{\gamma}_i$ é o valor estimado pelo modelo para a observação i, com $i = 1, \dots, n$;

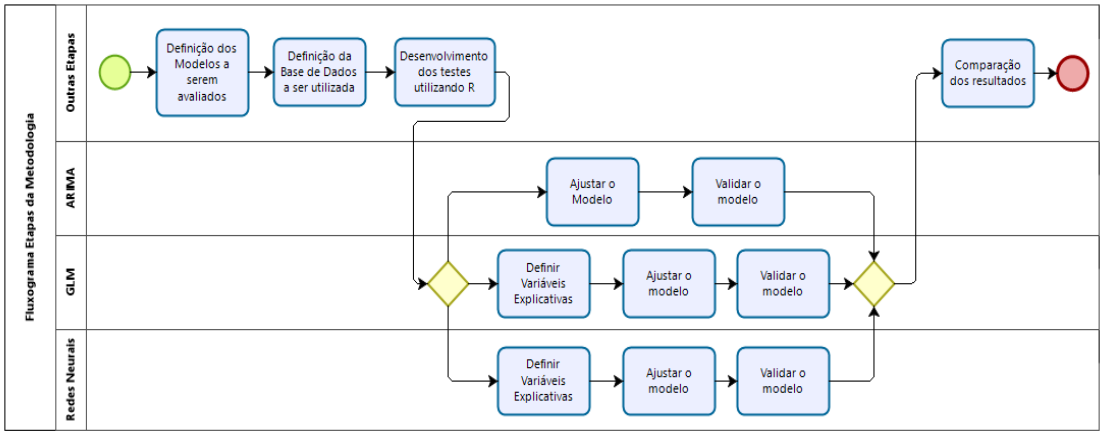
$\bar{y}$ é a média do valor real das n observações;

n é o número de observações.

3. METODOLOGIA

Para uma melhor organização e melhor desenvolvimento deste trabalho foi feita uma divisão em 6 etapas apresentadas na forma de fluxograma na FIGURA 3.

FIGURA 3 - FLUXOGRAMA DAS ETAPAS DA METODOLOGIA



FONTE: O Autor (2021)

- Etapa 1 – Definição dos modelos preditivos a serem avaliados

Primeiramente serão definidos os métodos preditivos a serem utilizados, bem como seu estudo, avaliando seu comportamento, quais seriam as restrições para sua aplicação e quais os testes necessários para validar o modelo gerado. Para o presente estudo foram utilizados os métodos GLM, RNA e ARIMA, os dois primeiros escolhidos por já serem métodos muito utilizados para modelos de previsão em atuária. Já o método RNA foi escolhido por ser um método que vem ganhando espaço no mercado com o aumento do uso de ferramentas da indústria 4.0. (FRONZETTI, 2019; KUMAR, 2015; ROCHA, 2020)

- Etapa 2 - Definição da base de dados a ser utilizada

Escolha da base de dados a ser utilizada, bem como quais as variáveis explicativas que poderão ser utilizadas. Para o presente estudo foi selecionada a base de dados de prêmio ganho e sinistro ocorrido obtida da base de dados do sistema de estatística da SUSEP, possui 6 variáveis e 1423380 observações. Além dessa base foi construído um banco de dados com variáveis econômicas obtidas do banco de dados do IPEADATA, com objetivo de incrementar os modelos construídos, essa base possui 252 observações e 12 variáveis. (SES SUSEP, 2021; IPEADATA, 2021)

- Etapa 3 - Desenvolvimento dos testes utilizando o *software* R;

Escolha dos pacotes existentes no R para criação dos modelos que serão comparados, estudo das variáveis de entrada e das variáveis de saída de cada função para poder aplicá-las corretamente. Os pacotes escolhidos foram *forecast* para os modelos ARIMA, *stats* para os modelos de GLM e *nnet* para Redes Neurais. (KLEINA, 2018; HYNDMAN, 2020; R CORE TEAM, 2020; VENABLE, 2002)

- Etapa 4 - Seleção das variáveis relevantes para criação do modelo

Uso de métodos estatísticos para definir quais variáveis são mais relevantes para o modelo. Os modelos de Redes Neurais e GLM utilizam, além da própria série histórica, variáveis externas, por esse motivo foram utilizados dois métodos (*stepwise* para os modelos GLM e árvore de decisão para os modelos de Redes Neurias) para a seleção das melhores variáveis a fim de maximizar o desempenho dos modelos gerados. (BARBOSA, 2010; HAYES, 2021)

- Etapa 5 – Validação dos modelos criados

Após a criação dos modelos é necessário que sejam realizados uma série de testes estatísticos para validar algumas premissas assumidas para a construção de cada um deles. Para a avaliação da normalidade dos resíduos foi utilizada a função no R *shapiro.test*, considerando um p-valor mínimo de 0.05 para aceitar a normalidade, para a homoscedasticidade utilizou-se a função *gqtest*, utilizando também um um p-valor mínimo de 0.05 e para a estacionariedade das séries ARIMA foi utilizada a função *kpss.test*, buscando um p-valor inferior a 0.05.

- Etapa 6 - Resultados e comparações entre os métodos

Após a validação dos modelos gerados, realizou-se a comparação dos resultados obtidos com os diferentes métodos. Para o presente estudo serão utilizados o MAPE e o MSE para avaliar a precisão dos modelos gerados, para avaliar a normalidade dos resíduos será utilizado o teste de Shapiro-Wilk, para avaliar a volatilidade das séries será utilizado o desvio padrão e por fim, para mensurar a estacionariedade das séries e dos períodos a serem previstos será utilizado o teste de Kwiatkowski–Phillips–Schmidt–Shin (KPSS).

4. RESULTADOS E DISCUSSÕES

Nesta seção serão apresentados os resultados obtidos ao gerar os modelos preditivos utilizando os três diferentes métodos abordados no presente estudo, bem como seu desempenho com as quatro diferentes séries históricas utilizadas.

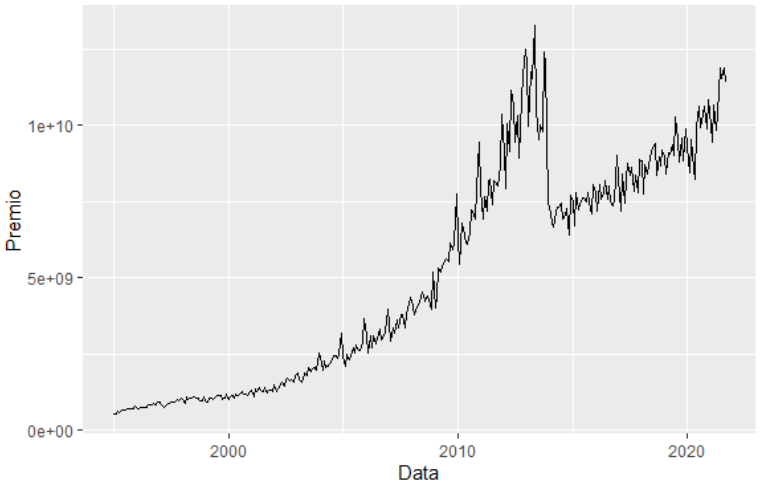
4.1. Análise descritiva das séries utilizadas

Nessa seção serão apresentadas as diferentes séries de dados que foram analisadas no presente estudo. Foram construídas 4 séries de dados com a base de dados obtida no Sistema de Estatística da SUSEP, duas de prêmio direto (valor pago pelo segurado para obter o direito ao seguro) e duas de sinistro ocorrido (valor pago pela seguradora para cobrir a materialização do risco segurado). (SUSEP, 2021)

4.1.1. Prêmio Direto e Sinistro Ocorrido Mercado

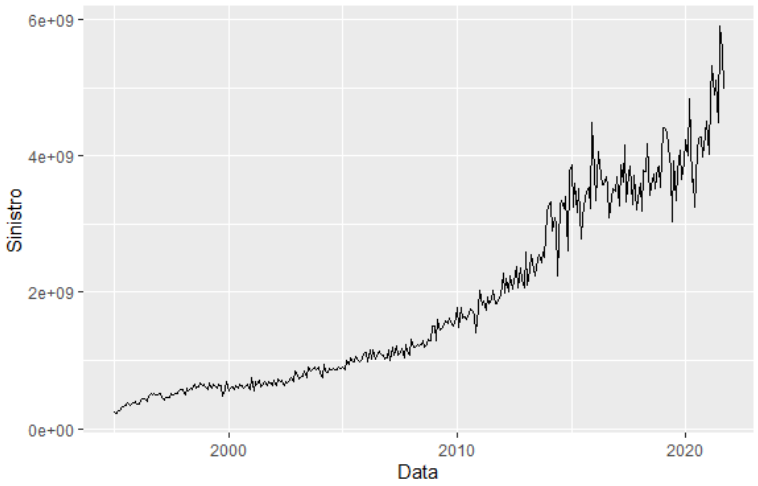
A série de Prêmio Direto Mercado e Sinistro Ocorrido Mercado representam o total de prêmio direto e sinistro ocorrido das 121 entidades existentes na base de dados em janeiro de 2021 (não foram consideradas todas as entidades presentes na base pois algumas deixaram de existir e a sua inclusão na análise poderia influenciar negativamente na previsão dos modelos). O histórico dessas séries está apresentado nas FIGURAS 4 e 5.

FIGURA 4 -HISTÓRICO PRÊMIO DIRETO MERCADO



FONTE: O Autor (2021)

FIGURA 5 – HISTÓRICO SINISTRO OCORRIDO MERCADO

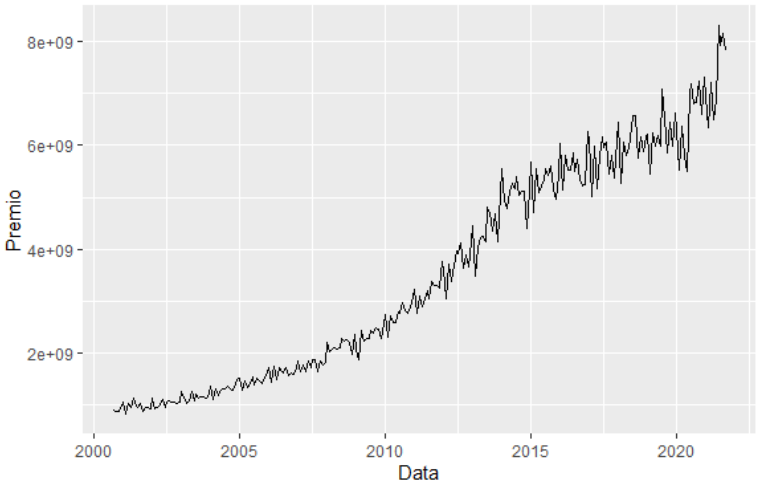


FONTE: O Autor (2021)

4.1.2. Prêmio Direto e Sinistro Ocorrido Danos Mercado

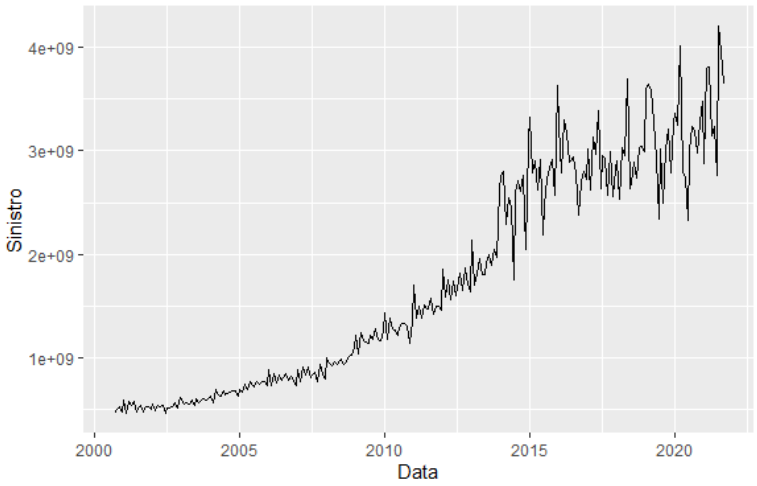
A série de Prêmio Direto Danos Mercado e Sinistro Ocorrido Danos Mercado (FIGURAS 7 e 8, respectivamente) representam o total de prêmio direto e sinistro ocorrido de 104 dos 208 ramos presentes nas das 121 entidades existentes na base de dados em janeiro de 2021 (não foram consideradas todas as entidades presentes na base pois algumas deixaram de existir e a sua inclusão na análise poderia influenciar negativamente na previsão dos modelos).

FIGURA 6 – HISTÓRICO PRÊMIO DIRETO DANOS MERCADO



FONTE: O Autor (2021)

FIGURA 7 - HISTÓRICO SINISTRO OCORRIDO MERCADO



FONTE: O Autor (2021)

4.1.3. Análise da volatilidade e estacionariedade das séries

Como o presente estudo tem como objetivo avaliar o desempenho dos métodos preditivos em diferentes séries de dados foi feita a análise de volatilidade e estacionariedade das quatro séries apresentadas anteriormente, os resultados estão apresentados na TABELA 3.

TABELA 3 – ANÁLISE DAS SÉRIES UTILIZADAS

Série	p-valor KPSS	Estatitica KPSS	Desvio Padrão / Média
Prêmio Mercado	< 0.01	4.9962	0.7178053
Sinistro Mercado	< 0.01	5.0799	0.738619
Prêmio Danos Mercado	< 0.01	4.2799	0.5956044
Sinistro Danos Mercado	< 0.01	4.2063	0.6070804

FONTE: O Autor (2021)

Apesar de, segundo teste de KPSS, nenhuma das séries avaliadas ser estacionária, pode-se, ao avaliar o valor resultante da estatística de KPSS, constatar que as séries de Prêmio e Sinistro Danos são mais estacionárias, já que apresentam um valor de KPSS menor que as demais. O mesmo acontece ao se avaliar a volatilidade das séries, os desvios em relação à média são menores para as séries de danos.

4.1.4. Séries Macroeconômicas

Como os métodos GLM e Redes Neurais podem utilizar variáveis explicativas construiu-se uma base de dados macroeconômicos utilizando séries disponibilizadas pelo IPEADATA, foram selecionadas 12 variáveis, apresentadas na TABELA 4.

TABELA 4 – VARIÁVEIS MACROECONÔMICAS

Código IPEA	Nome Variável	Explicação Variável
BM12_ERCF12	taxa_compra_real	Taxa de câmbio comercial para compra: real (R$) / dólar americano (US$) - fim período
FCESP12_IICA12	indices_condicoes_economicas	Índice de condições econômicas atuais (ICEA)
ANP12_CGASOL12	cosumo_gasol	Consumo aparente - gasolina - média - qde/dia
PAN12_DTSPY12	divida_publica	Dívida pública total
MPAS12_ARRBT12	fluxo_prev_arr_bruta	Fluxo de caixa da previdência - Arrecadação Bruta
MPAS12_ARRLIQ12	fluxo_prev_arr_liq	Fluxo de caixa da previdência - Arrecadação Líquida
MPAS12_BENPREV12	fluxo_prev_arr_benef_prev	Fluxo de caixa da previdência - Benefícios Previdenciários
MPAS12_RESPRGPS12	fluxo_prev_arr_result_prim	Fluxo de caixa da previdência - Resultado Primário do RGPS
BM12_RNDPO12	rend_poupanca	Poupança - rendimento nominal até 03.05.2012 - Rentabilidade no período (1º dia do mês)
PRECOS12_IPCASC12	var_saude	IPCA - saúde e cuidados pessoais - var.

FONTE: O Autor (2021)

4.2. Análise do desempenho dos modelos gerados

Nesta seção serão apresentados os modelos gerados para cada uma das séries em estudo, bem como os resultados para os testes estatísticos necessários para validá-los. Os resultados obtidos serão apresentados em dois grupos: o primeiro utilizando dados pré pandemia para o treinamento dos modelos (anterior a dezembro de 2019) e o segundo utilizando dados do início da pandemia para realizar o treinamento dos modelos (anterior a junho de 2020), para os dois grupos será realizada a previsão para 12 meses a partir de julho de 2020.

4.2.1. ARIMA

Como o modelo ARIMA não necessita de nenhuma variável explicativa para realizar a previsão não foi utilizada a base de dados do IPEADATA, para esse

modelo foram testados diferentes valores para p, d e q, bem como diferentes períodos para a construção dos modelos como apresentado.

4.2.1.1. Pré Pandemia

A TABELA 5 apresenta os principais resultados obtidos durante o teste dos diferentes parâmetros do modelo ARIMA e o resultado dos testes estatísticos necessários para a validação dos modelos.

TABELA 5 - RESULTADOS MODELO ARIMA PRÉ PANDEMIA

Série	p	q	d	Data Início	Data Fim	Shapiro	MSE	MAPE	P valor KPSS
Prem Danos Mercado	0	2	0	01/01/2010	31/12/2019	0,22	1,64E+19	0,998	0,1
Prem Danos Mercado	0	0	2	01/01/2012	31/12/2019	0,07	3,07E+19	1,031	0,1
Prem Danos Mercado	1	0	1	01/01/2016	31/12/2019	0,19	1,01E+18	1,846	0,1
Prem Mercado	0	2	0	01/01/2010	31/12/2019	0,22	1,64E+19	0,998	0,1
Prem Mercado	0	0	2	01/01/2012	31/12/2019	0,07	3,07E+19	1,031	0,1
Prem Mercado	3	1	1	01/01/2016	31/12/2019	0,09	1,08E+18	1,635	0,1
Sin Danos Mercado	1	0	1	01/01/2010	31/12/2019	0,12	1,86E+17	2,598	0,1
Sin Danos Mercado	0	0	1	01/01/2012	31/12/2019	0,38	1,89E+17	1,578	0,1
Sin Danos Mercado	1	0	1	01/01/2016	31/12/2019	0,44	1,14E+17	0,912	0,1
Sinistro Mercado	2	1	1	01/01/2010	31/12/2019	0,06	6,57E+17	0,842	0,1
Sinistro Mercado	1	2	1	01/01/2012	31/12/2019	0,42	6,63E+17	0,863	0,1
Sinistro Mercado	1	0	1	01/01/2016	31/12/2019	0,68	6,41E+17	0,858	0,1

FONTE: O Autor (2021)

Pode-se notar que os modelos com os melhores desempenhos foram os com menor intervalo de tempo utilizados para a construção do modelo e que a série mais estacionária (Sinistro Danos) obteve um desempenho superior as demais.

4.2.1.2. Durante Pandemia

A TABELA 6 apresenta os principais resultados obtidos durante o teste dos diferentes parâmetros do modelo ARIMA e o resultado dos testes estatísticos necessários para a validação dos modelos.

TABELA 6 RESULTADOS ARIMA PÓS PANDEMIA

Série	p	q	d	Data Início	Data Fim	Shapiro	MSE	MAPE	P valor KPSS
Prem Danos Mercado	0	1	0	01/01/2010	30/06/2021	0,56	1,00E+20	1,001	0,1
Prem Danos Mercado	0	1	0	01/01/2012	30/06/2021	0,74	1,00E+20	1,001	0,07
Prem Danos Mercado	4	0	2	01/01/2016	30/06/2021	0,20	3,90E+18	1,060	0,1
Prem Mercado	0	1	0	01/01/2010	30/06/2021	0,56	1,00E+20	1,001	0,1
Prem Mercado	0	0	2	01/01/2012	30/06/2021	0,11	1,70E+20	1,009	0,1
Prem Mercado	0	0	2	01/01/2016	30/06/2021	0,06	2,10E+20	1,009	0,1
Sin Danos Mercado	1	0	2	01/01/2010	30/06/2021	0,20	1,60E+18	1,024	0,1
Sin Danos Mercado	0	2	1	01/01/2012	30/06/2021	0,18	1,50E+17	1,395	0,1
Sin Danos Mercado	0	1	1	01/01/2016	30/06/2021	0,10	1,20E+17	1,426	0,1
Sinistro Mercado	1	0	1	01/01/2010	30/06/2021	0,18	9,30E+17	0,989	0,1
Sinistro Mercado	2	1	1	01/01/2012	30/06/2021	0,53	6,70E+17	0,955	0,1
Sinistro Mercado	0	1	1	01/01/2016	30/06/2021	0,26	6,00E+17	0,929	0,1

FONTE: O Autor (2021)

Pode-se notar, ao avaliar as medidas de erro, que os modelos que melhor se ajustaram foram os que utilizaram um intervalo de tempo menor e os piores desempenhos foram os que utilizaram um intervalo de tempo maior, além disso a série mais estacionária foi, mais uma vez, a que se melhor ajustou ao modelo ARIMA.

4.2.2. GLM

O modelo GLM utiliza, além da série histórica, variáveis explicativas para realizar a previsão, por isso utilizou-se a base de dados do IPEADATA. Foram testadas algumas seleções de variáveis para esse modelo, além disso foram testadas as três diferentes funções de ligação apresentadas na seção 2.1.2.2.1, bem como diferentes períodos para a construção dos modelos.

4.2.2.1. Pré Pandemia

A TABELA 7 apresenta os principais resultados obtidos durante o teste dos diferentes parâmetros do modelo GLM e o resultado dos testes estatísticos necessários para a validação dos modelos. Para cada uma das séries e períodos analisados realizou-se, utilizando o método *stepwise*, a seleção de variáveis explicativas como apresentado na TABELA 8.

TABELA 7 – RESULTADOS GLM PRÉ PANDEMIA

Série	Função de Ligação	Data início	Data fim	Shapiro	MSE	MAPE	Goldfeld Quandt
Prem Danos Mercado	identity	01/01/2010	30/06/2021	0,82	3,75E+17	17,08	0,13
Prem Danos Mercado	identity	01/01/2012	30/06/2021	0,84	4,06E+17	1,57	0,08
Prem Danos Mercado	identity	01/01/2016	30/06/2021	0,27	4,20E+18	1,11	0,12
Prem Mercado	identity	01/01/2010	30/06/2021	0,10	6,73E+17	1,96	1,00
Prem Mercado	inverse	01/01/2012	30/06/2021	0,67	6,43E+17	1,56	1,00
Prem Mercado	inverse	01/01/2016	30/06/2021	0,97	4,41E+19	1,04	0,15
Sin Danos Mercado	inverse	01/01/2010	30/06/2021	0,06	4,65E+18	0,99	0,00
Sin Danos Mercado	1/mu^2	01/01/2012	30/06/2021	0,35	1,85E+18	1,06	0,13
Sin Danos Mercado	identity	01/01/2016	30/06/2021	0,65	1,17E+18	1,00	0,08
Sinistro Mercado	inverse	01/01/2010	30/06/2021	0,30	6,58E+18	0,96	0,00
Sinistro Mercado	1/mu^2	01/01/2012	30/06/2021	0,05	2,38E+18	1,04	0,29
Sinistro Mercado	1/mu^2	01/01/2016	30/06/2021	0,93	2,01E+18	1,01	0,18

FONTE: O Autor (2021)

TABELA 8 – SELEÇÃO DE VARIÁVEIS GLM PRÉ PANDEMIA

Série	Data início	Variáveis selecionadas
Prem Danos Mercado	01/01/2010	taxa_compra_real + indices_condicoes_economicas + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca
Prem Danos Mercado	01/01/2012	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim
Prem Danos Mercado	01/01/2016	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca
Prem Mercado	01/01/2010	taxa_compra_real + indices_condicoes_economicas + cosumo_gasol + rend_poupanca
Prem Mercado	01/01/2012	indices_condicoes_economicas + rend_poupanca
Prem Mercado	01/01/2016	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca
Sin Danos Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim
Sin Danos Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca
Sin Danos Mercado	01/01/2016	indices_condicoes_economicas + cosumo_gasol + rend_poupanca

Sinistro Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim
Sinistro Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca
Sinistro Mercado	01/01/2016	indices_condicoes_economicas + cosumo_gasol + rend_poupanca

Fonte (O Autor, 2021)

Ao avaliar os resultados obtidos utilizando o modelo GLM pode-se notar que a série com menor volatilidade (Prêmio Direto Danos Mercado) foi a que, ao avaliar o indicador MAPE, pior se adequou ao modelo proposto quando utilizado um período histórico maior.

4.2.2.2. Durante Pandemia

A TABELA 9 apresenta os principais resultados obtidos durante o teste dos diferentes parâmetros do modelo GLM e o resultado dos testes estatísticos necessários para a validação dos modelos. Para cada uma das séries e períodos analisados realizou-se, utilizando o método *stepwise*, a seleção de variáveis explicativas como apresentado na TABELA 10.

TABELA 9 - RESULTADOS GLM DURANTE PANDEMIA

Série	Função de Ligação	Data início	Data fim	Shapiro	MSE	MAPE	Goldfeld Quandt
Prem Danos Mercado	identity	01/01/2010	30/06/2021	0,73	3,53E+17	5,96	0,04
Prem Danos Mercado	1/mu^2	01/01/2012	30/06/2021	0,00	4,78E+18	1,31	0,28
Prem Danos Mercado	inverse	01/01/2016	30/06/2021	0,94	2,05E+19	1,03	0,30
Prem Mercado	inverse	01/01/2010	30/06/2021	0,19	1,05E+18	3,12	1,00
Prem Mercado	inverse	01/01/2012	30/06/2021	0,51	1,61E+18	0,90	1,00
Prem Mercado	inverse	01/01/2016	30/06/2021	0,86	4,46E+19	1,09	0,40
Sin Danos Mercado	inverse	01/01/2010	30/06/2021	0,78	1,24E+19	0,95	0,30
Sin Danos Mercado	1/mu^2	01/01/2012	30/06/2021	0,06	7,94E+18	1,07	0,23
Sin Danos Mercado	1/mu^2	01/01/2016	30/06/2021	0,95	2,09E+18	1,01	0,22
Sinistro Mercado	inverse	01/01/2010	30/06/2021	0,78	1,24E+19	0,95	0,00
Sinistro Mercado	1/mu^2	01/01/2012	30/06/2021	0,06	7,94E+18	1,07	0,23
Sinistro Mercado	1/mu^2	01/01/2016	30/06/2021	0,95	2,09E+18	1,01	0,22

FONTE: O Autor (2021)

TABELA 10 – SELEÇÃO DE VARIÁVEIS GLM DURANTE PANDEMIA

Série	Data início	Variáveis selecionadas
Prem Danos Mercado	01/01/2010	taxa_venda_real + indices_condicoes_economicas + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca
Prem Danos Mercado	01/01/2012	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim
Prem Danos Mercado	01/01/2016	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca
Prem Mercado	01/01/2010	taxa_compra_real + indices_condicoes_economicas + cosumo_gasol + rend_poupanca
Prem Mercado	01/01/2012	indices_condicoes_economicas + rend_poupanca
Prem Mercado	01/01/2016	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca
Sin Danos Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca + var_saude
Sin Danos Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca + var_saude
Sin Danos Mercado	01/01/2016	indices_condicoes_economicas + cosumo_gasol + rend_poupanca + var_saude
Sinistro Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + var_saude
Sinistro Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_benef_prev + fluxo_prev_arr_result_prim + rend_poupanca + var_saude
Sinistro Mercado	01/01/2016	indices_condicoes_economicas + cosumo_gasol + rend_poupanca + var_saude

FONTE: O Autor (2021)

Ao avaliar os resultados obtidos utilizando o modelo GLM pode-se notar que a série com menor volatilidade (Prêmio Direto Danos Mercado) foi, mais uma vez, a que, ao avaliar o indicador MAPE pior se adequou ao modelo proposto quando utilizado um período histórico maior. Nota-se também que as séries que tiveram menor MSE foram as que utilizaram um maior período de tempo para o treinamento do modelo.

4.2.3. Redes Neurais

O modelo de Redes Neurais utiliza, além da série histórica, variáveis explicativas para realizar a previsão, por isso utilizou-se a base de dados do IPEADATA. Além dos diferentes períodos para a construção dos modelos como testado nos demais modelos, foram testados diferentes números de neurônios em cada camada escondida.

4.2.3.1. Pré Pandemia

A TABELA 11 apresenta os principais resultados obtidos durante o teste dos diferentes parâmetros do modelo de Redes Neurais e o resultado dos testes estatísticos necessários para a validação dos modelos. Para cada uma das séries e períodos analisados realizou-se, utilizando árvore de decisão, a seleção de variáveis explicativas como apresentado na TABELA 12.

TABELA 11 - RESULTADOS REDES NEURAIS PRÉ PANDEMIA

Série	Primeira Camada	Segunda Camada	Data início	Data fim	Shapiro	MSE	MAPE
Prem Danos Mercado	6	1	01/01/2010	30/06/2021	0,55	4,85E+17	1,33
Prem Danos Mercado	2	1	01/01/2012	30/06/2021	0,22	2,18E+18	1,00
Prem Danos Mercado	5	3	01/01/2016	30/06/2021	0,07	7,82E+17	0,95
Prem Mercado	7	4	01/01/2010	30/06/2021	0,001	8,35E+18	1,004
Prem Mercado	10	4	01/01/2012	30/06/2021	0,14	1,29E+19	1,00
Prem Mercado	9	2	01/01/2016	30/06/2021	0,68	2,47E+18	0,97
Sin Danos Mercado	10	2	01/01/2010	30/06/2021	0,08	1,46E+17	1,03
Sin Danos Mercado	7	3	01/01/2012	30/06/2021	0,21	5,43E+17	0,96
Sin Danos Mercado	3	5	01/01/2016	30/06/2021	0,10	1,80E+17	0,88
Sinistro Mercado	4	5	01/01/2010	30/06/2021	0,10	5,07E+17	0,99
Sinistro Mercado	10	3	01/01/2012	30/06/2021	0,23	7,58E+17	0,99
Sinistro Mercado	9	3	01/01/2016	30/06/2021	0,16	9,86E+17	0,98

FONTE: O Autor (2021)

TABELA 12 – SELEÇÃO DE VARIÁVEIS REDES NEURAIS PRÉ PANDEMIA

Série	Data início	Variáveis selecionadas
Prem Danos Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + fluxo_prev_arr_result_prim

Série	Data início	Variáveis selecionadas
Prem Danos Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + cosumo_gasol + fluxo_prev_arr_result_prim
Prem Danos Mercado	01/01/2016	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + cosumo_gasol + rend_poupanca + var_saude
Prem Mercado	01/01/2010	rend_poupanca
Prem Mercado	01/01/2012	taxa_venda_real + taxa_compra_real + rend_poupanca
Prem Mercado	01/01/2016	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + cosumo_gasol + rend_poupanca + var_saude
Sin Danos Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + fluxo_prev_arr_result_prim
Sin Danos Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + fluxo_prev_arr_result_prim
Sin Danos Mercado	01/01/2016	mes
Sinistro Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + fluxo_prev_arr_result_prim
Sinistro Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + cosumo_gasol + fluxo_prev_arr_result_prim
Sinistro Mercado	01/01/2016	taxa_compra_real + indices_condicoes_economicas + ano + cosumo_gasol + mes

FONTE: O Autor (2021)

Ao se analisar os erros das séries notam-se que as que melhor se ajustaram foram as duas séries de Sinistro – Sinistro Danos Mercado é a série mais estacionária e Sinistro Mercado a menor, porém Sinistro Mercado é a série com maior volatilidade. Tirando a série de Prêmio Mercado não houve grandes diferenças entre períodos como se nota nos outros dois modelos.

4.2.3.2. Durante Pandemia

A TABELA 13 apresenta os principais resultados obtidos durante o teste dos diferentes parâmetros do modelo de Redes Neurais e o resultado dos testes estatísticos necessários para a validação dos modelos. Para cada uma das séries e períodos analisados realizou-se, utilizando árvore de decisão, a seleção de variáveis explicativas como apresentado na TABELA 14.

TABELA 13 - RESULTADOS REDES NEURAIS DURANTE PANDEMIA

Série	Primeira Camada	Segunda Camada	Data início	Data fim	Shapiro	MSE	MAPE
Prem Danos Mercado	5	4	01/01/2010	30/06/2021	0,904	3,85E+17	1,091
Prem Danos Mercado	10	3	01/01/2012	30/06/2021	0,331	1,76E+18	0,953
Prem Danos Mercado	9	2	01/01/2016	30/06/2021	0,326	8,77E+17	0,901
Prem Mercado	8	5	01/01/2010	30/06/2021	0,001	4,73E+18	0,999
Prem Mercado	6	2	01/01/2012	30/06/2021	0,092	7,40E+18	0,997
Prem Mercado	9	4	01/01/2016	30/06/2021	0,062	7,69E+18	0,969
Sin Danos Mercado	7	1	01/01/2010	30/06/2021	0,110	2,70E+17	1,041
Sin Danos Mercado	6	4	01/01/2012	30/06/2021	0,117	2,60E+17	0,870
Sin Danos Mercado	8	2	01/01/2016	30/06/2021	0,083	1,86E+17	0,817
Sinistro Mercado	4	5	01/01/2010	30/06/2021	0,057	5,80E+17	0,977
Sinistro Mercado	6	4	01/01/2012	30/06/2021	0,083	2,30E+17	0,874
Sinistro Mercado	7	5	01/01/2016	30/06/2021	0,140	6,81E+17	0,983

FONTE: O Autor (2021)

TABELA 14 – SELEÇÃO DE VARIÁVEIS REDES NEURAIS DURANTE PANDEMIA

Série	Data início	Variáveis selecionadas
Prem Danos Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + fluxo_prev_arr_result_prim + rend_poupanca
Prem Danos Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + cosumo_gasol + fluxo_prev_arr_result_prim
Prem Danos Mercado	01/01/2016	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + cosumo_gasol + rend_poupanca + var_saude
Prem Mercado	01/01/2010	rend_poupanca

Série	Data início	Variáveis selecionadas
Prem Mercado	01/01/2012	rend_poupanca
Prem Mercado	01/01/2016	taxa_compra_real + indices_condicoes_economicas + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + cosumo_gasol + rend_poupanca + var_saude
Sin Danos Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + fluxo_prev_arr_result_prim
Sin Danos Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + cosumo_gasol + fluxo_prev_arr_result_prim
Sin Danos Mercado	01/01/2016	fluxo_prev_arr_result_prim + mes
Sinistro Mercado	01/01/2010	taxa_compra_real + cosumo_gasol + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + fluxo_prev_arr_result_prim
Sinistro Mercado	01/01/2012	taxa_compra_real + divida_publica + fluxo_prev_arr_bruta + fluxo_prev_arr_liq + fluxo_prev_arr_benef_prev + ano + indices_condicoes_economicas + cosumo_gasol + fluxo_prev_arr_result_prim
Sinistro Mercado	01/01/2016	ano + cosumo_gasol + mes

FONTE: O Autor (2021)

Assim como para os modelos pré pandemia as séries que melhor se ajustaram foram as de Sinistro – Sinistro Danos Mercado é a série mais estacionária e Sinistro Mercado a menor, porém Sinistro Mercado é a série com maior volatilidade. Tirando a série de Prêmio Mercado não houve grandes diferenças entre períodos como se nota nos outros dois modelos.

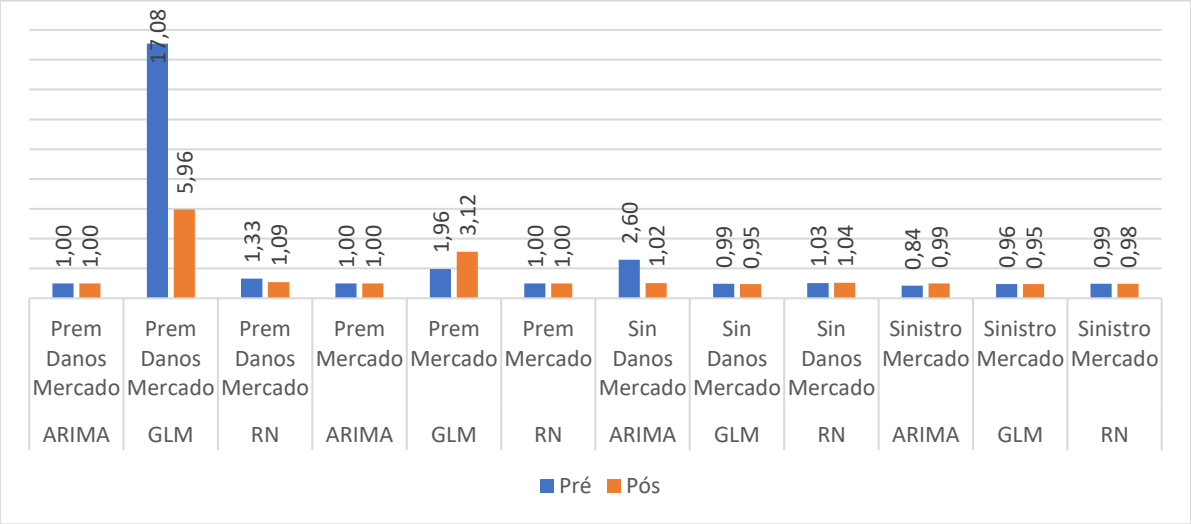
4.2.4. Comparativo de desempenho dos métodos nas diferentes datas

Além de avaliar o desempenho de cada modelo com diferentes parâmetros o presente estudo também busca comparar o desempenho dos melhores modelos gerados com cada um dos métodos propostos, para isso foram construídos gráficos comparando as medidas de erro avaliadas, nas três datas de início propostas, bem como com a divisão antes e durante pandemia.

4.2.4.1. Data Início 01/2010

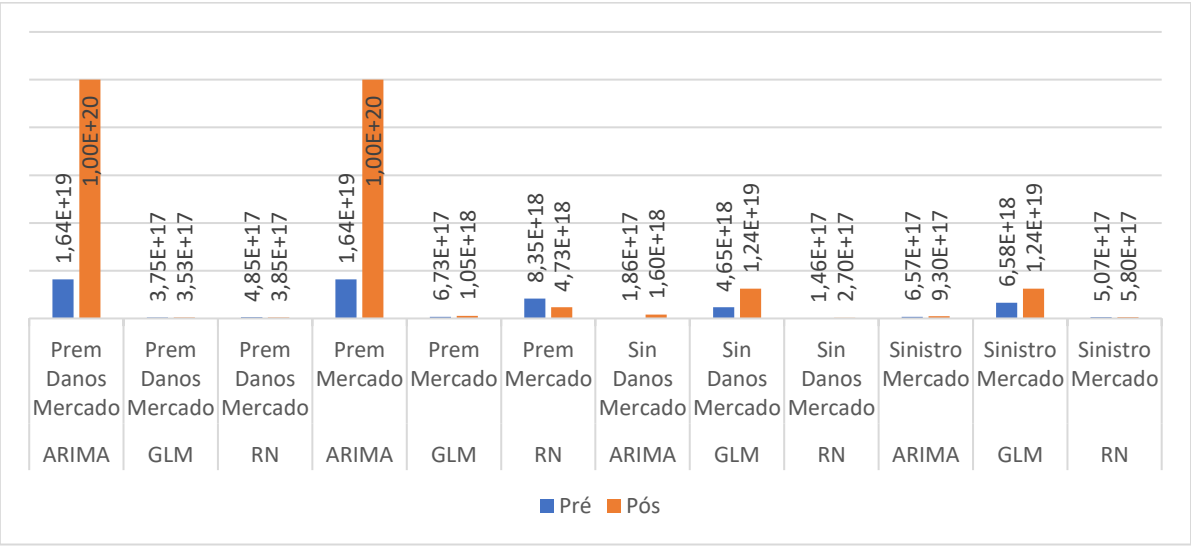
O maior período de tempo analisado foi o com a construção dos modelos com dados a partir de janeiro de 2010, as FIGURAS 8 e 9 apresentam os gráficos comparativos das medidas de erro MSE e MAPE.

FIGURA 8 – COMPARAÇÃO MAPE 01/2010



FONTE: O Autor (2021)

FIGURA 9 – COMPARAÇÃO MSE 01/2010



FONTE: O Autor (2021)

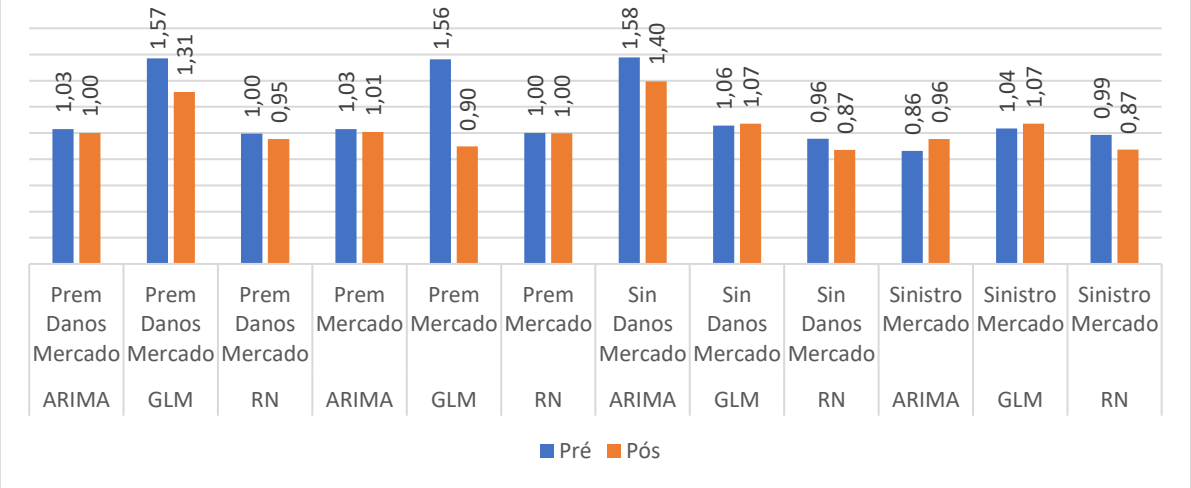
Ao avaliar os valores de MAPE de cada modelo observa-se que existe uma grande diferença entre os valores obtidos para a séries de Prêmio Danos Mercado utilizando GLM e Sinistro Danos Mercado utilizando ARIMA, além disso pode-se

notar grandes valores no MSE para os modelos ARIMA nas duas séries de prêmio além de uma grande variação entre o resultado pré e durante pandemia (crescimento de mais de 500% no erro de previsão). Já os valores obtidos pelo método de Redes Neurais se mostram consistente em todos os cenários apresentados.

4.2.4.2. Data Início 01/2012

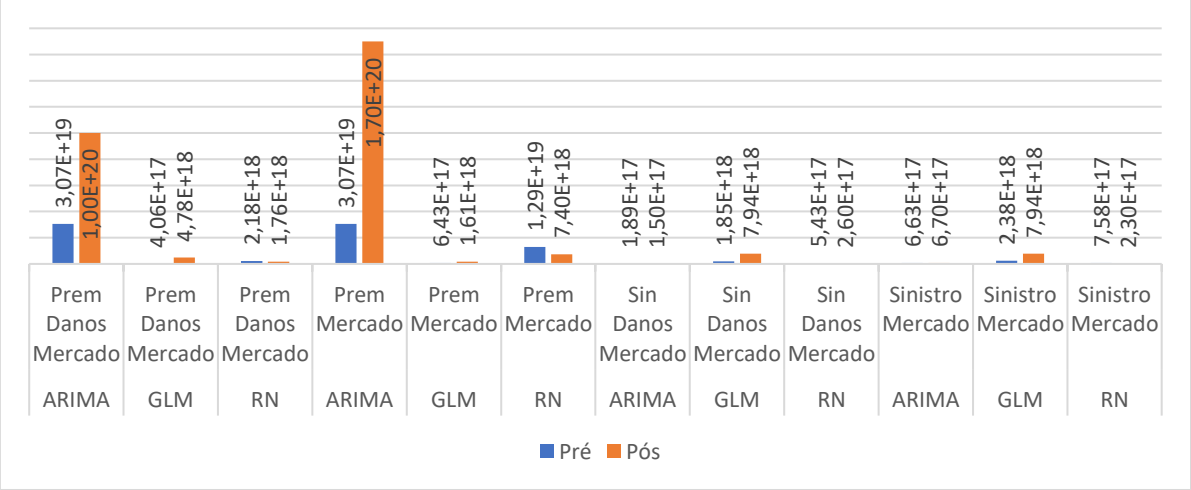
Foi realizada a construção dos modelos com dados a partir de janeiro de 2012, as FIGURAS 10 e 11 apresentam os gráficos comparativos das medidas de erro MSE e MAPE.

FIGURA 10 – COMPARAÇÃO MAPE 01/2012



FONTE: O Autor (2021)

FIGURA 11 - COMPARAÇÃO MSE - 01/2012



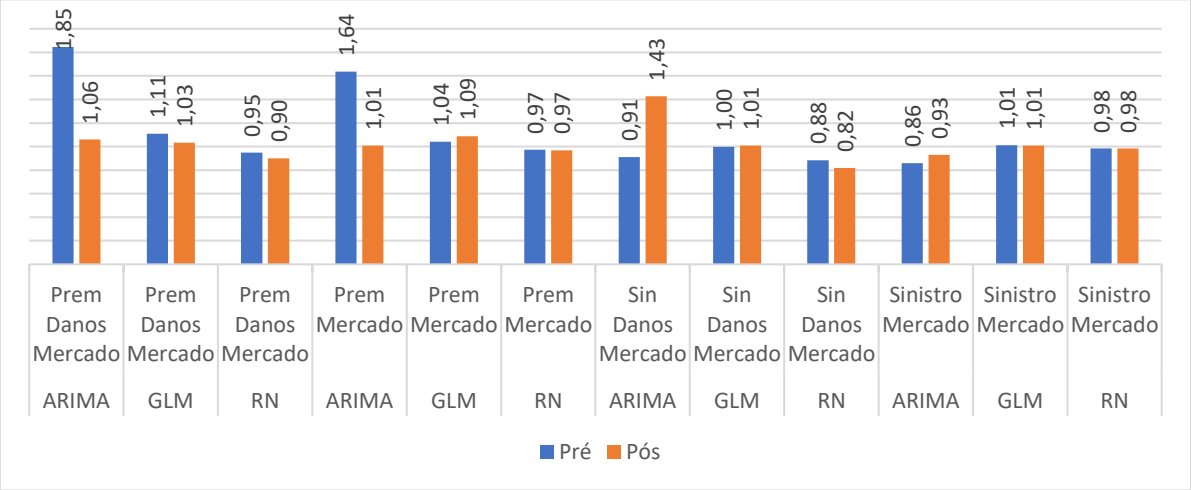
FONTE: O Autor (2021)

Pode-se notar que as mesmas séries que tiveram problemas de ajuste com modelo ARIMA com data de início 01/2010 tiveram, novamente, altos valores de MSE, reforçando a dificuldade de ajuste de séries ARIMA com grande quantidade de informação. Com exceção dos valores MSE para as séries de Prêmio utilizando o modelo ARIMA os resultados obtidos se mostraram consistentes, sem grandes variações de resultados entre os modelos.

4.2.4.3. Data Início 01/2012

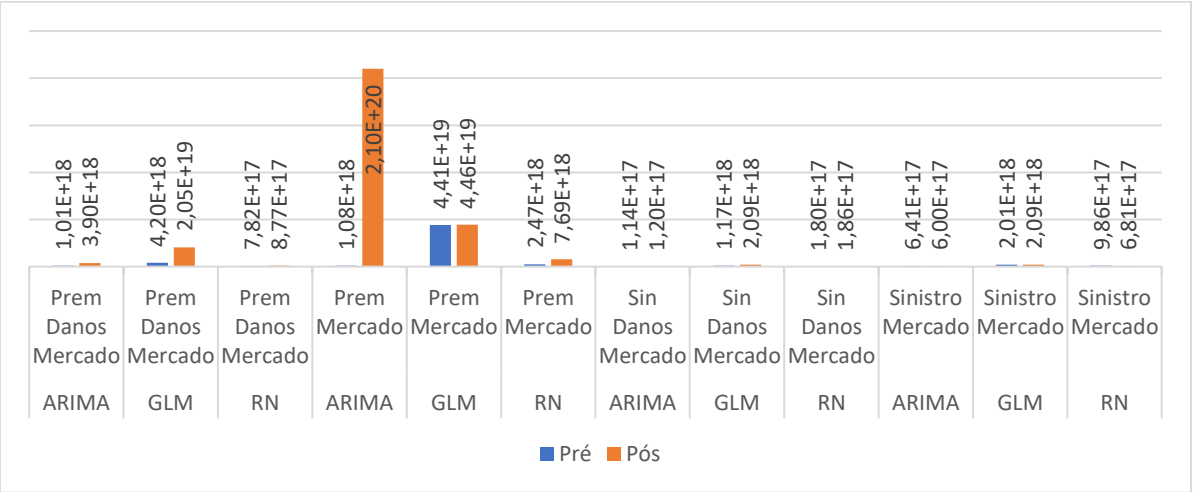
Por fim para a construção dos modelos com dados a partir de janeiro de 2016, tem-se as FIGURAS 12 e 13 com os gráficos comparativos das medidas de erro MSE e MAPE.

FIGURA 12 – COMPARAÇÃO MAPE - 01/2016



FONTE: O Autor (2021)

FIGURA 13 - COMPARAÇÃO MSE - 01/2016



FONTE: O Autor (2021)

Mesmo as mesmas séries de Prêmio tendo um maior erro quadrático utilizando o ARIMA este valor é 16 vezes menor que o valor obtido ao modelar a mesma série tendo a data de início em 01/2010. Outro ponto de atenção é o aumento do erro para a série de Prêmio Mercado utilizando GLM.

5. CONSIDERAÇÕES FINAIS

Esse estudo tinha como objetivo avaliar se o desempenho dos métodos ARIMA, GLM e Redes Neurais era satisfatório ao realizar previsões e simulações de possíveis cenários utilizando dados do mercado de seguros e variáveis macroeconômicas. Foram utilizadas funções do *software* R para realizar os testes com diversos parâmetros para cada um dos modelos.

Surpreendentemente a série com menor volatilidade (Prêmio Ganho Mercado) foi uma das que apresentou maiores valores de erro para os modelos gerados, além disso foi a série que teve maior variação do resultado entre modelos gerados pré pandemia e modelos gerados durante pandemia. Isso pode ser pelo fato de que a falta de volatilidade da série pode fazer com que qualquer variação (como uma pandemia, por exemplo) no comportamento da série acabe gerando erros muito grandes. Ao analisar as séries mais e menos estacionárias (Sinistro Ocorrido Danos Mercado e Sinistro Ocorrido Mercado), não se pode tirar conclusões a respeito da influência da estacionariedade no desempenho dos modelos estatísticos.

Pode-se perceber que os métodos preditivos acabam tendo uma taxa de erro maior quando se utiliza um período muito grande de tempo para realizar o ajuste dos modelos, isso pode ser causado tanto pelo *overfitting* dos modelos como pela própria mudança de comportamento das séries ao longo do tempo que acaba influenciando negativamente nos resultados obtidos. Dentre os três modelos testados o que mais teve influência do período de tempo utilizado foi o ARIMA, é notória a melhora do desempenho desse método, tendo uma redução de mais de 16 vezes no valor do erro calculado. Apesar de não ter sido tão preciso nas previsões o modelo ARIMA não utiliza outras séries para construção do modelo, visto que se trata de um modelo autorregressivo, o que pode ser uma vantagem quando não se tem um conjunto de variáveis explicativas.

Diferente do modelo ARIMA os modelos GLM e Redes Neurais podem utilizar variáveis explicativas para a construção dos seus modelos, o que pode possibilitar inúmeras análises de cenários diferentes dentro das empresas, possibilitando um melhor planejamento e auxiliando cada vez mais na tomada de decisão.

O modelo que se manteve mais estável entre todas as análises feitas foi o de Redes Neurais, mesmo utilizando diversos parâmetros, variáveis, e períodos de tempo o método se mostrou bem eficaz para realizar as previsões propostas.

Embora existam diversos fatores que podem influenciar na escolha de métodos preditivos, acredita-se que o trabalho tenha conseguido apresentar o desempenho dos métodos estatísticos propostos em diferentes situações e realçado os pontos fortes e fracos de cada um deles.

5.1. Trabalhos futuros

Deve-se atentar ao fato de que, como existe uma infinidade de parâmetros diferentes a serem testados para cada um dos métodos os resultados aqui obtidos podem não refletir em um estudo similar utilizando base de dados e parâmetros diferentes dos avaliados aqui. Por esse motivo sugere-se para outros trabalhos que sejam testados cada vez mais parâmetros a fim de consolidar os resultados obtidos.

Outro ponto para estudos futuros é o aumento do número séries a serem analisadas, com comportamentos mais distintos entre elas, bem como o estudo de outros métodos preditivos, visto que existem inúmeros métodos e modelos que possibilitam realizar previsões.

REFERÊNCIAS

BARBOSA, J. M. D. M. R. 2010. **Utilização de Árvores de Regressão Lineares para Avaliação de Segurança Dinâmica de Sistemas Interligados com Elevada Integração de Produção Eólica**. 2010. Dissertação de Mestrado - Faculdade de Engenharia da Universidade do Porto, Porto, 2010.

BAZIEWICZ, C. M., **Análise de uma metodologia de classificação de séries temporais para definição de métodos de previsão**. 2019. 135 f. Dissertação (Pós-Graduação em Métodos Numéricos em Engenharia). Universidade Federal do Paraná, Curitiba (PR), 2019

CORDEIRO, G. M., DEMÉTRIO,C. G. B., **Modelos Lineares Generalizados e Extensões**., Piracicaba, 2013

DUMAS, H., **COVID-19: Why Strategic Planning is Most Important in an Uncertain Economy?**. 2020. Não paginado. Disponível em: https://www.mntc.edu/about/news-publications/techtalkspost/~board/techtalks/post/covid-19-why-strategic-planning-is-most-important-in-an-uncertain-economy. Acesso em: 08 jun 2021

ESTATCAMP - Consultoria Estatística e Qualidade. Disponível em: http://www.portalaction.com.br/ . Acesso em: 05 mai 2021

FRONZETTI, Nicola. (2019). **Predictive Neural Network Applications for Insurance Processes.**

HAYES ADAMS. **Stepwise Regression**. Disponível em: https://www.investopedia.com/terms/s/stepwise-regression.asp . Acesso em: 03 dec 2021.

HYNDMAN, R.J., & ATHANASOPOULOS, G ,Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2020). _forecast: **Forecasting functions for time series and linear models**_. R package version 8.12, Disponível em: http://pkg.robjhyndman.com/forecast . Acesso em 08 mai 2021

HYNDMAN, R.J., & ATHANASOPOULOS, G. (2021) **Forecasting: principles and practice**, 3rd edition, OTexts: Melbourne, Australia. Disponível em: OTexts.com/fpp3. Acesso em 08 mai 2021

IPEADATA. Disponível em: http://ipeadata.gov.br/ . Acesso em: 08 jun 2021

IRB Brasil RE. Disponível em: https://www.irbre.com/ . Acesso em: 23 ago 2021

KLEINA, M. 2018. **PROGRAMAÇÃO EM R** . Curitiba : s.n., 2018.

MONTGOMERY, D. C. e RUNGER, G. C. . 1943. **Estatística Aplicada e Probabilidade para Engenheiros**. s.l. : LTC, 1943.

MOHD RAZALI, NORNADIAH & YAP, BEE. (2011). Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. J. Stat. Model. Analytics. 2.

NELDER, J. A. and WEDDERBURN, R. W. M. 1972. **Generalized linear models**. Journal of the Royal Statistical Society, Series A (Statistics in Society). 135 (3), pp. 370-384.

NELIZE F., **Estacionariedade das séries temporais do modelo matemático ARIMAX de Propulsoress Eletromecânicos**, 89 f, Dissertação de Mestrado (Mestrado em Modelagem Matemática), Departamento de Ciências Exatas e Engenharias, Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, 2018. Disponível em: https://bibliodigital.unijui.edu.br:8443/xmlui/bitstream/handle/123456789/5565/Nelize_Fracaro.pdf?sequence=1&isAllowed=y

OLIVEIRA, R. F. de, et al. 2016. **ANÁLISE DE MODELOS DE REGRESSÃO LINEAR MÚLTIPLA.** 2016. COBRAC 2016 - Anais. Florianópolis, UFSC, 2016

PAULA, A. G., **Modelos de Regressão com apoio computacional**, São Paulo, 2013

Pinho, Frank Magalhães, Modelos **SARIMA**, IBMEC/MG, 2019. Disponível em: https://rpubs.com/frank-pinho/535638 . Acesso em: 15 out 2021

POZZOLO, DAL A., **Comparison of Data Mining Techniques for Insurance Claim Prediction**. 2010. 81f. Tese (Mestrado em Sistemas de Informação para Negócios e Finanças) , Faculdade de Ciência Estatíctica, Bologna, 2010

R Core Team (2020). R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria. Disponível em: https://www.R-project.org/.

ROCHA, A. S., **Modelagem GLM aplicada à atuária: uma utilização dos Modelos Lineares Generalizados na precificação de seguros**. 64 f. Monografia de graduação (Bacharel em Ciências Atuariais) - Universidade Federal do Ceará, Fortaleza, 2015

SAMPAIO, I. G., BARNARDINI, F., PAES, A., ANDRADE, E. D. O., & Viterbo, J. **Avaliação de Modelos de Predição e Previsão Construídos por Algoritmos de Aprendizado de Máquina em Problemas de Cidades Inteligentes**. 2019. p. 81 – 113

SES SUSEP. Disponível em: https://www2.susep.gov.br/menuestatistica/SES/principal.aspx . Acesso em: 07 jun 2021.

SILVA, I.N.; SPATTI, D.H.; FLAUZINO, R.A. **Redes Neurais Artificiais para engenharia e ciências aplicadas.** Artliber, 2010

SIQUEIRA, P. H. **Metaheurísticas e Aplicações.** PPGMNE/UFPR. Curitiba, 2014
TURKMAN, M. A. A., SILVA, G. L., Modelos Lineares Generalizados - da teoria à prática, Lisboa, 2000

V, SELVKUMAR & SATPATHI, DIPAK & V., PRAVEEN & Vajjha, Haragopal. (2020). **Forecasting motor insurance claim amount using ARIMA model. AIP Conference Proceedings**. 2246. 020005. 10.1063/5.0014449.