INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Wallace Costa Faria

ANÁLISE PREDITIVA DA DEMANDA DE PASSAGEIROS POR TRANSPORTE AÉREO BRASILEIRO APÓS A PANDEMIA DO COVID-19

Trabalho de Graduação
2020

Curso de Engenharia Civil-Aeronáutica

CDU 656.71

Wallace Costa Faria

ANÁLISE PREDITIVA DA DEMANDA DE PASSAGEIROS POR TRANSPORTE AÉREO BRASILEIRO APÓS A PANDEMIA DO COVID-19

Orientador

Prof. Dr. Marcelo Xavier Guterres (ITA)

ENGENHARIA CIVIL-AERONÁUTICA

São José dos Campos
Instituto Tecnológico de Aeronáutica

2020

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Faria, Wallace Costa
Análise preditiva da demanda de passageiros por transporte aéreo brasileiro após a pandemia do COVID-19 / Wallace Costa Faria.
São José dos Campos, 2020.
49f.

Trabalho de Graduação – Curso de Engenharia Civil-Aeronáutica– Instituto Tecnológico de Aeronáutica, 2020. Orientador: Prof. Dr. Marcelo Xavier Guterres.

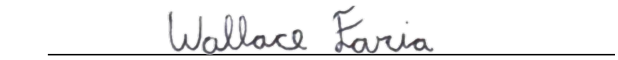
1. Transporte aéreo. 2. Transporte de passageiros. 3. Previsão. 4. Demanda (Economia). 5. Transportes. I. Instituto Tecnológico de Aeronáutica. II. Análise preditiva da demanda de passageiros por transporte aéreo brasileiro após a pandemia do COVID-19.

REFERÊNCIA BIBLIOGRÁFICA

FARIA, Wallace Costa. **Análise preditiva da demanda de passageiros por transporte aéreo brasileiro após a pandemia do COVID-19**. 2020. 49f. Trabalho de Conclusão de Curso (Graduação) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

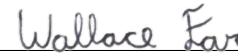
NOME DO AUTOR: Wallace Costa Faria
TITULO DO TRABALHO: Análise preditiva da demanda de passageiros por transporte aéreo brasileiro após a pandemia do COVID-19.
TIPO DO TRABALHO/ANO: Trabalho de Conclusão de Curso (Graduação) / 2020



Wallace Costa Faria
Rua H8B, 235
12228-461 – São josé dos Campos - SP

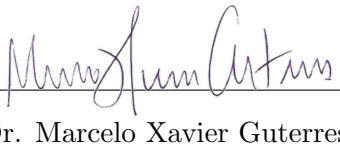
ANÁLISE PREDITIVA DA DEMANDA DE PASSAGEIROS POR TRANSPORTE AÉREO BRASILEIRO APÓS A PANDEMIA DO COVID-19

Essa publicação foi aceita como Relatório Final de Trabalho de Graduação



Wallace Costa Faria

Autor



Prof. Dr. Marcelo Xavier Guterres (ITA)

Orientador

Prof. Dr. João Cláudio Bassan de Moraes
Coordenador do Curso de Engenharia Civil-Aeronáutica

São José dos Campos, 13 de novembro de 2020.

Dedico esse trabalho ao professor Guterres por toda a orientação durante o ano.

Agradecimentos

Obrigado.

"Se queres prever o futuro, estuda o passado"
— Confúcio (552 a.c. - 489 a.c.)

Resumo

O setor aéreo brasileiro exerce um papel fundamental na economia do país e foi muito impactado pela pandemia causada pelo Covid-19, portanto a previsão para a retomada do setor é muito importante para planejamento das companhias aéreas. No presente estudo, foi utilizado modelos de Floresta Aleatória e SARIMA para estimar a demanda de passageiros no transporte aéreo no Brasil durante o período de pandemia causado pelo Covid-19. Para a construção do algoritmo de Floresta aleatória foram utilizadas variáveis macroeconômicas como o PIB, Ibovespa, taxa de câmbio do dólar, consumo de derivados do petróleo, taxa de desemprego e preço do petróleo. O modelo de SARIMA se mostrou mais eficiente do que o modelo de Floresta Aleatória em todos os cenários analisados, porém no período de crise ainda é necessário mais estudos para que possa ser aplicado.

Abstract

The Brazilian airline industry plays a key role in the country's economy and was greatly impacted by the pandemic caused by Covid-19, the forecast for the sector's recovery is very important for airline planning. In this study, Random Forest and SARIMA models were used to estimate the passenger air transport demand in Brazil during the pandemic period caused by Covid-19. For the construction of the Random Forest algorithm, the variables used were macroeconomic variables such as GDP, Ibovespa, dollar exchange rate, consumption of oil derivatives, unemployment rate and oil price. The SARIMA model proved to be more efficient than the Random Forest model in all analyzed scenarios, but in the crisis period more studies are still needed for it to be applied.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

O transporte aéreo exerce um papel importante na economia mundial, pois é o responsável por conectar as pessoas de diversas partes do país e do mundo por meio do turismo, comércio e cargas. A média de crescimento anual do número de passageiros é de 5,3% [11] se levado em consideração do ano de 2000 até o ano de 2018, sendo que em 2018 a quantidade total ultrapassou os 4,2 bilhões [11], como pode ser observado na Figura 1.1. O Brasil apresenta um dos maiores mercados domésticos do mundo em transporte aéreo. O setor teve um crescimento ainda maior do que a média mundial, sendo 6,8% a média do aumento do número de passageiros entre os mesmos anos, sendo em que 2018 a quantidade de passageiros do Brasil ultrapassou os 100 milhões [11].

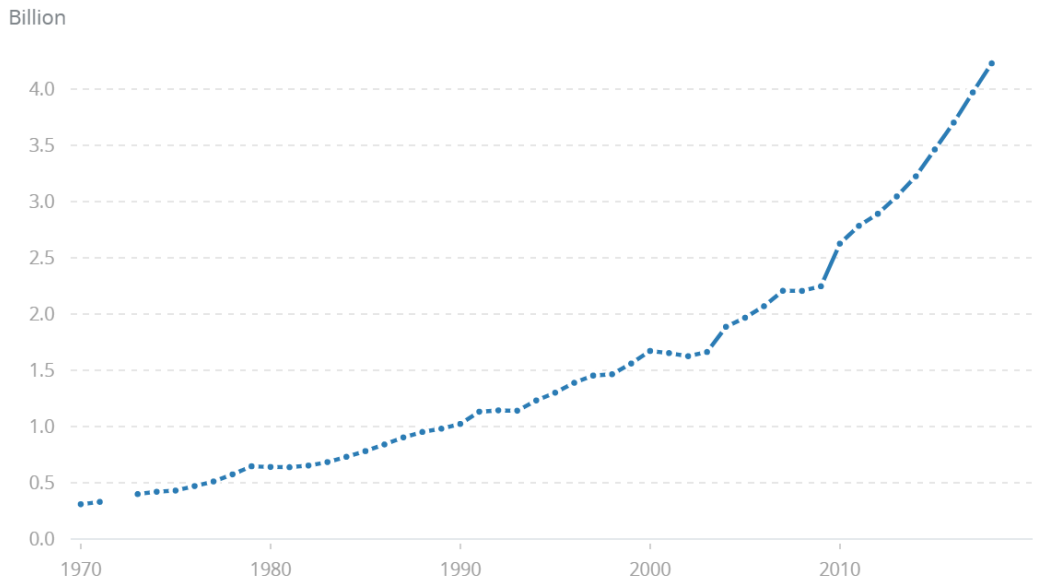


FIGURA 1.1 – Total de passageiros aéreos transportados por ano desde 1970.
Fonte: World Bank Data (2020) [11].

Devido ao alto custo de operação e outros fatores, o setor de transporte aéreo apresenta elevada complexidade, o que faz com que o mercado seja composto por poucas empresas. Atualmente no Brasil existem três empresas principais: Azul, GOL e LATAM. Sendo que

a quarta empresa, a Avianca, foi a falência ao final de 2019, o que acontece com muita frequência nesse mercado.

O surgimento da doença Covid-19, do coronavírus, ao final de 2019 gerou grandes consequências em toda a economia, principalmente no setor aéreo. No Brasil o vírus começou a ser alastrado no inicio de 2020, como resultado houve uma queda de 72% nas ações da LATAM, 72% nas ações da GOL e 76% nas ações da Azul ao se comparar o dia 5 de fevereiro de 2020 ao dia 2 de abril de 2020 [7], como pode ser observado pela Figura 1.2. E isso reforça o quanto essa área é afetada pelos fatores externos.

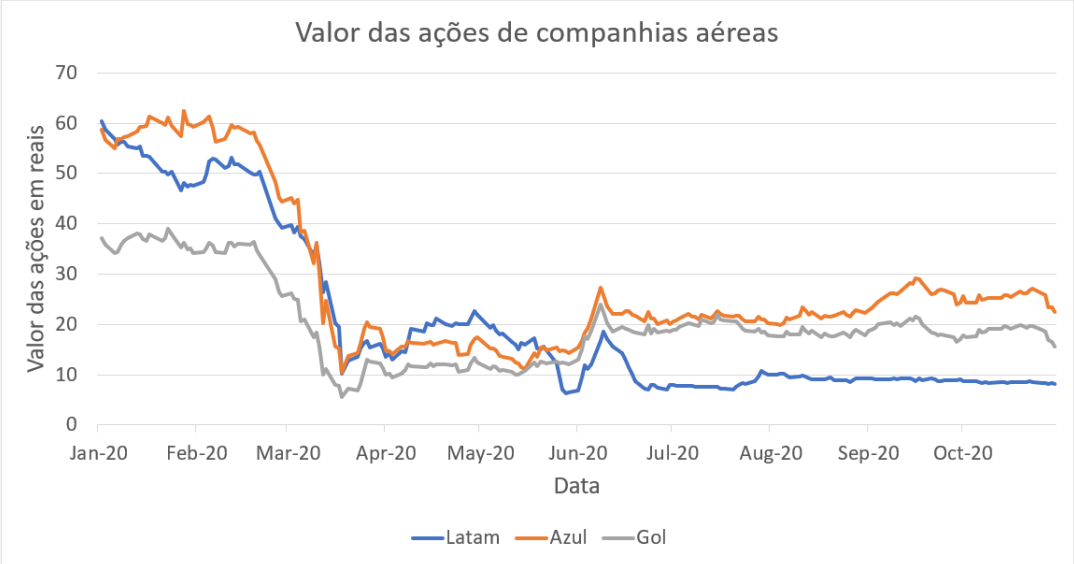


FIGURA 1.2 – Valor das ações das principais companhias aéreas do Brasil nos meses de fevereiro e março de 2020.

Fonte: Elaborado pelo autor a partir de dados da IBOV (2020) [7].

Conseguir prever a demanda de passageiros de determinada rota aérea sempre foi uma das principais métricas para auxiliar as companhias alocarem corretamente os seus recursos. Historicamente as decisões estratégicas de médio e longo prazo das empresas eram tomadas utilizando a econometria. Essa ferramenta nasceu como disciplina científica na década de 1930 e foi por muitas décadas a mais utilizada para estudar a relação entre variáveis de um sistema a partir da aplicação de um modelo matemático. Na década de 1960 foi introduzida a metodologia CAPM (Capital Asset Pricing Model), que originou-se no campo da economia como um modelo de para precificação de ativos. A teoria por trás desse modelo foi utilizado em vários outros estudos, inclusive pelo Instituto de Aviação Civil para o cálculo da demanda dos aeroportos feitos em 2005 [5]. Nas últimas décadas surgiram diversos modelos preditivos devido ao desenvolvimento recente da área de ciência de dados. O que possibilitou análises preditivas mais precisas, como exemplo temos as técnicas de rede neural e árvore de decisão.

Dada toda a complexidade do setor aéreo já citado e o momento atual da tecnologia em que é cada vez mais comum o uso de ciência de dados nas empresas de todas as áreas de atuação. Este trabalho visa criar uma ferramenta para auxiliar estudos e tomada de decisão dentro do setor aéreo brasileiro, sendo possível estimar a demanda de passageiros a partir de modelagens mais recentes, que nesse caso foi utilizado o modelo de SARIMA (*Seasonal Integrated Autoregressive Moving Average*) e Floresta Aleatória (*Random Forest*). O estudo também contempla o período de pandemia causada pelo coronavírus, em que a previsão da demanda se torna ainda mais complexa.

1.1 Problema

O setor aéreo requer um alto investimento para aplicações de operação e isso está associado a um alto risco, para se tomar uma decisão estratégica é preciso um estudo sobre a previsão do comportamento dos passageiros em diversas situações para maximizar o retorno da companhia aérea. Pois após tomada a decisão, como a compra de uma aeronave ou a criação de uma nova rota, é necessário investimento para fechar os contratos sobre a manutenção de toda a operação antes que esse investimento realmente passe a gerar lucros. Em um período de pandemia todas essas dificuldades do setor são amplificadas, gerando um grau ainda maior de incerteza. Portanto esse trabalho tem como objetivo auxiliar na solução do seguinte problema: qual é o modelo de previsão mais adequado para estimar a demanda de passageiros em um período de alto grau de incerteza?

1.2 Objetivo geral

O trabalho tem como objetivo criar uma ferramenta computacional capaz de prever a demanda de passageiros no transporte aéreo do Brasil em um período de alto grau de incerteza, que nesse estudo foi considerado a pandemia causada pela doença Covid-19, do coronavírus. Essa ferramenta pode ser utilizada tanto para as empresas aéreas a entender o comportamento dos passageiros para tomar decisões quanto para outros estudos acadêmicos na área, dado que esse assunto não é muito explorado na literatura brasileira. Para isso será criado uma modelagem de aprendizado de máquina utilizando o algoritmo de Floresta Aleatória e outra modelagem utilizando SARIMA para comparar o desempenho de cada uma na previsão de passageiros no período desejado.

1.3 Objetivos específicos

O trabalho tem como objetivo específico a exploração de ferramentas atuais de previsão para a previsão da demanda de passageiros do setor aéreo brasileiro. Para isso o trabalho pode ser separado nesses principais tópicos:

- Estudo do comportamento do setor aéreo brasileiro e das principais variáveis que regem esse mercado;
- Extração e tratamento da base de dados proveniente da ANAC e do Ipeadata;
- Modelagem matemática dos algoritmos de SARIMA e Floresta Aleatória para a previsão da demanda de passageiros no transporte aéreo brasileiro;
- Análise e interpretação de resultado de cada um dos modelos;
- Conclusão do trabalho.

1.4 Justificativas

A partir do uso de econometria e aprendizado de máquina é possível criar modelos que auxiliam a compreensão de determinado mercado, o que ajuda a tomada uma decisão estratégica. No setor aéreo em específico, em que é muito afetado pela economia e os riscos são altos devido ao elevado custo de manutenção de operação, uma análise específica desse mercado se torna essencial.

Além disso o Brasil sofre com uma falta de uma literatura abrangente em relação a aplicação de ciência de dados na área da aviação, portanto esse trabalho também tem como objetivo uma exploração inicial para ser utilizado como base para outros estudos acadêmicos.

1.5 Limitação do tema

Como o objetivo do trabalho é o estudo da demanda de passageiros de transporte aéreo no Brasil, todos os dados utilizados são de rotas aéreas nacionais, utilizando dados a partir do ano de 2000 fornecidos pela ANAC [4]. Durante a crise causada pelo Covid-19 muitas rotas aéreas deixaram de existir ou foram interrompidas, diante disso o trabalho utiliza como base os dados agregados mensais da soma da demanda de passageiros em todas as rotas. Sendo que, para a aplicação desses algoritmos em rotas específicas é necessário uma readequação dos modelos.

1.6 Estrutura do trabalho

Este trabalho de graduação é constituído por 6 capítulos sendo organizados da seguinte forma:

No Capítulo 1 é feito a introdução do assunto com o cenário atual do setor aéreo brasileiro, os desafios desse mercado, o objetivo do trabalho, a justificativa e as limitações.

O Capítulo 2 apresenta uma visão geral de previsão em séries temporais e é detalhado as modelagens utilizadas nesse trabalho: SARIMA e Floresta Aleatória. Para cada um dos modelos é feito uma introdução, histórico do desenvolvimento do modelo e o funcionamento analítico.

O Capítulo 3 apresenta a metodologia utilizada no trabalho. É apresentado todas as etapas que foram seguidas para a construção de ambas as modelagens, incluindo extração e tratamento de dados, criação e calibração dos parâmetros.

O Capítulo 4 apresenta os resultados encontrados, incluindo uma análise individual de cada uma das modelagens utilizadas para aplicações antes e depois da crise causada pelo Covid-19. Em seguida é feita a comparação entre esses modelos e a interpretação do resultado.

O Capítulo 5 apresenta a conclusão do trabalho, que é a análise final da viabilidade de aplicação de cada uma das modelagens utilizadas para previsão da demanda de passageiros no transporte aéreo brasileiro.

2 Modelagens preditivas em séries temporais

Série temporal é um conjunto de observações de um fenômeno ordenadas no tempo. Essas séries podem ser observadas em diversos campos da ciência, como as taxas de juros e produto interno bruto na economia, registros de temperatura e índice pluviométrico na meteorologia, quantidade de clientes e de faturamento de uma industria, etc. Como mostrado na Figura 2.1, uma série temporal pode ser genericamente composta em quatro itens [22]:

- **Tendência**: refere-se aos elementos de longo prazo relacionados com a série de tempo;
- **Ciclo**: refere-se à variações com certo grau de liberdade, porém com período não bem definido;
- **Sazonalidade**: refere-se à variações regulares na séria temporal
- **Resíduo**: refere-se à variações aleatórias, que não apresentam nenhum tipo de regularidade

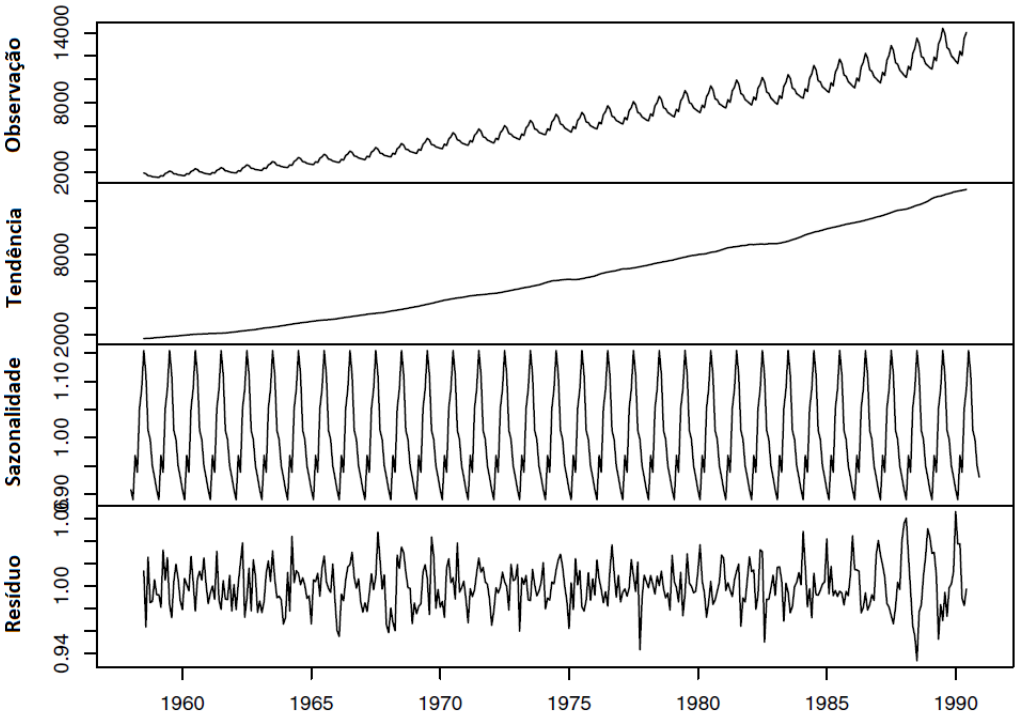


FIGURA 2.1 – Série temporal separada em tendência, sazonalidade e resíduo.
Fonte: Adaptado de Paul S. P. Cowpertwait, Andrew V. Matcalfe (2009) [23].

A análise de séries temporais tem o objetivo de identificar cada um desses itens para que seja possível a previsão de dados futuros da série. Hoje, existem diversas metodologias para auxiliar na previsão dessas séries.

Os modelos para previsão em série temporais podem ser divididos essencialmente em dois tipos: os estatísticos e os baseados em inteligência artificial. Como exemplo de modelagens estatísticas temos a família de algoritmo de Box & Jenkins, que são genericamente denominados de ARIMA, modelo autorregressivo integrado e de média móvel (*autoregressive integrated moving average*). Como exemplo de modelagens baseadas em inteligência artificial podemos citar os algoritmos de rede neural, árvore de decisão e todas as suas variações.

2.1 Análise de desempenho de modelos preditivos

A análise de desempenho de previsões em séries temporais apresenta algumas peculiaridades em comparação com as modelagens de categorização. A principal delas é que em série temporais não podemos separar os dados aleatoriamente entre base de treinamento e de teste, pois a ordem dos elementos fazem muita diferença na análise. Nesse caso ao fazer a previsão em um determinado período, não é possível utilizar quaisquer informações que originaram após esse mesmo período. Portanto a base de teste representa necessariamente

os dados iniciais e a base de teste os dados finais em relação a série temporal estudada. Normalmente a base de teste representa entre 20% a 30% do total de dados. Um esquema do cálculo pode ser visualizado em 2.2.

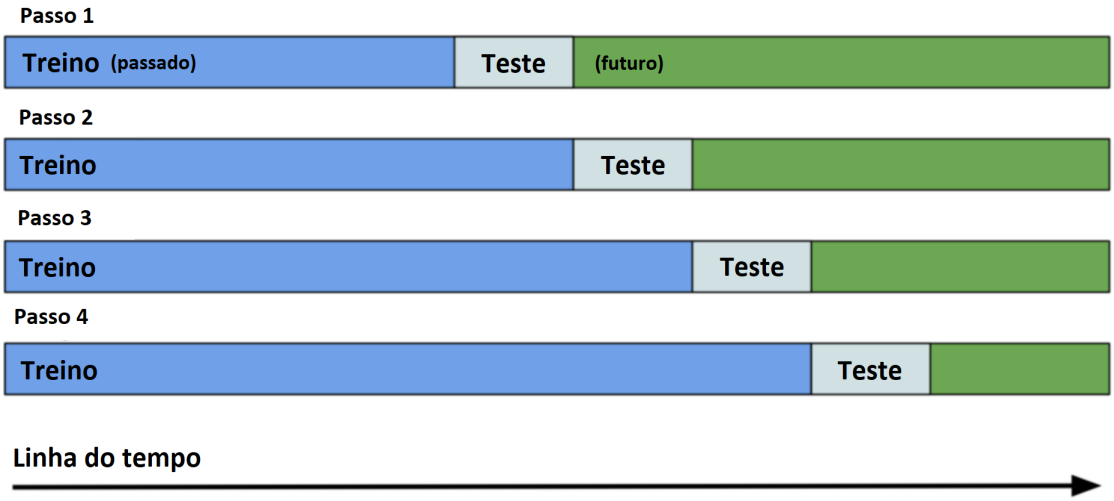


FIGURA 2.2 – Série temporal separada em tendência, sazonalidade e componente irregular.

Fonte: Elaborado pelo autor.

Existem diversos métodos que podem ser utilizados para metrificar a precisão de um modelo de previsão. O mais simples é o MAPE (*Mean Absolute Percentage Error*), que é simplesmente a média do módulo do erro percentual de cada elemento, a vantagem do método é resultado encontrado é simples de ser interpretado. O MAPE é definido pela expressão 2.1:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_{pred,i} - y_i}{y_i}\right| \tag{2.1}$$

Outro método utilizado é o RMSE (*root-mean-square error*), que é calculado pela raiz quadrada do erro quadrático médio, a vantagem do método é que consegue representar melhor a acurácia do modelo, sendo o mais utilizado para comparar a performance entre modelagens distintas, a desvantagem é que não é intuitivo a interpretação desse resultado, mesmo estando na mesma dimensão dos dados analisados. O RMSE é definido pela expressão 2.2:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{pred,i} - y_i\right)^2} \tag{2.2}$$

Além dessas métricas existem outras que podem ser utilizadas para medir a performance de um modelo preditivo, como a MSE (*Mean Squared Error*) e MAE (*Mean Abso-*

lute Error). Porém essas métricas não são muito utilizadas nesse contexto e por isso não foram exploradas.

2.2 Floresta Aleatória

Floresta Aleatória é um algoritmo de aprendizado supervisionado capaz se executar tarefas de regressão e de classificação. Como o próprio nome indica, o método se baseia na combinação de árvores de decisão, utilizando o resultado individual de cada árvore de decisão em conjunto para obter um resultado final.

Esse algoritmo é muito utilizado devido à sua simplicidade em comparação com os outros métodos de machine learning existentes, além de obter bons resultados para diversos tipos de aplicações. Como exemplos de usos podemos citar a previsão de churn de clientes em seguro de saúde, previsão de vendas de uma empresa, classificações de tipos de plantas a partir das suas dimensões e a identificação de spam em uma caixa de e-mails.

O modelo de Floresta Aleatória apresenta algumas vantagens em relação á outros métodos como a habilidade para a modelagem de relações dimensionais altamente não lineares; a utilização de variáveis categóricas e contínuas; a resistência ao sobreajuste (*overfitting*); a relativa robustez ante a presença de "ruídos"nos dados; e a exigência de poucos parâmetros para ser implementado [6].

Por outro lado uma das desvantagens dessa dessa aplicação é a limitação da interpretação dos resultados, já que as relações entre os preditores e as respostas não podem ser examinadas individualmente [6]. Outra limitação da modelagem é que uma grande quantidade de árvores, para deixar o método mais preciso, pode tornar o algoritmo ineficiente para as predições em tempo real.

Outra característica relevante do modelo de Floresta Aleatória é a possibilidade de verificar as correlações que as variáveis de entrada têm com o resultado previsto a partir do uso de importância de característica (*feature importance*), assim como outros modelos de aprendizado de máquina. Pois, existem casos em que é igualmente importante não apenas ter um modelo preciso, mas também interpretável. Para que seja possível além de conseguir prever o resultado também entender as variáveis que devemos alterar para obter um resultado melhor

2.2.1 Histórico

O método geral de florestas de decisão aleatória foi proposto pela primeira vez por Ho em 1995 [18]. Ho estabeleceu que as florestas de árvores que se dividem com hiperplanos oblíquos podem ganhar precisão à medida que crescem, desde que as florestas sejam

aleatoriamente restritas para serem sensíveis apenas a dimensões de recursos selecionados. Um trabalho subsequente [19] nas mesmas linhas concluiu que outros métodos de divisão se comportam de maneira semelhante, desde que sejam forçados aleatoriamente a serem insensíveis a algumas dimensões de recursos.

A introdução de florestas aleatórias propriamente dita foi feita pela primeira vez em um artigo de Leo Breiman [20]. O desenvolvimento inicial da noção de Breiman de florestas aleatórias foi influenciado pelo trabalho de Amit e Geman [2], que introduziram a ideia de pesquisar um subconjunto aleatório das decisões disponíveis ao dividir um nó, no contexto do crescimento de uma única árvore. A ideia de seleção aleatória de subespaços de Ho [19] também foi influente no projeto de florestas aleatórias. Neste método, uma floresta de árvores é cultivada e a variação entre as árvores é introduzida projetando os dados de treinamento em um subespaço escolhido aleatoriamente antes de ajustar cada árvore ou cada nó. Finalmente, a ideia de otimização de nó aleatório, onde a decisão em cada nó é selecionada por um procedimento aleatório, ao invés de uma otimização determinística foi introduzida pela primeira vez por Dietterich [13]

2.2.2 Funcionamento

Árvores de decisão é um método que utiliza uma representação gráfica baseada em árvores, em que o objetivo é identificar grupos de indivíduos com características em comum. Para isso, é utilizado um método recursivo que divide a amostra inicial em subamostras, baseando-se em resultados observados das variáveis preditoras. Dessa forma formam-se grupos para os quais a variável resposta apresenta comportamento homogêneo dentro dos grupos e heterogêneo entre eles [**friedman**].

Existem dois tipos de árvores de decisão, a árvore de classificação, se a variável de resposta for categórica, e árvore de regressão, se a variável de resposta for numérica [29]. Nesse trabalho será abordado apenas a árvore de regressão, pois ela será a base para o modelo de Floresta Aleatória utilizado.

Para utilizar o método de árvore de regressão é necessário ter dados de treinamento sobre o objeto de estudo para que seja ajustado os parâmetros utilizados no modelo. Para isso é necessário que para cada observação seja conhecida as variáveis preditoras e a variável de resposta. Sendo que, no geral, quanto mais observações forem selecionadas para o treinamento da árvore de regressão, maior vai ser a precisão da modelagem.

O processo de indução de árvores é iniciado por meio de uma amostra, denominada nó raiz, que é dividida em subamostras, denominadas nós intermediários ou nó filhos. Essas subamostras quando subdivididas são chamadas de nós pais, pois geram nós filhos. Quando uma subamostra não for mais dividida por um critério de parada, é então deno-

minada de nó final ou nó folha. Esse processo é considerado como recursivo devido a cada subamostra gerar novas subamostras. A estrutura de uma árvore de decisão genérica é exemplificada em 2.3.

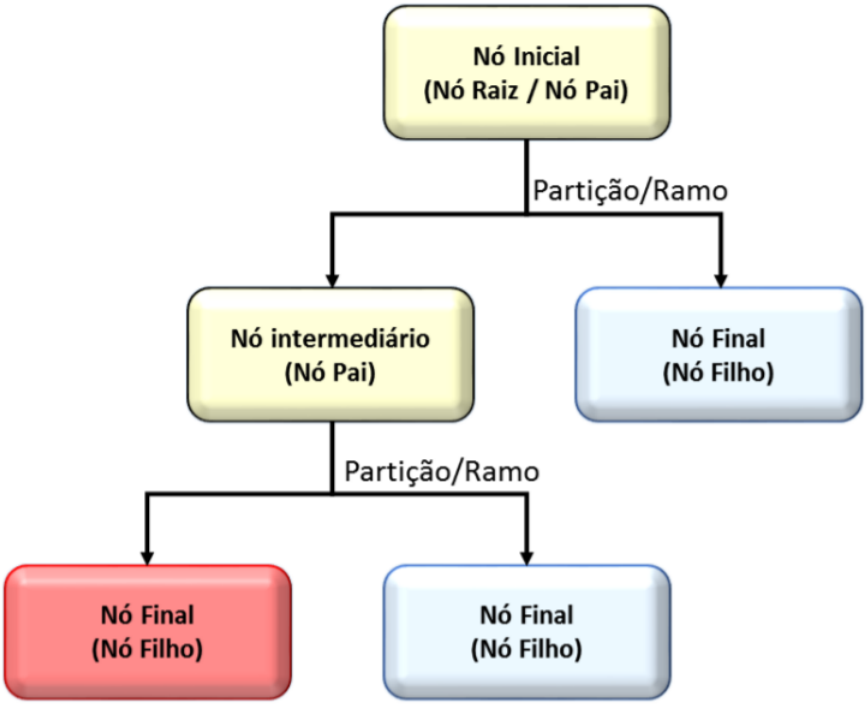


FIGURA 2.3 – Estrutura de uma árvore de decisão genérica.

Fonte: Donadia, D. D. E (2013) [14].

Portanto temos que os itens principais da construção de uma árvore de regressão se resume nos seguintes pontos [3]:

1. Determinação de todas as possíveis divisões de um nó para cada variável do espaço de predição;
2. Seleção da divisão que melhor atende os critérios utilizados;
3. Determinação de quando se deve considerar o nó como terminal;
4. Atribuição de um valor resposta a cada nó terminal.

Analiticamente podemos descrever a variável de resposta como $Y \in \mathbb{R}^1$ e as variáveis preditoras como $X = (X_1, X_2, ..., X_p) \in \mathbb{R}^p$, sendo p a quantidade de variáveis preditoras para cada observação. Na Tabela 2.1 é possível visualizar os dados necessários para a utilização da modelagem de árvore de regressão, sendo n a quantidade de observações utilizadas no treinamento.

TABELA 2.1 – Dados genéricos para treinamento de uma árvore de regressão.

Observação	X_1	X_2	$\cdots$	X_p	Y
1	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,p}$	Y_1
2	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{2,p}$	Y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	$X_{n,1}$	$X_{n,2}$	$\cdots$	$X_{n,p}$	Y_n

Uma outra forma de interpretar uma árvore de regressão é através da região formada por cada uma das folhas finais da árvore dentro do espaço $\mathbb{R}^P$, formando as regiões R_j [**insper**]. Sendo que, assim como nas folhas, para cada região é atribuído um valor para a variável resultado. A exemplo de visualização a Figura 2.4 representa as regiões formadas por uma árvore de regressão em que existem apenas 2 variáveis preditivas.

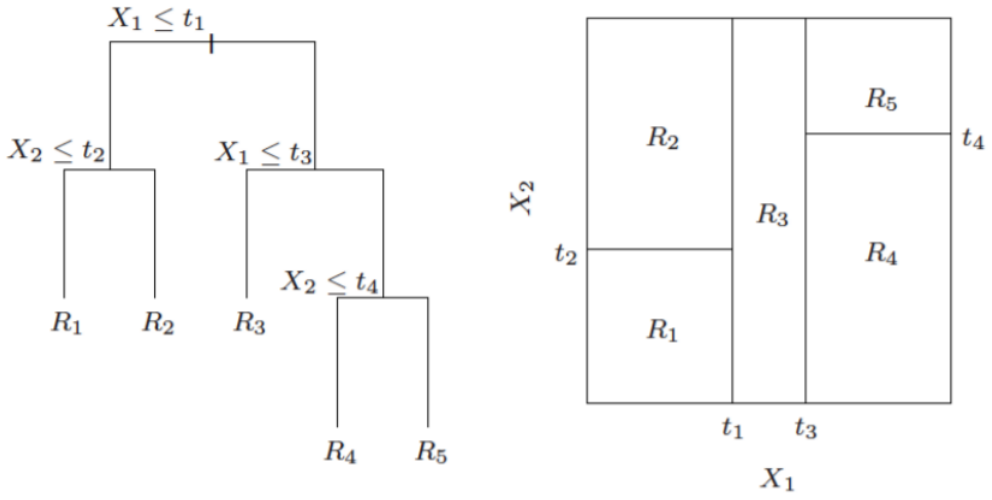


FIGURA 2.4 – Árvore de regressão e a partição do espaço das preditoras nas regiões correspondentes.

Fonte: Hastie, T. Tibshirane, R. Friedman, J. (2017) [**insper**].

Dessa forma podemos interpretar a aplicação de um método de árvore de decisão como sendo a identificação da região R_j de determinada observação , seguida da atribuição da constante relacionada a essa região, o que equivale a uma das folhas finais da estrutura da árvore. Isso pode ser expresso através da Equação 2.3.

$$f(x) = c_j, \ x \in R_j \tag{2.3}$$

A forma de determinar a constante c_j para cada região R_j é um fator relevante para precisão da árvore. Para árvores de regressão, é usual considerar a minimização da soma de quadrados de resíduos (SQR) como critério para a definição desse valor [30], como pode

ser observado na Equação 2.4.

$$SQR = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{2.4}$$

A partir desse critério de divisão é possível calcular que a melhor escolha para o valor da constante c_j da região R_j é simplesmente a média aritmética da variável resposta de todas as observações que pertencem a essa região [30], como é representado na Equação 2.5.

$$c_j = \frac{1}{n_j} \sum_{x_i \in R_j} y_i \tag{2.5}$$

Portanto, temos que a maneira com que um nó é dividido depende exclusivamente da escolha da variável preditora que será utilizada. Podemos representar de forma genérica a divisão de uma região em outras duas pela Equação 2.6.

$$R_1 = \{X \in \mathbb{R}^p : X_k \leq c_j\}, \; R_2 = \{X \in \mathbb{R}^p : X_k > c_j\} \tag{2.6}$$

A avaliação da partição de um nó se dá pelo cálculo da soma de quadrados de resíduos para cada uma das variáveis preditoras disponíveis e escolhida a variável que apresentar o menor resultado [30]. O que pode ser representado pela Equação 2.7.

$$k = arg\ min \left(\sum_{x_i \in R_1} (y_i - c_1)^2 + \sum_{x_i \in R_2} (y_i - c_2)^2 \right) \tag{2.7}$$

A regra de partição de um nó é aplicada sucessivamente aos novos nós até que atinja algum critério de parada, que normalmente é definido pelo usuário que está aplicando o modelo, como número mínimo de observações por no, altura máxima da árvore, etc.

Floresta Aleatória, como o nome sugere, é a utilização de várias árvores de decisão para encontrar um único resultado a partir do resultado de cada árvore individualmente. Cada árvore é treinada a partir de entradas selecionadas aleatoriamente, isso é possibilitado pela utilização da metodologia *bagging* (*bootstrap aggregating*) para gerar os subconjuntos de treinamento. Dado um conjunto de características M de tamanho t, o *bagging* produz subconjuntos de características M_i, cada um com tamanho t', amostrando M uniformemente e com reposição, o que significa que uma característica $m \in M$ pode ser uma amostra em mais de um M_i. A utilização de *bagging* tem o objetivo de reduzir a variância de previsões e evitar o sobreajuste, que é uma característica comum nos métodos de Floresta Aleatória.

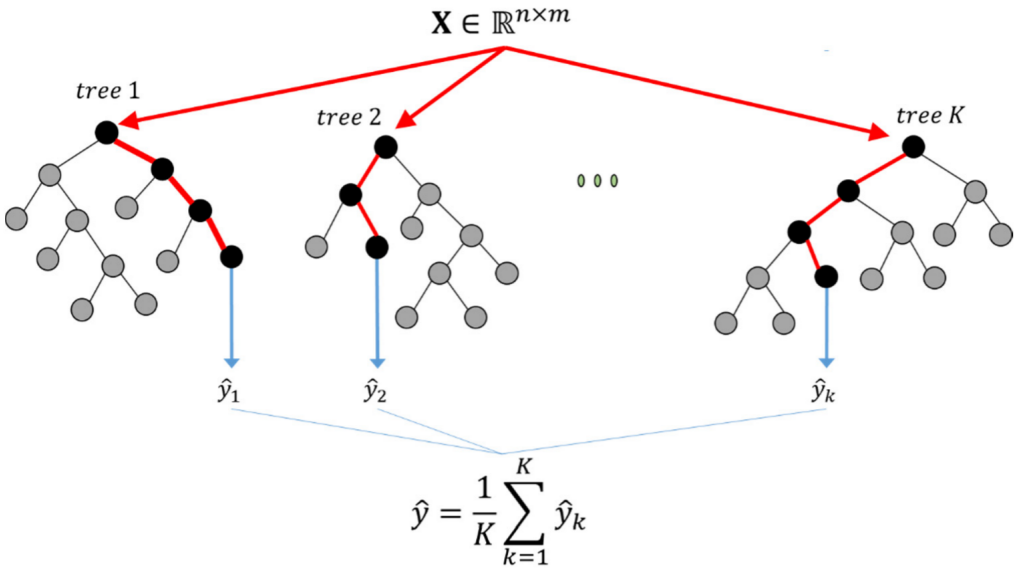


FIGURA 2.5 – Representação de uma árvore aleatória.
Fonte: Aldrich, Chris (2020) [1].

Após aplicado o método de *bagging* para a criação das subamostras e o treinamento de cada árvore de decisão, o resultado do modelo se dá pela combinação de resultados de cada árvore individual, como é representado na Figura 2.5. No caso de problemas de regressão o resultado final costuma ser a média dos resultados de cada uma das árvores.

2.3 SARIMA

O modelo de SARIMA, autorregressivo integrado de médias móveis sazonais (*Seasonal Autoregressive Integrated Moving Average*), é um caso especial do modelo ARIMA, em que é adicionador termos referente a sazonalidade da série. O ARIMA é uma generalização de um modelo de ARMA, autorregressivo de média móveis. Esses modelos são utilizados para prever dados futuros em uma série temporal utilizando modelos estatísticos lineares [17]. Como exemplos de aplicações em artigos científicos podemos citar a previsão da tarifa de consumo de energia elétrica [10], previsão da velocidade do vento no dia seguinte [27] e a previsão da inflação da moeda de um país [**aidan**].

O SARIMA é uma ferramenta muito conhecida e utilizada pela praticidade do uso e pelo bom resultado obtido na maioria dos casos, além de ser simples os testes formais para testar a adequação de um modelo. Como desvantagem é necessário uma quantidade relativamente grande de dados para a utilização do modelo e não existem métodos simples para recalcular os parâmetros na inclusão de novos dados, sendo necessário desenvolver um novo modelo.

2.3.1 Histórico

A associação de dados em séries temporais é objeto de estudo há mais de um século. O primeiro modelo denominado autorregressivo (AR) foi desenvolvido por Yule [31] em 1927, no qual o valor previsto dependia dos valores anteriores. Porém a modelagem era limitada devido a dependência da existência de uma relação linear entre os elementos da sequência, sendo que a maioria das séries reais apresentam uma forte tendência de não linearidade [9]. Os estudos só passaram a serem frequente a partir da década de 1950 com o surgimento do computador.

Nessa busca por modelos mais precisos para previsão em séries temporais o Box e o Jenkins [8] publicaram um livro demonstrando um sistema matemático que tem como base a Teoria Geral de Sistemas Lineares em 1976, que levou a criação de um modelo que leva o nome de Box & Jenkins (B&J). A principal evolução em relação aos modelos anteriores foi a suposição de que a passagem de um ruído branco por um filtro linear de memória infinita gera um processo estacionário de segunda ordem, o que já demonstrava um resultado satisfatório. De forma genérica esses modelos são conhecidos como ARIMA, porém outras variações do modelo, como o SARIMA, também foram desenvolvidos pelo Box & Jenkins. Com o desenvolvimento dessa área foram criados modelos que permitem a inclusão de variáveis exógenas, como o exemplo do NARIMAX (*Non-linear Autoregressive with Moving Average and Exogenous inputs*) [15]. Esses conjuntos de modelos formam uma família que pode ser representada pela Figura 2.6.

2.3.2 Funcionamento

O modelo de SARIMA é um caso especial do modelo de ARIMA em que é adicionado a sazonalidade da série. O modelo de ARIMA tem como premissa básica que a série temporal é gerada por um processo estocástico cuja natureza pode ser representada através de um modelo. A notação empregada para designação do modelo é normalmente ARIMA (p, d, q) onde p representa o número de parâmetros autorregressivos, d o número de diferenciações para que a série torne-se estacionaria e q o número de parâmetros de médias móveis. Casos particulares são o modelo ARMA(p, q), o modelo autorregressivo AR(p) e o modelo de médias móveis MA(q), todos para séries temporais estacionárias ($d = 0$) [21].

O modelo autorregressivo de ordem p, AR(p), é usado quando há autocorrelações entre as observações [16], ou seja, o processo autorregressivo é usado quando o valor da observação t, representado por y_t, é gerada pela média ponderada das p primeiras observações próximas anteriores da variável acrescida de um erro aleatório ε_t, podemos

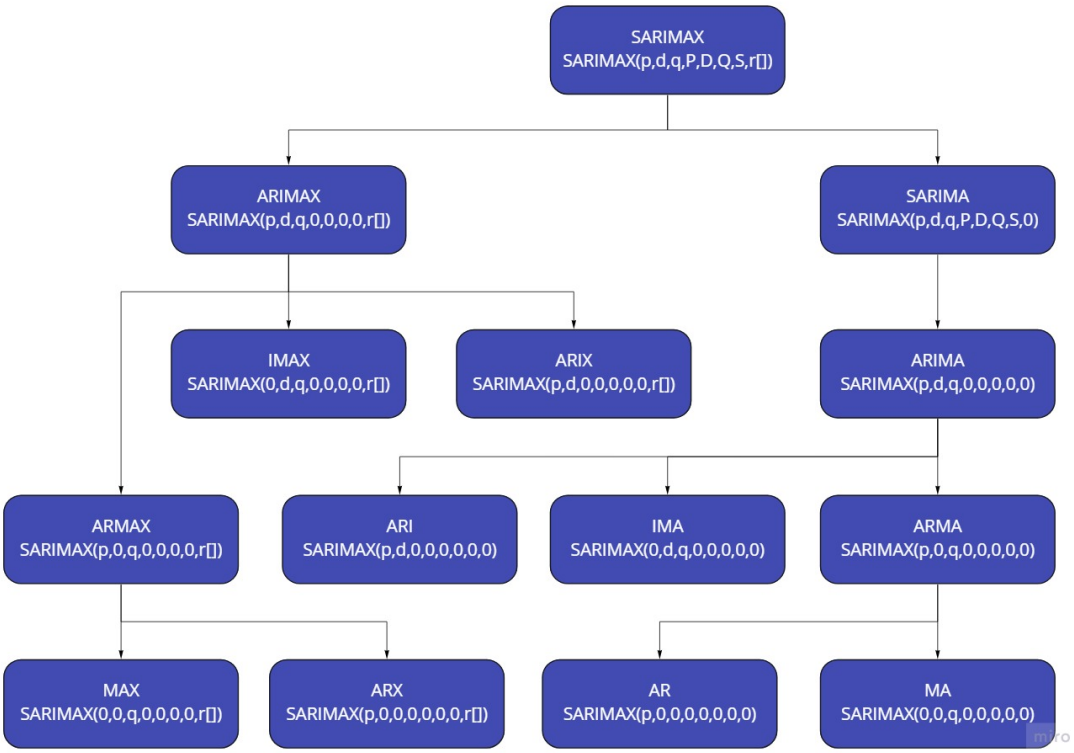


FIGURA 2.6 – Diagrama de subdivisões do modelo ARIMAX.
Fonte: Elaborado pelo autor.

representar o modelo AR(p) pela Equação 2.8.

$$y_t = \phi_1.y_{t-1} + \phi_2.y_{t-2} + \cdots + \phi_p.y_{t-p} + \varepsilon_t \tag{2.8}$$

Onde ϕ é um parâmetro do modelo, o ε_t é o termo do erro e y_i representa o valor da observação no instante i. Para o processo ser estacionário, a média (μ) deve ser constante e a condição necessária (mas não suficiente) de estacionariedade é dada por:

$$y_t = \phi_1 + \phi_2 + \cdots + \phi_p < 1 \tag{2.9}$$

O modelo de média móvel de ordem q, MA(q), é usado quando há autocorrelação entre os resíduos. Ou seja, a observação y_t é gerada pela média ponderada dos q primeiros valores passados de um processo de "ruído branco"[12], podemos representar o modelo AR(p) pela Equação 2.10.

$$y_t = \varepsilon_t - \theta_1.\varepsilon_{t-1} - \theta_2.\varepsilon_{t-2} - \cdots - \theta_q.\varepsilon_{t-q} \tag{2.10}$$

Onde θ é um parâmetro do modelo e ε_i é o ruído do período i. Nesse caso a condição de estacionariedade é verificada se o valor de q for finito.

O modelo autorregressivo de média móvel, ARMA (p, q), é usado quando há autocorrelação entre as observações e autocorrelação entre os resíduos. Nesse caso temos que a observação y_t é gerada tanto pela média ponderada das p primeiras observações próximas anteriores da variável, quanto pela média ponderada dos q primeiros valores passados de um processo de "ruído branco". Podemos representar o modelo ARMA (p, q) pela equação 2.11.

$$y_t = \phi_1.y_{t-1} + \cdots + \phi_p.y_{t-p} + \varepsilon_t - \theta_1.\varepsilon_{t-1} - \cdots - \theta_q.\varepsilon_{t-q} \tag{2.11}$$

Ou, escrevendo de outra forma:

$$(1 - \phi_1.B - \cdots - \phi_p.B^p).y_t = (1 - \theta_1.B - \cdots - \theta_q.B^q).\varepsilon_t \tag{2.12}$$

Onde B é o operador de defasagem ($B.y_t = y_{t-1}$), ϕ e θ são parâmetros do modelo e ε_t o termo de erro. Em que as variáveis defasadas de y_t representam a parte AR do modelo enquanto que as defasagens de ϵ_t representam a parte MA do modelo. Nesse caso, a média (μ) e as condições de estacionariedade são dadas pela parcela AR do processo, sendo que para a aplicação desse modelo é necessário que o processo seja estacionário.

O modelo ARIMA surge da generalização do modelo ARMA para as séries não necessariamente estacionárias. Portanto se as observações y_t forem geradas tanto pela média ponderada das p primeiras observações próximas anteriores da variável, quanto pela média ponderada dos q primeiros valores passados de um processo de "ruído branco", e ainda pertencerem a um processo não estacionário, será necessário diferenciar a série original dos dados d vezes até se obter uma série estacionária. Podemos representar o modelo ARIMA (p, d, q) pela equação 2.13.

$$\phi_p(B).\Delta^d.y_t = \theta_q(B).\varepsilon_t \tag{2.13}$$

Em que $\Delta^d = (1 - B)^d$. A interpretação do valor de d na série representa o comportamento de tendência da curva. No caso em que $d = 1$ a série não é estacionária quanto ao nível de processo, ou seja, quando a média durante um período é diferente da média de um outro período da série. No caso em que $d = 2$ a série também não é estacionária em relação a inclinação, isso é, quando a direção da série se altera entre um período e outro. Não é comum o uso dessa modelagem com $d >= 3$ [21].

O modelo de SARIMA surge da adição da sazonalidade no modelo de ARIMA. Para isso é adicionado um termo referente a sazonalidade para cada expressão que compõem o modelo de ARIMA, portanto o termo Φ_P é o operador sazonal de AR(p), o termo Θ_Q é o operador sazonal de MA(s) e o termo Δ_S^D é o operador sazonal de de diferenciação (d).

Os termos sazonais são representados pelas Equações 2.14, 2.15 e 2.16.

$$\Delta_S^D = (1 - B^S)^D \tag{2.14}$$

$$\Phi_P(B^S) = 1 - \Phi_1(B^S) - \Phi_2(B^{2S}) - \cdots - \Phi_P(B^{PS}) \tag{2.15}$$

$$\Theta_Q(B^S) = 1 - \Theta_1(B^S) - \Theta_2(B^{2S}) - \cdots - \Theta_Q(B^{QS}) \tag{2.16}$$

Sendo que S é o período da sazonalidade. Utilizando a Equação 2.13 do ARIMA e adicionando o termos sazonais chegamos a modelagem de SARIMA $(p, d, q)(P, D, Q)_S$, que pode ser representada pela Equação 2.17.

$$\phi_p(B).\Phi_P(B^S).\Delta^d.\Delta_S^D.y_t = \theta_q(B).\Theta_Q(B^S).\varepsilon_t \tag{2.17}$$

3 Metodologia

Inicialmente foi feito uma revisão de literatura em relação à demanda de passageiros no transporte aéreo brasileiro e sobre as previsões de dados em séries temporais, em seguida foi feito as escolhas das modelagens utilizadas. Como os modelos podem ser divididos essencialmente em estatísticos e os que utilizam inteligência artificial foi definido que seria utilizado um tipo de cada modelo. Das modelagens estatísticas foi escolhido um modelo da família ARIMA que mais se adequava a situação, por se tratar de uma demanda altamente sazonal foi escolhido o método de SARIMA, como pode ser observado pela Figura 3.1.

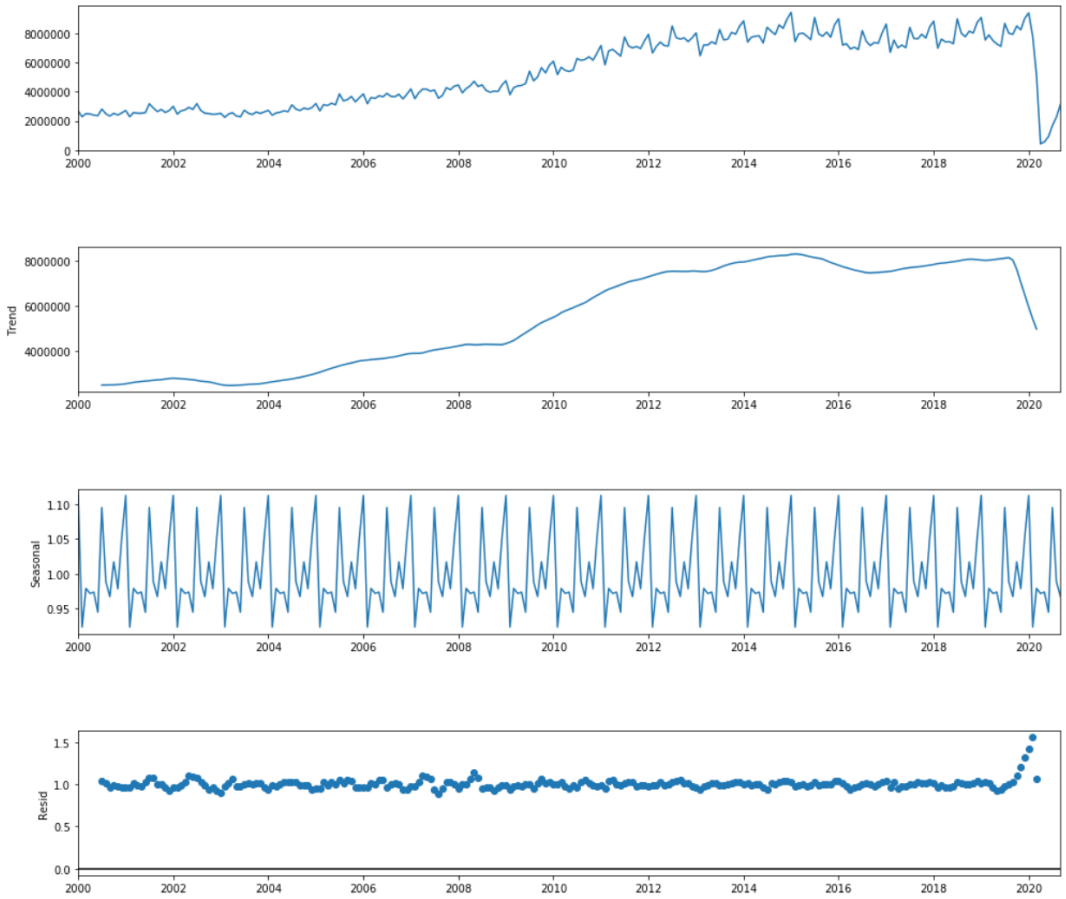


FIGURA 3.1 – Série da demanda de passageiros no transporte aéreo separada em tendência, sazonalidade e resíduo.

Fonte: Elaborado pelo autor.

Das modelagens baseadas em inteligência artificial foram feitos alguns testes prévios sem otimizações de parâmetros para avaliar a adequação de cada um. Ao final foram testados os algoritmos de Floresta Aleatória, LightGBM e Redes Neurais LSTM (*Long Short Term Memory*). Dentre esses algoritmos o de Floresta Aleatória foi o que demonstrou uma melhor adaptação ao problema proposto.

Após a escolha das modelagens utilizadas, foi aplicado a metodologia normalmente utilizada em ciência de dados, que se resume a extração e tratamento de dados, desenvolvimento das modelagens, otimizações de parâmetros, análise de resultados e conclusões. Para a análise de resultado foi aplicado cada um dos modelos a base de teste e comparado o resultado encontrado com o valor real do período utilizando o RMSE, que foi detalhado na seção 2.1. O fluxograma da Figura 3.2 apresenta a sequência metodológica do trabalho e em seguida é detalhado cada uma dessas etapas.

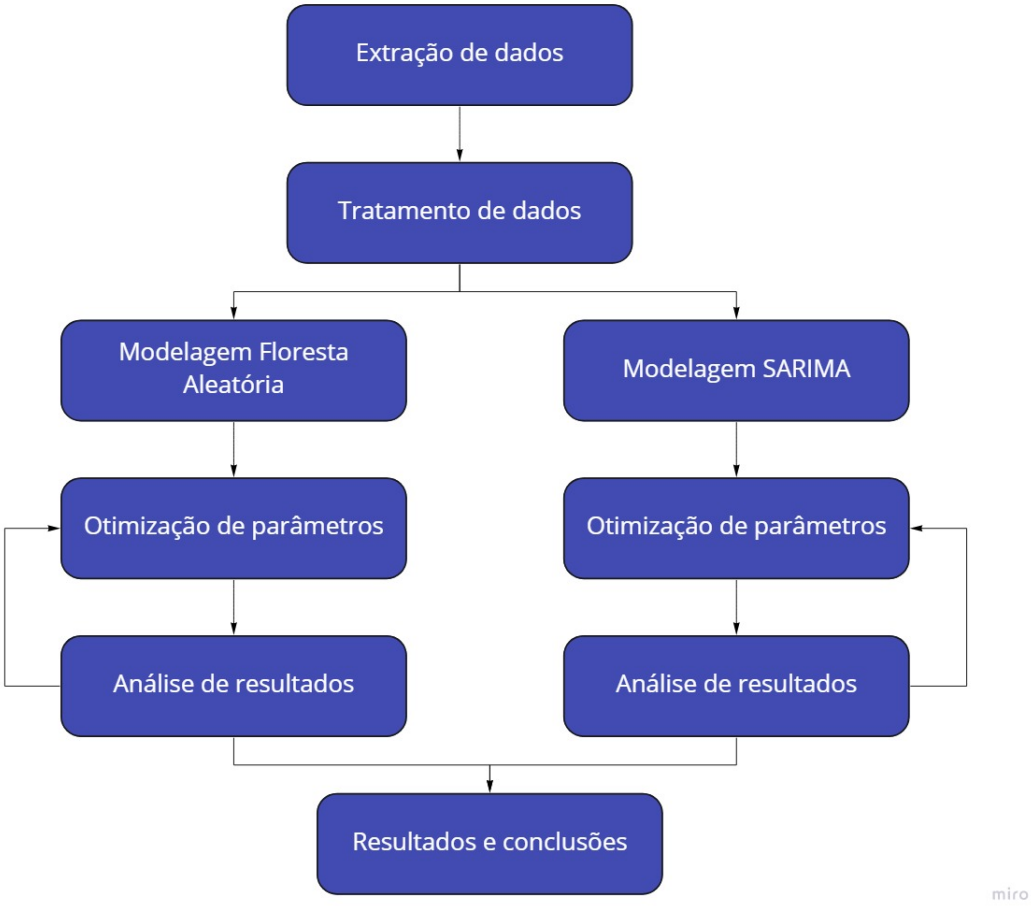


FIGURA 3.2 – Fluxograma do trabalho.

Fonte: Elaborado pelo autor.

3.1 Extração de dados

Para a construção dos modelos utilizados na previsão, faz-se necessário os dados referentes ao histórico da demanda de passageiros no transporte aéreo além de outros dados externos, principalmente econômicos, que exercem influência na demanda de passageiros, para serem utilizados no treinamento da modelagem de Floresta Aleatória. Portanto no presente trabalho foi utilizado as seguintes fontes de informações:

- Agência Nacional de Aviação Civil (ANAC) : base de dados presente na seção "dados e estatística"[4];
- Instituto de Pesquisa Econômica Aplicada (Ipea): base de dados do Ipeadata - Dados Macroeconômicos [24].

A extração dos dados referente a demanda de passageiros no transporte aéreo foi feito a partir do site da ANAC, essa base de dados contem informações desde janeiro de 2000 até hoje e os dados são agrupados por mês, rota aérea e companhia aérea. Para a utilização de variáveis exógenas na modelagem de Floresta Aleatória, foi extraído informações econômicas e sociais a nível federal do Ipeadata. A seguir é apresentado os códigos do Ipeadata a descrição de cada uma dos itens extraídos:

- BM12_PIB12: PIB;
- ANBIMA12_IBVSP12: Índice de ações - Ibovespa;
- BM12_ERCF12: Taxa de câmbio - R$ / US$ - comercial - compra;
- PRECOS12_IPCATC12: IPCA;
- ANP12_CODP12: Consumo aparente - derivados de petróleo;
- ECONMI12_JPU12: Taxa de desemprego - força de trabalho;
- EIA366_PBRENT366: Preço - petróleo bruto.

Com exceção do preço do petróleo, que a atualização dos dados é diária, o restante dos dados é atualizado mensalmente. Todos os dados foram extraídos do Ipeadata utilizando a biblioteca em R disponibilizada por eles.

3.2 Tratamento de dados

Para o tratamento da base extraída da ANAC, foi selecionada todas as rotas cujo aeroporto de origem e de destino fosse localizado no Brasil e em seguida os dados foram agrupados por mês, somando a quantidade de passageiros que decolaram em cada período. Para a aplicação do modelo de SARIMA, os dados provenientes da ANAC foram os únicos utilizados, pois essa modelagem utiliza apenas os dados históricos da série e não permite a inserção de variáveis exógenas.

Para a aplicação do algoritmo de Floresta Aleatória foi necessário um tratamento mais extensivo dos dados extraídos. Das informações obtidas pela base do Ipeadata, os dados referente ao preço do petróleo foi tratado a parte pelo fato de ter atualização diária, foi selecionado apenas o último valor registrado em cada um dos meses. Em seguida foi feito uma defasagem de um mês em todos os dados extraídos do Ipeadata e foi criada uma coluna nova com a demanda de passageiros no mês anterior. Pois o algoritmo de Floresta Aleatória apenas utilizará dados referente ao mês anterior para estimar a demanda de determinado mês.

Ao final foi unificado os dados referente a demanda de passageiros no transporte aéreo e todos os dados extraídos do Ipeadata, resultando em uma base de dados contendo 247 linhas, uma para cada mês (de 02/2000 até 08/2020), e 11 colunas: uma referente ao mês e ano, uma referente apenas ao mês (utilizada no modelo de Floresta Aleatória), uma referente a demanda de passageiros no mês, uma referente a demanda de passageiros no mês anterior, e outras sete contendo informações econômicas e sociais do mês anterior, extraídas do Ipeadata. Os dados normalizados podem ser visualizados pela Figura 3.3.

Após testes utilizando a coluna mês como uma variável categórica na modelagem de Floresta Aleatória verificou-se que o modelo não tem um bom comportamento com esse tipo de variável em comparação com as variáveis contínuas. Dessa forma, exclusivamente para a modelagem de Floresta Aleatória, foi feito uma transformação dessa coluna categórica em doze colunas contínuas: mes_i, com i variando de 1 até 12, utilizando a seguinte lei de formação:

$$\begin{cases} 1, se\ i = n \\ 0, se\ i \neq n \end{cases} \tag{3.1}$$

Sendo n o número do mês já contido na coluna "mes" da base de dados e i o índice da coluna mes_i. A vantagem dessa transformação é que dessa forma o modelo de Floresta Aleatória consegue trabalhar com cada mês de forma independente, utilizando apenas os meses em que há relevância para o resultado final.

Com o objetivo de metrificar e comparar o desempenho dos dois modelos utilizados, foi

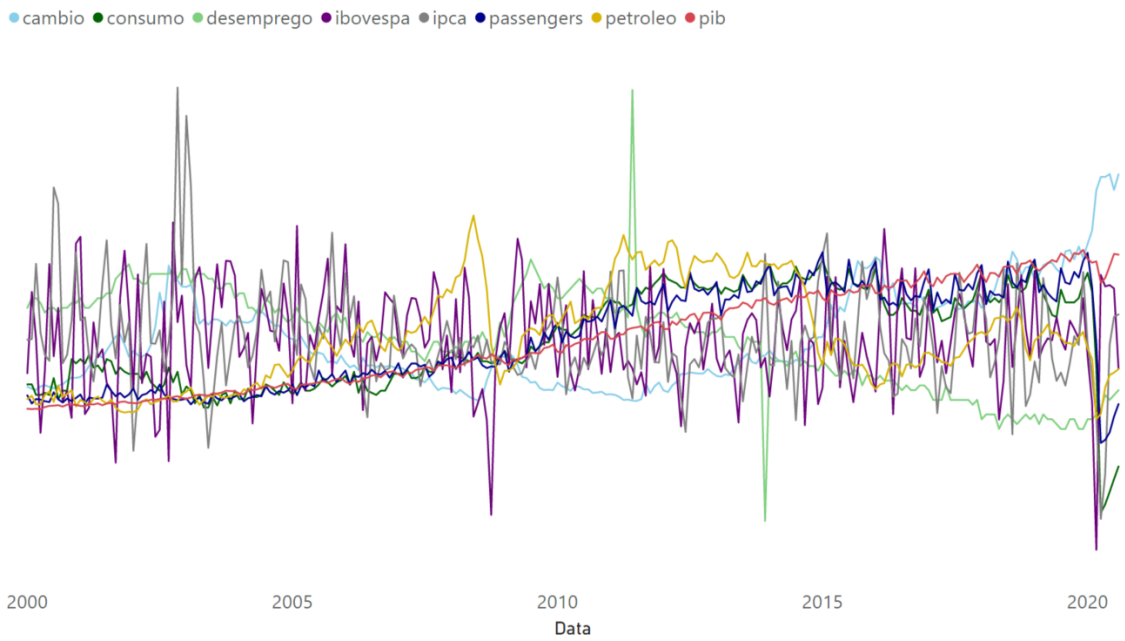


FIGURA 3.3 – Evolução de cada variável selecionada para a modelagem de Floresta Aleatória.

Fonte: Elaborado pelo autor.

definido a mesma base de treinamento e de teste para ambos os modelos. Para a divisão da base de dados foi considerada a proporção de 70% de treinamento e 30% de teste.

- Base de treinamento: de 02/2000 até 05/2014
- Base de teste: de 06/2014 até 08/2020

Para medir o desempenho de cada um dos modelos será utilizado o período da base teste, em que será comparado o resultado obtido pelos modelos e o valor real. Esse período pode ser dividido em três partes que apresentam comportamentos distintos devido ao Covid-19: período pré pandemia (de 06/2014 até 02/2020), período de início de pandemia (de 03/2020 até 04/2020) e período de pandemia (de 05/2020 até 08/2020). A Figura 3.4 representa o gráfico da divisão de cada período. Sendo que a análise de resultado será feita separada para o período de pré pandemia e o período de pandemia.

3.3 Floresta Aleatória

Para a aplicação do algoritmo de Floresta Aleatória foi utilizada a linguagem de programação python na versão 3.8.6 e a biblioteca sklearn [25], que já apresenta a estrutura básica da modelagem. Na modelagem de Floresta Aleatória de regressão do sklearn existem 18 parâmetros que podem ser alterados, desses pode-se destacar 8 principais:

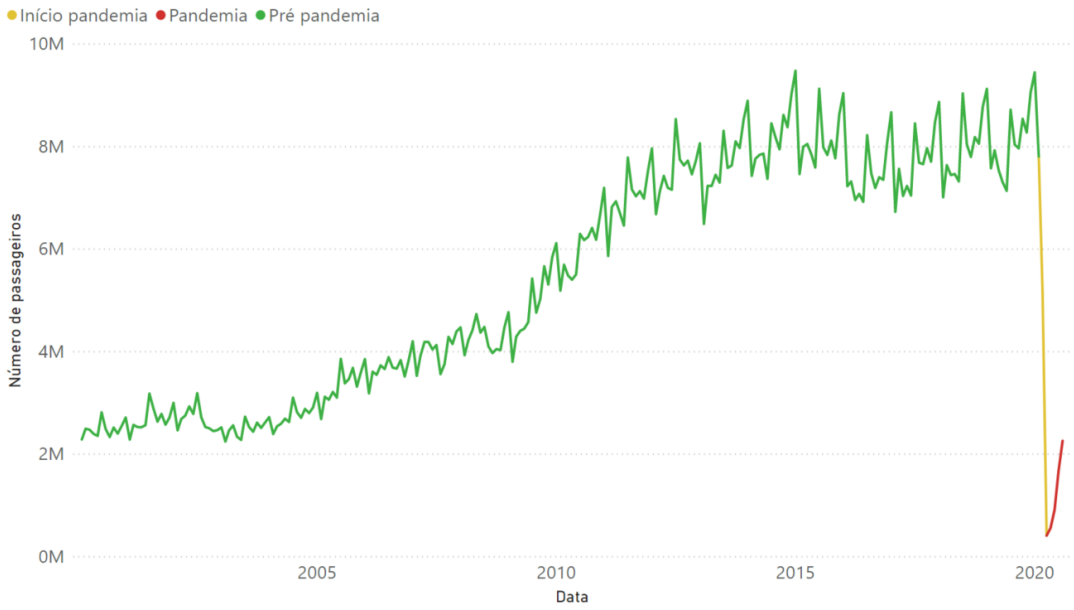


FIGURA 3.4 – Divisão do período estudado em pré pandemia, inicio de pandemia e pandemia.

Fonte: Elaborado pelo autor.

- $n\_estimators$: número de árvores na floresta;
- max_depth: profundidade máxima de cada árvore;
- $min\_samples\_split$: quantidade mínima de amostras em cada nó para permitir uma nova divisão;
- $min\_samples\_leaf$: quantidade mínima de amostras por nó;
- $max\_features$: quantidade máxima de características por árvore;
- max_leaf_nodes: quantidade máxima de folhas na árvore;
- *bootstrap*: aplicação ou não de *bagging*;
- *criterion*: critério de divisão do nó, pode ser MSE ou MAE.

Dentre os parâmetros citados foram fixados o *bootstrap* como sendo verdadeiro, para garantir a variabilidade das árvores da floresta, e o *criterion* como RSE, pois é o critério mais utilizado para árvores de decisão, como detalhado na seção 2.1. Foram feitos alguns testes com a variação do parâmetro $n\_estimators$ e foi identificado que para um número de árvores acima de 500 não há uma diferença expressiva no resultado do modelo, portanto foi fixado a quantidade de 500 árvores na floresta. O restante dos parâmetros foi definido a partir de um processo de otimização definido pelo fluxograma da Figura 3.5. O objetivo desse processo é encontrar uma modelagem de Floresta Aleatória que não apresente em sobreajuste em cima da base de dados e que mantenha uma boa precisão final. Para esse

trabalho foi definido que a modelagem não está sobreajustada se nenhuma característica do modelo tiver uma relevância de mais de 40%.

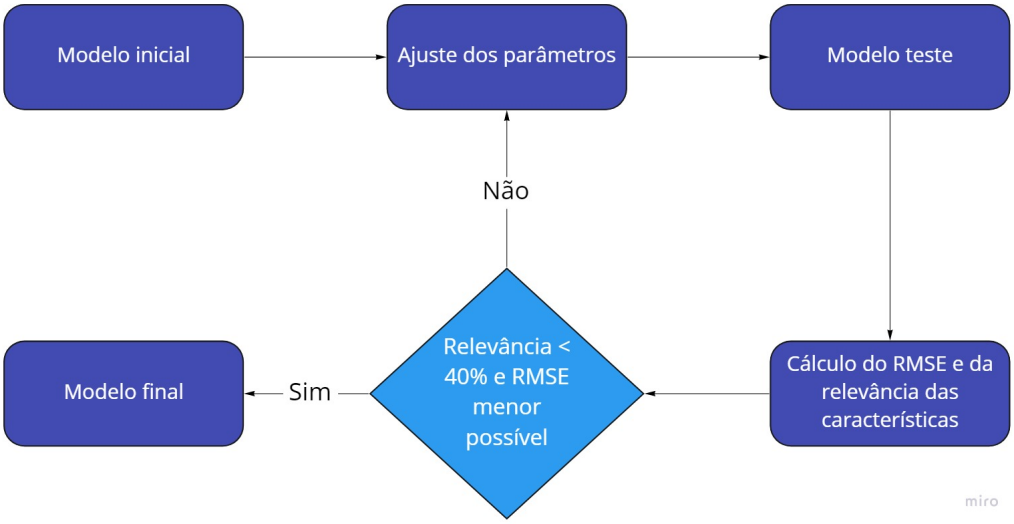


FIGURA 3.5 – Fluxograma de otimização dos parâmetros da modelagem de Floresta Aleatória.

Fonte: Elaborado pelo autor.

O principal parâmetro alterado para atingir a relevância de característica de no máximo 40% foi o $max\_features = 8$, pois limita a quantidade de características por árvore o que aumenta a distribuição a relevância de cada característica no modelo. Dos parâmetros utilizados como critério de parada na divisão dos nós foram utilizados os parâmetros $min\_samples\_split = 5$, para não dividir nós que já tem poucas observações, e $max\_depth = 9$, para que a árvore não tenha uma profundidade elevada e utilize apenas as características mais relevantes. Os outros parâmetros do modelo não foram necessários. A Figura 3.6 apresenta um trecho de uma das árvores de regressão obtidas pela modelagem final da Floresta Aleatória.

3.4 SARIMA

Para a aplicação do algoritmo de SARIMA foi utilizada a linguagem de programação python na versão 3.8.6 e a biblioteca statsmodels [26], que já apresenta a estrutura básica da modelagem. Na modelagem de SARIMA é necessário definir sete parâmetros que foram apresentados na Seção 2.3.2: p, q, d, P, Q, D e S. Desses parâmetros é possível definir o valor de S, pois é o período de sazonalidade, portanto temos que $S = 12$.

Para a otimização dos demais parâmetros foi utilizado uma função de autoajuste presente na biblioteca pmdarima [28]. O funcionamento do algoritmo é com base em testes de todas as possibilidades possíveis dentro dos limites estabelecidos, inicialmente é feito

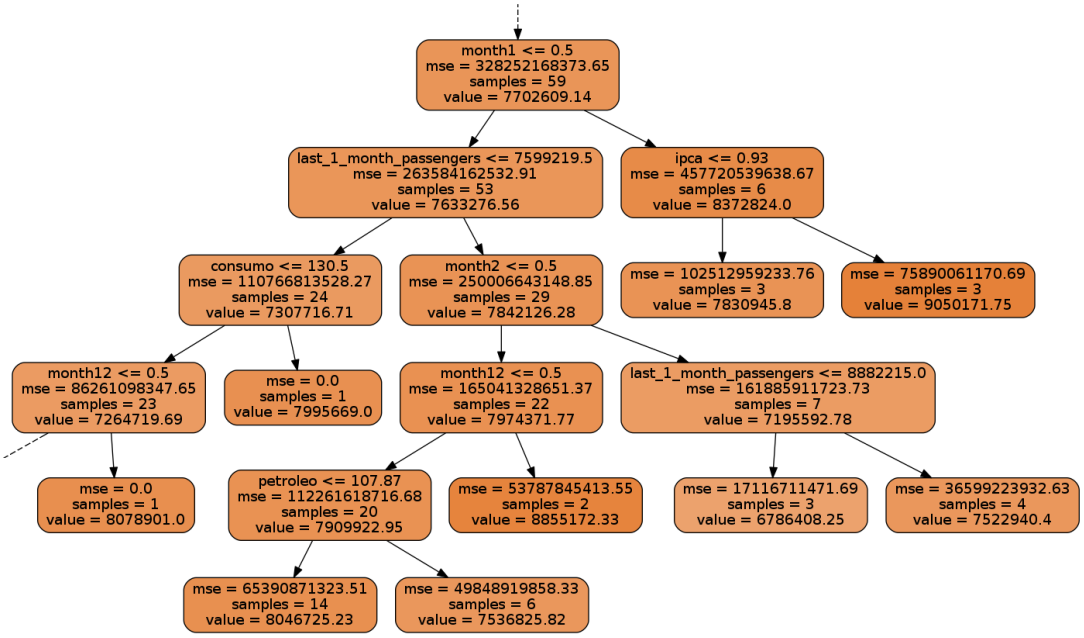


FIGURA 3.6 – Trecho de uma árvore de regressão presente na modelagem final da Floresta Aleatória

Fonte: Elaborado pelo autor.

testes de diferenciação para determinar o parâmetro d e em seguida é determinado os parâmetros p e q. Nessa aplicação, em que a componente sazonal está habilitada, é encontrado os parâmetros D, P e Q utilizando o mesmo método. Como resultado foi obtido os seguintes parâmetros: $p = 2$, $d = 1$, $q = 2$, $P = 1$, $D = 0$ e $Q = 1$. Portanto o modelo final obtido é o SARIMA$(2, 1, 2)(1, 0, 1)_{12}$.

4 Resultados e discussões

Na modelagem de Floresta Aleatória, em que foi utilizado dados macroeconômicos como variáveis exógenas, é possível identificar a importância de cada uma das características utilizadas para obtenção do resultado final, como observado na 4.1. Pode-se identificar que as variáveis mais relevantes no modelo são o consumo de derivados do petróleo do mês anterior, demanda de passageiros no mês anterior e PIB do mês anterior. A curva do consumo de petróleo segue um comportamento muito similar ao da própria curva de demanda de passageiros, como pode ser observado na Figura 3.3, por isso a alta correlação. O PIB do país está muito relacionado com o poder de compra da população, então a variação desse índice econômico reflete diretamente na demanda de passageiros, essa correlação já é descrita na literatura.

TABELA 4.1 – Importância das características no modelo de Floresta Aleatória.

Característica	**Importância**
Consumo de derivados do petróleo	0.381
Demanda de passageiros	0.286
PIB	0.180
Preço do Petróleo	0.060
Taxa de desemprego	0.040
Taxa de câmbio - R$ / US$	0.030
IPCA	0.006
Mês de janeiro	0.005
Ibovespa	0.004
Mês de Julho	0.003
Mês de dezembro	0.002
Mês de fevereiro	0.002

Após o ajuste dos parâmetros dos algoritmos de SARIMA e Floresta aleatória, ambos os modelos foram aplicados para a previsão da demanda de passageiros no período de teste, definido entre junho de 2014 e agosto de 2020. A curva da demanda história de passageiros e de previsão de cada uma das modelagens podem ser visualizadas pela Figura

4.1.

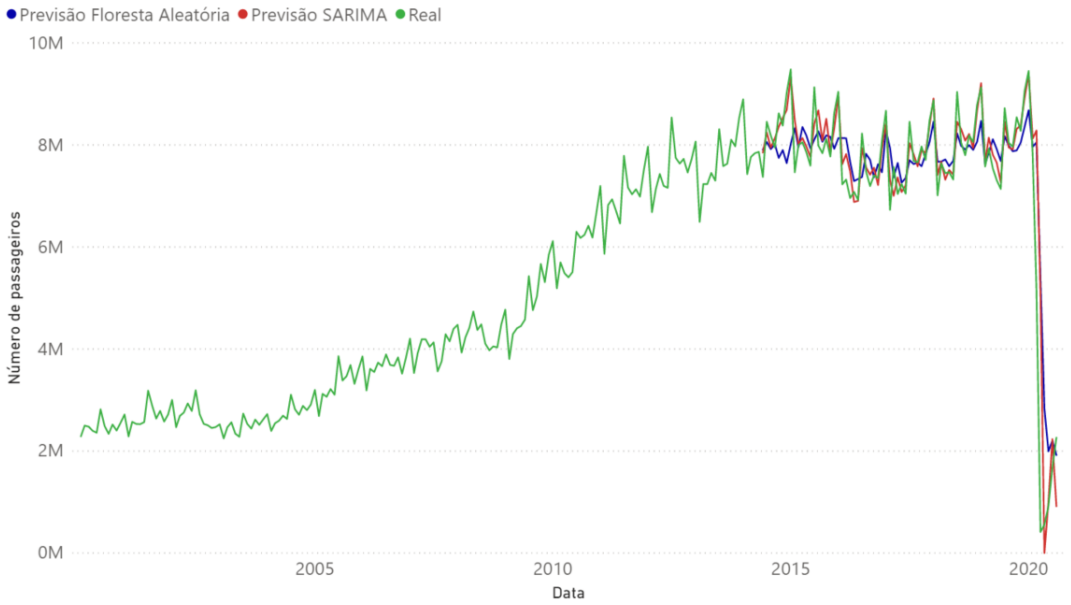


FIGURA 4.1 – Gráfico da evolução da demanda de passageiro no transporte aéreo e as previsões obtidas pelos modelos de SARIMA e Floresta Aleatória.

Fonte: Elaborado pelo autor.

Para a análise de resultados, o intervalo estudado foi dividido entre o período de pré pandemia e o de pandemia, para melhor interpretar o resultado em cada um dos cenários. A partir da Figura 4.2 é possível avaliar o comportamento de ambos os modelos quando submetidos a previsão de demanda antes do período de crise. O modelo de SARIMA apresenta um melhor ajuste em relação a curva real, principalmente nas regiões de picos e de vales, enquanto que o resultado do algoritmo de Floresta Aleatória normalmente é restrito a um intervalo menor de valores.

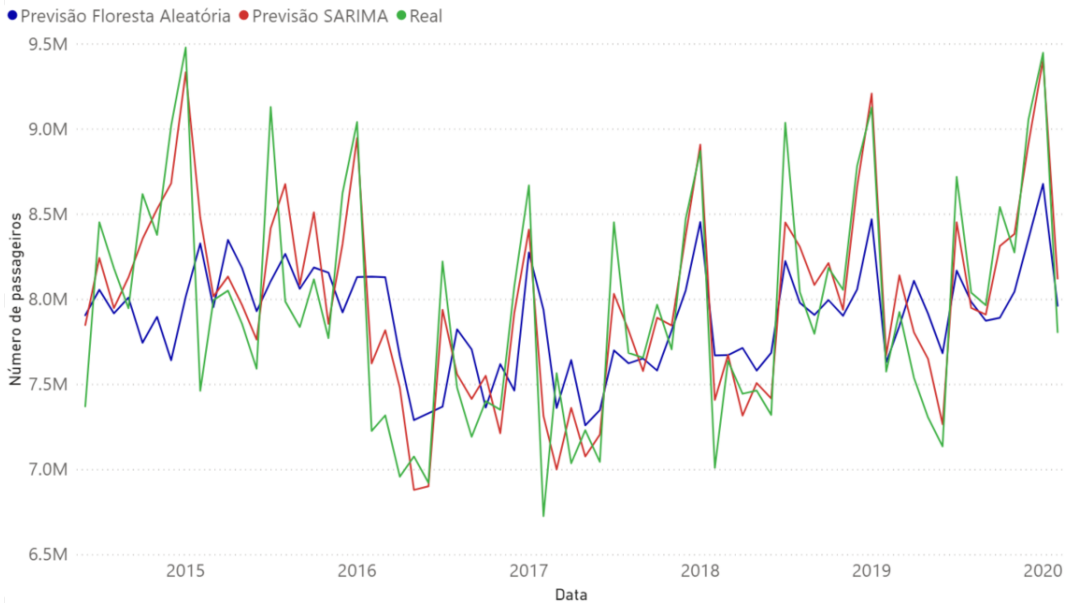


FIGURA 4.2 – Gráfico das previsões obtidas pelos modelos no período anterior ao início da crise causada pelo Covid-19.

Fonte: Elaborado pelo autor.

A Figura 4.3 representa o desempenho de ambos os modelos a partir do momento em que a crise causada pelo Covid-19 se inicia no Brasil, a linha amarela representa a transição do período de início de pandemia para o período de pandemia. É possível observar que, nos meses de início de crise (março e abril de 2020), ambos os modelos retornaram valores próximos entre si, porém distante do valor original. No período de crise é que observamos a maior diferença entre as modelagens, em que o modelo de Floresta Aleatória se mostrou mais conservador em relação a mudança, reduzindo pouco a previsão de demanda de um mês para o mês seguinte, mesmo na situação em que os valores reais dos meses anteriores tenham sido muito mais baixos que os previstos. O modelo de SARIMA se mostrou mais adaptável e, apesar que prever uma demanda próxima de 0 no mês de maio, acompanhou melhor a curva da demanda real, logo no segundo mês de pandemia o algoritmo de SARIMA obteve uma previsão muito próxima ao valor real.

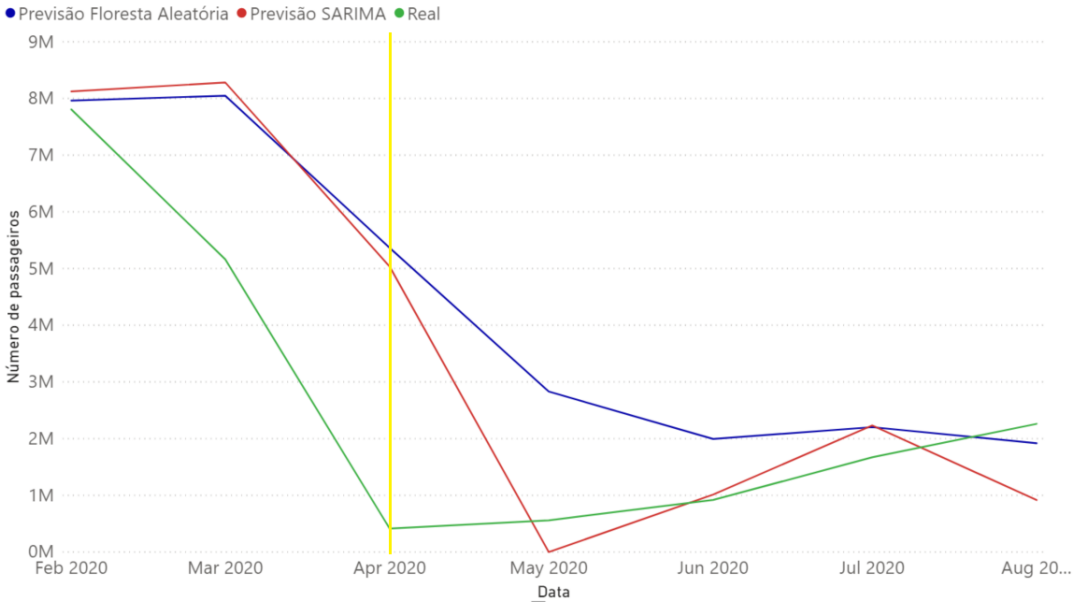


FIGURA 4.3 – Gráfico das previsões obtidas pelos modelos no período posterior ao início da crise causada pelo Covid-19.

Fonte: Elaborado pelo autor.

A Tabela 4.2 representa o resumo dos resultados obtidos por cada uma das modelagens para diferentes períodos de previsão. Ambos os modelos se mostraram viáveis para a aplicação em um período de normalidade, apresentando uma média de erros percentuais absolutos de 5,5% para o modelo de Floresta Aleatória e 3,0% para o modelo de SARIMA. Porém para o período de crise ainda é necessário uma otimização dos algoritmos para a aplicação, pois o melhor desempenho obtido foi pelo modelo de SARIMA e a média de erros percentuais absolutos ultrapassou os 50%. O modelo de SARIMA se mostrou mais preciso que o modelo de Floresta Aleatória em todos os cenários, sendo que o RMSE do modelo de SARIMA foi de 39,6% a 45,9% menor do que o RMSE do modelo de Floresta Aleatória.

TABELA 4.2 – Resultados obtidos pelas modelagens de SARIMA e Floresta Aleatória.

Período	**Modelo**	**RMSE**	**MAPE**
Pré crise	FA	557828	5,5%
	SARIMA	301560	3,0%
	Δ	-45,9%	-45,5%
Pós crise	FA	1295285	142,4%
	SARIMA	782059	50,8%
	Δ	-39,6%	-64,3%
Todo período	FA	621332	13,0%
	SARIMA	345642	5,6%
	Δ	-44,4%	-56,9%

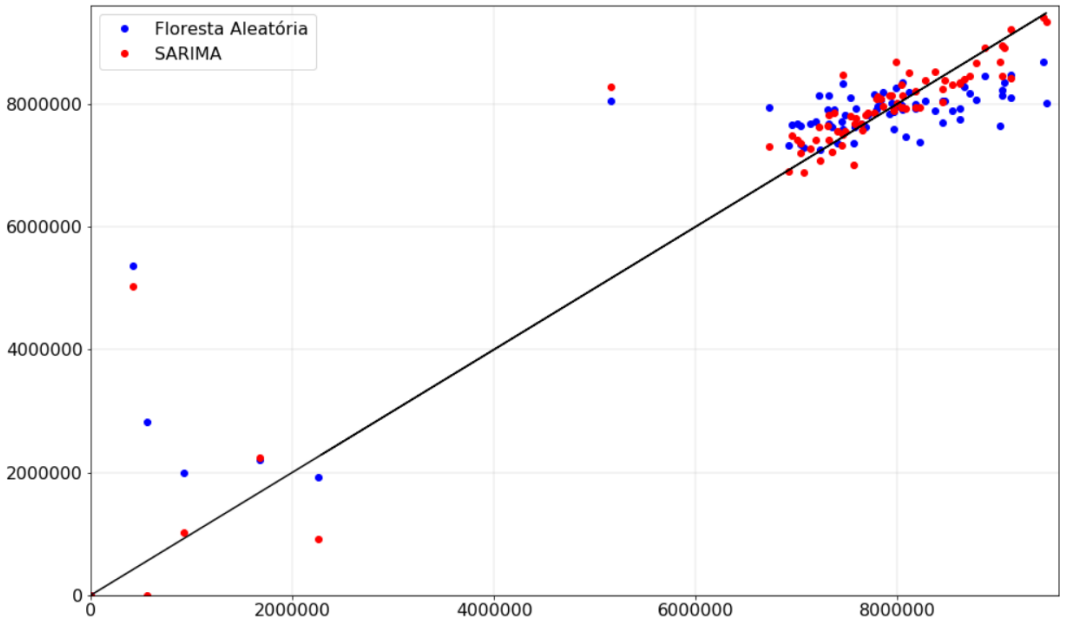


FIGURA 4.4 – Relação entre os resultados obtido pelas modelagens de SARIMA e Floresta aleatória e os dados reais.

Fonte: Elaborado pelo autor.

A Figura 4.4 representa o desvio de cada um dos modelos em relação a demanda real de passageiros. Os pontos na direita superior são referentes ao período pré pandemia, em que a demanda mensal ultrapassava os 6 milhões de passageiros, os pontos na região inferior esquerda do gráfico são referentes ao período de pandemia e os pontos muito acima da reta diagonal são referente aos dois meses de transição, março e abril de 2020. Sendo que, nesse tipo de análise, quanto maior a proximidade dos pontos com a reta diagonal,

maior é a precisão da previsão desse mês. Pela imagem é possível identificar que, no período de pré pandemia, os pontos obtidos pelo modelo de SARIMA apresentam grande proximidade com os valores reais, principalmente nos valores extremos, enquanto que o modelo de Floresta Aleatória apresenta bastante dificuldade para previsão nesses mesmos meses, o que fica evidente na quantidade de pontos abaixo da diagonal principal quando os valores reais de demanda são superior a 8 milhões de passageiros.

5 Conclusão

Os modelos construídos tinham como objetivo prever a demanda de passageiros no transporte aéreo no período de pandemia causada pelo Covid-19. O resultado pode ser utilizado para auxílio de tomadas de decisões no setor no período de crise e também para dar suporte para previsões de longo prazo, ao estimar como será a retomada da demanda de passageiros após a crise. Ambos os modelos se mostraram eficientes na previsão no período anterior ao da pandemia, porém no período de pandemia é necessário melhorias para que possam ser aplicados. Dentre os modelos estudados o modelo de SARIMA alcançou os melhores resultados, conclui-se que o objetivo foi parcialmente atingido.

O modelo de SARIMA alcançou os melhores resultados tanto no período da crise causada pelo Covid-19, alcançando um resultado de MAPE = 50,8%, quanto para o período anterior, apresentando MAPE = 3,0%. A estrutura utilizada foi o SARIMA$(2,1,2)(1,0,1)_{12}$. O modelo de Floresta aleatória também apresentou um resultado razoável para a previsão no período anterior a crise, alcançando um resultado de MAPE = 5,5%, porém no período de crise houve uma divergência maior em relação ao resultado real, obtendo MAPE = 142,4%. A floresta utilizada contém 500 árvores de regressões, número máximo de característica por árvore igual a 8, profundidade máxima da árvore igual a 9, número mínimo de amostras para divisão do nó igual a 5, *bootstrap* ativo e RSE como critério de divisão do nó.

Para aperfeiçoar o trabalho pode ser feito ajustes nos modelos para permitir uma melhor comparação entre eles. O modelo de SARIMA pode ser substituído pelo SARIMAX, para também incluir as variáveis externas macroeconômicas no modelo estatístico, e pode ser adicionado os parâmetros criados pelo algoritmo de SARIMA na modelagem de Floresta aleatória. Dessa forma ambos os modelos utilizariam os mesmos dados de entrada.

Este estudo possibilitou uma visão sobre alguns dos fatores que influenciam a demanda de passageiros no transporte aéreo no Brasil e permitiu identificar a viabilidade de se empregar os modelos SARIMA e Floresta Aleatória na previsão dessa demanda em diferentes cenários. O conhecimento construído e adquirido durante a confecção do trabalho poderá ser aproveitado para estudos posteriores.

Bibliografia

[1] Chris Aldrich. *Process Variable Importance Analysis by Use of Random Forests in a Shapley Regression Framework*. Western Australian School of Mines: Minerals, Energy e Chemical Engineering, Curtin University, GPO Box U1987, Perth, WA 6845, Australia, 2020.

[2] Geman D Amit Y. *Shape quantization and recognition with randomized trees*. Neural Computation. 9 (7): 1545–1588, 1997.

[3] Juliana Luz Passos Argenton. *Árvore de regressão para dados censurados e correlacionados*. Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica, Campinas, SP., 2013.

[4] Agência Nacional de Aviação Civil. *Dados e Estatística*. Disponível em: https://www.anac.gov.br. Acesso em: 24 mai, 2020.

[5] IAC - Instituto de Aviação Civil. *Demanda detalhada dos aeroportos brasileiros vol.1 2005*. 2005.

[6] Silvio Barge BHERING. *Mapeamento digital de areia, argila e carbono orgânico por modelos Random Forest sob diferentes resoluções espaciais*. Brasília, v. 51, n. 9, p. 1359-1370, 2016.

[7] IBOV - Índice Bovespa. *Bolsa de Valores*. Disponível em: br.advfn.com. Acesso em: 24 mai, 2020.

[8] G.E.P. BOX e G.M. JENKINS. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1976.

[9] A. N. CHAVES. *Bootstrap em Séries Temporais*. Rio de Janeiro. 208f. Tese (Doutorado em Engenharia Elétrica), Universidade Federal do Rio de Janeiro, 1991.

[10] J. Contreras; R. Espinola; F.J. Nogales; A.J. Conejo. *RIMA models to predict next-day electricity prices*. IEEE Transactions on Power Systems, 2003.

[11] World Bank Data. *Air transport, passengers carried*. Disponível em: https://data.worldbank.org/indicator/is.air.psgr. Acesso em 4 abr, 2020.

[12] S. A. Delurgio. *Forecasting principles and applications*. 1st Edition. Singapore: McGraw-Hill. 802p., 1997.

[13] Thomas Dietterich. *An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization*. Machine Learning. 40 (2): 139–157, 2000.

[14] D. D. E. Donadia. *Comparação entre as técnicas de regressão logística, árvore de decisão, bagging e random forest aplicadas a um estudo de concessão de crédito*. Universidade Federal do Paraná, Laboratório de Estatística do curso de Estatística do Setor de Ciências Exatas, Curitiba, PR, 2013.

[15] Goodhart S. G.; Burnham K. J. & James D. J. G. *Bilinear Self-tuning Control of a high temperature Heat treatment Plant*. IEEE Control Theory Appl.: Vol. 141, nº.1, 1994.

[16] W. H. Greene. *Previsão para o preço futuro do cacau através de uma série univariada de tempo: uma abordagem utilizando o método ARIMA*. 5.ed. New Jersey: Prentice Hall, 2003.

[17] J. L. C. GUTIERREZ. *Monitoring of the Corumbá*. i Dam Instrumentation Neural Networks e the Box & Jenkins null Model, 2003.

[18] Tin Kam Ho. *Random Decision Forests*. Proceedings of the 3rd International Conference on Document Analysis e Recognition, Montreal, QC, 14–16, 1995.

[19] Tin Kam Ho. *The Random Subspace Method for Constructing Decision Forests*. IEEE Transactions on Pattern Analysis e Machine Intelligence. 20 (8): 832–844, 1998.

[20] Breiman L. *Random Forests*. Machine Learning. 45 (1): 5–32, 2001.

[21] Aldous Pereira Moraes Marcel Castro de; Albuquerque. *Previsão para o preço futuro do cacau através de uma série univariada de tempo: uma abordagem utilizando o método ARIMA*. Sociedade Brasileira de Economia, Administração e Sociologia Rural (SOBER), 2006.

[22] C. M. C. MORETTIN P. A.; TOLOI. *Análise de Séries Temporais*. São Paulo: Edgard Blücher, 2004.

[23] Andrew V. Matcalfe Paul S. P. Cowpertwait. *Backtesting Time Series models — Weekend of a Data Scientist*. Springer, 2009.

[24] Instituto de Pesquisa Econômica Aplicada. *Dados macroeconômicos*. Disponível em: http://www.ipeadata.gov.br. Acesso em 4 nov., 2020.

[25] scikit-learn. *sklearn.ensemble.RandomForestRegressor*. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html. Acesso em 5 nov, 2020.

[26] Skipper Seabold e Josef Perktold. *statsmodels: Econometric and statistical modeling with python*. Proceedings of the 9th Python in Science Conference., 2010.

[27] Rajesh G. Kavasseri; Krithika Seetharaman. *Day-ahead wind speed forecasting using f-ARIMA models*. Renewable Energy Volume 34, Issue 5, 2009.

[28] Taylor G. Smith. *pmdarima: ARIMA estimators for Python, 2017*. Disponível em: http://www.alkaline-ml.com/pmdarima. Acesso em: 6 nov, 2020.

[29] C. A. Taconeli. *Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia*. 2008. Tese (Doutorado), Escola Superior de Agricultura "Luiz de Queiroz" – USP, Piracicaba, 2008.

[30] LFRA Torgo. *Inductive learning of tree-based regression models*. PhD Thesis, Faculty of Sciences, University of Porto, 1999.

[31] G. U. Yule. *On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers*. Philosophical Transactions of the Royal Society A: Mathematical, Physical e Engineering Sciences. 226 (636–646): 267–298., 1927.