CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
CAMPUS TIMÓTEO

Atos Ferreira Machado

ESTUDO COMPARATIVO DOS MODELOS PREDITIVOS SARIMA E RANDOM FOREST EM BASES DE DADOS DE BAIXA AMOSTRAGEM

Timóteo

2020

Atos Ferreira Machado

ESTUDO COMPARATIVO DOS MODELOS PREDITIVOS SARIMA E RANDOM FOREST EM BASES DE DADOS DE BAIXA AMOSTRAGEM

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Aléssio Miranda Júnior

Timóteo

2020

Atos Ferreira Machado

ESTUDO COMPARATIVO DOS MODELOS PREDITIVOS SARIMA E RANDOM FOREST EM BASES DE DADOS DE BAIXA AMOSTRAGEM

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Computação do Centro Federal de Educação Tecnológica de Minas Gerais, campus Timóteo, como requisito parcial para obtenção do título de Engenheiro de Computação.

Trabalho aprovado. Timóteo, 10 de dezembro de 2020:

_____

Prof. Me. Aléssio Miranda Júnior
Orientador

_____

Prof. Dr. Lucas Pantuza Amorim
Professor Convidado

_____

Prof. Me. Odilon Corrêa da Silva
Professor Convidado

Timóteo
2020

Emitido em 10/12/2020

FOLHA DE ROSTO (PLATAFORMA BRASIL) Nº 2/2020 - DCCTM (11.63.05)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 10/12/2020 17:29)
ALESSIO MIRANDA JUNIOR
PROFESSOR ENS BASICO TECN TECNOLOGICO
DCCTM (11.63.05)
Matrícula: 1713470

(Assinado digitalmente em 11/12/2020 13:37)
LUCAS PANTUZA AMORIM
PROFESSOR ENS BASICO TECN TECNOLOGICO
DCCTM (11.63.05)
Matrícula: 2897411

(Assinado digitalmente em 11/12/2020 16:32)
ODILON CORREA DA SILVA
PROFESSOR ENS BASICO TECN TECNOLOGICO
DCCTM (11.63.05)
Matrícula: 2794495

Para verificar a autenticidade deste documento entre em https://sig.cefetmg.br/documentos/ informando seu número: **2**, ano: **2020**, tipo: **FOLHA DE ROSTO (PLATAFORMA BRASIL)**, data de emissão: **10/12/2020** e o código de verificação: **a5993c7f89**

.

.

.

Dedico a todas as pessoas que um dia se dispuseram a me ajudar para que eu pudesse chegar até aqui.

Agradecimentos

Agradeço aos meus pais por me educarem durante toda a vida, além de poderem me dar aquilo que eles não tiveram acesso na juventude. Agradeço aos meus amigos em especial ao Tiago por estar presente em momentos difíceis e me apoiar sempre que possível. Agradeço ao meu Orientador Aléssio pela paciência e apoio. Agradeço a todos colegas da faculdade e do trabalho que me ajudaram até aqui.

“O homem é a medida de todas as coisas.”.
Protágoras

Resumo

O uso de dados para predição está cada vez mais presente. Um tipo de dado amplamente usado para fazer essas previsões são as séries temporais. Para lidar com esse tipo de série, há uma variedade de tipos de modelos preditivos. Um dos mais conhecidos é o SARIMA, que lida com características de séries temporais como sazonalidade e tendência. Outro modelo possível de ser utilizado é o *Random Forest*, que é um algoritmo de aprendizado de máquina que pode ser adaptado para uma variedade de situações. Um estudo comparativo da qualidade dos resultados desses dois modelos foi proposto usando bases com baixa amostragem. O método RMSD foi utilizado para comparar o desempenho desses dois modelos. Ambos modelos tiveram bom desempenho, cada um se sobressaindo num cenário específico.

Palavras-chave: Séries temporais, SARIMA, Random Forest, RMSD.

Abstract

The use of data for prediction is increasingly present. One kind of data widely used to make these predictions are time series. To deal with this type of series, there are a range of types of predictive models. One of the best known is SARIMA, which handles time series characteristics like seasonality and trend. Another possible model to be used is the Random Forest, which is a machine learning algorithm that can be adapted for a variety of situations. A comparative study of the quality of the results of these two models was proposed using bases with low sampling. The RMSD method was used to compare the performance of these two models. Both models performed well, each standing out in a specific scenario.

Keywords: Time Series, SARIMA, Random Forest, RMSD.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ANN	Artificial Neural Network
ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
AR	Autoregressive
ARFIMA	Autoregressive Fractionally Integrated Moving Average
ARIMA	Autoregressive Integrated Moving Average
FAC	Função de Autocorrelação
FACP	Função de Autocorrelação Parcial
MA	Moving Average
IQR	Variação Interquartil
Q1	Primeiro Quartil
Q3	Terceiro Quartil
RF	Random forest
RMSD	Root Mean Square Deviatian
SARIMA	Seasonal AutoregressiveMoving Average
SNIF	Sistema Nacional de Informações Florestais
SVM	Support Vector Machines

Sumário

1 Introdução

Diversas áreas do conhecimento hoje produzem muitos dados, variando de dados meteorológicos, ações da bolsa de valores, dados de produtos produzidos em uma fábrica, surtos de doenças numa região e usuários em redes sociais. Esses dados se tornam interessantes a medida em que podem ser utilizados pra entender e otimizar tomada de decisões.

Com o crescimento do volume de dados se faz necessário a pesquisa e o uso de métodos mais eficientes para a análise dos dados (BERTHOLD; HAND, 2003). Esses dados podem ser utilizados na produção de novos produtos bem como na forma de alocar recursos. Outra importante forma de se utilizar esses dados é na predição de valores.

Predizer valores pode ter impacto importante nos negócios de uma empresa, pois dependendo da natureza do dado, pode ajudar na tomada consciente de decisões. Alguns exemplos são compra ou venda de produtos e ativos, escolha da época que promoções podem ser realizadas, a criação de uma escala de trabalho de forma a otimizar os recursos humanos disponíveis além de possíveis áreas pra reduzir custos.

Predizer situações futuras também tem utilidade para governos e outros tipos de organizações que também podem otimizar seus recursos e se planejar, dependendo da natureza dos dados. Alguns exemplos são ocorrência de eventos climáticos que afetam regiões, problemas sanitários como surtos de doenças e problemas ambientais como queimadas. Um tipo de dado que se utiliza muito para predições são séries temporais.

Segundo Box et al. (2015), uma série temporal pode ser considerada um conjunto de valores ordenados num período de tempo. Pela simplicidade da forma, existem modelos que podem ser utilizados a fim de prever os valores dessa série. Porém pela diversidade de características se faz necessário estudos de todos os tipos de modelos bem como testes pra entender os que se melhor encaixam em determinado contexto.

Alguns modelos como o Seasonal Autoregressive Moving Average (SARIMA) são utilizados em bases de dados de séries temporais mesmo com características como variações constantes dos valores e um direcionamento para os valores futuros. Outros modelos são mais abrangentes mas podem ser adaptados para o uso de séries temporais.

1.1 Problema

Existem diversos modelos e algoritmos que podem realizar a predição de valores em séries temporais, mas nem sempre se sabe qual aplicar em determinado caso. Como os dados podem assumir diversas formas, entender a base de dados e as características de diversos modelos se faz necessário para selecionar o algoritmo que melhor se adequará aos dados.

1.2 Justificativa

A partir de um estudo de dois modelos preditivos é possível construir uma base de conhecimento para que a tomada de decisão e consequentemente o uso correto de determinados modelos possa ser feita mesmo em bases de dados com menor amostragem. Dessa forma as predições podem ser utilizadas para importantes tomadas de decisões.

1.3 Objetivos

Para a realização deste trabalho decidiu-se comparar os modelos SARIMA e *Random forest* utilizando bases de dados com baixa amostragem, alguma relevância e diferentes características, de modo a demonstrar o melhor modelo dependendo do tipo de dado e das condições de uso. Também, objetivam-se mais especificamente:

- Analisar e preparar as bases de dados escolhidas;
- Aplicar os modelos preditivos nas bases de dados;
- Avaliar os resultados obtidos.

1.4 Estrutura do texto

O texto está estruturado em seis capítulos, sendo os próximos a seguir na seguinte ordem:

- O Capítulo 2 apresenta a fundamentação teórica dos conceitos utilizados pra realização do trabalho incluindo a definição de séries temporais, o que são os valores *outliers* e como tratá-los, os métodos de predição utilizado (SARIMA e Random Forest) e a forma de comparar os resultados obtidos desses métodos utilizando o método de comparação RMSD;
- O Capítulo 3 lista alguns trabalhos que realizaram a comparação de diferente modelos preditivos, entre eles os utilizados nesse trabalho como SARIMA e Random forest;
- O Capítulo 4 apresenta os materiais utilizados na pesquisa como as bases de dados e a linguaguem de programação e suas bibliotecas. Também lista os procedimentos metodológicos através dos quais este trabalho se desenvolve;
- O Capítulo 5 Descreve como foram seguidos os passos metodológicos, apresentando o tratamento dado as bases de dados, como se deu a execução dos modelos e quais resultados foram alcançados;
- Por fim, o Capítulo 6 descreve as conclusões finais dos resultados, as limitações durante o desenvolvimento do trabalho bem como sugestões de trabalhos futuros.

2 Fundamentação Teórica

2.1 Séries Temporais

Uma série temporal é uma sequência de observações realizadas sequencialmente no tempo (BOX et al., 2015). Pode-se citar como exemplos de séries temporais: a quantidade de acidentes de trânsito por semana, a quantidade de produtos produzidos em uma fábrica num mês e a temperatura média mensal. Diversas áreas do conhecimento geram e ou trabalham com essas séries e fazem aplicações com as mesmas. Alguma dessas áreas são: ciências sociais, economia, engenharia e ciências da natureza como meteorologia.

As séries temporais são muito úteis, pois com o uso de métodos de predição é possível estimar os seus valores futuros baseado no histórico dos valores anteriores. Para realizar essas estimativas é preciso entender primeiro as características que a definem, dessa forma é viável definir o melhor método a ser utilizado na predição. Cada série possui características distintas como tendência, sazonalidade ou ruído.

A tendência pode ser entendida como a direção que os valores caminham ao longo do tempo. Geralmente essa direção é bem definida, embora os valores talvez possuam alguma variância. Os tipos de tendência podem ser:

- Linear: os valores da série possuem um comportamento constante de crescimento ou decrescimento;
- Quadrática: os valores da série possuem um comportamento que não é constante e que se aproxima de uma curva quadrática;
- Sem tendência: não possui uma direção bem definida geralmente variando em torno de um valor constante;
- Outros: alguns termos da séries podem estar elevados a potências maiores.

A sazonalidade é a frequência com que os termos da série se repetem num determinado intervalo de tempo, e o ruído são as flutuações dos valores mais destoantes. Para uma melhor análise de uma série, é preciso primeiro decompor a tendência, a sazonalidade e o ruído.

Para a decomposição das características da série, utiliza-se do método de decomposição sazonal. O produto desse método é um gráfico para cada uma dessas caraterísticas. Uma vez gerados os gráficos é possível inferir o valor da sazonalidade, se a série possui tendência e qual ela é bem como a interferência do ruído nos valores.

2.2 Outliers

Para o uso de alguns modelos preditivos é desejável que as bases de dados possuam uma distribuição normal. No entanto, nem sempre isso é possível, geralmente devido a valores muito destoantes na distribuição da série. Esses valores são conhecidos como *outliers*, e existem por uma série de fatores como dados inexistentes, dados corrompidos e erros na medição.

A identificação inicial dos possíveis valores *outliers* pode ser realizada por um especialista na área da base de dados, já que somente com um domínio específico é possível determinar critérios de classificação para valores destoantes. Outra maneira é a utilização de métodos estatísticos para encontrar esse valores, como a Variação Interquartil (IQR).

A Variação Interquartil é definida pela subtração do terceiro quartil (Q3) pelo primeiro quartil (Q1) conforme a Equação 2.1.

$$IQR = Q3 - Q1 \tag{2.1}$$

O primeiro quartil (Q1) representa a mediana dos valores abaixo da mediana do conjunto completo. O terceiro quartil (Q3) representa a mediana dos valores acima da mediana do conjunto completo. A partir desse cálculo são considerados *outliers* os valores x no conjunto da base de dados que respeitem as relações das Equações 2.2 e 2.3.

$$x < Q1 - 1.5 * IQR \tag{2.2}$$

$$x > Q3 + 1.5 * IQR \tag{2.3}$$

Em ambas Equações, 2.2 e 2.3, percebe-se o uso da constante 1.5. Essa constante poderia assumir outros valores, mas esse é o valor mais utilizado por captar até 99% dos *outliers* (MORETTIN, 2010). Uma vez identificados, esses valores podem ser removidos a fim de tornar a distribuição mais próxima de uma gaussiana.

2.3 Modelos ARIMA

2.3.1 ARIMA

O modelo ARIMA (*Autoregressive Integrated Moving Average*) foi proposto por Box e Jenkins (1976) e representa uma combinação dos modelos AR (*Autoregressive*) e MA (*Moving Average*) adicionando um passo de diferenciação dos valores da série. Para entender melhor o ARIMA é preciso entender primeiro os modelos AR e MA.

O modelo AR é um modelo estocástico em que o valor atual pode ser descrito como uma combinação linear entre os valores anteriores acrescido de um erro, como descrito pela Equação 2.4:

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + ... + \phi_p z_{t-p} + a_t \tag{2.4}$$

A variável z_t representa o valor atual, as variáveis z_{t-1}, z_{t-2}, ..., z_{t-p} são os valores anteriores, os coeficientes ϕ_1, ϕ_2,...,ϕ_p representam a autocorrelação dos termos anteriores, e a variável a_t representa o ruído. Pode-se definir que é um processo de ordem p porque depende dos p termos anteriores.

O modelo MA pode ser definido como o cálculo do valor atual a partir da soma do erro das observações anteriores, como descrito pela Equação 2.5:

$$z_t = \theta_1 a_{t-1} + \theta_2 a_{t-2} + ... + \theta_p a_{t-q} + a_t \tag{2.5}$$

A variável z_t representa o valor atual, as variáveis a_t, a_{t-1}, a_{t-2}, ..., a_{t-q} são os valores dos ruídos e os coeficientes θ_1, θ_2,...,θ_q representam os parâmetros do modelo. Pode-se definir que é um processo de ordem q porque depende dos q termos anteriores.

A partir da combinação do modelos AR e MA surgiu o modelo ARMA (*Autoregressive Moving Average*). Como uma melhoria de tal modelo, aplicou-se uma diferenciação dos termos da série. Essa diferenciação é a subtração de um termo da série pelo valor anterior do mesmo, como na Equação 2.6:

$$z_t^{'} = z_t - z_{t-1} \tag{2.6}$$

Dessa melhoria surgiu o ARIMA, já que a diferenciação faz com que a série fique mais estacionária. Para o cálculo do ARIMA ele utiliza de 3 parâmetros (p, d, q) que podem ser definidos como:

- p: A ordem do modelo AR, ou seja, o número de observações passadas(*lags*) levadas em consideração no processo;
- d: O número de vezes que os termos da série são diferenciados;
- q: A ordem do modelo MA.

Para definir os valores utilizados pelos parâmetros p e q são utilizadas as funções de autocorrelação parcial e as funções de autocorrelação respectivamente. A Função de Autocorrelação Parcial (FACP) é a função que calcula a correlação direta entre os valores da série sem levar em consideração os valores passados. Ela serve pra determinar a ordem p do modelo de auto-regressão.

A Função de Autocorrelação (FAC) é a função que calcula a correlação direta entre os valores da série levando em consideração os valores passados. Ela serve pra determinar a ordem q do modelo de média móvel.

2.3.2 SARIMA

Embora o ARIMA fosse um interessante método para prever valores futuros de uma série, ele apresenta limitações quando confrontado com séries que possuem uma alta sazonalidade. Dessa forma, como uma evolução do modelo surge o *Seasonal Autoregressive Moving Average* (SARIMA).

O SARIMA é um modelo que, por ser uma extensão do ARIMA, utiliza dos mesmos parâmetros (p, d, q) para lidar com fatores de tendência, mas acrescenta outras variáveis (P, D, Q, m) para lidar com fatores sazonais. As variáveis (P, D, Q, m) significam:

- P: A ordem sazonal de autoregressão;
- D: A ordem sazonal de diferenciação;
- Q: A ordem sazonal de média movel;
- m: O intervalo para um único período sazonal.

A combinação de todos esses parâmetros (p, d, q) (P, D, Q, m) é o que define esse método. Uma das formas de estimar esses valores é utilizando as funções de Autocorrelação Parcial e de Autocorrelação, no entanto para a interpretação desses resultados pode ser necessário conhecimentos específicos da base e do tipo de dado analisado, necessitando assim de um especialista da área para fazer uma leitura correta.

2.4 Random Forest

Random Forest é um algoritmo de aprendizado de máquina supervisionado que consiste no conjunto de múltiplas árvores de decisão geralmente treinadas e que votam entre si pela mais popular. Ele é uma melhoria do algoritmo de árvore de decisão que utiliza apenas uma árvore pra prever os resultados. Essas melhorias significativas na precisão da classificação são resultado do uso de um conjunto de árvores, permitindo que elas votem na mais popular. Para crescer esses conjuntos, geralmente são gerados vetores aleatórios que governam o crescimento de cada árvore no conjunto, como o *bagging*, que para o crescimento de cada árvore há uma seleção aleatória sem substituição feita a partir do exemplos no conjunto de treinamento (BREIMAN, 1996).

Por ser um algoritmo de aprendizado de máquina, pode ser tanto aplicado em problemas de classificação ou regressão. Sua popularidade se dá pela simplicidade no uso, uma vez que necessita de ajuste de poucos parâmetros, tem ampla aplicabilidade em problemas de previsão e consegue lidar com pequenas amostras.

Uma definição mais formal é dada por Breiman (2001) diz que Random forest é um classificador que consiste de uma coleção de classificadores estruturados em árvore { $h(x, \theta_k)$, $k = 1$, ...} onde os $\{\theta_k\}$ são vetores aleatórios distribuídos de forma idêntica, e cada árvore lança um voto unitário para a classe mais popular no parâmetro x.

2.5 RMSD

Diferentes métodos de predição podem ser utilizados numa base de dados. Algumas bases de dados possuem características distintas que podem tanto facilitar como prejudicar os resultados gerados. Para se fazer uma comparação entre os resultados de diferentes métodos se faz necessário o uso de alguma medida de equiparação como o *Root Mean Square Deviatian* (RMSD).

O RMSD, cuja tradução é *Desvio da Raiz do erro quadrático médio*, é uma medida de performance que contrasta os resultados encontrados pelos modelos preditivos em relação aos valores esperados na base de treino. A Equação 2.7 define o cálculo realizado pelo RMSD:

$$R = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \tag{2.7}$$

Nessa equação, o valor de P_i é o resultado alcançado da aplicação do modelo de predição, enquanto que o valor de O_i é o resultado esperado baseado nos valores da sua base de treino. Quanto mais próximo de zero for o valor calculado de R melhor pode ser considerado o modelo utilizado.

O RMSD pode ser aplicado tanto para comparação dos resultados de um modelo dentro de uma mesma base de dados como em bases distintas. Ele pode ser considerado como uma forma preferível de equiparar diferentes modelos, uma vez que o erro achado é retratado na mesma proporção das variáveis analisadas (HYNDMAN; KOEHLER, 2006).

3 Trabalhos Correlatos

Os presentes trabalhos tem como base a comparação entre modelos preditivos. Assim como neste trabalho, alguns artigos utilizaram de modelos de análise de séries temporais como ARIMA e SARIMA, além de modelos de aprendizado de máquina como o *Random Forest*.

Noureen et al. (2019) fizeram um estudo de caso sobre a previsão de demanda de energia numa escala industrial pequena. Fizeram uma análise comparativa de desempenho entre o modelo SARIMA e a técnica de aprendizado de máquina *Random forest*. Esses dois modelos foram aplicados para a previsão de demandas de curto e longo prazo. Para demandas carga de curto e longo prazo o *Random forest* demonstrou desempenho superior em relação ao SARIMA em termos de precisão, no entanto não foram levados em consideração as diferenças entre complexidade e velocidade de execução, que ficaram como sugestões de trabalhos futuros.

Tyralis e Papacharalampous (2017) concentraram em avaliar o desempenho do *Random forest* utilizando dois grandes conjuntos de séries temporais curtas, com o objetivo de sugerir um conjunto ótimo de variáveis utilizadas para a predição. Utilizaram como conjunto de dados 16.000 séries temporais simuladas a partir de uma variedade de modelos *Autoregressive Fractionally Integrated Moving Average* (ARFIMA). Também utilizaram dados de 135 séries temporais de temperatura média anual. O desempenho preditivo mais alto do *Random forest* foi observado ao utilizar um número baixo de variáveis defasadas recentes. Este resultado pode ser útil em aplicações futuras relevantes, com a perspectiva de alcançar maior precisão preditiva.

Kane et al. (2014) aplicaram modelos de séries temporais ARIMA e *Random forest* aos dados de incidência de surtos de alta influenza aviária patogênica (H5N1) no Egito. Utilizaram de dados disponíveis no sistema EMPRES-I. Durante os testes perceberam que o modelo *Random forest* superou o modelo ARIMA em capacidade preditiva. Desta forma puderam concluir que o *Random forest* é eficaz para prever surtos de H5N1 no Egito.

Kumar e Thenmozhi (2014) desenvolveram um artigo para identificar o melhor modelo híbrido para prever os retornos do índice de ações. Utilizaram três modelos híbridos diferentes combinando ARIMA linear e modelos não lineares, como o *Support Vector Machines* (SVM), *Artificial Neural Network* (ANN) e *Random Forest* (RF). Os modelos SVM, ANN e RF foram avaliados em termos de métricas estatísticas e critérios de desempenho de negociação. Após a análise concluiram que o modelo híbrido ARIMA-SVM é o melhor modelo de previsão para alcançar alta precisão.

4 Materiais e Método

Neste capítulo serão apresentados os materiais utilizados durante a realização da pesquisa bem como o método utilizado para comparação das bases de dados.

4.1 Materiais

4.1.1 Bases de dados

Esta seção abordará as bases de dados utilizadas para as comparações dos métodos de predição, explorando algumas de suas caraterísticas como a sazonalidade e a tendência. Para a realização do trabalho foram escolhidas duas bases de dados, todas no formato de séries temporais.

A primeira base de dados é uma série histórica mensal do número de focos de calor no Brasil entre 1998 e 2017 e a Figura 1 representa um gráfico com os valores da série:

Figura 1 – Focos de Calor entre 1998 e 2017

Fonte: Adaptado de SNIF (2018)

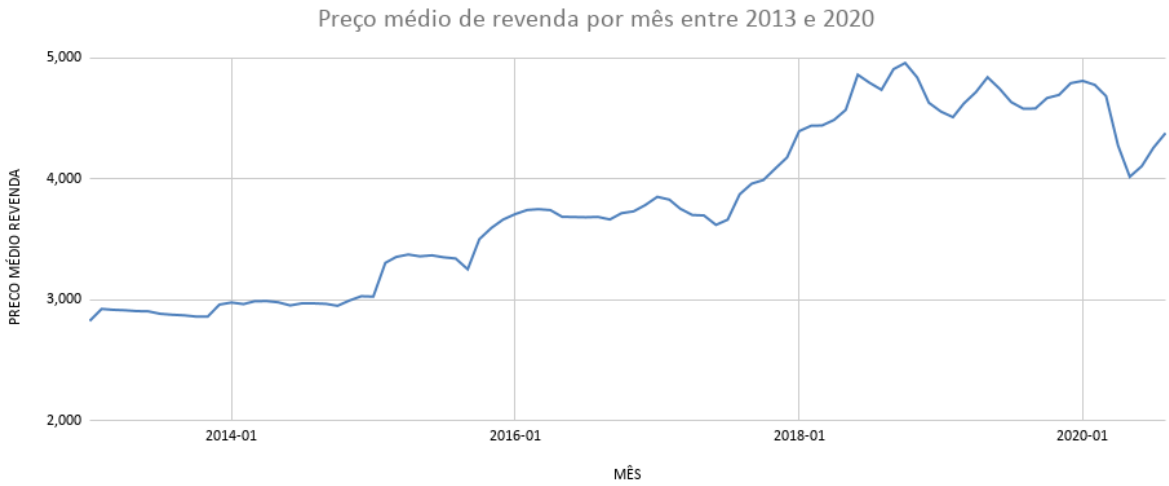
Essa base foi criada a partir dos dados disponíveis no Portal Brasileiro de Dados Abertos com base na planilha de *Incêndios Florestais - Focos de Calor - Brasil* do Sistema Nacional de Informações Florestais (SNIF). A partir dessa planilha foram selecionadas as colunas de “Número“ e “Período“ que representam respectivamente o número de focos de calor no Brasil naquele mês e o respectivo ano e mês da ocorrência.

A base de dados final possui 239 registros que estão dispostos em duas colunas “período“ e o “número“. A coluna “período“ possui o formato “ano-mês“ com as datas em formato numérico e a coluna “numero“ possui os valores do número de focos. Na Figura 1 é possível

observar um padrão de sazonalidade anual, uma vez que notando a variação dos valores. Percebe-se que os vales nas curvas do gráfico ocorrem geralmente no início de cada ano.

A segunda base de dados é uma série histórica mensal do preço médio de revenda de gasolina no estado de Minas Gerais entre os anos de 2013 e 2020. Na Figura 2 é possível observar o gráfico com os valores dessa série temporal:

Figura 2 – Série temporal do preço médio de revenda de gasolina



Fonte: Adaptado de ANP (2020)

Essa base foi criada a partir dos dados disponíveis no portal da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) na seção *Série histórica mensal*. Da planilha disponível foram primeiro filtradas as informações baseadas na coluna no estado e no produto e depois selecionadas as colunas “MÊS“ e “PREÇO MÉDIO REVENDA“ que representam respectivamente o ano e o mês do registro, o valor do preço médio de revenda de gasolina.

A base de dados final possui 92 registros que estão dispostos em duas colunas, “mês“ e “preco“. A coluna de “mês“ possui o formato "ano-mês"com as datas em formato numérico e a coluna “preço“ possui os valores do preço médio da gasolina com precisão de 3 casas decimais. Na Figura 2 é possível observar uma tendência crescente dos valores, com variações mais abruptas nas observações finais do preço.

4.1.2 Python

A linguagem Python se tornou referência nas áreas de análise de dados, processamento e *machine learning*. Isso se deve ao ecossistema criado entorno da linguagem com uma série de bibliotecas implementadas, tornando o preparo das bases de dados bem como o uso de modelos preditivos mais simples. Dessa forma, profissionais e pesquisadores na área podem focar mais nos resultados (RASCHKA, 2019).

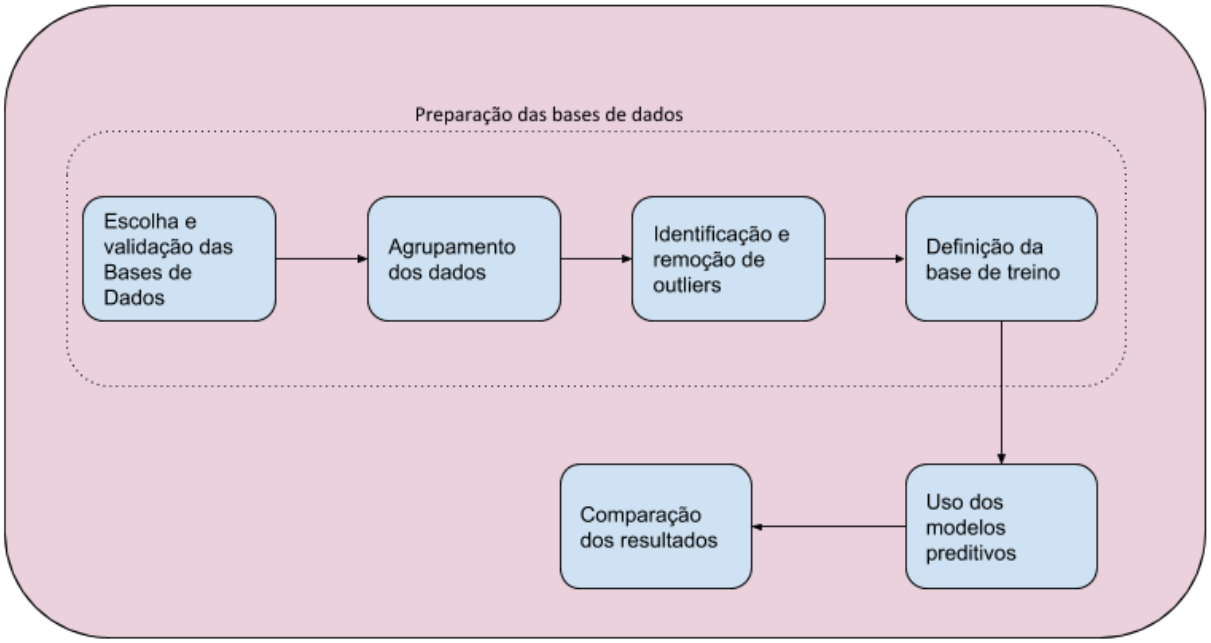
Para a realização deste trabalho serão utilizadas principalmente as bibliotecas *Statsmodels*, *Sklearn* e *Pandas*. A biblioteca *Statsmodels* é para a utilização do modelo SARIMA, a biblioteca *Sklearn* para utilização do modelo Random Forest e a biblioteca *Pandas* será utili-

zada na conversão dos arquivos de extensão *csv* em estruturas de dados específicas conhecidas como *Dataframes*. Essas estruturas permitem uma manipulação mais simples do dados em memória para a realização de cálculos. O uso dessas bibliotecas possibilitou a análise de séries temporais.

4.2 Método

Os passos metodológicos utilizados no trabalho são descritos brevemente nos parágrafos subsequentes. No capítulo seguinte serão explorados com mais clareza como cada um deles foi de fato realizado na prática. A Figura 3 lista todos passos seguidos em ordem de execução.

Figura 3 – Passos metodológicos seguidos neste trabalho



Fonte: Elaborada pelo autor

Como etapa inicial, foi necessário escolher algumas bases de dados para a realização do trabalho, descritas na Seção 4.1.1. Essas bases necessitam de análise para checar se os dados estão corretos, se não estão corrompidos e se as datas correspondem a um formato de séries temporais. Caso algum dos registros possua algum desses problemas é necessário corrigi-lo ou remove-lo.

Uma vez que os dados sejam validados é necessário checar se é preciso agrupa-los e organiza-los de modo que a série seja normalizada. Desta maneira pode-se garantir a integridade dos valores evitando que existam valores inconsistentes ou errados.

Depois de garantido a integridade dos dados, é necessário remover os valores *outliers*. O método utilizado para isso é o IQR descrito na Seção 2.2. Desta forma todos os valores que não obedecem as fórmulas 2.2 e 2.3 serão descartados. Posteriormente serão avaliados os impactos dessa mudança nas bases de dados.

A partir da base de dados original será definida a base de dados de treino. Essa base de treino será composta pelos n últimos valores da série temporal analisada. Esse valor de n pode ser determinado a partir de múltiplos do valor da sazonalidade.

Após as etapas anteriores de preparação das bases originais e a definição das bases de treino, serão aplicados os modelos de predição SARIMA e Random Forest descritos na Seções 2.3.2 e 2.4 respectivamente. Por fim, será utilizado o método RMSD descrito na Seção 2.5 para comparar os valores encontrados após aplicação dos modelos com os valores da base de treino. Desta maneira pode-se inferir e entender qual modelo obteve melhores resultados para cada base de dados.

5 Resultados e Discussões

Neste capítulo serão discutidos os resultados obtidos aplicando-se os passos descritos na Seção 4.2 do Capítulo 4. Escolheu-se dividir em três seções: a Seção 5.1 discute a aplicação dos passos referentes as modificações necessárias para preparar as bases de dados. A Seção 5.2 discute a aplicação dos modelos preditivos. Por fim, a Seção 5.3 discute os resultados encontrados após a aplicação dos passos metodológicos.

5.1 Preparação das Bases de dados

5.1.1 Criação e Validação das Bases de Dados

Conforme descrito na Seção 4.1.1 foram escolhidas duas bases de dados. A primeira base de dados é uma série histórica mensal do número de focos de calor no Brasil entre 1998 e 2017, e a segunda base de dados é uma série histórica mensal do preço médio de revenda de gasolina no estado de Minas Gerais entre os anos de 2013 e 2020. Ambas as bases necessitaram de modificações pra se adequarem a um formato de séries temporais.

A primeira base foi extraída com base numa planilha de *Incêndios Florestais - Focos de Calor - Brasil*. Originalmente os dados estavam agrupados por mês e ordenados pelo ano de ocorrência, conforme observa-se na Figura 4.

Figura 4 – Planilha original de focos de calor no Brasil

Ano	Mês	Número	Período
1998	Janeiro	0	01/01/1998
1999	Janeiro	1081	01/01/1999
2000	Janeiro	778	01/01/2000
2001	Janeiro	547	01/01/2001
2002	Janeiro	1654	01/01/2002
2003	Janeiro	5091	01/01/2003
2004	Janeiro	2705	01/01/2004
2005	Janeiro	4990	01/01/2005
2006	Janeiro	3255	01/01/2006
2007	Janeiro	3055	01/01/2007
2008	Janeiro	2125	01/01/2008
2009	Janeiro	2848	01/01/2009
2010	Janeiro	2851	01/01/2010
2011	Janeiro	1416	01/01/2011
2012	Janeiro	2491	01/01/2012
2013	Janeiro	2049	01/01/2013
2014	Janeiro	2634	01/01/2014
2015	Janeiro	4637	01/01/2015
2016	Janeiro	5983	01/01/2016
2017	Janeiro	2370	01/01/2017
Máximo	Janeiro	5983	
Média	Janeiro	2766	
Mínimo	Janeiro	547	
1998	Fevereiro	0	01/02/1998
1999	Fevereiro	1284	01/02/1999
2000	Fevereiro	562	01/02/2000
2001	Fevereiro	1059	01/02/2001
2002	Fevereiro	1570	01/02/2002
2003	Fevereiro	2398	01/02/2003

Fonte: Adaptado de SNIF (2018)

Primeiramente foram removidas as linhas com as informações de valores “Máximo“, “Média“ e “Mínimo“ para cada mês. Depois foi necessário agrupar os valores por ano de ocorrência e ordenar por meses. Dessa forma, os dados ficaram ordenados da forma correta. Após ordenar os valores a partir do período, foram selecionadas as colunas de “Número“ e “Período“ para a criação da base final utilizada. Por fim os valores da coluna *Período* foram adequados para o formato “ano-mês“. O formato da base final ficou conforme a Figura 5.

Figura 5 – Formato final da base de focos de calor no Brasil.

periodo	numero
1998-01	0
1998-02	0
1998-03	0
1998-04	0
1998-05	0
1998-06	3551
1998-07	8067
1998-08	35551
1998-09	41974
1998-10	23498
1998-11	6804
1998-12	4449
1999-01	1081
1999-02	1284
1999-03	667
1999-04	717
1999-05	1812
1999-06	3632
1999-07	8758
1999-08	39487
1999-09	36914
1999-10	27014

Fonte: Adaptado de SNIF (2018)

A Segunda base de dados foi criada a partir da planilha de levantamento de preços mensal por estado brasileiro disponível no portal da ANP. A planilha original contêm diversas informações como mês, estado, região, diversos tipos de produtos combustíveis como gasolina e álcool, número de postos pesquisados, unidade de medida, preço médio observado entre outros, conforme observado na Figura 6.

Figura 6 – Formato inicial da base de preço da gasolina em MG

MÊS	PRODUTO	REGIÃO	ESTADO	ÚMERO DE POSTOS PESQUISADO	NIDADE DE MEDID	RECO MÉDIO REVENDA	DESVIO PADRÃO REVENDA
abr.-02	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4865	R$/l	1,682	0,102
mai.-02	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4928	R$/l	1,679	0,104
jun.-02	GASOLINA COMUM	SUDESTE	MINAS GERAIS	3921	R$/l	1,661	0,110
jul.-02	GASOLINA COMUM	SUDESTE	MINAS GERAIS	5884	R$/l	1,711	0,118
ago.-02	GASOLINA COMUM	SUDESTE	MINAS GERAIS	5225	R$/l	1,682	0,126
set.-02	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4880	R$/l	1,673	0,118
out.-02	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4801	R$/l	1,689	0,105
nov.-02	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4286	R$/l	1,900	0,097
dez.-02	GASOLINA COMUM	SUDESTE	MINAS GERAIS	5271	R$/l	1,944	0,117
jan.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4372	R$/l	2,094	0,104
fev.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4279	R$/l	2,171	0,104
mar.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4558	R$/l	2,164	0,110
abr.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	5104	R$/l	2,145	0,114
mai.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4303	R$/l	2,071	0,111
jun.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4679	R$/l	1,982	0,110
jul.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4999	R$/l	1,918	0,113
ago.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4244	R$/l	1,935	0,104
set.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4949	R$/l	1,966	0,111
out.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4604	R$/l	1,963	0,103
nov.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4279	R$/l	1,963	0,106
dez.-03	GASOLINA COMUM	SUDESTE	MINAS GERAIS	5336	R$/l	1,963	0,105
jan.-04	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4256	R$/l	1,967	0,106
fev.-04	GASOLINA COMUM	SUDESTE	MINAS GERAIS	4250	R$/l	1,962	0,112

Fonte: Adaptado de ANP (2020)

Após filtrar as informações por estado “MINAS GERAIS“ e por produto “GASOLINA COMUM“, foram selecionadas as colunas “MÊS“ e “PREÇO MÉDIO REVENDA“. A coluna “MÊS“ apresentava um formato “mês-ano“, com o mês na forma escrita abreviada e o ano representado por dois dígitos, por isso teve de sofrer uma formatação para o padrão “ano-mês“ com o ano tendo 4 dígitos e o mês tendo dois dígitos afim de se adequar ao formato de séries temporais. O resultado final pode ser observado conforme a Figura 7. Como ultima modificação, o caractere de separação das casas decimais foi modificado para ponto após a exportação do arquivo para o formato CSV.

Figura 7 – Formato final da base de preço da gasolina em MG

MÊS	PRECO MÉDIO REVENDA
2013-01	2,824
2013-02	2,923
2013-03	2,916
2013-04	2,912
2013-05	2,905
2013-06	2,903
2013-07	2,883
2013-08	2,875
2013-09	2,870
2013-10	2,860
2013-11	2,860
2013-12	2,959
2014-01	2,976
2014-02	2,963
2014-03	2,986
2014-04	2,988
2014-05	2,978

Fonte: Adaptado de ANP (2020)

Após a criação das bases não se observou nenhuma irregularidade nos dados. Ambas bases apresentam dados consistentes e registros completos. Também não foi observada necessidade de agrupamento dos dados, pois os dados já possuem uma periodicidade mensal.

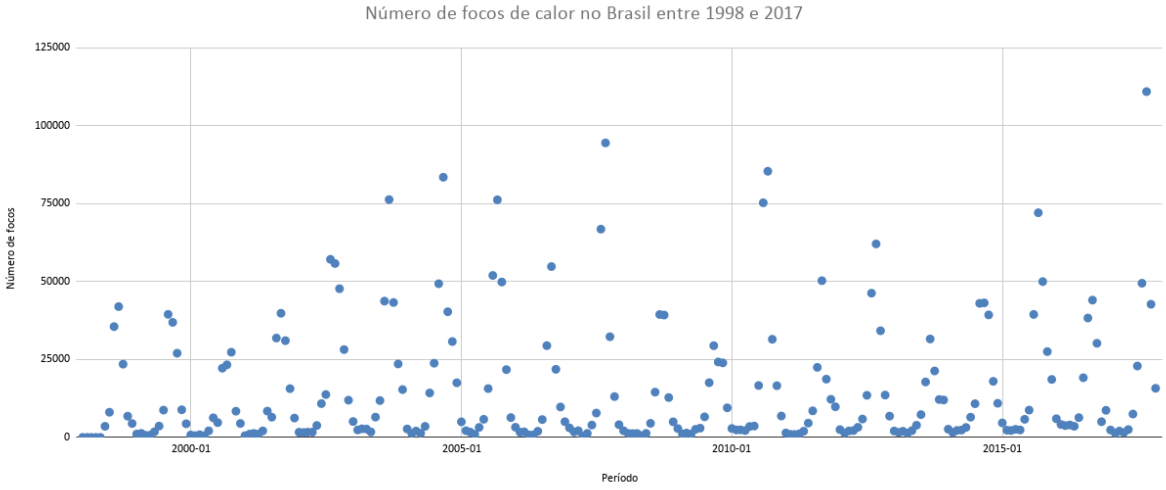
5.1.2 Remoção dos *outliers*

Após a criação e validação das bases de dados foi aplicado o método de Variação Interquartil em todas elas afim de remover os valores considerados *outliers*. A segunda base referente ao preço médio da gasolina em MG não sofreu alterações, então todos valores permaneceram durante a execução dos algoritmos. Por outro lado, a primeira base que é referente ao número de focos de calor no Brasil por mês, foram encontrados 15 registros considerados *outliers*.

O gráfico de dispersão dos valores da base antes da remoção dos *outliers* pode ser observado na Figura 8. Nele é possível observar que os valores da série estavam mais dispersos e os picos tinham valores muito destoantes do restante. Os picos geralmente acontecem

em setembro.

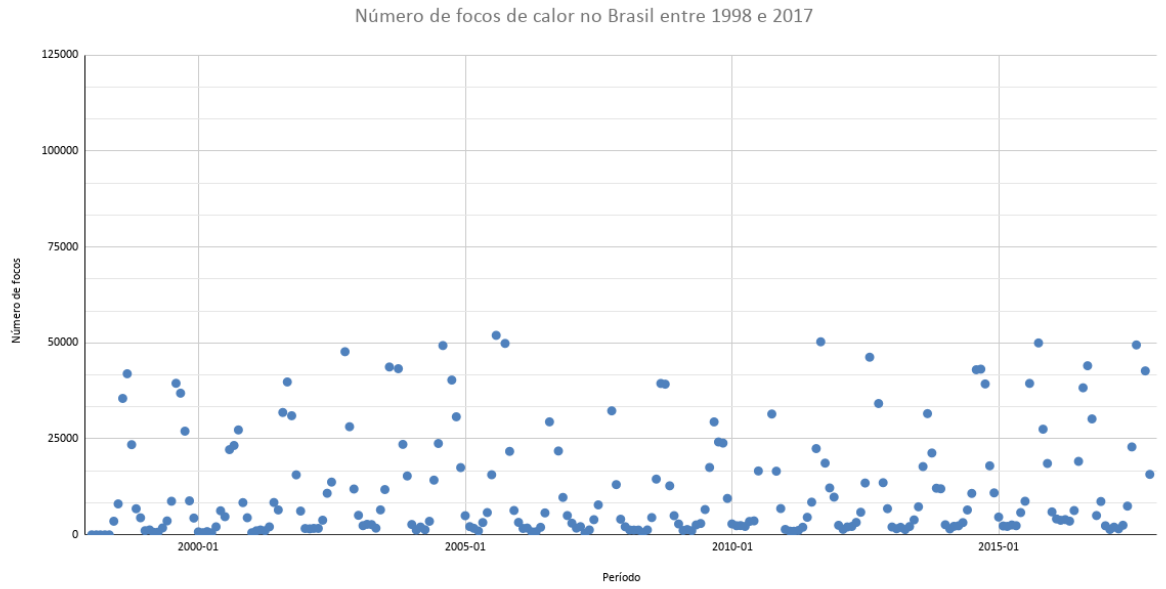
Figura 8 – Série temporal de focos de calor no Brasil antes da remoção de *outliers*



Fonte: Elaborada pelo autor

O gráfico da base após a remoção dos *outliers* pode ser observado na Figura 9. Nele é possível observar que os valores da série estão mais próximos e os picos apresentam padrões mais regulares.

Figura 9 – Série temporal de focos de calor no Brasil após remoção de *outliers*



Fonte: Elaborada pelo autor

Os *outliers* dessa base foram removidos antes da execução dos algoritmos e da definição das bases de treino.

5.1.3 Definição das bases de treinamento

Após a remoção dos *outliers* definiu-se um valor para a quantidade de registros na base de treinamento. As bases de treinamento nada mais são que os últimos n valores da base original que são separados para comparar com os valores estimados durante os cálculos dos modelos de predição. Considerando uma base de tamanho T e um valor de n elementos pra base de teste, são formados dois conjuntos de valores: os $T - n$ primeiros registros que serão utilizados na execução da predição e os n últimos elementos pra comparação entre o resultado encontrado e o esperado. Para a primeira base escolheu-se o valor de 60 registros e para a segunda base 24 registros. Já que a sazonalidade das séries é anual escolheu-se valores múltiplos de 12.

5.2 Execução

Uma vez que foram definidas as bases de treinamento para a execução dos algoritmos o próximo passo foi definir o valor dos seus parâmetros de execução. Para a execução do SARIMA se faz necessário os sete parâmetros $(p, d, q)(P, D, Q, m)$, sendo os parâmetros (p, d, q) os parâmetros que lidam com fatores de tendência e os parâmetros (P, D, Q, m) que lidam com os fatores de sazonalidade.

Esses parâmetros podem ser configurados analisando gráficos de autocorrelação e autocorrelação parcial, mas isso pode exigir domínio técnico sobre as informações da base. Uma vez que a escolha dos parâmetros influencia diretamente no resultado encontrado daquele modelo, foi necessário definir uma forma de buscar os parâmetros com melhores resultados, e a forma escolhida foi o *GridSearch*.

O *GridSearch* é uma busca extensiva entre combinações de diferentes valores definidos para cada um dos parâmetros. Para cada combinação de parâmetros o modelo deve ser executado completamente, ou seja, estimar todos os valores definidos na base de teste. Porém ao final da execução de todas as combinações, foram obtidos resultados mais otimizados, já que uma análise das séries temporais não garante a melhor escolha pra definição dos parâmetros e os melhores resultados encontrados podem ser não intuitivos.

O *GridSearch* é um método altamente paralelizável uma vez que pra cada combinação do conjunto de parâmetros $(p, d, q)(P, D, Q, m)$ a execução é independente e, por tanto, pode ser delegada a outros processadores ou núcleos de processamento dependendo do computador utilizado. Para gerar diferentes combinações foram utilizadas variações entre 0 e 2 para os parâmetros (p, q) e (P, Q) e variações entre 0 e 1 para os parâmetros d e D. O parâmetro m foi definido com valor 12 por conta da periodicidade mensal da série e sazonalidade observada.

Durante cada execução independente para uma determinada combinação de parâmetros, a predição foi realizada, utilizando os valores definidos para base de teste. Conforme descrito na Subseção 5.1.3 são gerados dois conjuntos: o conjunto de valores a ser utilizado pela predição e a base de teste para comparação. A cada valor do conjunto da base de teste é feito a predição do mesmo e após isso esse valor da base de teste passa a integrar o conjunto de valores a ser utilizado na próxima predição. Esse processo se repete com todos os n valo-

res da base de teste. A esse esquema de validação é chamado de *walk forward* já que a os valores vão “movendo“ adiante a medida que as predições são feitas (II; DAHLQUIST, 2010). Uma vez finalizada a iteração pelos n valores, o RMSD é calculado utilizando o conjunto dos valores estimados e os valores da base de teste.

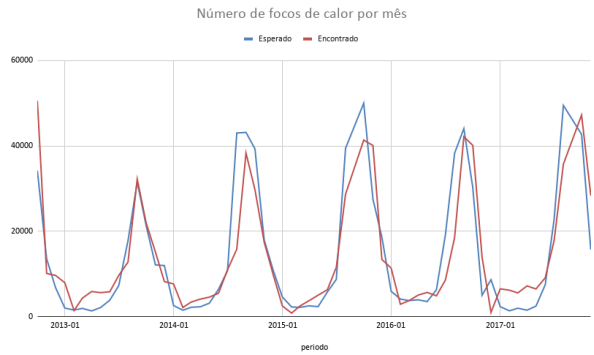
Para a execução do *Random Forest* foi necessário somente a definição do número de árvores de decisão utilizadas para realização da estimativa. Foram testados diferentes valores crescentes no intuito de encontrar um valor mais otimizado que apresentasse boas estimativas num tempo de processamento razoável. Durante a execução do mesmo também foi utilizada a estratégia *walk forward* para o treinamento e ao final de todas iterações era calculado o RMSD utilizando o conjunto dos valores estimados e os valores da base de teste.

5.3 Resultados

Para comparar os resultados encontrados pelos modelos SARIMA e *Random Forest* foi utilizado o valor calculado do RMSD para cada base de dados. O menor valor calculado representa que o modelo alcançou melhores resultados que o outro. Foram gerados gráficos comparando os valores encontrados nas predições e os valores esperados da base de teste, afim de facilitar o entendimento desses resultados proporcionando uma visão mais detalhada. Optou-se por mostrar somente esse recorte das séries temporais para facilitar a visualização dos gráficos.

Utilizando o modelo SARIMA para a execução da primeira base, a melhor combinação de parâmetros encontrada foi [(2, 0, 1), (2, 1, 2, 12)] apresentando um RMSD de 7054.908 aproximadamente. Na Figura 10 observa-se que os valores estimados no geral tem um desenho próximo da curva esperada. Os vales das curvas se aproximam do valor esperado no entanto as maiores discrepâncias notadas foram nos picos.

Figura 10 – Predição do número de focos de calor por mês (SARIMA)



Fonte: Elaborada pelo autor

No caso do *Random Forest*, o número de árvores de decisão utilizado foi de 1000 e o valor encontrado do RMSD foi de 6809.522. Os testes para definir o valor do parâmetro variaram entre 1000 e 10000, mas apresentaram pouca diferença no resultado, variando o valor do RMSD em 200 tanto para mais quanto para menos. Devido a natureza do algoritmo

que estima resultados diferentes a cada execução, optou-se pelo uso do valor 1000 pela maior velocidade de execução com precisão semelhante a de uso de valores maiores como 10000. A Figura 11 representa o resultado encontrado para esse modelo:

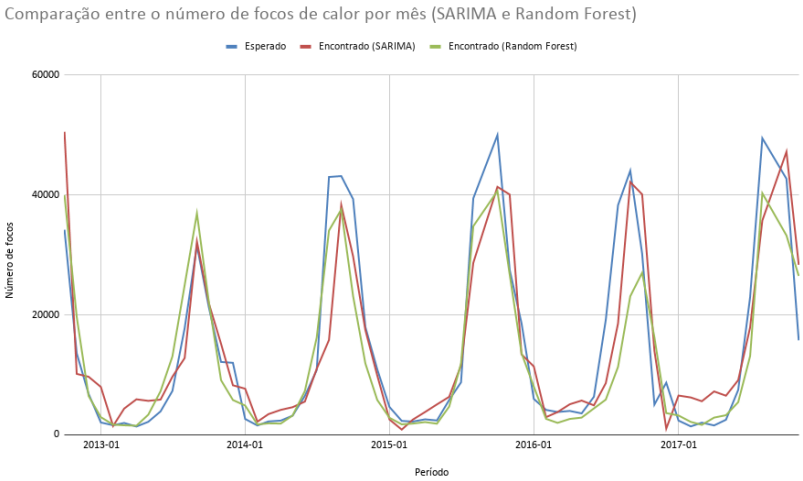
Figura 11 – Predição do número de focos de calor por mês (*Random Forest*)

Fonte: Elaborada pelo autor

Como se observa na Figura 11 o *Random Forest* se aproximou ainda mais das curvas tendo resultados muito próximos, principalmente nos vales que demarcam a sazonalidade. Consequentemente, obteve um resultado um pouco melhor que o SARIMA, tendo uma diferença de RMSD de 245,386. Mesmo assim ainda teve dificuldades de predizer com maior precisão os picos dos valores, mesmo após a remoção dos *outliers*, assim como o SARIMA.

A Figura 12 permite uma visualização mais precisa desses detalhes das diferenças dos resultados entre os dois modelos:

Figura 12 – Predição do número de focos de calor por mês (SARIMA e *Random Forest*)



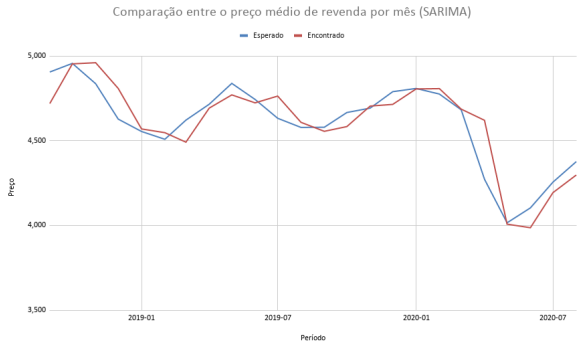
Fonte: Elaborada pelo autor

Através da Figura 12 é possível perceber que a remoção dos *outliers* pode ter contribuído pra qualidade da predição em ambos modelos, uma vez que ambos modelos tiveram maior dificuldade em predizer esses valores. Picos ainda mais destoantes do que os espera-

dos talvez pudessem prejudicar o restante das predições. No entanto as diferenças nos picos ainda se mostrou por vezes significativa, especialmente no caso do *Random Forest*, embora ele tenha tido melhor desempenho no geral.

Para a execução da segunda base, a melhor combinação de parâmetros encontrada para o modelo SARIMA foi [(0, 1, 1), (1, 0, 2, 12)] apresentando um RMSD de 0.109. Na Figura 13 é possível observar que os resultados previstos em relação aos resultados esperados levando em consideração que a escala está mais aproximada pra facilitar a visualização. No geral os resultados foram bons, uma vez que o formato da curva ficou relativamente próximo do esperado. Os maiores erros podem ser notados nas partes descendentes da curva que ficaram um pouco fora do esperado.

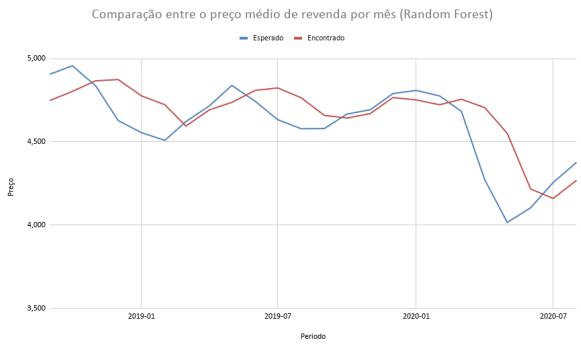
Figura 13 – Predição do preço médio de revenda da gasolina por mês (SARIMA)



Fonte: Elaborada pelo autor

Para a execução do *Random Forest* na segunda base, o número de árvores de decisão utilizadas foi 1000 e o RMSD encontrado foi 0,185. Como se observa na Figura 14, o *Random Forest* obteve uma curva com formato parecido com o esperado mas com dificuldades de prever as quedas. As predições não acompanharam tão bem seja as quedas mais suaves ou a queda mais acentuada dos valores no final. Por conta disso obteve um resultado consideravelmente pior que o SARIMA.

Figura 14 – Predição do preço médio de revenda da gasolina por mês (*Random Forest*)

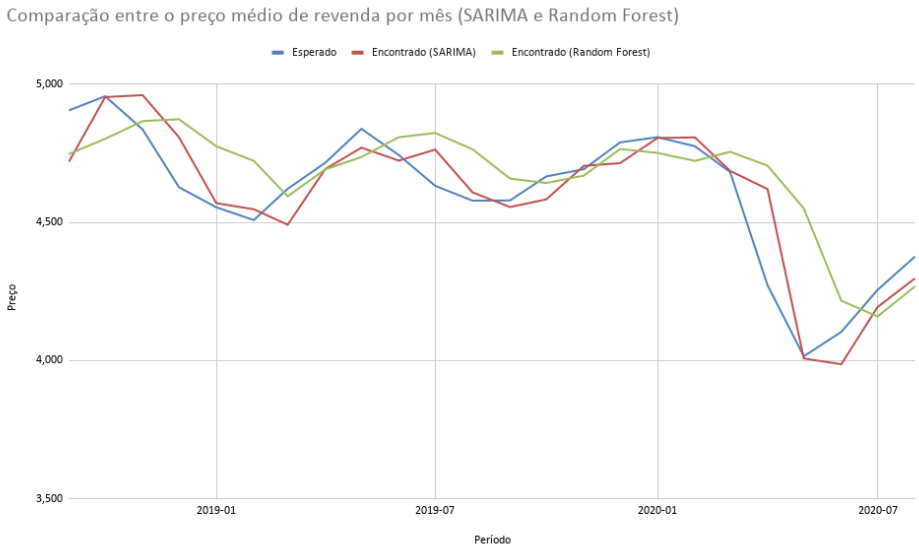


Fonte: Elaborada pelo autor

A Figura 15 demonstra com mais clareza as diferenças entre os dois modelos e a

superioridade do SARIMA nesse caso. Visto que a diferença do valor do RMSD foi de 0,74, o *Random Forest* não foi tão bom para prever comportamentos mais aleatórios da base.

Figura 15 – Predição do preço médio de revenda da gasolina por mês (SARIMA e *Random Forest*)



Fonte: Elaborada pelo autor

Por fim a Tabela 1 apresenta um compilado dos resultados encontrados para ambas as bases de dados e os modelos utilizados.

Tabela 1 – Compilado dos resultados encontrados

Modelo de predição	**Base de dados**	**Parâmetro utilizado**	**RMSD**
SARIMA	Focos de Calor	(2, 0, 1), (2, 1, 2, 12)	7054,908
RANDOM FOREST	Focos de Calor	1000	6809,522
SARIMA	Preço médio Gasolina em MG	(0, 1, 1), (1, 0, 2, 12)	0,109
RANDOM FOREST	Preço médio Gasolina em MG	1000	0,185

Fonte: Elaborada pelo autor.

6 Conclusão

6.1 Resultados e considerações

Ambos modelos tiveram bom desempenho. SARIMA, mesmo sendo voltado principalmente pra bases de dados com sazonalidade, teve um desempenho um pouco pior do que o Random Forest no caso da base com sazonalidade explícita. Por outro lado a série que apresentava uma leve tendência de crescimento o SARIMA teve desempenho muito bom principalmente em relação ao Random Forest.

Também observou-se que o Random Forest de fato alcançou bons resultados mesmo com amostras relativamente pequenas como as bases de dados utilizadas. Porém a quantidade de árvores de decisão não foi o fator crucial pra qualidade de resultados nas medidas utilizadas. Ele se saiu melhor no caso em que se separou uma quantidade maior de valores pra base de teste.

Por fim nenhum dos dois modelos se apresentou evidentemente superior ao outro. Cada um deles um obteve desempenho melhor num cenário específico.

Uma das grandes limitações pra uma pesquisa dessa natureza é o poder computacional disponível. Uma vez definidos os parâmetros de execução, ambos modelos executam com relativa velocidade em bases menores como as utilizadas, mas definir tais parâmetros pode ser uma árdua tarefa, especialmente em bases de dados gigantes.

Devido a limitação de poder de processamento disponível para a realização do trabalho, optou-se pelo uso de bases de dados menores. Porém mesmo levando em consideração esse aspecto, as bases utilizadas possuem relevância pra possíveis usos empresariais e ou governamentais.

A base de preços médios de gasolina é um bom exemplo de como o uso de métodos preditivos com a devida antecedência se demonstram ferramentas poderosas na tomada de decisões mesmo utilizando máquinas de uso pessoal para a execução desses modelos. Prever tais informações pode ter um impacto direto nos gastos e projeções de uma empresa que depende do uso de carros ou outros tipos de automóveis. Além disso os preços de produtos combustíveis impacta os preços de entrega e distribuição e consequentemente afeta o valor final das mercadorias. No entanto é preciso sempre ter em mente que no caso pode haver interferência de variáveis externas, como decisões políticas, e isso afeta o modelo e os resultados.

No caso da base de focos de calor, o governo pode se planejar seu orçamento com antecedência bem como as medidas cabíveis para remediar os incêndios. Especialmente levando em consideração a natureza da base, possibilita-se que os recursos sejam alocados de forma mais dinâmica a depender da época do ano. E para superar as questões de poder computacional, governos e grandes empresas podem utilizar do processamento distribuído na

nuvem.

6.2 Trabalhos futuros

Para um maior aprofundamento na gama de modelos preditivos para se comparar com o SARIMA e o Random Forest pode-se utilizar o *Support Vector Machines* e o *Artificial Neural Network* como utilizados por Kumar e Thenmozhi (2014). Também pode-se utilizar outros tipo de modelos como modelos híbridos, bem como outros métodos de comparação como o *Normalised Mean Squared Error*(NMSE). Além disso, é interessante explorar bases de dados com outras características como a presença de sazonalidade e tendência ao mesmo tempo.

Referências

ANP. *Agência Nacional do Petróleo, Gás Natural e Biocombustíveis*. 2020. Disponível em: <http://www.anp.gov.br/precos-e-defesa-da-concorrencia/precos/levantamento-de-precos/serie-historica-levantamento-precos>. Acesso em: 20 ago. 2020. Citado nas páginas 20, 24 e 25.

BERTHOLD, M.; HAND, D. J. *Intelligent Data Analysis: An Introduction*. 2. ed. [S.l.]: Springer-Verlag Berlin Heidelberg, 2003. ISBN 978-3-540-43060-5,978-3-540-48625-1. Citado na página 11.

BOX, G. E.; JENKINS, G. M. Time series analysis: Forecasting and control san francisco. *Calif: Holden-Day*, 1976. Citado na página 14.

BOX, G. E. et al. *Time Series Analysis: Forecasting and Control*. 5. ed. [S.l.]: Wiley, 2015. (Wiley Series in Probability and Statistics). ISBN 1118675029,9781118675021. Citado nas páginas 11 e 13.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996. Citado na página 16.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 16.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International journal of forecasting*, Elsevier, v. 22, n. 4, p. 679–688, 2006. Citado na página 17.

II, C. D. K.; DAHLQUIST, J. A. *Technical analysis: the complete resource for financial market technicians*. [S.l.]: FT press, 2010. Citado na página 28.

KANE, M. J. et al. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC bioinformatics*, Springer, v. 15, n. 1, p. 276, 2014. Citado na página 18.

KUMAR, M.; THENMOZHI, M. Forecasting stock index returns using arima-svm, arima-ann, and arima-random forest hybrid models. *International Journal of Banking, Accounting and Finance*, Inderscience Publishers Ltd, v. 5, n. 3, p. 284–308, 2014. Citado nas páginas 18 e 33.

MORETTIN, W. d. O. B. P. A. *Estatística Básica*. 6. ed. [S.l.]: Saraiva, 2010. ISBN 978-85-02-08177-2. Citado na página 14.

NOUREEN, S. et al. A comparative forecasting analysis of arima model vs random forest algorithm for a case study of small-scale industrial load. 2019. Citado na página 18.

RASCHKA, V. M. S. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. 3. ed. [S.l.]: Packt Publishing - ebooks Account, 2019. ISBN 1789955750, 978-1789955750. Citado na página 20.

SNIF. *Sistema Nacional de Informações Florestais - SNIF*. 2018. Disponível em: <https://dados.gov.br/dataset/sistema-nacional-de-informacoes-florestais-snif>. Acesso em: 20 ago. 2020. Citado nas páginas 19, 23 e 24.

TYRALIS, H.; PAPACHARALAMPOUS, G. Variable selection in time series forecasting using random forests. *Algorithms*, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 114, 2017. Citado na página 18.