UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO MECÂNICA

David Teles Eller

PREVISÃO DE DEMANDA COM O USO DE MODELOS AUTO REGRESSIVOS E CLUSTERIZAÇÃO DE SÉRIES TEMPORAIS: UM ESTUDO DE CASO DE UMA EMPRESA DE VAREJO

Trabalho Conclusão do Curso de Graduação em Engenharia de Produção Mecânica do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito para a obtenção do Título de Engenheiro Mecânico, habilitado em Produção.
Orientador: Prof. Dr. Carlos Ernani Fries

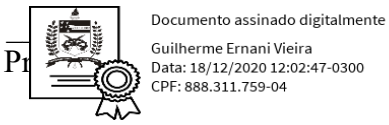
Florianópolis
2020

David Teles Eller

PREVISÃO DE DEMANDA COM O USO DE MODELOS AUTO REGRESSIVOS E CLUSTERIZAÇÃO DE SÉRIES TEMPORAIS: UM ESTUDO DE CASO DE UMA EMPRESA DE VAREJO

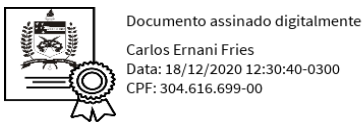
Este Trabalho de Conclusão de Curso foi julgado adequado e aprovado, em sua forma final, pelo Curso de Graduação em Engenharia de Produção Mecânica, da Universidade Federal de Santa Catarina.

Florianópolis, 18 de dezembro de 2020.

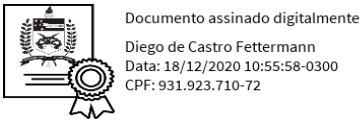


Pr

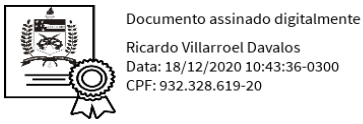
Banca Examinadora:



Prof. Carlos Ernani Fries, Dr.
Orientador
Universidade Federal de Santa Catarina



Prof. Diego de Castro Fetterman, Dr.
Universidade Federal de Santa Catarina



Prof. Ricardo Villarroel Dávalos, Dr.
Universidade Federal de Santa Catarina

AGRADECIMENTOS

Agradeço aos meus pais, Clayre Teles Eller e Daniel Eduardo Eller Junior, por todo o amor, suporte e carinho ao longo da jornada. Sem o apoio recebido, nada disso teria sido possível.

Aos meus irmãos, Daniel Eduardo Eller Neto, Filipe Teles Eller e Lilith Eller. Perto ou longe, a amizade e o companheirismo foram fundamentais para me trazer até aqui.

Aos demais familiares, pelos conselhos, ensinamentos e momentos compartilhados.

Ao meu orientador, Professor Dr. Carlos Ernani Fries, pelos conhecimentos compartilhados e pelo auxílio na construção deste trabalho.

Aos amigos da graduação, com quem nos últimos anos vivi momentos de realizações, dúvidas, decepções, alegrias e tristezas. Em especial, agradeço à Catarina, Daniel, João, Jorge, Kalina e Manuela pela amizade e pelo suporte emocional para o desenvolvimento deste trabalho.

À Empresa Júnior de Engenharia de Produção (EJEP), pelo grande desenvolvimento pessoal, profissional e pelas portas que me foram abertas. Em especial, agradeço à Alissa e Gabriela por terem sido referências.

Por fim, agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela oportunidade de realizar intercâmbio acadêmico, período em que tive o primeiro contato com o tema deste trabalho.

RESUMO

O processo de previsão de demanda tem por objetivo reduzir a incerteza em relação aos valores de demandas futuras. A redução da incerteza de demandas futuras permite às empresas reduzirem custos de estocagem, custos de perda de pedidos, aumentarem o capital de giro e tornarem mais eficiente a gestão da cadeia de suprimentos. O presente trabalho tem por objetivo analisar modelos de previsão para produtos de um comércio varejista localizado no Norte do país, que possui o garimpo como principal mercado alvo. Os modelos de previsão utilizados são pertencentes à família de modelos auto regressivos SARIMA. A parametrização dos modelos foi realizada por três métodos: *Grid Search*, método de otimização *Auto Arima* e proposição de um modelo de generalização dos parâmetros por meio da clusterização das séries temporais. O modelo de generalização dos parâmetros consistiu na utilização da métrica de dissimilaridade *Dynamic Time Warping* e do método *Ward* de clusterização aglomerativa para a generalização dos parâmetros dos modelos de previsão para séries de um cluster. Os resultados foram analisados pelas métricas RMSE e MAE e indicam que os parâmetros obtidos por meio da técnica *Grid Search* geram previsões com menor RMSE e MAE. A parametrização pelo cluster obteve resultados com menor RMSE e MAE para 10 das 15 séries temporais quando comparadas com os resultados obtidos com o método *Auto Arima*, o que indica que a clusterização pode auxiliar no processo de parametrização de modelos de previsão. Os tempos de processamento da técnica *Grid Search* foram significativamente superiores em relação aos tempos de processamento dos demais métodos e ressaltam a importância da definição do objetivo do modelo de previsão.

Palavras-chave: previsão de demanda, SARIMA, clusterização de séries temporais.

ABSTRACT

The demand forecasting process aims to reduce uncertainty in relation to the values of future demands. Reducing the uncertainty of future demands allows companies to reduce stocking costs, stockout costs, increase working capital and make supply chain management more efficient. The present work aims to propose forecasting models for products of a retail trade located in the North of the country, which has gold mining industry as the main target market. The forecast models used are from the SARIMA class of regressive models. The parameterization of the models was made by three methods: Grid Search, Auto Arima optimization method and a proposal of generalization of the parameters through time series clustering. The generalization model of the parameters consisted of using the dissimilarity metric Dynamic Time Warping and the Ward method of agglomerative clustering to generalize the parameters of the forecasting models for the series of a cluster. The results were analyzed using the RMSE and MAE metric and indicated that the parameters obtained through the Grid Search method generate forecasts with lower RMSE and MAE. Parameterization by the cluster obtained results with lower RMSE and MAE for 10 of the 15 time series when compared with the Auto Arima method, which indicates that clustering can help in the parameterization process of forecasting models. The processing times of the Grid Search method were significantly higher than the processing times of the other methods and highlight the importance of defining the objective of a forecasting model.

Keywords: demand forecasting, SARIMA, time series clusterization.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O garimpo, definido na Constituição de 1988 como a localidade onde é realizada a atividade de extração substâncias mineráveis garimpáveis, dentre as quais incluem-se minérios como o ouro, diamante, cassiterita, entre outros, como atividade econômica, desempenhou um papel fundamental na estruturação econômica, geográfica e social da região Amazônica.

Para Coelho, Wanderley e Costa (2017), sob ótica social, o garimpo caracterizou-se como uma opção para os imigrantes excluídos dos enfraquecidos projetos de colonização e como ameaça aos índios e ribeirinhos que habitavam a região antes da chegada das atividades capitalistas. Nos aglomerados populacionais criados para o apoio das atividades garimpeiras, instalaram-se os comércios para o abastecimento local e financiamento dos garimpeiros com mercadorias e subsídios diretos de capital. Os grupos sociais que compunham estes aglomerados, compostos por comerciantes, compradores de ouro, pecuaristas, garimpeiros e índios não conviviam de forma harmoniosa, caracterizando o período por conflitos e lutas.

Ainda segundo os autores, o perfil dos atores pode ser classificado em três grupos. O primeiro grupo é composto por garimpeiros artesanais, caracterizados pelo baixo uso de tecnologia, utilização de mão de obra familiar, produção em baixa escala, reduzido poder político e realização da atividade na informalidade. O segundo, antagônico ao primeiro, é constituído por empresas mineradoras com o emprego de tecnologia de ponta, elevado apoio político e financiados por bancos. O terceiro, intermediário entre os anteriores, são os "dragueiros" ou "balseiros", que possuem meios de produção para a extração dos minerais, pagam aos trabalhadores percentuais do montante extraído e são financiados por comerciantes da região por meios de vendas a prazo. Uma parte dos balseiros realiza a atividade na informalidade e outra parte, procura formalizar-se por meio de empresas ou na forma de cooperativas.

Do ponto de vista legal, a atividade de garimpagem é regulamentada pelo Governo Federal e é considerada ilegal quando realizada em áreas determinadas por lei, como por exemplo, terras indígenas, áreas de preservação, zonas de fronteira, de riscos ambientais, entre outras. O Departamento Nacional de Produção Mineral (DNPM) é o órgão responsável pela emissão da Lavra Garimpeira, que formaliza e concede a permissão para a realização da atividade, priorizando a concessão da permissão para empresas mineradoras e cooperativas, gerando uma significativa taxa de informalidade. Nesse contexto, o Estado atua de forma dúbia, ora com repressões ostensivas, ora, devido às pressões sociais, de forma tolerante, todavia, sem

tornar mais flexíveis as normas de regularização ou atender as demandas das reivindicações da classe (BAÍA JUNIOR, 2014; KOLEN; THEIJE; MATHIS, 2013; SOUSA *et al.,* 2011).

Na esfera econômica, fatores como a cotação do preço do ouro possuem influência sobre a produção de ouro na região. Além disso, fatores geoclimáticos como o regime de chuvas da região afetam o volume de ouro extraído (CAHETÉ, 1998).

No cenário em que o volume de ouro garimpado é fortemente influenciado por diversos fatores (legais, econômicos, geoclimáticos, sociais e políticos), existem os comércios varejistas e atacadistas fornecedores de equipamentos, materiais e maquinários sujeitos às fortes variações de demanda dos produtos comercializados.

1.1 IMPORTÂNCIA

Historicamente, a previsão de demanda na empresa em questão é realizada de forma qualitativa com base na *expertise* dos gestores do negócio. Apesar de possuir histórico de vendas de longos períodos de tempo, que permitem a previsão de períodos futuros, estes não são utilizados sistematicamente. Associado a isso, os fatores externos influenciadores na demanda do negócio são de complexa previsibilidade e possuem forte impacto sobre a variação da demanda. Com o intuito de reduzir o impacto deste conjunto de fatores a cerca da demanda, a empresa adota a política de manter altos níveis de estoque para evitar rupturas de estoque. Em contrapartida, a política adotada acarreta em ineficiências no processo de planejamento de compras e alto capital imobilizado.

Neste contexto, identificou-se a oportunidade de desenvolver modelos de previsão de demanda a fim de tornar mais eficaz processo de planejamento de compras e reduzir o capital imobilizado da unidade de negócio.

1.2 OBJETIVO GERAL

O objetivo geral do trabalho consiste em analisar modelos de previsão de demanda para um comércio varejista focado em mercadorias para garimpo, localizado na região Norte do país.

1.3 OBJETIVOS ESPECÍFICOS

Para atingir o objetivo geral, o trabalho deve atingir os seguintes objetivos específicos:

a) consolidar os dados históricos da demanda de uma seleção de mercadorias comercializadas;
b) definir os modelos de previsão para a amostra considerada de mercadorias;
c) analisar a acurácia dos modelos com os dados obtidos;
d) avaliar a adequação dos modelos de previsão para aqueles artigos não incluídos no processo inicial de seleção de modelos, referido no item (b).

1.4 ESTRUTURA DO TRABALHO

Este trabalho está dividido em 6 capítulos. O primeiro capítulo contextualiza o problema, define os objetivos do trabalho bem como sua importância, limitações e estrutura.

O segundo capítulo constitui a fundamentação teórica do trabalho. Nesse capítulo são abordados os temas: séries temporais, transformação de séries temporais, métodos de previsão em séries temporais, e clusterização de séries temporais,

No segundo capítulo, são apresentados conceitos relevantes que suportam o desenvolvimento do trabalho.

O terceiro capítulo discorre sobre os procedimentos metodológicos utilizados para o desenvolvimento do trabalho.

O quarto capítulo apresenta detalhadamente o desenvolvimento dos procedimentos metodológicos adotados.

Os resultados são apresentados e analisados no capítulo cinco.

Por fim, no sexto capítulo são apresentadas conclusões do trabalho e sugestões para trabalhos futuros.

1.5 LIMITAÇÕES DO TRABALHO

O presente trabalho teve por objetivo propor modelos de previsão de demanda para uma seleção de mercadorias, baseando-se exclusivamente no histórico de vendas. Não foram levadas em consideração variáveis exógenas como a evolução do preço do ouro, oferta de concorrentes ou dados geoclimáticos.

Realizou-se a previsão de demanda para uma amostra de 22 mercadorias em um total de 16.000 produtos devido aos fatores de tempo e recursos computacionais limitados. Entretanto, entende-se que a metodologia utilizada poderia ser expandida para a análise dos demais produtos não analisados.

2 FUNDAMENTAÇÃO TEÓRICA

Previsão de demanda é um processo que permite estimar volumes de vendas futuros e influencia no ganho de competitividade de uma organização. Para Martins e Laugeni (2015), a previsão de demanda é um processo metodológico para determinação de valores futuros, fundamentado em modelos estatísticos, econométricos, matemáticos ou em modelos subjetivos.

Os métodos de previsão podem ser classificados em quantitativos e qualitativos (SLACK; CHAMBERS; JOHNSTON, 2009). Dentre os métodos quantitativos, os principais são os causais, que avaliam relações de causa e efeito entre variáveis, e os de séries temporais, que consistem no tratamento de séries temporais para a previsão de valores futuros (ARCHER, 1980).

Neste capítulo serão apresentados conceitos relacionados a tratamento de séries temporais, dentre eles, a definição de séries temporais e seus componentes, estacionariedade em séries temporais, transformação de séries e métodos de previsão.

Serão abordados também conceitos envolvendo clusterização de séries temporais, métricas de dissimilaridade, clusterização hierárquica aglomerativa e métodos de clusterização.

2.1 SÉRIES TEMPORAIS

Em uma definição ampla, Morettin e Toloi (2006) e Box, Jenkins e Reinsel (1994) conceituam séries temporais como um conjunto de dados ordenados sequencialmente ao longo do tempo. Clark e Downing (2005) acrescenta ao conceito o fato de que tal conjunto de observações deve medir uma mesma grandeza.

Uma série temporal é expressa matematicamente por Fonseca, Martins e Toledo (1985) como um conjunto de valores y_1, y_2, y_3, …, y_n nos tempos t_1, t_2, t_3, …, t_n. Sendo y uma junção de T, expressa na forma $y = f(t)$..

Para Morettin e Toloi (2006) e Ehlers (2007), as séries temporais podem ser classificadas em séries contínuas, quando as observações são feitas continuamente no tempo, ou discretas, quando as observações são realizadas em tempos específicos. As séries contínuas podem ser discretizadas por meio do registro dos valores a certos intervalos de tempo.

Box, Jenkins e Reinsel (1994) ressaltam que as séries temporais possuem como característica inerente a dependência temporal entre os valores observados e que a análise de séries temporais é constituída pela análise de tais dependências. A investigação e identificação

do mecanismo gerador da série temporal permite a previsão de valores futuros da série, dando utilidade ao estudo (MORETTIN; TOLOI, 2006).

Silva e Silva (1999) destacam que as séries temporais são ordenadas cronologicamente e que a alteração na ordem dos dados pode modificar a informação contida na série.

Séries temporais são exemplificadas por Morettin e Toloi (2006) como:

a) Valores diários de poluição na cidade de São Paulo;
b) Valores mensais de temperatura na cidade de Cananéia-SP;
c) Índices diários da Bolsa de Valores de São Paulo;
d) Precipitação atmosférica anual na cidade de Fortaleza;
e) Número médio anual de manchas solares;
f) Registro de marés no porto de Santos.

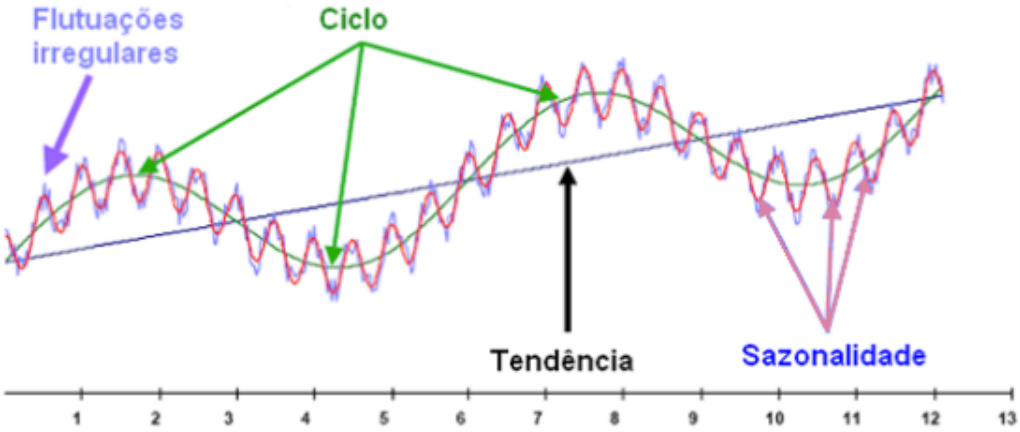
2.1.1 Componentes de uma série temporal

Uma série temporal é composta pela combinação de sinal e ruído. O sinal é definido como o componente previsível, enquanto o ruído, o componente aleatório. O ruído, também é denominado de variáveis aleatórias independentes e identicamente distribuídas (NAU, 2014).

Para Clark e Downing (2005), uma série temporal pode ser decomposta em componentes de tendências, cíclicos, sazonais e residuais. Os componentes de tendência estão relacionados a orientação contínua de crescimento ou decrescimento dos valores da série e está relacionada com o comportamento da série à longo prazo. Os componentes de sazonalidade são descritos como a repetição de um padrão dos valores da série no curto prazo. Milone (2006) descreve os ciclos como movimentos oscilantes dos valores em torno da tendência. A componente aleatória, por sua vez, é descrita por Fonseca, Martins e Toledo (1985) como alterações imprevisíveis sem regularidade nos valores da variável analisada.

A Figura 1 ilustra os conceitos abordados.

Figura 1 - Componentes de uma série temporal



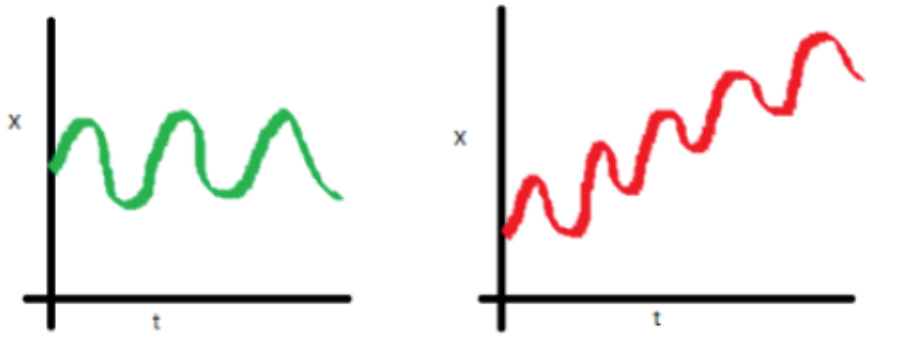
Fonte: Traduzido de Genesis (2018).

2.1.2 Estacionariedade em séries temporais

As séries temporais podem ser classificadas em estacionárias ou não estacionárias. Séries estacionárias possuem média e variância constantes ao longo do tempo, contrariamente às séries não estacionárias (BUENO, 2011). As séries estacionárias desenvolvem-se de forma aleatória em torno de uma média fixa, manifestando um comportamento equilibrado e estável (MORETTIN; TOLOI, 2006).

De forma complementar, Sean Abu (2016) e Box, Jenkins e Reinsel (1994) acrescentam o critério de covariância constante para a caracterização de uma série estacionária. A classificação quanto à estacionariedade pode ser realizada de duas formas: visualmente ou por meio de testes estatísticos. Verifica-se na Figura 2, por meio da inspeção visual, que o gráfico b representa uma série não estacionária pois a média não é constante ao longo do tempo.

Figura 2 - Teste visual de estacionariedade de média não constante

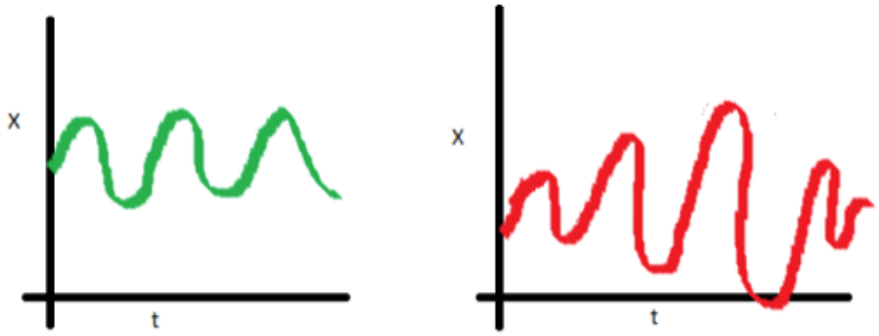


a) Série estacionária b) Série não estacionária

Fonte: Sean Abu (2016)

Na Figura 3 nota-se no gráfico b que a variância não é constante ao longo do tempo devido às fortes variações dos valores de x.

Figura 3 - Teste visual de estacionariedade de variância não constante

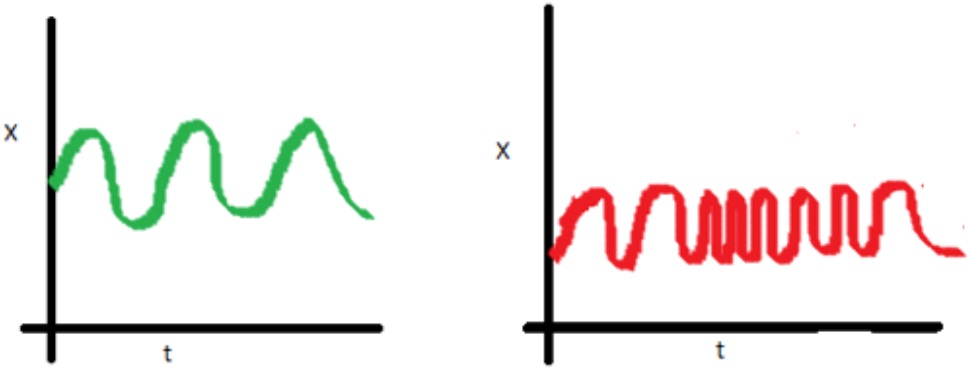


a)Série estacionária b) Série não estacionária

Fonte: Sean Abu (2016)

Na Figura 4, nota-se no gráfico b que a covariância não é constante, o que caracteriza a série como não estacionária.

Figura 4 - Teste visual de estacionariedade de covariância não constante



a)Série estacionária b) Série não estacionária

Fonte: Sean Abu (2016)

2.1.3 Testes de raíz unitária

A inspeção visual de séries dificilmente possibilita diferenciar uma série não estacionária de uma série estacionária (BUENO, 2011). A verificação da não estacionaridade por meio de testes estatísticos é realizada por meio dos testes de raiz unitária. Os testes mais populares são o Teste de Dickey-Fuller, Teste de Dickey-Fuller Aumentado e Teste de Phillips-Perron (MATTOS, 2018). É possível encontrar outros testes na literatura como o KPSS, ERS,

Critério de Informação e Janela Ótima, Raízes Unitárias Sazonais, Quebra Estrutural e Múltiplas Raízes (BUENO, 2011).

2.1.3.1 Teste de Dickey Fuller

Segundo Mattos, o teste de Dickey Fuller verifica se uma série é estacionária testando se ela possui raiz unitária. Para isso, considerando um processo estocástico

$$Y_t = TD_t + Z_t$$
$$TD_t = \psi_0 + \psi_1 t$$

Em que ψ_0, ψ_1 são constantes reais. Z_t é um processo autorregressivo do tipo

$$Z_t = \rho Z_{t-1} + u_t$$

Reescreve-se

$$Y_t = \psi_0(1-\rho) + \psi_1\rho + \psi_1(1-\rho)t + \rho Y_{t-1} + u_t \, .$$

Define-se

$$a = \psi_0(1-\rho) + \psi_1\rho$$
$$b = \psi_1(1-\rho)t$$

Dessa forma

$$Y_t = a + bt + \rho Y_{t-1} + u_t$$

Para o caso particular em que $\psi_0 = \psi_1 = 0$, $a = b = 0$, então:

$$Y_t = \rho Y_{t-1} + u_t$$

Então, o processo será estacionário se $|\rho| < 1$ e não estacionário se $|\rho| \geq 1$. Testa-se a hipótese nula H_0 de $\rho = 1$.

2.1.3.2 Teste aumentado de Dickey Fuller (ADF)

O teste de Dickey Fuller considera o erro um ruído branco, ou seja, deconsidera que o erro seja um processo estacionário qualquer (BUENO 2011). A versão aumentada do teste considera a existência de estruturas de autocorrelação para os erros da equação de teste. São adicionados termos ΔY_t no lado direito da equação.

$$Y_t = a + bt + \lambda Y_{t-1} + \sum_{j=1}^{p} \lambda \Delta Y_{t-j} + \varepsilon_t$$

Onde λ_j $(j = 1, \dots, p)$ são parâmetros e ε_t é um ruído branco.

De forma similar, testa-se a hipótese nula $H_0 : \lambda = 0$ (MATTOS, 2018).

2.2 TRANSFORMAÇÃO DE SÉRIES

As séries não estacionárias, ao contrário das estacionárias, não podem ser estimadas trivialmente (BUENO, 2011). Neste tipo de série a variável y_t é aleatória e impossível de prevê-la perfeitamente (HILL; JUDGE; GRIFFITHS, 2010).

Robert Nau ([201-]) destaca que a maioria dos métodos estatísticos para previsão são baseados na premissa de que séries não estacionárias podem ser "estacionarizadas" por meio de transformações matemáticas. Tal premissa também inclui o fato de que as séries podem ser revertidas às originais ao aplicar o inverso da primeira transformação matemática.

Tornar uma série estacionária também é útil pois torna possível a análise de parâmetros estatísticos como média, variâncias e covariâncias.

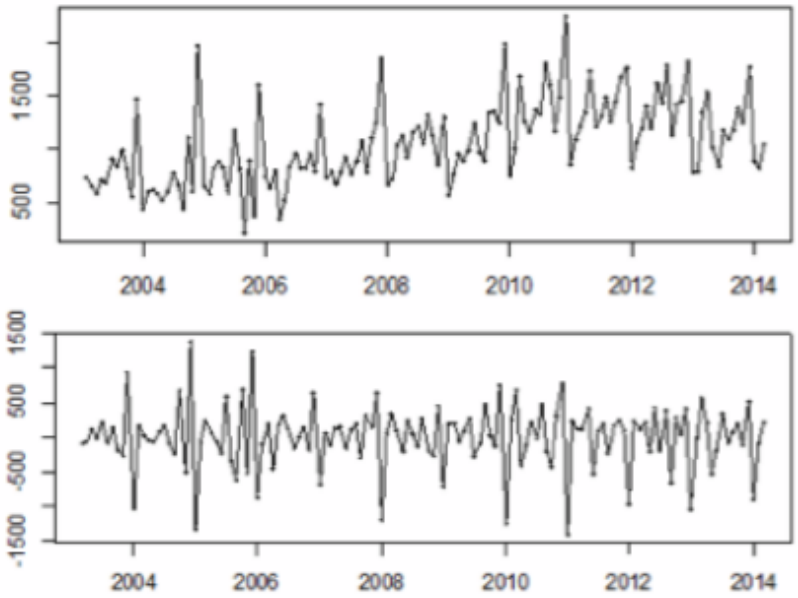
Com o objetivo de transformar uma série não estacionária em estacionária, existem diversas operações que podem ser utilizadas de forma isolada ou em conjunto, sendo as principais a diferenciação e o ajuste logarítimico.

Dada uma série Z_t não estacionária, Morettin e Toloi (2006) definem o processo de diferenciação de primeira ordem como:

$$W_t = Z_t - Z_{t-1}$$

De forma complementar, Robert Nau (2015) ressalta que o processo de diferenciação de séries é indicado para séries de caminhos aleatórios e/ou com tendência acentuada. Visualmente verifica-se o resultado do processo de diferenciação primária na Figura 5, sendo o gráfico superior não estacionário e o inferior, estacionário após o procedimento de diferenciação.

Figura 5 - Processo de diferenciação



Fonte: Martin, Henning, Walter e Konrath (2016).

Morettin e Toloi (2006) citam a Transformação de Box-Cox (1994) como alternativa para a estacionarização de séries econômicas e financeiras, representada da forma

$$Z_t^{\lambda} \begin{cases} \dfrac{Z_t^{\lambda} - c}{\lambda}, & se\ \lambda \neq 0 \\ \log Z_t & se\ \lambda = 0 \end{cases}$$

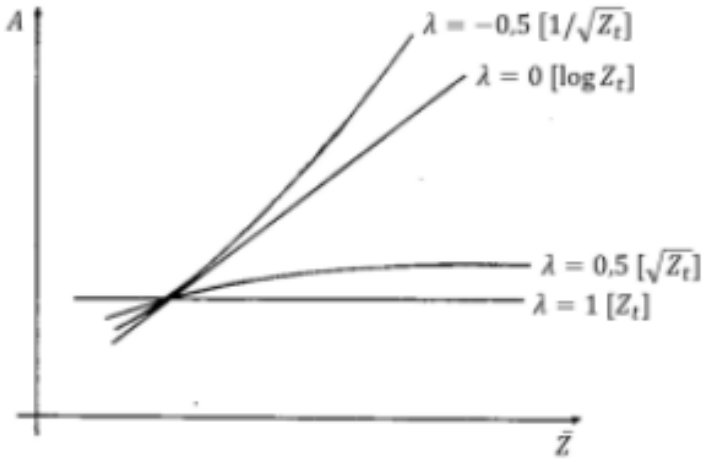
Em que λ e c são parâmetros a serem estimados. Para verificar se a transformação logarítimica é apropriada, utiliza-se o gráfico com a média ($\bar{Z}$) de k observações no eixo das abcissas e a amplitude w no eixo das ordenadas, dadas por:

$$\bar{Z} = \frac{1}{k} \sum_{i=1}^{k} Z_{ti}$$

$$w = \max(Z_{ti}) - \min(Z_{ti}).$$

Se w independer de ($\bar{Z}$), os pontos estarão sob uma mesma reta e não há necessidade de transformação. Caso w seja diretamente proporcional a ($\bar{Z}$), a transformação logarítmica é apropriada.

Figura 6 - Gráfico amplitude x média, ilustrando alguns possíveis valores de λ.



Fonte: Box e Jenkins (1979)

2.3 MÉTODOS DE PREVISÃO EM SÉRIES TEMPORAIS

Previsões em séries temporais utilizam o histórico de demanda da variável analisada para prever a demanda futura (ELSAYED; BOUCHER 1994).

Robert Nau (2014) define:

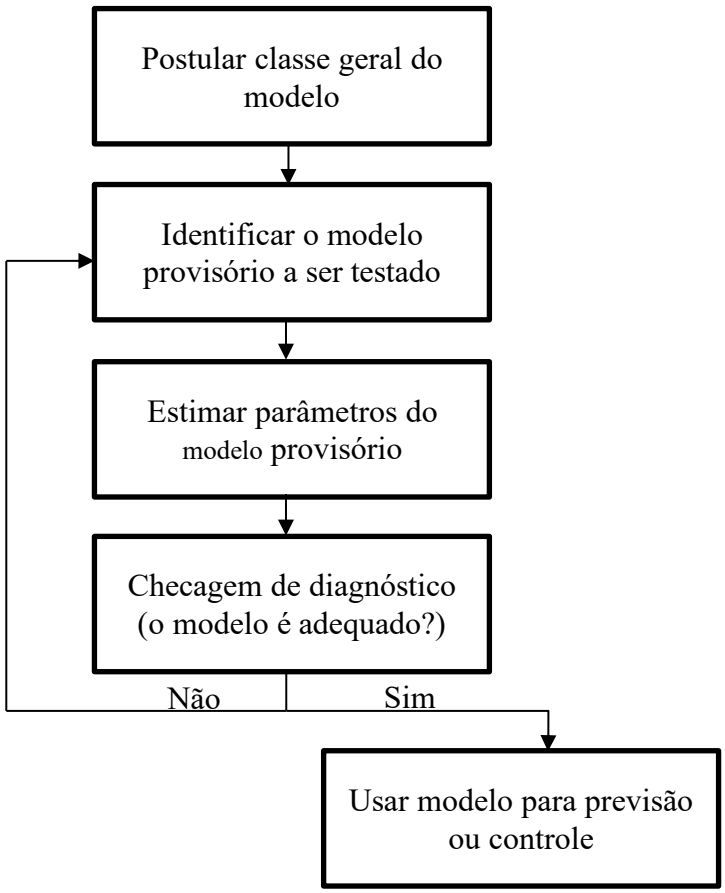
> Previsões podem se dar de diferentes formas - olhando uma bola de cristal ou canecas com folhas de chá, combinando opiniões de *experts*, *brainstorming*, análise de cenários, análise de hipóteses, simulação de Monte Carlo, resolvendo equações que são ditadas por leis físicas ou teorias econômicas - mas previsões estatísticas, tópico principal a ser discutido, é a arte e a ciência de se fazer previsão com dados, com ou sem saber a priori qual equação usar. A ideia é simples: buscar padrões estatísticos nos dados disponíveis que você acredita que continuarão no futuro. Em outras palavras, perceber que a forma que o futuro será se parecerá muito com o presente.

Morettin e Toloi (2006) acrescentam que previsões são insumos para a tomada de decisão, visando objetivos específicos. Para Robert Nau, às previsões estão associados três tipos de riscos que devem ser monitorados e minimizados ao longo do processo de construção:

a) Risco intrínseco: está relacionado com o ruído nas séries. Trata-se da variação aleatória e incerteza do futuro, podendo ser medido pelo "Erro Padrão do Modelo". A escolha do método de previsão está relacionada magnitude do erro intrínseco, modelos mais robustos podem identificar padrões nos ruídos de modelo mais simples, aperfeiçoando a análise da série temporal;
b) Risco de parâmetro: trata-se do erro associado à estimação errônea dos parâmetros do modelo utilizado, como por exemplo, o da inclinação da reta em uma série com tendência. Este tipo de erro é medido pelo "Erro Padrão do Coeficiente" e, em alguns casos, pode ser corrigido aumentando o tamanho da amostra de dados. Tal solução pode não funcionar em outros casos pois a inclusão de valores antigos pode estar associada à inclusão de valores não condizentes com a situação atual da série;
c) Risco do modelo: trata-se da modelagem inadequada da série para a previsão dos modelos futuros e é o que tem maior impacto sobre a qualidade da previsão. Tal risco pode ser reduzido seguindo boas práticas de modelagem.

Box, Jenkins e Reinsel (1994) propõem uma abordagem iterativa para a construção de um modelo de previsão, conforme a Figura 7.

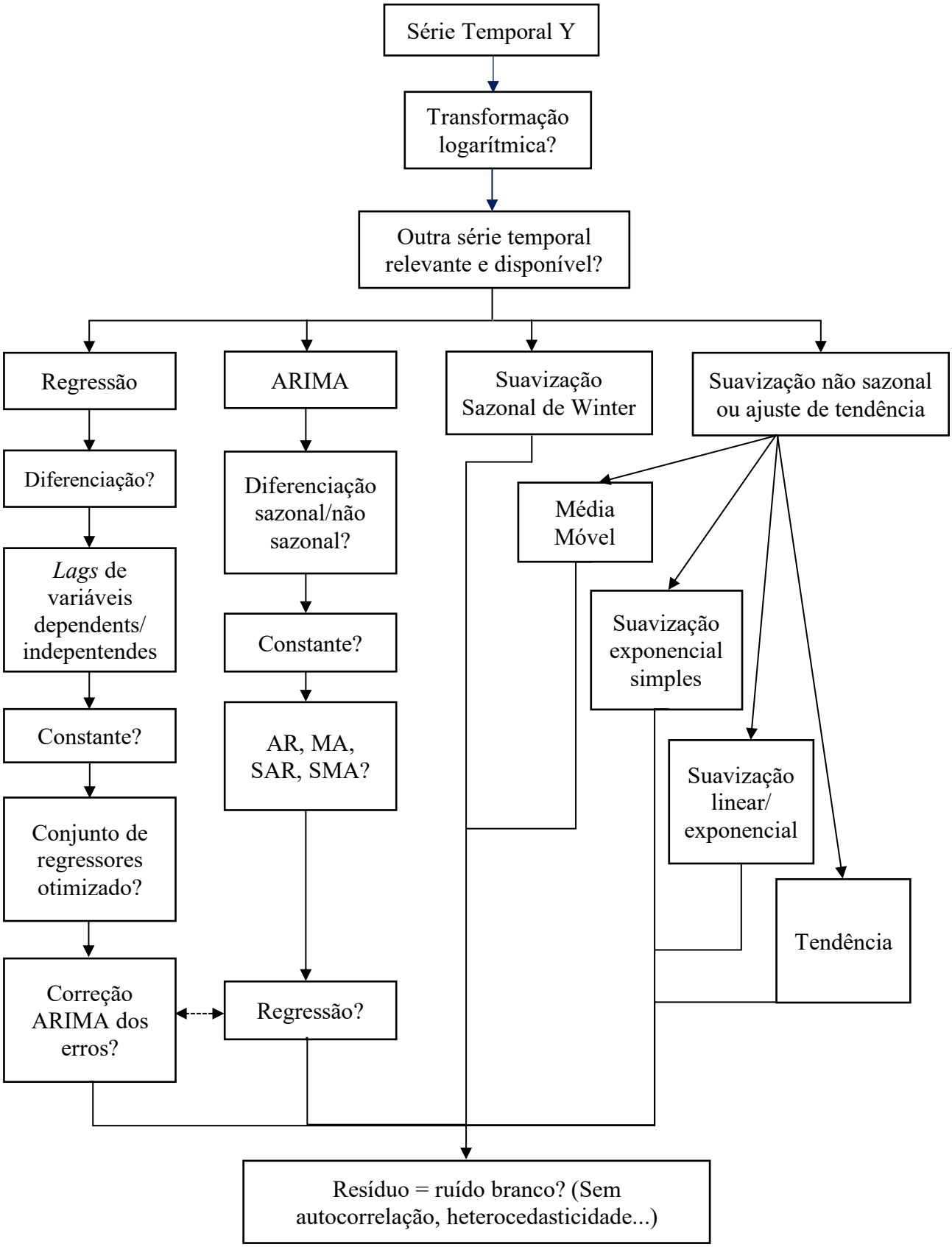
Figura 7 - Fluxograma das etapas do método Box/Jenkins para a construção de um modelo de previsão



Fonte: Adaptado de Box, Jenkins e Reinsel (1994)

Robert Nau (2015) descreve aspectos gerais a serem considerados na escolha da classe do modelo de previsão, sintetizados na Figura 8.

Figura 8 - Modelos de previsão



Fonte: Adaptado (NAU *apud* DUKE UNIVERSITY, 2016).

Para o autor, transformações logarítmicas devem ser aplicadas em séries temporais detentoras de componentes de crescimento geométrico e/ou padrões sazonais multiplicativos com o objetivo de torná-las modeláveis por modelos lineares e transformar padrões sazonais multiplicativos em padrões sazonais aditivos (NAU *apud* DUKE UNIVERSITY, 2016).

Ajustes sazonais devem ser considerados em séries temporais que possuem padrões repetidos ao longo dos períodos considerados, de forma a estimar os valores futuros por meio da extrapolação dos padrões de sazonalidade.

Caso existam séries com variáveis independentes potencialmente relacionadas com a série de interesse, deve-se considerar modelos de regressão para a previsão. Esta classe de modelos também deve ser considerada caso os resíduos de modelos auto regressivos possuam relação cruzada com outras variáveis.

Para séries que não possuem componentes de sazonalidade e tendência, indica-se a utilização da classe de modelos de médias móveis e de suavização exponencial, que assumem que a melhor previsão de um valor futuro é um valor próximo ao da média dos valores anteriores. Ressalta-se que estes modelos são casos particulares da classe de modelos ARIMA.

A Suavização Exponencial de Holt Winters é uma extensão da suavização exponencial que permite a inclusão de componentes de tendência e sazonalidade por meio de equações recursivas. Os parâmetros de suavização, alfa, beta e gama, devem ser estimados simultaneamente, o que dificulta o ajuste do modelo.

Os modelos ARIMA são uma classe abrangente de modelos que incluem caminhos aleatórios, tendências aleatórias, suavização exponencial e modelos auto regressivos. Suas propriedades são apresentadas com maior profundidade nos tópicos abaixo.

2.3.1 ARIMA (p, d, q)

O modelo ARIMA (Auto Regressivo Integrado de Média Móvel) é adequado para modelagem de séries não estacionárias homogêneas (BOX; JENKINS, 1979). Fava (2000) descreve o modelo ARIMA como o resultado da combinação de três filtros: o componente Auto Regressivo (AR), o componente de Integração (I) e o componente de Médias Móveis (MA).

Uma série não estacionária homogênea torna-se estacionária depois de d diferenças, é descrita pelo modelo ARIMA (p, q, d), representada por:

$$w_t = \emptyset_1 w_{t-1} + \cdots + \emptyset_p w_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

onde $w_t = \Delta^d y_t$.

Define-se *B* como o operador de defasagem definido por:

$$B^p y_t = y_{t-p}.$$

Utiliza-se o operador de defasagem

$$\left(1 - \emptyset_1 B - \cdots - \emptyset_p B^P\right) w_t = (1 - \theta_1 B - \cdots - \theta_q B^q)\varepsilon_t$$

Onde

$$w_t = (1 - \mathrm{B})^d y_t$$

que resulta em

$$(1 - B)^d \emptyset(\mathrm{B}) y_t = \theta(B)_{\varepsilon t} .$$

2.3.2 SARIMA (p, d, q) × (P, D, Q)$_s$

Segundo Box, Jenkins e Reinsel (1994) o modelo SARIMA possui além dos componentes (p, d, q) do modelo ARIMA, três componentes sazonais (P, D, Q)$_s$ o que resulta no modelo multiplicativo sazonal SARIMA (p, d, q) x (P, D, Q)$_s$ representado por

$$\emptyset_p(B)\Phi_p(B^s)\nabla^d \nabla_s^D z_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t$$

Em que s é a periodicidade sazonal, $\emptyset_p(B)$ é o coeficiente auto regressivo não sazonal, $\Phi_p(B^s)$ é o coeficiente auto regressivo sazonal, $\Theta_Q(B^s)$ é o coeficiente da média móvel sazonal, $\theta_q(B)$ é o coeficiente da média móvel não sazonal, B é o operador de retardo, ∇^d é o

diferenciador não sazonal, ∇_s^D é o diferenciador sazonal, z_t é a série não diferenciada e ε_té o ruído branco.

2.3.3 SARIMAX

Chen e Tjandra (2014) descrevem o modelo SARIMAX como uma amplificação do modelo SARIMA que inclui variáveis exógenas.

Seja $\hat{y}_t$ a estimativa do modelo SARIMA sem a inclusão de variáveis exógenas e x_{reg} a série temporal externa. A estimativa é modificada, tornando-se $\hat{y}_{t,reg}$ e tem a forma

$$\hat{y}_{t,reg} = \hat{y}_t + \gamma \phi_p(B)\beta_p(B^s)(1-B)^d(1-B^s)^D x_t$$

onde:

$$\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$$

$$\beta_p(B^s) = (1 - \beta_1 B^s - \beta_2 B^{2s} - \cdots - \beta_P B^{Ps})$$

sendo γ o parâmetro de x_{reg} a ser estimado.

2.4 CLUSTERIZAÇÃO DE SÉRIES TEMPORAIS

A tarefa de clusterização consiste em agrupar objetos em grupos cuja a similaridade entre os objetos de um mesmo grupo seja maximizada e a similaridade destes objetos com objetos de grupos diferentes seja minimizada (LI; WU; ZHANG, 2020).

A clusterização pode ser realizada com diversos tipos de dados, como dados binários, categóricos, numéricos, ordinais, relacionais, textuais, espaciais, temporais, espaço-temporais, com imagens, multimídias ou combinações dos mencionados (LIAO, 2005).

Utilizam-se técnicas de agrupamento em conjuntos de dados, que *a priori*, não possuem rotulação, ou seja, não se sabe à qual grupo cada objeto pertence, com o objetivo de extrair conhecimento acerca da estrutura dos dados (SILVA, 2016).

Formalmente, define-se o problema de clusterização como um conjunto de *n* objetos $X = \{X_1, X_2, ..., X_n\}$ em que cada $X_i \in \mathbb{R}^p$ é um vetor de *p* medidas reais que dão dimensão as

características do objeto e que devem ser agrupados em k clusters disjuntos $C = \{C_1, C_2, ..., C_k\}$ respeitando as condições:

1. $C_1 \cup C_2 \cup \ldots \cup C_k = X$;
2. $C_i \neq \emptyset, \forall i, 1 \leq i \leq k$;
3. $C_i \cap C_j = \emptyset, \forall i \neq j, 1 \leq i \leq k, 1 \leq j \leq k$.

Essas condições garantem que um objeto pertença somente à um cluster e que cada cluster contem pelo menos um objeto (HRUSCHKA; EBECKEN, 2003).

2.4.1 Métricas de dissimilaridade

Métricas de dissimilaridade são utilizadas para comparar duas séries temporais e podem ser calculadas de diversas formas, possuindo diferenças no que se refere à custo computacional, sensibilidade à ruídos, entre outros (SILVA, 2016).

Os dados de problemas de clusterização são organizados na forma matricial, em que linhas representam os objetos a serem agrupados e colunas representam *p* atributos dos objetos como em:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

A partir da matriz dos objetos e atributos, calcula-se a matriz de dissimilaridade, na qual são representadas as distâncias entre os pares de objetos da seguinte forma:

$$D = \begin{bmatrix} 0 & \cdots & d_{1,m} \\ \vdots & \ddots & \vdots \\ d_{n,1} & \cdots & 0 \end{bmatrix}$$

Onde *d(i, j)* a dissimilaridade entre o objeto *i* e o *j* (HAN; KAMBER; PEI, c2012).

2.4.1.1 Distância euclidiana

A distância euclidiana é uma métrica de dissimilaridade amplamente utilizada devido à sua simplicidade e rapidez (YE *et al.,* 2019). É uma das métricas de dissimilaridade utilizadas com maior frequência em tarefas de clusterização, porém, no contexto em que os dados são compostos por séries temporais a distância euclidiana pode gerar resultados incoerentes (SILVA, 2016).

A distância euclidiana pode ser calculada por:

$$d_{L_n}(x, y) = \left(\sum_{i=1}^{M} (x_i - y_i)^n \right)^{\frac{1}{n}}$$

Em que *n* é um número inteiro positivo, *M* é o tamanho da série temporal, x_i e y_i são o *i*-ésimo elemento das séries x e y (KEOGH; KASETTY, 2003).

A distância euclidiana possui a limitação em comparar apenas séries temporais com o mesmo número de observações e ser altamente sensível à *outliers*, distorções e ruídos em séries temporais (YE *et al.,* 2019).

2.4.1.2 Dynamic Time Warping

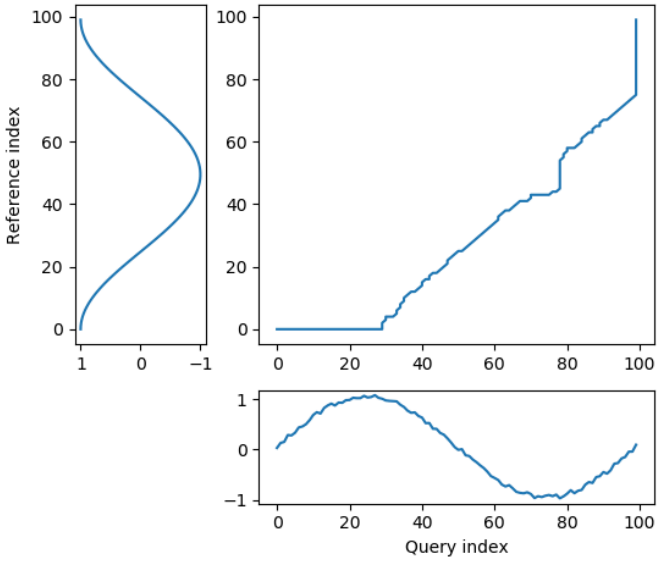
Dynamic time warping (DTW) é um método de otimização de alinhamento de séries temporais que permite o cálculo de métrica de dissimilaridade entre duas sequências (LERATO; NIESLER, 2019).

Segundo Kate (2016), dadas duas séries temporais Q = q_1, q_2, .., q_i, ..., q_n e C = c_1, c_2, .., c_i, ..., c_m a distância DTW é calculada a partir do melhor alinhamento entre as duas séries temporais. O alinhamento das séries é feito através de uma matriz n x m, sendo n o número de termos da série Q e m o número de termos da série C, na qual os termos (i,j) são dados por $(q_i - c_j)^2$, termo que representa o custo do alinhamento do ponto q_i com o ponto c_j. O melhor alinhamento é representado por um caminho W = w_1, w_2, .., w_k, ..., w_K que comece no canto inferior esquerdo da matriz e chegue ao canto superior direito, de forma contínua e que minimize o custo total da distância DTW, dada por:

$$DTW\ (Q,C) = argmin\ _{W=w_1,\ldots,w_k,\ldots,w_K} = \sqrt{\sum_{k=1,w_k=(i,j)}^{K} (q_i - c_j)^2}$$

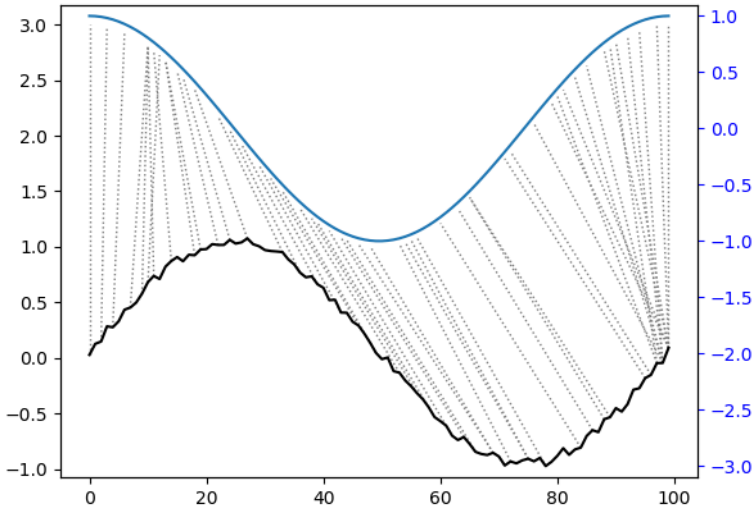
A matriz de distâncias e o resultado do alinhamento de duas séries temporais podem ser observados nas Figuras 9 e 10.

Figura 9 - Matriz de distâncias *Dynamic Time Warping*



Fonte: Github (THE DTW SUÍTE, 2019).

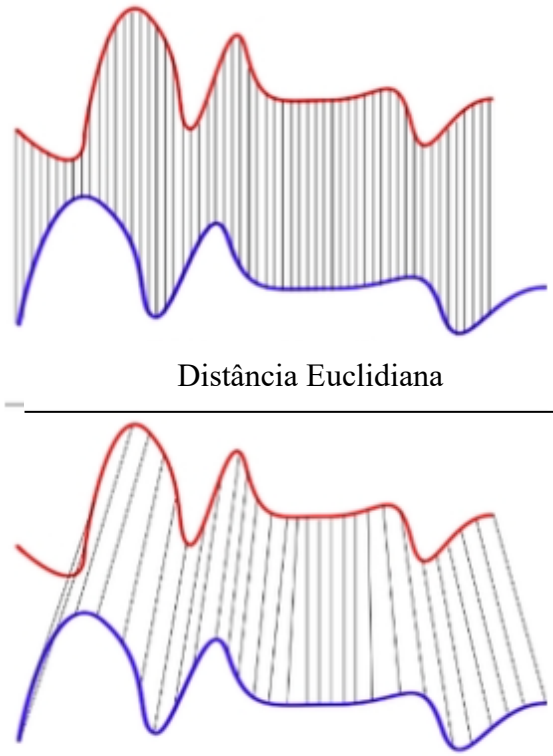
Figura 10 - Resultado do alinhamento de séries temporais



Fonte: Github (THE DTW SUÍTE, 2019).

A Figura 11 mostra a diferença entre o cálculo da métrica de dissimilaridade utilizando a distância euclidiana e o método *Dynamic Time Warping*.

Figura 11 - Comparação entre distância euclidiana e DTW



Distância Euclidiana

Dynamic Time Warping

Fonte: Traduzido de The DTW suíte (2019).

2.4.2 Clusterização hierárquica

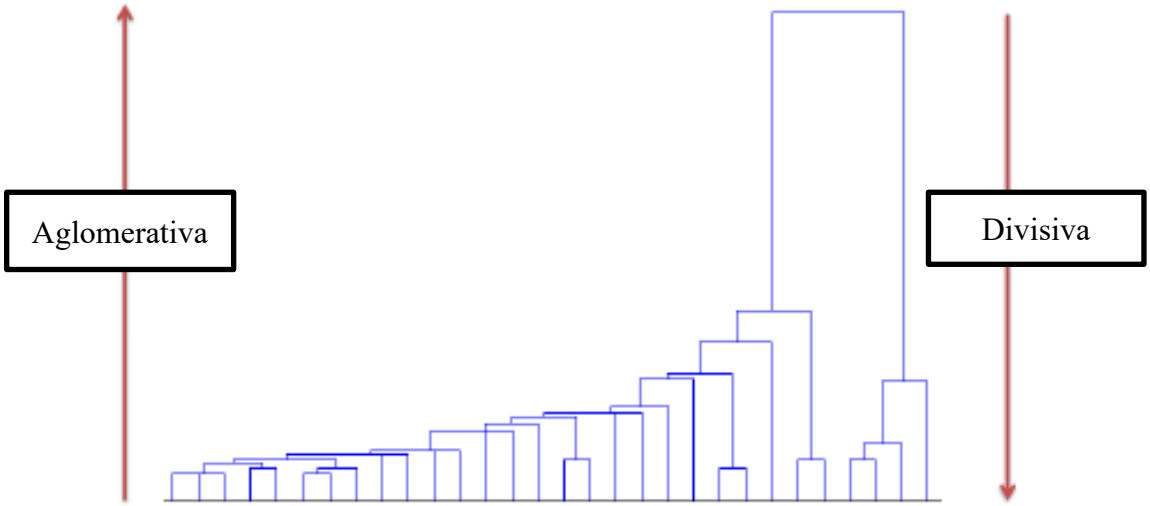
Os algoritmos de clusterização hierárquica agrupam dados em árvores de *clusters*, representados pela forma de árvores binárias ou dendogramas. Podem ser classificados em métodos divisivos, que possuem uma abordagem *top-down* ou métodos aglomerativos, que possuem uma abordagem *bottom-up*.

Na abordagem *top-down* proposta pelos métodos divisivos, a clusterização é iniciada com um único grupo que contem os *n* objetos a serem agrupados e, recursivamente divide o cluster inicial em *n* grupos até que cada objeto pertença à um grupo. Já na abordagem *bottom-up* dos métodos aglomerativos, é realizado o caminho inverso, no qual a clusterização inicia com *n* grupos, cada um com um objeto e recursivamente divide o conjunto de dados em uma estrutura de árvore baseada em operações que calculam a proximidade baseados em métrica de

dissimilaridade, e, por fim, agrupa todos os objetos em um único grupo (SREEDHAR KUMAR *et al.,* 2019).

A Figura 12 mostra a representação de um dendograma e o sentido (*bottom-up* ou *top-down*) proposto por cada classe de método de clusterização.

Figura 12 - Clusterização hierárquica divisiva e aglomerativa



Fonte: Traduzido de Sayad (2020).

A sistemática geral do processo de clusterização hierárquica aglomerativa pode ser descrita pelas etapas:

- Computar a matriz de dissimilaridade;
- Considerar cada objeto um cluster;
- Encontrar o objeto mais próximo com base na medida de dissimilaridade e agrupá-los em um único cluster;
- Atualizar a matriz de dissimilaridade e calcular as distâncias entre o novo cluster e cada um dos clusters antigos;
- Repetir as etapas três e quatro até que todos os clusters sejam agrupados em um único cluster.

A atualização da matriz de dissimilaridade de cálculo das novas distâncias (etapa 4) pode ser realizada por diferentes métodos, sendo os mais comuns:

- Ligação simples (*Sigle-Linkage/Nearest Neighbor*): a distância entre o cluster A e o cluster B é a menor distância de qualquer um dos objetos do cluster A para qualquer um dos objetos do cluster B. Matematicamente pode ser representada por:

$$d_s(A, B) = min_{i \in A, jB}\, d_{ij}$$

- Ligação completa (*Complete-Linkage/Furthest Neighbor*): a distância entre o cluster A e o cluster B é a maior distância de qualquer um dos objetos do cluster A para qualquer um dos objetos do cluster B. Matematicamente pode ser representada por:

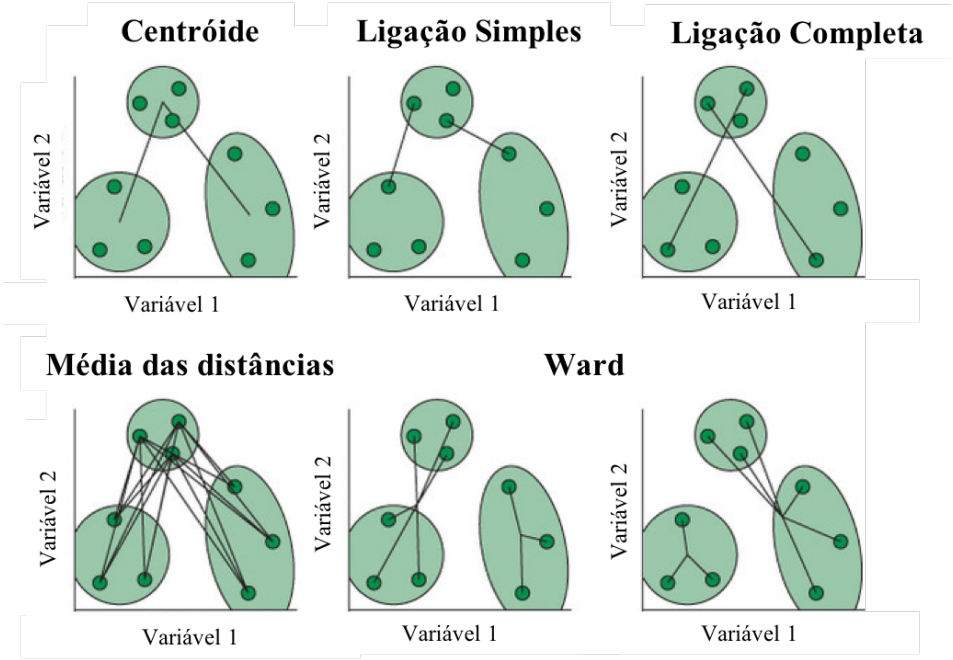
$$d_C(A, B) = max_{i \in A, jB}\, d_{ij}$$

- Média das distâncias: a distância entre o cluster A e o cluster B é a média de todas as distâncias em cada objeto do cluster A para cada objeto do cluster B. Matematicamente pode ser representada por:

$$d_A(A, B) = \frac{1}{N_A N_B} \sum_{i \in A} \quad \sum_{i \in B} d_{ij}$$

- Centroide: a distância entre o cluster A e o cluster B é a distância do centroide de A para o centroide de B (BARBAKH; WU; FYFE, 2009).
- Ward: o método de Ward se difere dos demais métodos por não medir a distância dos clusters de forma única, neste método, as combinações entre clusters são feitas de modo que a soma dos quadrados das distâncias entre os objetos dentro de um agrupamento seja minimizada (HAIR JÚNIOR *et al*., 2009).

Figura 13 - Métodos de ligação de clusters



Fonte: Traduzido de Rhys (2020), Cap. 17.

3 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo são apresentados as metodologias e os procedimentos utilizados para o desenvolvimento do trabalho.

3.1 CARACTERIZAÇÃO DA PESQUISA

Gil (1994) propõe a classificação de um trabalho de pesquisa segundo quatro pontos de vista. São eles: do ponto de vista de sua natureza, do ponto de vista da forma de abordagem ao problema, do ponto de vista dos propósitos e do ponto de vista dos procedimentos técnicos.

Com base na proposta de autor, do ponto de vista da forma de abordagem, trata-se de uma pesquisa quantitativa pois baseia-se na utilização de dados, estatísticas e métricas de avaliações numéricas.

Sob o ponto de vista da natureza, classifica-se este trabalho como uma pesquisa aplicada, visto que o conhecimento discutido tem por objetivo solucionar um problema específico de uma empresa.

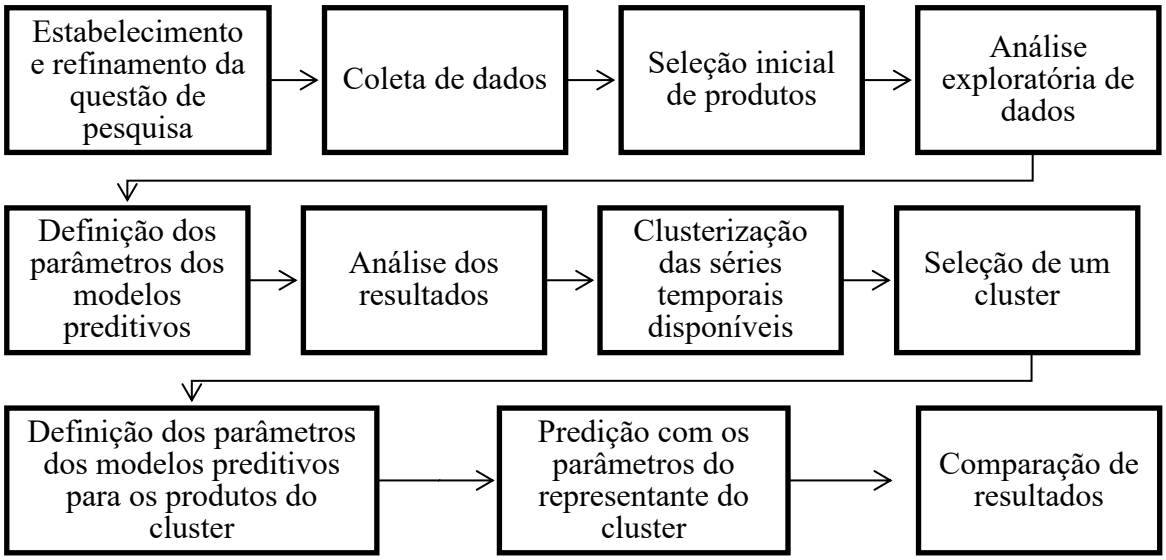
Quanto ao ponto de vista dos propósitos, classifica-se como um trabalho descritivo uma vez que se busca descrever as características do fenômeno estudado.

Sob o ponto de vista dos procedimentos técnicos, trata-se de um estudo de caso, pois realiza o estudo exaustivo do objeto em questão.

3.2 ROTEIRO METODOLÓGICO

O presente trabalho, que tem por objetivo propor modelos de previsão de demanda para um comércio varejista, é caracterizado como um estudo de caso. As etapas que conduziram à realização do trabalho são apresentadas no roteiro metodológico da Figura 4.

Figura 14 - Roteiro metodológico



Fonte: Elaborado pelo autor.

A primeira etapa consiste na elaboração clara e concisa da questão de pesquisa, de forma a delimitar os objetivos e abrangência da pesquisa. Nesta etapa define-se o período de análise e a granularidade dos modelos de previsão.

A partir da elaboração da questão de pesquisa, coleta-se o histórico de venda de todos os produtos no banco de dados da empresa utilizando a linguagem SQL (*Structured Query Language*), representando os dados brutos da pesquisa.

Após a coleta de dados, seleciona-se um subgrupo de produtos para a realização das previsões. De forma a garantir a consolidação dos dados históricos de demanda, realiza-se uma análise exploratória dos dados, na qual são tratados possíveis dados faltantes ou *outliers.* A análise também permite o entendimento de características estatísticas gerais dos dados, viabilizando a visualização e familiarização com a natureza dos dados.

Em seguida, definem-se os parâmetros dos modelos preditivos através de diferentes métodos de parametrização. Para cada produto e cada método de parametrização, analisam-se a métrica de erro resultante das previsões.

Após a modelagem das séries temporais do subgrupo inicial, realiza-se a clusterização das séries temporais não consideradas na seleção inicial de produtos com o uso de algoritmos de clusterização hierárquica. A partir do agrupamento, seleciona-se um cluster e uma série temporal representante do cluster (pertencente à seleção inicial de produtos). Em seguida, definem-se os parâmetros dos modelos de previsão para os produtos do cluster escolhido

utilizando os mesmos métodos empregados na definição dos parâmetros dos modelos dos produtos da seleção inicial. Por fim, comparam-se os resultados da modelagem das séries temporais obtidas por meio dos diferentes métodos.

4 CONTEXTUALIZAÇÃO E DESENVOLVIMENTO DO TRABALHO

O capítulo apresenta a empresa e os itens da amostra inicial de produtos selecionados para a construção dos modelos de previsão. O capítulo apresenta também o desenvolvimento dos modelos de previsão de demanda e abrange as etapas da análise exploratória de dados, parametrização dos modelos, clusterização das séries temporais e análise de generalização dos modelos de previsão. Por fim, são apresentadas considerações finais sobre o capítulo.

4.1 A EMPRESA

A empresa objeto de estudo foi fundada em 1985 na cidade de Porto Velho – RO com objetivo de comercializar mercadorias para as empresas responsáveis pelo asfaltamento da BR 364, que liga Limeira – SP a Rio Branco - AC. Nos anos iniciais do negócio, notou-se que a atividade econômica em destaque na região era a extração de ouro, e que o mercado de artigos para o garimpo caracterizava-se como um mercado atrativo, o que acarretou na mudança do portfólio de produtos da empresa para atender as necessidades do setor em destaque.

Atualmente, o negócio oferece uma vasta gama de produtos e abrange soluções em plásticos, borrachas e acessórios para manutenção de equipamentos de diversos setores da economia, sendo o principal destes, o setor garimpeiro.

4.2 OS PRODUTOS

A definição da amostra de produtos para a análise consistiu na seleção de produtos de famílias de produtos diversas com o objetivo de obter séries temporais com comportamentos variados.

Os itens selecionados foram:

- Item I: bota de segurança Usafe;
- Item II: luva de malha pigmentada Carbografite;
- Item III: eletrodo de carvão para corte Carbografite 1/4x12”;
- Item IV: capa prensada Skive R1/R2;
- Item V: pneu Levorim 400x8;
- Item VI: mangueira industrial água e ar 300 Psi 3/8”;
- Item VII: carpete canelado com borracha.

Os itens listados possuem diversos usos e públicos alvos, de modo que os fatores que influenciam na demanda também são variados.

Os itens I (bota de segurança) e II (luva de malha pigmentada) são equipamentos de segurança e possuem como principal público alvo empresas da construção civil.

Os itens III (eletrodo de carvão), IV (capa prensada) e VII (carpete canelado) são destinados majoritariamente ao garimpo. O item III é utilizado na manutenção geral de itens de aço das dragas. O item IV possui uso na parte hidráulica das dragas, responsável pela extração de material do fundo do rio. O item VII é utilizado para forrar as calhas e bicas das dragas, onde separa-se o ouro.

O item V (pneu) é utilizado na indústria automotiva como componente de carretas para motos e, na indústria de cerâmica, como componente de carrinhos de transporte industriais.

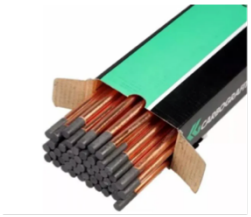
O item VI (mangueira industrial água e ar) é utilizado em compressores e possui como principal público alvo postos de gasolina e borracheiros.

A Figura 15 apresenta as imagens dos produtos selecionados para análise.

Figura 15 - Produtos selecionados para análise

Item I – Bota de segurança

Item II – Luva de malha pigmentada



Item III – Eletrodo de carvão para corte

Item IV – Capa prensada

Item V – Pneu

Item VI – Mangueira Industrial

Item VII - Carpete

Fonte: Empresa objeto de estudo, 2020.

4.3 ANÁLISE EXPLORATÓRIA DE DADOS

A análise exploratória de dados consiste na manipulação e visualização inicial dos dados, com o objetivo de explorar suas principais propriedades. Nesta etapa, utilizam-se métodos visuais e estatísticos que permitem a identificação de padrões a cerca da estrutura dos dados e relacionamento entre variáveis.

A leitura e manipulação dos dados foi realizada através da biblioteca *Pandas* da linguagem *Python*, enquanto para a visualização dos dados, utilizou-se a biblioteca *Matplotlib*.

O conjunto de dados selecionado é composto por 72 observações mensais, compreendendo o período de janeiro de 2014 a dezembro de 2019. As estatísticas gerais dos produtos são sintetizadas na Tabela 1.

Tabela 1 - Estatísticas gerais dos produtos.

Item	Média	Desvio Padrão	Coeficiente de Variação	Min	25%	50%	75%	Máx
I	86,4	55,7	0,64	3,0	48,5	80,0	109,5	355,0
II	110,8	52,9	0,47	0,0	77,0	112,0	135,3	261,0
III	2149,7	1123,9	0,52	150,0	1401,5	1905,0	2786,0	6700,0
IV	47,8	26,3	0,55	0,0	31,7	45,0	60,5	121,0
V	25,9	11,6	0,44	0,0	18	26,0	32,3	49,0
VI	51,0	41,65	0,81	1,0	20,9	39,9	67,9	218,1
VII	126,6	65,9	0,52	16,05	84,9	113,8	154,0	381,5

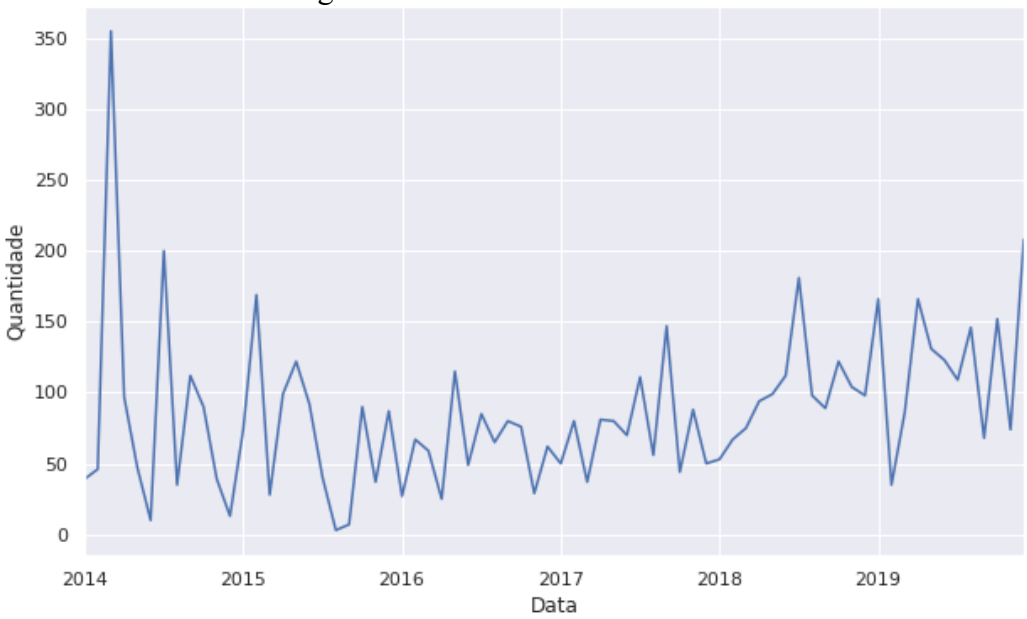
Fonte: Elaborado pelo autor.

Os itens I e VI são caracterizados por possuírem alto coeficiente de variação em relação aos demais itens, enquanto o item V possui o menor coeficiente de variação.

O comportamento das séries temporais pode ser caracterizado por meio da inspeção visual dos dados. As demandas dos itens I e VII são caracterizadas por possuírem leve tendência positiva de crescimento ao longo do período considerado enquanto as demandas dos itens III e V possuem tendência negativa. As séries temporais dos itens II, IV e VI não mostraram tendências no período considerado.

A série temporal do item I, além de leve tendência, possui forte variação de valores nos meses iniciais do período considerado, como observa-se na Figura 16.

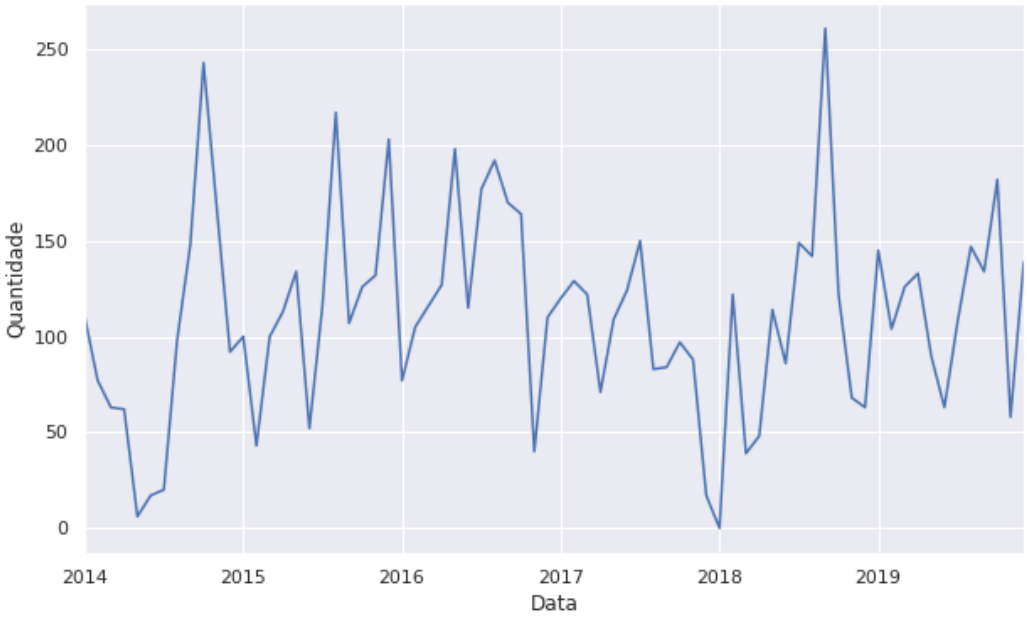
Figura 16 - Demanda mensal do item I



Fonte: Elaborado pelo autor.

A série temporal do item II, que não apresenta comportamento de tendência, é caracterizada por possuir forte variação entre meses consecutivos, como observa-se na Figura 17.

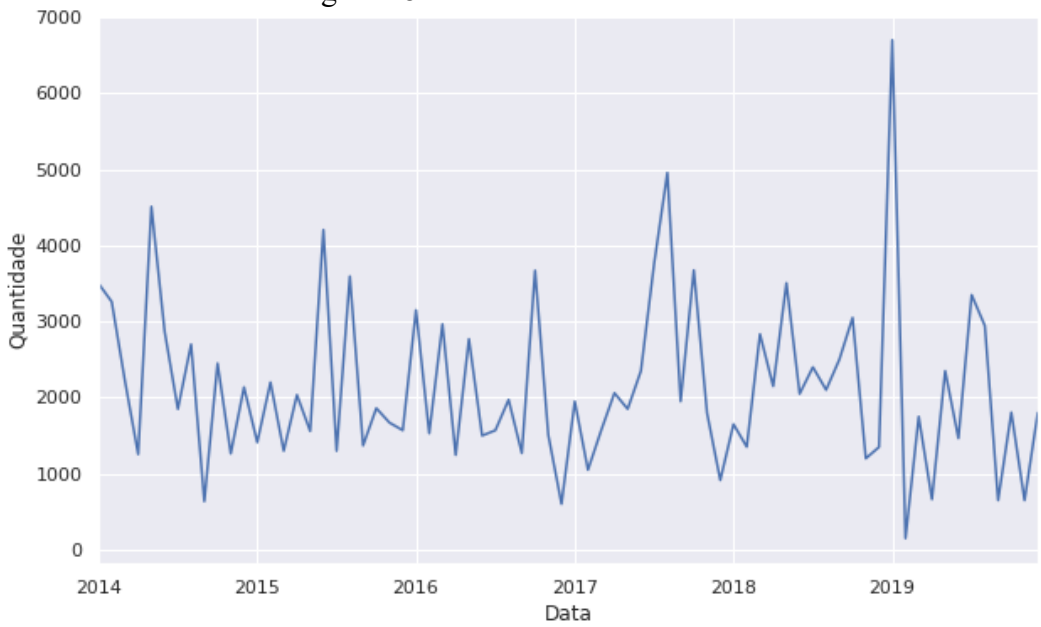
Figura 17 - Demanda mensal do item II



Fonte: Elaborado pelo autor.

A série temporal do item III, que possui a maior amplitude entre as séries consideradas, apresenta um pico de demanda no final do ano de 2018 como pode ser observado na Figura 18.

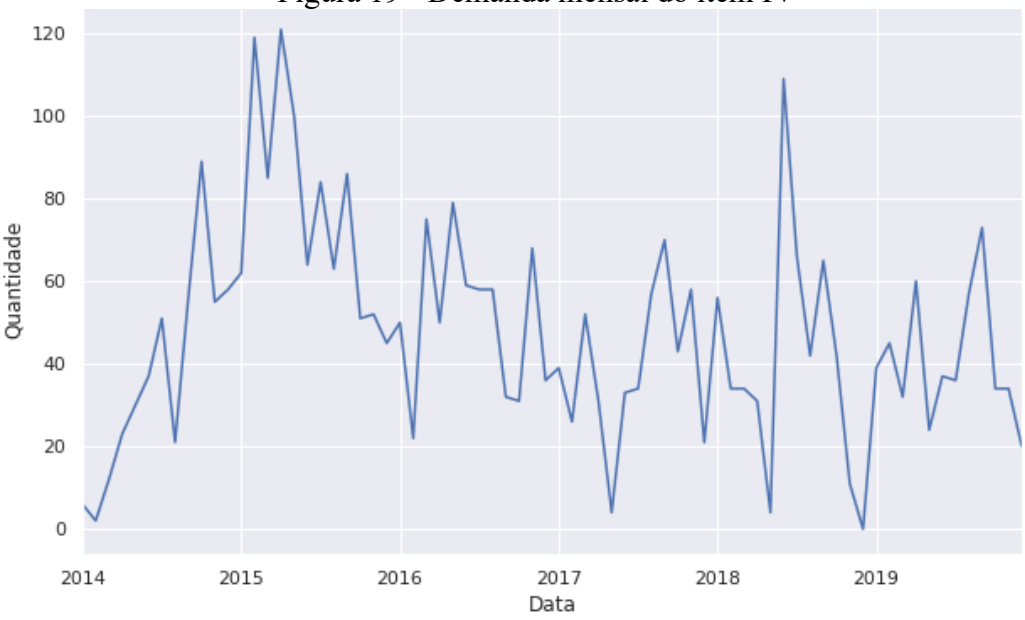
Figura 18 - Demanda mensal do item III



Fonte: Elaborado pelo autor.

A série temporal do item IV, que não apresenta tendência de crescimento, possui quatro vales, nos quais a demanda é nula ou se aproxima de 0. A série pode ser observada na Figura 19.

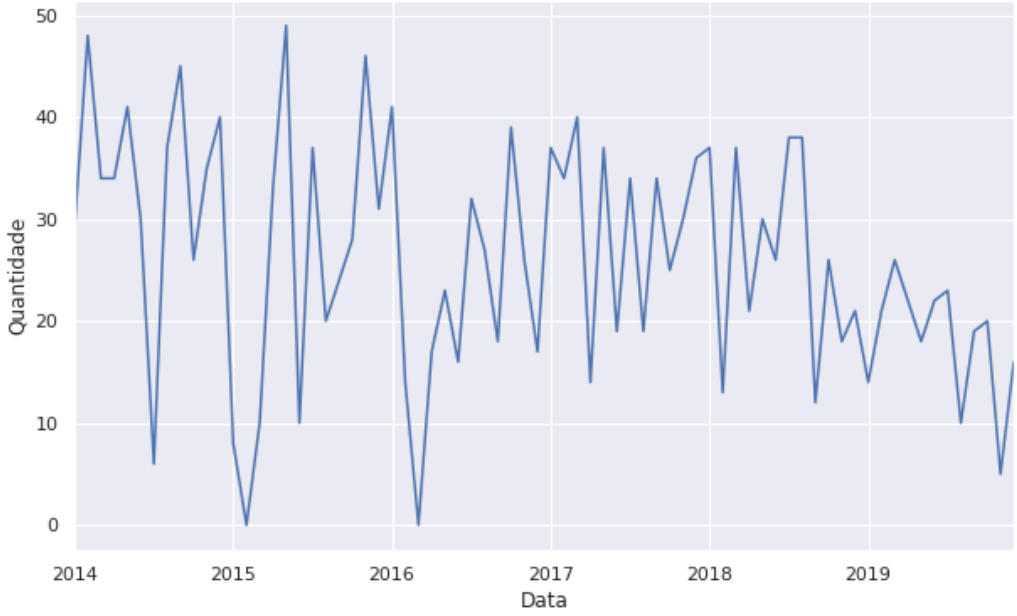
Figura 19 - Demanda mensal do item IV



Fonte: Elaborado pelo autor.

A série temporal do item V, que possui tendência negativa de crescimento, também apresenta vales sendo estes concentrados nos meses da primeira metade do período considerado, como observa-se na Figura 20.

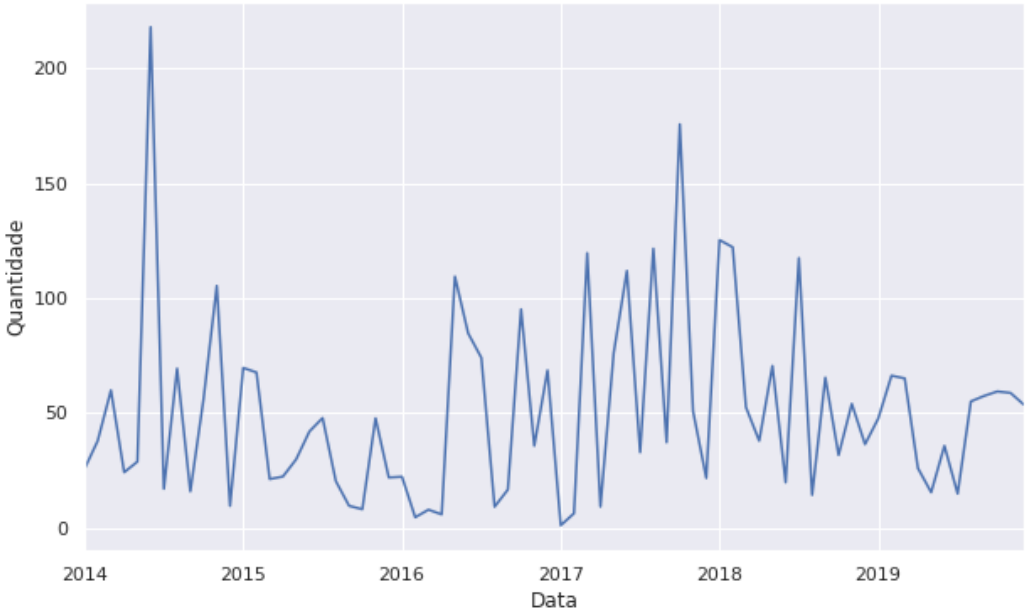
Figura 20 - Demanda mensal do item V



Fonte: Elaborado pelo autor.

A série temporal do item VI, que não possui tendência, apresenta um pico no início do período e fortes variações entre meses consecutivos entre o ano de 2017 e 2018. A série temporal pode ser observada na Figura 21.

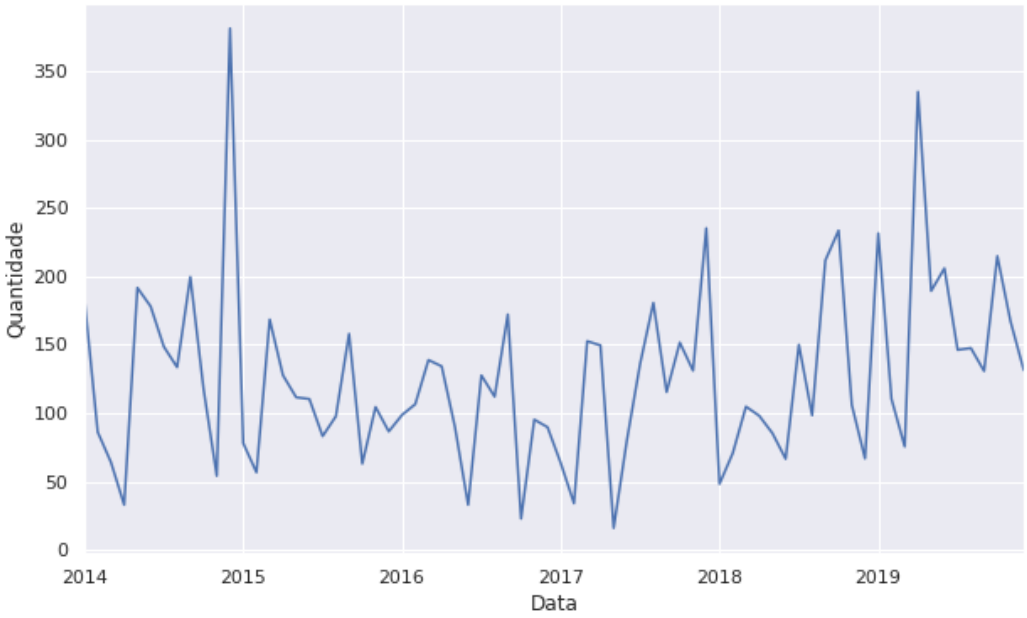
Figura 21 - Demanda mensal do item VI



Fonte: Elaborado pelo autor.

A série temporal do item VII, caracterizada pela leve tendência de crescimento, possui seu maior pico no primeiro ano do período considerado, como pode ser observado na Figura 22.

Figura 22 - Demanda mensal do item VII



Fonte: Elaborado pelo autor.

A inspeção visual permitiu identificação das estruturas gerais das séries temporais, como o comportamento da tendência e a presença de picos e vales nas demandas dos itens analisados.

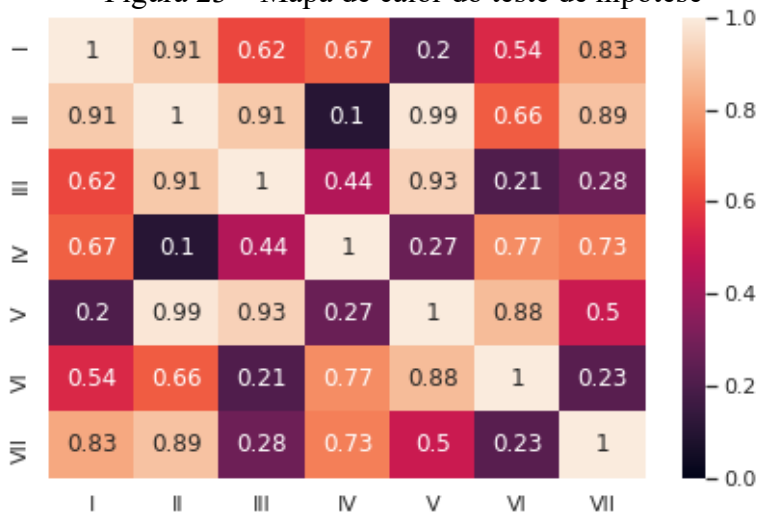
Para avaliar o grau de relacionamento entre séries temporais, utilizou-se a correlação de Pearson, na qual as séries temporais são comparadas em pares e avalia-se a correlação entre os valores das demandas mensais.

O coeficiente de correlação de Pearson, resultado da análise de correlação, varia entre -1 e 1 e quanto mais próximo dos extremos da escala maior é o grau de relacionamento entre as variáveis consideradas. Valores próximos a 0 indicam a ausência de relacionamento linear (SAMOHYL, 2009).

A verificação do grau de relacionamento entre as variáveis é avaliada a partir do teste de hipótese, no qual o valor *p-value* é comparado com o nível de significância de 0,05. Valores de *p-value* abaixo do nível de significância definido indicam a existência de correlação estatística significativa.

Os resultados da análise podem ser representados na forma de um mapa de calor, mostrado na Figura 23.

Figura 23 – Mapa de calor do teste de hipótese



Fonte: Elaborado pelo autor.

Os valores *p-value* calculados estão acima do nível de significância definido o que permite afirmar que as séries não possuem relação linear significativa.

4.3.1 Teste de estacionariedade

A análise de estacionariedade das séries temporais é utilizada na análise exploratória de dados como ferramenta para a compreensão da estrutura dos dados e permite identificar se propriedades como média, variância e covariância são constantes ao longo do tempo.

O teste de Dickey-Fuller Aumentado é um teste de hipótese e possui o nível de significância como parâmetro do teste. A hipótese nula, caso não seja rejeitada, sugere que a série temporal não é estacionária.

Utilizou-se o teste Dickey-Fuller Aumentado com nível de significância de 1%, ou seja, considerou-se não estacionária a série cujo valor *p-value* fosse menor que 0,01. O teste foi aplicado nas séries originais e em suas respectivas séries diferenciadas. Os resultados obtidos estão sintetizados na Tabela 2.

Tabela 2 - Teste Dickey-Fuller Aumentado

Item	Produto	p-value	Classificação	p-value diff	Classificação
I	825	2,84 E-12	Estacionária	4,58 E-19	Estacionária
II	794	9,35 E-16	Estacionária	4,55 E-10	Estacionária
III	643	3,61 E-15	Estacionária	1,03 E-7	Estacionária
IV	28802	2,10 E-5	Estacionária	2,81 E-5	Estacionária
V	9089	2,45 E-10	Estacionária	3,09 E-7	Estacionária
VI	16100	6,65 E-14	Estacionária	4,59 E-14	Estacionária
VII	3636	3,23 E-12	Estacionária	2,79 E-16	Estacionária

Fonte: Elaborado pelo autor.

Os resultados dos testes indicam que as séries temporais são estacionárias, isto é, possuem média, variância e covariância constantes ao longo do tempo.

A estacionariedade das séries temporais também possui implicação na parametrização dos modelos de previsão da família SARIMA pois indicam que não é necessário o processo de diferenciação da série para torna-la estacionária.

4.3.2 Decomposição das séries temporais

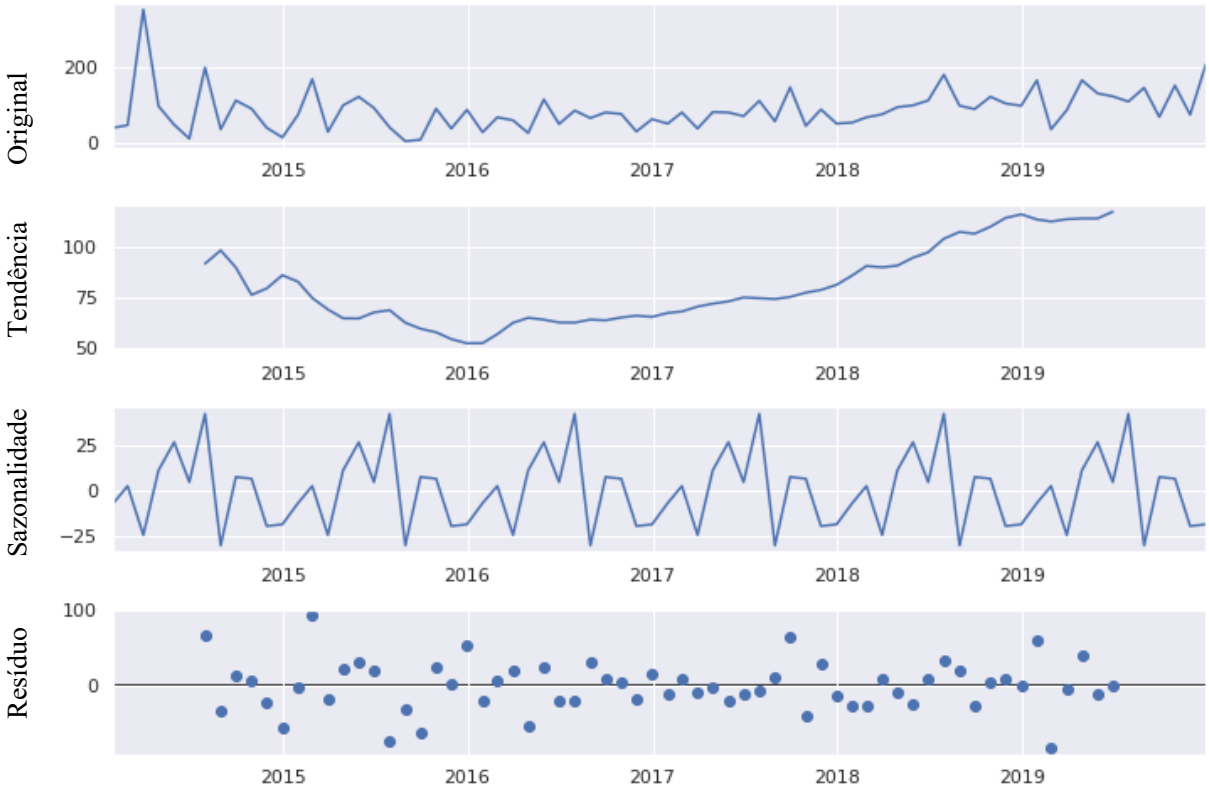
A decomposição de séries temporais auxilia na compreensão da estrutura dos dados de uma série temporal e podem ser decompostas em componentes de tendência, sazonalidade e resíduo.

O modelo de decomposição de séries temporais pode ser aditivo, no qual o valor da série é dado pela soma dos componentes, ou multiplicativo, no qual o valor da série é dado pela multiplicação da componente de tendência e sazonalidade e posterior soma do resíduo.

A decomposição das séries temporais foi realizada utilizando o método aditivo, escolha justificada pelo comportamento padrão limitado do crescimento dos valores das séries.

A decomposição da série temporal do item I mostra que as principais componentes de maior amplitude da série são a tendência e o resíduo, enquanto a sazonalidade não possui grande expressividade. A componente de tendência é negativa no primeiro terço do período e positiva nos demais meses, como mostra a Figura 24.

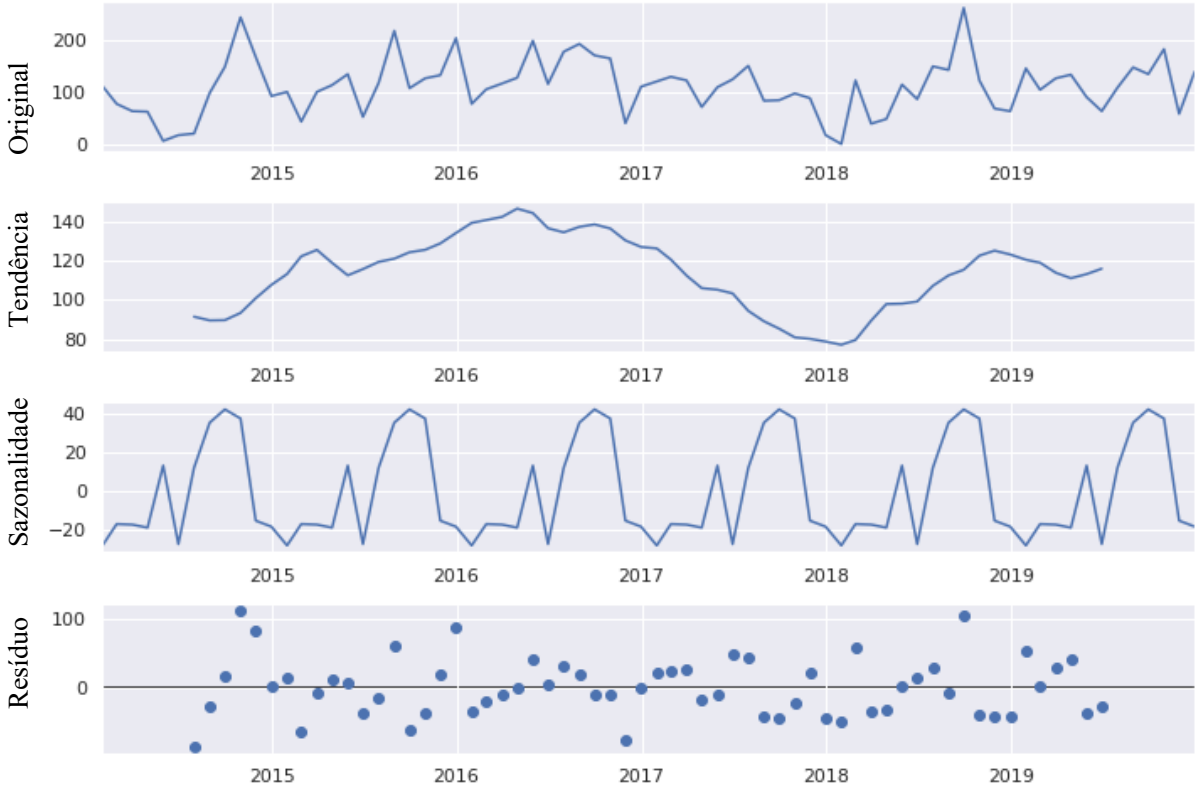
Figura 24 - Decomposição da série temporal do item I



Fonte: Elaborado pelo autor.

A Figura 25 mostra que a série temporal do item II possui a tendência como componente de maior amplitude. A componente de tendência é positiva no primeiro terço do período, negativa no segundo terço e positiva no terceiro terço.

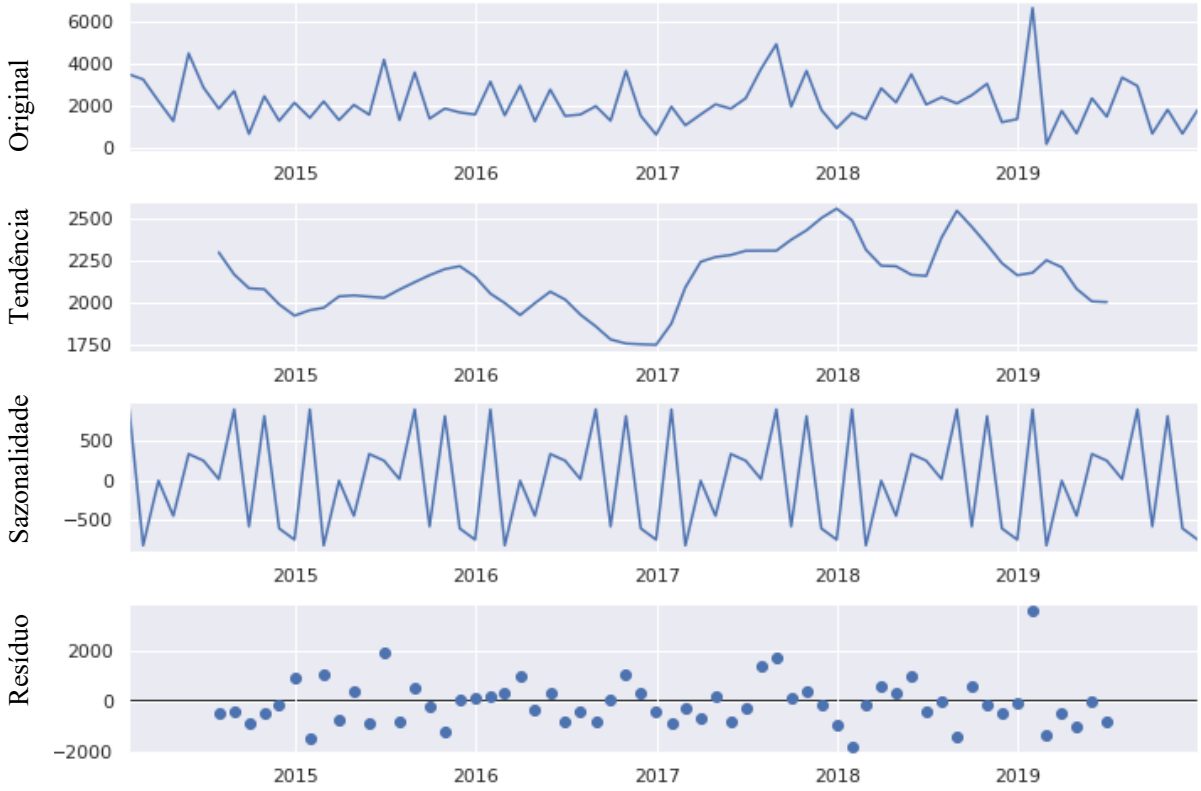
Figura 25 - Decomposição da série temporal do item II



Fonte: Elaborado pelo autor.

De forma similar aos itens anteriores, a decomposição da série temporal do item III indica que as componentes de maior amplitude são as componentes de tendência e resíduos. A Figura 26 mostra que a série não possui tendência de crescimento estável no primeiro terço do período, tendência positiva no segundo terço e negativa no terceiro terço do período.

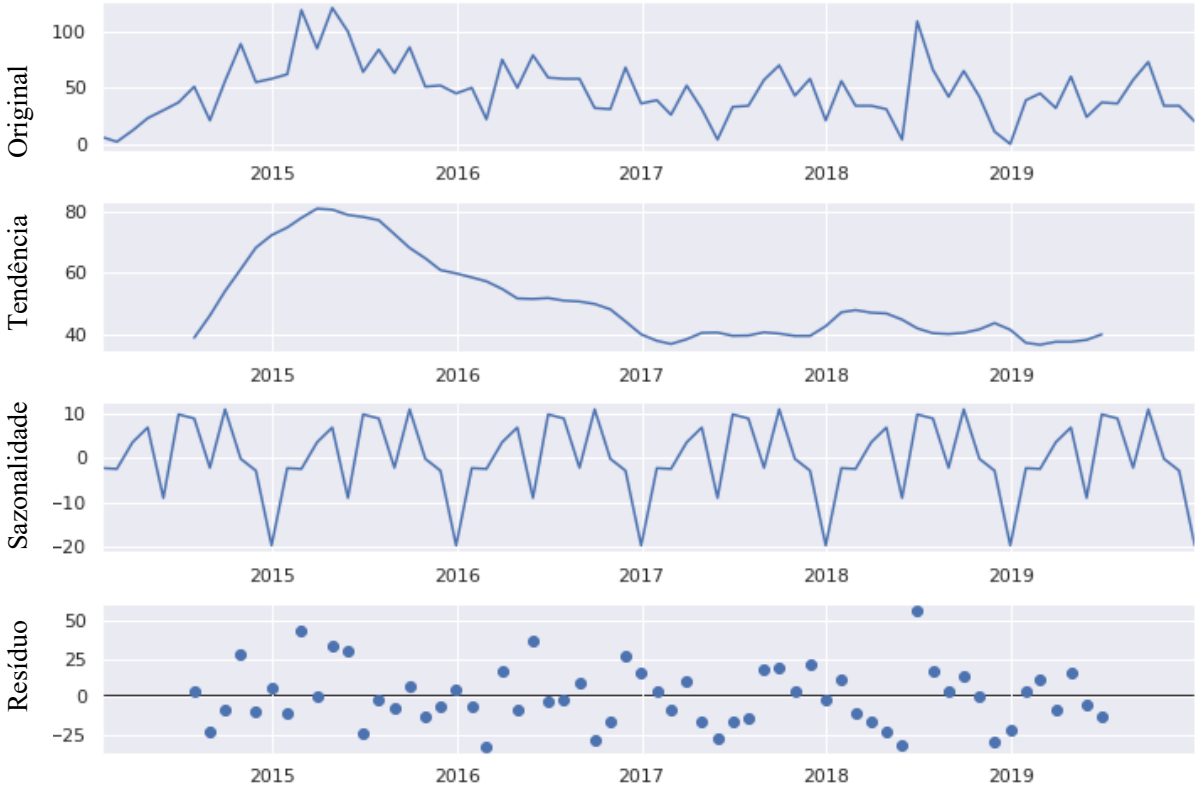
Figura 26 - Decomposição da série temporal do item III



Fonte: Elaborado pelo autor.

A decomposição da série do item IV, ilustrada na Figura 27, indica que a série possui a tendência como componente de maior amplitude. A componente de tendência varia entre positivo e negativo no primeiro terço do período e é negativa nos demais meses.

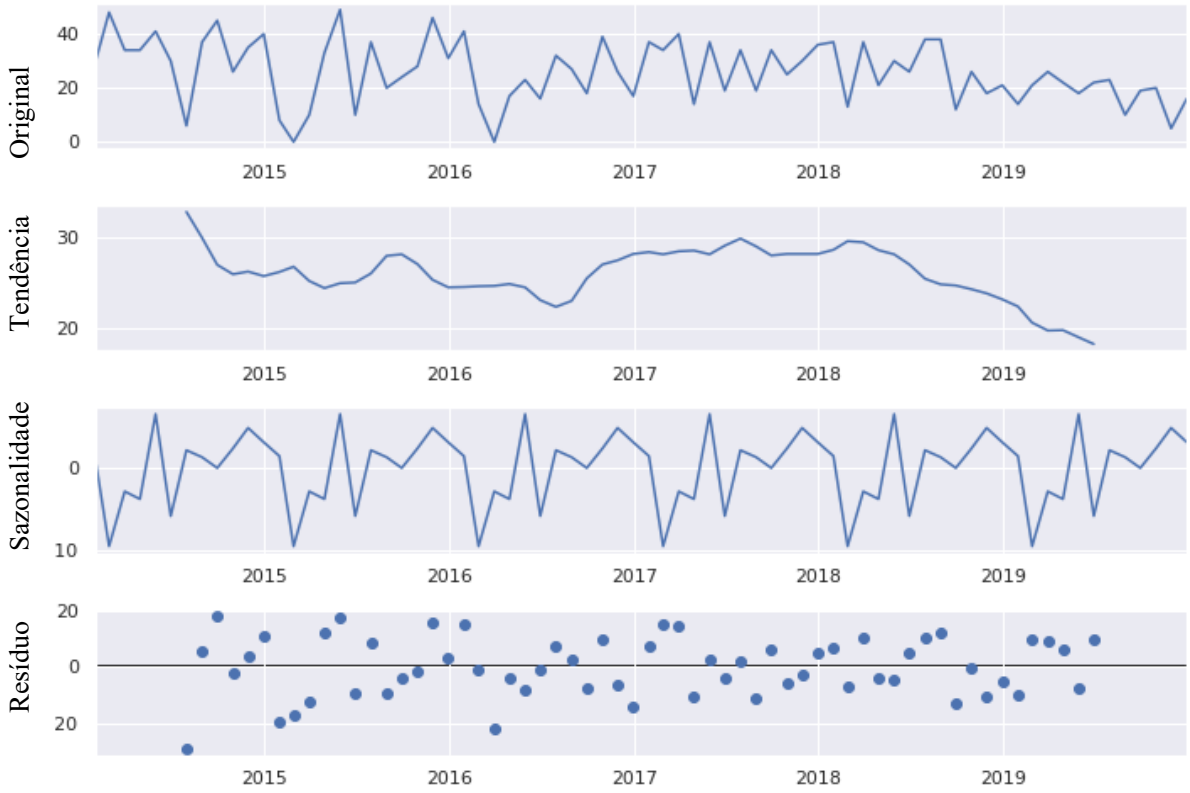
Figura 27 - Decomposição da série temporal do item IV



Fonte: Elaborado pelo autor.

O produto V também possui a componente de tendência como componente de maior amplitude. Nos dois primeiros terços do período é estável e no último é negativa, como mostra a Figura 28.

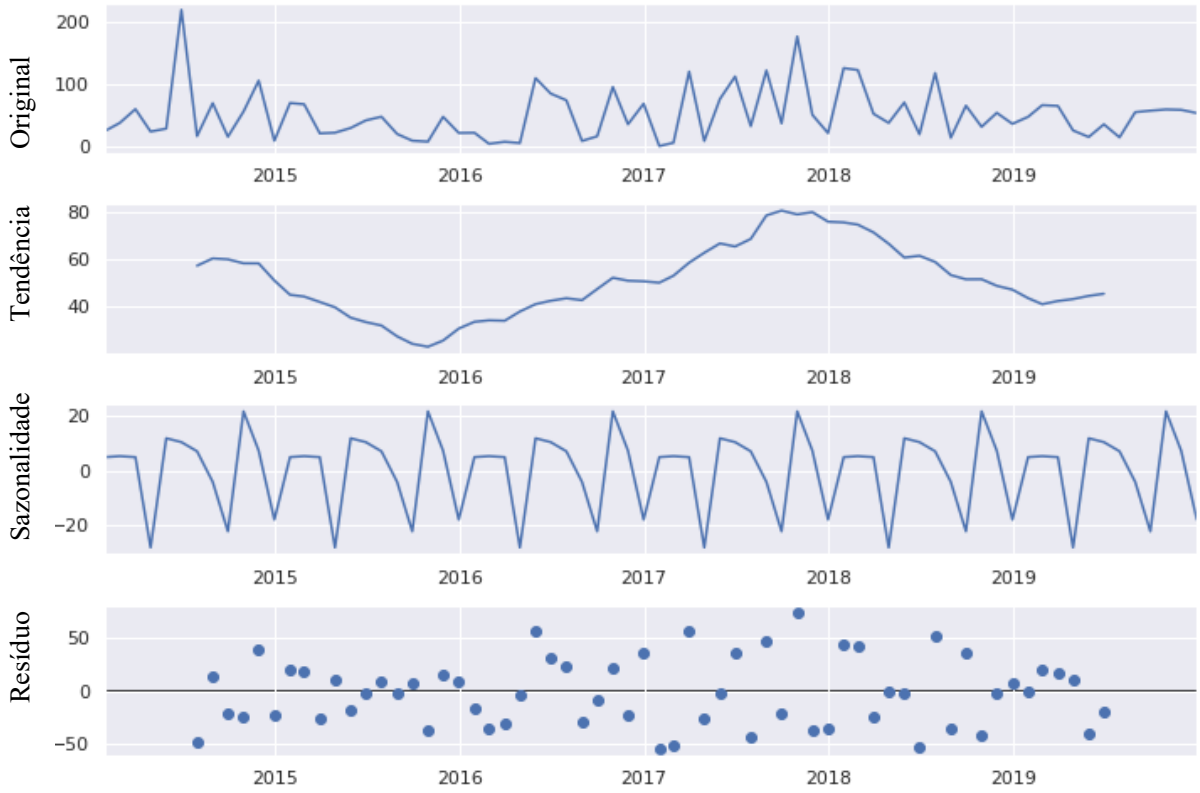
Figura 28 - Decomposição da série temporal do item V



Fonte: A Elaborado pelo autor.

O item VI possui um padrão de tendência similar ao padrão do item II, no qual o componente varia entre períodos de positividade e negatividade ao longo do período. No primeiro e terceiro terços do período, a componente é negativa, enquanto no segundo terço, positiva. A decomposição da série temporal do item VI é representada na Figura 29.

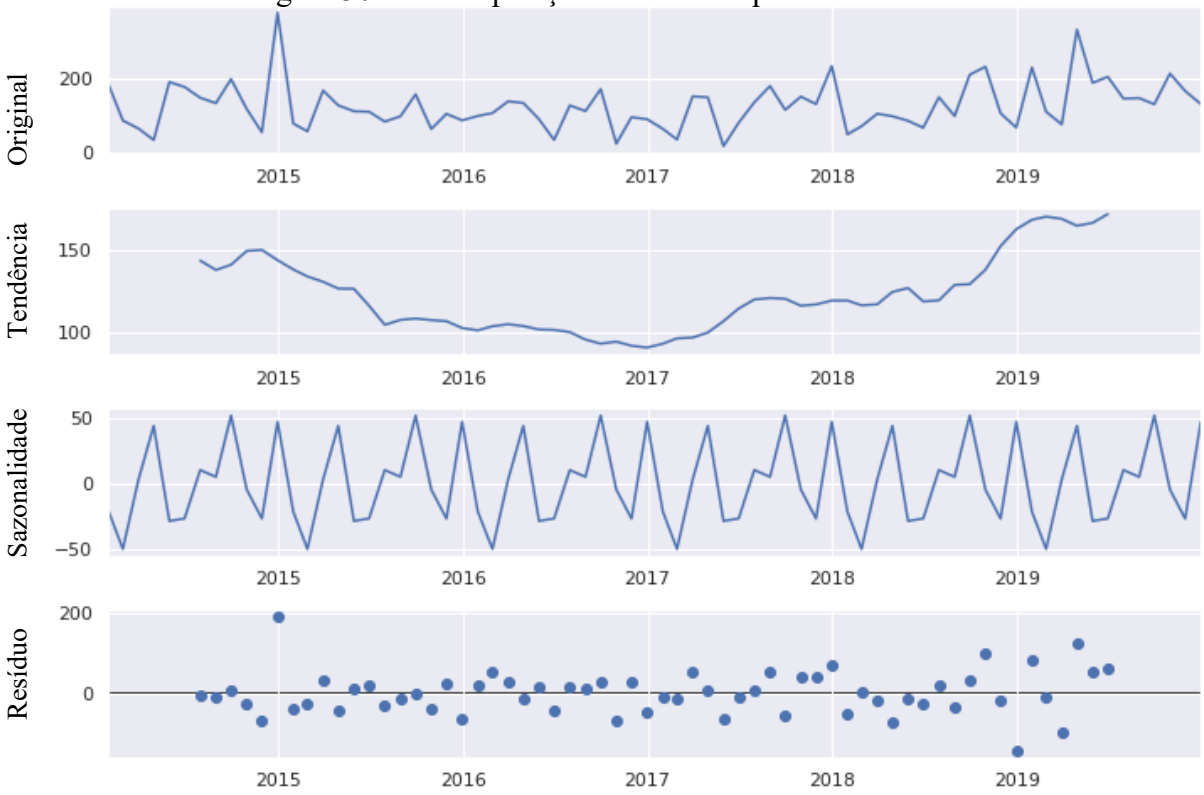
Figura 29 - Decomposição da série temporal do item VI



Fonte: Elaborado pelo autor.

A decomposição da série temporal do item VII, representada na Figura 30, indica a componente de resíduo como componente de maior amplitude. A tendência é negativa no primeiro terço do período, estável no segundo terço e positiva no terceiro terço do período.

Figura 30 - Decomposição da série temporal do item VII



Fonte: Elaborado pelo autor.

As tendências gerais das séries estão sintetizadas na Tabela 3.

Tabela 3 - Comportamento da tendência por períodos

Item	1º terço	2º terço	3º terço
I	Negativa	Positiva	Positiva
II	Positiva	Negativa	Positiva
III	Estável	Positiva	Negativa
IV	Positiva/Negativa	Negativa	Negativa
V	Estável	Estável	Negativa
VI	Negativa	Positiva	Negativa
VII	Negativa	Estável	Positiva

Fonte: Elaborado pelo autor.

4.4 ESTIMATIVA DE PARÂMETROS DOS MODELOS

A estimativa de parâmetros dos modelos consiste em testar combinações de parâmetros da família de modelos escolhida e avaliá-los com base métricas de erro.

O modelo SARIMA, representado por SARIMA(p, d, q)×(P, D, Q, s) possui os parâmetros p e P sazonal, que indicam o número de termos auto regressivos, ou seja, a quantia de *lags* da série estacionária; d e D sazonal, que representam o número de diferenciações feitas para tornar a série estacionária; q e Q sazonal, representantes do número de termos da média móvel, e s, a magnitude da sazonalidade dos dados.

Os parâmetros podem ser estimados através de diferentes métodos, sendo os principais a análise de gráficos de auto correlação e auto correlação parcial; método *Grid Search* ou utilizando métodos de otimização como o disponível na biblioteca *pyramid.arima* em *Python*.

Neste trabalho, optou-se por realizar a estimativa através dos métodos *Grid Search* e de otimização pois estes incluem, além dos valores que seriam estimados por meio da análise gráfica, outras combinações de parâmetros, o que permite uma análise mais abrangente.

4.4.1 *Auto Arima*

A função *Auto Arima* busca identificar o conjunto de parâmetros ótimo para um modelo ARIMA. O processo de busca se inicia pela condução de testes de diferenciação (ex.: Kwiatkowski-Phillips-Schmidt-Shin, Dickey-Fuller Aumentado ou Phillips Perron) para determinar a ordem de diferenciação, ou seja, o parâmetro “d” do modelo. Em seguida a função testa valores de “p” e “q” dentro de limites definidos *a priori*, calculando o AIC para cada combinação.

A função oferece a possibilidade de extensão do modelo ARIMA para o modelo SARIMA, ou seja, inclusão dos parâmetros P, Q e D sazonais. Neste caso, a função otimiza o processo de busca conduzindo o teste de raiz unitária sazonal denominado Canova-Hansen.

Os resultados da parametrização dos modelos com base na função *Auto Arima* estão sintetizados na Tabela 4.

Tabela 4 - Parâmetros dos modelos obtidos via *Auto Arima*

Item	(p, d, q)	(P, D, Q, s)
I	(0, 1, 2)	(0, 0, 0, 0)
II	(1, 0, 0)	(0, 0, 0, 0)
III	(1, 0, 0)	(0, 0, 0, 0)
IV	(0, 1, 1)	(0, 0, 0, 0)
V	(1, 0, 0)	(0, 0, 0, 0)
VI	(0, 1, 1)	(0, 0, 0, 0)
VII	(0, 1, 1)	(0, 0, 0, 0)

Fonte: Elaborado pelo autor.

4.4.2 *Grid Search*

O método *Grid Search*, também denominado como varredura de parâmetros, consiste na busca exaustiva de um subconjunto do espaço de hiperparâmetros que minimize uma determinada métrica de erro.

O método é iniciado com a definição da grade de parâmetros a ser considerada, na qual são definidos os intervalos nos quais os parâmetros serão avaliados. Em seguida, são feitas as combinações, testados os modelos e calculados seus respectivos erros. Escolhe-se então a combinação que minimize o erro da métrica considerada.

Os resultados da parametrização por meio da pesquisa estão sintetizados na Tabela 5.

Tabela 5 - Parâmetros dos modelos obtidos via *Grid Search*

Item	(p, d, q)	(P, D, Q, s)
I	(2, 0, 1)	(2, 0, 0, 12)
II	(0, 0, 0)	(2, 1, 1, 12)
III	(1, 0, 0)	(2, 1, 2, 12)
IV	(0, 0, 1)	(1, 0, 1, 12)
V	(2, 1, 0)	(0, 0, 0, 12)
VI	(0, 1, 0)	(1, 0, 1, 12)
VII	(1, 0, 1)	(1, 0, 0, 12)

Fonte: Elaborado pelo autor.

4.5 GENERALIZAÇÃO DOS MODELOS

A generalização dos modelos tem por objetivo simplificar o processo de parametrização. A abordagem de generalização proposta consistiu em clusterizar as séries temporais e estimar os parâmetros para uma série representante do cluster de modo que os parâmetros da série representante pudessem ser generalizados para as demais séries do cluster.

Nesta etapa, foi selecionado um subconjunto de 272 itens vendidos pela empresa. Na seleção, escolheram-se itens cuja demanda fosse diferente de 0 em no mínimo 95% das observações mensais. O critério adotado justifica-se por: alta recorrência de quebra de estoque, o que prejudica a análise de demanda real e a inclusão ou descontinuação da venda de itens no portfólio da empresa.

4.5.1 Clusterização das séries temporais

A clusterização de séries temporais consiste em agrupar séries com base em uma métrica de dissimilaridade, que representa o grau de similaridade das formas entre séries temporais.

O cálculo da métrica de dissimilaridade foi realizado por meio da aplicação do algoritmo *Dynamic Time Warping,* disponível no pacote *dtw-python,* nos dados das demandas mensais históricas dos itens. O algoritmo calcula a métrica para pares de séries, o que resulta na matriz de dissimilaridade. A matriz de dissimilaridade em sua forma truncada pode ser observada na Figura 31.

Figura 31 - Matriz de dissimilaridade

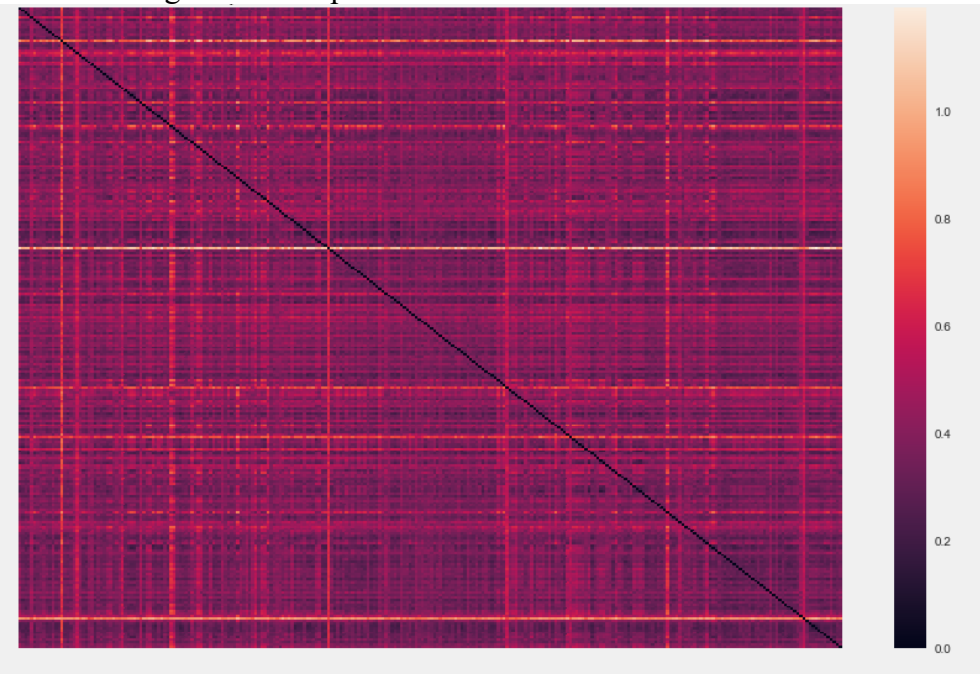
	10037	1015	1017	1020	10280	...	974	975	976	978	991
10037	0.000000	0.277471	0.315104	0.284127	0.487277	...	0.361420	0.307213	0.292363	0.378602	0.378187
1015	0.388403	0.000000	0.343415	0.303712	0.479740	...	0.397634	0.411058	0.356338	0.333851	0.356241
1017	0.380476	0.315580	0.000000	0.273397	0.430770	...	0.304702	0.353841	0.304077	0.292227	0.338124
1020	0.346620	0.265626	0.283859	0.000000	0.430848	...	0.350250	0.339451	0.312094	0.288362	0.356114
10280	0.677575	0.600799	0.569376	0.564018	0.000000	...	0.601153	0.627683	0.563820	0.532892	0.506200
...	...	...	...	...	...	...	...	...	...	...	...
974	0.407031	0.309545	0.258229	0.332981	0.447811	...	0.000000	0.311349	0.224382	0.322064	0.298358
975	0.357335	0.364397	0.321164	0.350507	0.492214	...	0.344517	0.000000	0.316888	0.401807	0.406508
976	0.333871	0.295402	0.286105	0.317320	0.440686	...	0.238912	0.325613	0.000000	0.314058	0.313174
978	0.477016	0.376250	0.321733	0.339106	0.413722	...	0.393553	0.449361	0.379435	0.000000	0.391620
991	0.464530	0.398695	0.377027	0.383668	0.424382	...	0.396135	0.428685	0.378927	0.365437	0.000000

Fonte: Elaborado pelo autor.

A matriz de dissimilaridade pode ser visualizada na forma de mapa de calor no qual cores mais escuras indicam maior grau de similaridade e cores mais claras indicam menor grau de similaridade.

A diagonal da matriz é em tom mais escuro que indica a comparação de cada série temporal com ela mesma, resultando em distância 0. As linhas do mapa que são mais claras em toda sua extensão indicam a presença de séries temporais que não possuem forte similaridade com as demais séries consideradas.

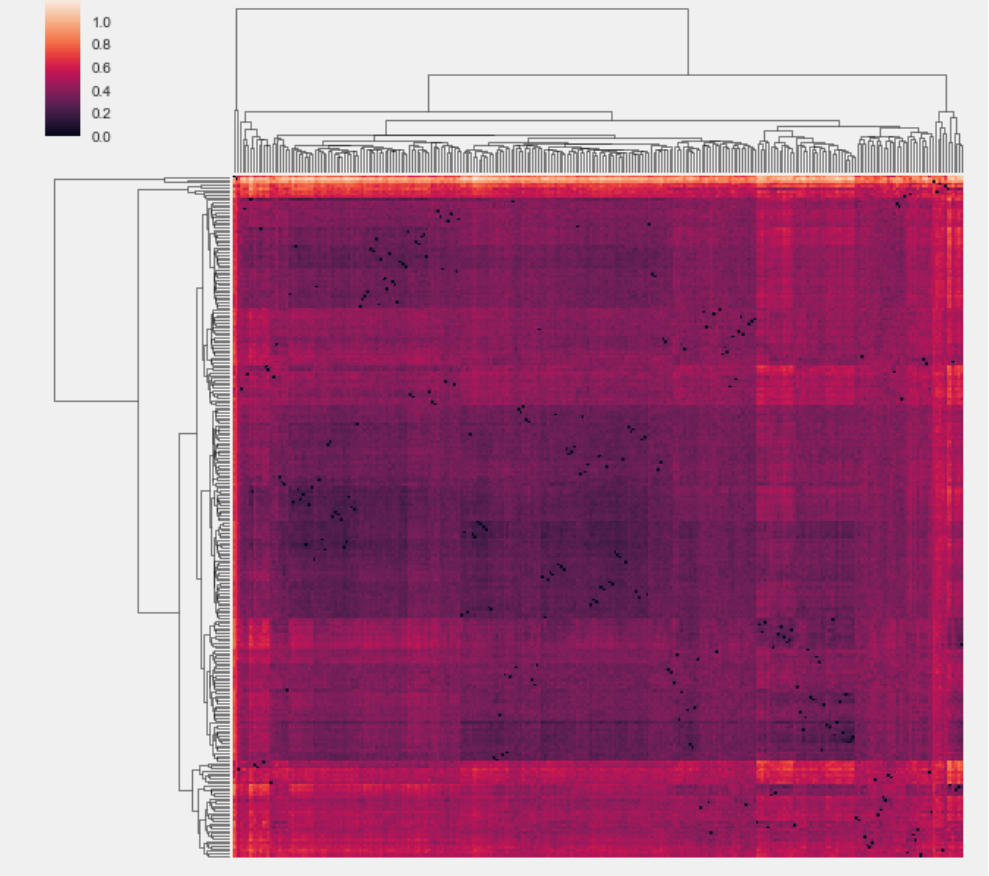
Figura 32 - Mapa de calor da matriz de dissimilaridade



Fonte: Elaborado pelo autor.

Uma terceira visualização foi elaborada através do método *.clustermap()* disponível na biblioteca *seaborn* com implementação em *Python*. Nesta visualização, as séries temporais são organizadas numa matriz de forma que as séries com maior índice de dissimilaridade sejam dispostas próximas umas das outras no mapa. A Figura 33 mostra o mapa em conjunto com o dendograma da clusterização hierárquica.

Figura 33 - Mapa de clusters da matriz de dissimilaridade



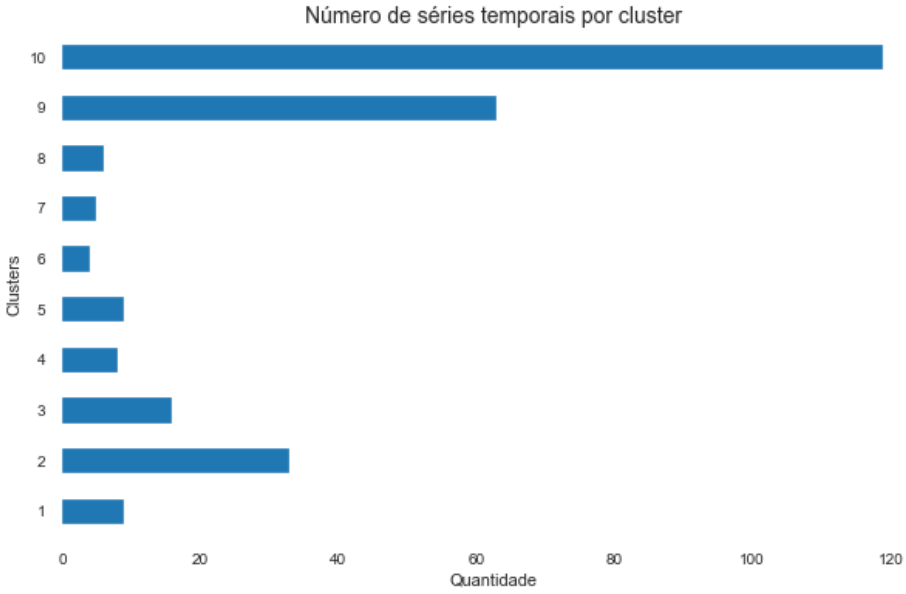
Fonte: Elaborado pelo autor.

O dendograma e os quadriláteros que se formam pela densidade de cores possibilitam uma estimativa inicial da número de clusters que pode ser obtido pelo processo de clusterização.

Os ramos do dendograma obtido são caracterizados por uma alta densidade na parte inferior do dendograma o que indica que pequenas variações na distância de corte da árvore resultam em grandes variações do número de clusters.

O número de clusters foi obtido de forma iterativa ao variar o parâmetro da distância de corte do dendograma. Durante o processo, buscou-se balancear a troca entre: ter um grande número de clusters e uma alta capacidade de generalização para um pequeno número de séries; e um pequeno número de clusters com uma baixa capacidade de generalização devido à baixa similaridade entre os itens do cluster. A Figura 34 mostra como estão distribuídas as séries temporais nos clusters.

Figura 34 - Distribuição das séries temporais por cluster



Fonte: Elaborado pelo autor.

A análise visual das séries temporais de cada cluster permite a identificação de padrões de propriedades do cluster.

As séries temporais do cluster 1, que totalizam 9 das 272 séries, demonstraram, durante os períodos iniciais, grande amplitude de valores.

As 33 séries do cluster 2 apresentaram menor amplitude e uma leve tendência de crescimento nos meses finais do período considerado.

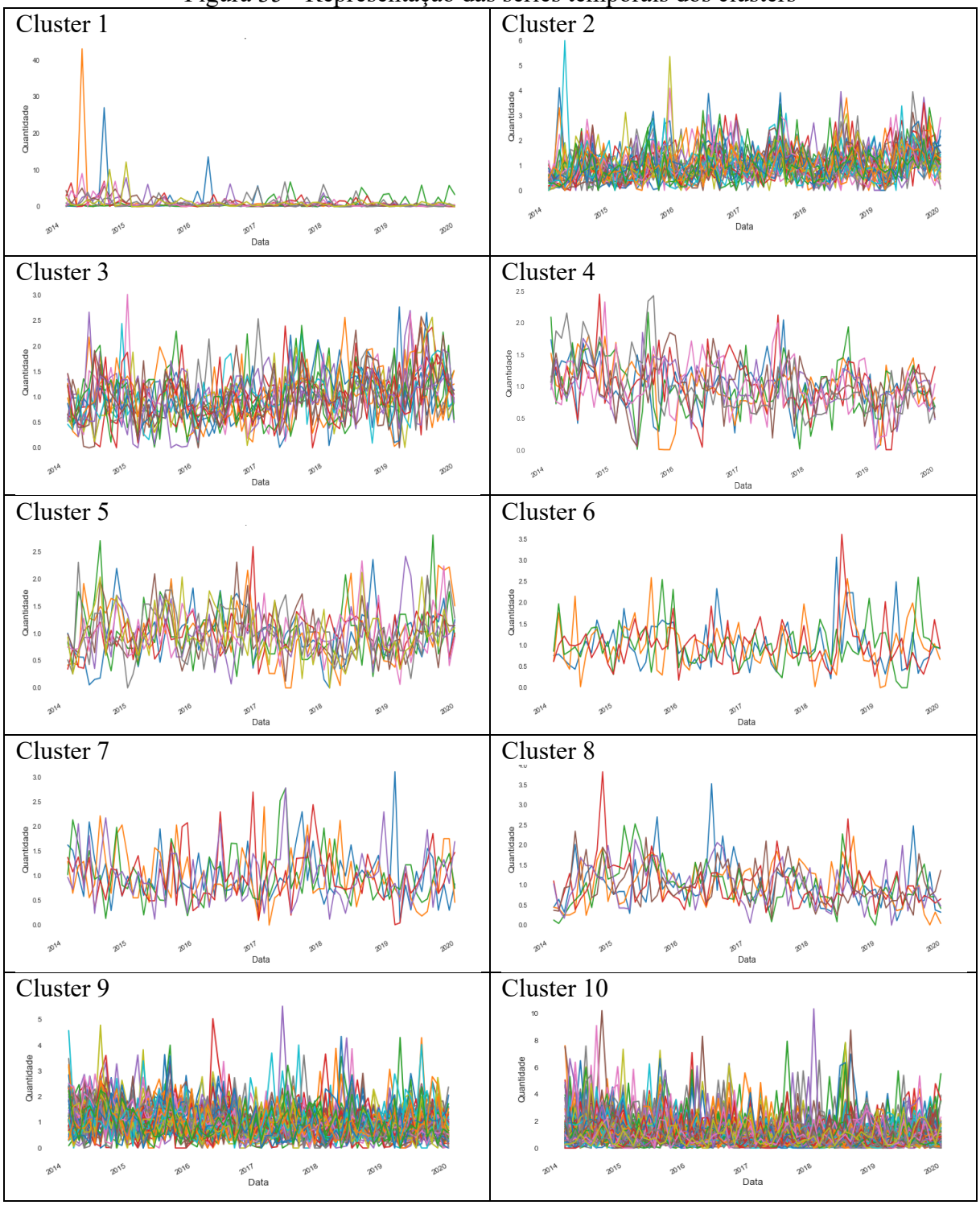
As 16 séries do cluster 3, assim como as séries do cluster 2, apresentaram um leve crescimento no fim do período; entretanto apresentam amplitudes menores ao longo dos meses.

As 8 séries do cluster 4 apresentaram tendência negativa de crescimento e baixa amplitude de valores.

As séries dos clusters 5, 6, 7, 8 e 9 demonstraram ausência de tendência e baixa amplitude

As séries do cluster 10 apresentaram maiores amplitudes e tendência negativa de crescimento. Os comportamentos descritos podem ser observados nos gráficos da Figura 35.

Figura 35 - Representação das séries temporais dos clusters



Fonte: Elaborado pelo autor.

Com os clusters definidos, identificou-se a qual cluster as séries temporais inicialmente selecionadas pertenciam. Os resultados são sintetizados na Tabela 6.

Tabela 6 - Clusters da seleção inicial de artigos

Item	Cluster
I	2
II	5
III	7
IV	8
V	9
VI	10
VII	3

Fonte: Elaborado pelo autor.

A clusterização das séries temporais agrupou as séries da amostra original em clusters distintos, conforme mostra a Tabela 6. Este resultado sugere que a seleção da amostra inicial foi adequada para a clusterização pois as séries temporais escolhidas possuem comportamentos diversos, conforme avaliado na Tabela 3, e possuem potencial para a generalização de séries não inclusas na amostra inicial.

4.5.2 Avaliação da generalização dos modelos

Para verificar a possibilidade de generalização dos parâmetros dos modelos definidos, foi selecionado o cluster 3 para o qual determinou-se os parâmetros do modelo de previsão para os artigos pertencentes ao cluster por meio do método *Grid Search* e do método *Auto Arima*. Ou seja, procedeu-se de forma similar a realizada para a definição dos parâmetros para a seleção inicial de artigos.

Os parâmetros obtidos por meio do método *Grid Search* e do método *Auto Arima* podem ser visualizados na Tabela 7. Os parâmetros do representante do cluster são representados pelos parâmetros do modelo do item VII.

Tabela 7 - Parâmetros dos modelos dos itens do cluster 3

Item	*Grid Search*	*Auto Arima*
VIII	(0, 1, 1) (2, 0, 2, 12)	(0, 1, 1) (0, 0, 0, 0)
IX	(2, 0, 1) (0, 1, 0, 12)	(2, 0, 0) (0, 0, 0, 0)
X	(0, 0, 0) (2, 0, 1, 12)	(0, 0, 0) (0, 0, 0, 0)
XI	(2, 0, 1) (1, 1, 2, 12)	(1, 1, 1) (0, 0, 0, 0)
XII	(2, 0, 1) (1, 0, 1, 12)	(1, 0, 0) (0, 0, 0, 0)
XIII	(1, 0, 1) (0, 0, 0, 12)	(0, 1, 1) (0, 0, 0, 0)
XIV	(0, 0, 0) (0, 1, 1, 12)	(1, 0, 0) (0, 0, 0, 0)
XV	(2, 1, 2) (0, 0, 0, 12)	(0, 1, 1) (0, 0, 0, 0)
XVI	(1, 0, 1) (0, 0, 0, 12)	(1, 0, 1) (0, 0, 0, 0)
XVII	(0, 1, 1) (0, 1, 0, 12)	(1, 0, 0) (0, 0, 0, 0)
XVIII	(2, 1, 1) (2, 0, 1, 12)	(0, 1, 1) (0, 0, 0, 0)
XIX	(2, 0, 2) (0, 0, 2, 12)	(1, 0, 0) (0, 0, 0, 0)
XX	(0, 1, 1) (1, 1, 1, 12)	(1, 0, 0) (0, 0, 0, 0)
XI	(1, 0, 1) (1, 0, 0, 12)	(1, 0, 0) (0, 0, 0, 0)
XII	(1, 0, 1) (1, 0, 0, 12)	(1, 0, 0) (0, 0, 0, 0)
XIII	(2, 1, 1) (2, 0, 1, 12)	(0, 1, 1) (0, 0, 0, 0)

Fonte: Elaborado pelo autor.

4.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

O presente capítulo apresentou a empresa e os itens da amostra inicial de produtos selecionada para a análise. O capítulo também abordou a análise exploratória dos dados, na qual foram analisadas características e propriedades das séries temporais, como estatísticas gerais (média, desvio padrão, coeficiente de variação e distribuição dos quartis), estacionariedade, correlação entre as séries, decomposição das séries temporais e comportamento da tendência das séries ao longo do período considerado. A análise exploratória de dados da amostra inicial de produtos indica que as séries temporais possuem comportamentos distintos no que tange às

estatísticas gerais e ao comportamento da componente de tendência, enquanto possuem a mesma característica de estacionariedade.

A estimativa de parâmetros dos modelos de previsão, apresentada neste capítulo, foi realizada por meio de 3 métodos: *Auto Arima*, *Grid Search* e parametrização pelo cluster. Os parâmetros apresentados mostram diferentes configurações. Enquanto os parâmetros obtidos pelo método *Auto Arima* não indicam sazonalidade (parâmetros P, D, Q e s são iguais a 0), os parâmetros obtidos pelo *Grid Search* e pela clusterização indicam a presença de sazonalidade nas séries temporais. Ressalta-se que, apesar das séries temporais da amostra inicial de produtos serem estacionárias, os métodos de parametrização indicaram a diferenciação para estacionarização das séries, representada pelo parâmetro "d" do modelo.

A avaliação da generalização dos modelos consistiu em clusterizar as séries temporais com base na métrica de dissimilaridade *Dynamic Time Warping*. O resultado da clusterização das séries mostra que as séries selecionadas para a amostra inicial de produtos foram agrupadas em clusters distintos, o que corrobora com as características observadas na análise exploratória de dados e sugere que a seleção inicial de produtos foi adequada.

5 RESULTADOS E DISCUSSÃO

O presente capítulo apresenta a comparação dos resultados da modelagem das séries temporais por meio do método *Auto Arima*, *Grid Search* e generalização dos modelos por clusterização. Os resultados são analisados a partir das métricas RMSE e MAE. Por fim, são apresentadas considerações finais sobre o capítulo.

5.1 ANÁLISE DOS MODELOS

Os dados foram divididos entre treino e teste (respectivamente 80% e 20% dos dados) e os valores do conjunto de teste alimentaram o conjunto de treino a medida em que fora realizada uma nova iteração. A partir desta divisão, o período de treino foi delimitado entre janeiro de 2014 a outubro de 2018, enquanto o período de testes entre novembro de 2018 a dezembro de 2019. Os resultados são mostrados na Tabela 8.

Tabela 8 - Resultados dos modelos avaliados por RMSE e MAE.

	RMSE		MAE	
Item	*Grid Search*	*Auto Arima*	*Grid Search*	*Auto Arima*
I	47,42	47,71	35,14	37,48
II	31,88	60,43	27,74	49,18
III	1637,85	2342,02	1206,24	1716,92
IV	16,56	22,08	12,89	19,00
V	5,74	8,29	4,52	7,10
VI	17,82	20,33	13,65	16,03
VII	76,94	79,21	55,52	57,98

Fonte: Elaborado pelo autor.

Os resultados indicam que, na totalidade dos casos avaliados, as previsões geradas pelos parâmetros obtidos via *Grid Search* obtiveram resultados superiores do que as previsões geradas pelos parâmetros obtidos pelo o método *Auto Arima*. Os resultados superiores podem ser explicados devido ao fato de o método *Grid Search* realizar todas as combinações possíveis de parâmetros, enquanto a função *Auto Arima* testa apenas uma parcela destas combinações.

Ressalta-se a significativa diferença no esforço computacional entre a utilização do método *Grid Search* e do método *Auto Arima*. O tempo médio de parametrização pelo método

Auto Arima é de 0,60 segundo enquanto o tempo médio de parametrização pelo método *Grid Search* é de 18.155,30 segundos. As métricas de erro, RMSE e MAE, que comparam os valores originais das séries com os valores previstos pelos modelos, foram reduzidas em 22,8% e 23,88% respectivamente, ao comparar as previsões do *Auto Arima* com as previsões do *Grid Search.* A redução % do erro é calculada por

$$Redução\ \%\ do\ erro = \frac{(Erro\ Auto\ Arima - Erro\ Grid\ Search)}{Erro\ Auto\ Arima}$$

A Tabela 9 sintetiza os tempos de processamento para a modelagem das séries temporais dos itens da amostra inicial e a redução percentual do erro para as métricas.

Tabela 9 - Tempos de processamento

Item	*Auto Arima* (s)	*Grid Search* (s)	Redução % do erro (RMSE)	Redução % do erro (MAE)
I	0,86	18.135,60	11,09%	6,24%
II	0,33	18.257,92	47,24%	43,59%
III	0,23	18.035,70	30,07%	29,74%
IV	0,38	17.989,34	25,00%	32,16%
V	0,26	18.402,51	30,76%	36,34%
VI	1,64	18.255,56	12,35%	14,85%
VII	0,51	18.010,12	2,87%	4,24%
Média	0,60	18.155,30	22,8%	23,88%

Fonte: Elaborado pelo autor.

5.2 ANÁLISE DA GENERALIZAÇÃO DOS MODELOS

A análise da generalização foi dividida em duas etapas: análise dos parâmetros e análise dos erros. Inicialmente, buscou-se identificar se os parâmetros definidos pelo método *Grid Search* eram similares entre itens de um mesmo cluster, similares ao representante do cluster e distintos de itens de outros clusters. Em seguida, foram analisados o RMSE e o MAE

associados às previsões com os parâmetros obtidos pelo método *Grid Search*, *Auto Arima* e parâmetros do representante do cluster.

5.2.1 Análise dos parâmetros

A análise visual dos parâmetros obtidos por meio do método *Grid Search* entre produtos do cluster 3 e dos demais clusters, sugere que não há um padrão nos parâmetros entre produtos de um mesmo cluster. Tal fato pode ser verificado observando a alta variabilidade dos valores dos parâmetros na Figura 36.

Figura 36 - Parâmetros dos modelos de previsão

	p	d	q	Q	P	D
Cluster 3	2	1	2	0	0	0
Cluster 3	2	1	1	2	2	0
Cluster 3	2	1	1	2	0	0
Cluster 9	2	1	0	0	0	0
Cluster 3	2	0	1	2	2	1
Cluster 3	2	0	1	2	2	0
Cluster 2	2	0	1	2	2	0
Cluster 3	2	0	1	0	0	1
Cluster 3	1	0	1	0	2	0
Cluster 3	1	0	1	0	2	0
Cluster 3	1	0	1	0	2	0
Cluster 3	1	0	1	0	0	0
Cluster 3	1	0	1	0	0	0
Cluster 7	1	0	0	2	2	1
Cluster 3	0	1	1	2	2	1
Cluster 3	0	1	1	2	2	0
Cluster 3	0	1	1	0	0	1
Cluster 10	0	1	0	2	2	0
Cluster 8	0	0	1	2	2	0
Cluster 5	0	0	0	2	2	1
Cluster 3	0	0	0	2	2	0
Cluster 3	0	0	0	2	0	1

Fonte: Elaborado pelo autor.

O parâmetro “d”, que indica o número de vezes que uma série deve ser diferenciada para se tornar estacionária, oscilou em 0 ou 1, o que indica que séries de um mesmo cluster podem apresentar diferentes características quanto à estacionariedade.

Os parâmetros Q, P e D são utilizados para modelagem de séries que possuem componentes de sazonalidade. A diferença na parametrização das séries de um mesmo cluster indica que séries de um mesmo cluster podem possuir características diferentes quanto à sazonalidade.

5.2.2 Análise dos erros

A análise dos erros indica que o método *Grid Search* gerou resultados com maior acurácia em relação aos outros métodos de parametrização.

O método de parametrização pelo cluster obteve resultados com maior acurácia para 10 dos 15 itens sendo dois iguais aos resultados obtidos pelo método *Grid Search*. Os resultados estão apresentados na Tabela 10.

Tabela 10 – RMSE e MAE dos itens do cluster 3

	RMSE			MAE		
Item	*Grid Search*	*Auto Arima*	*Cluster*	*Grid Search*	*Auto Arima*	*Cluster*
VIII	7,29	7,61	7,93	5,82	5,83	6,18
IX	66,94	74,95	70,55	56,87	59,81	59,19
X	34,27	90,81	40,32	26,85	83,56	33,71
XI	33,07	47,41	57,92	24,62	37,89	47,04
XII	12,83	16,85	14,35	9,82	14,89	11,28
XIII	64,09	66,38	64,73	48,89	49,72	49,56
XIV	43,04	58,86	51,43	37,10	50,71	43,01
XV	68,3	70,16	76,34	51,00	51,19	58,20
XVI	80,64	80,64	83,18	61,41	61,41	63,23
XVII	47,48	51,57	48,62	34,55	40,66	39,53
XVIII	39,62	44,78	46,24	31,12	35,65	35,90
XIX	11,4	16,16	11,62	9,30	13,33	9,48
XX	9,18	12,66	9,89	7,68	9,10	8,30
XXI	9,46	14,7	9,46	7,50	12,52	7,50
XXII	6,17	7,89	6,17	4,70	6,64	4,70

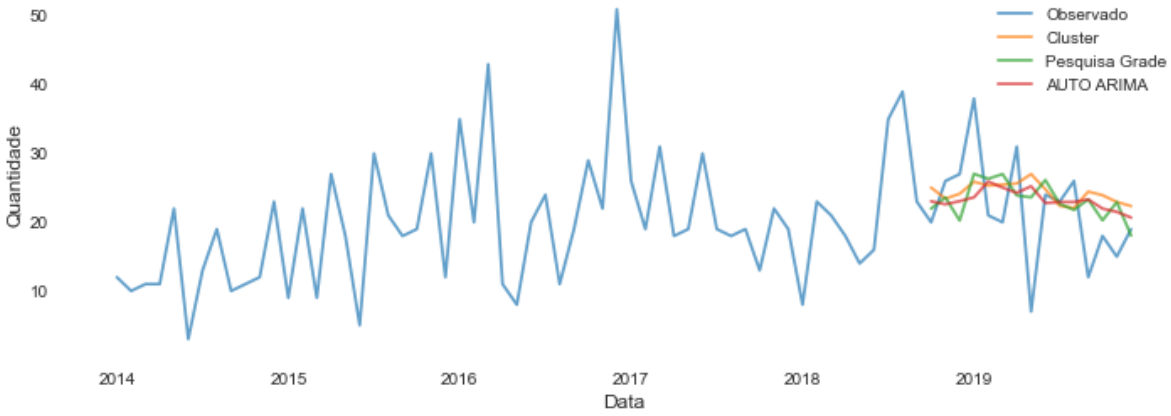
Fonte: Elaborado pelo autor.

5.2.3 Análise visual dos resultados

A análise visual dos resultados permite a compreensão do comportamento dos modelos frente às séries originais. Neste tópico são analisados 5 dos 15 itens do cluster 3.

Os modelos do item VIII demonstram comportamento similar, variando com baixa amplitude e falhando em detectar alterações bruscas nos valores da série temporal.

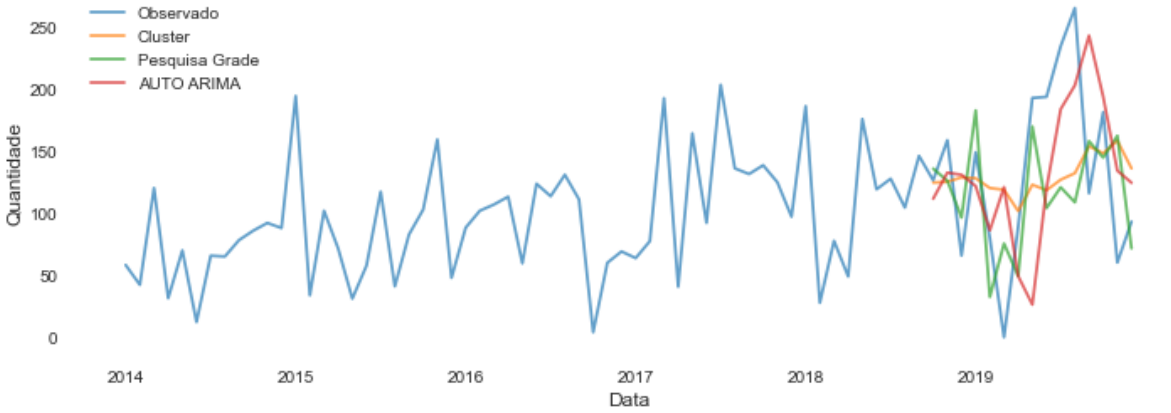
Figura 37 - Análise visual dos resultados do item VIII



Fonte: Elaborado pelo autor.

O modelo *Auto Arima* do item IX demonstra atraso em relação à série temporal original enquanto a parametrização obtida pelo cluster tende a prever valores medianos e não acompanhar as mudanças na série. A modelagem pelo método *Grid Search* mostrou maior capacidade em acompanhar os valores da série original.

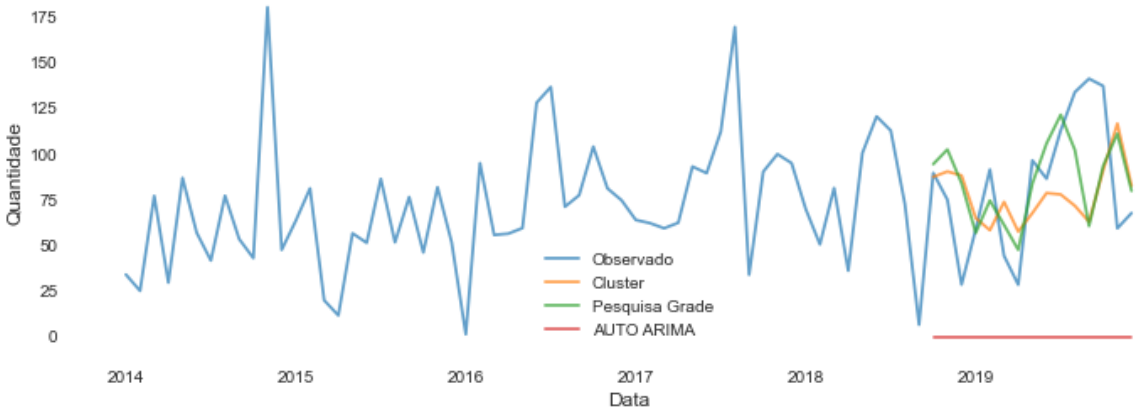
Figura 38 - Análise visual dos resultados do item IX



Fonte: Elaborado pelo autor.

Para o item X, o método *Auto Arima* falhou em identificar um parâmetro adequado para o modelo e realizou previsões zeradas para o período. Os parâmetros do cluster novamente geraram previsões medianas enquanto o método *Grid Search* demonstrou maior aderência à série original.

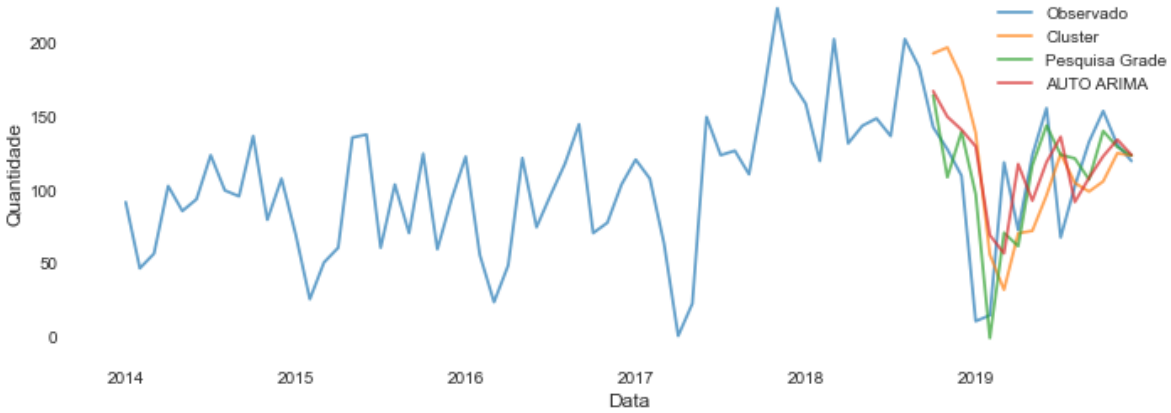
Figura 39 - Análise visual dos resultados do item X



Fonte: Elaborado pelo autor.

As previsões geradas pelos parâmetros obtidos pelo método *Grid Search* demostraram maior capacidade em acompanhar as mudanças da série do item XI enquanto os parâmetros dos demais modelos apresentam atraso em relação à série original.

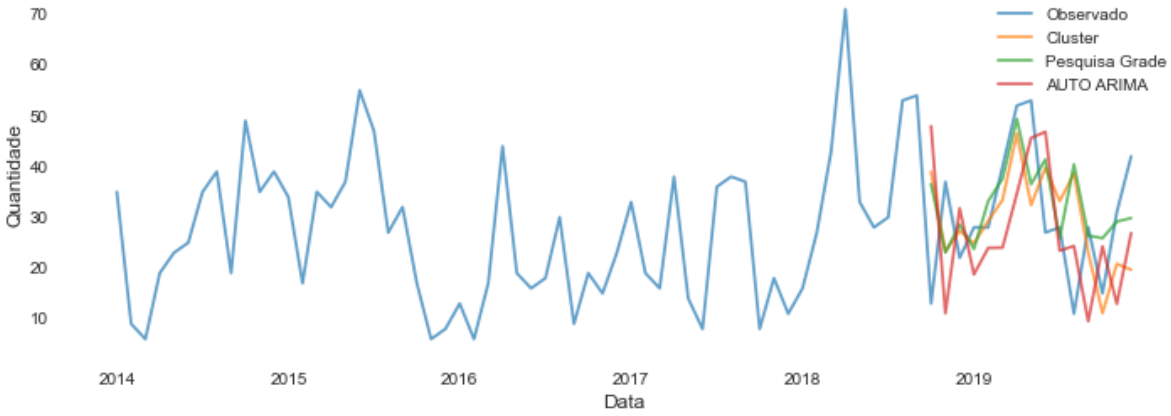
Figura 40 - Análise visual dos resultados do item XI



Fonte: Elaborado pelo autor.

As previsões geradas pelo método *Auto Arima* para o item XII apresentam atraso em relação à série original enquanto os demais modelos demonstraram maior capacidade em acompanhar a série original.

Figura 41 - Análise visual dos resultados do item XII



Fonte: Elaborado pelo autor.

5.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

O presente capítulo apresentou a comparação dos resultados da modelagem das séries temporais por meio do método *Auto Arima*, *Grid Search* e generalização dos modelos por clusterização.

A análise dos erros mostra que o método *Grid Search* conduz a previsões com menores erros quando avaliados pelas métricas RMSE e MAE. As previsões com menores erros são justificadas devido ao fato de que o método utilizado testa todas as combinações de parâmetros, dado um intervalo de parâmetros, de modo que a métrica RMSE seja minimizada, enquanto o método *Auto Arima* testa uma combinação reduzida de parâmetros com base em heurísticas. Ressalta-se que, o método *Grid Search*, apesar de conduzir a previsões com maior acurácia, possui tempo de processamento significativamente superior ao método *Auto Arima*.

A avaliação da generalização dos modelos foi realizada através da análise dos parâmetros e da análise dos erros. A análise dos parâmetros não permitiu a identificação de padrões de parâmetros entre produtos de um mesmo cluster, tampouco padrões de sazonalidade e estacionariedade. A análise dos erros mostra que a parametrização pelo cluster conduziu a previsões com maior acurácia para 10 dos 15 produtos quando comparado com a parametrização pelo método *Auto Arima*.

6 CONCLUSÕES E RECOMENDAÇÕES

O presente trabalho teve por objetivo analisar modelos de previsão de demanda para produtos de um comércio varejista localizado no norte do país.

O trabalho foi iniciado com a extração dos dados históricos de demanda de 72 meses, entre janeiro de 2014 e dezembro de 2019. Após a extração, realizou-se a análise exploratória dos dados com o objetivo de visualizar e consolidar os dados históricos. A modelagem das séries temporais foi realizada por três métodos de parametrização: a) *Grid Search*, no qual é realizada uma busca exaustiva de parâmetros; b) *Auto Arima*, que consiste em uma função de otimização e c) parametrização pelo cluster, na qual os parâmetros da série representante do cluster são generalizados para as demais séries do cluster. Os resultados das previsões foram analisados pelas métricas RMSE e MAE.

Os resultados obtidos indicam que o método *Grid Search* conduziu a resultados com menor RMSE e MAE em comparação aos resultados obtidos pelo *Auto Arima* e pela clusterização. O método da clusterização, por outro lado, gerou resultados com menor RMSE e MAE para 10 dos 15 produtos considerados no cluster enquanto, o método *Auto Arima* resultou em previsões com menor acurácia.

Os tempos de processamento para a parametrização dos modelos por meio dos diferentes métodos foram discrepantes. O tempo médio de processamento do método *Grid Search* foi de 18.155,30 segundos enquanto o tempo médio do método *Auto Arima* foi de 0,60 segundos. As diferenças entre os tempos de processamento e acurácias obtidas ressaltam a importância da definição dos objetivos de um modelo de previsão, nos quais são especificados os níveis de acurácia desejados e a disponibilidade de recursos computacionais.

A parametrização dos modelos por meio do método *Grid Search*, quando avaliada pela métrica RMSE, pode gerar resultados conflituosos com a parametrização sugerida pela literatura. A literatura indica que o parâmetro “d” representa o número vezes que uma série deve ser diferenciada para tornar-se estacionária. Dentre as sete séries temporais estacionárias da seleção inicial, duas apresentaram o parâmetro “d” diferente de 0 quando parametrizados pelo método *Grid Search*, o que indica a inconsistência com a parametrização proposta na literatura. Apesar da inconsistência descrita, a diferenciação de séries estacionárias pode melhorar os resultados dos modelos de previsão enquanto a clusterização pode auxiliar na parametrização de modelos de previsão de demanda.

Recomenda-se para trabalhos futuros, a modelagem dos demais produtos não inclusos na seleção inicial. Recomenda-se também o teste de modelos que incluam séries temporais exógenas como a cotação e volume vendido de ouro, índices pluviométricos, taxa de juros, tais como o modelo SARIMAX, modelos de aprendizado de máquina e redes neurais.

REFERÊNCIAS

ABU, Sean. Seasonal arima with python: time series forecasting: creating a seasonal arima model using python and statsmodel. **SeanAbu**, 22 mar. 2016. Disponível em: https://www.seanabu.com/2016/03/22/time-series-seasonal-ARIMA-model-in-python/. Acesso em: 01 out. 2019.

ARCHER, B. H. Forecasting demand: quantitative and intuitive techniques. **International Journal of Tourism Management**, [S.l.], v. 1, n. 1, p. 5-12, mar. 1980. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/014325168090016X. Acesso em: 03 dez. 2019.

BAÍA JUNIOR, P. **Entre o ouro e a biodiversidade**: garimpos e unidades de conservação na região de Itaituba, Pará, Brasil, 2014. 211 f. Tese (Doutorado) – Universidade Federal do Pará, Núcleo de Altos Estudos Amazônicos, Programa de Pós-Graduação em Desenvolvimento Sustentável do Trópico Úmido. Belém, 2014. Disponível em: http://repositorio.ufpa.br/jspui/handle/2011/7774. Acesso em: 03 dez. 2019.

BARBAKH, W. A.; WU, Y.; FYFE, C. **Non-standard parameter adaptation for exploratory data analysis**. [S.l.]: Springer, 2009.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time series analysis**: forecasting and control. Englewood Clifs: Prentice Hall, 1994.

BOX, G. E. P.; JENKINS, G. M. **Time series analysis**: forecasting and control. [S.l.]: Holden-Day, 1979.

BRASIL. Lei nº 11.685, de 2 de junho de 2018. Institui o Estatuto do Garimpeiro e dá outras providências. **Diário Oficial da União**, Brasília, DF, 2 jun. 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2008/Lei/L11685.htm. Acesso em: 01 out. 2019.

BUENO, R. De L. da S. **Econometria de séries temporais**. 2. ed. São Paulo: Cengage Learning, 2011.

CAHETÉ, F. S. A extração do ouro na Amazônia e implicações para o meio ambiente. Novos **Cadernos NAEA**, Belém, v. 1, n. 1, 1998. Disponível em: https://periodicos.ufpa.br/index.php/ncn/article/view/14. Acesso em: 05 dez. 2019.

CLARK, J.; DOWNING, D. **Estatística Aplicada**. 2. ed. São Paulo: Saraiva, 2005

CHEN, Y.; TJANDRA, S. Daily collision prediction with SARIMAX and generalized linear models on the basis of temporal and weather variables. **Transportation Research Record: Journal of the Transportation Research Board**, [S.l.], v. 2432, n. 1, p. 26-36, 2014. Disponível em: http://saiv.espaceweb.usherbrooke.ca/References/347_2014_DailyCollissionPrevisionBasisOnOfTemporalAndWeatherVariables_11p.pdf. Acesso em: 10 dez. 2019.

COELHO, M. C.; WANDERLEY, L. J.; COSTA, Reinaldo. Garimpeiros de ouro e cooperativismo no século XXI: exemplos nos rios Tapajós, Juma e Madeira no sudoeste da Amazônia Brasileira. **Confins**, São Paulo, v. 33, n. 33, 2017. Disponível em: http://journals.openedition.org/confins/12445. Acesso em: 03 dez. 2019.

DUKE UNIVERSITY. People.duke.edu. **Forecasting flow chart**. 2016. Disponível em: https://people.duke.edu/~rnau/411flow.gif. Acesso em: 08 dez. 2019.

EHLERS, Ricardo S. **Análise de séries temporais**. 4. ed. Curitiba: Universidade Federal do Paraná, 2007. Disponível em: https://www.docsity.com/pt/analise-de-series-temporais/4708722/. Acesso em: 10 dez. 2019.

ELSAYED, A. E.; BOUCHER, T. O. **Analysis and control of production systems**. 2. ed. [S.l.]: Prentice Hall, 1993.

FAVA, V. L. Análise de séries de tempo. In: VASCONSELOS, M. A. S.; ALVES, D. (Org.). **Manual de econometria**: nível intermediário. São Paulo: Atlas, 2000.

FONSECA, J. S. da; MARTINS, G. de A.; TOLEDO, G. L. **Estatística aplicada**. São Paulo: Editora Atlas, 1985.

GENESIS. **Components of time series**. 27 set. 2018. Disponível em: https://www.fromthegenesis.com/components-of-time-series/. Acesso em: 13 dez. 2019.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 4 ed. São Paulo: Atlas, 1994.

HAIR JÚNIOR, J. F. *et al.* **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009.

HILL, R. C.; JUDGE, G. G.; GRIFFITHS, W. E. **Econometria**. 3. ed. São Paulo: editora Saraiva, 2010.

HRUSCHKA, E. R.; EBECKEN, N. A Genetic algorithm for cluster analysis. **Intelligent Data Analysis**, [S.l.], v. 7, n. 1 p. 15-25, fev. 2003. Disponível em: https://www.researchgate.net/publication/220571471_A_genetic_algorithm_for_cluster_analysis. Acesso em: 02 out. 2019.

KATE, R. J. Using dynamic time warping distances as features for improved time series classification. **Data Mining and Knowledge Discovery**, [S.l.], n. 30, p. 283-312, 2016. Disponível em: https://link.springer.com/article/10.1007/s10618-015-0418-x. Acesso e: 14 dez. 2019.

KEOGH, E. J.; KASETTY, S. On the need for time series data mining benchmarks: a survey and empirical demonstration. **Data Mining & Knowledge Discovery**, [S.l.]. v. 7, n. 4, p. 349–371, out. 2003. Disponível em: https://insights.ovid.com/data-mining-knowledge-discovery/dmkd/2003/07/040/need-time-series-data-mining-benchmarks-survey/2/00124422. Acesso em: 20 out.2019.

KOLEN, J.; THEIJE, M.; MATHIS, A. Formalized small-scale gold mining in the brazilian Amazon: an activity surrounded by informality. In: CREMERS, L.; KOLEN, J.; THEIJE, M. (Eds.). Small-scale gold mining in the Amazon. the cases of Bolivia, Brazil, Colombia, Peru and Suriname. **Cuadernos del CEDLA**, Amsterdam, n. 26, p. 31-45, 2013. Disponível em: http://www.cedla.uva.nl/50_publications/pdf/cuadernos/cuad26.pdf. Acesso em: 01 out. 2019.

LERATO, L.; NIESLER, T. Feature trajectory dynamic time warping for clustering of speech segments. **EURASIP Journal on Audio, Speech, and Music Processing**, [S.l.], 2019. Disponível em: https://www.researchgate.net/publication/328628324_Feature_Trajectory_Dynamic_Time_Warping_for_Clustering_of_Speech_Segments. Acesso em: 12 dez. 2019.

LI, T.; WU, X.; ZHANG, J. Time series clustering model based on DTW for classifying car parks. **Algorithms**, [S.l.], V. 13, n. 3, mar. 2020. Diponível em: https://www.researchgate.net/publication/339630992_Time_Series_Clustering_Model_Based_on_DTW_for_Classifying_Car_Parks. Acesso em: 04 abr. 2020.

LIAO, T. W. Clustering of time series data: a survey. **Pattern Recognition**, [S.l.], v. 38, n. 11, p. 1857-1874, nov. 2005. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0031320305001305. Acesso em: 20 dez. 2019.

MARTIN, A. C.; HENNING, E.; WALTER, O. M. F. C.; KONRATH, A. C. Análise de séries temporais para previsão da evolução do número de automóveis no Município de Joinville. **Revista Espacios**, Caracas, Venezuela, v. 37, n. 6, 2016. Disponível em: https://qualimetria.ufsc.br/files/2016/05/Revista-ESPACIOS-_-Vol.pdf. Acesso em: 12 dez. 2019.

MARTINS, P. G.; LAUGENI, F. P. **Administração da produção**. 3. ed. São Paulo: Saraiva, 2015.

MATHIS, A. Garimpos de ouro na Amazônia: fatores sociais, relações de trabalho e condições de vida. **Papers do NAEA**, Belém, n. 37, p. 3-16, abr. 1995. Disponível em: file:///C:/Users/Grazi/Downloads/037.pdf. Acesso em: 10 dez. 2019.

MATTOS, R. S. de. **Tendências e raízes unitárias**. Juiz de Fora: Universidade Federal de Juiz de Fora, 2018. [Texto didático]. Disponível em: https://www.ufjf.br/wilson_rotatori/files/2011/05/Tendencias-e-Raizes-Unitarias-2018.pdf. Acesso em: 10 out. 2019.

MILONE, G. **Estatística geral e aplicada**. 2. ed. São Paulo: Thomson Learning, 2006.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de series temporais**. 2. ed. São Paulo: Egard Blucher, 2006.

NAU, R. **Principles and risks of forecasting**. Durham: Duke University, 2014. Disponível em: https://people.duke.edu/~rnau/Principles_and_risks_of_forecasting--Robert_Nau.pdf. Acesso em: 10 dez. 2019.

NAU, R. **Stationarity and differencing**. Durham: Duke University, [201-]. Disponível em: https://people.duke.edu/~rnau/411diff.htm. Acesso em: 03 dez. 2019.

NAU, R. **Statistical forecasting**: notes on regression and time series analysis. Durham: Duke University, 2015. Disponível em: https://people.duke.edu/~rnau/411home.htm. Acesso em: 02 dez. 2019.

PYRAMID. API Reference. **Pyramid.arima.auto_arima**. 2020. Disponível em: http://alkaline-ml.com/pmdarima/0.9.0/modules/generated/pyramid.arima.auto_arima.html. Acesso em: 20 jan. 2020.

RHYS, H. **Machine Learning with R, the tidyverse, and mlr**. Shelter Island: Manning Publications, 2020. Disponível em: https://livebook.manning.com/book/machine-learning-for-mortals-mere-and-otherwise/chapter-17/. Acesso em: 20 set. 2020.

SAMOHYL, R. W. **Controle estatístico de qualidade**. Amsterdã: Elsevier, 2009.

SCIPY.ORG. **Scipy.cluster.hierarchy.ward**. 2020. Diponível em: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.ward.html. Acesso em: 10 dez. 2019.

SEABORN. **Seaborn.clustermap**. 2020. Disponível em: https://seaborn.pydata.org/generated/seaborn.clustermap.html. Acesso em: 10 dez. 2019.

SERRÀ, J; ARCOS, J. L. An empirical evaluation of similarity measures for time series classification. **Knowledge-Based Systems**, [S.l.] n. 67, p. 305-314, 2014. Disponível em: https://arxiv.org/abs/1401.3973. Acesso em: 10 out. 2019.

SILVA, E. M; SILVA, E. M. **Matemática e estatística aplicada**. São Paulo: Atlas, 1999.

SILVA, P. L. P. da. **Um estudo sobre o agrupamento de séries temporais e sua aplicação em curvas de cargas residenciais**. 2016. 171 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte, 2016. Disponível em: https://repositorio.ufmg.br/handle/1843/BUOS-APWMJD. Acesso em: 10 dez. 2019.

SLACK, N.; CHAMBERS, S.; JOHNSTON, R. **Administração da produção**. 3. ed. São Paulo: Editora Atlas, 2009.

SOUSA, R. *et al.* Policies and regulations for Brazil's artisanal gold mining sector: analysis and recommendations. **Journal of Cleaner Production**, [S.l.], v. 19, n. 6-7, p. 742-750, abr./maio 2011. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S095965261000452X. Acesso em: 01/out. 2019.

SREEDHAR KUMAR, S. *et al*. A brief survey of unsupervised agglomerative hierarchical clustering schemes Sreedhar. **International Journal of Engineering & Technology**, [S.l.], v.

8, n. 1, p. 29-37, 2019. Disponível em: https://www.sciencepubco.com/index.php/ijet/article/view/13971. Acesso em: 08 dez. 2019.
STATSMODELS. **Statsmodels v0.12.0**. 2020. Disponível em: https://www.statsmodels.org/stable/index.html. Acesso em: 02 mar.2020.

THE DTW SUITE. **Github**. 2019. Disponível em: https://dynamictimewarping.github.io/python/. Acesso em: 02 out. 2019.

TONI, G. Computing and visualizing dynamic time warping alignments in R: the dtw package. **J. Stat. Soft.**, [S.l.], n. 31, 2009. Disponível em: https://cran.r-project.org/web/packages/dtw/vignettes/dtw.pdf. Acesso em: 10 nov. 2019.

YE, Y. *et al.* Similarity measures for time series data classification using grid representation and matrix distance. **Knowledge and Information System**, [S.l.], v. 60, n. 2, ago. 2019. Disponível em: https://dl.acm.org/doi/10.1007/s10115-018-1264-0. Acesso em: 15 out.