



Universidade Federal de Uberlândia
Faculdade de Engenharia Elétrica
Graduação em Engenharia Biomédica

DHARA LOUISE CAMPOS DE ASSIS

**ANÁLISE TEMPORAL PREDITIVA DE ACIDENTES NO TRABALHO:
UMA ABORDAGEM DO MODELO SARIMA**

Uberlândia
2019

DHARA LOUISE CAMPOS DE ASSIS

**ANÁLISE TEMPORAL PREDITIVA DE ACIDENTES NO TRABALHO:
UMA ABORDAGEM DO MODELO SARIMA**

Trabalho apresentado como requisito parcial de avaliação na disciplina Trabalho de Conclusão de Curso de Engenharia Biomédica da Universidade Federal de Uberlândia.

Orientador: Prof. Adriano de Oliveira Andrade, PhD

Assinatura do Orientador

Uberlândia
2019

Dedico este trabalho aos meus pais, Fátima e Sérgio, e ao meu irmão, Guga, pela paciência, suporte e carinho demonstrados em toda a minha trajetória acadêmica.

AGRADECIMENTOS

Agradeço a Deus, por ser quem Ele é: aquele que guia e sempre guiará meus passos e cuja alegria é a minha força. Por aquietar meu coração em momentos difíceis e por me proporcionar conquistas e oportunidades únicas.

Agradeço especialmente aos meus pais e irmão, por serem meu suporte em absolutamente todas as etapas da minha vida e por sacrificarem tudo o que podem para o meu sucesso pessoal, acadêmico e profissional.

Ao meu namorado Odelmo, por me inspirar a ser a melhor versão de mim mesma, por sempre me encorajar a defender minhas ideias além de ser meu companheiro fiel em todos os momentos.

Ao meu professor orientador Adriano, pela maestria em compartilhar conhecimentos, pelo incentivo e apoio não apenas na elaboração deste trabalho mas também durante a minha graduação.

Ao meu gestor, Carlos, pela confiança, pelos feedbacks essenciais e por me apresentar um grande propósito em minha vida: a Engenharia de Segurança.

Aos professores e técnicos da FEELT pelas lições e por contribuírem para meu crescimento.

Aos meus amigos, dentro e fora da faculdade, que são essenciais e sempre estiveram presentes todo o tempo.

RESUMO

O protagonismo do pilar segurança em ambientes de trabalho é um fator de extrema importância no desempenho e reputação de uma empresa, medidos por indicadores estabelecidos por órgãos internacionais, além do fato de uma cultura de segurança bem estabelecida ser decisiva na manutenção da saúde do trabalhador. Para desafiar o status quo em segurança do trabalho, ou seja, o caráter reativo dado as tratativas de mitigação de risco acionadas circunstancialmente perante ocorrência pontual de incidentes, é necessária a tomada de decisão em caráter preventivo, baseada em dados. A estatística então torna-se ferramenta poderosa para a mudança de mindset. Nesse sentido, o presente trabalho apresenta uma aplicação do modelo estatístico SARIMA (Seasonal Auto Regressive Integrated Moving Average) na predição de acidentes de trabalho com dados em um período de janeiro de 2016 a julho de 2019, para previsão até dezembro de 2019, baseados em um indicador calculado a partir da definição da OSHA (Occupational Safety and Health Administration): Taxa de incidentes reportáveis (TRIR). O melhor modelo foi o SARIMA(1,2,1)(1,1,1) e obteve-se uma previsão de 0.21 para dezembro. Foram utilizados os testes de média percentual absoluta do erro (MAPE) com resultado de 5.71% comprovando equivalências do resultado do modelo e do valor real separado para a validação.

Palavras-chave: SARIMA, previsão, acidentes do trabalho, Box-Jenkins, Engenharia de Segurança.

ABSTRACT

The role of the safety pillar in the workplace is an extremely important factor in a company's performance and reputation, as measured by indicators used by international agencies, as well as the fact that a successful safety culture is crucial to maintaining the health of workers. Data-based proactive decision-making is important to challenge the status quo of occupational safety, i.e. the reactive character provided as circumstantially induced risk mitigation transactions in the occasional incidence of accidents. In this sense, this work presents an application of the statistical model SARIMA (Seasonal Autoregressive Integrated Moving Average) in the prediction of occupational accidents with data from January 2016 to July 2019, forecast for December 2019, The absolute mean percent error (MAPE) tests with a result of 5.71 percent were used and the model result and the real value of the following months were founded equivalences.

Keywords: SARIMA, forecast, occupational accident, Box-Jenkins, Safety Engineering.

LISTA DE TABELAS

Tabela 1 – Comportamento das funções de auto correlação e auto correlação parcial para cada parâmetro do modelo.	23
Tabela 2 – Coeficientes do ajuste da variável de teste SARIMA(1,2,1)(1,1,1).....	33
Tabela 3 – Coeficientes do ajuste na variável de teste SARIMA(1,2,1)(1,1,1).....	34
Tabela 4 – Previsão da série temporal de teste SARIMA(1,2,1)(1,1,1).....	35
Tabela 5 – Resultados da previsão próximos 5 meses SARIMA(1,2,1)(1,1,1).	36
Tabela 6 - Coeficientes do ajuste da variável de teste SARIMA(3,1,0)(1,0,0).	38
Tabela 7 – Coeficientes do ajuste na variável de teste SARIMA(3,1,0)(1,0,0).....	40
Tabela 8 – Previsão da série temporal de teste SARIMA(3,1,0)(1,0,0).....	40
Tabela 9- Resultados da previsão próximos 5 meses SARIMA(3,1,0)(1,0,0).	42
Tabela 10 - Coeficientes do ajuste da variável de teste SARIMA(3,2,0)(1,0,0).	44
Tabela 11 – Coeficientes do ajuste na variável de teste SARIMA(3,2,0)(1,0,0).....	46
Tabela 12 - Previsão da série temporal de teste SARIMA(3,2,0)(1,0,0).....	46
Tabela 13 - Resultados da previsão próximos 5 meses SARIMA(3,2,0)(1,0,0).	48
Tabela 14 – Comparação dos valores de MAPE para os três modelos propostos.	51

LISTA DE ILUSTRAÇÕES

Figura 1 – Dados semanais de passageiros da classe econômica na companhia de voo Ansett.	16
Figura 2 – Dados mensais de venda de medicamentos diabéticos na Austrália.	17
Figura 3 – Correlograma de uma série representando o comportamento de produção de cerveja durante um período do ano.	19
Figura 4 – Ruído branco.	20
Figura 5 – Função de auto correlação de uma série representando o comportamento do ruído branco da Figura 4.	20
Figura 6 – Média móvel últimos 12 meses TRIR de janeiro de 2016 a julho 2019 para 5 sites de manufatura de sementes.	28
Figura 7 – Análise da série original, suas funções de auto correlação e auto correlação parcial.	30
Figura 8 – Série temporal analisada pela função decompose.	31
Figura 9 - Serie temporal com uma diferenciação.	32
Figura 10 – Série temporal com duas diferenciações.	32
Figura 11 – Plotagem da variável de teste (cinza) e ajuste (azul) SARIMA(1,2,1)(1,1,1).	33
Figura 12 – Predição utilizando a variável de teste SARIMA(1,2,1)(1,1,1).	34
Figura 13 – Plotagem dos valores de predição da variável de teste (cinza) e valores de validação (azul) SARIMA(1,2,1)(1,1,1).	34
Figura 14– Análise de resíduos de teste SARIMA(1,2,1)(1,1,1).	35
Figura 15 – Normal QQ Plot SARIMA(1,2,1)(1,1,1).	36
Figura 16 – Previsão para próximos 5 meses utilizando SARIMA(1,2,1)(1,1,1).	36
Figura 17 – Padrão dos resíduos da previsão próximos 5 meses SARIMA(1,2,1)(1,1,1).	37
Figura 18 – Função de auto correlação para os resíduos SARIMA(1,2,1)(1,1,1).	37
Figura 19 – Função de auto correlação parcial para os resíduos SARIMA(1,2,1)(1,1,1).	38
Figura 20 - Plotagem da variável de teste (cinza) e ajuste (azul) SARIMA(3,1,0)(1,0,0).	38
Figura 21 - Predição utilizando a variável de teste SARIMA(3,1,0)(1,0,0).	39
Figura 22- Plotagem dos valores de predição da variável de teste (cinza) e valores de validação (azul) SARIMA(3,1,0)(1,0,0).	40
Figura 23- Análise de resíduos de teste SARIMA(3,1,0)(1,0,0).	41
Figura 24 - Normal QQ Plot SARIMA(3,1,0)(1,0,0).	41
Figura 25– Previsão para próximos 5 meses utilizando SARIMA(3,1,0)(1,0,0).	42
Figura 26 –Padrão dos resíduos do valor ajustado SARIMA(3,1,0)(1,0,0).	43
Figura 27 - Função de auto correlação para os resíduos SARIMA(3,1,0)(1,0,0).	43
Figura 28 - Função de auto correlação parcial para os resíduos SARIMA(3,1,0)(1,0,0).	44
Figura 29 - Plotagem da variável de teste (cinza) e ajuste (azul) SARIMA(3,2,0)(1,0,0).	44
Figura 30 - Predição utilizando a variável de teste SARIMA(3,2,0)(1,0,0).	45
Figura 31 - Plotagem dos valores de predição da variável de teste (cinza) e valores de validação (azul) SARIMA(3,2,0)(1,0,0).	45
Figura 32 - Análise de resíduos de teste SARIMA(3,2,0)(1,0,0).	46
Figura 33 - Normal QQ Plot SARIMA(3,2,0)(1,0,0).	47

Figura 34 – Previsão para próximos 5 meses utilizando SARIMA(3,2,0)(1,0,0).	47
Figura 35 - Padrão dos resíduos do valor ajustado SARIMA(3,2,0)(1,0,0).	48
Figura 36 - Padrão dos resíduos do valor ajustado SARIMA(3,2,0)(1,0,0).	49
Figura 37 - Função de auto correlação para os resíduos SARIMA(3,2,0)(1,0,0).	49
Figura 38 – Comparação dos valores previstos pelos modelos (coloridos) e valores reais (preto)	50

LISTA DE ABREVIATURAS E SIGLAS

UFU – Universidade Federal de Uberlândia

FEELT – Faculdade de Engenharia Elétrica

SARIMA – Seasonal auto regressive integrated moving average

MAPE – Median Absolute Percentual Error

TRIR – Total recordable incident rate

OSHA – Occupational Safety Health Administration

Full time – Tempo integral

CONFEA – Conselho Federal de Engenharia e Agronomia

ACF – Autocorrelation function

PACF – Partial autocorrelation function

AR – Auto regressive

I – Integrated

MA – Moving average

KPSS - KwiatkowskiPhillips-Schmidt-Shin

AIC – Akaike information criterion

BIC – Bayesian information criterion

HHT – Homem horas trabalhadas

SUMÁRIO

1	INTRODUÇÃO	13
1.1	OBJETIVOS.....	15
1.1.1	Objetivo geral.....	15
1.1.2	Objetivos específicos	15
2	DESENVOLVIMENTO	16
2.1	FUNDAMENTAÇÃO TEÓRICA.....	16
2.1.1	Análises de séries temporais.....	16
2.1.2	Modelo auto regressivo integrado de médias móveis (ARIMA) (p,d,q)	21
2.1.3	AR (Auto regressivo) (p)	21
2.1.4	I (Integrado) (d)	22
2.1.5	MA (Média móvel) (q)	22
2.1.6	Modelo sazonal auto regressivo integrado de médias móveis (SARIMA) (p,d,q)(P,D,Q) 23	
2.1.7	Média absoluta percentual do erro (MAPE)	24
2.1.8	Teste Kwiatkowski, Phillips, Schmidt and Shin (KPSS).....	24
2.1.9	Teste de Fisher	24
2.1.10	Teste de Box-Pierce	25
2.1.11	Critério de informação Akaike (AIC).....	25
2.1.12	Critério de informação Bayesiano (BIC)	26
2.2	ARIMA COMO MODELO DE PREDIÇÃO.....	26
2.3	METODOLOGIA	27
2.3.1	Característica dos dados.....	27
2.3.2	Aplicação do método Box & Jenkins	29
3	RESULTADOS.....	30
4	DISCUSSÃO	49
5	CONCLUSÕES	52
6	REFERÊNCIAS.....	53

1 INTRODUÇÃO

O Brasil é o quarto país com o maior número de incidentes de trabalho no mundo com um registro de 16.455 mortes e 4.5 milhões de acidentes, em um período de 2012 a 2018, com uma estimativa de um óbito a cada 3h40 e de um acidente de trabalho a cada 48 segundos. (Fundacentro, 2019; ANAMT, 2018).

A baixa cultura de segurança de uma companhia é variável de grande influência para a ocorrência de incidentes e acidentes na mesma. Por definição, cultura consiste nos valores que os membros de um determinado grupo mantêm, nas normas que eles seguem e nos bens materiais que eles criam (Giddens, 1991). Tais valores e normas descritos são aprendidos pela socialização - humanos aprendendo com os outros em grupos; portanto, a cultura é aprendida pela socialização, ou seja, líderes comprometidos com a segurança de seus empregados devem disseminar uma cultura de segurança a partir de valores e normativas claras e bem estabelecidas.

Diante a este fato, é crucial a abordagem efetiva de segurança no ambiente de trabalho e a criação de uma sistemática diante de uma situação de baixa cultura de segurança, retirando principalmente o caráter reativo perante incidentes e tomando medidas que verdadeiramente mitiguem os riscos associados ao trabalho antes que incidentes e acidentes ocorram.

Um incidente é, por definição, a ocorrência de qualquer evento inesperado que resulte em consequências desfavoráveis. Um incidente pode ocorrer na forma de um incêndio, derramamento químico e líquido, liberação de fumaça ou gases tóxicos, uma explosão, qualquer anormalidade na operação regular que possa danificar o equipamento ou o equivalente a uma lesão reportável, tornando-se um acidente (Haight et al. 2001a).

Para estabelecer uma sistemática de gerenciamento de acidentes e a promoção de um sistema de gestão em segurança eficaz é necessária a definição de indicadores de avaliação de performance em segurança. Nesse sentido, agências governamentais estabelecem parâmetros para definição de métricas que medem a performance em segurança de uma empresa. O estudo em questão adota um indicador reativo definido pela Administração de Segurança Ocupacional e Saúde (OSHA), a taxa total de incidentes reportáveis (TRIR – total recordable incidente rate), descrito em (1).

$$TRIR = \frac{\text{quantidade de incidentes reportáveis} \times 200.000}{\text{total de homens horas trabalhadas}} \quad (1)$$

De acordo com a OSHA, este indicador define a taxa de incidentes ocorridos por 200.000 horas de exposição humana. O fato de ser uma taxa total é que este leva em conta todos os tipos de contratados em uma empresa, tanto os trabalhadores sazonais, empregados *full time*, contratados supervisionados e não supervisionados. O enquadramento em incidente reportável segue as seguintes diretrizes definidas pela agência (OSHA, 1992):

- Morte;
- Dias afastados do trabalho;
- Afastamento ou transferência de cargo;
- Tratamentos médicos além de primeiros socorros;
- Perda da consciência;
- Lesão ou doença diagnosticada por um profissional da saúde licenciado.

A engenharia de segurança é a área de estudo que aborda questões relacionadas à segurança e prevenção de incidentes no ambiente de trabalho. De acordo com a resolução Nº 325, DE 27 NOV 1987 do CONFEA, são algumas das atribuições do engenheiro de segurança:

Art. 4º - As atividades dos Engenheiros e Arquitetos na especialidade de Engenharia de Segurança do Trabalho são as seguintes: 5- Analisar riscos, acidentes e falhas, investigando causas, propondo medidas preventivas e corretivas e orientando trabalhos estatísticos, inclusive com respeito a custos; 13- Elaborar planos destinados a criar e desenvolver a prevenção de acidentes, promovendo a instalação de comissões e assessorando-lhes o funcionamento; 17- Propor medidas preventivas no campo de Segurança do Trabalho, em face do conhecimento da natureza e gravidade das lesões provenientes do Acidente de Trabalho, incluídas as doenças do trabalho. (CREA, 1987)

Oyewole (2009) reforça a importância de mudança de mindset ao lidar com incidentes e o quanto é necessária a mudança de estrutura das ações de promoção de segurança, de uma percepção puramente sentimental para decisões baseadas em dados. Uma vez que os dados estão disponíveis cabe então propor alternativas de manuseio dos mesmos, apoiados por ferramentas estatísticas poderosas.

Além disso, o autor também destaca o quanto é importante a criação de medidas eficazes orientadas pelos dados e a necessidade de alocação efetiva de recursos em uma realidade de mercado cada vez mais competitiva onde muitas vezes a segurança não é protagonista.

Apesar de LINGARD et al. (2017) propor a utilização de indicadores preventivos para medição de performance em segurança, esse reforça a importância da métrica TRIR (total recordable incident rate) como variável dependente da sua análise devido a sua distribuição normal e o seu reporte ser feito de maneira sistemática e padronizada quando comparado a pequenos desvios relatados. Adicional a isto, como a métrica de homens horas trabalhadas é elevada, estatisticamente, o TRIR é considerado uma medida aceitável para medição de cultura de segurança.

Ao que se refere a reputação de uma empresa, os indicadores de segurança são grande indicativo de performance. A repercussão de lesões e doenças recorrentes no ambiente de trabalho pode causar queda nos lucros, multas, pressão de grupos ativistas além de exposição negativa na mídia. A prevenção de acidentes auxilia na manutenção de uma boa reputação com os clientes, trabalhadores, mídia e acionistas (HSE, 2019).

Nessa perspectiva, o presente trabalho visa empoderar o caráter preventivo na tratativa de acidentes através da inserção de manipulações estatísticas, predizendo o valor da métrica de segurança para o ano de modo a suportar a tomada de decisão efetiva mediante a possível ocorrência de acidentes.

1.1 Objetivos

1.1.1 Objetivo geral

O presente estudo tem como objetivo realizar a análise da série temporal da métrica de segurança TRIR, suavizada em média móvel dos últimos 12 meses, em um período de janeiro de 2016 a julho de 2019, para aplicação do modelo SARIMA na mesma, suportada pela metodologia Box-Jenkins, de modo a prever o valor da métrica no final do ano, no contexto operacional de 5 localidades de uma mesma indústria de produção de sementes.

1.1.2 Objetivos específicos

Tem-se como objetivos específicos:

- Realizar revisão bibliográfica a respeito da aplicação do modelo SARIMA tanto em aspectos gerais quanto em previsões de acidentes de trabalho;
- Analisar séries no domínio do tempo;
- Descrever o modelo SARIMA;
- Aplicar o método Box Jenkins para previsão da série temporal;
- Impulsionar a tomada de decisão baseada em dados e ações preventivas para evitar acidentes de trabalho.

2 DESENVOLVIMENTO

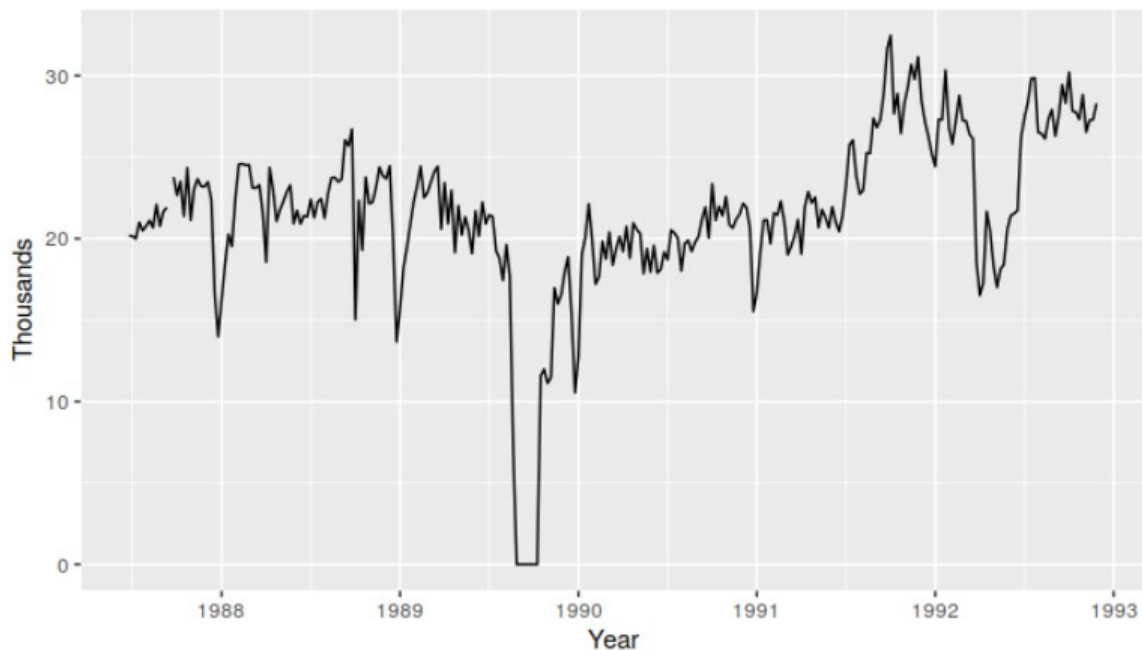
2.1 Fundamentação teórica

2.1.1 Análises de séries temporais

Uma série temporal é definida como uma observação de variáveis indexadas de acordo com a ordem em que estas são obtidas em diferentes pontos no tempo (SHUMWAY, 2000) e representam diversas problemáticas reais, como por exemplo: medição de qualidade do ar, números mensais de passageiros em uma companhia aérea, dados de vendas com indicadores preventivos, médias de temperaturas mensais em uma determinada cidade, entre outros (R Core Team, 2018).

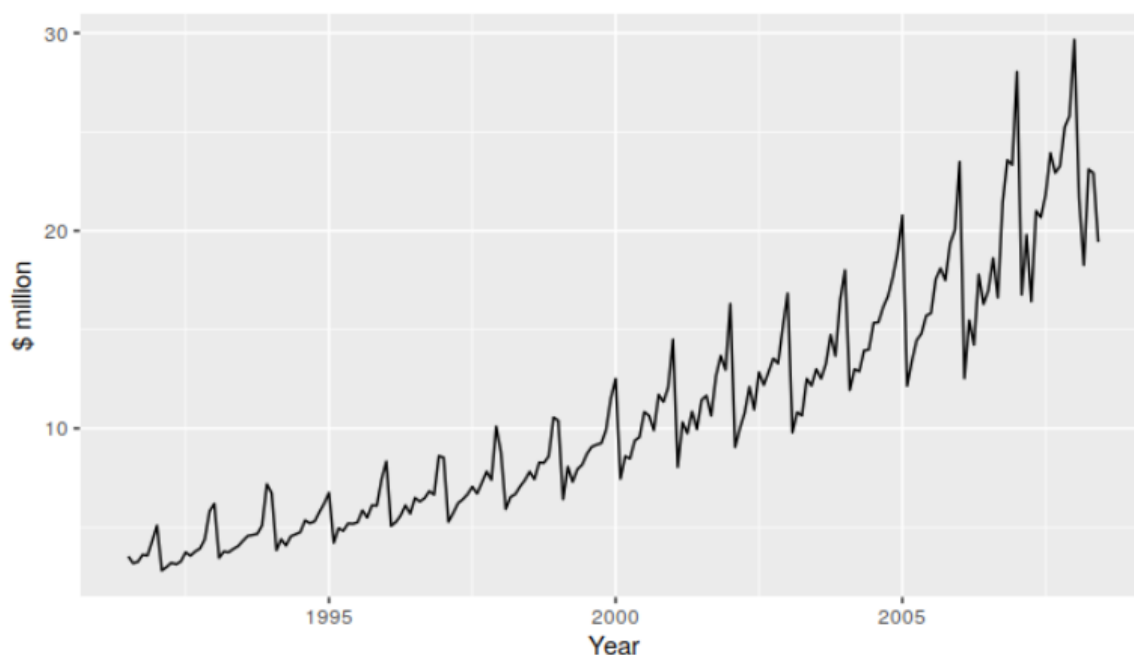
As Figuras 1 e 2 apresentam alguns exemplos de séries temporais

Figura 1 – Dados semanais de passageiros da classe econômica na companhia de voo Ansett.



Fonte: Hyndman, Athanasopoulos (2018)

Figura 2 – Dados mensais de venda de medicamentos diabéticos na Austrália.



Fonte: Hyndman, Athanasopoulos (2018)

Em ambos os casos é possível analisar as séries temporais e identificar padrões e comportamentos nas mesmas que são explicadas por fatores relacionados com cada tema. Por exemplo, analisando a Figura 1 percebe-se a existência de algumas grandes quedas de carga no início de cada ano causadas pelas férias além de uma flutuação de longo prazo no nível da série que aumenta em 1987, diminui em 1989 e aumenta novamente de 1990 a 1991. Já na análise da Figura 2, além de padrões claros de tendência crescente, há queda repentina no início de cada ano causada por um esquema de subsídios do governo que torna econômico para os pacientes o armazenamento de medicamentos no final do ano civil (HYNDMAN, 2018). Além disso, observa-se o padrão multiplicativo ao analisar como tendência e sazonalidade são proporcionais ao longo do tempo, ou seja, a componente sazonal tende a acompanhar o crescimento da tendência em séries multiplicativas o que não ocorre com séries aditivas, as quais componente sazonal possuem magnitude constante.

Tais análises só são possíveis e fundamentadas devido ao alto conhecimento da problemática, reafirmando a necessidade de um estudo prévio do tema de análise para entender como cada parâmetro no tempo responde à influência dos acontecimentos relacionados a questão.

Existem padrões em séries temporais que auxiliam fortemente na interpretação do seu comportamento. São essas:

- **Tendência:** Um comportamento de longo prazo de crescimento ou decréscimo nos dados, normalmente referido como mudança de direção quando esta vai de uma tendência de crescimento para uma tendência de decréscimo.
- **Sazonalidade:** Um padrão sazonal ocorre toda vez que o comportamento de uma série temporal é influenciado por um fatores sazonais como dia da semana, meses do ano, etc. Uma característica importante da sazonalidade é que esta é sempre fixa em uma frequência específica. A sazonalidade pode ser aditiva, onde as flutuações são mais ou menos constantes, ou multiplicativas, quando o tamanho das flutuações variam de acordo com o nível global da série.
- **Ciclicidade:** Um padrão cíclico refere-se a um comportamento de subida e descida que não são explicados por uma frequência definida.
- **Estacionariedade:** Padrão o qual não são observadas variações no tempo de média, variância e covariância.

Uma grande diferença nas séries temporais mostradas nas figuras acima são seus diferentes graus de suavização. Isso se relaciona fortemente ao fato de cada ponto no tempo ser auto correlacionado com o ponto anterior, ou seja, uma observação no tempo x_t é dependente dos valores observados em tempo $x_{t-1}, x_{t-2}, \dots$ (SHUMWAY, 2000)

A autocorrelação pode ser medida pelos coeficientes de autocorrelação, que variam de -1 (relacionamento negativo) a 1 (relacionamento positivo) expressados pela equação (2):

$$r_k = \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2} \quad (2)$$

Onde T é a quantidade de amostras da série temporal.

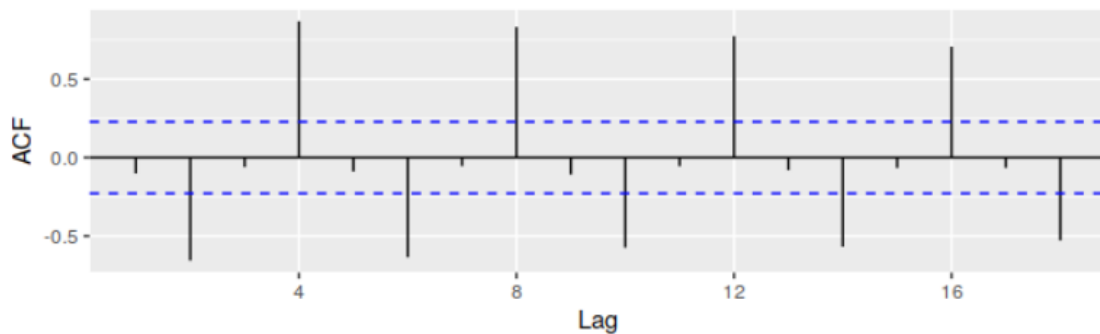
Uma boa forma de visualização de tais coeficientes é através das funções de auto correlação (ACF) plotadas em um correlograma, que nada mais é do que a comparação da série observada com ela mesma deslocada. A representação em um correlograma tem como eixo das abcissas os pontos dos coeficientes de auto correlação e como eixo das ordenadas os valores correspondentes aos mesmos e, auxilia na identificação dos comportamentos de tendência e sazonalidade em uma série.

Quando há tendência, quanto menor o lag, maiores e mais positivas são as auto correlações que decrescem lentamente conforme os lags aumentam. Já quando há comportamento sazonal, as

auto correlações serão grandes para lags correspondentes a sazonalidade e também para suas frequências múltiplas. A figura abaixo representa os coeficientes de uma série temporal representando a quantidade de cerveja produzida durante o ano.

Observa-se que r_4 possui maior valor, devido ao padrão sazonal dos dados, onde os picos e os vales estão separados por quatro trimestres. O valor mais negativo está em r_2 pois os vales estão dois trimestres atrás dos picos e as linhas azuis indicam quando as correlações são significativamente diferentes de zero.

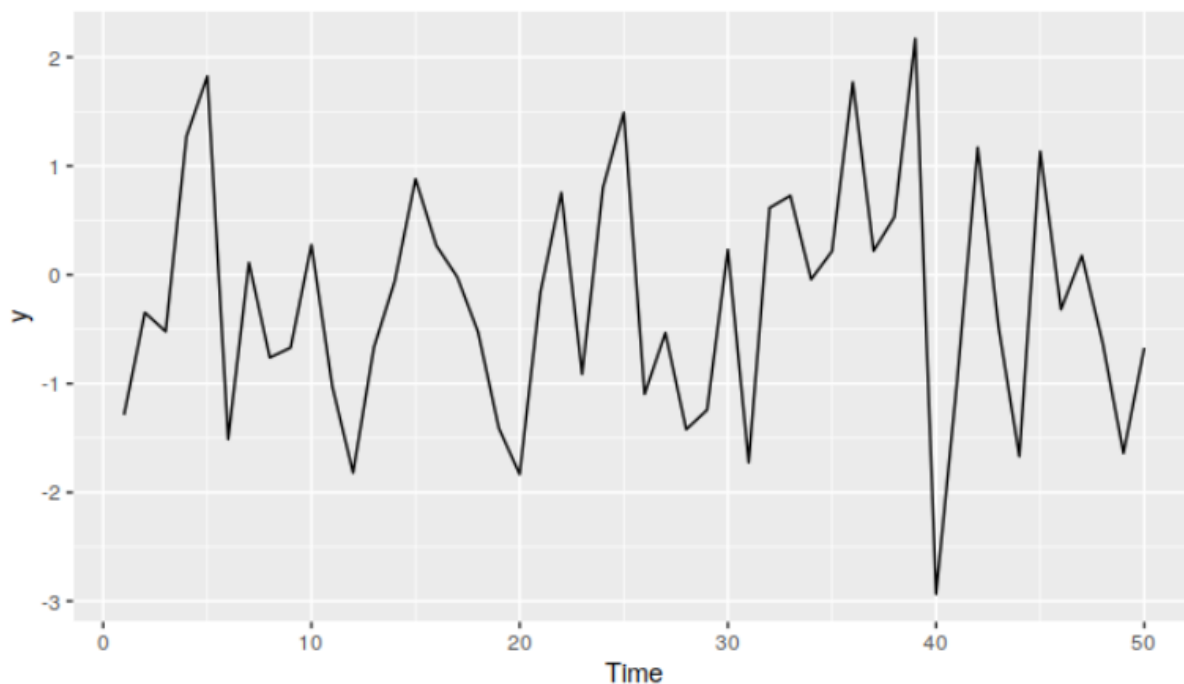
Figura 3 – Correlograma de uma série representando o comportamento de produção de cerveja durante um período do ano.



Fonte: Hyndman, Athanasopoulos (2018)

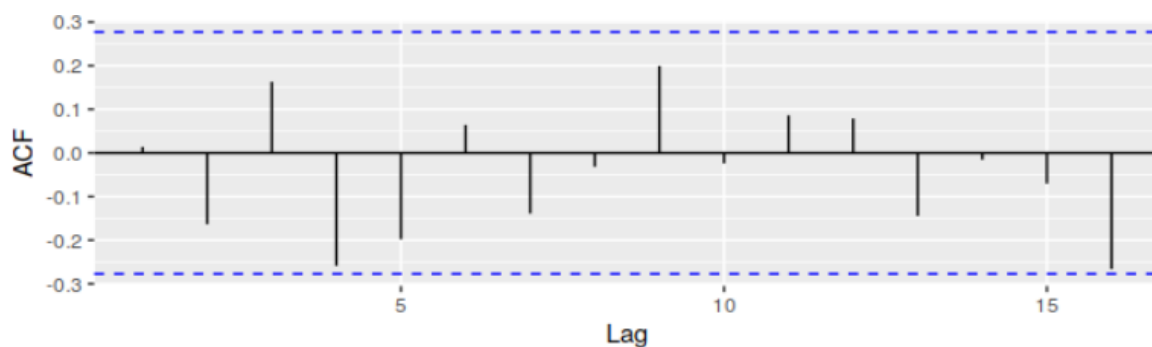
As séries temporais podem ser ainda uma série de variáveis randômicas e não correlacionadas, com média nula e variância finita. Estas são chamadas de ruído branco e a grande questão por trás da análise das mesmas como uma série temporal é a característica instável devido a inexistência de correlação entre os pontos adjacentes (SHUMWAY, 2000).

Figura 4 – Ruído branco.



Fonte: Hyndman, Athanasopoulos (2018)

Figura 5 – Função de auto correlação de uma série representando o comportamento do ruído branco da Figura 4.



Fonte: Hyndman, Athanasopoulos (2018)

A auto correlação esperada para ruídos brancos é quase nula, com os picos dos lags em $\pm \frac{2}{\sqrt{T}}$ e tal limite é exemplificado pelas linhas tracejadas azuis no gráfico. No exemplo acima, temos $T = 50$, e obtemos ± 0.28 para os coeficientes e como todos os picos estão neste intervalo pode-se afirmar que a série é ruído branco (HYNDMAN, 2018).

2.1.2 Modelo auto regressivo integrado de médias móveis (ARIMA) (p,d,q)

Devido ao fato das séries temporais possuírem acentuada correlação entre seus pontos adjacentes, as análises de tais dados não são satisfatoriamente bem contempladas por métodos estatísticos tradicionais que consideram os pontos observados como independentes e igualmente distribuídos. Nesse sentido, faz-se necessária a aplicação de métodos mais sofisticados para explicarem as dinâmicas entre as séries temporais, especialmente o fato de auto correlações serem explicadas por relacionamento lineares passados (SHUMWAY, 2000).

O modelo estatístico ARIMA – Auto regressivo integrado de médias móveis é uma ferramenta poderosa na análise de séries temporais correlacionadas não estacionárias. Para melhor compreensão de tal, é necessário abordar cada termo referente ao modelo.

2.1.3 AR (Auto regressivo) (p)

Os modelos auto regressivos são baseados na ideia de que valores atuais de uma série x_t podem ser explicados como função de p valores passados, $x_{t-1}, x_{t-2}, \dots, x_{t-p}$ onde p determina o número de parâmetros passados necessários para prever o valor atual.

Este pode ser descrito através da seguinte equação (3):

$$\phi(B)x_t = w_t \quad (3)$$

Onde, x_t é uma série estacionária, w_t é um ruído branco com média nula e variância finita, ϕ é constante e $\phi(B)$ é o operador auto regressivo, que por sua vez é definido como:

$$\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p \quad (4)$$

E, a função de auto correlação (ACF) para auto regressão é expressa por:

$$p(h) = \phi^h, h \geq 0 \quad (5)$$

2.1.4 I (Integrado) (d)

Em muitos casos reais, as séries temporais são não estacionárias. Essas apresentam comportamento de tendência ou possuem média, variância ou covariância variantes no tempo. A diferenciação é um método importante para torná-las estacionárias pois essa característica é essencial para obtenção de predições assertivas e efetivas. Mas, em muitos casos, as séries temporais podem ser decompostas em duas componentes: uma estacionária e outra não estacionária:

$$x_t = \mu_t + y_t \quad (6)$$

Onde, y_t é a componente estacionária e μ_t é a componente não estacionária ($\mu_t = \beta_0 + \beta_1 t$).

$$\nabla x_t = \beta_1 + \nabla y_t \quad (7)$$

Outras vezes, a estacionariedade é garantida com uma ordem maior de diferenciação. Por exemplo, suponha-se que $\mu_t = \mu_{t-1} + v_t$ e $v_t = v_{t-1} + e_t$

Diferenciando a equação 7, em termos das definições acima tem-se:

$$\nabla^2 x_t = e_t + \nabla^2 y_t \quad (8)$$

2.1.5 MA (Média móvel) (q)

Quando o ruído branco é combinado linearmente para a formação da série, tem-se o modelo de média móvel. Este é definido por:

$$x_t = w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q} \quad (9)$$

Onde, q representa os lags da média móvel e θ_q são o quadrado dos parâmetros. O operador de média móvel é descrito como:

$$\theta(B) = 1 + \theta_1 B - \dots - \theta_q B^q \quad (10)$$

E, a função de auto correlação (ACF):

$$p(h) = \begin{cases} \frac{\theta}{(1 + \theta^2)}, & h = 0 \\ 0, & h > 0 \end{cases} \quad (11)$$

2.1.6 Modelo sazonal auto regressivo integrado de médias móveis (SARIMA) (p,d,q)(P,D,Q)

É comum que a variabilidade de processos físicos, econômicos e biológicos tendem a ocorrer conforme uma flutuação sazonal. Nesse sentido, algumas séries temporais exibem um padrão periódico pois suas dependências tendem a ocorrer com maior significância em múltiplos de lags subjacentes (Shumway, 2000). Para esses casos, uma variação do modelo ARIMA é empregada para obtenção dos parâmetros sazonais, denominada SARIMA (sazonal auto regressivo integrado de médias móveis) e tem equação descrita por:

$$\Phi_P(B^S)\phi(B)\nabla_s^D\nabla_{x_t}^d x_t = \alpha + \Theta_Q(B^S)\theta(B)w_t \quad (12)$$

Onde, s é a ordem do componente sazonal, $\phi(B)$ é o componente autoregressivo de ordem p , $\theta(B)$ é o componente de média móvel de ordem q , $\Phi_P(B^S)$ é o componente sazonal autoregressivo de ordem P , $\Theta_Q(B^S)$ é o componente sazonal de média móvel de ordem Q e componentes ordinal e sazonal diferenciais $\nabla^d = (1 - B)^d$ e $\nabla_s^D = (1 - B^S)^D$, sucessivamente.

A análise dos coeficientes sazonais através das funções de auto correlação e auto correlação parcial devem ser descritas a partir da tabela a seguir:

Tabela 1 – Comportamento das funções de auto correlação e auto correlação parcial para cada parâmetro do modelo.

	AR(P)s			MA(Q)s		ARMA(P,Q)s		
ACF	Diminui	nos	lags	Intercepta	após o lag	Diminui	nos	lags
	múltiplos de s			Qs		múltiplos de s		

PACF	Intercepta após o lag Ps	Diminui nos múltiplos de s	lags	Diminui nos múltiplos de s	lags
------	--------------------------	----------------------------	------	----------------------------	------

2.1.7 Média absoluta percentual do erro (MAPE)

A média absoluta percentual do erro é uma medida de acurácia da previsão do modelo e mede o tamanho do erro em porcentagem. Esta medida não é afetada por valores extremos, utiliza percentuais do erro e não depende da unidade dos dados, ou seja, é adimensional. Quanto menor o MAPE, melhor é o modelo e sua equação é dada por:

$$MAPE = \frac{1}{T} \sum_{t=1}^T \frac{|y_t - \hat{y}_t|}{y_t} 100\% \quad (13)$$

Onde, t é o período do tempo, y_t é o valor observado no tempo t e $\hat{y}_t$ é o valor predito no tempo t .

2.1.8 Teste Kwiatkowski, Phillips, Schmidt and Shin (KPSS)

A estimação correta do modelo depende de uma análise realizada em séries temporais estacionárias. A não estacionariedade implica na representação pouco realística e em uma estimação mais difícil, devido a média e variância variarem ao longo do tempo (Manuca & Savit, 1996). O teste de KPSS é uma maneira de verificar a estacionariedade da série temporal, com hipótese nula de não estacionariedade e tem como representação matemática o seguinte:

$$KPSS = \frac{(T^{-2} \sum_{t=1}^T \hat{S}_t^2)}{\hat{\lambda}^2} \quad (14)$$

Onde, $\hat{S}_t^2$ é a somatória de resíduos de regressão de uma série temporal e $\hat{\lambda}^2$ é o termo consistente estimado da variância da componente de ruído branco.

2.1.9 Teste de Fisher

O teste de Fisher avalia a presença de sazonalidade na série temporal. Este é baseado na análise espectral dos componentes discretos e finitos estimados através da Transformada de Fourier Discreta e então avaliada a presença de sazonalidade através do periodograma. Tem-se como hipótese nula a não existência de sazonalidade na série e sua estatística é descrita por:

$$g = \frac{\max I_j^{(T)}}{\sum_{j=1}^T I_j^{(T)}} \quad (15)$$

Onde j é o índice de cada observação da transformada de Fourier Discreta e $I_j^{(T)}$ é o valor do periodograma em j .

2.1.10 Teste de Box-Pierce

A análise de um bom ajuste em uma série temporal tem como premissa a análise dos ruídos. Estes devem possuir comportamento de ruído branco, com média nula, variância constante e distribuição normal além de não possuir auto correlação. Para estimação do padrão de ruído branco, aplica-se o teste de Box-Pierce que possui hipótese nula de resíduos aleatórios independentes e identicamente distribuídos e possui a seguinte representação estatística:

$$Q(k) = T(T+2) \sum_{j=1}^k \frac{\hat{p}_j^2}{(T-j)} \quad (16)$$

Onde, k é o número de defasagem, p é a ordem do termo auto regressivo e q é a ordem do termo de médias móveis.

2.1.11 Critério de informação Akaike (AIC)

A representação de um evento real pode ser conseguida a partir da manipulação correta de parâmetros e a medida quantitativa de qual boa é essa representação é dada através de critérios de informação. Akaike (1974) avaliou e reconheceu a dificuldade de se construir um modelo adequado baseado em um número de dados finitos, como a análise de séries temporais, e propôs um critério baseado no log verossimilhança comparando a magnitude da função suporte maximizada $L(\hat{\theta}_i)$. O log verossimilhança é utilizado para expressar os valores ideais dos coeficientes estimados, mais plausível ao real.

Assim, o AIC é dado por:

$$AIC = -2 \log L(\hat{\theta}_i) + 2(n) \quad (17)$$

Onde, n é o número de parâmetros para serem estimados pelo modelo (Akaike, 1974).

Quanto menor o valor de AIC, em uma comparação de modelos ajustados em um mesmo período amostral, melhor é o modelo.

2.1.12 Critério de informação Bayesiano (BIC)

Schwarz (1978) propôs uma alternativa a abordagem de Akaike e tem como pressuposto a existência de um “modelo verdadeiro” que descreve a relação entre a variável dependente e as diversas variáveis explanatórias entre os diversos modelos sob seleção. Assim o critério é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados. O valor do critério BIC para um determinado modelo é dado por:

$$BIC = -2 \log f(x_n|\theta) + n \log T \quad (18)$$

Onde, $f(x_n|\theta)$ representa o modelo escolhido e T o número de observações da amostra.

Semelhante ao AIC, quanto menor o valor de BIC melhor é ajuste do modelo.

2.2 ARIMA como modelo de predição

Utilizando a metodologia de proposta por Box & Jenkins para o modelo ARIMA, o autor Svs (2018) propôs a estruturação de um programa de prevenção através da predição do número total de acidentes fatais que ocorrerão no período 2014 a 2019 em fábricas na Índia, analisando os dados em um período de 1980 a 2013. O autor propõe uma melhor alocação de recursos e uma forma de avaliação da efetividade das iniciativas de segurança tomadas.

O autor também compara dois modelos de predição estimando os coeficientes auto regressivo e de média móvel, escolhendo o modelo que representa ambos coeficientes menores que o intervalo de confiança (α) e plotando o correlograma dos resíduos. Ao final, o autor compara o valor real de incidentes com a previsão em um período conhecido e mostra a proximidade de ambos valores, dentro de um erro definido pelo mesmo.

Sarr (2015) propôs uma comparação entre o consumo de álcool e a taxa de incidentes veiculares relacionados com o mesmo na Estônia. O autor adota os dados temporais em um período de 1998 a 2013 e utiliza como método de regressão o modelo ARIMA para explicar as correlações nos resíduos além de utilizar os dados de incidentes veiculares não relacionados com consumo de álcool como argumento contra factual para verificação de robustez no resultado. Os resultados mostram uma significância fortemente negativa entre as duas variáveis onde, por exemplo, com o ARIMA estimado para (0,1,1)(0,1,1) um aumento de uma unidade no consumo de álcool é

associado com um decréscimo de 1.6% no nível dos acidentes a cada 100.000 motoristas com sinais de alcoolismo.

Fattah et. al. (2018) aplica o modelo ARIMA na predição de demanda para os próximos 10 meses em uma indústria alimentícia de forma a melhorar a tomada de decisão de gestores ao criarem um planejamento de negócio a partir de tais informações, melhorando seu sistema de gestão. Os dados foram coletados de 2010 a 2015 em uma indústria de Marrocos e métodos como erro padrão e critério Akaike (AIC) foram aplicados como critérios de performance do modelo escolhido ARIMA (1,0,1).

Brito (2019) estuda o comportamento temporal dos atendimentos a pacientes cardíacos além de atendimentos em pronto socorro da Rede de Assistência à Saúde de um município mineiro e faz uma previsão para os próximos 6 meses da demanda dos mesmos, possibilitando aos gestores tomada de decisão assertiva no contexto de elaboração de medidas, planejamento e avaliação de políticas públicas. Os dados foram colhidos de janeiro de 2014 à março de 2019 e o melhor modelo estimado foi ARIMA(3,1,1) para a previsão de atendimentos cardiológicos e ARIMA(2,1,1) para o atendimento de urgência e emergência.

2.3 Metodologia

2.3.1 Característica dos dados

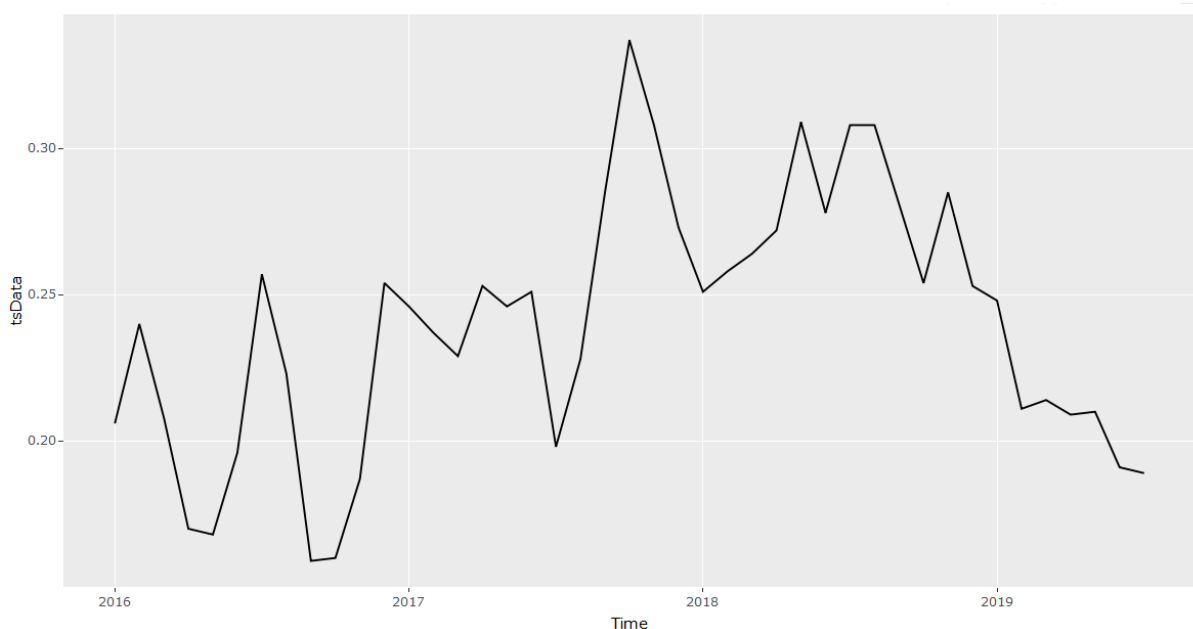
Duas características principais do indicador TRIR devem ser sinalizadas para maior efetividade no manuseio do mesmo. A primeira consiste na representatividade do mesmo em relação ao seu espaço amostral, ou seja, como o valor do indicador é inversamente proporcional as homens horas trabalhadas a análise do mesmo se torna relevante quando considerado um contexto operacional como um todo e não apenas uma localidade isolada, pois quanto menor a HHT, maior o indicador se torna e interpretações errôneas podem ser tiradas disso devido ao fato de metas isoladas por site não serem estabelecidas pelo negócio. Isso justifica a consolidação dos dados considerando 5 sites de sementes em localidades distintas, porém com o mesmo contexto de operação de manufatura.

A outra característica é a nulidade do indicador na maioria dos meses pois não é de praxe que ocorram acidentes reportáveis todos os meses, no contexto operacional dos 5 sites. Como este fato gera uma série temporal praticamente nula ao longo do tempo, foi adotada como técnica de pré-processamento a média móvel dos últimos 12 meses, ou seja, os pontos da série são representados por:

$$x(t) = \frac{\sum n^{\circ} \text{ de incidentes reportáveis últimos 12 meses} * 200.000}{\sum \text{homem horas trabalhadas últimos 12 meses}} \quad (19)$$

Exemplificando, o ponto janeiro de 2016 é representado pela somatória do número dos incidentes reportáveis de janeiro 2015 a dezembro de 2015 sobre a somatória dos últimos 12 meses da hht do mesmo período multiplicado por 200.000.

Figura 6 – Média móvel últimos 12 meses TRIR de janeiro de 2016 a julho 2019 para 5 sites de manufatura de sementes.



Para fins de comparação da predição do modelo com valores reais, foi realizada a separação do dataset entre teste e validação. A série temporal foi dividida em variável teste, que contempla dados de janeiro de 2016 a dezembro de 2018 e em variável validação, contemplando dados de janeiro de 2019 a julho de 2019. Três modelos foram ajustados e então comparados através da metodologia a seguir:

1. Ajuste na variável de teste, levantando os critérios de informação AIC e BIC
2. Comparação do ajuste com a respectiva variável de teste;
3. Previsão utilizando a variável de teste;
4. Comparação dos valores previstos da variável de teste com os valores da variável de validação;
5. Previsão utilizando o dataset original;

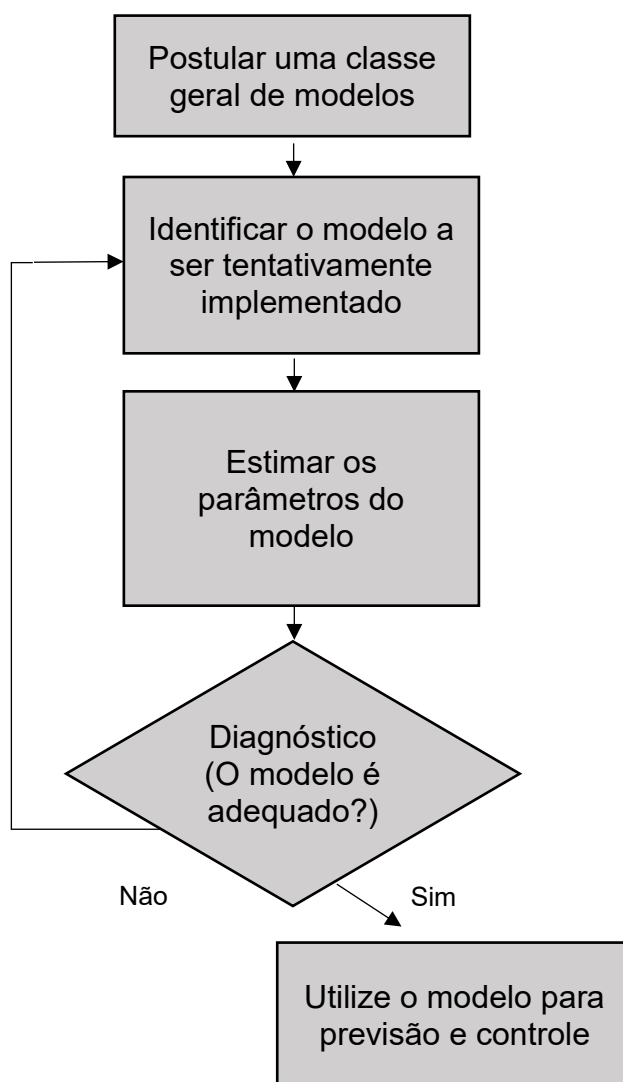
6. Consolidado da média absoluta percentual do erro e análise dos resíduos para os passos anteriores.
7. Comparação dos três modelos através dos critérios de informação, MAPE e análise residual.

Todas as análises foram realizadas na linguagem de programação R, na IDE *RStudio*.

2.3.2 Aplicação do método Box & Jenkins

BOX; JENKINS; REINSEL (2013) propuseram um método a ser seguido para construir modelos para previsão e controle, base para a estruturação deste trabalho:

Esquema 1 – Passos para construção de um modelo.



Baseado no esquema acima, tem-se primeiramente a fase de identificação do modelo onde analisa-se a característica da série determinando comportamentos estacionários, sazonais e são determinados também os coeficientes do modelo (p,d,q) através das funções de autocorrelação e

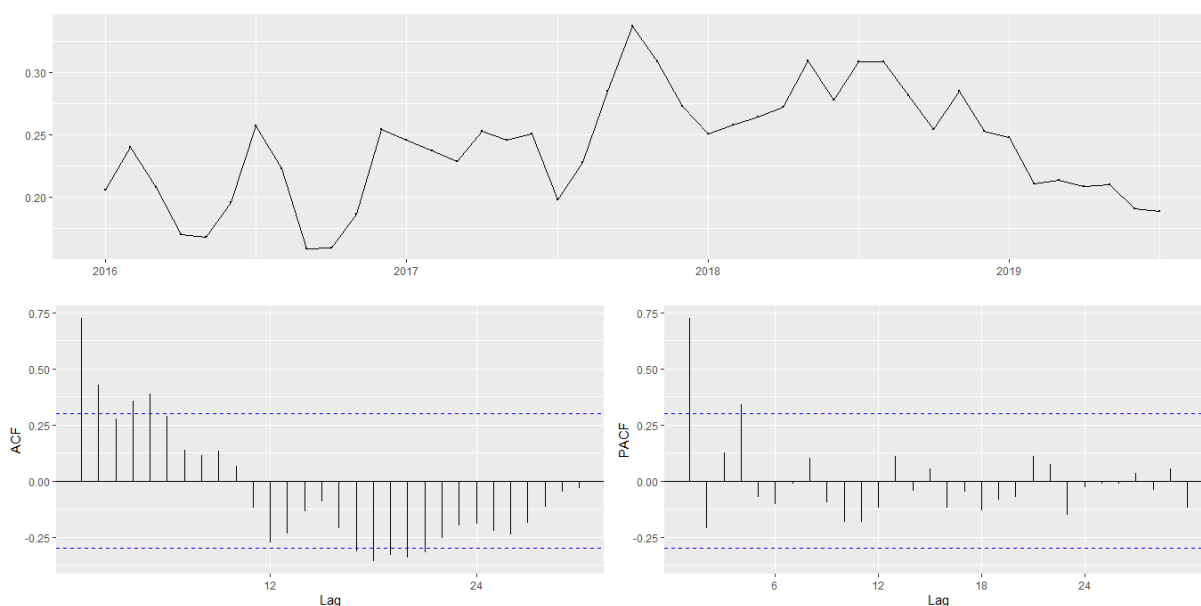
autocorrelação parcial. Uma vez identificado, é necessário estimar os parâmetros autoregressivo e de média móvel, seja através da separação dos dados de validação ou também pelo diagnóstico dos resíduos.

Uma vez comprovada a adequabilidade do modelo, realiza-se a previsão comparando modelos estimados através de medidas de acurácia como por exemplo o MAPE. (Morettin, 1981)

3 RESULTADOS

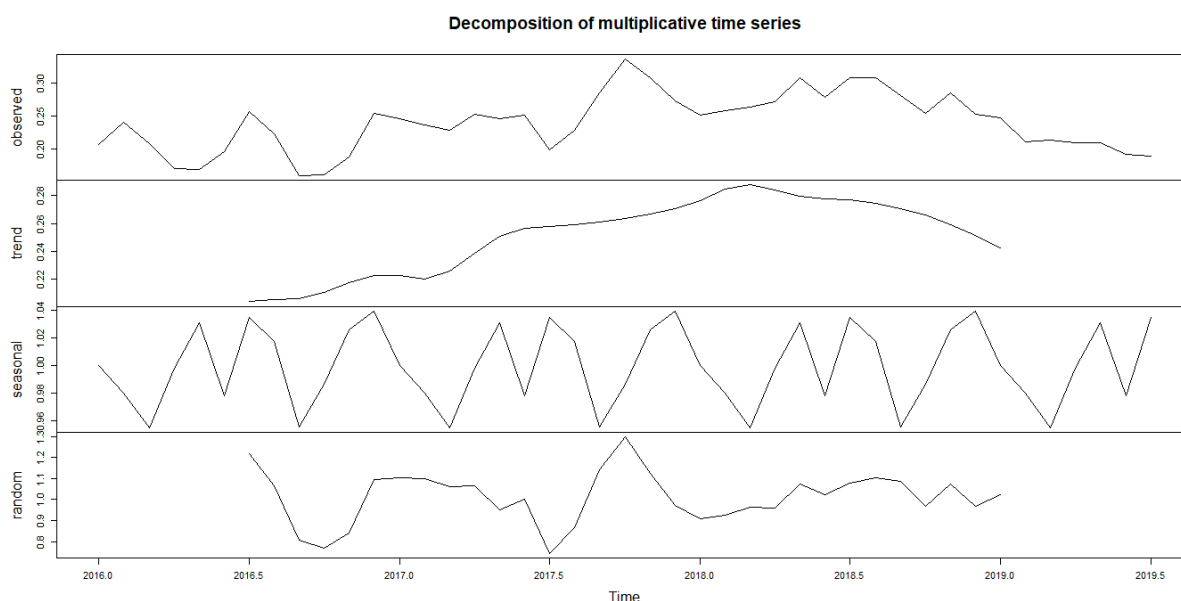
Com base na metodologia de Box & Jenkins, foram adotadas técnicas para melhor entendimento do comportamento da série e assim estimar a melhor manipulação estatística para a mesma. A priori a série é plotada juntamente com sua função de auto correlação e auto correlação parcial. Observa-se um comportamento relativamente lento de decréscimo nos lags da função de auto correlação e um pico no lag 1, indicando não estacionariedade. (Shumway, 2000)

Figura 7 – Análise da série original, suas funções de auto correlação e auto correlação parcial.



Foi realizada a plotagem do gráfico através da função *decompose* para decomposição multiplicativa, devido a característica da magnitude da sazonalidade acompanhar o crescimento ou decréscimo da tendência.

Figura 8 – Série temporal analisada pela função decompose.

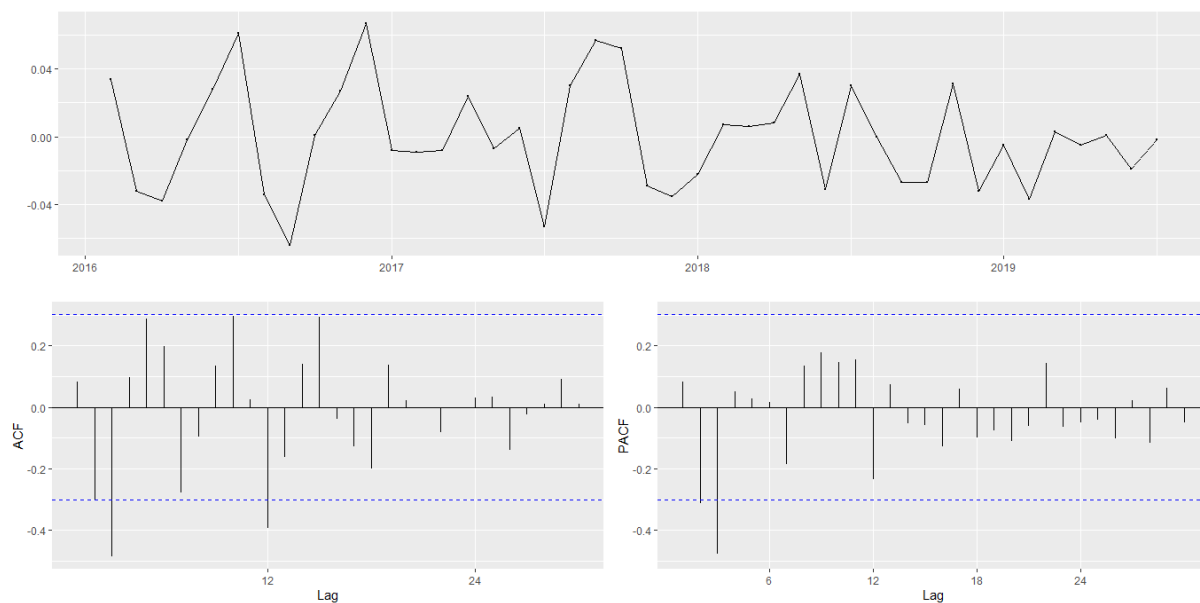


Através desta função, observa-se a priori um comportamento de tendência. Este comportamento, além da sazonalidade, também são grandes indicativos de séries não estacionárias (Morettin, 1981). Para embasar a observação, tomou-se o teste de KPSS onde a hipótese nula (a série observada é estacionaria) é rejeitada com o p-value = 0,01, ao nível de significância de 5%.

Ao observar a componente sazonal conta-se 10 pontos até a próxima repetição caracterizando assim um padrão repetitivo baseado em uma frequência, vide definição de sazonalidade. Realiza-se o teste de Fisher, tem-se p-value = $3,32^{-0.7}$, rejeitando a hipótese nula de ausência de sazonalidade.

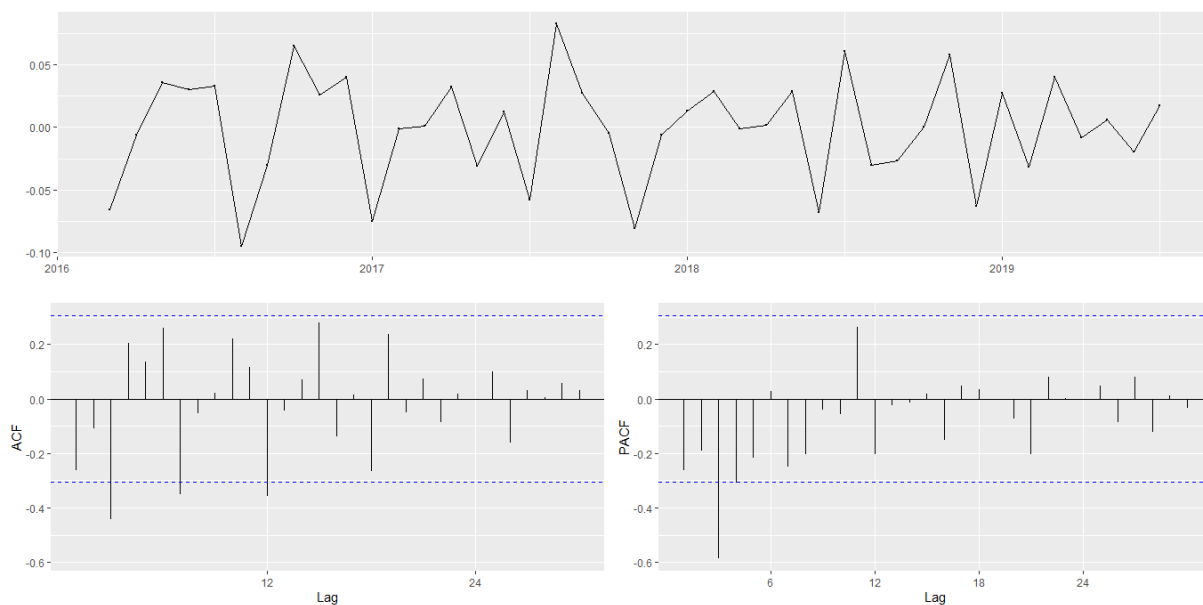
Comprovada a ausência de estacionariedade além da presença de sazonalidade, vide metodologia Box-Jenkins, é identificado o modelo SARIMA (p,d,q)(P,D,Q). Com o intuito de tornar os dados estacionários, é realizada a primeira diferenciação da série e então é repetido o teste de KPSS. Tem-se agora um p-value = 0,10, acima de $\alpha = 0,05$ que é traduzido na hipótese nula não rejeitada, ou seja, a série torna-se estacionária. Entretanto analisando o gráfico da série temporal, bem como a função de autocorrelação, observa-se que uma diferenciação apenas não foi o suficiente pois a média da série apresenta comportamento variado ao longo do tempo.

Figura 9 - Série temporal com uma diferenciação.



E então é realizada uma segunda diferenciação e realizada a plotagem da série após bem como seus gráficos de autocorrelação e autocorrelação parcial.

Figura 10 – Série temporal com duas diferenciações.



Uma vez identificado os padrões da série, realizando análises de estacionariedade e sazonalidade é necessário estimar os parâmetros. Analisando série, esta apresenta pico significativo no lag 3 em sua função de autocorrelação e pico no lag 3 e 11 também na função de autocorrelação parcial.

Como ponto de partida, tomou-se o modelo SARIMA(1,2,1)(1,1,1) para o intervalo de teste e obtém-se o gráfico da série (em cinza) e o ajuste (azul). É realizado a predição para os próximos 7 meses para fins comparativos dos dados de validação.

Figura 11 – Plotagem da variável de teste (cinza) e ajuste (azul) SARIMA(1,2,1)(1,1,1).

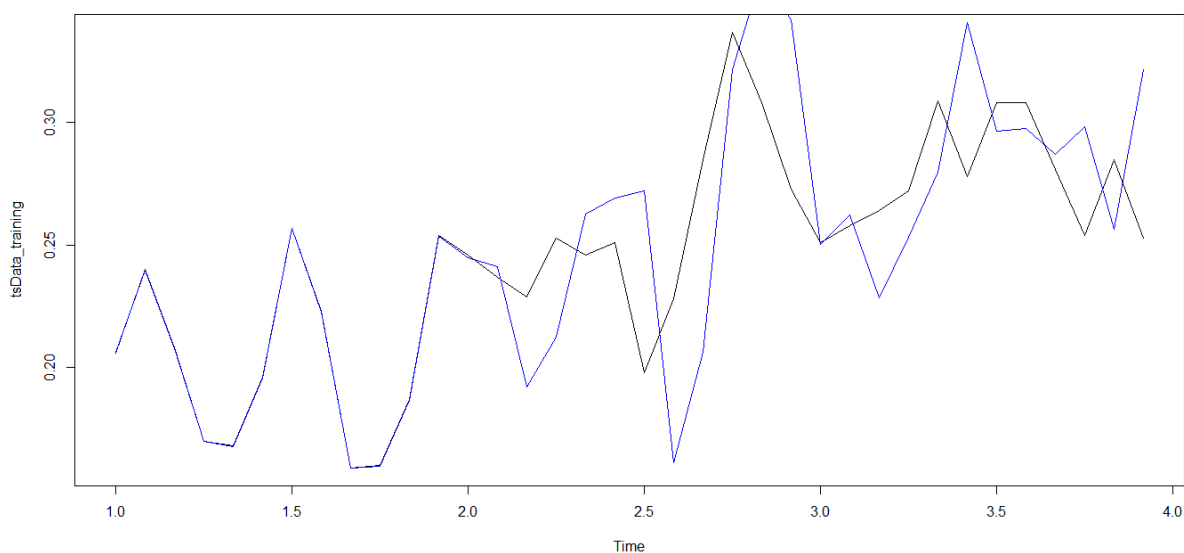


Tabela 2 – Coeficientes do ajuste da variável de teste SARIMA(1,2,1)(1,1,1).

	Estimado	Erro padrão	Z value	Pz z
ar1	0,152615	0,222662	0,6854	0,4931
ma1	-0,999226	0,184631	-5,4120	6,232e-08
sar1	-0,647543	NA	NA	NA
sma1	-0,062754	NA	NA	NA

Figura 12 – Predição utilizando a variável de teste SARIMA(1,2,1)(1,1,1).

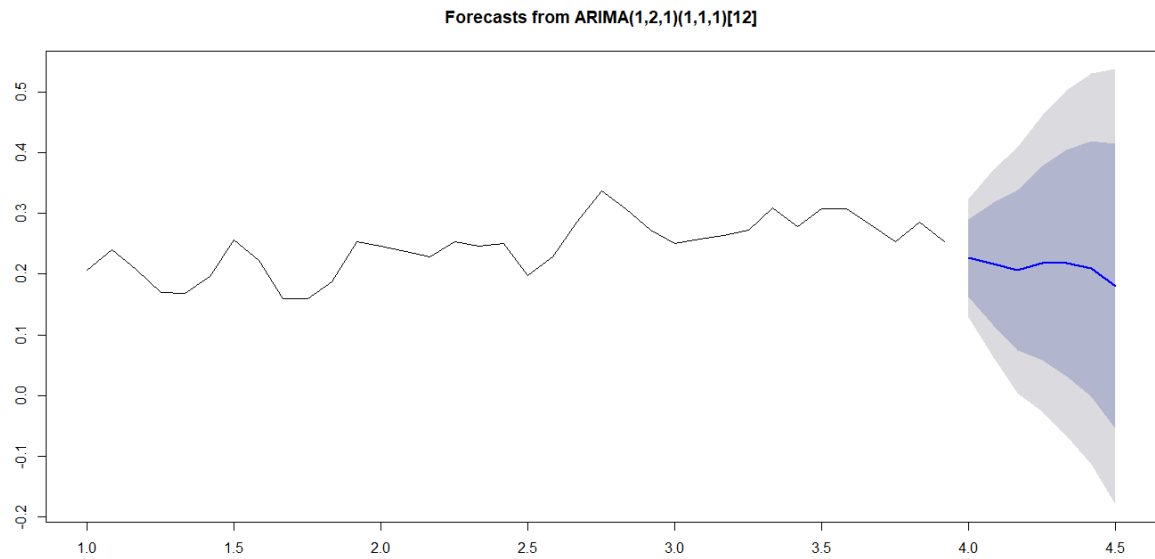


Figura 13 – Plotagem dos valores de predição da variável de teste (cinza) e valores de validação (azul) SARIMA(1,2,1)(1,1,1).

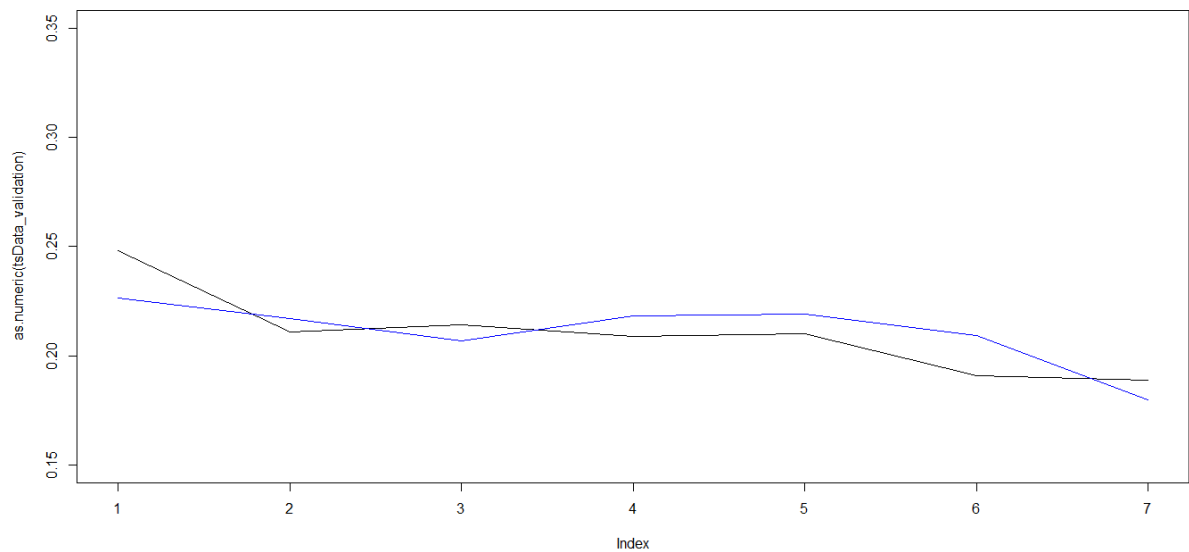


Tabela 3 – Coeficientes do ajuste na variável de teste SARIMA(1,2,1)(1,1,1).

AIC	BIC	MAPE ajuste	MAPE valores do ajuste e série teste	MAPE variável de validação e valores predição
-49,42	-43,97	8,61939	8,69273	5,71000

Tabela 4 – Previsão da série temporal de teste SARIMA(1,2,1)(1,1,1).

Mês	Dados da variável de validação	Resultado da predição
Janeiro	0,248	0,22632
Fevereiro	0,211	0,21715
Março	0,214	0,20675
Abril	0,209	0,21822
Maiο	0,210	0,21896
Junho	0,191	0,20929
Julho	0,189	0,18000

Figura 14– Análise de resíduos de teste SARIMA(1,2,1)(1,1,1).

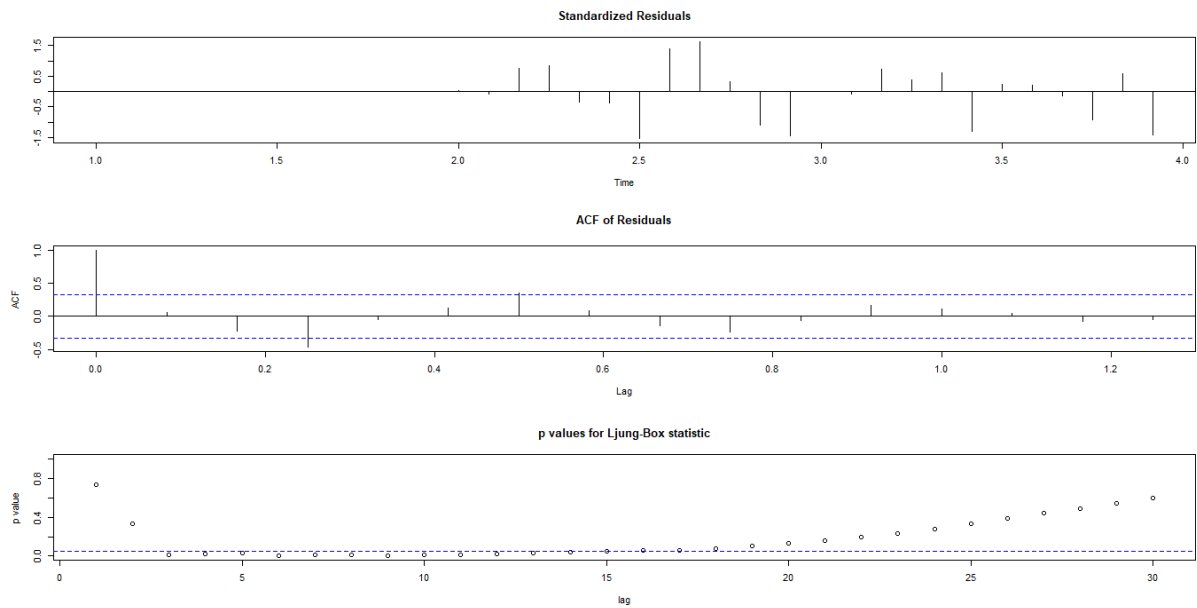
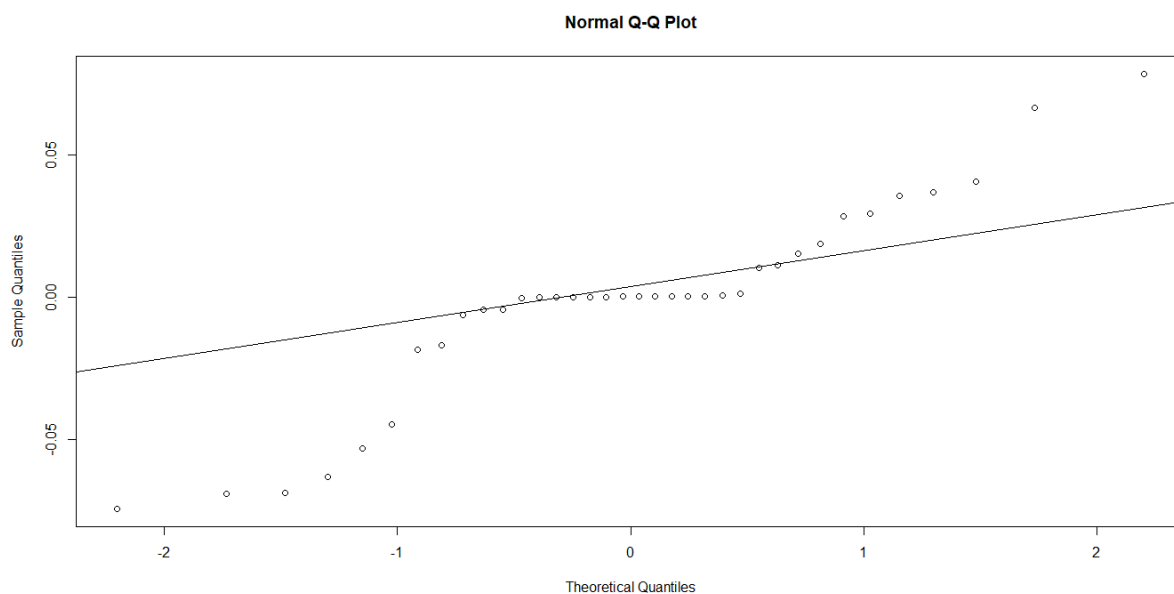


Figura 15 – Normal QQ Plot SARIMA(1,2,1)(1,1,1).



E então é realizada a previsão, utilizando-se a variável inicial, até dezembro de 2019. Com a análise de Box-Pierce para os resíduos obteve-se o p-value de 0,9094, aceitando a hipótese nula de resíduo com padrão de ruído branco.

Figura 16 – Previsão para próximos 5 meses utilizando SARIMA(1,2,1)(1,1,1).

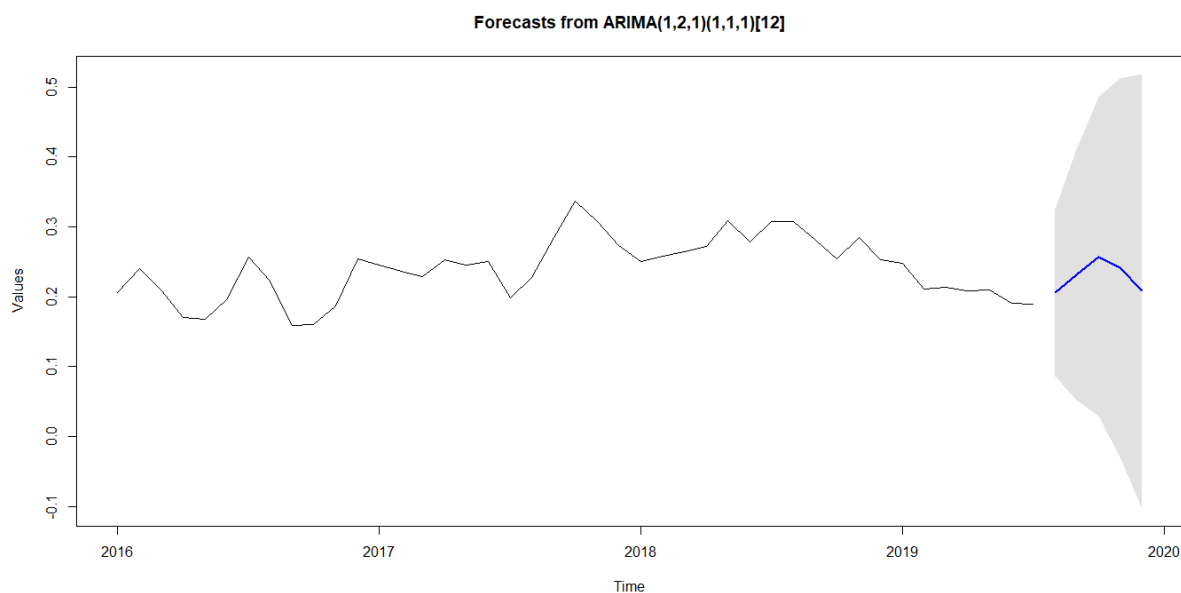


Tabela 5 – Resultados da previsão próximos 5 meses SARIMA(1,2,1)(1,1,1).

Mês	Resultado da previsão
Agosto	0,20604
Setembro	0,23173

Outubro	0,25715
Novembro	0,24143
Dezembro	0,20805

Figura 17 – Padrão dos resíduos da previsão próximos 5 meses SARIMA(1,2,1)(1,1,1).

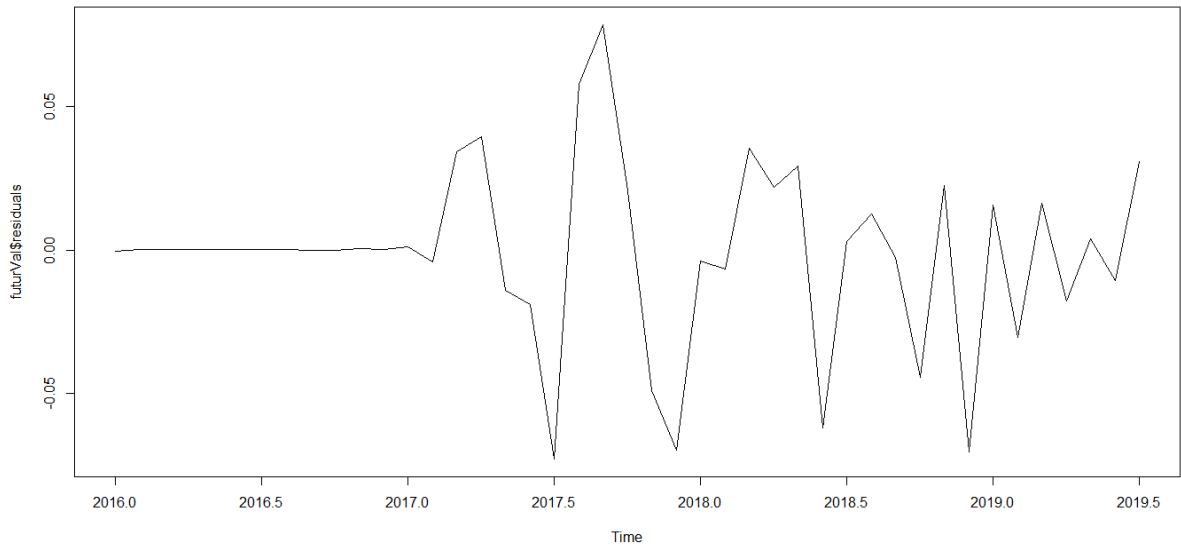


Figura 18 – Função de auto correlação para os resíduos SARIMA(1,2,1)(1,1,1).

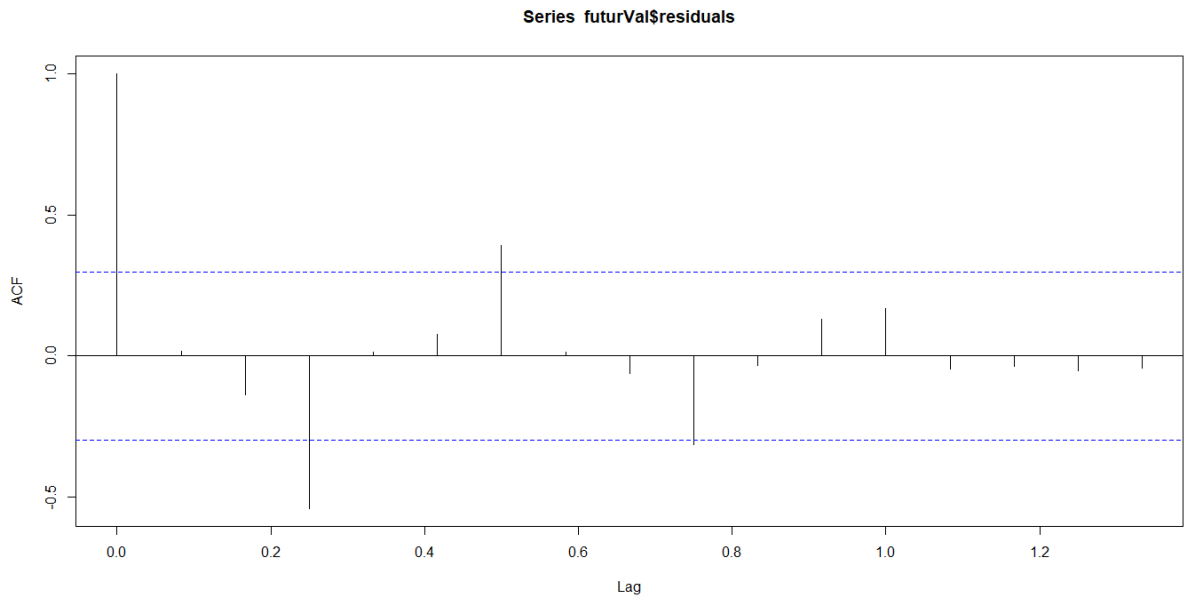
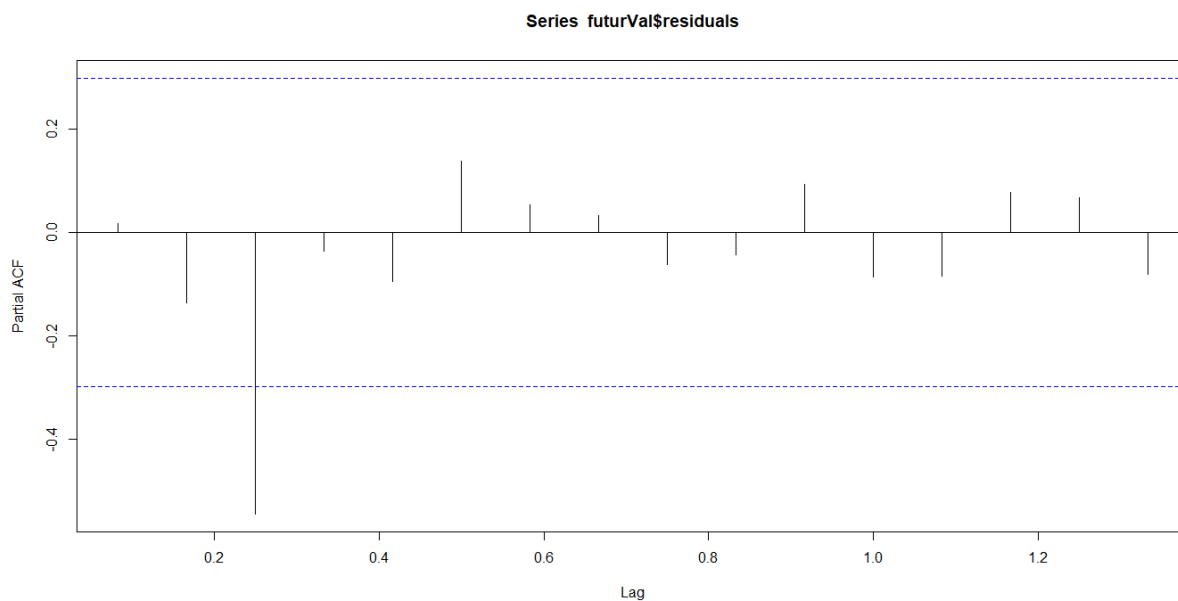


Figura 19 – Função de auto correlação parcial para os resíduos SARIMA(1,2,1)(1,1,1).



O segundo modelo proposto é SARIMA(3,1,0)(1,0,0).

Figura 20 - Plotagem da variável de teste (cinza) e ajuste (azul) SARIMA(3,1,0)(1,0,0).

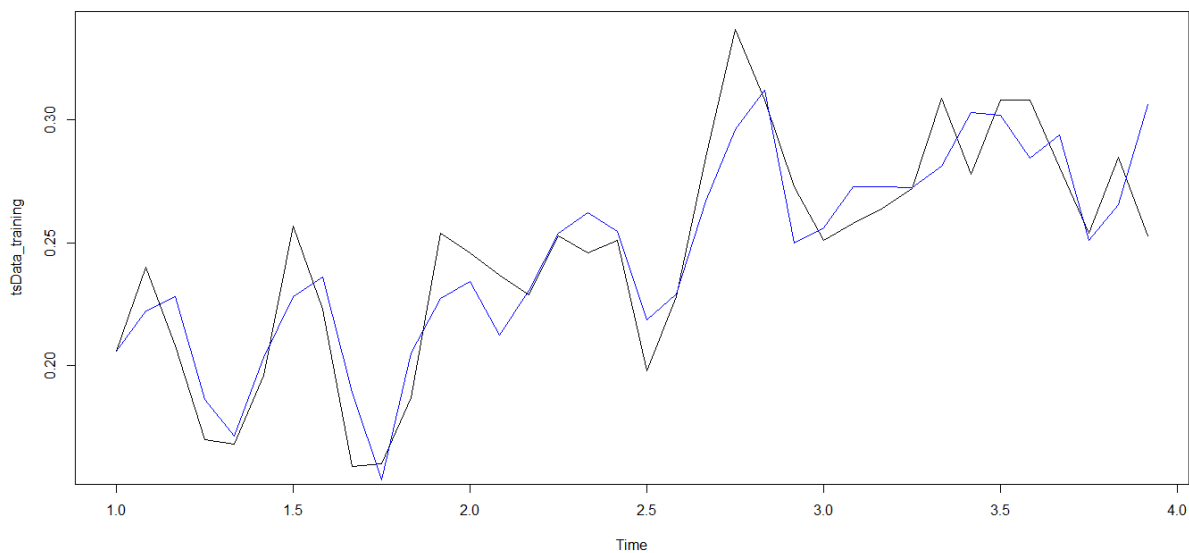


Tabela 6 - Coeficientes do ajuste da variável de teste SARIMA(3,1,0)(1,0,0).

	Estimado	Erro padrão	Z value	Pz z
ar1	-0,12837	0,15731	-0,81600	0,41447
ar2	-0,26872	0,13843	-1,94130	0,05223

ar3	-0,60846	0,16262	-3,74160	0,00018
sa1	-0,66818	0,13414	-4,98140	6,313e-07
Drift	0,00247	0,00126	1,96430	0,04949

Figura 21 - Predição utilizando a variável de teste SARIMA(3,1,0)(1,0,0).

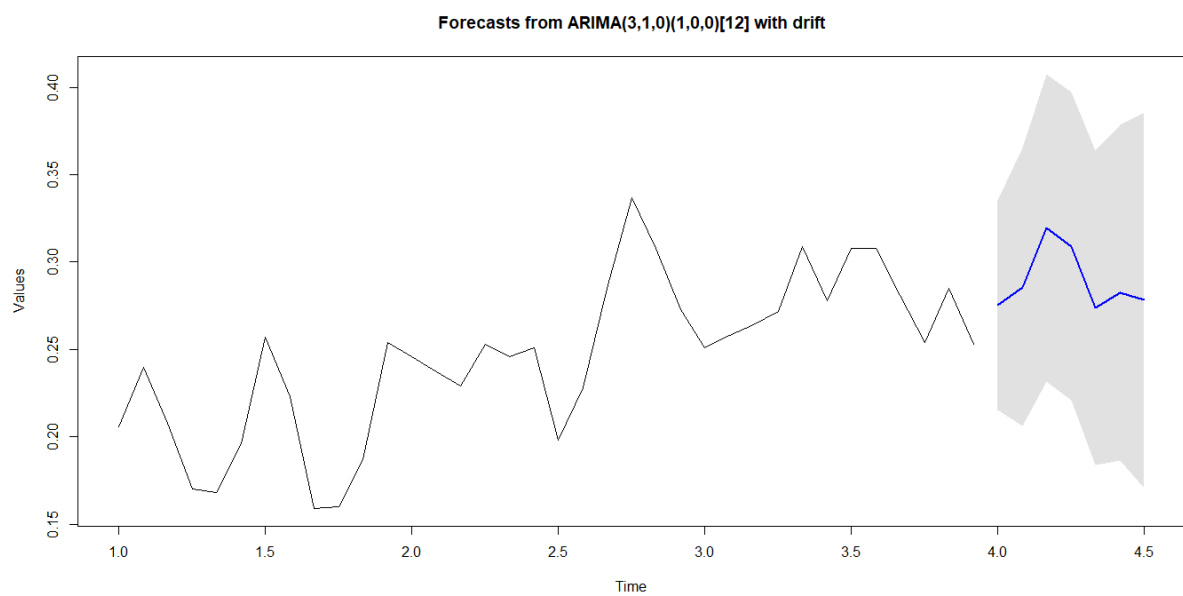


Figura 22- Plotagem dos valores de predição da variável de teste (cinza) e valores de validação (azul) SARIMA(3,1,0)(1,0,0).

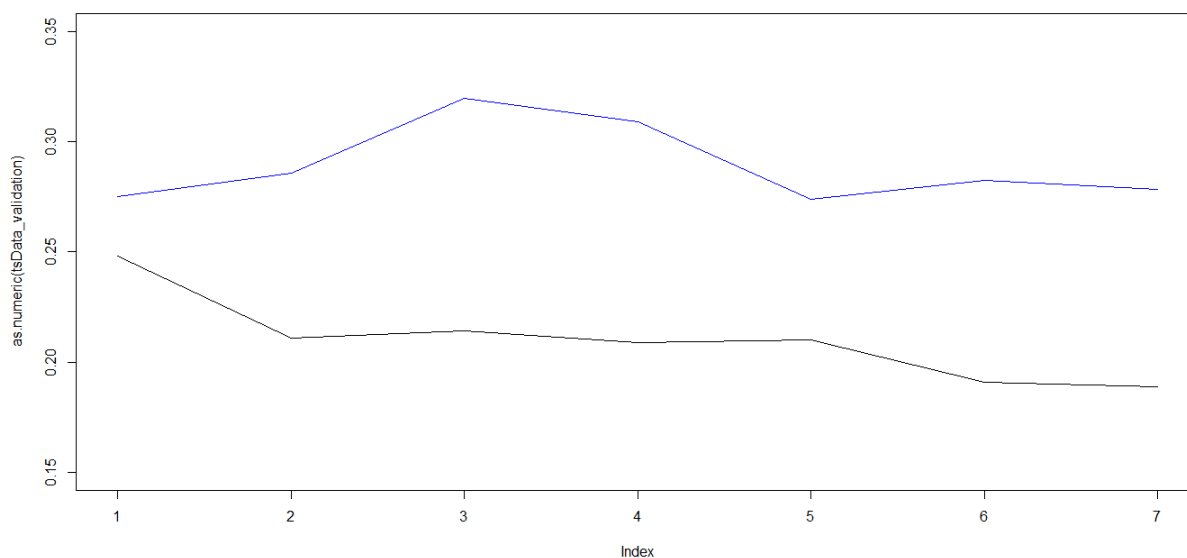


Tabela 7 – Coeficientes do ajuste na variável de teste SARIMA(3,1,0)(1,0,0).

AIC	BIC	MAPE ajuste	MAPE valores do ajuste e série teste	MAPE variável de validação e valores predição
-154,55	-145,22	6,35685	6,27790	27,03339

Tabela 8 – Previsão da série temporal de teste SARIMA(3,1,0)(1,0,0).

Mês	Dados da variável de validação	Resultado da predição
Janeiro	0,248	0,27523
Fevereiro	0,211	0,28566
Março	0,214	0,31965
Abril	0,209	0,30905
Maio	0,210	0,27386
Junho	0,191	0,28247
Julho	0,189	0,27825

Figura 23- Análise de resíduos de teste SARIMA(3,1,0)(1,0,0).

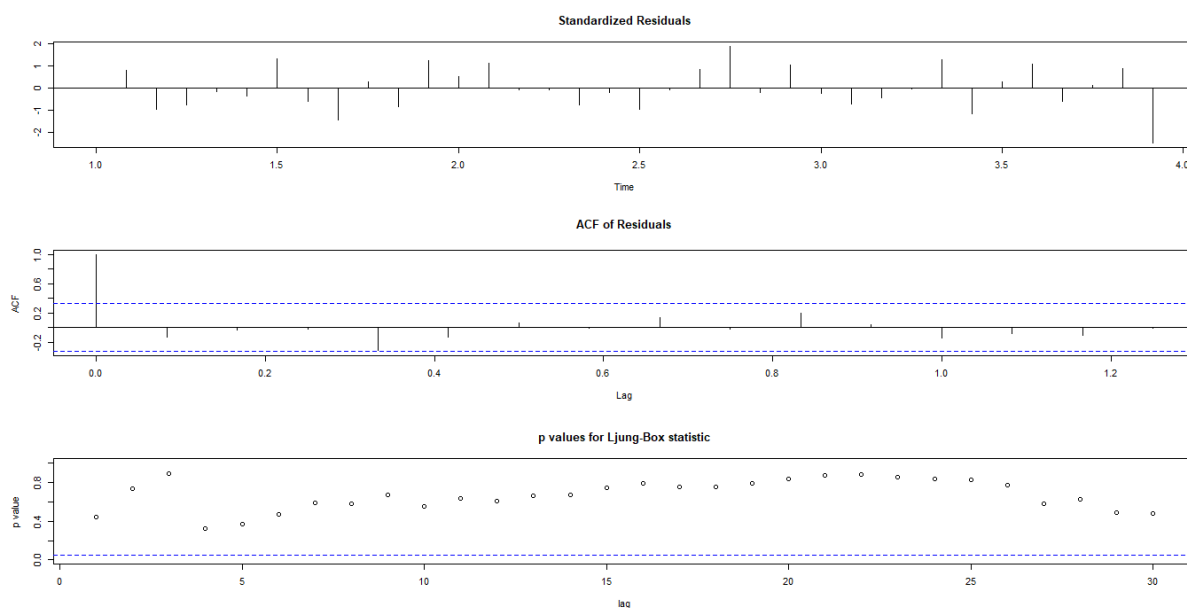
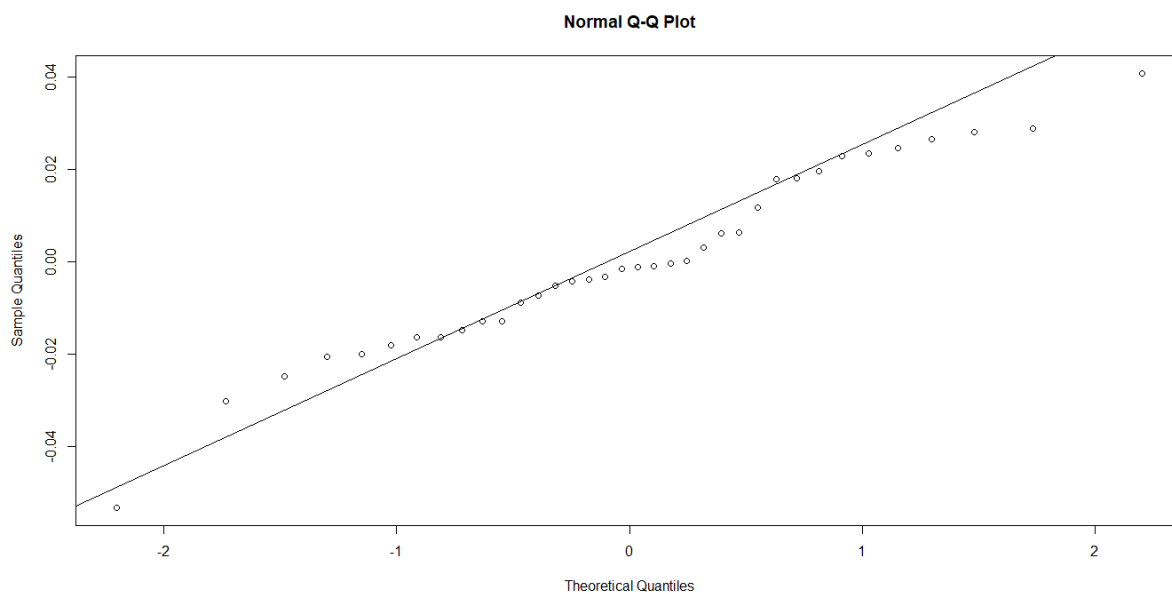


Figura 24 - Normal QQ Plot SARIMA(3,1,0)(1,0,0).



E então é realizada a predição, utilizando-se a variável inicial, até dezembro de 2019. Com a análise de Box-Pierce para os resíduos obteve-se o p-value de 0,8789, aceitando a hipótese nula de resíduo com padrão de ruído branco.

Figura 25– Previsão para próximos 5 meses utilizando SARIMA(3,1,0)(1,0,0).

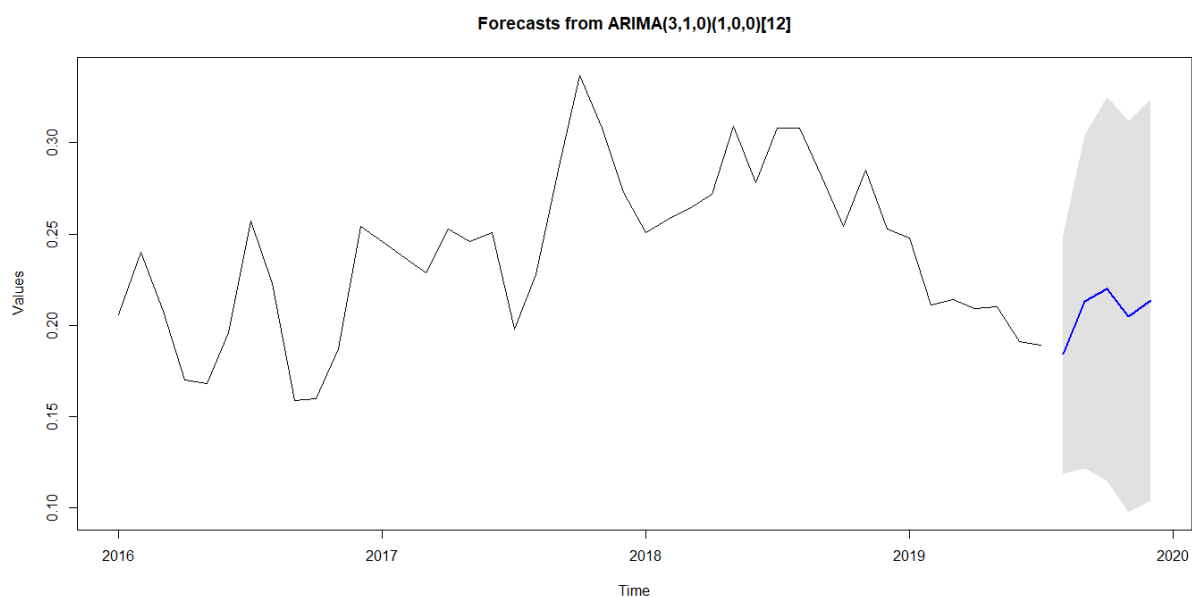


Tabela 9- Resultados da previsão próximos 5 meses SARIMA(3,1,0)(1,0,0).

Mês	Resultado da predição
Agosto	0,18425
Setembro	0,21292
Outubro	0,21989
Novembro	0,20486
Dezembro	0,21384

Figura 26 –Padrão dos resíduos do valor ajustado SARIMA(3,1,0)(1,0,0).

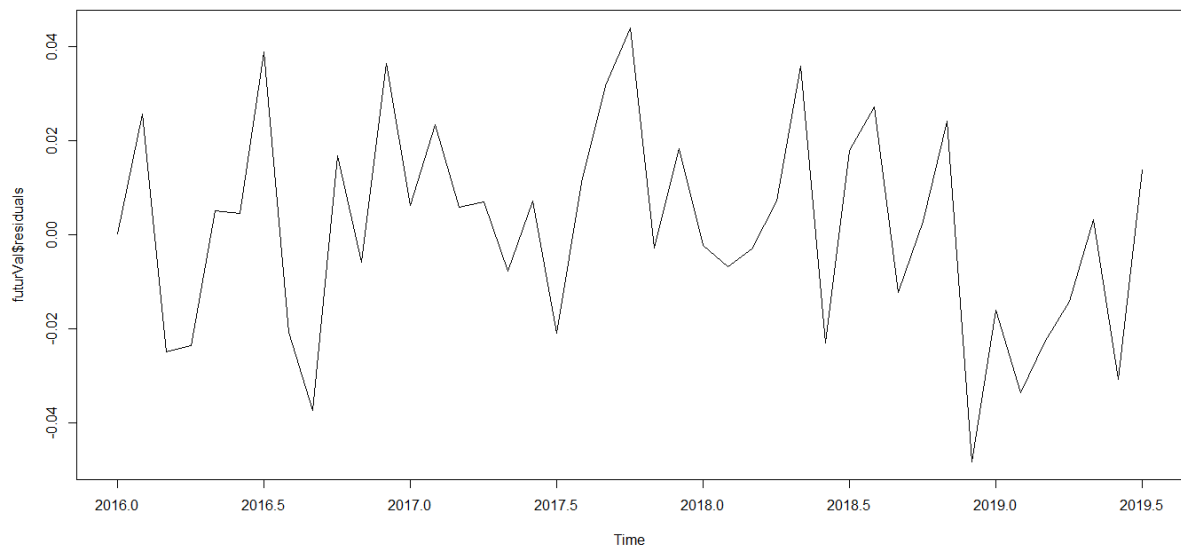


Figura 27 - Função de auto correlação para os resíduos SARIMA(3,1,0)(1,0,0).

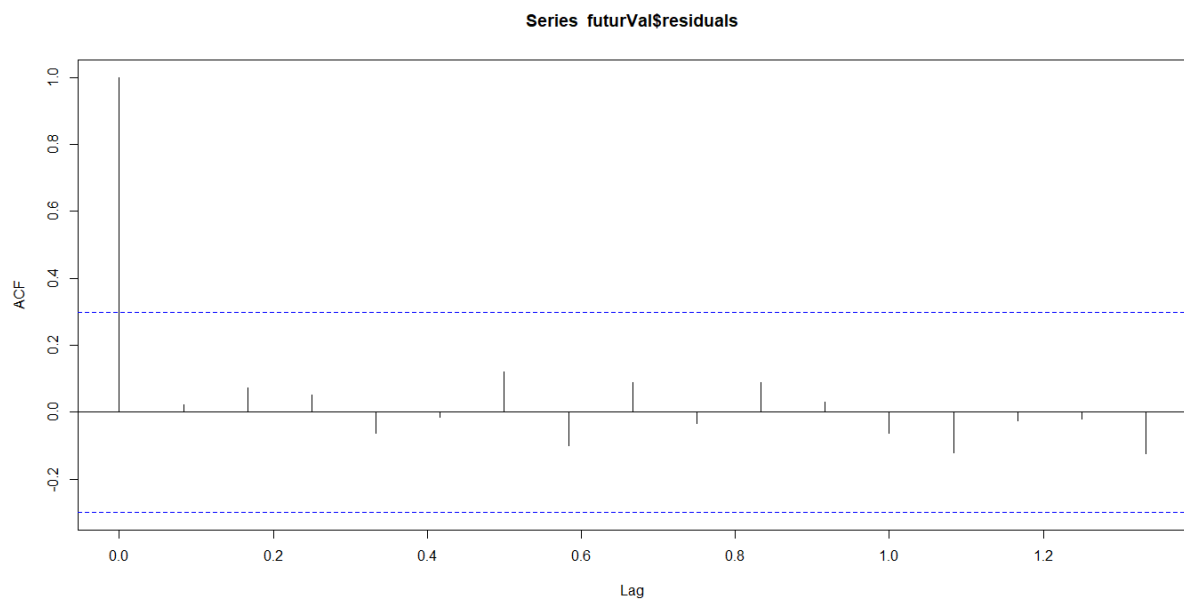
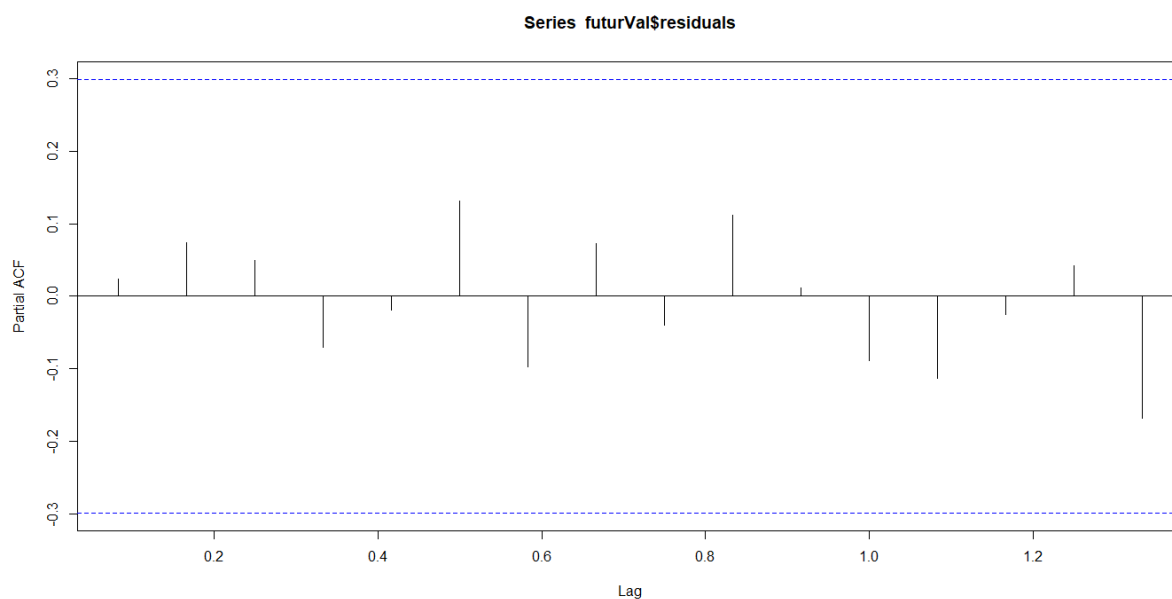


Figura 28 - Função de auto correlação parcial para os resíduos SARIMA(3,1,0)(1,0,0).



O terceiro modelo proposto é SARIMA(3,2,0)(1,0,0).

Figura 29 - Plotagem da variável de teste (cinza) e ajuste (azul) SARIMA(3,2,0)(1,0,0).

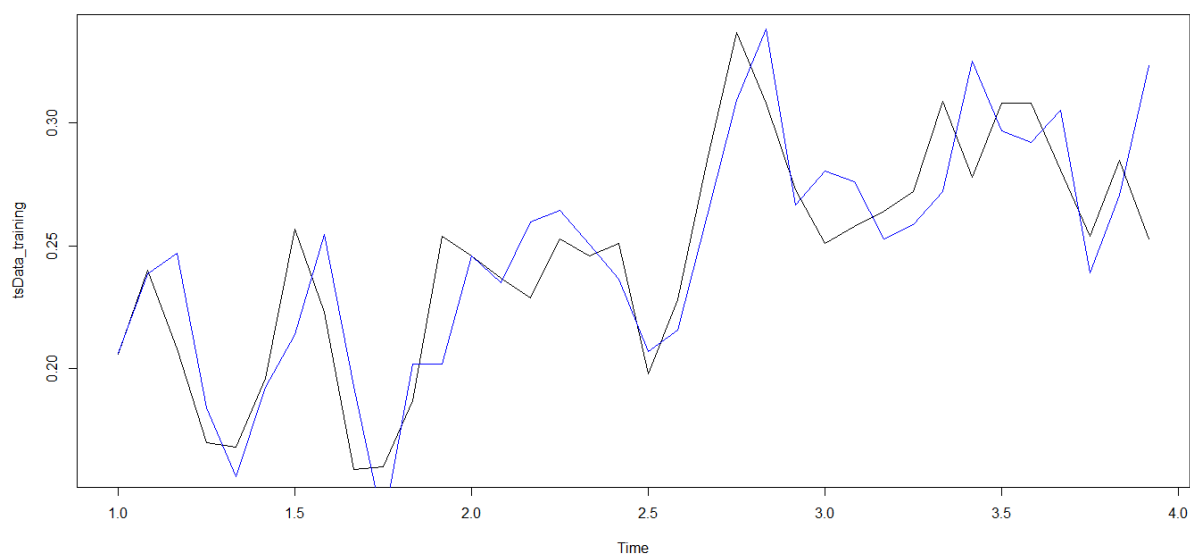


Tabela 10 - Coeficientes do ajuste da variável de teste SARIMA(3,2,0)(1,0,0).

	Estimado	Erro padrão	Z value	Pz z
ar1	-0,49528	0,16169	-3,06310	0,00219
ar2	-0,39696	0,16824	-2,35960	0,01830

ar3	-0,57731	0,14256	-4,04960	5,131e-05
sa1	-0,62994	0,14394	-4,37640	1,207e-05

Figura 30 - Predição utilizando a variável de teste SARIMA(3,2,0)(1,0,0).

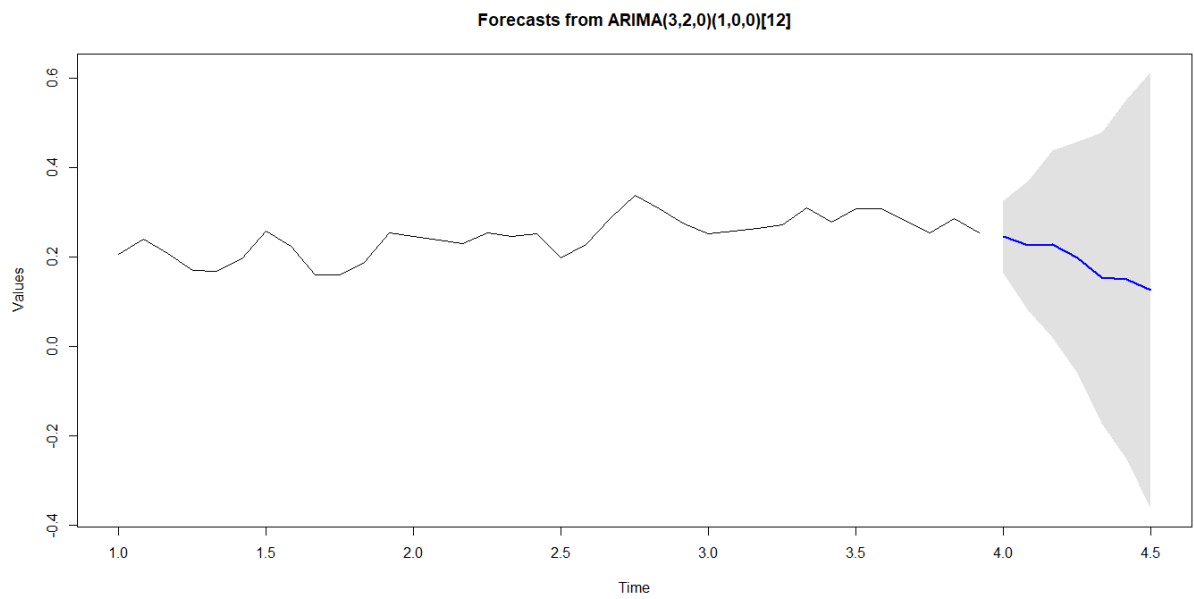


Figura 31 - Plotagem dos valores de predição da variável de teste (cinza) e valores de validação (azul) SARIMA(3,2,0)(1,0,0).

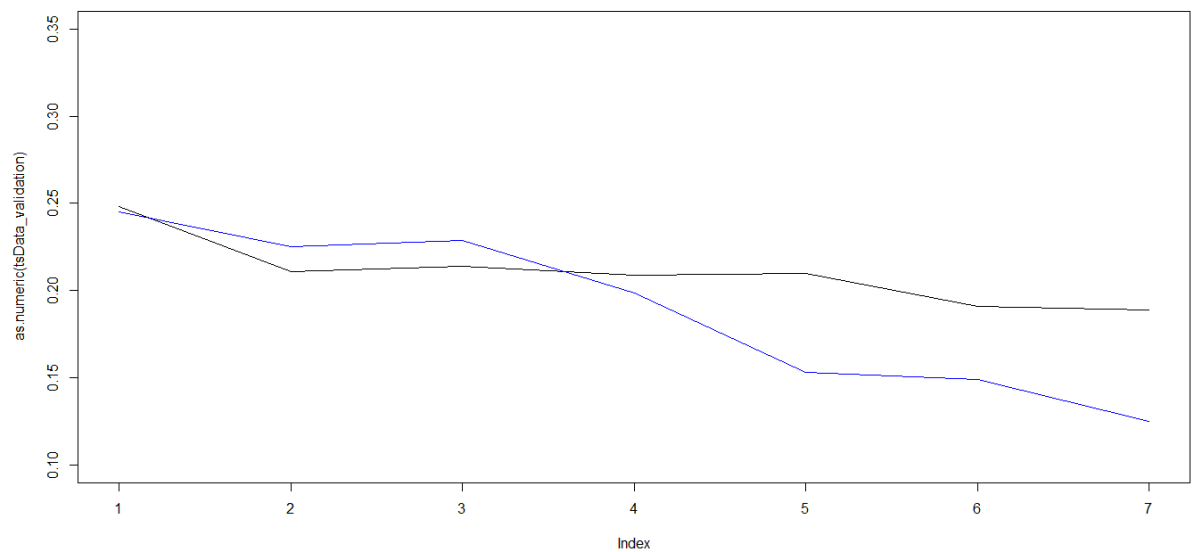


Tabela 11 – Coeficientes do ajuste na variável de teste SARIMA(3,2,0)(1,0,0).

AIC	BIC	MAPE ajuste	MAPE valores do ajuste e série treinamento	MAPE variável de validação e valores predição
-131,88	-124,25	8,60347	8,42312	19,40207

Tabela 12 - Previsão da série temporal de teste SARIMA(3,2,0)(1,0,0).

Mês	Dados de validação	Resultado da predição
Janeiro	0,248	0,24493
Fevereiro	0,211	0,22517
Março	0,214	0,22859
Abril	0,209	0,19842
Maio	0,210	0,15325
Junho	0,191	0,14911
Julho	0,189	0,12481

Figura 32 - Análise de resíduos de teste SARIMA(3,2,0)(1,0,0).

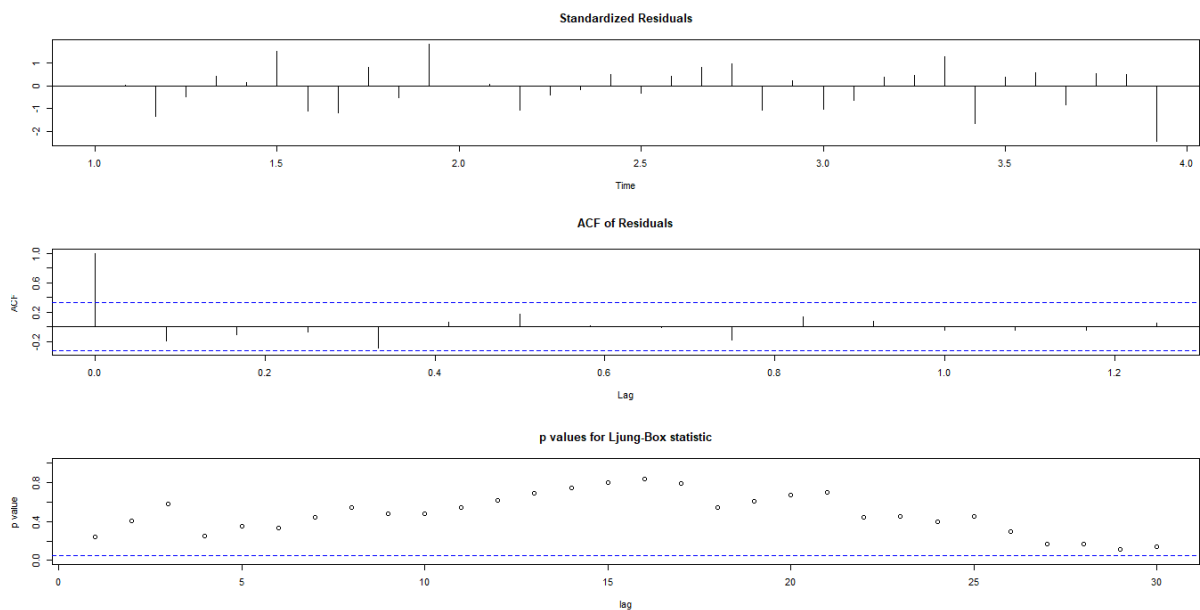
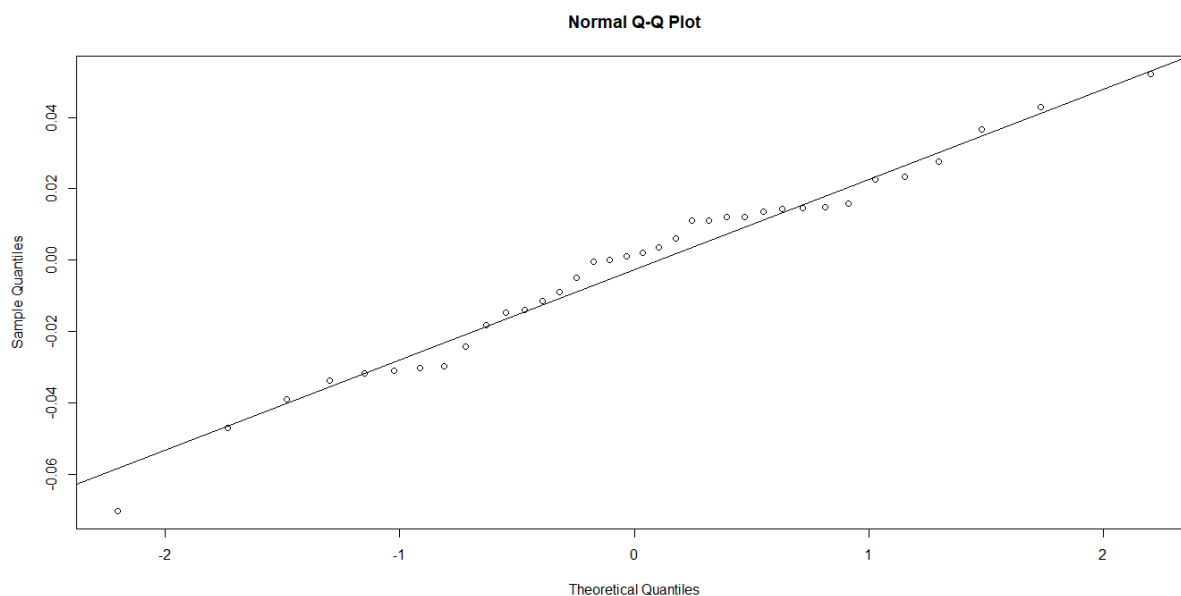


Figura 33 - Normal QQ Plot SARIMA(3,2,0)(1,0,0).



E então é realizada a predição, utilizando-se a variável inicial, até dezembro de 2019. Com a análise de Box-Pierce para os resíduos obteve-se o p-value de 0,2331, aceitando a hipótese nula de resíduo com padrão de ruído branco.

Figura 34 – Previsão para próximos 5 meses utilizando SARIMA(3,2,0)(1,0,0).

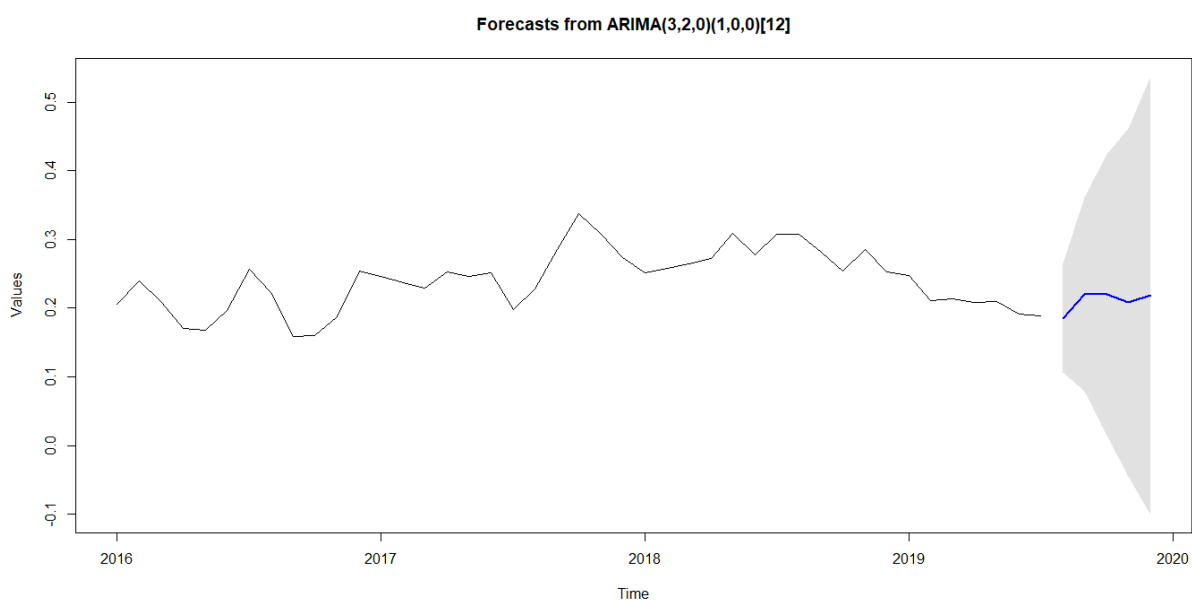


Tabela 13 - Resultados da previsão próximos 5 meses SARIMA(3,2,0)(1,0,0).

Mês	Resultado da predição
Agosto	0,18501
Setembro	0,22023
Outubro	0,21963
Novembro	0,20839
Dezembro	0,21828

Figura 35 - Padrão dos resíduos do valor ajustado SARIMA(3,2,0)(1,0,0).

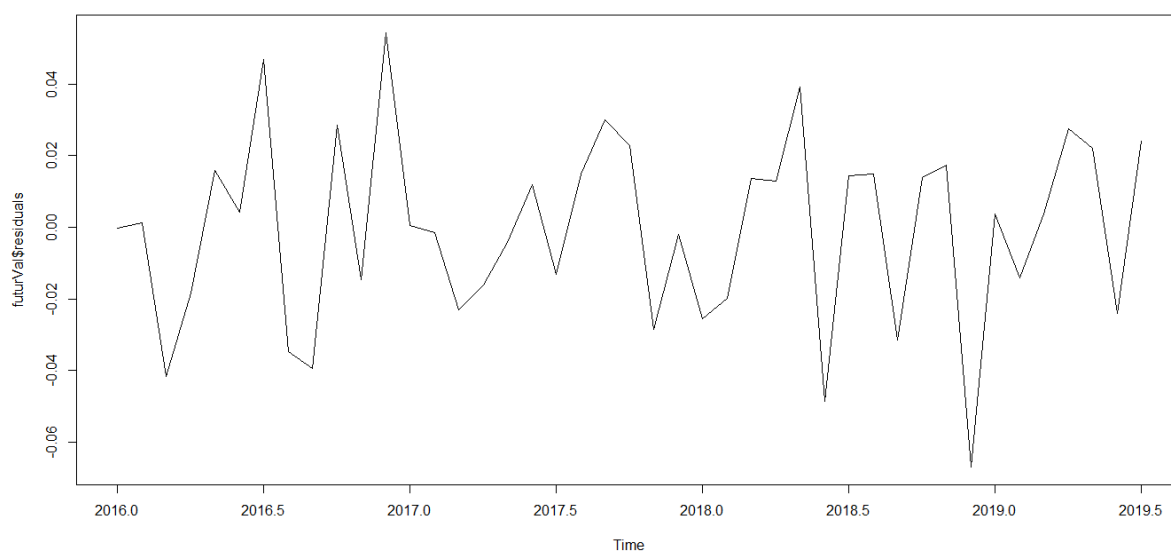


Figura 36 - Padrão dos resíduos do valor ajustado SARIMA(3,2,0)(1,0,0).

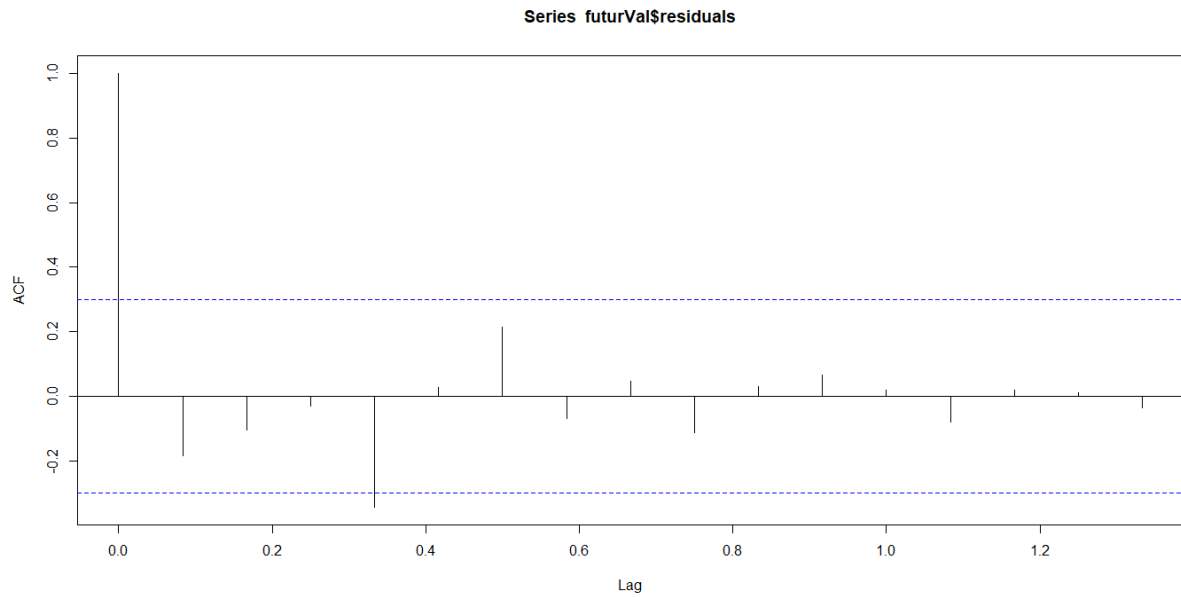
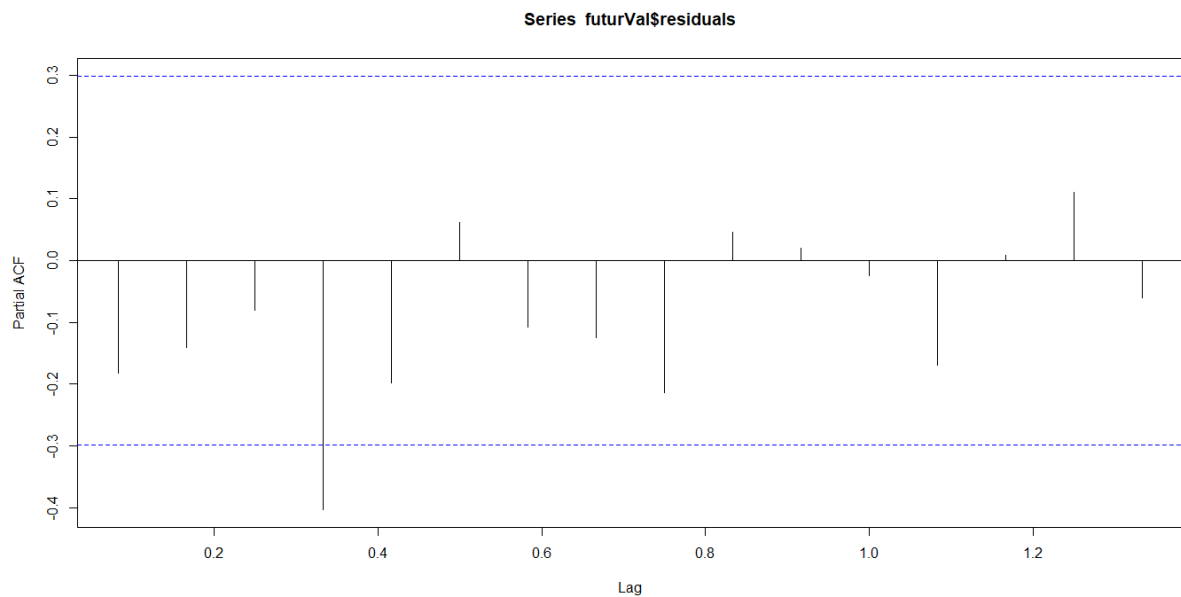


Figura 37 - Função de auto correlação para os resíduos SARIMA(3,2,0)(1,0,0).



4 Discussão

É importante salientar que cada medida de erro ou critério de informação possui vantagens e desvantagens ao ser considerado para decidir a melhor combinação de parâmetros que descreve a série. (Bal et. Al., 2016) realiza a previsão em séries temporais através da técnica de redes neurais e explicita cada medida de erro (média, mediana e critérios de informação) apontando suas vantagens e desvantagens.

O autor também afirma que a maior vantagem na escolha da medida de MAPE para definir o melhor modelo está na sua independência de escala, tornando possível a comparação de métodos além de ressaltar que a experimentação mostra a melhor performance de modelos selecionados por valores de erro em datasets de validação do que valores de erro em datasets de teste.

Para esse estudo foi realizada um tipo de validação cruzada, que divide os valores de entrada em teste e validação diferindo assim do estudo tradicional de séries temporais onde analisa-se somente os critérios de informação para o melhor modelo, restringindo a escolha para modelos com os menores valores de AIC e BIC.

Então, como escolha do melhor modelo, foi estabelecido como critério de seleção aquele que apresentou menor MAPE nos valores de validação, ou seja, maior acurácia comparando os dados previstos e os valores reais da série.

A figura 39 mostra os valores de cada previsão comparado com os dados de validação. Para cada modelo, os valores de erro médio absoluto percentual em relação a previsão são mostrados na tabela 13.

Figura 38 – Comparação dos valores previstos pelos modelos (coloridos) e valores reais (preto)

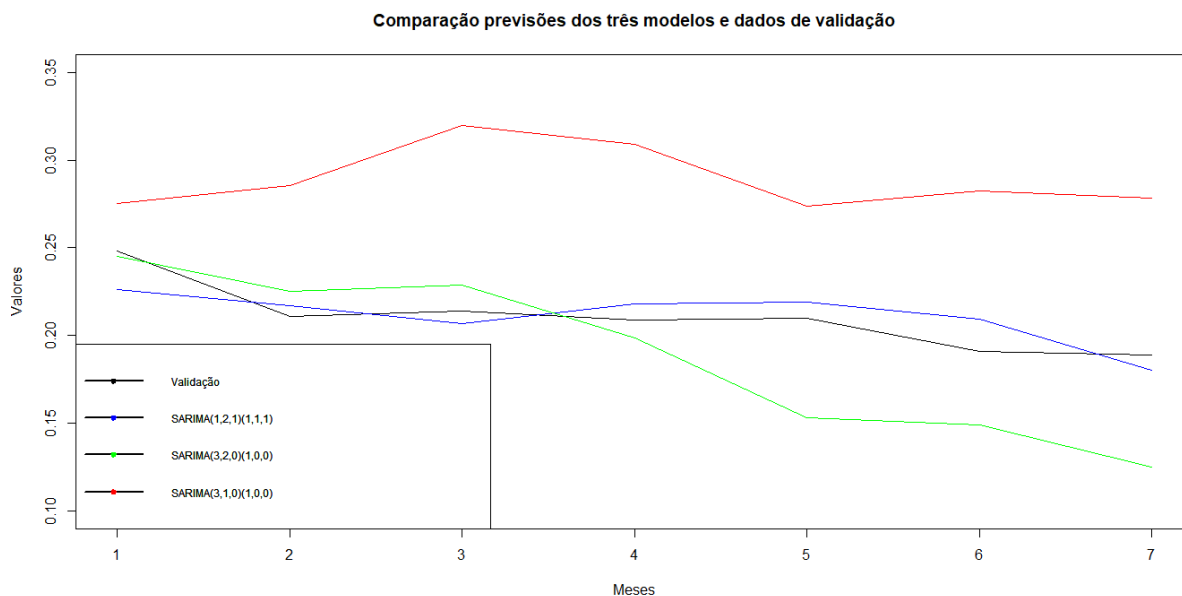


Tabela 14 – Comparação dos valores de MAPE para os três modelos propostos.

Modelo	MAPE previsto e variável de validação (%)
SARIMA(1,2,1)(1,1,1)	5,71
SARIMA(3,1,0)(1,0,0)	27,03
SARIMA(3,2,0)(1,0,0)	19,40

Assim, o melhor modelo é o SARIMA (1,2,1)(1,1,1) e, partir da equação 12, obtém-se os parâmetros para o modelo escolhido:

$$x_t = \frac{(1 - \theta_1 B)(1 - \Theta B^{12})w_t}{(1 - \Phi_1 B)(1 - \phi B^{12})(1 - B)(1 - B^{12})(1 - B^{12})^2} \quad (20)$$

Substituindo os valores da tabela 3 na equação 20, tem-se:

$$x_t = \frac{(1 + 0.999226B)(1 + 0.647543B^{12})w_t}{(1 + 0.647543B)(1 - 0.152615B^{12})(1 - B)(1 - B^{12})(1 - B^{12})^2} \quad (21)$$

5 CONCLUSÕES

A melhor tratativa para acidentes é evitar que esses ocorram, logo a abordagem preventiva no contexto de segurança no trabalho é a melhor opção para a preservação da integridade física e psicológica do empregado, e também para a manutenção da reputação da empresa. Assim, os esforços concentrados na elaboração de métodos e tratativas provenientes de análises apoiadas pela estatística dos dados históricos reforçam a necessidade de prevenir-se mediante um indicativo de ocorrência de acidentes, o que mais tarde também é refletido positivamente em indicadores reativos de performance como o TRIR.

O trabalho obteve êxito pois atingiu o objetivo proposto ao encontrar o melhor ajuste de parâmetros mediante a série temporal proposta. A inclusão de dados consolidados em 5 localidades distintas, porém com operações semelhantes bem como a aplicação de uma técnica de pré-processamento de média móvel dos últimos 12 meses proporcionaram a robustez necessária para as análises visto a baixa representatividade do dado mediante a baixa quantidade de homens horas trabalhadas e a alta nulidade do indicador ao longo dos meses.

Seguindo a metodologia de Box Jenkins, o melhor modelo ajustado foi SARIMA $(1,2,1)(1,1,1)$ com acurácia de 94,29% na variável prevista comparada com a variável de validação, com resíduos em padrão de ruído branco.

Como trabalhos futuros, é importante a inserção de variáveis exógenas, melhor ajustadas pelo modelo ARIMAX como, por exemplo, quantidade de exposição humana ao longo da série entre outras variáveis que produzam influência da série temporal analisadas.

6 REFERÊNCIAS

FUNDACENTRO. **Brasil registra 17 mil mortes e 4 milhões de acidentes de trabalho.** Disponível em: <http://www.fundacentro.gov.br/noticias/detalhe-da-noticia/2019/4/acoes-regressivas-gestao-de-riscos-e-impacto-dos-acidentes-de-trabalho-foram-temas-de-debate>. Acesso em 24 de novembro de 2019.

ANAMT. **A cada 3h40 um trabalhador morre de causas relacionadas à atividade laboral.** Disponível em: <https://www.anamt.org.br/portal/2019/05/03/a-cada-3h40-um-morre-por-acidente-de-trabalho-no-brasil/>. Acesso em 24 de novembro de 2019.

GIDDENS, Anthony et al. **Introduction to sociology.** New York: Norton, 1991.

HAIGHT, J. M., R. E. THOMAS, L. A. SMITH, R. L. BULFIN JR. AND B. K. HOPKINS. **Developing a Mathematical Relationship between Interventions and Incident Rates for the Design of a Loss Prevention System.** American Society of Safety Engineers Technical Forum: 38-44, 2001.

U.S. DEPARTMENT OF LABOR. **Recording and Reporting Occupational Injuries and Illness.** U.S. Department of Labor, Washington, DC, 1992.

CREA. CREA/CONFEA **Resolucao Nº 325.** , CREA § (1987).

OYEWOLE, S. A. **The implementation of statistical and forecasting techniques in the assessment of safety intervention effectiveness and optimization of resource allocation.** 2009, 173 f. Dissertação (Engenharia Industrial) - The Pennsylvania State University, Pennsylvania, 2009.

LINGARD, H., HALLOWELL, M., SALAS, R., & PIRZADEH, P. **Leading or lagging? Temporal analysis of safety indicators on a large infrastructure construction project.** *Safety Science*. <https://doi.org/10.1016/j.ssci.2016.08.020>, 2017

HEALTH AND SAFETY EXECUTIVE. **Keep your reputation**. Disponível em: <<http://www.hse.gov.uk/business/costs-damage.htm>>. Acesso em: 13 de novembro de 2019.

SHUMWAY, H. R.; STOFFER, S. D. **Time series analysis and its applications**. 2000.

R CORE TEAM (2018). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.R-project.org/>. Acesso em: 13 de novembro de 2019.

HYNDMAN, R.J., & ATHANASOPOULOS, G. **Forecasting: principles and practice**, 2nd edition, OTexts: Melbourne, Australia. 2018.

MANUCA, R., & SAVIT, R. **Stationarity and nonstationarity in time series analysis**. Physica D: Nonlinear Phenomena. [https://doi.org/10.1016/S0167-2789\(96\)00139-X](https://doi.org/10.1016/S0167-2789(96)00139-X), 1996.

MORETTIN, P.A. E TOLOI, C. M. **Modelos para Previsão de Séries Temporais**, Edgard Blucher, 1981.

AKAIKE, H. **A new look at the statistical model identification**. IEEE Transactions on Automatic Control., Boston, v.19, n.6, p.716-723, Dec. 1974.

SCHWARZ, G. **Estimating the dimensional of a model**. **Annals of Statistics**, Hayward, v.6, n.2, p.461-464, Mar. 1978.

SVS, R. Prediction of fatal accidents in Indian factories based on ARIMA. **Production Engineering Archives**, 2018.

SAAR, I. **Do Alcohol Excise Taxes Affect Traffic Accidents? Evidence From Estonia**. Traffic Injury Prevention. <https://doi.org/10.1080/15389588.2014.933817>, 2015.

FATTAH, J., EZZINE, L., AMAN, Z. **Forecasting of demand using ARIMA model.** International Journal of Engineering Business Management, v. 10, p. 1-9, 2018.

BRITO, G. F. **Aplicação do modelo arima na previsão de atendimentos em pontos de atenção com alta demanda da rede de assistência à saúde do município de monte carmelo, mg.** 2019. 101 f. Tese (Doutorado em Ciências) – Faculdade de Engenharia Elétrica, Universidade Federal de Uberlândia, Uberlândia, 2019.

BOX, G. E. P., JENKINS, G. M., & REINSEL, G. C. **Time series analysis: Forecasting and control:** Fourth edition. In Time Series Analysis: Forecasting and Control: Fourth Edition. <https://doi.org/10.1002/9781118619193>, 2013.

BAL, C., DEMIR, S. & ALADAG, C. H. **A Comparison of Different Model Selection Criteria for Forecasting EURO/USD Exchange Rates by Feed Forward Neural Network.** Journal of Computing, Communications & Instrumentation Eng. v.3, 2016.