UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

João Vitor da Silva

Suporte a decisão no setor sucroalcooleiro

Recife

2019

João Vitor da Silva

Suporte a decisão no setor sucroalcooleiro

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Estatística e Informática

Curso de Bacharelado em Sistemas de Informação

Orientador: Glauco Estácio Gonçalves

Recife

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

S586s Silva, João, João Vitor
Suporte a decisão no setor sucroalcooleiro / João, João Vitor Silva. - 2019.
42 f. : il.

Orientador: Glauco Estacio Goncalves.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2020.

1. Açúcar. 2. FDA. 3. Previsão. 4. Resíduos. 5. Séries temporias. I. Goncalves, Glauco Estacio, orient. II. Título

CDD 004

João Vitor da Silva

Suporte a decisão no setor sucroalcooleiro

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 11 de Dezembro de 2019.

BANCA EXAMINADORA

Glauco Estácio Gonçalves
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Silvana Bocanegra
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Gledson Luiz Pontes de Almeida
Departamento de Tecnologia Rural
Universidade Federal Rural de Pernambuco

Aos meus pais, que apesar de tudo, sempre se esforçaram ao máximo para me dar todo o suporte para chegar até aqui …

Agradecimentos

Agradeço primeiramente a Deus, que em seus planos permitiu que eu tivesse a oportunidade de chegar até aqui nesse momento.

Agradeço de todo o coração a cada um que contribuiu para minha formação profissional e pessoal, desde as tias da educação infantil, aos mestres e doutores da universidade. Sem vocês nada disso seria possível.

Agradeço à toda minha família, em especial aos meus pais, Maria Luzia e João José por seu empenho para oferecer para mim e meu irmão a melhor educação possível. Agradeço também a minha namorada Sabrina Emily, que buscou sempre me ajudar no que fosse preciso para a construção desse trabalho.

Agradeço à todos amigos que fiz ao longo do curso, em especial o grupinho: Daniel, Filipe, Jonathan, Romero, Vinícius e Udney. Das reuniões incansáveis que viravam a noite, vai ser impossível esquecer. Vocês tiveram papel fundamental na minha formação.

Agradeço ao professor Glauco Gonçalves, um mestre incrível que compartilhou muito de seu conhecimento comigo. Muito obrigado, pela sua orientação professor. Ao mestrando Diego Bezerra, que também compartilhou muito do que sabe comigo, só tenho o que agradecer. Sem vocês a construção desse trabalho seria impossível.

“Parte da jornada é o fim.”
(Tony Stark)

Resumo

O setor sucroalcooleiro é um dos maiores setores agrícolas do Brasil. A cada safra milhões de litros de etanol e milhares de toneladas de açúcar são importados mundo a fora. Apesar da grandeza do setor, existem diversos problemas que assombram o produtor de cana-de-açúcar. Um deles é a redução de produção provocando paradas na produção do açúcar e etanol.

Este trabalho tem como objetivo realizar um estudo comparativo de métodos de previsão de séries temporais, em dados históricos de produção de cana de açúcar, junto com a construção de indicadores operacionais para ajuda na tomada de decisão. A base de dados foi retirada dos resultados trimestrais publicados pela São Martinho para os seus investidores. A São Martinho é uma empresa de capital aberto e uma das maiores usinas de produção de açúcar, álcool e energia do Brasil. Para a realização do estudo foi utilizada a linguagem R. Os experimentos deste trabalho utilizaram o modelo preditivo SARIMA, por sua quase unanimidade na previsão de produções agrícola. Para a escolha do melhor modelo SARIMA foi utilizado o AICC do modelo junto com as métricas de erro RMSE, ME, e MAE. No desenvolvimento dos indicadores operacionais foi utilizada a função de distribuição dos resíduos do modelo SARIMA definido junto com as previsões do próprio modelo.

Ao final de todo o trabalho foi obtido o melhor modelo SARIMA para os dados de produção de cana-de-açúcar trimestrais junto com os indicadores de redução da produção: probabilidade de redução da produção em um determinado percentual de produção ou mais e probabilidade de redução da produção ser acima da média de produção trimestral.

Palavras-chave: açúcar, FDA, previsão, resíduos, séries temporais.

Abstract

The sugar and alcohol sector is one of the largest agricultural sectors in Brazil. Each harvest millions of liters of ethanol and thousands of tons of sugar are imported worldwide. Despite the size of the sector, there are several problems that haunt the sugarcane producer. One is the drop in production causing sugar and ethanol production stops.

This paper aims to carry out a comparative study of time series forecasting methods in historical sugarcane production data, together with the construction of operational indicators to aid in decision making. The database was taken from the quarterly results published by São Martinho for its investors. São Martinho is a publicly traded company and one of the largest sugar, alcohol and energy production plants in Brazil. The R language was used to carry out the study. The experiments of this work used the predictive model SARIMA, for its almost unanimity in the forecast of agricultural yields. RMSE, ME, and MAE. In the development of the operational indicators, the waste distribution function of the SARIMA model defined along with the forecasts of the model itself was used.

At the end of all the work, the best SARIMA model was obtained for the quarterly sugarcane production data together with the indicators of fall in production: probability of fall in production by 30 % and probability of fall in production below quarterly average production.

Keywords: sugarcane, time series,residuals, SARIMA, forecast, CDF.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AICC	Corrected Akaike Information Criterion
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
FDA	Função de Distribuição Acumulada
TS	Temporal Series
ME	Mean Error
MAE	Mean Absolute Error
ME	Mean Percentege Error
MAPE	Mean Absolute Percentege Error
RMSE	Root-Mean-Square Error

Sumário

1 Introdução

Com o passar dos anos, as agroindústrias vem se tornando cada vez mais importante no cenário econômico mundial. Só no Brasil, segundo a EMBRAPA, a agroindústria move aproximadamente 5,9% do PIB do pais[1]. Dividido em diversos tipos de fábricas, esse setor abrange desde moinhos de trigo até vinícolas. Dentre às diversas sub-áreas presentes no Brasil, existe uma agroindústria conhecida pela seu tempo presente no país, a primeira cultura instalada no Brasil, as **Usinas de cana-de-açúcar** (FONSECA, 2003).

Desde seu início em terras brasileiras, o setor sofreu diversas modificações. O processo de moagem da cana que antes era feito por animais, hoje, é feito por grandes máquinas em um processo totalmente automatizado. O setor se modificou, e o que antes eram pequenos negócios em engenhos regionais, hoje são grandes indústrias que exportam açúcar e etanol para o mundo todo. Somente na safra 2018/2019, o Brasil exportou 209.854 toneladas de açúcar e 73.996.000 litros de etanol, apenas para União Europeia, um negócio que moveu aproximadamente US$ 135.646.000 [2].

Um setor que move mais de US$ 100.000.000 pode ser considerado de fato grande. Ao total são aproximadamente 410 usinas espalhadas por todo o território nacional onde, aproximadamente 150 produzem açúcar, 100% produzem etanol e 99% produzem bioenergia. Na Figura 1 temos a localização de todas as usinas presentes em território nacional com clara concentração na Zona da Mata, da região Nordeste, e no interior do estado de São Paulo.

Figura 1 – Localização das usinas de açúcar e bioetanol no Brasil. Fonte: CTC – NIPE (2005)

1 <https://www.embrapa.br/grandes-contribuicoes-para-a-agricultura-brasileira/agroindustria>
2 <https://unicadata.com.br>

1.1 Justificativa e Relevância do Tema

Apesar do seu tamanho, o setor sucroalcooleiro possui dificuldades e problemas como qualquer outro. Um dos problemas que os grandes produtores do setor sucroalcooleiro buscam resolver é o da falta de chegada de cana na indústria provocando a parada de toda linha de produção.

Pela indústria só trabalhar em parte do ano, no período de safra, é fundamental garantir que o tempo de produção seja o máximo possível. Num período de aproximadamente 6 meses, de setembro a março no Norte-Nordeste, e de abril a novembro no Centro-Sul, toda a cana é colhida e processada, gerando açúcar e etanol que são estocados e vendidos ao longo do ano.[3] A redução da produção pode acarretar além das paradas da indústrias por falta de recurso, a redução da produção de açúcar e do etanol.

A busca por previsões de produção é de fato importante para o produtor de cana-de-açúcar, através dela é possível prever prováveis paradas de produção, e com isso tomar decisões estratégicas. E essa busca não é algo novo, na Índia, segundo maior produtor de cana-de-açúcar do mundo, atrás apenas do Brasil, diversos trabalhos já foram desenvolvidos em busca da previsão de produção de cana-de-açúcar [4]. Em 2011 (SURESH; PRIYA, 2011) publicaram um trabalho que abordava a previsão de área, produção e produtividade de cana-de-açúcar em Tamil Nadu, Índia. Já mais recentemente em 2015, (HOSSAIN; ABDULLA, 2015) utilizaram modelos de previsões nos dados de produção anual de cana-de-açúcar em Bangladesh no período de 1971 a 2013 para prever a produção do ano seguinte.

1.2 Objetivos

Este trabalho tem como objetivo realizar um estudo comparativo de métodos de previsão de séries temporais em dados históricos de produção de cana de açúcar, junto com a construção de indicadores operacionais para ajuda na tomada de decisão. Especificamente, espera-se:

- Definir um modelo de previsão para dados de produção de cana-de-açúcar com menor erro;
- Elaborar um indicador operacional que aponte a probabilidade da produção de cana-de-açúcar ser acima da média de produção com base no valor previsto pelo modelo;

[3] <https://www.novacana.com/cana/producao-cana-de-acucar-brasil-e-mundo>
[4] <https://www.novacana.com/cana/producao-cana-de-acucar-brasil-e-mundo>

- Elaborar um indicador operacional que aponte a probabilidade da produção prevista ser reduzida em um dado percentual p ou mais.

1.3 Organização do trabalho

O presente trabalho foi organizado em 6 capítulos. Além do presente capítulo de Introdução, o capítulo 2 aborda as técnicas utilizadas na execução do trabalho, o que são séries temporais, o método de previsão ARIMA, tanto as métricas de desempenhos utilizadas quanto os testes de hipóteses aplicados. Também no capítulo 2 foram expostos os conceitos de resíduos de um modelo estatístico, simulação Monte Carlo e função de distribuição acumulada.

Já no capítulo 3 foram apresentados os trabalhos correlacionados com o estudo feito. No capítulo 4 está apresentado o método utilizado na montagem dos experimentos. No capítulo 5 estão apresentados os resultados obtidos. Por fim, no capítulo 6 apresenta as considerações finais deste trabalho junto com sugestões de trabalhos futuros na mesma linha.

2 Referencial Teórico

2.1 Séries Temporais

Para prever valores futuros de determinada variável é necessário o conhecimento de como ela funciona. A melhor forma de conhecer como uma variável funciona é observar seu comportamento no longo do tempo. A coleção de observações de uma variável feitas sequencialmente ao longo do tempo é denominada de **série temporal**. Para (WOOLDRIDGE, 2003), uma série temporal é uma sequência de observações de uma variável ao longo do tempo.

Para uma coleção de observações ser de fato uma série temporal, é necessário que ela seja uma coleção de medidas de um fato através de um determinado intervalo de tempo. Outra característica fundamental de uma série temporal é o de que os dados estejam ordenados por ordem de acontecimento. Na Tabela 1 temos exemplos do que são séries temporais e do que não são séries temporais.

Séries Temporais	**NÃO são séries temporais**
Evasão de alunos por mês numa escola	Conjunto de dados sobre a chuva de PE
Internações por dia em um hospital	Número de acidentes de carro em 2017
Manchas solares por mês	Os salário dos habitantes de uma cidade

Tabela 1 – Exemplos de Séries Temporais.

2.1.1 Decomposição de uma Série Temporal

Segundo (SHIKIDA; MARGARIDO, 2009), uma série temporal pode ser genericamente decomposta em um ou mais componentes são eles: a Tendência, a Sazonalidade, o Ciclo e o Erro (Random). Esses são os componentes padrões que podemos encontrar numa série temporal. Com um gráfico de decomposição é possível observar essas 4 partes de uma série de forma visual.

A **Tendência** é a tendência de uma série subir ou descer. A série que possui uma tendência é considerada uma série não estável ao longo do tempo.

Outro ponto de elemento importante de uma série temporal é a **Sazonalidade**. A sazonalidade é definida como padrões que ocorrem em intervalos fixos de tempo, como dias, meses ou anos. O número de observações entre esses intervalos é chamado de **padrão sazonal**. Vários modelos de séries temporais estão sujeitos a sazonalidade.

Mais um elemento que uma série temporal pode ter é o **Ciclo**. O ato do aumento ou redução da frequência da ocorrência do fenômeno, mas que não que não ocorre

em intervalo fixo é denominado de Ciclo.

Por fim, outro elemento que encontramos numa série temporal é o **Erro**. O erro é justamente o resto da série, ou seja, removendo todos o componentes anteriores (Tendência, Sazonalidade e o Ciclo) o restante é justamente ele. São os dados que ocorrem devido a causas aleatórias que não são explicadas numa série temporal, que não estão nem na tendência, nem na sazonalidade e nem no ciclo.

Uma característica que difere as séries temporais de outros modelos estatísticos é a influência da ordem dos registros. Por conta disso, a **Autocorrelação** é uma ferramenta crucial na análise de séries. A autocorrelação consiste na correlação entre os registos de uma série temporal de acordo com seu intervalo de tempo.

A **Estacionariedade** é mais uma das características importantes das séries temporais. Uma série estácionaria é aquela em que a média, variância e estrutura de autocorrelação não mudam no decorrer do tempo. Se a série temporal não for estacionária, é possível transformá-la em estacionatravés de diversos métodos. A transformação mais popular é através da diferenciação sucessiva da série, expressa pela equação 2.1 onde, $Z(t)$ é a série temporal,e n o número de diferenças realizadas.

$$\Delta^n Z(t) = \Delta[\Delta^{n-1} Z(t)] \tag{2.1}$$

2.2 ARIMA

Proposto por (BOX; JENKINS, 1990) o ARIMA(Autoregressive Integrated Moving Average) é uma generalização do modelo ARMA (Autoregressive Moving Average), proposto pelos mesmos autores. O ARIMA é o modelo geral para processos não estacionários que possuem homogeneidade, ou seja, a série temporal pode apresentar traços de não estacionariedade, havendo a necessidade de se realizar diferenciações para torná-la estacionária. O ARIMA é composto de três características, são elas:

- **AR** : Autorregreção
  Relaciona os valores passados através de pesos para estimar os valores futuros.
- **I** : Integração
  Transformar a série em estacionária.
- **MA** : Médias Móveis
  Suaviza a série temporal, eliminando o não determinismo e características aleatórias.

O ARIMA possui três parâmetros e geralmente é apresentado da seguinte forma: **ARIMA(p,d,q)**, o **p** é a ordem (número de defasagens) do modelo auto-regressivo **d** é o grau de diferenciação (o número de vezes em que a série foi diferenciada) e **q** é a ordem do modelo de média móvel (BOX; JENKINS, 1990).

O ARIMA(p,d,q) pode ser expresso pela equação 2.2, onde p, d e q são os parâmetros do ARIMA, L é o operador defasagem, ϕ é o polinômio ligado ao operador autorregressivo de ordem p, θ é o polinômio ligado ao operador de média móveis de ordem q, e ε_t é o ruído branco.

$$(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i \varepsilon_t) \tag{2.2}$$

O ARIMA não funciona com a sazonalidade, para previsões de séries temporais com sazonalidade é preciso o uso o **SARIMA**, que também foi proposto por (BOX; JENKINS, 1990), o ARIMA que considera sazonalidade. Já o SARIMA por trabalhar com sazonalidade possui um conjunto de parâmetros adicionais e é apresentado da seguinte forma: **SARIMA(p,d,q)(P,D,Q)**, o **p** é a ordem (número de defasagens) do modelo auto-regressivo **d** é o grau de diferenciação (o número de vezes em que a série foi diferenciada), **q** é a ordem do modelo de média móvel, o **P** é a ordem (número de defasagens) do modelo auto-regressivo da parte sazonal **D** d é o grau de diferenciação (o número de vezes em que a série foi diferenciada) da parte sazonal e **q** é a ordem do modelo de média móvel da parte sazonal.

Segundo (DENGEN et al., 2016), o SARIMA(p,d,q)(P,D,Q) pode ser expresso pela equação 2.3, onde $\theta()$ é a autoregressão não-sazonal, $\Phi_p B^S$ é a autoregressão sazonal, $(1 - B)^d$ é a diferenciação não sazonal, $(1 - B^S)^D$ é a diferenciação sazonal, $\theta_q(B)$ é a média móvel não-sazonal e $\Theta_q(B^S)a_t$ a média móvel sazonal.

$$\Phi_p B^S \phi_p(B)(1 - B)^d(1 - B^S)^D Z_t = \theta_q(B)\Theta_q(B^S)a_t \tag{2.3}$$

2.3 Métricas de Desempenho

2.3.1 Critérios de Informação

Selecionar os parâmetros para um modelo SARIMA é uma tarefa complexa e existem vários caminhos para isso. Um dos métodos mais recomendados de seleção dos modelos ARIMA é utilizar os valores de referência, denominados **Critérios de Informação (CI)**. Em seu artigo, (MONDAL; SHIT; GOSWAMI, 2014) afirmaram que os Critérios de Informação (CI) são as uma das medidas mais usadas para avaliar modelos estatísticos. Por exemplo, o AIC (Akaike Information Criterion) é usado para

quantificar a qualidade do ajuste do modelo. Ao comparar dois ou mais modelos, o modelo com o AIC mais baixo é geralmente considerado o modelo que apresenta valores mais próximos dos dados reais. Isso também se aplica aos demais CI como o AICc (Corrected Akaike Information Criterion) e o BIC (Bayesian Information Criterion). O AIC pode ser expresso pela equação 2.4, o AICc pela equação 2.5, e o BIC pela equação 2.6, onde k é o número de parâmetros estimados no modelo, L é o valor máximo da função de verossimilhança para o modelo, e n é o tamanho da amostra, m é o número de observações.

$$AIC = 2k - 2\ln(\hat{L}) \tag{2.4}$$

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1} \tag{2.5}$$

$$BIC = \ln(m)k - 2\ln(\hat{L}) \tag{2.6}$$

2.3.2 Métricas de Erro

Outra forma de selecionar os melhores parâmetros do ARIMA é através do erro. As métricas de erro basicamente mensuram a diferença entre duas variáveis continuas (o previsto, e o relizado) e para assim criar um indicador do erro do modelo. Em 2017 (KAUR; AHUJA, 2017) usaram o **RMSE (Root-Mean-Square Error)** e **MPE (Mean Percentage Error)** como métrica para selecionar o melhor modelo ARIMA, o modelo que o obteve o menor valor dos indicadores foi considerado o vencedor. As principais métricas de erro aplicadas em séries temporais são o **ME (Mean Error)**, **MEA (Mean Absolute Error)**, **RMSE (Root-Mean-Square Error)**, **MPE (Mean Percentage Error)** e o **MAPE (Mean Absolute Percentage Error)**.

O ME (Mean Error) é a média da diferença entre o realizado e o previsto. A equação 2.7 apresenta a fórmula do ME.

$$ME = \frac{\sum_{i=1}^{n} y_i - x_i}{n} \tag{2.7}$$

O MAE (Mean Absolute Error) é a média da diferença absoluta entre o realizado e o previsto. A equação 2.8 apresenta a fórmula do MAE.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{2.8}$$

O RMSE (Root-Mean-Square Error) é o desvio padrão da amostra da diferença entre o previsto e realizado. A equação 2.9 apresenta a fórmula do RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \tag{2.9}$$

O MPE (Mean Percentage Error) é a média da diferença percentual do erro. A equação 2.10 apresenta a fórmula do MPE.

$$MAPE = \frac{1}{n} * \sum_{i=1}^{n} \frac{e_i}{y_i} \tag{2.10}$$

O MAPE (Mean Absolute Percentage Error) é a média da diferença absoluta percentual do erro. A equação 2.11 apresenta a fórmula do MAPE.

$$MAPE = \frac{1}{n} * \sum_{i=1}^{n} \frac{e_i}{y_i} \tag{2.11}$$

2.4 Testes de Hipótese

Quando se trabalha com estatística é necessário validar algumas hipóteses sobre a amostra de dados. Por exemplo, para afirmamos que uma amostra de dados é de fato normalmente distribuída, executamos um teste que aplica uma regra de decisão para aceitar ou rejeitar a hipótese de que a distribuição é normal. A classe de testes que aplicam uma regra de decisão para aceitar ou rejeitar uma hipótese estatística com base nos elementos amostrais é denominada de Testes de Hipóteses (MONTGOMERY; RUNGER; CALADO, 2000).

Os testes de hipóteses possuem duas hipóteses, a nula e a alternativa. A hipótese nula é a declaração que está sendo testada. Normalmente, a hipótese nula é uma declaração de ”nenhum efeito”ou ”nenhuma diferença”. A hipótese alternativa é a declaração que espera-se ser capaz de concluir que é verdadeira com base em evidências fornecidas pelos dados da amostra (MONTGOMERY; RUNGER; CALADO, 2000).

Dentro da classe de Testes de Hipóteses existem diversos testes, de testes de estacionalidade até testes de normalidade. Apesar de serem testes bem distintos um dos outros, os testes de hipótese possuem características em comum, uma delas é a de que o retorno se a hipótese nula foi aceita ou não pode ser expressa através de um **valor-p**. O valor-p é a probabilidade de alcançar uma estatística de teste igual ou mais extrema que a hipótese nula (WASSERSTEIN; LAZAR, 2016).

Considerando uma margem de confiança de 95 % ,pode-se rejeitar a hipótese nula à 5 %, neste caso o valor-p para deve ser menor ou igual à 0,05. Uma outra interpretação para o valor-p é que este é o menor nível de significância com que se pode rejeitar a hipótese nula. Em outras palavras, com um valor-p $\leq$ 0,05 , ou $\leq$ 0,01 (a depender da margem de confiança escolhida), a hipótese nula é rejeitada, logo a hipótese alternativa é verdadeira (WASSERSTEIN; LAZAR, 2016).

O Teste Shapiro–Wilk é um teste de normalidade publicado por (SHAPIRO; WILK, 1965). O teste de Shapiro-Wilk testa a hipótese nula é de que dada uma amostra $x_1, ..., x_n$ sua população é normalmente distribuída. A fórmula deste teste de hipótese é dada pela equação 2.12, onde $x_{(i)}$ é o menor número da amostra e $\bar{x} = \frac{(x_1+...+x_n)}{n}$ é a média da amostra.

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_1 - \bar{x})^2} \tag{2.12}$$

O coeficiente a_i é dado por pela equação 2.13, onde $C = \|V^{-1}m\| = (m^T V^{-1} V^{-1} m)^{\frac{1}{2}}$, e $m = (m_1, ..., m_n)^T$, sendo $(m_1, ..., m_n)^T$ constituído pelos valores médios da estatística da ordenada, de variáveis aleatórias independentes e identicamente distribuídas, amostradas a partir da distribuição normal padrão, e V é a matriz de covariância das estatísticas de ordem normal.

$$(a_1, ..., a_n) = \frac{m^T V^{-1}}{C} \tag{2.13}$$

Sendo a hipótese nula de que a população é normalmente distribuída, se o valor-p for $\leq$ 0,05 , ou $\leq$ 0,01 (a depender da margem de confiança escolhida), a hipótese nula será rejeitada (conclui-se que os dados não provêm de uma distribuição normal). Se não, conclui-se que esta hipótese não pôde ser rejeitada.

Proposto por (DICKEY; FULLER, 1979), o teste de Dick-Fuller Aumentado é um teste de estacionariedade em séries temporais. Sua hipótese alternativa é a de que a série é estacionária enquanto sua hipótese nula é a de que a série possui raiz unitária. A fórmula deste teste de hipótese é dada pela equação 2.14, onde y_t é a amosta, α é uma constante, β o coeficiente em uma tendência temporal, e p é a ordem de atraso do processo autoregressivo.

$$\Delta_{y_t} = \alpha + \beta t + \gamma y_{y_{t-1}} + \delta_1 \Delta y_{t-1} + + \delta_2 \Delta y_{t-2} + .. \tag{2.14}$$

Sendo a hipótese alternativa de que a a série é estacionária , se o valor-p for $>$ 0,05 , ou 0,01 (a depender da margem de confiança escolhida), a hipótese alternativa será aceita (conclui-se que a série é não estacionaria). Se não, conclui-se que a hipótese nula é verdadeira e a série possui raiz unitária, logo, a série não é estacionária.

2.5 Resíduos

Dado um modelo estatístico de regressão linear, como os modelos de previsão em séries temporais, existe uma diferença entre os valores ajustados que formam a função de previsão e os valores reais. O conjunto da diferença entre os valores ajustados pelo modelo e os valores reais é denominada de **Resíduos**. Na Figura 2 temos uma ilustração gráfica com um exemplo dos resíduos em uma regressão linear (DEKKING et al., 2005).

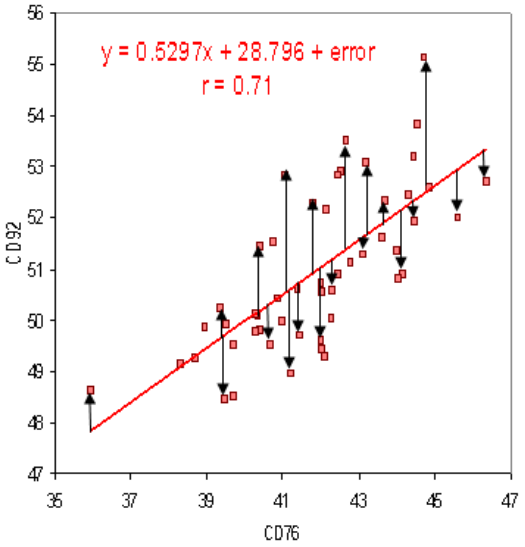

Figura 2 – Demostração gráfica dos resíduos de uma regressão linear. Fonte: nws.noaa.gov

2.6 Simulação Monte Carlo

A Simulação Monte Carlo é uma método de simulação baseada no **Método de Monte Carlo (MMC)**, conhecido por seu método de simulação estocástica. Em 2014 (KROESE et al., 2014) desenvolveram um trabalho abordando o motivo da importância do Método Monte Carlo hoje por sua característica estocástica.

A ideia principal por trás desse método é que os resultados sejam calculados com base em amostragem aleatória repetida junto a análise estatística. A simulação de Monte Carlo é, de fato, experimentações aleatórias, em que os resultados dessas experiências não são bem conhecidos. Como afirmado no trabalho de (SHOJAEEFARD; KHALKHALI; YARMOHAMMADISATRI, 2017), as simulações de Monte Carlo são tipicamente caracterizadas por muitos parâmetros desconhecidos, muitos dos quais são difíceis de obter experimentalmente.

O Método de Monte Carlo pode variar dependendo da implementação, mas tende a seguir o seguinte padrão específico:

- Definir um domínio de possíveis entradas;
- Gerar entradas aleatoriamente a partir da distribuição de probabilidade no domínio;
- Executar um cálculo determinístico nas entradas;
- Agregar o cálculo aos resultados.

2.7 Função de Distribuição Acumulada

Na descrição de modelos estatísticos, a **distribuição de probabilidade** é fundamental para descobrir o comportamento aleatório do fenômeno estudado. A Função de Distribuição Aculumada (FDA) descreve a distribuição de uma variável aleatória X. Para um dado número real x, a FDA é dada pela equação 2.15, onde X é a probabilidade de que a variável esteja num intervalo a entre a e b se a ≤ b (WALCK, 1996).

$$F(x) = P(X \leq x) \tag{2.15}$$

A função distribuição pode ser facilmente obtida a partir da função de probabilidade respectiva. No caso duma variável aleatória discreta a função de distribuição é dada pela equação 2.16.

$$\int_{-\infty}^{x} f(x_i)dx \tag{2.16}$$

3 Trabalhos Correlatos

Na literatura o uso de séries temporais para previsões de produções agrícolas não é nenhuma novidade, com o passar dos anos o número de estudos envolvendo esse assunto vem crescendo devido à grande preocupação da demanda de produção que aumenta a cada dia. Como (SAATH; FACHINELLO, 2018) abordaram e seu trabalho, considerando-se os limites da fronteira agrícola definidos no Brasil pela Embrapa em 2014, conclui-se que, embora exista uma pequena área legalmente disponível para a expansão agrícola e pecuária no Brasil, as novas demandas deverão ser atendidas com aumentos de produtividade e/ou substituição de cultura, justificando a necessidade de trabalhos como esses possuírem relevância para a sociedade . Neste capítulo serão abordados projetos e estudos que utilizam métodos de previsão de produção agrícola, que foram utilizados para a fundamentação deste trabalho.

Em 2011 (SURESH; PRIYA, 2011) publicaram um trabalho que abordava a previsão de área, produção e produtividade de cana-de-açúcar em Tamil Nadu, Índia. Com dados coletados entre 1950 e 2007 , Suresh, K. K. e Krishna Priya, S. R. modelaram dois modelos ARIMA capazes de preverem área, produção e produtividade de cana-de-açúcar. Para a previsão de área e produtividade de cana-de-açúcar foi selecionado o modelo ARIMA (1, 1, 1) e para a produção o ARIMA (2, 1, 2) . Os desempenhos dos modelos foram validados por comparação com valores reais. Com os modelos desenvolvidos, foram previstos os valores para a área de cana, produção e produtividade para os anos seguintes. Este trabalho pretende assim como (SURESH; PRIYA, 2011), desenvolver modelos ARIMA capazes de preverem a produtividade de cana-de-açúcar e escolher através de seu desempenho qual dos modelos é o melhor para o problema.

Já mais recentemente em 2015, (HOSSAIN; ABDULLA, 2015) utilizaram modelos ARIMA nos dados de produção anual de cana-de-açúcar em Bangladesh no período de 1971 a 2013 para efetuar previsões. O melhor modelo selecionado para prever as produções de cana-de-açúcar em Bangladesh foi o ARIMA (0,2,1). A comparação entre a série original e a série prevista mostraram similaridades, indicando que o modelo ajustado é estatisticamente bem comportado para prever produções de cana-de-açúcar em Bangladesh, ou seja, os modelos preveem bem durante e além da estimativa do período a um nível satisfatório. Neste trabalho as previsões geradas pelo modelo ARIMA serão comparadas com o padrão sazonal da série e com os dados de teste e treino através da métricas de erro, a fim de definir qual modelo selecionar.

No ano de 2005 o trabalho de (MURTA et al., 2005) objetivou-se em testar o modelo de Distribuição Gama na estimativa da precipitação pluvial mensal para Itapetinga

e Vitória da Conquista-BA através de bases de dados pluviais mensais do período coletadas entre 1978 a 1997. Para cada base de dados determinaram-se os parâmetros a e b da Distribuição Gama para a probabilidade mensal de chuva. As distribuições ajustadas foram usadas para estimar as probabilidades de chuva para cada mês. Este trabalho se propõe a fazer algo parecido com o desenvolvido por (MURTA et al., 2005), mas ao invés do uso da bases de precipitação pluvial mensal serão utilizados os resíduos do modelo ARIMA. Através dos resíduos do modelo será executado o cálculo das probabilidades por meio da função de distribuição, para obter as funções dos indicadores de redução de produção de cana-de-açúcar.

No Brasil, o uso do ARIMA no setor agrícola se restringe a previsão de preços de commodities. Como o trabalho de (PINTO et al., 2008), que analisaram o comportamento dos preços recebidos pelo produtor das principais commodities agrícolas brasileiros: cacau, café, cana de açúcar, laranja e soja. Prevendo preços através do modelo ARIMA, os resultados obtidos forneceram uma ferramenta de análise para o mercado destas commodities, na medida em que demonstram a tendência dos preços para um horizonte de curto prazo, servindo de auxílio à tomada de decisão de agentes que comercializam estes bens.

Também no Brasil, (FELIPE, 2012) desenvolveu um estudo com o objetivo de analisar a série de preços diária da soja do Norte do Paraná e descrever seu comportamento com previsões a curto prazo. Para responder a tal questionamento, o mesmo utilizou a metodologia ARIMA para as previsões. A manipulação dos dados utilizada foi baseada na análise gráfica e em testes estatísticos da própria metodologia. Ao final do trabalho, (FELIPE, 2012) observou que o modelo ARIMA (5,0,0) ou simplesmente AR (5), respondeu como o melhor modelo dentre o conjunto de modelos testados para prever o preço da soja e assim efetuou as previsões.

4 Método de Experimentação

Esta seção descreve as ferramentas, base de dados e etapas dos experimentos desenvolvidos no estudo. O tema foi escolhido pela grande relevância do setor sucro-alcooleiro na economia do país e a esperança de retomada do categoria nos próximos anos (UNICA, 2016). Um estudo que proponha auxiliar tomadas de decisões cruciais no ramo é fundamental para o crescimento dessa esfera que emprega dezenas de milhares de pessoas.

4.1 Ferramental de Software

Os experimentos abordados nesse trabalho tem como objetivo realizar o calculo de probabilidade de redução de produção nas previsões de toneladas de cana-de-açúcar previstas por um modelo SARIMA. Todas as análises foram desenvolvidas usando a linguagem R por sua eficiência e disponibilidade de bibliotecas para análise de dados e modelagem de séries temporais, facilitando o uso de funções de manipulação de dados e métodos estatísticos.

Para este trabalho foram utilizadas diversas bibliotecas para manipulação de séries temporais, são elas: `forecast`[1], usada na análise e visualização das previsões de séries temporais com ARIMA; `tseries`[2], que possui funções comumente usadas na análise de séries temporais; `ggplot2`[3], usada para a construção de gráficos; e `dplyr`[4], usada para a construção de *data frames* que permitem melhor manipulação da série temporal.

4.2 Base de Dados

A base de dados foi retirada dos resultados trimestrais publicados pela São Martinho para os seus investidores[5]. A São Martinho é uma empresa de capital aberto e um dos maiores grupos de produção de açúcar, álcool e energia do Brasil. Localizado no sudeste e constituído por 4 usinas, o grupo possui ao todo 300 mil hectares de área agrícola de colheita onde, aproximadamente 100% da colheita é mecanizada, e uma capacidade de moagem de 24 milhões de toneladas por safra. A empresa publica seus

[1] <https://cran.r-project.org/web/packages/forecast/index.html>
[2] <https://cran.r-project.org/web/packages/tseries/index.html>
[3] <https://cran.r-project.org/web/packages/ggplot2/index.html>
[4] <https://cran.r-project.org/web/packages/dplyr/index.html>
[5] Central de Resultados da São Martinho - <https://ri.saomartinho.com.br/>

resultados a cada trimestre a fim de informar os seus investidores do seu rendimento e também para atrair possíveis novos investidores.

A partir dos relatórios trimestrais foi o montante de cana processada em milhões de toneladas dos relatórios de produção do primeiro trimestre do ano safra 06/07 até o primeiro trimestre do ano safra 19/20 (atual), totalizando 53 observações.

É importante saber que a medida de ano safra é diferente de um ano convencional do calendário gregoriano. Geralmente o ano safra se inicia no começo da safra e acaba no final da entre safra. Com safra entre abril(inicio do ano safra) e novembro é possível observar a produção da São Martinho nos três primeiros trimestres de cada ano safra até o quarto trimestre que não possui produção, por ser justamente o período de entre safra. Na Tabela 2 temos a base de dados com as observações retiradas dos relatórios.

4.3 Seleção dos Métodos de Previsão

Para os experimentos deste trabalho foi utilizado o modelo preditivo ARIMA, por sua quase unanimidade na previsão de produções agrícolas, como visto na Seção 3. Como a série temporal possui certo padrão sazonal, preferiu-se empregar o modelo SARIMA. A implementação do SARIMA utilizada foi a da biblioteca `forecast` da linguagem R.

Para realização dos experimentos, foi utilizada a técnica de Hold-out, dividindo a série temporal em duas partes: treino e teste. Devido a base de dados ser relativamente pequena, o intervalo de previsão definido foi de 8 trimestres (2 anos). Sendo assim a base foi dividida em treino (45 amostras) e teste (8 amostras). Já para a previsão foi utilizada a função `forecast` da biblioteca da linguagem R com mesmo nome.

O modelo SARIMA, como visto na Seção 2.2, pode ser considerado um modelo-base já que possui seis parâmetros livres, que podem ser combinados para gerar diferentes modelos. Neste trabalho, os parâmetros p, q, P e Q variaram de 0 a 5, enquanto que os parâmetros d e D foram configurados para o valor 1, já que a análise de estacionariedade, como mostrado na Seção 5.1.2, indicaram que este valor é suficiente para estacionar a série e sua sazonalidade.

Ao todo foram testados 1296 diferentes modelos SARIMA os quais foram avaliados utilizando o AICc de cada modelo junto com as métricas de erro RMSE, ME, e MAE.

N°	Safra	Trimestre	Produção
1	06/07	1	3754
2	06/07	2	4882
3	06/07	3	640
4	06/07	4	0
5	07/08	1	3199
6	07/08	2	4679
7	07/08	3	2340
8	07/08	4	0
9	08/09	1	2996
10	08/09	2	5674
11	08/09	3	3331
12	08/09	4	0
13	09/10	1	4484
14	09/10	2	4962
15	09/10	3	3477
16	09/10	4	0
17	10/11	1	5252
18	10/11	2	5561
19	10/11	3	2254
20	10/11	4	0
21	11/12	1	3648
22	11/12	2	5693
23	11/12	3	2071
24	11/12	4	0
25	12/13	1	2918
26	12/13	2	6036
27	12/13	3	1636
28	12/13	4	0
29	13/14	1	5543
30	13/14	2	6097
31	13/14	3	3952
32	13/14	4	0
33	14/15	1	6467
34	14/15	2	8691
35	14/15	3	3559
36	14/15	4	0
37	15/16	1	7409
38	15/16	2	7628
39	15/16	3	4987
40	15/16	4	0
41	16/17	1	8186
42	16/17	2	8346
43	16/17	3	2749
44	16/17	4	0
45	17/18	1	8739
46	17/18	2	9933
47	17/18	3	3534
48	17/18	4	0
49	18/19	1	9508
50	18/19	2	8921
51	18/19	3	2021
52	18/19	4	0
53	19/20	1	9042

Tabela 2 – Base de dados retirada dos relatórios trimestrais da São Martinho. Na coluna *Safra* tem-se o ano safra referente a produção, na coluna *Trimestre* tem-se o trimestre referente do ano safra e na coluna *Produção* a quantidade de cana processada em milhares de toneladas.

4.4 Desenvolvimento dos indicadores

Como abordado na seção 1.2, o presente trabalho tem por objetivo a criação de dois indicadores operacionais, denominados de **I1** e **I2**, definidos da seguinte forma:

- **I1** - Aponta a probabilidade da produção de cana-de-açúcar ser acima da média de produção histórica com base no valor previsto pelo modelo SARIMA;
- **I2** - Aponta a probabilidade da produção prevista ser reduzida em um dado percentual p ou mais.

O método de cálculo de cada indicador parte, igualmente, da distribuição dos resíduos do modelo SARIMA. Esta distribuição é elaborada com o auxílio da simulação Monte Carlo para reamostragem uniforme com reposição da pequena amostra de resíduos (45 observações). Por meio da simulação a amostra inicial de resíduos é reamostrada para compor uma nova amostra de 1000 observações. Assim, é possível ter uma representação mais acurada da distribuição dos resíduos.

A Figura 3 apresenta os passos para a construção do indicador I1. Após a simulação dos resíduos, adiciona-se a média histórica de produção de cana-de-açúcar – obtida a partir dos dados de produção – aos resíduos simulados. Desta forma, a função de distribuição acumulada capturará o comportamento estocástico da média de produção histórica. A partir dos dados simulados, a função de distribuição acumulada pode ser obtida empiricamente.

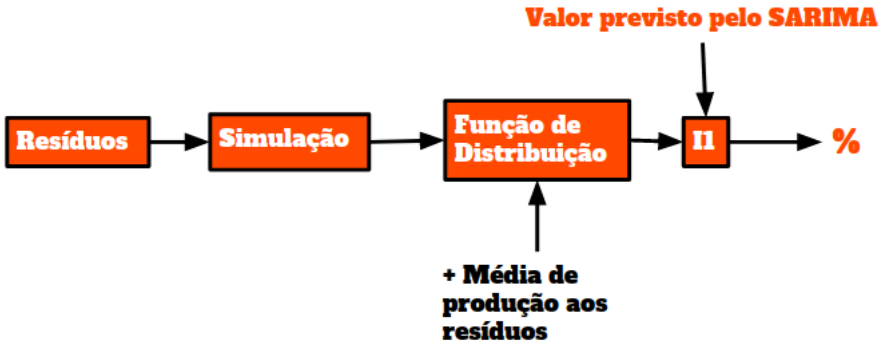

Figura 3 – Algoritmo para cálculo de I1.

O último passo na construção do indicador I1 é calcular a probabilidade do valor previsto pelo modelo SARIMA ser maior que a média de produção histórica. Assumindo que a média de produção histórica é uma variável aleatória P_h, que a função acumulada empírica é dada por $\widehat{F_{P_h}(x)}$ e que o valor previsto pelo modelo SARIMA para um trimestre futuro é $\hat{p}$ então a probabilidade da produção ser acima da média (p_m) é dada por

$$p_m = \widehat{F_{P_h}}(\hat{p})$$

Para melhor entendimento desta definição, note-se que, conforme Equação 2.15, a função de probabilidade acumulada calcula a probabilidade de que a variável aleatória assuma um valor menor ou igual ao argumento da função, que é o mesmo que dizer que é a probabilidade de que o argumento da função é maior ou igual à variável aleatória. Formalmente, $F(x) = P(X \leq x) = P(x \geq X)$.

A Figura 4 mostra o passo a passo para a construção do indicador I2. Após a simulação Monte Carlo dos resíduos do modelo SARIMA, cria-se a função de distribuição acumulada empírica destes resíduos.

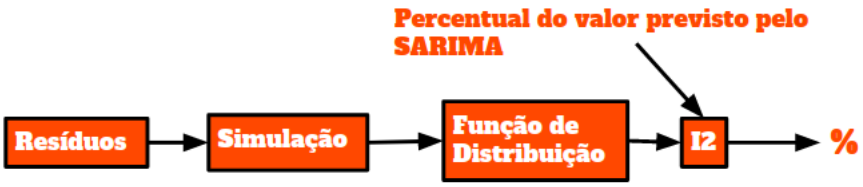

Figura 4 – Algoritmo para cálculo de I2.

Considerando que os resíduos são uma variável aleatória R, que a função acumulada empírica é dada por $\widehat{F_R}$ e que o valor previsto pelo modelo SARIMA para um trimestre futuro é $\hat{p}$, então a probabilidade de redução da produção $p_r(p)$ em um dado percentual p ou mais é dada por

$$p_r(p) = \widehat{F_R}(-p \times \hat{p})$$

Como a distribuição dos resíduos do modelo SARIMA tende a ter um conjunto de dados centrados em torno de 0, o uso do valor negativo para o percentual permite capturar os casos em que há a redução da produção de cana-de-açúcar. Além disso, a própria definição da função acumulada de probabilidade captura os casos que vão além do percentual de redução p definido. Neste trabalho, para fins de experimentação e demonstração do funcionamento do indicador I2, p foi definido arbitrariamente como 10%, 20% e 30%.

4.5 Estrutura

Os experimentos desenvolvidos ao longo deste trabalho foram divididos em duas partes:

- Definição do modelo SARIMA;
- Previsões e construção dos indicadores.

Na Definição do modelo SARIMA foram analisadas as combinações de parâmetros de modelo e seu desempenho, selecionando um modelo para ser usado no trabalho. Na segunda parte foram realizadas as previsões e construção dos indicadores com base no modelo selecionado na primeira parte do trabalho.

5 Resultados

Este capítulo apresenta os resultados para escolha do modelo de previsão SARIMA adequado aos dados de produção da Usina São Martinho (Seção 5.1), bem como apresenta, na Seção 5.2, os resultados dos indicadores de estimativa da produção de cana-de-açúcar.

5.1 Definição do modelo de previsão

A Figura 5 mostra a decomposição da série temporal de produção trimestral de cana-de-açúcar. No gráfico é possível observar a própria série (*observed*), a tendência da série (*trend*), os padrões de sazonalidade (*seasonal*) e o erro (*remainder*).

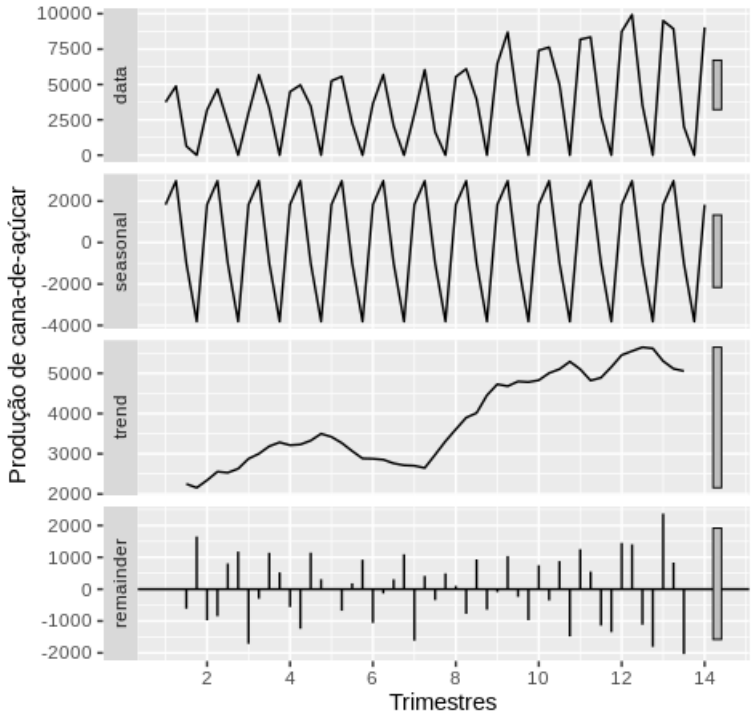

Figura 5 – Decomposição da série temporal de produção de cana de açúcar

Analisando a decomposição, pode-se observar o claro padrão sazonal da série de produção de cana-de-açúcar, tanto na própria série quanto, com mais clareza, na parte sazonal da decomposição. Como uma das premissas para o uso do ARIMA é o de que os dados não apresentem sazonalidade foi definido o uso do **SARIMA(p,d,q)(P,D,Q)**, para cobrir o aspecto da sazonalidade presente nos dados.

Para definir qual o conjunto mais adequado de parâmetros para o SARIMA utilizou-se, primeiramente, uma abordagem baseada em testes de hipótese e análise gráfica para determinação dos parâmetros de diferenciação da série (*d*) e de seu padrão sazonal (*D*). Em seguida, realizou-se uma busca exaustiva nos demais parâmetros do modelo, sendo a escolha determinada pela métrica AICc junto às métricas de erro (RMSE, ME, MAE). O modelo que apresentou os menores valores nestas métricas foi escolhido para o restante das análises do trabalho.

5.1.1 Estacionariedade e Estacionariedade Sazonal

A fim de diminuir a quantidade de interações para descobrir o modelo mais adequado, decidiu-se, inicialmente, encontrar valores para *d* e *D*.

Para determinar o valor de *d*, utilizou-se o teste de hipótese de Dickey-Fuller, abordado na Seção 2.4, que informa se uma série é estacionária ou não. Foram aplicadas diferenciações até o teste de hipótese informar que a série se tornou estacionária, desta forma definindo a quantidade de diferenciações para tornar a série estacionária.

Ao submeter-se a série temporal ao teste de *Dickey-Fuller* sem qualquer diferenciação, obteve-se um valor-p de 0.71, i.e., não foi possível refutar a hipótese de não-estacionariedade. Aplicando-se uma diferenciação à série temporal e reaplicando o teste de *Dickey-Fuller*, obteve-se um valor-p de 0,01, assim demonstrando que a série é estacionaria. Desta forma, concluiu-se que é preciso apenas uma diferenciação para a série se tornar estacionária, ou seja, $d = 1$.

Já para definir o valor de *D* foi utilizado o decaimento do gráfico do função de autocorrelação da série temporal (*Autocorrelation Function* - ACF) como indicador da estacionariedade na sazonalidade. Foram aplicadas diferenciações sazonais até o decaimento da ACF se tornar aparentemente exponencial, informando que a série se tornou estacionária na parte sazonal e definindo assim o número de diferenciações sazonais para tornar a série estacionária.

Na Figura 6 tem-se o gráfico da ACF da série temporal de produção de cana de açúcar. A análise gráfica demonstra a presença de um claro padrão sazonal, posto que a autocorrelação nos *lags* (atrasos) sazonais é significativa. Além disso, a função de autocorrelação tem um decaimento linear, indicando que a série temporal é não-estacionária na parte sazonal, sendo, portanto, necessário deixá-la estacionária.

Ao aplicar-se uma diferenciação sazonal à série, obteve-se uma nova função de autocorreção em que o padrão de sazonalidade foi eliminado, conforme mostra a Figura 7. Note-se que o decaimento apresentado é exponencial. Sendo assim, conclui-se que só é preciso uma diferenciação sazonal para a série se tornar estacionária e, portanto, $D = 1$.

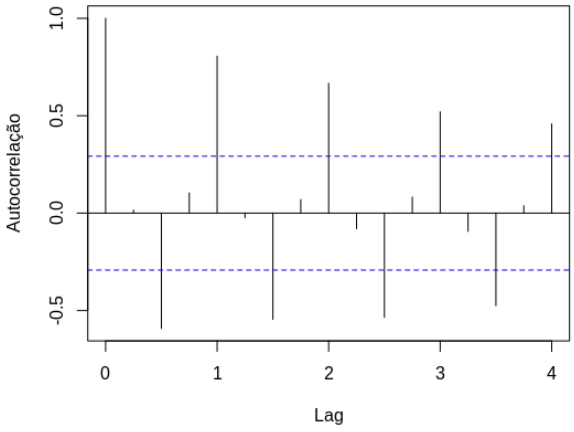

Figura 6 – Gráfico da função de autocorrelação dos dados de produção de cana-de-açúcar antes da diferenciação sazonal

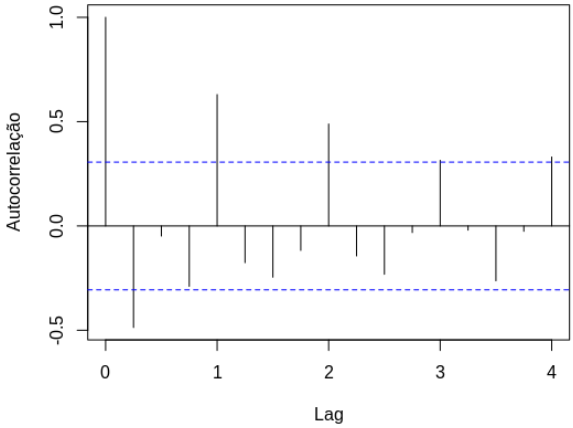

Figura 7 – Gráfico da função de autocorrelação dos dados de produção de cana de açúcar após a diferenciação sazonal

5.1.2 Escolha dos demais parâmetros do modelo SARIMA

Com os valores de d e D definidos, foram testadas todas as combinações possíveis para os valores de p, q, P e Q variando-os de 0 à 5. Para cada modelo gerado foram calculados os valores de AICc e RMSE e, a partir disso foram selecionados os modelos que obtiveram o melhor desempenho nas duas métricas.

Após análise dos resultados de cada combinação de parâmetros do modelo, foram selecionados 13 modelos, de um total de 1296, que apresentaram a melhor precisão considerando as duas métricas. Para uma decisão criteriosa do modelo mais adequado, foram calculadas mais duas métricas de precisão, o ME e o MAE. Estas médias foram determinadas tanto para o erro de treino quanto para o erro de teste. Os

resultados obtidos são apresentados na Tabela 3.

(p,d,q)	(P,D,Q)	RMSE	AICC	ME 1	ME 2	MAE 1	MAE 2
(0,1,1)	(0,1,0)	962,28	675,81	49,12	25,11	662,08	582,00
(0,1,2)	(0,1,0)	954,01	677,66	46,87	28,92	656,86	588,26
(0,1,1)	(1,1,0)	940,82	676,82	60,26	-77,58	664,62	605,24
(0,1,2)	(1,1,0)	959,54	677,99	14,05	375,01	696,90	683,77
(0,1,1)	(3,1,0)	867,88	676,71	86,75	-11,81	633,12	861,93
(0,1,2)	(3,1,0)	879,21	678,92	40,48	-27,84	637,96	916,76
(0,1,1)	(4,1,0)	854,18	678,75	78,75	-90,12	622,05	805,88
(0,1,1)	(0,1,1)	933,61	676,54	77,15	-42,83	679,28	746,64
(0,1,2)	(1,1,1)	930,10	678,68	11,79	532,51	697,01	951,18
(0,1,1)	(3,1,1)	823,27	677,61	75,25	-3,70	609,84	772,01
(0,1,1)	(0,1,2)	924,71	678,63	80,13	-93,40	682,85	682,83
(0,1,1)	(1,1,3)	816,30	677,81	69,27	66,02	610,58	821,36

Tabela 3 – Modelos com os menores AICC (Corrected Akaike Information Criterion), RMSE (Root-Mean-Square Error), ME (Mean Error) de treino (1) e teste (2), MAE (Mean Absolute Error) de treino (1) e teste (2).

Analisando o desempenho de cada modelo, com base no RMSE, AICC, ME (teste), ME (treino), MAE (teste) e MAE (treino), conclui-se que o conjunto de parâmetros mais adequado foi o **(0,1,1)(0,1,0)**. Sendo assim, o modelo **SARIMA(0,1,1)(0,1,0)** foi selecionado para as demais análises.

5.1.3 Previsões

Como explicado na Seção 4.3, definiu-se uma janela de previsão de dois anos à frente, ou seja 8 trimestres, para teste. Sendo assim, foram previstos 8 trimestres de produção de cana de açúcar com o modelo **SARIMA(0,1,1)(0,1,0)**. Na Tabela 4 temos os valores previstos pelo modelo SARIMA, em cada trimestre, comparados com os valores reais de produção. Os mesmos dados são apresentados na Figura 8, para melhor visualização.

Trimestre	Previsão	Teste
T1	8603,51	9933
T2	3006,51	3534
T3	257,51	0
T4	8996,51	9508
T5	8861,02	8921
T6	3264,02	2021
T7	515,02	0
T8	9254,02	9042

Tabela 4 – Previsões do Modelo SARIMA(0,1,1)(0,1,0) versus Amostra de Teste. Os valores são dados em milhares de toneladas de cana-de-açúcar.

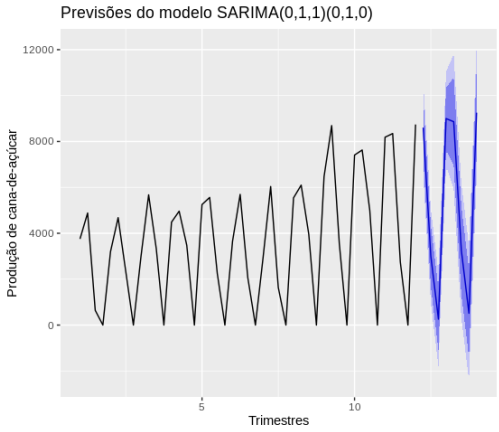

Figura 8 – Gráfico de previsão dos dados de produção de cana de açúcar. Em azul, os trimestres previstos pelo modelo SARIMA(0,1,1)(0,1,0); Em preto, a série temporal (treino).

A partir da Tabela 4 depreende-se que o modelo de previsão aproximou razoavelmente os valores de teste, alcançando um erro médio absoluto de aproximadamente 662 milhares de toneladas. Além disso, a Figura 8 mostra que o padrão sazonal da produção de cana-de-açúcar do Grupo São Martinho também foi preservado no modelo.

Outro aspecto que pode ser observado deste resultado é que o modelo SARIMA escolhido tende a subestimar os valores reais. De certa forma, o modelo é pessimista e tal característica dá segurança ao tomador de decisão de que o modelo não produzirá previsões de produção muito acima da realidade.

5.2 Resultados dos Indicadores

Para o desenvolvimento dos indicadores I1 e I2 foram retirados os resíduos do modelo SARIMA selecionado. Um dos indicadores de que um modelo estatístico é de boa qualidade é que os resíduos do modelo sigam distribuição normal, com média zero, desvio padrão constante e que estes sejam aleatórios.

A Figura 9 exibe os resíduos gerados pelo modelo SARIMA para cada trimestre. No gráfico é possível observar que os resíduos apresentam características de que são aleatórios. Já a Figura 10 mostra que a distribuição dos resíduos gerados pelo modelo se aproxima da distribuição normal.

Conforme determinado na Seção 4.4, após a coleta dos resíduos do modelo SARIMA, deve-se realizar a simulação de reamostragem dos resíduos para melhor representação da distribuição dos resíduos. A Figura 11 mostra a distribuição gerada pela simulação Monte Carlo dos resíduos.

Os dados dos resíduos já simulados foram submetidos ao teste de hipótese *Shapiro-Wilk*, abordado na seção 2.4, para verificar se estes seguiam uma distribui-

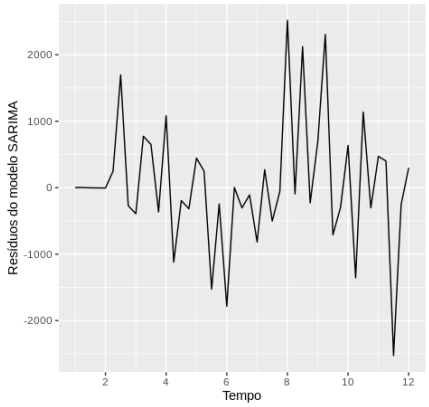

Figura 9 – Gráfico dos resíduos do modelo SARIMA(0,1,1)(0,1,0).

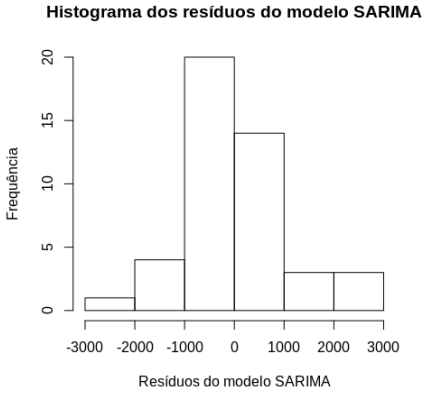

Figura 10 – Gráfico de distribuição dos resíduos do modelo SARIMA.

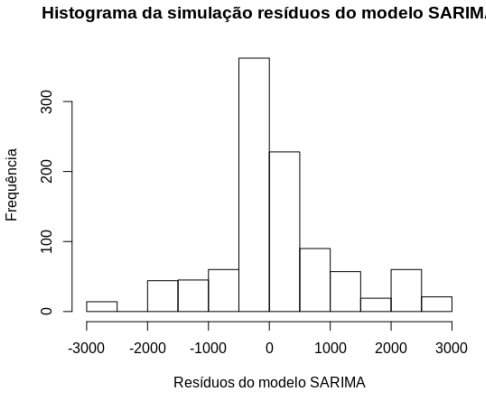

Figura 11 – Gráfico de distribuição dos resíduos simulados do modelo SARIMA.

ção normal. Como o valor-p foi menor que 0,05, obteve-se a função de distribuição acumulada empiricamente a partir dos dados resíduos.

Os resultados da análise dos indicadores estão apresentados na Tabela 5. A Tabela mostra a produção prevista em cada trimestre de teste, o valor da probabilidade do indicador I1 em cada trimestre – considerando uma média de produção histórica de aproximadamente 3.9 milhões de toneladas de cana-de-açúcar – e as probabilidades referentes ao indicador I2 com o percentual de redução configurado para 10%, 20% e 30%.

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**
Valor previsto	8603,51	3006,51	257,51	8996,51	8861,02	3264,02	515,02	9254,02
I1 - Probabilidade da produção de cana prevista ser acima da média de produção (3864,132).	100%	10%	0%	100%	100%	15%	0%	100%
I2 - Probabilidade de redução da produção em 10 % ou mais do valor previsto.	11,3%	29,2%	48,7%	11,3%	11,3%	21,1%	48,7%	11,3%
I2 - Probabilidade de redução da produção em 20 % ou mais do valor previsto.	4,3%	14,5%	48.7%	1,7%	4,3%	14,5%	44,4%	1,7%
I2 - Probabilidade de redução da produção em 30 % ou mais do valor previsto	0%	11,3%	46,2%	0%	0%	11,3%	42,5%	0%

Tabela 5 – Previsões do Modelo SARIMA(0,1,1)(0,1,0) e indicadores I1 e I2 (com percentual de redução de 10%, 20% e 30%) por trimestre.

Para análise do indicador I1 considere-se que, para sua correta operação, a usina necessita de uma produção no mínimo igual à média histórica. Assim, analisando a Tabela 5, pode-se perceber que nos trimestres T1, T4, T5 e T8 a produção de cana-de-açúcar garante funcionamento da usina em condições satisfatórias, já que a probabilidade da produção ser acima da média é de 100%. Já nos trimestres T2, T3, T6 e T7 as probabilidades de funcionamento são bem desfavoráveis, provavelmente por serem trimestres de final de safra onde a produção é pequena ou inexistente.

Já ao analisar o indicador I2, pode-se perceber que as probabilidade nos trimestres T1, T4, T5 e T8 apontam que provavelmente a produção irá sofrer pequena ou nenhuma redução. Nos trimestres T2, T3, T6 e T7, diferentemente, as probabilidades se mantêm maiores, principalmente no T3 e T7 onde a probabilidade passa de 40%. Este resultado tem relação com o já explicado no indicador I1: trimestres finais de safra tem sua produção muito baixa ou quase nula, provocando assim a probabilidade de redução de produção maior que em outros trimestres.

Outro aspecto a se observar nestes resultados é que as probabilidades de redu-

ção da produção diminuíram conforme o aumento do percentual de produção aplicado, apontando que a probabilidade de produção de cana-de-açúcar reduzir mais do que 30% é muito baixa.

6 Conclusão

Buscou-se elaborar com esse trabalho um estudo sobre o uso de séries temporais e modelos de previsão em dados de produção de cana-de-açúcar, elaborando indicadores operacionais de estimativas de produção. Os dados utilizados foram retirados dos relatórios trimestrais de produção publicados pela Usina São Martinho desde 2007 até o inicio de 2019.

Para a previsão, foi utilizado sobre os dados de produção de cana-de-açúcar o modelo SARIMA, uma vertente do modelo ARIMA que consegue capturar a sazonalidade em séries temporais. Para selecionar o melhor modelo foram utilizados o AICc, junto com as métricas de erro RMSE, ME, e MAE, como indicadores de desempenho do modelo. Ao final das análises foi selecionado o modelo SARIMA(0,1,1)(0,1,0) para executar as previsões de produção.

Com base no modelo escolhido, foram desenvolvidos dois indicadores operacionais. O primeiro indicador (I1) aponta a probabilidade da produção ser acima da média da série histórica de produção de cana, enquanto que o segundo (I2), indica a probabilidade da produção prevista ser reduzida em um dado percentual p ou mais. Para fins de experimentação, no I2 foram definidos arbitrariamente 10 %, 20 % e 30 % como percentual redução de produção.

Ambos indicadores utilizam, como base, a distribuição de probabilidade dos resíduos do modelo, obtida através da função de distribuição acumulada. Como os resíduos são uma amostra, de fato, pequena (45 amostras) – por conta do tamanho da base de dados de produção da Usina São Martinho disponível –, utilizou-se a simulação Monte Carlo para aumentar o tamanho da amostra dos resíduos do modelo SARIMA, para assim ser aplicada a função de distribuição acumulada. Foram desenvolvidas duas funções diferentes, uma para cada um dos indicadores desenvolvidos.

Todas as previsões trimestrais de produção foram calculadas em cima dos valores previstos pelo modelo SARIMA. Foram utilizados 45 trimestres para treinamento e foram previstos 8 trimestres. Para cada previsão, foram aplicados os indicadores operacionais de produção. Observou-se que há uma alta probabilidade (acima de 40%) de que a produção possa reduzir em até 30% nos trimestres finais da safra. Isto decorre do fato de que a produção nos trimestres finais tende a ter um valor global menor. Também pode-se observar que as probabilidades de redução da produção diminuíram conforme o aumento do percentual de produção aplicado, apontando que a probabilidade de produção de cana-de-açúcar reduzir mais que 30%, é muito baixa.

Os métodos apresentados ao longo do trabalho possuem grande relevância

para o processo de tomada de decisões estratégicas. Através dele o tomador de decisão é capaz de prever possíveis reduções de produção que possam levar a paradas na indústria. Apesar dos indicadores desenvolvidos no trabalho se limitarem a informar a probabilidade de redução de produção em 10%, 20% e 30% da probabilidade da produção trimestral e a probabilidade da produção ser acima da média, estes indicadores podem ser modificados para investigar outros valores que tomador de decisão tenha interesse.

Além disso, as técnicas aqui apresentadas podem ser aplicadas a quais quer tipos de dados de produção, além da produção de cana-de-açúcar, e em outras escalas de tempo, tais como diária, semanal ou mensal. Desta forma, o método aqui aplicado e os indicadores desenvolvidos poderiam ser tomados como base para decisões melhoria dos processos operacionais de uma usina de açúcar, de uma produtora de leite, de um moinho de trigo ou de qualquer outra produção agropecuária.

Para o futuro, espera-se poder aplicar o método abordado em dados que possuam um intervalo de tempo menor do que o estudado. Apesar de a previsão de produção trimestral ser de fato algo inovador, o método aplicado neste trabalho pode ser usado para a previsão do operacional em intervalos mais breves, como semanal ou diário, agregando valor para o tomador de decisão.

Referências

BOX, G. E. P.; JENKINS, G. *Time Series Analysis, Forecasting and Control*. San Francisco, CA, USA: Holden-Day, Inc., 1990. ISBN 0816211043. Citado 2 vezes nas páginas 16 e 17.

DEKKING, F. M. et al. *A Modern Introduction to Probability and Statistics: Understanding why and how*. [S.l.]: Springer Science & Business Media, 2005. Citado na página 21.

DENGEN, N. et al. Comparison of sarima, narx and bpnn models in forecasting time series data of network traffic. In: IEEE. *2016 2nd International Conference on Science in Information Technology (ICSITech)*. [S.l.], 2016. p. 264–269. Citado na página 17.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, [American Statistical Association, Taylor & Francis, Ltd.], v. 74, n. 366, p. 427–431, 1979. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2286348>. Citado na página 20.

FELIPE, I. J. dos S. Aplicação de modelos arima em séries de preços de soja no norte do paraná. *Tekhne e Logos*, v. 3, n. 3, p. 16–32, 2012. Citado na página 24.

FONSECA, P. C. D. Baer, werner. a economia brasileira. são paulo: Nobel, 1996, 416p. *Análise Econômica*, v. 14, n. 25 e 26, 2003. Citado na página 12.

HOSSAIN, M. M.; ABDULLA, F. Forecasting the sugarcane production in bangladesh by arima model. *Journal of Statistics Applications & Probability*, Natural Sciences Publishing Corp, v. 4, n. 2, p. 297, 2015. Citado 2 vezes nas páginas 13 e 23.

KAUR, H.; AHUJA, S. Time series analysis and prediction of electricity consumption of health care institution using arima model. In: DEEP, K. et al. (Ed.). *Proceedings of Sixth International Conference on Soft Computing for Problem Solving*. Singapore: Springer Singapore, 2017. p. 347–358. Citado na página 18.

KROESE, D. P. et al. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 6, n. 6, p. 386–392, 2014. Citado na página 21.

MONDAL, P.; SHIT, L.; GOSWAMI, S. Study of effectiveness of time series modeling (arima) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, Academy & Industry Research Collaboration Center (AIRCC), v. 4, n. 2, p. 13, 2014. Citado na página 17.

MONTGOMERY, D. C.; RUNGER, G. C.; CALADO, V. *Estatística Aplicada E Probabilidade Para Engenheiros* . [S.l.]: Grupo Gen-LTC, 2000. Citado na página 19.

MURTA, R. M. et al. Precipitação pluvial mensal em níveis de probabilidade pela distribuição gama para duas localidades do sudoeste da bahia. *Ciência e Agrotecnologia*, v. 29, n. 5, p. 988–994, 2005. Citado 2 vezes nas páginas 23 e 24.

PINTO, P. A. L. D. A. et al. *Aplicação do modelo ARIMA à previsão do preço das commodities agrícolas brasileiras*. [S.l.], 2008. Citado na página 24.

SAATH, K. C. d. O.; FACHINELLO, A. L. Crescimento da demanda mundial de alimentos e restrições do fator terra no Brasil. *Revista de Economia e Sociologia Rural*, scielo, v. 56, p. 195 – 212, 06 2018. ISSN 0103-2003. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-20032018000200195&nrm=iso>. Citado na página 23.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples)†. *Biometrika*, v. 52, n. 3-4, p. 591–611, 12 1965. ISSN 0006-3444. Disponível em: <https://doi.org/10.1093/biomet/52.3-4.591>. Citado na página 20.

SHIKIDA, P.; MARGARIDO, M. Uma análise econométrica de sazonalidade dos preços da canade-açúcar, estado do paraná, 2001-2007. *Informações Econômicas*, p. 13, 2009. Citado na página 15.

SHOJAEEFARD, M. H.; KHALKHALI, A.; YARMOHAMMADISATRI, S. An efficient sensitivity analysis method for modified geometry of macpherson suspension based on pearson correlation coefficient. *Vehicle System Dynamics*, Taylor & Francis, v. 55, n. 6, p. 827–852, 2017. Disponível em: <https://doi.org/10.1080/00423114.2017.1283046>. Citado na página 21.

SURESH, K. K.; PRIYA, S. R. K. Forecasting sugarcane yield of tamilnadu using arima models. *Sugar Tech*, v. 13, n. 1, p. 23–26, Mar 2011. ISSN 0974-0740. Disponível em: <https://doi.org/10.1007/s12355-011-0071-7>. Citado 2 vezes nas páginas 13 e 23.

UNICA, U. da Indústria de Cana-de-açúcar. Setor sucroenergético no brasil uma visão para 2030. p. 1–8, 2016. ISSN 0034-723X. Disponível em: <http://www.mme.gov.br/documents/10584/7948692/UNICA-CEISE_Setor+Sucroenerg%C3%A9tico+no+Brasil_Uma+Vis%C3%A3o+para+2030.pdf/80da9580-60c7-4f53-afaf-030ad01f3ebf;jsessionid=AC802B166C93389BED1AB445EAB7CD10.srv155>. Acesso em: 18 nov. 2019. Citado na página 25.

WALCK, C. *Hand-book on statistical distributions for experimentalists*. [S.l.], 1996. Citado na página 22.

WASSERSTEIN, R. L.; LAZAR, N. A. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, Taylor & Francis, v. 70, n. 2, p. 129–133, 2016. Disponível em: <https://doi.org/10.1080/00031305.2016.1154108>. Citado 2 vezes nas páginas 19 e 20.

WOOLDRIDGE, J. *Introductory Econometrics: A Modern Approach*. [S.l.: s.n.], 2003. v. 5th edition. 8 p. Citado na página 15.