Universidade Federal de Ouro Preto

Escola de Minas

Programa de Pós-Graduação em Engenharia das Construções
Mestrado Profissional em Engenharia das Construções

Dissertação

ANÁLISE ESTATÍSTICA DA PRODUÇÃO DE CIMENTO NO BRASIL E SUA RELAÇÃO COM O PIB DA CONSTRUÇÃO CIVIL.

ANA CAROLINA RODRIGUES DA ROCHA SOUZA

Ouro Preto
2021

MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Escola de Minas – Departamento de Engenharia Civil
Programa de Pós-Graduação em Engenharia das Construções
Mestrado Profissional em Engenharia das Construções – MECON

ANÁLISE ESTATÍSTICA DA PRODUÇÃO DE CIMENTO NO BRASIL E SUA RELAÇÃO COM O PIB DA CONSTRUÇÃO CIVIL

OURO PRETO
2021

MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Escola de Minas – Departamento de Engenharia Civil
Programa de Pós-Graduação em Engenharia das Construções
Mestrado Profissional em Engenharia das Construções – MECON

Ana Carolina Rodrigues da Rocha Souza

ANÁLISE ESTATÍSTICA DA PRODUÇÃO DE CIMENTO NO BRASIL E SUA RELAÇÃO COM O PIB DA CONSTRUÇÃO CIVIL

Dissertação de Mestrado apresentado ao Programa de Pós- Graduação em Engenharia das Construções do Departamento de Engenharia Civil da Escola de Minas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do título de Mestre em Engenharia das Construções.

Orientador: Prof. Helton Cristiano Gomes, D. Sc.
Coorientadora: Profª. Irce Fernandes Gomes Guimarães, D. Sc.

OURO PRETO
2021

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S729a Souza, Ana Carolina Rodrigues da Rocha .
Análise estatística da produção de cimento no Brasil e sua relação com o PIB da construção civil. [manuscrito] / Ana Carolina Rodrigues da Rocha Souza. - 2021.
78 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Helton Cristiano Gomes.
Coorientadora: Profa. Dra. Irce Fernandes Gomes Guimarães.
Dissertação (Mestrado Profissional). Universidade Federal de Ouro Preto. Departamento de Engenharia Civil. Programa de Pós-Graduação em Engenharia das Construções.
Área de Concentração: Engenharia das Construções.

1. Construção Civil. 2. Cimento - Produção. 3. Produto interno bruto (PIB). 4. Economia - Ciência de Dados. I. Gomes, Helton Cristiano. II. Guimarães, Irce Fernandes Gomes. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 624

Bibliotecário(a) Responsável: Maristela Sanches Lima Mesquita - CRB-1716

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
DEPARTAMENTO DE ENGENHARIA CIVIL

UFOP

FOLHA DE APROVAÇÃO

Ana Carolina Rodrigues da Rocha Souza

Análise estatística da produção de cimento no Brasil e sua relação com o PIB da construção civil

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia das Construções da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de mestre

Aprovada em 18 de junho de 2021

Membros da banca

Doutor - Helton Cristiano Gomes - Orientador - Universidade Federal de Ouro Preto
Doutora - Irce Fernandes Gomes Guimarães - Universidade Federal de Ouro Preto
Doutora - Clarisse da Silva Vieira Camelo de Souza - Universidade Federal de Ouro Preto
Doutor - Leandro Reis Muniz - Universidade Federal de São João del-Rei

Helton Cristiano Gomes, orientador do trabalho, aprovou a versão final e autorizou seu depósito no Repositório Institucional da UFOP em 14/07/2021

Documento assinado eletronicamente por **Irce Fernandes Gomes Guimaraes**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 14/07/2021, às 16:50, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Helton Cristiano Gomes**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 15/07/2021, às 07:26, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0193663** e o código CRC **D13E7E35**.

Referência: Caso responda este documento, indicar expressamente o Processo nº 23109.007034/2021-55 SEI nº 0193663

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: 3135591546 - www.ufop.br

DEDICATÓRIA

Dedico esta dissertação a minha mãe, que sempre me incentivou a estudar e me mostrou que o melhor caminho é a educação.

AGRADECIMENTOS

Primeiramente agradeço à Deus, que sem dúvida guiou e abençoou meus passos durante esta trajetória para a concretização desta conquista. Agradeço ao meu marido e colega de mestrado Rodrigo Souza, que me acompanhou nas viagens para Ouro Preto e sempre foi meu alicerce nos momentos exaustivos que enfrentamos durante esta jornada, acreditou incansavelmente em mim, fato este, que me motivou a nunca desistir. Gostaria de agradecer a minha Avó Geralda e minha Tia Cecília que sempre rezaram e entenderam minha ausência durante este período e sempre me ofertaram amor e carinho. Não poderia deixar de agradecer minhas irmãs Mayara Rodrigues e Maria Alice Rodrigues e também minha sobrinha Laura Barbosa que sempre traziam palavras de apoio repletas de muito amor e carinho. Agradeço ao meu pai Vilmar, que sempre se orgulhou do meu esforço e sempre me incentivou. Agradeço a minha mãe Judith, que dentro do meu coração sempre me motivou a seguir em frente e nunca desistir. Agradeço aos meus sogros Silmar e Fátima por todo apoio nesta etapa. Agradeço também a você meu filho João, tão querido, que tanto sonhei em tê-lo em meus braços e hoje faz parte desta conquista, você serve como força motor para que eu não desista dos meus sonhos. Agradeço a todos os professores do mestrado que sempre dividiram seus vastos conhecimentos, proporcionado um grande crescimento acadêmico durante este período. Em especial agradeço ao orientador professor Helton Gomes e a coorientadora professora Irce Guimarães por todo apoio, paciência e por sempre dividirem seus conhecimentos.

RESUMO

A Indústria da Construção Civil (ICC) possui considerável participação no PIB brasileiro. O PIB deste setor mantém-se, geralmente, acima dos 5% ao ano, o que resulta em um forte impacto na economia do país. Como há elevada necessidade de mão de obra e insumos, a ICC gera renda devido à oferta de empregos e movimenta a economia dos demais setores. O cimento é um recurso altamente utilizado na construção civil, sendo empregado em quase todas as obras de infraestrutura. Devido ao fato de ser o principal componente do concreto, faz com que este insumo seja amplamente utilizado, com isso, ocupa o segundo lugar entre os materiais mais utilizados no mundo, perdendo apenas para a água, e o Brasil encontra-se entre os 10 maiores produtores deste insumo. Estudos têm mostrado como os dados desempenham papel fundamental dentro de uma organização, oferecendo *insights* capazes de auxiliar nas tomadas de decisões estratégicas, assegurando competitividade no mercado. As organizações geram com alta velocidade, grandes volumes de dados que não possuem padrões em sua estrutura e podem ser captados em tempo real. Porém, esses dados tornam-se inúteis caso informações valiosas não sejam extraídas, fazendo-se necessária a utilização da Mineração de Dados (MD) e da Aprendizagem de Máquina (AM), que possuem métodos capazes de identificar padrões e correlações, realizar associações e fazer predições. Essa dissertação teve como objetivo avaliar o comportamento da produção de cimento no Brasil, baseando-se em seus dados históricos. Com isso foi possível identificar tendências e períodos de sazonalidade na série temporal (ST), bem como fazer previsões para períodos futuros. Feito isso, analisou-se a existência de correlação entre a produção de cimento e o PIB da ICC no Brasil, sendo essa hipótese confirmada por testes estatísticos. Dada a forte correlação positiva entre as ST´s, foi possível propor modelos de AM para tentar predizer o PIB da ICC com base na produção anual de cimento no Brasil. Os modelos utilizados mostraram-se eficientes, apresentando elevada acurácia, isto é, coeficientes de determinação superiores a 80%. Com base nos erros de previsão, foi possível concluir que os métodos de *Ensemble Learning* melhor se adaptaram aos dados, com destaque para o *Random Forest*. Os resultados obtidos podem auxiliar os gestores da ICC a tomarem melhores decisões, permitindo a eles prepararem-se para as oscilações do mercado.

Palavras-chave: Construção Civil, Produção de Cimento, PIB, Ciência de Dados, Aprendizagem de Máquina.

ABSTRACT

The Civil Construction Industry (CCI) has a considerable participation in the Brazilian GDP. The GDP of this sector generally remains above 5% per year, which results in a strong impact on the country's economy. As there is a high need for manpower and inputs, the ICC generates income due to the offer of jobs and drives the economy of the other sectors. Cement is a resource widely used in civil construction, being used in almost all infrastructure works. In addition, it is the main component of concrete, the second most used material in the world, second only to water, and Brazil is among the 10 largest producers of this input. Studies have shown how data plays a key role within an organization, offering insights capable of assisting in strategic decision making, ensuring competitiveness in the market. Organizations generate large volumes of data at high speed that have no standards in their structure and can be captured in real time. However, this data becomes useless if valuable information is not extracted, making it necessary to use Data Mining (DM) and Machine Learning (ML), which have methods capable of identifying patterns and correlations, making associations and make predictions. This dissertation aimed to evaluate the behavior of cement production in Brazil, based on its historical data. Therewith, it was possible to identify trends and seasonality periods in the time series (TS), as well as to make forecasts for future periods. After that, the existence of a correlation between the cement production and the GDP of the CCI in Brazil was analyzed, and this hypothesis was confirmed by statistical tests. Given the strong positive correlation between the TS´s, it was possible to propose models of ML to try to predict the GDP of the CCI based on the annual cement production in Brazil. The models used proved to be efficient, presenting high accuracy, in other words, determination coefficients greater than 80%. Based on the forecasting errors, it was possible to conclude that the methods of Ensemble Learning were better adapted to the data, with emphasis on the Random Forest. The results obtained can help CCI managers to make better decisions, allowing them to prepare for market fluctuations.

Keywords**:** Civil Construction, Cement Production, GDP, Data Science, Machine Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE SIGLAS

ABPC - Associação Brasileira de Cimento Portland

ADF - Dickey–Fuller Aumentado

AM - Aprendizado de Máquina

AR - A*utorregressive*

ARIMA - *Autoregressive Integrated Moving Average*

ARMA - Autorregressivos de Médias Móveis

BNH - Banco Nacional de Habitação

CBIC - Câmara Brasileira da Indústria da Construção

CNI - Confederação Nacional da Indústria

DM - *Data Mining*

EDA - *Exploratory Data Analysis*

EQM - Erro Quadrático Médio

FAC - Função de autocorrelação

FACP - Função de autocorrelação parcial

FGTS - Fundo de Garantia do Tempo de Serviço

FIFA - *Fédération Internationale* de *Football Association*

GPS - *Global Positioning System*

GB - *Gigabytes*

IA - Inteligência Artificial

ICC - Indústria da Construção Civil

IPEA - Instituto de Pesquisa Econômica Aplicada

KPSS - Kwiatkowski, Phillips, Schmidt e Shin

MA - *Moving Average*

MAE - *Mean Absolute Error*

MAPE - *Mean Absolute Percentage Error*

MD - Mineração de Dados

ML - *Machine Learning*

MQO - Mínimo Quadrado Ordinário

MSE - *Mean squared error*

PAC - Programa de Aceleração do Crescimento

PIB - Produto Interno Bruto

PP - Phillips e Perron

OGU - Orçamento Geral da União

RNA - Redes Neurais Artificiais

RMSE - *Root Mean squared error*

ST - Série Temporal

VABpb - Valor Adicionado Bruto a preços básicos

3 V’s - Volume, Variedade e Velocidade

SUMÁRIO

CAPÍTULO 1. INTRODUÇÃO

O advento das novas tecnologias possibilitou o aumento na geração, transmissão, armazenamento e disponibilização de dados. Este avanço na captação e geração de grandes volumes de dados em tempo real teve como desdobramento o termo *Big Data*, que representa grandes volumes de informações dentro de conjuntos de dados (ESPÍNDOLA *et al.*, 2016). Atualmente, organizações governamentais e não governamentais possuem grandes conjuntos de dados, tais como, banco de dados comerciais, governamentais e científicos, o que tornou mais complexo ao homem a interpretação minuciosa dos dados. Devido ao alto volume de dados, os métodos tradicionais de análise como planilhas ou consultas *Ad HOC*[1], inviabilizam uma análise mais aprofundada. Podem criar relatórios informativos, mas não são capazes de realizar a análise do conteúdo desses relatórios (REFFAT, GERO e PENG, 2017).

O *Big Data*, sem a devida análise das informações implícitas nos conjuntos de dados, não apresenta vantagem competitiva para as corporações. Para esse fim, utiliza-se da metodologia Mineração de Dados (*Data Mining* - *DM*), a qual é capaz de aproveitar este alto volume de dados e extrair conhecimento para transformá-los em informações úteis e estratégicas para as organizações (REZENDE e ABREU, 2013, p. 201). Por meio do MD é possível identificar padrões, correlações e informações, que combinado ao uso de modelos de Aprendizado de Máquina (*Machine Learning* ou *ML*), são capazes de realizar estimações e auxiliar na predição de cenários oferecendo suporte as corporações de maneira estratégica.

No Brasil, a indústria da construção civil (ICC) passou a impulsionar o crescimento do país após a Segunda Guerra Mundial (1939-1945), por meio da construção de infraestruturas como ferrovias, rodovias, aeroportos, usinas hidroelétricas, obras de urbanização e saneamento. A este setor, é dada a responsabilidade de fornecer edificações para impulsionar o crescimento de outras áreas, como escolas, hospitais e áreas de lazer, os quais são considerados seus produtos. Contudo, estes produtos são considerados importantes pois fazem parte de um setor de intensa fonte de atividade econômica e que possui grande contribuição para o Produto Interno Bruto (PIB). Ressaltando que a

[1] **Ad Hoc** São consultas com acesso casual único e tratamento dos dados com parâmetros nunca usados, geralmente de maneira iterativa e heurística. Consiste no próprio usuário, que gera consultas de acordo com suas necessidades de cruzar as informações de uma forma não vista e com métodos que o levem a descoberta do que se procura. (MUSARDO, 2008)

participação da ICC no PIB brasileiro mantém-se, em média, acima de 5% ao ano (FARAH, 1996).

Assim como o petróleo e aço, o cimento detém um vasto mercado. Ele é um dos principais insumos empregados na ICC, sendo utilizado em edificações, estradas e obras de infraestrutura. O setor produtivo de cimento movimenta, no mundo, cerca de U$ 250 bilhões por ano. A grande produção mundial deste produto encontra-se na China, sendo as maiores indústrias desse país a *China National Building Materials* (CNBM) **(**200Mt/ano**),** *Anhui Conch* **(**180Mt/ano**) e** *Jidong Cement* **(**100Mt/ano**).** (ARAUJO, 2020). "O cimento é o principal componente do concreto, segundo produto mais consumido no mundo, perdendo apenas para a água" (SLACK, BRANDON-JONES e JOHNSTON, 2018, p. 47).

Pressupõe-se que o cimento possui relação direta com o crescimento da ICC. Normalmente, quando há aumento no consumo da matéria-prima, há crescimento no PIB deste setor. Devido a este fato, neste estudo foi proposta uma análise do comportamento da produção de cimento no Brasil e de sua relação com o PIB da ICC. Foi possível identificar tendências, bem como propor modelos de AM para fazer previsões para períodos futuros. Comprovada a correlação entre as séries temporais (ST's), foram, também, utilizados modelos de AM para prever o PIB da ICC com base na produção de cimento. Com isso, pôde-se estimar o crescimento ou recessão do PIB da ICC, auxiliando a tomada de decisões estratégicas do setor. Para as análises foram coletados, em duas bases, dados dos anos de 2003 a 2019.

1.1 Justificativa

Atualmente, há uma crescente demanda pelo uso de tecnologias, seja para uso pessoal ou pelas corporações. Com isso, cada vez mais gera-se elevados volumes de dados que não obedecem a um padrão, podendo ser classificados como estruturados, não estruturados e semiestruturados. Um dos desdobramentos do desenvolvimento de sistemas de armazenamento de dados, é o *Big Data*, que tem como característica principal elevados volumes, velocidade e variedade de tipos de dados. Entende-se que as organizações necessitam de informações confiáveis e que sejam estratégicas para se manterem competitivas no mercado. Porém, apenas o *Big Data* não possibilita a compreensão do cenário passado e não permite a previsão de como se dará o futuro. Para extrair informações de um conjunto de dados, utiliza-se de modelos de MD, os quais possuem a

capacidade de encontrar tendências e identificar correlações e padrões em grandes conjuntos de dados. Previsões confiáveis, baseadas em dados passados, podem ser obtidas por modelos de AM. Os modelos preditivos podem ser utilizados em diversas áreas, tais como medicina, marketing, logística, bancos etc (BIECEK, 2018).

Utilizando-se a MD é possível identificar tendências de crescimento ou recessão na produção de cimento ao longo do tempo, bem como períodos sazonais. Com os modelos de AM pode-se prever o PIB da ICC, baseando-se na produção de cimento, caso haja correlação entre suas ST's. Essas informações e previsões podem auxiliar empresas da ICC no estabelecimento de estratégias e setores públicos responsáveis por infraestrutura no direcionamento de investimentos.

1.2 Objetivos

Este trabalho tem como objetivo geral analisar, por meio da MD, o comportamento da produção de cimento no Brasil e sua relação com o PIB da ICC. Após a análise, utilizar modelos de AM para prever o PIB da ICC, baseando-se na produção de cimento.

Tem-se como objetivos específicos:

- Coletar em bases confiáveis, dados relacionados à produção de cimento e ao PIB da ICC no Brasil;
- Comparar métodos para realizar previsões com base em séries temporais;
- Verificar a existência de correlação entre a produção de cimento e o PIB da ICC;
- Testar e comparar modelos de AM para prever o PIB da ICC, baseando-se na produção de cimento.

Esta dissertação está estruturada da seguinte maneira: no Capítulo 1 expõe-se a introdução, a justificativa e objetivos gerais e específicos desta dissertação. No Capítulo 2 apresenta-se uma revisão bibliográfica relacionada à ICC. O Capítulo 3 aborda o tema *Data Science* e suas áreas. No Capítulo 4 é apresentada a metodologia utilizada nesta dissertação e, no Capítulo 5 são apresentados e analisados os resultados obtidos. Por fim, são feitas as considerações finais, que se encontram no Capítulo 6.

CAPÍTULO 2. A CONSTRUÇÃO CIVIL E SUA IMPORTÂNCIA NA ECONOMIA BRASILEIRA

De acordo com Cunha (2012), a indústria da construção civil brasileira ganhou maior relevância na década de 50 no governo de Juscelino Kubitscheck. Durante este período, houve grandes investimentos em projetos para a construção de indústrias siderúrgicas, petrolíferas e de transportes e também para a construção da nova capital do país, Brasília. Durante a década de 60, a estratégia de crescimento no setor foi através da oferta de créditos para a produção imobiliária e, também para clientes adquirirem imóveis pelo Banco Nacional de Habitação (BNH).

O mesmo autor reforça ainda, que vinte anos mais tarde, em 1980, o nível de crescimento não foi compatível com os dos anos anteriores causando um declínio entre o período de 1990 a 2003. Já em 2006, houve uma tímida retomada no setor, porém a crise econômica de 2008 interferiu novamente no aumento esperado para época. Posteriormente, entre os anos de 2009 e 2010, o mercado imobiliário retomou o crescimento por meio do programa habitacional "Minha Casa, Minha Vida" que teve como principal fonte de financiamento o Plano de Aceleração do Crescimento – PAC, implantado pela gestão pública da época. Até na atualidade, houve grandes impulsos para o crescimento deste setor. Segundo CNI (2019), o setor não conseguiu sustentar o crescimento, esta desaceleração iniciou-se em 2014, devido à crise econômica e política que se instalou no país, e oferta impactos negativos na ICC até na atualidade.

Este setor teve o melhor desempenho em 2010, em que o PIB da indústria da construção civil cresceu 13,1%. Neste período foram criados alguns incentivos, tais como créditos para habitação onde o Fundo de Garantia do Tempo de Serviço (FGTS) e subsídios do Orçamento Geral da União (OGU) poderiam ser usados como principais fontes de recursos, porém os demais setores econômicos não acompanharam este expressivo crescimento. Com isso, nos anos seguintes, a participação da ICC na economia entrou em declínio. A renda real das famílias brasileiras não eram compatíveis a nova realidade, causando assim, a partir de 2012, alto endividamento e comprometendo, em 2015, uma média de 46% da renda familiar (CNI, 2019).

O setor da ICC é subdividido em três subsetores: construção pesada, montagem industrial e edificações. O subsetor da construção pesada realiza atividades voltadas a

infraestruturas viárias, urbana e industrial como serviços de terraplanagem, construção de rodovias, portos, construção de obras estruturais e de arte como pontes, contenção de encostas, obras de saneamento como redes de água e esgoto, barragens hidrelétricas e perfuração de poços de petróleo. O subsetor de montagem industrial desempenha atividades como montagem de estruturas para instalação de indústrias, sistemas de geração e transmissão de energia elétrica, sistemas de telecomunicação e montagem de sistemas de exploração de recursos naturais. O subsetor de edificações encontra-se atividades como construção de edifícios residenciais, comerciais, institucionais ou industriais e também atividades de reforma (FARAH, 1996).

A ICC desempenha um papel expressivo na economia brasileira, visto que este setor, é o segundo que mais gera emprego no país, perdendo apenas para o setor de atacado e varejo (CNI, 2019). Segundo Teixeira e Carvalho (2005), o impacto direto neste ambiente de trabalho, se dá por meio de investimentos em infraestruturas essenciais para o suporte e desenvolvimento de uma sociedade. O setor primário impulsiona as atividades secundárias e terciárias fechando assim no setor quaternário. Assim, facilita o desenvolvimento de outros setores.

Há também uma relação de interdependência com os demais setores, pois a cada um milhão produzido, gera-se cerca de 20 postos de trabalho diretos e indiretos ao ano (CNI, 2019). Este ambiente propicia ofertas de emprego para várias camadas da sociedade, devido à necessidade de grande diversidade de mão de obra. Visto que, a importância para a economia do país se dá por meio da geração de renda, advinda da oferta de empregos e necessidade de suprimentos para a construção advindos de outros setores e arrecadação de tributos (NUNES, 2019).

A maior contribuição para o PIB brasileiro é relativa ao setor da agropecuária, porém, 65% da ocupação gerada se concentra em trabalho não remunerado, ou seja, grande parte da mão-de-obra empregada é familiar, o que não eleva a remuneração paga. O mesmo não acontece na ICC, registra-se neste setor que 10% da mão-de-obra empregada não é remunerada (TEIXEIRA e CARVALHO, 2005).

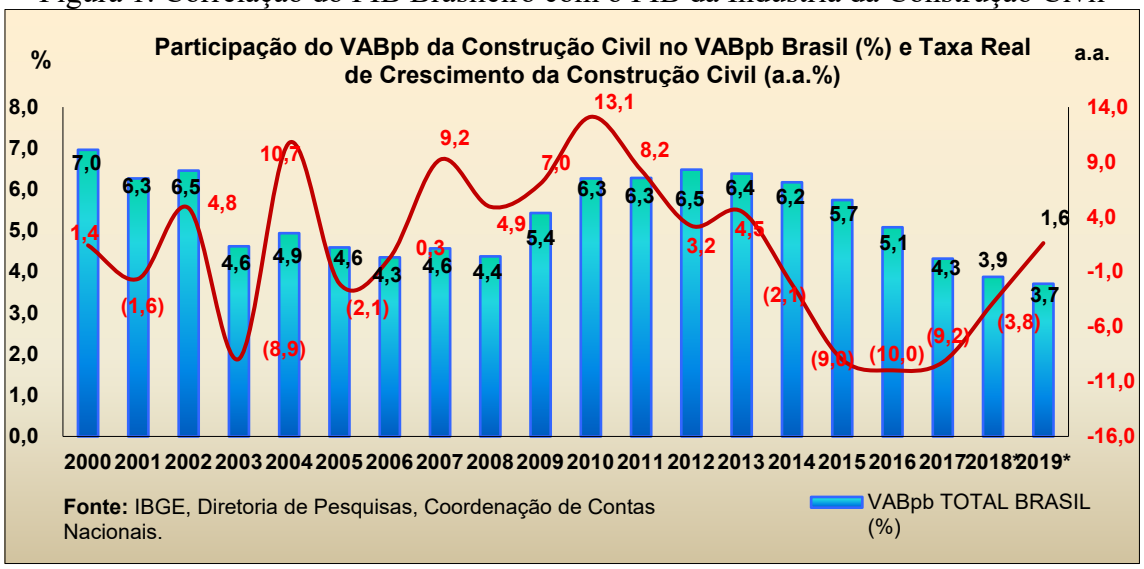
2.2 O PIB brasileiro e a sua correlação com o PIB da indústria da Construção Civil

Uma das maneiras de mensurar a riqueza acumulada de um país durante determinado período, é calcular o valor que cada setor da economia agrega ao produto final (PIB).

Este valor é determinado por meio do acúmulo de valores em três setores, são eles: agropecuária, indústria e serviços. O PIB pode alterar entre dois tipos de variações, positiva, se houve crescimento na economia e negativa, caso haja uma recessão (SOUZA *et al.*, 2015). As estatísticas do PIB podem ofertar informações importantes para economistas, exemplos são períodos de recessão que se aproximam ou sinalização de consumo de bens e/ou serviços de uma empresa de maneira irrestrita (MOYER e DUNN, 2020).

> Essa produção é medida com a soma do total do valor adicionado bruto gerado por todas as atividades econômicas do país que abrange os setores agropecuário (agricultura, extração vegetal e pecuário), industrial (extração mineral, transformação, serviços industriais de utilidade pública e construção civil) e serviços (comércio, transporte, comunicação, serviços da administração pública e outros serviços) (RIBEIRO *et al.*, 2010, p.1).

Figura 1: Correlação do PIB Brasileiro com o PIB da Indústria da Construção Civil



Fonte: CBIC (2020)

A ICC é um dos vinte subsetores vinculado a três principais atividades, está vinculada ao setor da indústria. De acordo com as informações obtidas por meio da Câmara Brasileira da Indústria da Construção Civil (CBIC, 2020), no gráfico da Figura 1 é sinalizada a forte contribuição do PIB da indústria da construção civil para a taxa real de crescimento do PIB Brasileiro, os valores apresentados nas barras em azul representam o Valor Absoluto

Bruto (VABpb[2]) total do Brasil por ano e as curvas em vermelho representam o Valor Absoluto Bruto (VABpb) da Construção Civil por ano.

É possível observar que, nos anos em que houve crescimento no PIB do país, os valores do PIB da ICC também obtiveram valores positivos. O ano de 2000 demonstra essa correlação, fechou com uma taxa positiva de 7,0% no PIB brasileiro e crescimento de 1,4% na ICC, já no ano posterior (2001), houve crescimento abaixo do esperado, com uma queda de 0,7% no PIB brasileiro, e contração de 1,6% na taxa de crescimento da indústria da construção civil. Em 2003, houve uma notória retração no PIB brasileiro (1,9%), com enorme queda de participação do PIB da indústria civil (8,9%), para o ano de 2016 em que houve a queda de 10% na participação do PIB da ICC. Em 2017, o PIB decresceu 0,8% em relação a 2016. O ano de 2019 teve um crescimento de 1,6% em relação ao ano anterior no PIB da ICC.

2.3 A origem do cimento e sua empregabilidade na Construção Civil

O cimento é um dos compostos mais utilizados no mundo, ao adicionar água ou outros materiais empregados na construção civil tais como areia, pedra-britada, pó-de-pedra e outros compostos, este material se transforma em concreto ou argamassas utilizados para a construção de casas, pontes, barragens, edifícios e estradas. Este pó fino, possui propriedades aglutinantes e aglomerantes, que endurece com a adição da água. Após endurecer, ele ganha alta resistência e durabilidade e pode ser novamente exposto à água, que não ocorre a sua decomposição (ABCP- Associação Brasileira de Cimento Portland, 2002).

O cimento é produzido pela combinação de substâncias minerais não metálicas advindas de processos de extração que passam por um forno com a temperatura média de 1.450°C, posteriormente, ocorre o processo de moagem e mistura com demais materiais. Em sua composição, o clínquer que é considerado o produto intermediário do cimento, é acrescido de adições que derivam em vários tipos de cimento. Esta matéria-prima, tem

[2] Conforme (CBIC, 2020) O Valor Absoluto Bruto a preços básicos corresponde ao valor que a atividade econômica acrescenta aos bens e serviços consumidos no seu processo produtivo. Ou seja, é a contribuição ao Produto Interno Bruto pelas diversas atividades econômicas. E, neste sentido, é considerado uma boa medida do Produto Interno Bruto setorial. É obtido por saldo entre o Valor da Produção e o Consumo Intermediário das atividades.

em sua composição, calcário e argila que segue uma proporção de 75% - 80% e 20% - 25% (CNI, 2017).

Figura 2: Maiores Produtores de Cimento

Fonte: SNIC (2019)

O Brasil possui um parque que detém 100 unidades produtoras de cimento, em que 64 são fábricas e 36 realizam o processo de moagem. Como é mostrado na Figura 2, os pontos na cor azul representam as fábricas integradas e os pontos na cor vermelha são as fábricas de moagem. Nota-se que as fábricas se concentram nas regiões Nordeste, Sudeste e Sul do Brasil. O país se encontra entre os principais produtores do mundo e ocupou a 5° posição neste *ranking* em 2013, como é destacado na Tabela 1.

Tabela 1: Maiores Produtores de Cimento

PRODUÇÃO DE CIMENTO DO MUNDO 2013					
Ranking	**Ranking**	**País / Região**	**Mil Toneladas**	**Mil Toneladas**	**Evolução**
2005	**2013**		**2005**	**2013**	**%**
1º	1º	China	1.079,60	2.300	113%
2º	2º	Índia	146,8	280	90%
3º	3º	Estados Unidos	99,4	77,8	**-22%**
5º	4º	Irã	32,7	75	129%
13º	**5º**	**Brasil**	**39,2**	**70**	**78,60%**
10º	6º	Turquia	45,6	70	54%
8º	7º	Rússia	49,5	65	31,30%
17º	8º	Vietnã	30,8	65	110.4%
4º	9º	Japão	72,7	53	**-27%**
18º	10º	Arábia Saudita	26,1	50	92%
6º	11º	Coréia do Sul	49,1	49	**-0,20%**
16º	12º	Egito	38,9	46	18%
14º	13º	México	35,4	36	1,70%
12º	14º	Indonésia	36,1	35	-3%
11º	15º	Tailândia	37	35	**-5,40%**
15º	16º	Alemanha	30	34	13%
21º	17º	Paquistão	18	32	78%
9º	18º	Itália	46	29	-37%
7º	19º	Espanha	50	20,7	**-59%**
Outros países			381,9	577,5	51%
Produção Total Mundo			**2.344,80**	**4.000**	**71%**

Fonte: Adaptado (CIMENTO.ORG, 2014)

Em 2015, o setor fabricou 65,3 milhões de toneladas de cimento, em que houve o consumo aparente do mesmo valor produzido. Houve uma queda de 9% na quantidade produzida em comparação ao ano de 2014. Este declínio justifica-se pelo fato da crise econômica instalada no país, que desencadeou falta de investimentos no setor da indústria da construção civil e também pela falta de renda devido a elevado número de desempregos, fatores que acarretaram na desaceleração do setor da ICC. Sabe-se que esta matéria-prima possui relação direta com o desenvolvimento da construção civil (CNI, 2017).

CAPÍTULO 3. CIÊNCIA DE DADOS

Atualmente, a captação e interpretação dos diferentes tipos de dados tornou-se primordial para a elaboração de estratégias eficientes para se determinar o futuro da ciência, tecnologia, economia e possivelmente tudo em nosso mundo hoje e amanhã (CAO, 2017). O universo utiliza desta estratégia de maneira intensiva e a considera como um ativo crítico. A ciência de dados é o núcleo interdisciplinar que impulsiona novas pesquisas, educação e economia em áreas distintas. Há diferentes definições e maneiras de interpretar esta ciência. Ela é um campo científico que desenvolve metodologias, teorias, tecnologias e aplicações relevantes para um conjunto de dados. Há maior abrangência nas áreas de estatística, análise de dados, AM, gerenciamento de *Big Data* e outras disciplinas, incluindo sistemas complexos, comunicações e ciências sociais (CAO, 2016).

Nos dias atuais, tem surgido áreas que desempenham atividades essencialmente relacionadas a análise de dados, nos campos de estatística e matemática. Nestas áreas é possível sintetizar uma série de parâmetros e campos de conhecimento relevantes, que inclui além dos campos citados, a comunicação, gestão e sociologia, para estudar dados seguindo o "pensamento da ciência de dados". O objetivo fundamental é utilizar os dados para transformá-los em *insight* e também o discernimento, que várias ferramentas de análise pode ofertar com maior rapidez no tempo de resposta, facilitando a tomada de decisões (CAO, 2017).

Com processos de decisão das organizações cada vez mais orientados por dados, essa ciência, oferta metodologias capazes de processar e interpretar grandes volumes de dados que são coletados por alguns recursos tecnológicos. Sua interdisciplinaridade, inclui métodos matemáticos, desenvolvimento de algoritmos, análise qualitativa, ciência da computação e, não menos importante, uma abordagem prática, com a intenção de extrair informações úteis de dados, quer estruturados em termos de informações quantitativas em um formato definido ou não estruturado, como relatórios, imagens e sons (VICARIO e COLEMAN, 2020).

3.1 Análise de Dados Exploratória - EDA

O conceito de EDA (do inglês, *Exploratory Data Analysis*) teve como pioneiro Tukey (1977) e baseia-se na análise dos dados de maneira simples por meio do uso de tabelas, gráficos, papel, caneta e fórmulas matemáticas, como soma ou subtração. A técnica trata

de ofertar maior autonomia aos dados, consiste em analisá-los de maneira livre, para que forneçam respostas e não os limita a testar apenas uma única hipótese pré-definida ou um único conjunto de cálculos isolados (TUKEY, 1977).

A maioria das técnicas de EDA são de natureza gráfica e algumas quantitativas. Acredita-se que esta é uma maneira de que os analistas possam melhor explorar e analisar os dados. Assim, sendo capaz de revelar segredos estruturais e obter alguma visão nova. Em combinação com os recursos naturais de reconhecimento de padrões que os seres humanos possuem, os gráficos fornecem, um poder incomparável para essa realização em tempos cada vez menores de respostas (CROARKIN e TOBIAS, 2012).

Atualmente, o EDA possuiu uma ampla variedade de abordagens, metodologias e técnicas para sua implementação de maneira simples, tais como, ferramentas estatísticas e bibliotecas de código aberto para análise de dados (BEZERRA *et al.*, 2019).

3.1.1 Estatística descritiva

A estatística descritiva é capaz de sintetizar as principais características observadas em um conjunto de dados através da construção de tabelas, gráficos e medidas-resumo, o que facilita a compreensão do comportamento dos dados pelo pesquisador (FÁVERO e BELFIORE, 2017). Para o entendimento e compreensão dos grandes conjuntos de dados, a técnica realiza a redução dos dados, para que possam ser facilmente compreendidos, assim, permite a visualização instantânea de todos os dados. É também atributo desta área, a obtenção de informações como médias, proporções, dispersões, tendências, índices, taxas e coeficientes (SILVA *et al.* 2018).

3.2 *Big Data*

Atualmente, o mundo lida diariamente com a análise e gestão de grandes quantidade de dados, isso é possível devido a grandes e rápidos avanços nas tecnologias existentes. A ferramenta de pesquisa Google processa cerca de 24 petabytes[3] de dados todos os dias. Já a rede social *Facebook*, obtém cerca de 10 mil fotos diariamente. Em 2012, a quantidade de dados diários aumentou, lida em média com cerca de 2,5 quintilhões bytes (BILAL *et al.*, 2016). Essa explosão de dados se dá pelas tecnologias utilizadas diariamente, como

[3] Segundo URFB (s.d) Um petabyte é uma unidade de armazenamento que tem o símbolo PB e equivale a 1024 terabytes, Um terabyte, possui 1024 Gigabytes em que 1 GB equivale a 1024 Megabytes.

vídeos digitais, músicas, *smartphones* e internet. Os dados advêm de pesquisas em websites, sensores, transações comerciais, interações em mídias sociais, áudio e *uploads* de vídeos e sinais de GPS (*Global Positioning System*) (SANTOVENA, 2013). Há previsões, que no ano de 2020, os dados gerados chegarão a 40 zettabytes (CAI e ZHU, 2015).

Em geral, o termo *Big Data* é utilizado para um conjunto imenso de dados, que possuem uma variedade de tipos, o que torna complexo o processamento por meio de técnicas tradicionais de processamento de dados. Usualmente, pode-se denominar como Big Data um conjunto de dados em que há maior complexidade de armazená-lo, processá-lo e visualizá-lo por meio de técnicas de processamento usuais (TOMAR *et al.*, 2016).

Com esta diversidade de conjunto de dados, que são capazes de se complementarem e preencher as lacunas existentes, consegue-se ofertar informações mais precisas, com isso, as empresas conseguem melhorar suas operações e são capazes de tomar decisões mais assertivas (SANTOVENA, 2013). Pesquisadores e tomadores de decisões das corporações, conseguiram perceber que através da enorme quantidade de conjuntos de dados ofertados através do Big Data, há inúmeros benefícios, dentre eles estão, entender as necessidades dos clientes e consequentemente melhorar a oferta e qualidade dos serviços. (CAI e ZHU, 2015).

3.2.1 Características do *Big Data*

O *Big Data* pode ser considerado como elevados volumes de dados armazenados. O volume de dados possui relevância, mas a definição não se limita somente a este fato, há outros atributos importantes como variedade de dados e velocidade de aquisição destes dados. Na Figura 3 é mostrado os 3 V's do *Big Data* (volume, variedade e velocidade dos dados) e suas ramificações (RUSSOM, 2011).

- Volume: refere-se a *terabytes* e, às vezes, *pentabytes*, a enorme quantidade de dados advém de registros, transações, tabelas ou arquivos. Muitas empresas armazenam dados por determinado período, assim o volume também pode ser quantificado de acordo com o tempo de armazenagem.
- Variedade de dados: este atributo advém da variedade maior de fontes como textos, sensores, áudio, vídeo, gráficos e outros dispositivos. As empresas

anteriormente lidavam com dados estruturados, hoje este cenário mudou, lida-se com dados semiestruturados e não estruturados (texto e linguagem humana),

- Velocidade: é a frequência em que os dados são gerados. Hoje a coleta se tornou um desafio, a todo tempo gera-se dados advindos de qualquer tipo de dispositivos ou sensores, máquinas de fabricação de robótica, termômetros que detectam a temperatura, microfones que monitoram barulho em uma área segura ou câmeras de vídeo. O acompanhamento deste fluxo de informações e entender os dados gerados.

Figura 3: Os 3 V's do *Big Data*

Fonte: Adaptado de (RUSSOM, 2011)

3.3 Séries Temporais

A humanidade sempre se encantou com a condição de antever o futuro, buscar formas de prever eventos antes de sua ocorrência pode ofertar benefícios. Muitos destes benefícios são, melhor aproveitamento dos eventos futuros ou preparar de forma antecipada para eventos com efeitos adversos. As séries temporais são definidas como sequência de dados quantitativos relativos a momentos específicos e estudados segundo sua distribuição no tempo (ANTUNES e CARDOSO, 2015). Já Morettin e Toloi (2018) definem como um conjunto de observações ordenadas no tempo. São exemplos de séries temporais, valores diários de poluição da cidade de São Paulo, precipitação atmosférica anual da cidade de Fortaleza e registro das marés no porto de Santos. Verifica-se no conjunto de dados dos exemplos citados, que há séries temporais discretas e contínuas. Muitas vezes as séries

discretas são transformadas em contínuas em intervalos de tempos iguais, (Δt). Assim, para realizar a análise dos dados de registro das marés do porto de Santos, converte a análise contínua observada no intervalo [0, T], em uma série discreta com *N* pontos, onde $N = \frac{\mathrm{T}}{\Delta \mathrm{t}}$. Outro exemplo que pode ser citado é, como a série temporal da precipitação atmosférica anual da cidade de Fortaleza, em que o valor da série em dado instante é obtido, acumulando-se (ou agregando) valores em tempos iguais.

Segundo Ehlers (2009), a coleção de observações feita de forma sequencial ao longo do tempo, possui como característica mais importante, em que os tipos de dados das observações vizinhas encontradas na ST, possuem dependência. Com isso, é interessante analisar e modelar o grau de dependência. Nas séries temporais a ordem das observações possui relevância para a análise, diferentemente ao que acontece nas regressões.

Há dois enfoques basicamente utilizados na análise de séries temporais. O objetivo desses enfoques é a construção de modelos para as séries, com propósitos determinados. No primeiro, a análise é feita no domínio temporal e os modelos propostos são modelos paramétricos (com um número finito de parâmetros). No segundo, a análise é conduzida no domínio de frequências e os modelos propostos são modelos não paramétricos (MORETTIN e TOLOI, 2018).

De acordo com Ehlers (2009) o objetivo principal da análise pode ser a realização de previsões de valores futuros ou verificar a estrutura da série ou sua relação com outras séries. Os principais objetivos de se realizar o estudo de uma série temporal podem ser:

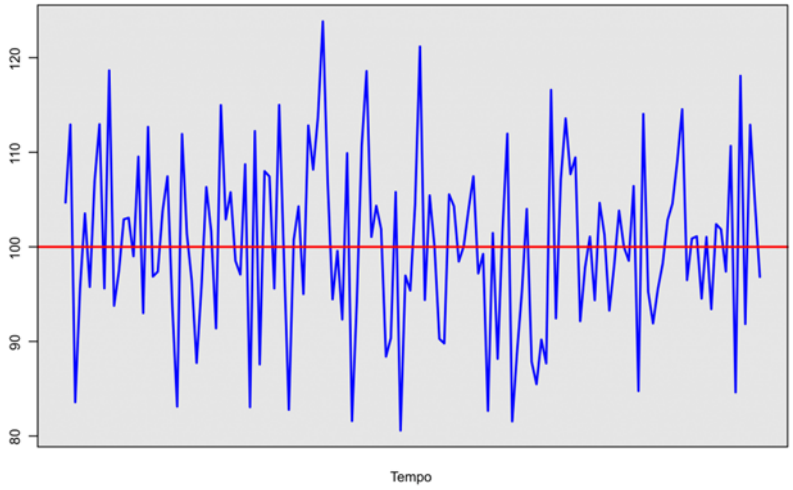
- Descrição: descrever propriedades da série tais como padrão de tendência, existência de variação sazonal ou cíclica, observações discrepantes (*outliers*) e alterações estruturais.
- Explicação: usar a ocorrência de variação de uma série para explicar a variação em outra.
- Predição: predizer valores futuros baseado em valores passados
- Controle: os valores da série temporal medem a qualidade de um processo de manufatura em que o objetivo é o controle do processo.

Pelo fato do uso crescente de séries temporais, várias pesquisas foram desenvolvidas no ramo de MD em que geralmente utiliza-se tarefas como classificação, o agrupamento, detecção de anomalias e predição (FU, 2011).

3.3.1 Estacionariedade

Frequentemente, supõe que uma série temporal é estacionária, ou seja, que ela se desenvolve no tempo aleatoriamente ao redor de uma média constante, assim, refletindo alguma forma de equilíbrio estável. Mas, grande parte das séries existentes apresentam alguma forma de não estacionariedade. Um exemplo são as séries econômicas e financeiras, a série flutua ao redor de uma reta, apresentando tendência linear através de uma inclinação positiva ou negativa. Há também casos em que apresenta uma forma de não estacionariedade explosiva, como o crescimento de uma colônia de bactérias. Um exemplo pode ser apresentado pela Figura 4, onde é apresentada uma série temporal com comportamento estacionário, observe que seus valores oscilam em torno de um nível fixo (MORETTIN e TOLOI, 2018).

Figura 4: Série de Temporal com comportamento estacionário



Fonte: (OLIVEIRA, 2019)

De acordo com Gurajati e Porter (2011) quando uma série não é estacionária, pode-se apenas estudar seu comportamento pelo período de consideração. Com isso, cada conjunto de dados, ficará específico a cada episódio. Não é possível generalizá-la e para o propósito de previsão, a série poderá ter pouco valor prático. Segundo Morettin e Toloi (2018), pelo fato da maioria de procedimentos de análise estatística de séries temporais entender que são estacionárias, há a necessidade de transformar os dados, para que a série se torne estacionária. Essa transformação, também mais usual, consiste em tomar diferenças sucessivas da série original, até torná-la estacionária. A primeira diferença da série observada (Z(t)) é definida pela Equação (1):

$$\Delta Z(t) = Z(t) - Z(t-1) \tag{1}$$

A segunda diferença, pela Equação (2):

$$\Delta^2 Z(t) = \Delta[\Delta Z(t)] = \Delta[Z(t) - Z(t-1)] \tag{2}$$

ou seja,

$$\Delta^2 Z(t) = Z(t) - 2Z(t-1) + Z(t-2) \tag{3}$$

De modo geral, a *n*-ésima diferença de Z(t) pode ser definida pela Equação (4):

$$\Delta^n Z(t) = \Delta[\Delta^{n-1} Z(t)] \tag{4}$$

3.3.2 Componentes das Séries Temporais

Geralmente o objetivo central em séries temporais, é modelar as principais características dos dados. Parte das séries tem como principais componentes a tendência, sazonalidade e resíduo. Como apresentam as equações (5) e (6):

$$Z(t) = T_t + S_t + N_t \tag{5}$$

$$Z(t) = T_t \, x \, S_t \, x \, N_t \tag{6}$$

Em que:

Z(*t*) é a série observada, T_t é a tendência, S_t é o efeito sazonal e N_t é o resíduo respectivamente. Na equação, é apresentado o método aditivo, o valor da variável de interesse é constituído pela soma dos componentes, que possuem a mesma unidade de observação Z(t). O modelo multiplicativo é geralmente empregado quando o efeito sazonal tende a aumentar de acordo com o efeito da tendência (COWPERTWAIT e METCALFE, 2009).

Para melhor compreensão dos eventos representados por uma ST, utiliza-se o conceito de decomposição da série, que consiste em escrever Z(*t*) como a soma de três componentes não observáveis. As componentes T_t e S_t geralmente são bastante relacionadas e a tendência possui influência sobre a componente sazonal. Esta situação ocorre devido

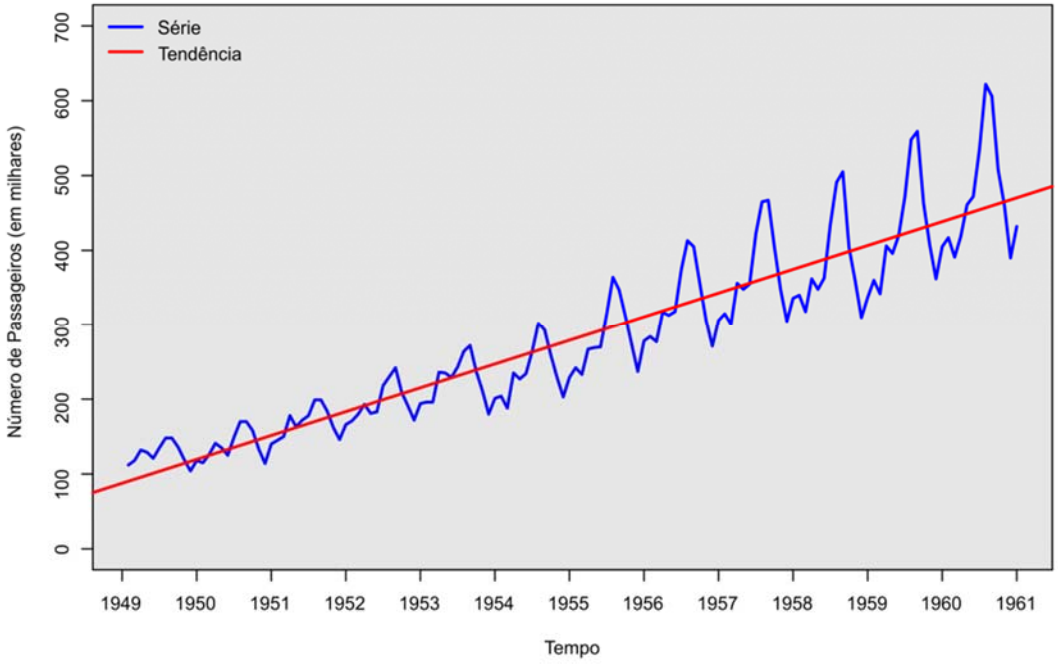
duas condições, os métodos de estimação da S_t podem ser bastante afetados caso não seja considerado a tendência ou a especificação de S_t depende da especificação de T_t.

3.3.2.1 Tendência

A tendência pode ser definida como uma mudança que ocorre lentamente a longo prazo no nível médio da série. A componente tendência pode adotar tais comportamentos como (EHLERS, 2009):

Crescimento linear: a taxa de crescimento dos dados adota um comportamento constante, a qual obedece a uma proporção linear. Na Figura 5, é mostrado o crescimento de uma ST linear em que considera os dados referentes ao número de passageiros em voos internacionais entre os anos de 1949 e 1961.

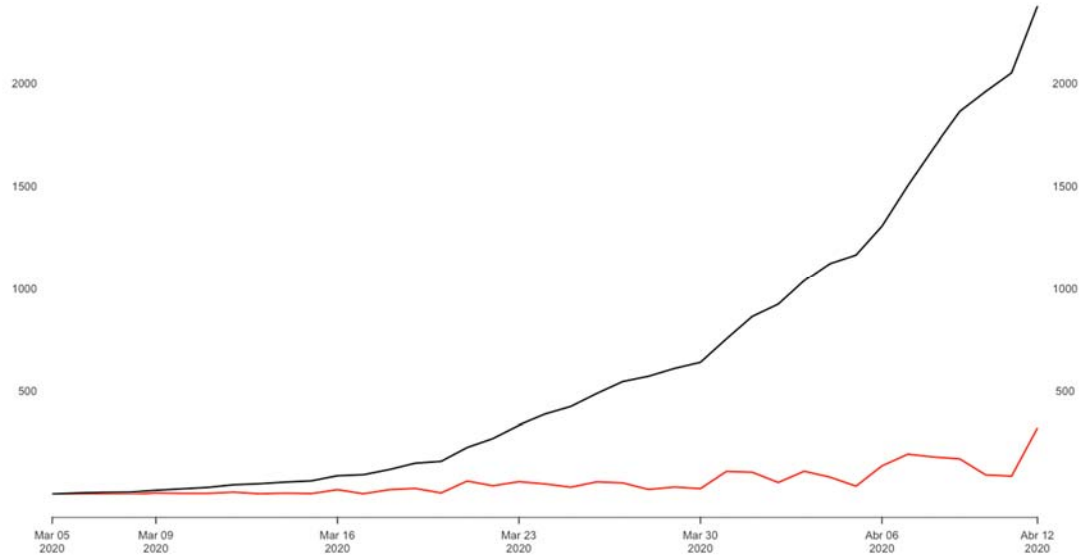
Figura 5: Número de passageiros em voos internacionais entre os anos de 1949 e 1961



Fonte: (OLIVEIRA, 2019)

Crescimento exponencial: caracteriza-se por um crescimento progressivo no percentual dos dados por período de tempo. As taxas de crescimento equivalem às propriedades de uma função exponencial. Na Figura 6 a linha preta representa o crescimento exponencial de uma ST dos casos de Covid-19 na cidade do Rio de Janeiro no período de 14 de abril a 04 de maio de 2020, enquanto a linha em vermelho, demonstra a reprodução do comportamento esperado para os próximos 21 dias, considerando a base histórica.

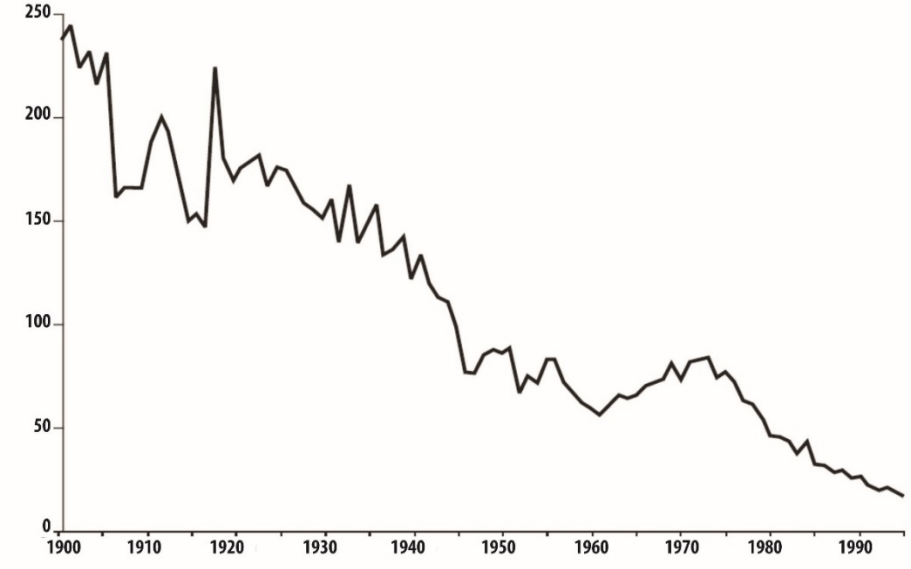
Figura 6: Série Temporal com crescimento exponencial do casos de Covid-19 na cidade do Rio de Janeiro



Fonte: Adaptado (ASSAD, 2020)

Crescimento amortecido: ocorre quando a taxa de crescimento de dados futuros é menor que os dados atuais, como mostrado na Figura 7 que apresenta a taxa de mortalidade infantil entre os anos de 1900 à 1994 na cidade de São Paulo. Observa-se no gráfico um decrescimento a partir de 1920.

Figura 7: Taxa de mortalidade infantil na cidade de São Paulo, estado de São Paulo. Brasil,1900-1994



Fonte: Antunes e Cardoso (2015)

Alguns exemplos de ST visualizados atualmente são tendência crescentes advém de observações de fenômenos de crescimento demográfico, mudança gradual de hábitos de

consumo e demanda do uso de tecnologias pela sociedade. Já para a ST que apresentam tendência decrescente cita-se séries que observam a taxa de mortalidade, epidemias e desemprego (PARMEZAN, 2016).

Segundo Reis (s.d.), há três objetivos básicos na identificação da tendência em uma ST, são eles, avaliar seu comportamento para a realização de previsões, remover a tendência T_t da ST para facilitar a visualização dos demais componentes ou utilizá-la para identificar o nível da série. Neste último caso, o nível refere-se ao valor ou faixa típica de valores que a variável pode assumir, caso não seja observado comportamento crescente ou decrescente a longo prazo.

3.3.2.2 Sazonalidade

São considerados como sazonais, fenômenos que ocorrem regularmente de ano para ano, exemplo são, o aumento da produção de leite no Brasil entre os meses de novembro a janeiro, o aumento das vendas de passagens aéreas durante o mês de janeiro ou o aumento de vendas no comércio durante o período do Natal. As relações analisadas nas séries sazonais são a ocorrência de relações entre as observações para meses sucessivos em um ano particular ou entre observações para o mesmo mês em anos sucessivos (MORETTIN e TOLOI, 2018).

As variações sazonais são oscilações de curto prazo que ocorrem dentro de um ano e se repetem de maneira sistêmica ano após ano.(REIS, [s.d.]) Este componente sazonal pode ser categorizado segundo sua variação sazonal, em dois tipos (EHLERS, 2009):

Sazonalidade Aditiva: a ST apresenta flutuações sazonais mais ou menos constantes não importando o nível global da série.

Sazonalidade Multiplicativa: Neste caso, o tamanho das flutuações sazonais varia dependendo do nível global da série.

3.3.2.3 Resíduo

O resíduo de uma ST é um componente que contém todos os movimentos que não pertencem à tendência ou à componente sazonal. São movimentos aleatórios que não são regulares e não se repetem em padrão regular (KIRCHGÄSSNER e WOLTERS, 2007).

3.4 Modelos de previsão de Séries Temporais

Prever os valores futuros baseados em uma série temporal é importante em várias áreas, tais como economia, planejamento de produção, previsão de vendas e controle de estoque (CHATFIELD, 2013). Os modelos de previsão de séries temporais são utilizados para reduzir a incerteza nos processos de tomada de decisões econômicas. Esses modelos buscam estimar valores futuros com base em valores passados. Entre os modelos mais utilizados destacam a suavização exponencial, média móvel, modelos Box & Jenkins, modelos estruturais, modelos Bayesianos e de redes neurais artificiais (RNA)(GASPAR, GONÇALVES e MATIAS, 2018).

3.4.1 Média Móvel

A técnica de média móvel simples, consiste em calcular a média aritmética de *r* observações mais recentes, como visualizado nas Equações (7) e (8):

$$M_t = \frac{Z_t + Z_{t-1} + \cdots + Z_{t-r+1}}{r} \tag{7}$$

Ou

$$M_t = M_{t-1} + \frac{Z_t - Z_{t-r}}{r} \tag{8}$$

Desta maneira, M_t é uma estimativa do nível μ_t em que o parâmetro é desconhecido e pode variar lentamente com o tempo, o r é o número de observações incluídas na média. Denomina-se média móvel, devido a cada período de observação, os dados antigos são substituídos e calcula-se uma nova média, com isso, o parâmetro varia suavemente ao longo do tempo (MORETTIN e TOLOI, 2018).

3.4.2 ARIMA

O modelo estatístico ARIMA também conhecido como abordagem de Box e Jekins (1970), é uma das metodologias mais empregadas para análise de series temporais. Consiste na análise probabilística ou estocástica da própria série temporal. O ARIMA (*p*, *d*, *q*) é uma série autorregressiva integrada de médias móveis, em que ***p*** denota o número de termos autorregressivos, ***d*** o número de vezes que a série deve ser diferenciada antes de tornar-se estacionária e ***q*** o número de termos de média móvel (GUJARATI e PORTER, 2011).

Os modelos ARIMA envolvem três processos estatísticos, são eles o (AR) autorregressão, (I) integração e (MA) médias-moveis. A autorregressão consiste em expressar a

autocorrelação das observações, ou seja, o quanto a observação anterior é capaz de influenciar no valor da próxima. O valor de (I) indica o número de diferenças que serão necessárias para assegurar a estacionariedade da ST. Já a média-móvel consiste na compreensão fatores desconhecidos que não podem ser explicados pelos valores passados (BOX *et al.*, 2015).

3.4.3 PROPHET

O algoritmo *PROPHET* foi desenvolvido para a predição de séries temporais e possui capacidade de suavizar e prever dados. Possui parâmetros intuitivos que podem ser ajustados ao modelo sem a necessidade de conhecer detalhes. Devido a esta característica, consegue lidar com características comuns da ST, tais como tendência e sazonalidade. A técnica utiliza um modelo de série temporal decomponível em que os três principais componentes são tendência, sazonalidade e feriados (TAYLOR e LETHAM, 2017).

$$\mathcal{Y}(t) = g(t) + s(t) + h(t) + \epsilon t \tag{9}$$

Em que:

- $g(t)$ é a tendência que modela as mudanças não periódicas no valor da série temporal;
- $s(t)$ representa as mudanças periódicas, sazonalidade;
- $h(t)$ representa o efeito dos feriados, que ocorrem em horários potencialmente irregulares durante um ou mais dias;
- ϵt representa quaisquer mudanças idiossincráticas que não são acomodadas pelo modelo.

3.5. Mineração de Dados

Diariamente, a quantidade de dados gerados e armazenados advém de diferentes fontes tais como, empresas, sociedade, ciências e engenharia, medicina e diversos outros aspectos do cotidiano da sociedade (HAN, KAMBER e PEI, 2011). Existe uma crescente preocupação destas instituições em coletar e armazenar este alto volume de dados, para isso, investem em recursos para a implantação de tecnologias capazes de realizar esses dois processos. Mas, nota-se que, ainda uma pequena quantidade de informações é extraída destes grandes conjuntos de dados. As empresas não conseguem analisá-los eficientemente, seja pelo fato de possuírem um alto volume ou devido suas estruturas serem restritas de recursos para tal análise, o que dificulta este processo. Cada vez mais,

torna-se necessário que as corporações busquem conhecimento por meio dos dados, que ofertam informações importantes, que são consideradas como recursos estratégicos para as organizações e asseguram maior competitividade no mercado (KANTARDZIC, 2011, p. 2).

A MD é capaz de realizar esta extração de informações de grandes volumes de dados, por meio do uso de técnicas que utilizam modelos estatísticos, algoritmos matemáticos e métodos de aprendizagem de máquina (DUA e DU, 2016, p. 5). Sabe-se que na estatística tradicional utiliza-se de aproximação já predeterminada, pois há um conjunto de resultados já esperados como resposta, o que difere da mineração de dados, que utiliza da descoberta de conhecimento sem esta prévia supervisão, apenas por uso de meios automáticos (OLSON e DELEN, 2008).

Mas, não se limita apenas a meios automáticos, pode-se utilizar de meios manuais para a descoberta de informações contidas nos dados. Essa técnica é empregada para análise exploratória de cenários em que não se sabe quais serão os resultados de saída gerados. Para isso, há a participação humana que faz junção dos esforços entre homem e máquina, em que por meio da *expertise* humana os conhecimentos gerados tornam-se de maior relevância. Esta técnica busca por informações novas, valiosas e não-triviais em grandes volumes de dados (KANTARDZIC, 2011, p. 2).

3.5.1 Tarefas desempenhadas pela mineração de dados

A MD tem como tarefas primárias a predição e a descrição. O emprego de cada uma delas, varia de acordo com a especificação do tipo de informação que deseja-se extrair, quando busca a predição, o conjunto de dados irá envolver algumas variáveis ou áreas de atuação com objetivo de predizer algum valor desconhecido de outras variáveis de interesse, já a descrição tem como objetivo descrever padrões encontrados como resposta e podem ser interpretados pelos humanos (KANTARDZIC, 2011).

A tarefas empregadas para a mineração de dados são distintas, conforme demonstra a Figura 8, quando o objetivo requerido é a predição, o atributo a ser previsto é conhecido como variável de destino ou variável dependente, e os atributos utilizados para realizar a previsão são conhecidos como variáveis explicativas ou independentes, utiliza-se as tarefas de classificação ou regressão. Quando o alvo é a descrição, é necessário derivar

padrões, para isso, emprega-se as tarefas de correlação, clusterização e sumarização (KUMAR, 2014).

Figura 8: Tarefas desempenhadas pela Mineração de Dados (*Data Mining*)

Fonte: Própria autora (2020)

3.5.1.1 Classificação

A tarefa de mineração de dados denominada classificação, é um processo que consiste em encontrar um conjunto de modelos (funções) que descrevem e, também, são capazes de distinguir classes ou conceitos. O modelo, deriva da análise de um conjunto de dados de treinamento ou um conjunto de dados de amostragem, em que já ocorreu a correta classificação de objetos. Há alguns modelos que podem ser empregados como as regras de classificação, árvore de decisão, fórmulas matemáticas ou redes neurais (HAN, KAMBER e PEI, 2011). Este processo é empregado na tarefa de predição da mineração de dados e especificadamente é utilizada para valores discretos. Os rótulos dos dados de treinamento são conhecidos a priori e os modelos são ajustados ao modelo de predição, é denominado treinamento supervisionado, também conhecido como aprendizagem supervisionada (CASTRO e FERRARI, 2016).

3.5.1.2 Regressão

A tarefa de regressão também conhecida como estimação é similar a classificação, o que difere é o tipo de atributo empregado, neste caso são valores contínuos, em que a entrada é um valor numérico e não categórico (CAMILO e SILVA, 2009).

A estimação realiza a aproximação dos valores, utilizando de uma variável numérica alvo através de um conjunto numérico e/ou variáveis categoricamente preditoras. Esses

modelos são construídos usando “completos” registros, que provém de variáveis de valor alvo. Para novas observações, utiliza-se destes valores das variáveis alvo para obter novos valores. Este tipo de tarefa pode ser empregada para a estimativa da pressão arterial sistólica de um paciente, em que o modelo é construído utilizando dados como idade, gênero, índice de massa corporal e níveis de sódio presentes no sangue, que são considerados variáveis preditoras e possui relação com a pressão arterial sistólica. Com isso, consegue-se construir um modelo de estimativa utilizando de conjunto de dados de treinamento, que poderá ser utilizado para a predição em novos casos (LAROSE e LAROSE, 2014).

3.5.1.3 Agrupamento

O agrupamento, também conhecido como *clustering*, não necessita de supervisão, ocorre através da aprendizagem não supervisionada, o que difere das tarefas de classificação e regressão, pois o conjunto de entrada não é rotulado, e a qual classe pertencerá não é conhecido a *priori*. O processo ocorre da seguinte maneira, há a separação dos dados em grupos que possuem similaridade. O processo de agrupamento (*clusterização*) é utilizado para identificar tais grupo, e cada um dos grupos formados são vistos como classes, há também como objetivo maximizar a distância interclasse e a similaridade intraclasse. Um *cluster* é conhecido como uma coleção de objetos que possuem similaridades uns com os outros e dissimilaridade com os objetos pertencentes a outros *clusters* (CASTRO e FERRARI, 2016).

3.5.2 Técnicas de Mineração de Dados

As principais atividades realizadas pela MD, divide-se em dois tipos de análises, em que a tarefa desempenhada pode ser descritiva ou preditiva. A descritiva desempenha o papel de analisar um estado passado ou atual, assim sendo capaz de descrever tendências ou padrões que residem nos dados e fornecer relatórios. A análise preditiva tem como objetivo determinar o provável resultado futuro de um evento ou a probabilidade de um estado atual que se encontra desconhecido (MCCUE, 2007).

3.5.3 Aprendizado de Máquina

A Inteligência Artificial (IA) entende que agentes são capazes de executar ações por meio de percepções externas. Turing (1950) já defendia o uso de computadores para a realização de atividades que antes, eram estritamente realizadas pelos seres humanos, acreditava-se que os computadores, caso fossem previamente treinados para a realização

destas operações, são capazes de realizá-las de maneira eficaz. Segundo Fernandes e Chiavegatto Filho (2019) este campo discute como agentes físicos (máquina) e lógicos (programas de computador) são capazes de realizar a tomada de decisão com base em dados captados por sensores ou alimentados por intervenção humana.

Com o advento do uso da IA, a subárea AM vem ganhando destaque, esta área utiliza de algoritmos de aprendizado para realizar predições. Ao invés do humano construir modelos manuais e alterar regras para análise de grandes quantidades de dados, o aprendizado de máquina consiste no uso de algoritmos que capturam o conhecimento de maneira gradual, com isso, constantemente melhoram a performance dos modelos preditivos (RASCHKA e MIRJALILI, 2017).

O AM é uma das principais áreas de pesquisa de IA e há constante busca por programas de computadores que sejam capazes de aprender automaticamente padrões complexos e a tomar decisões inteligentes baseadas em dados. Há quatro tipos de aprendizado (HAN, KAMBER e PEI, 2011):

- Aprendizado supervisionado: é sinônimo das tarefas de classificação e regressão, em que utiliza-se da supervisão humana para a criação de dados de treinamento rotulados.
- Aprendizado não supervisionado: os algoritmos buscam padrões e desempenham o papel de realizar a análise do conjunto de dados, é empregado na tarefa de agrupamento.
- Aprendizado semi-supervisionado: é uma classe de técnicas que utiliza de dados rotulado e não rotulados ao aprender um modelo. Os exemplos rotulados são usados para aprender modelos de classe e exemplos não rotulados são empregados para refinar os limites entre as classes.
- Aprendizado por reforço: permite aos usuários desempenhar um papel ativo no processo de aprendizado. Pode pedir a um usuário para rotular um exemplo, que pode ser de um conjunto de exemplos não rotulados ou sintetizado pelo programa de aprendizagem.

3.6 Técnicas de Previsão de Aprendizado de Máquina

3.6.1. Regressão Linear

Através da regressão, consegue-se modelar a relação das variáveis de respostas, conhecidas como variáveis de saída ou dependentes (variável *y*) com os preditores também denominada como variáveis independentes (variável *x*). A forma de relacionamento de uma regressão pode ser denominada de três maneiras, quando o relacionamento funcional entre as variáveis dependentes e independentes é conhecido, mas há existência de valores desconhecidos, que podem ser estimados por meio do conjunto de treinamento, denomina-se a regressão como paramétrica. Quando não há conhecimento prévio sobre a função que está sendo estimada, é conhecida como não paramétrica. Mesmo que a função estimada passe por alguns ajustes de parâmetros livres, o conjunto de formas com classes de funções é bastante amplo. Em alguns casos, a relação entre as variáveis de entrada e saída possuem uma relação determinística, através da variável independente x, consegue-se determinar sem erro a variável dependente y. Um exemplo que demonstra este tipo de regressão é a utilização da formula física para o cálculo do peso $P = m\ x\ g$, sabe-se que g = 9,8 m/s², ao determinar o valor da massa m de uma pessoa consegue determinar o peso com acurácia (CASTRO e FERRARI, 2016).

3.6.2 Regressão Linear Simples

A regressão linear simples é uma função que pode ser descrita em linha reta, onde *x* é a variável independente e *y* é a variável dependente, os modelos em que se utiliza uma única variável regressora ou preditora, no caso x para a variável dependente y, é denominado como regressão linear simples. O modelo de regressão linear é descrito pela Equação (10):

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{10}$$

As variáveis β_0 e β_1 são constantes desconhecidas e ε é um componente de erro aleatório. O regressor *x* é controlado pelo analista de dados e medido com erro desprezível, a reposta é encontrada através da variável aleatória *y* (MONTGOMERY, PECK e GEOFFREY, 2012).

3.6.3 Regressão Linear Múltipla

No mundo real, frequentemente os cenários não se limitam há uma única variável preditora, ou seja, utiliza mais de uma variável independente (*x*) para explicar a variação na variável dependente (*y*). Esta técnica estatística de análise de dados denominada como regressão linear múltipla. Sua formulação básica é definida pela Equação (11):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (11)$$

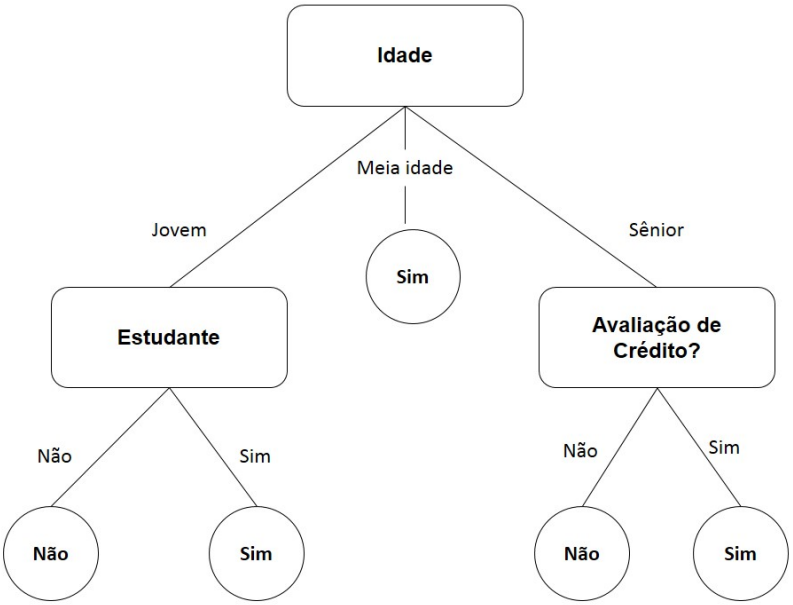
Em que y apresenta o acumulado de x e x_2, as variáveis β_0 e β_1 e β_2 são constantes desconhecidas e ε é um componente de erro aleatório (GUPTA e GUTTMAN, 2018).

3.6.4 Árvore de Decisão

Árvore de decisão, também conhecida como *Decision Tree* (DT), é uma técnica que se baseia na lógica, o processo de tomada de decisões, ocorre por meio da dedução de um conjunto de regras. É realizado de maneira particionada, representada pelos nós da árvore, até a chegada do nó folha, em que a decisão é tomada. A vantagem consiste na simplicidade para serem atendidas e compreendidas (TAN *et al.*, 2019). Consiste em uma técnica de modelagem preditiva empregada em tarefas como classificação e clusterização. A estrutura de uma árvore se assemelha com um fluxograma, onde cada nó denota um novo teste em um valor atributo, cada ramificação representa um resultado do teste e as folhas da árvore representam classes ou distribuições de classe (HAN, KAMBER e PEI, 2011).

A Figura 9 esboça uma árvore de decisão que prevê se um cliente de uma empresa irá ou não comprar um computador. Os nós internos são denotados por retângulos e os nós folha são denotados por círculos. Cada nó interno (não folha) representa um teste em um atributo. Cada nó folha representa uma classe (compra computador sim ou compra computador não) (HAN, KAMBER e PEI, 2011). Nesta árvore a raiz e cada nó é rotulado com uma pergunta, os arcos que provêm de cada nó representam cada resposta possível para a questão associada. Cada nó folha representa uma previsão de uma solução para o problema em consideração (DUNHAM, 2003).

Figura 9: Exemplo de uma árvore de decisão



Fonte: Adaptado HAN, KAMBER e PEI (2011)

3.6.5 *Ensemble Learning*

Os métodos de *ensemble learning*, também conhecido como aprendizado por agrupamento, consiste em um modelo composto pela combinação de diversos modelos de predição que tem como objetivo aumentar a acurácia da técnica aplicada (HAN, KAMBER e PEI, 2011).

3.6.5.1 *Random Forest*

A precisão de um classificador está relacionada a menor taxa de erro que ele pode ofertar nas suposições. Por isso, métodos de *ensemble learning* ofertam maior acuracidade ao modelo de predição, deve-se ao fato que um conjunto de classificadores são utilizados e cujas as decisões individuais são combinadas de alguma forma, normalmente por votação ponderada ou não ponderada para classificar novos exemplos (DIETTERICH, 2000). A técnica de *Random Forest* (florestas aleatórias) é a combinação de várias árvores, o que forma uma “floresta”, em que depende dos valores de um vetor aleatório amostrado de forma independente e que a distribuição é a mesma para todas as árvores da floresta. Durante a classificação, cada árvore vota e a classe mais popular é retornada (KANTARDZIC, 2011).

3.6.5.2 *Gradient Boosting*

O poder preditivo de uma árvore de decisão é normalmente inferior, mas pode então, ser aprimorado através da formação de um comitê de previsão, que possui como ideia central o uso das técnicas de *boosting*. Esta técnica segue o paradigma sequencial, em que busca atuar nos erros obtidos da etapa anterior, com o objetivo de reduzir os resíduos da previsão (MAYRINK, 2016). O *Gradient Boosting* combina estrategicamente árvores de decisão adicionais corrigindo erros cometidos por seus modelos de base anteriores, portanto, melhora potencialmente a precisão da previsão (ZHANG e HAGHANI, 2015).

3.7. Métricas de Avaliação de Desempenho de modelos para estimação

Entre as fases de modelagem e implantação, há a avaliação dos modelos, onde é feita a análise quanto à eficácia e qualidade deles. Durante a fase de modelagem gera-se mais de um modelo candidato e, nesta etapa, é determinado o que apresenta melhores resultados para posteriormente ser implantado (LAROSE e LAROSE, 2014).

3.7.1 Coeficiente de determinação

O coeficiente de determinação (R^2) é uma medida amplamente utilizada para analisar a adequação de um modelo de regressão. Definido pela Equação (12), indica a porcentagem da variabilidade total que é explicada pelo modelo. Equivale à proporção da variância dos valores de *y* que pode ser atribuída à regressão com a variável *x*, logo é um indicador do tipo quanto maior, melhor. Quanto mais perto R^2 se aproxima de 0, menor é a relação entre a variável dependente e as variáveis independentes. Na Equação (12), SQ_R é a soma dos quadrados da regressão, calculado pela Equação (13), e SQ_E é soma dos quadrados dos erros apresentados, calculado pela Equação (14), e SQ_T soma total corrigida dos quadrados de y, calculado, pela Equação (15) (MONTGOMERY e RUNGER, 2018).

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T} \tag{12}$$

$$SQ_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{13}$$

$$SQ_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (14)$$

$$SQ_T = SQ_R + SQ_E \quad (15)$$

3.7.2 Erro Médio Quadrático

O erro quadrático médio (*Mean Squared Error* - MSE) é uma medida da eficácia de uma estimativa, e é definido como o valor esperado da diferença quadrática entre a estimativa e o valor real. É utilizado para determinar a previsão específica e mensurar a precisão, ao invés de observar a diferença média (KUMAR, 2014). O MSE é definido pela Equação (16).

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (\hat{\theta} - \theta)^2 \quad (16)$$

Em que n é o número de dados da população de teste, $\hat{\theta}$ é o valor encontrado pelo método de estimação e θ é o valor real obtido pela Base de Dados (OLIVEIRA, 2017).

3.7.3 Erro Médio Absoluto

O erro médio absoluto (*Mean Absolute Error* - MAE) é uma métrica de avaliação utilizada em modelos de regressão. O MAE, calculado pela Equação (17), de um modelo em relação a um conjunto de teste é a média dos valores absolutos dos erros de predição individuais em todas as instâncias do conjunto de teste. Cada erro de predição é a diferença entre o valor verdadeiro e o valor predito para a instância.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{\theta} - \theta| \quad (17)$$

Em que n é o número de instâncias de teste, $\hat{\theta}$ é o valor encontrado pelo método de estimação e θ é o valor real obtido pela Base de Dados (SAMMUT e WEBB, 2011).

3.7.4 Erro Médio Absoluto Percentual

O erro percentual médio absoluto (*Mean Absolute Percentage Error* - MAPE) é calculado utilizando o erro absoluto para cada período dividido pelo valor real observado para esse período. Em seguida, calcula a média desses erros percentuais absolutos. O MAPE é uma medida de erro que calcula o desvio percentual entre os dados reais e os dados de previsão. O valor MAPE pode ser calculado pela Equação (18) (KRISMA, AZHARI e WIDAGDO, 2019).

$$MAPE = \left(\frac{100}{n}\right)\sum_{t=1}^{n}\frac{\left|Y_t - \widehat{Y_t}\right|}{Y_t} \tag{18}$$

Em que *n* é o número de observações para as quais as previsões foram feitas, $\widehat{Y_t}$ é o valor encontrado pelo método de estimação e Y_t é o valor real obtido pela Base de Dados (MONTGOMERY, JENNINGS e KULAHCI,2008)

3.7.5 Raiz do erro quadrático médio

A métrica Raiz do erro quadrático médio também conhecido como *Root Mean Squared Error (RMSE),* coloca em ênfase os erros absolutos maiores e é fornecida pela Equação 19:

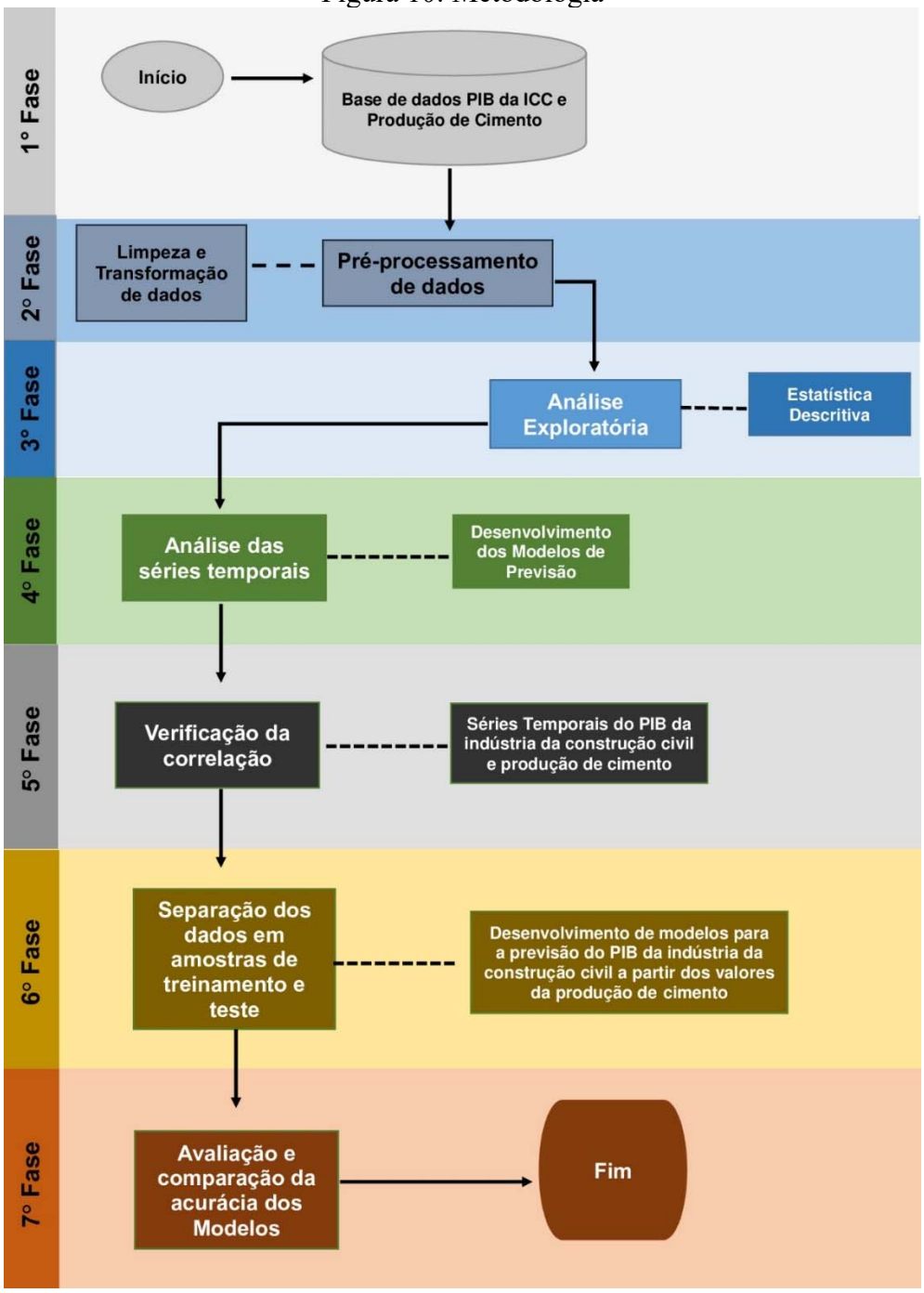
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta} - \theta)^2} \tag{19}$$

Em que *n* é o número de instâncias de teste, $\hat{\theta}$ é o valor encontrado pelo método de estimação e θ é o valor real obtido pela Base de Dados (SAMMUT e WEBB, 2011). O RMSE demonstra a raiz quadrada das diferenças entre os valores preditos e os valores reais observados ou a média quadrática dessas diferenças (NABAVI-PELESARAEI *et al.* 2021)

CAPÍTULO 4. MATERIAIS E MÉTODOS

Neste capítulo são apresentadas as etapas realizadas no desenvolvimento desta dissertação. A metodologia proposta neste trabalho foi particionada em 7 etapas, conforme ilustra a Figura 10.

Figura 10: Metodologia



Fonte: Própria autora (2020)

Na 1° fase foi realizada a coleta de dados para a análise e criação dos modelos preditivos, onde utilizou-se duas bases distintas contendo informações sobre o PIB da ICC e o

quantitativo das toneladas de produção de cimento no Brasil. Os dados referem-se às ST´s correspondentes ao período entre os anos de 2003 e 2019 e foram obtidas dos bancos de dados do Instituto de Pesquisa Econômica Aplicada (IPEA) e da Câmara Brasileira da Indústria da Construção (CBIC).

Para uma melhor análise e desenvolvimento dos modelos, a 2° fase consistiu no pré-processamento dos dados. Foi realizada uma análise, limpeza e transformação dos dados, bem como a detecção de *outliers* e identificação de valores faltantes (*missing values*). Nesta dissertação não foram detectados dados faltantes dentro do conjunto coletado e o ano de 2019 foi descartado por não possuir os dados dos doze meses. Devido ao fato de a coleta ter sido realizada em novembro de 2019, o quantitativo das toneladas de cimento produzido nos últimos meses do ano não estava disponibilizado na Base de Dados. A Tabela 2 descreve o número de dados empregados na análise.

Tabela 2: Perfil do conjunto de dados do cimento

Quantidade de anos	**Quantidade de dados por ano**	**Tamanho da amostra**
16	12	192

Fonte: Própria autora (2021)

Como pode ser visto na Tabela 2, foram coletados dados referentes a 16 anos, 2003 a 2018, totalizando uma amostra de tamanho 192.

A 3° fase corresponde a uma análise exploratória descritiva dos dados coletados, em que foram construídos gráficos e tabelas, com o intuito de gerar informações relevantes e identificar anormalidades no comportamento das ST´s. Foram analisadas ocorrências de acontecimentos que podem ter causado impactos significativos na produção de cimento e no PIB da ICC ao longo do período, buscando, assim, justificativas para as possíveis alterações.

Na 4° fase foi feita uma análise preliminar da ST da produção de cimento brasileira em relação à estacionariedade, tendência e sazonalidade. Para o estudo da estacionariedade emprega-se testes estatísticos, como por exemplo, os de raiz unitária, sendo os principais, o teste Dickey–Fuller Aumentado (ADF), o teste Phillips e Perron (PP) e o teste Kwiatkowski, Phillips, Schmidt e Shin (KPSS). O teste PP tende a aceitar a hipótese de não estacionariedade com maior frequência, mesmo quando esta é falsa. Dessa forma, a

fim de dar maior robustez aos resultados, geralmente realiza-se o teste clássico ADF e o teste de KPSS (FREITAS e SÁFADI, 2015).

O Dickey–Fuller (DF) é um teste estatístico de hipótese nula. Quando a ST possui raiz unitária, ela é não estacionária, caso não possua, é considerada estacionária e pode ser calculada pela Equação (20).

$$y_t = \rho y_{t-1} + u_t \qquad -1 \leq \rho \leq 1 \tag{20}$$

Na Equação (20), o u_t é um termo de ruído branco, quando o ρ =1 há a ocorrência da raiz unitária, ou seja, a ST é não estacionária (GUJARATI e PORTER, 2011).

O teste ADF difere apenas pelo fato de considerar a existência de alguma estrutura de autocorrelação para os erros da equação de teste. O teste ADF é calculado através da Equação (21). Caso a estrutura não seja considerada, há perda de eficiência do estimador de Mínimo Quadrado Ordinário (MQO) (GUJARATI e PORTER, 2011).

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \sum_{i=1}^{m} \alpha_i \, \Delta Y_{t-1} + \varepsilon_t \tag{21}$$

Em que β_1 é o termo independente, β_2 o coeficiente de tendência, ε_t um termo de erro de ruído branco puro, δ o coeficiente da presença de raiz unitária, m o número de atrasos utilizados da série, α_i os coeficientes de ΔY_{t-1} para aproximar a estrutura de modelos autorregressivos de médias móveis (ARMA) dos erros (FRACARO, 2018).

O teste PP utiliza métodos estatísticos não paramétricos para tratar da correlação serial nos termos de erro sem adicionar os termos de diferença defasados. A distribuição assintótica do teste PP é a mesma da estatística do teste ADF (GUJARATI e PORTER, 2011). O teste PP é descrito a partir da regressão dada pela Equação (21), a mesma utilizada para o teste ADF (FRACARO, 2018).

O teste KPSS tem como objetivo determinar a estacionariedade em uma ST, considerando como hipótese nula (H_0) que a série é estacionária ou estacionária em torno de uma tendência determinística. A hipótese alternativa (H_1) afirma que há um caminho aleatório na ST (KWIATKOWSKI *et al*.,1992).

As hipóteses dos testes KPSS e ADF não são iguais, no teste KPSS a hipótese nula é de que a série seja estacionária, já no teste ADF, quando a ST possui raiz unitária, ela é não

estacionária (FRACARO, 2018). Silveira, Mattos e Konrath (2017) propõem avaliar a estacionariedade considerando um modelo com tendência, passeio aleatório e erro, conforme apresenta a Equação (22), em que r_t é um passeio aleatório calculado pela Equação (23).

$$Y_t = \xi D_t + r_t + \varepsilon_t \tag{22}$$

$$r_t = r_{t-1} + \mu_t \tag{23}$$

O valor inicial r_0 é fixo e serve como intercepto e μ_t é uma Distribuição Normal.

Após a análise de estacionariedade, através do Teste de Dickey–Fuller, foi feita a predição da produção de cimento para os anos futuros. Para isso, foram selecionados os métodos Média Móvel, ARIMA e PROPHET. A partir das previsões obtidas, juntamente com o cálculo dos seus erros, foi possível identificar qual método apresentou o melhor desempenho.

Em estudos que envolvem duas ou mais variáveis é comum haver o interesse de identificar se há relação entre o comportamento delas. A medida que apresenta o grau de relacionamento entre variáveis quantitativas denomina-se coeficiente de correlação (BOUGARD e GOMES, 2017). A 5° fase buscou identificar se há relação, positiva ou negativa, entre as ST´s analisadas, ou seja, determinar a influência no valor do PIB da ICC gerada pela variação na produção de cimento. Para investigar a relação entre as variáveis foram utilizadas as correlações de Pearson e de Spearman, descritas pelas Equações (24) e (25), respectivamente.

O coeficiente de Pearson (ρ_p) mede o grau da correlação linear entre duas variáveis quantitativas ***x*** e ***y***, sendo calculado pela Equação (23).

$$\rho_p = \frac{\sum_{i=1}^{n}(x_i - \bar{x}\,)(y_i - \bar{y}\,)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}\,)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}\,)^2}} \tag{24}$$

Em que x_i e y_i são pares de *n* observações das variáveis *x* e *y* e ρ_p varia entre -1 e 1 (ORIGUELA, 2018). O valor zero indica que não há relação linear entre as duas variáveis e, quanto mais próximo o valor absoluto for de 1, mais forte é a relação linear entre as duas variáveis. O sinal indica o sentido da relação entre as variáveis, de forma que, sinal

positivo indica que as duas variáveis variam no mesmo sentido e, sinal negativo indica que as variáveis variam em sentido inverso (SOUSA, 2019).

O coeficiente de Spearman (ρ_s) é uma medida de correlação não-paramétrica. Ao contrário do coeficiente de Pearson, ρ_s pode indicar uma associação não necessariamente linear entre as variáveis, mas sempre indicará que a relação entre as variáveis será crescente ou decrescente. O coeficiente de Spearman é calculado pela Equação (24).

$$\rho_S = \frac{\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right)}{\frac{n(n^2-1)}{12}} \tag{25}$$

Na Equação (24), R_i e S_i são os postos das variáveis e n é o número de observações. O coeficiente de Spearman também varia entre -1 e 1, de forma que indicará uma relação crescente ou decrescente (ORIGUELA, 2018).

Na 6° fase foram utilizados métodos de AM, foram empregados os algoritmos de Regressão Linear, Árvore de Decisão, *Randon Forest* e *Gradient Boosting,* para a predição de valores futuros para o PIB da ICC, com base na produção anual de cimento no Brasil. Para isso, primeiramente, os dados foram divididos em dois conjuntos, treino e teste, buscando, assim, evitar o sobreajuste (*overfitting*) nos métodos de predição. Segundo Ishizaki (2018), o *overfitting* pode acontecer quando um método apresenta bom desempenho para o conjunto de dados utilizado na sua criação, e não apresenta a mesma acurácia para outros conjuntos de dados. A Tabela 3 apresenta o particionamento final dos conjuntos de dados em que empregou 180 entradas (93,75%) para treinamento e 12 entradas (6,25%) para teste.

Tabela 3: Particionamento dos dados - Cimento

Partição	**%Relativo**	**Qtde de entradas**
Treinamento	93,75%	180
Teste	6,25%	12

Fonte: Própria autora (2021)

Na 7° fase foram comparadas as acurácias dos métodos de AM, para isso, utilizou-se as métricas coeficiente de determinação (R^2) e RMSE, que avaliam a diferença entre os valores preditos e os reais. Feito isso, determinou-se o melhor método de predição.

As análises e implementações computacionais feitas neste trabalho foram realizadas utilizando a linguagem de programação *Python*, que apresenta uma diversidade de bibliotecas que auxiliam neste processo.

CAPITÚLO 5. RESULTADOS E ANÁLISES

Neste capítulo são apresentados as análises e os resultados das previsões obtidas pelos modelos.

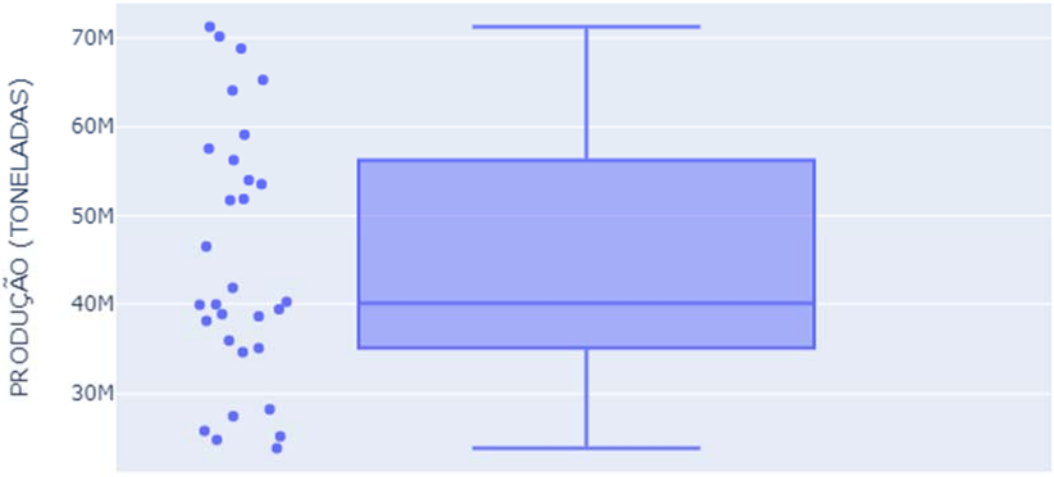
5.1 Análise de valores discrepantes - *outliers*

Feita a limpeza e a transformação dos dados e a identificação de valores faltantes é necessário analisar a ocorrência de *outliers*. Uma ferramenta importante para a análise de uma ST são os gráficos, com eles é possível identificar características como tendência, sazonalidade, variabilidade e valores discrepantes (*outliers*) (MORETTIN e TOLOI, 2018). Para analisar a ocorrência de *outliers* nos conjuntos de dados relacionados à produção de cimento e do PIB da ICC foram construídos *boxplots*.

De acordo com Doane e Seward (2014) os Quartis são denominados por Q_1, Q_2, Q_3, referem-se aos pontos que dividem os dados em quatro grupos de tamanhos aproximadamente iguais, que respectivamente, correspondem a 25%, 50% e 75% do conjunto de dados. O gráfico *box-plot* é representado da seguinte maneira, o centro é o valor de Q_2, a variabilidade é a largura da caixa que inicia-se em Q_1e finaliza em Q_3, a amplitude inicia-se na cerca inferior, representada por x_{min} e finaliza na cerca superior, representada por $x_{máx}$. Valores que se encontram fora do limite das cercas representadas por x_{min} e $x_{máx}$, são considerados como valores discrepantes que representam *outliers.*

O Gráfico 1 apresenta um *boxplot* para a ST da produção anual de cimento no período entre 2003 e 2018. Observa-se que não há presença de *outliers*, ou seja, não há dados além das extremidades dos x_{min} e $x_{máx}$. O *boxplot* apresenta as seguintes medidas estatísticas com os valores de: x_{min} é de aproximadamente 23 mil toneladas, o primeiro quartil (Q_1) é aproximadamente 35 mil toneladas, a mediana/segundo quartil (Q_2) é aproximadamente 40 mil toneladas, o terceiro quartil (Q_3) é aproximadamente, 56 mil toneladas e $x_{máx}$ é aproximadamente 71 mil toneladas.

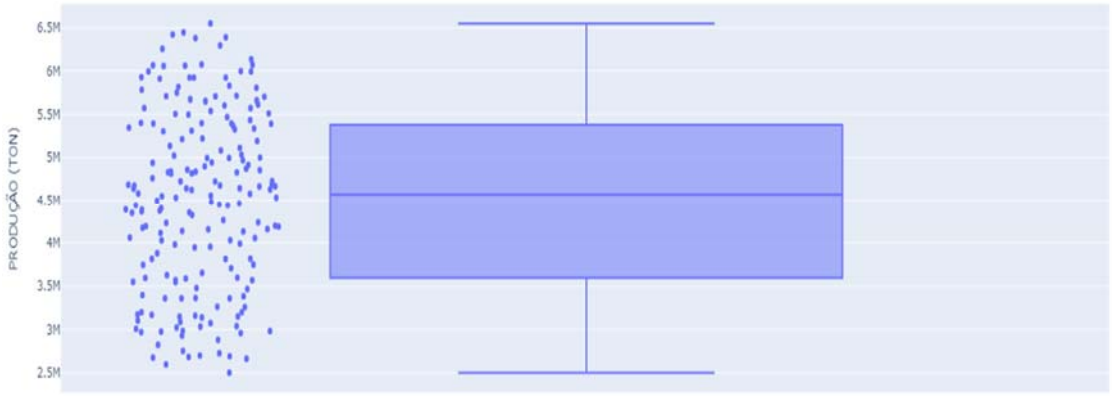
Gráfico 1: *Boxplot* da produção anual de cimento entre 2003 e 2018



Fonte: Própria autora (2021)

O Gráfico 2 apresenta um *boxplot* para a ST da produção mensal de cimento entre os anos de 2003 e 2018. Observa-se que não há presença de *outliers*, ou seja, não há dados além das extremidades dos x_{min} e $x_{máx}$. O *boxplot* apresenta as seguintes medidas estatísticas com os valores de: x_{min} é de aproximadamente 2,5 mil toneladas, o primeiro quartil (Q_1) é aproximadamente 3,5 mil toneladas, a mediana/segundo quartil (Q_2) é aproximadamente 4,5 mil toneladas, o terceiro quartil (Q_3) é aproximadamente, 5,4 mil toneladas e $x_{máx}$ é aproximadamente 6,6 mil toneladas.

Gráfico 2: *Boxplot* da produção de cimento mensal entre os anos de 2003 e 2018

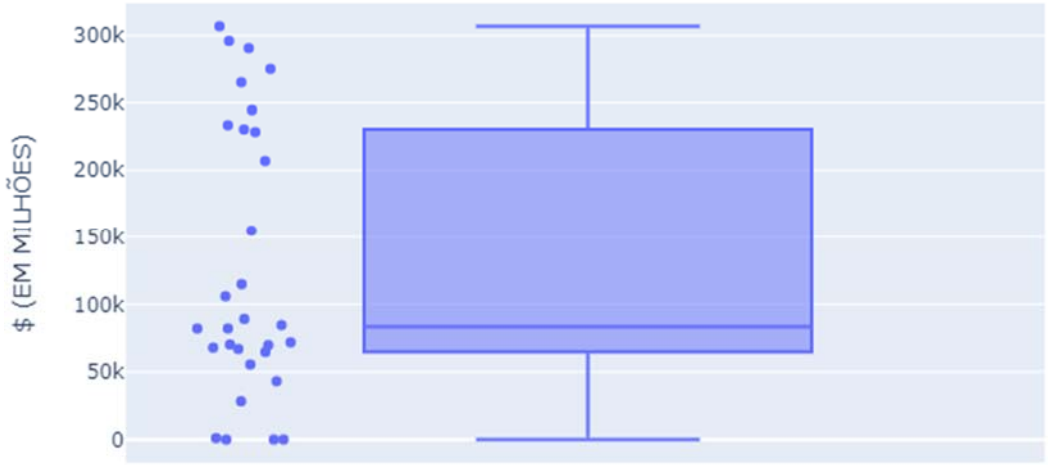


Fonte: Própria autora (2021)

O Gráfico 3 apresenta um *boxplot* para a ST do PIB anual da indústria da construção civil entre 2003 e 2018. Observa-se que não há presença de *outliers*, ou seja, não há dados além das extremidades dos x_{min} e $x_{máx}$. O *boxplot* apresenta as seguintes medidas estatísticas com os valores de: x_{min} é de aproximadamente $0,77 milhões, o primeiro quartil (Q_1) é aproximadamente $ 65 milhões, a mediana/segundo quartil (Q_2) é

aproximadamente $ 83 milhões, o terceiro quartil (Q_3) é aproximadamente, $ 230 milhões e $x_{máx}$ é aproximadamente $ 306 milhões.

Gráfico 3: *Boxplot* do PIB da Construção Civil entre os anos de 2003 e 2018



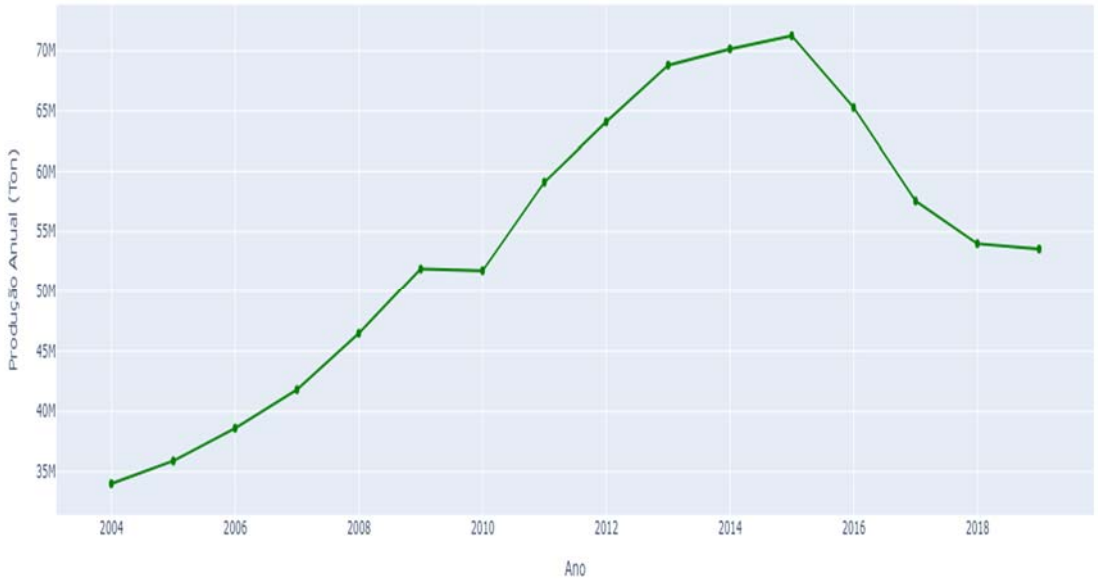
Fonte: Própria autora (2021)

5.2 Análise de Dados Exploratória da ST da produção anual de cimento no Brasil

Nesta seção são apresentados os resultados da análise de dados exploratória da ST da produção anual de cimento.

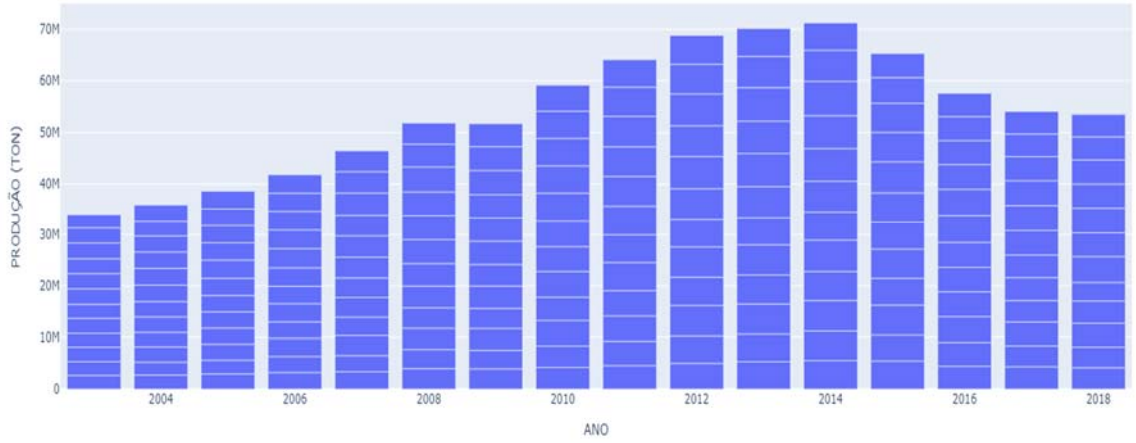
Os Gráficos 4 (linha) e 5 (barra) apresentam a evolução da produção anual de cimento, em toneladas, entre os anos de 2003 e 2018.

Gráfico 4: Evolução da Produção de Cimento por ano – 2003 a 2018



Fonte: Própria autora (2021)

Gráfico 5: Evolução da Produção de Cimento por ano – 2003 a 2018



Fonte: Própria autora (2021)

Ao analisar os Gráficos 4 e 5, verifica-se que no ano de 2003 a produção de cimento apresentou o menor valor dentre os dados históricos, totalizando 32,5 mil toneladas produzidas. Neste ano o país passava por um momento de baixo crescimento do setor da ICC, não havendo investimentos para expansão ou melhorias em infraestruturas (CUNHA, 2012).

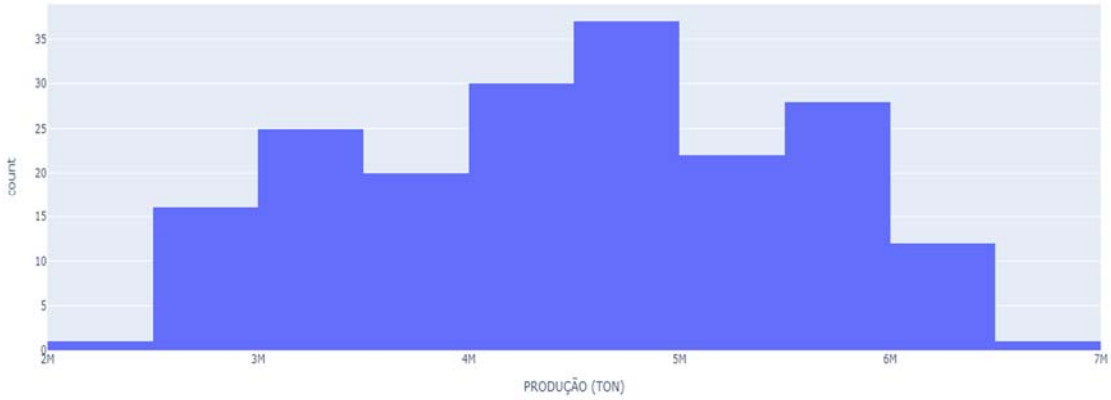
Nos anos de 2009 e 2010 houve um significativo crescimento na produção de cimento no Brasil, respectivamente, 50 mil e 60 mil toneladas. Segundo Monteiro e Veras (2017), em 2009 o governo do então presidente Luís Inácio Lula da Silva lançou o "Programa Minha Casa, Minha Vida", que tinha como objetivo expandir o número de habitações para as camadas sociais mais baixas, detentoras de renda máxima de até 10 salários mínimos. De acordo com o PT na Câmara (2020), entre os anos de 2009 e 2014 foram entregues 1,7 milhões de casas e apartamentos, beneficiando um total de 6,8 milhões de pessoas.

Nos anos que sucedem 2010 houve uma ascensão na produção de cimento devido a políticas públicas que destinaram investimentos para o setor de infraestrutura e também para o programa habitacional. Segundo Leão, Ferreira e Gomes (2016), os megaeventos esportivos desempenharam um papel importante para o aquecimento do setor da ICC. Em 2007 ocorreram os Jogos Pan-Americanos, sediados na cidade do Rio de Janeiro. Em 2014 ocorreu a Copa do Mundo da Federação Internacional de Futebol – FIFA, sediada em 12 capitais brasileiras, o que demandou grandes investimentos por parte do governo para a reforma e construção dos estádios, bem como a melhoria da infraestrutura das cidades que sediaram os jogos. Posteriormente, em 2016 ocorreram os Jogos Olímpicos na cidade do Rio de Janeiro.

Pode-se observar nos Gráficos 4 e 5, que quando há investimentos do governo no setor da ICC, como em obras de infraestrutura ou com políticas públicas para auxílio a programas habitacionais, consequentemente há maior produção do cimento. Nota-se, também, que este insumo serve como um termômetro para o setor, pois quando aquecido, aumenta sua produção.

O Gráfico 6 apresenta um histograma com a produção mensal de cimento, em toneladas, entre os anos de 2003 e 2018.

Gráfico 6: Distribuição da produção mensal de cimento entre 2003 e 2018

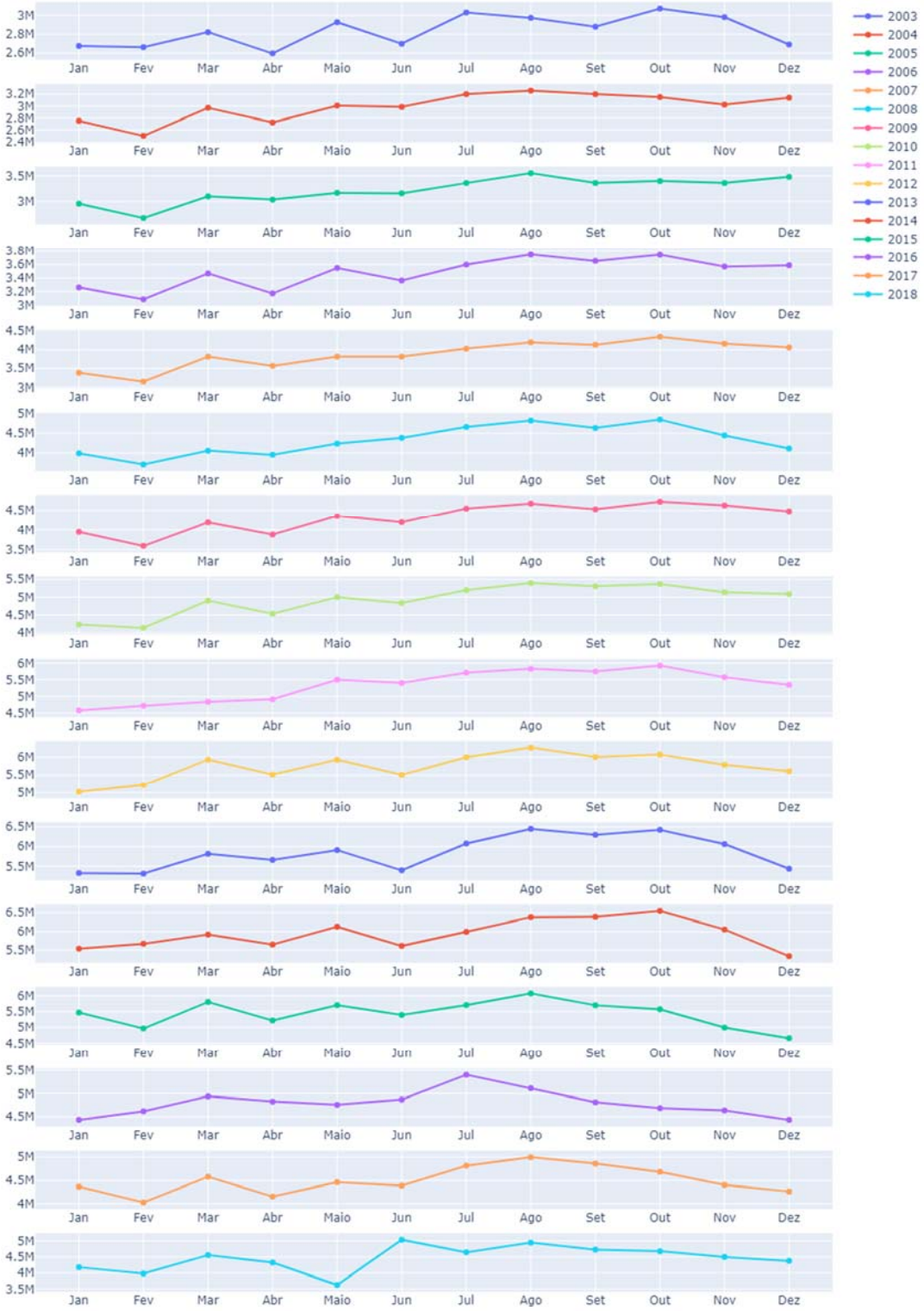


Fonte: Própria autora (2021)

Observa-se pelo Gráfico 6 que a faixa de valores para a produção de cimento que obteve a maior frequência, em 37 meses, foi a dos 4.500.000 a 5.000.000 de toneladas. Em seguida foi a faixa dos 4.000.000 e 4.500.000 de toneladas, em 30 meses. A menor produção ocorreu em fevereiro de 2004, com 2.499.959 de toneladas, e a maior em outubro de 2014, com 6.551.524 de toneladas. A produção média ficou na casa das 4.510.272 de toneladas.

O Gráfico 7 apresenta a evolução da produção mensal de cimento, em toneladas, para os anos de 2003 a 2018. Analisando-se o Gráfico 7, é possível identificar uma sazonalidade na produção de cimento. Entre os meses de maio a novembro há um aumento significativo na produção. Em pesquisa realizada em campo, chegou-se à conclusão de que esta característica observada deve-se ao período de estiagem em grande maioria das regiões do Brasil. O período de dezembro a abril é composto por estações chuvosas, o que acaba por refletir em paradas e atrasos no andamento das obras, levando a consequente redução do consumo de cimento.

Gráfico 7: Produção mensal de cimento nos anos de 2003 a 2018



Fonte: Própria autora (2021)

De acordo com Sindicato Nacional da Indústria do Cimento-SNIC (2018), o ano de 2018 fechou com uma queda de 1,2% na produção de cimento em relação ao ano anterior. Nota-se no Gráfico 7 que o mês de maio de 2018 apresentou o pior desempenho do ano, sendo sua produção de 3.500 toneladas, contrariando o ocorrido nos demais anos. Esta redução

significativa possui forte correlação com a greve dos caminhoneiros que ocorreu neste mesmo mês, pois o fato dos fretes terem sidos paralisados afetou a logística, reduzindo a oferta do produto nas prateleiras próximas ao cliente final e impactando no processo produtivo do cimento devido à falta de matéria-prima nas fábricas.

5.3 Análise preliminar da ST da produção de cimento

Para averiguar se a ST da produção de cimento apresenta comportamento tendencioso, foi realizado o Teste de Dickey-Fuller, descrito no Capítulo 4. A Tabela 4 apresenta os resultados do teste, permitindo verificar se a ST é estacionaria. Caso o *p-value* seja menor ou igual a 0,05 (95% de significância), rejeita-se a hipótese nula e a ST pode ser considerada estacionária. Caso contrário, aceita-se a hipótese nula e a ST é considerada não estacionária. Segundo Gujarati e Porter (2011), outra análise deve ser feita. De acordo com os autores, se o valor computado para a estatística de teste exceder os valores críticos, a ST é considerada não estacionária.

Tabela 4: Resultados do Teste de Dickey-Fuller para a ST da produção de cimento

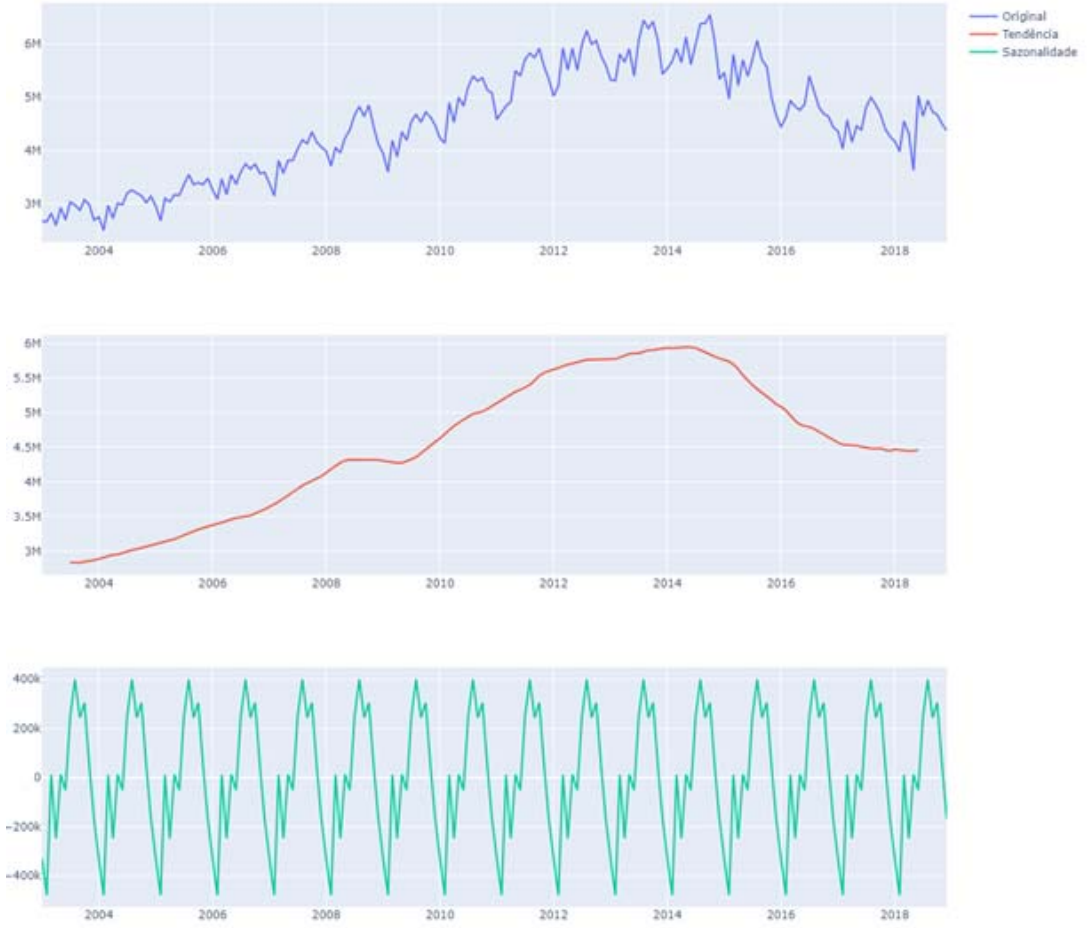
Estatística de Teste	-2.057.684
p-value	0,261867
Valor Crítico (1%)	-3.468.062
Valor Crítico (5%)	-2.878.106
Valor Crítico (10%)	-2.575.602

Fonte: Própria autora (2021)

Como pode ser visto na Tabela 4, o *p-value* encontrado foi 0,261867, a estatística de teste -2.057.684 e os valores críticos -3.468.062, -2.878.106 e -2.575.602, respectivamente. Verifica-se, portanto, que *p*-value > 0,05 e que a estatística de teste excedeu os valores críticos, ou seja, existem evidências estatísticas de que a ST da produção de cimento entre os anos de 2003 e 2018 seja uma não estacionária.

O Gráfico 8 apresenta a decomposição da ST da produção de cimento em relação à tendência e à sazonalidade. Observando o Gráfico 8, pode-se concluir que a produção de cimento no Brasil apresentou uma tendência de crescimento de 2003 até 2015. Entretanto, a partir de 2016 a tendência passou a ser de queda, perdurando até os dias atuais. Em relação à sazonalidade, observa-se que há um aumento na produção de cimento entre os meses de maio a novembro em todos os anos, corroborando com o que foi observado no Gráfico 7.

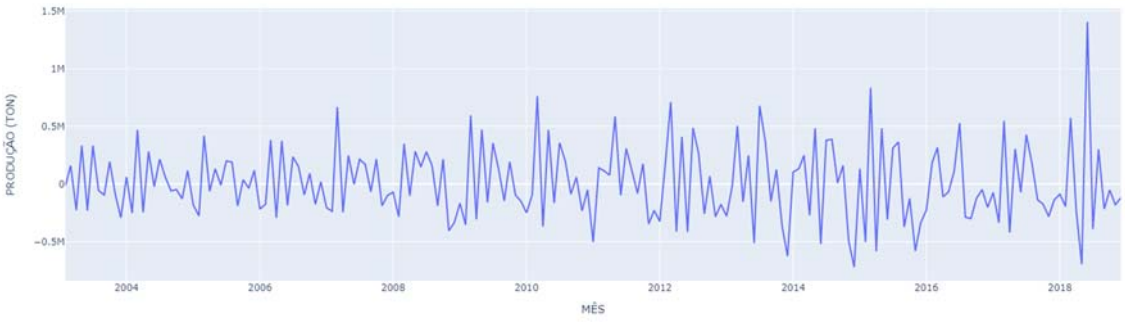
Gráfico 8: Decomposição da ST da produção de cimento



Fonte: Própria autora (2021)

O Gráfico 9 apresenta a ST da produção de cimento após a remoção das componentes tendência e sazonalidade.

Gráfico 9: ST da produção de cimento após a remoção da tendência e da sazonalidade



Fonte: Própria autora (2021)

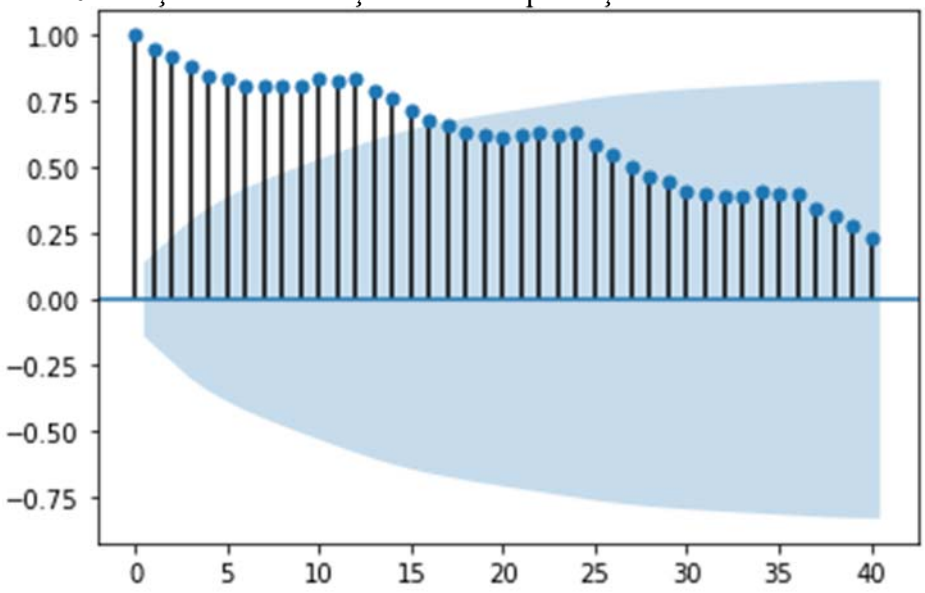
Como pode ser visto no Gráfico 9, com a remoção da tenência e da sazonalidade a ST passa a apresentar um comportamento estacionário, o que auxilia na construção dos modelos preditivos, aumentando a acurácia.

5.4 Modelos de predição para a produção de cimento no Brasil

Para a construção de um modelo estatístico é fundamental identificar os filtros que irão o compor, isto é, a presença e o número de componentes autorregressivos e de médias móveis. Para isso, emprega-se dois tipos de técnicas: a análise da função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP) (JACOBS, ZANINI e COSTA, 2014). O FAC e o FACP são capazes de verificar a estacionariedade da ST, ou seja, se a média, a variância e a estrutura de correlação não se alteram com o tempo. A FAC avalia quanto um valor tomado em um tempo *t* depende do outro em um tempo *t* - *k*, enquanto a FACP é uma extensão da FAC e mede a correlação entre dois valores eliminando-se a dependência dos termos entre eles (ROSA, CHRISTO e COSTA, 2017).

Os Gráficos 10 e 11 apresentam, respectivamente, a FAC e a FACP para a ST da produção de cimento.

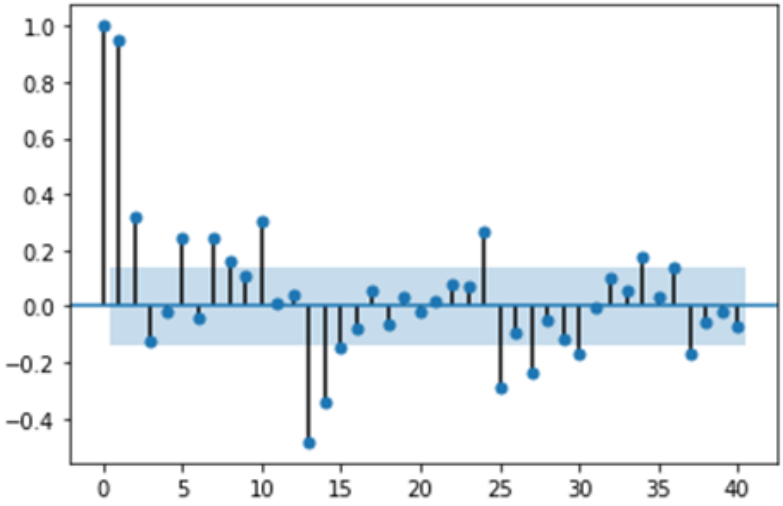
Gráfico 10: Função autocorrelação da ST da produção de toneladas de cimento



Fonte: Própria autora (2021)

Pode-se observar no Gráfico 10 que os coeficientes de autocorrelação apresentam um comportamento não estacionário dentro da ST, ou seja, eles se alteram no decorrer do tempo. Observa-se também que existe uma forte dependência nos valores da ST para até o t - 10 períodos.

Gráfico 11: Função autocorrelação parcial da ST da produção de toneladas de cimento

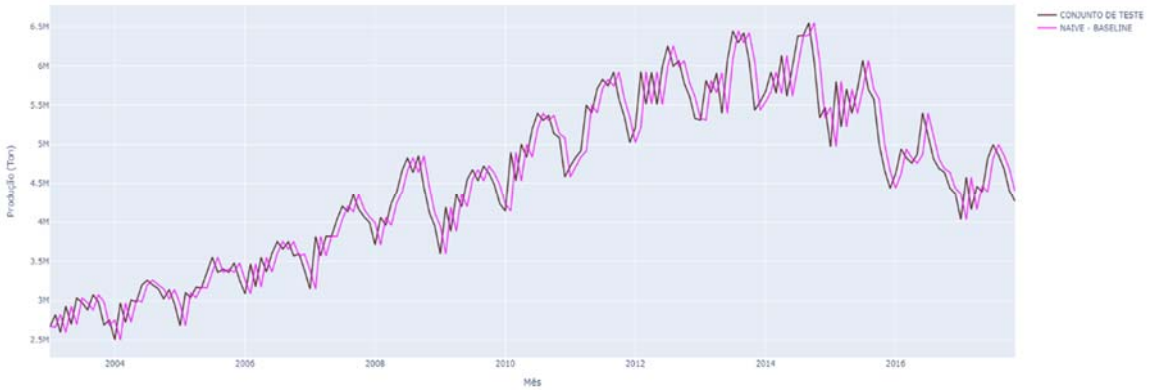


Fonte: Própria autora (2021)

No Gráfico 11 é possível observar que alguns coeficientes de autocorrelação encontram-se fora do intervalo de aceitação, o que pode ser considerado uma situação indesejável para a construção do modelo estatístico ARIMA. Isto deve-se ao fato de que alguns padrões podem não ser capturados pelo modelo.

Para efeito de comparação entre as acurácias dos modelos de predição utilizados, foi proposto neste trabalho um modelo "ingênuo" (*naive*) para ser tomado como base. Este modelo é denominado *Baseline*. No *Baseline* o valor previsto para o tempo *t* é igual ao valor real do tempo *t* - 1. O Gráfico 12 apresenta os valores reais da ST da produção de cimento (linha preta) e os valores previstos pelo *Baseline* (linha rosa).

Gráfico 12: Valores reais e preditos pelo *Baseline*



Fonte: Própria autora (2021)

Já o Gráfico 13 apresenta os valores reais da ST da produção de cimento (linha preta) e os valores previstos pelo modelo ARIMA (3,1,1) (linha rosa). Os valores 3, 1 e 1 para os parâmetros do ARIMA foram obtidos utilizando a biblioteca auto_arima do Python.

Gráfico 13: Observações reais e do modelo de predição ARIMA

Fonte: Própria autora (2021)

O Gráfico 14 apresenta os valores reais da ST da produção de cimento (linha preta) e os valores previstos pelo modelo Média Móvel considerando o período de 3 anos (linha rosa).

Gráfico 14: Observações reais e do modelo de predição Média Móvel

Fonte: Própria autora (2021)

A Tabela 5 apresenta os erros de previsão gerados pelos modelos utilizados. Para isso foi utilizada a raiz quadrada do erro médio (*Root Mean Squared Error* - RMSE).

Tabela 5: Erros gerados pelos modelos de previsão

Técnicas de Previsão	**RMSE**
Baseline	303.243,52
ARIMA	327.361,63
Média Móvel	421.182,67
PROPHET	169.565,83

Fonte: Própria autora (2021)

Com base na Tabela 5, o PROPHET foi o que apresentou o menor RMSE, comparado aos demais modelos, com valor igual a 169.565,83. Esse resultado deve-se à capacidade do PROPHET de melhor se ajustar a conjuntos de dados que apresentam comportamento

não estacionário. O ARIMA e a Média Móvel obtiveram resultados piores que o *Baseline*, o que justifica-se pelo fato de os modelos não conseguirem lidar bem com uma ST que apresenta tendência e sazonalidade.

5.5 Análise de correlação entre a produção de cimento e o PIB da ICC no Brasil

Para verificar o quanto uma variável influência no comportamento de outra, é necessário calcular o coeficiente de correlação entre elas. Isso também vale para ST's. Sendo assim, visando verificar a existência de correlação entre as ST´s da produção anual de cimento e do PIB da ICC, foram calculados os coeficientes de Pearson e de Spearman entre as duas. A Tabela 6 apresenta os valores obtidos para os coeficientes.

Tabela 6: Análise dos coeficientes de correlação

Coeficiente	Valor
Pearson	$\rho_p = 0{,}96$
Spearman	$\rho_s = 0{,}96$

Fonte: Própria autora (2021)

Como pode ser visto na Tabela 6, tanto o coeficiente de correlação de Pearson (ρ_p) quanto o de Spearman (ρ_s) apresentaram valor igual a 0,96. O ρ_p mede a intensidade e a direção de uma relação de linearidade. Quanto mais o próximo de 1 o valor de ρ_p, mais forte é a correlação positiva. Já o ρ_s indica se uma relação é crescente ou decrescente, sendo que, valores próximos de 1 indicam relação crescente. Pode-se, então, concluir que as ST's da produção de cimento e do PIB da ICC apresentam uma forte correlação positiva e crescente. O Gráfico 15 apresenta a correlação entre as ST's da produção de cimento e do PIB da ICC.

Como pode ser visto no Gráfico 15, há forte correlação positiva entre as ST's, ou seja, quando há um aumento na produção de cimento, há, também, um crescimento no PIB da ICC. No Gráfico 16 pode ser visto um mapa de calor representando a correlação entre as ST's e, mais uma vez, observa-se que há forte correlação positiva entre elas.

Gráfico 15: Correlação entre as ST’s da produção de cimento e do PIB da ICC
Fonte: Própria autora (2021)

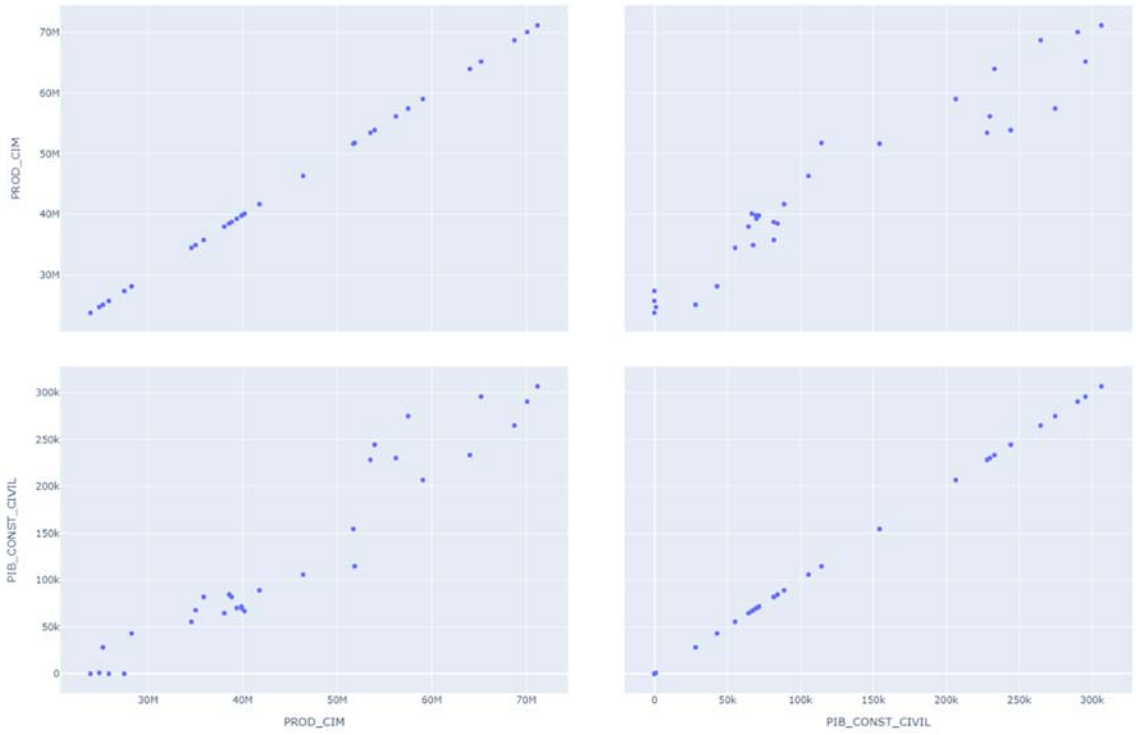
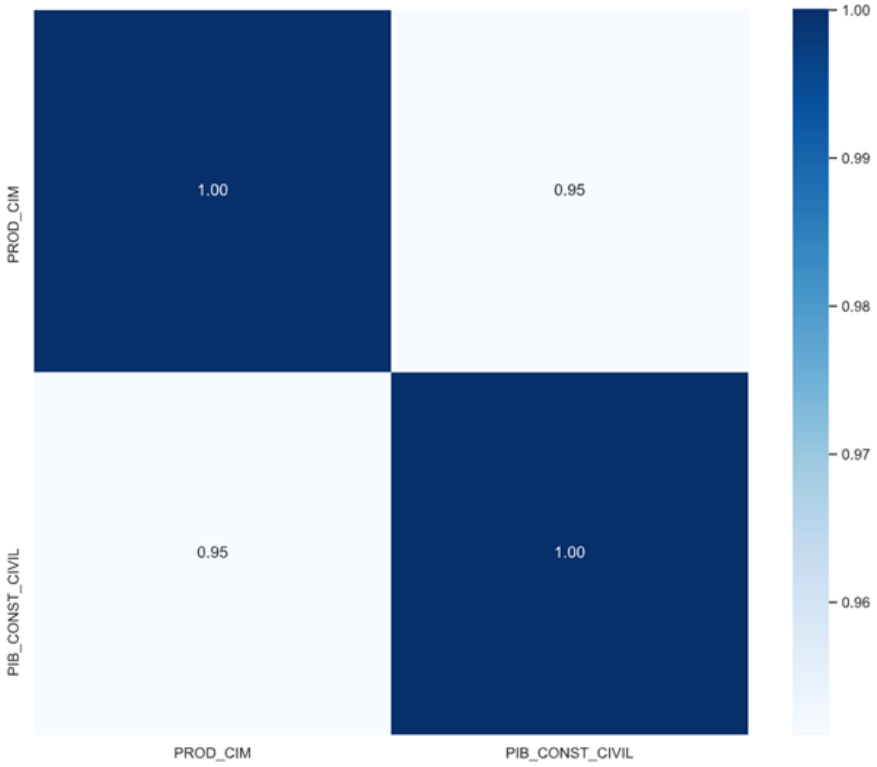


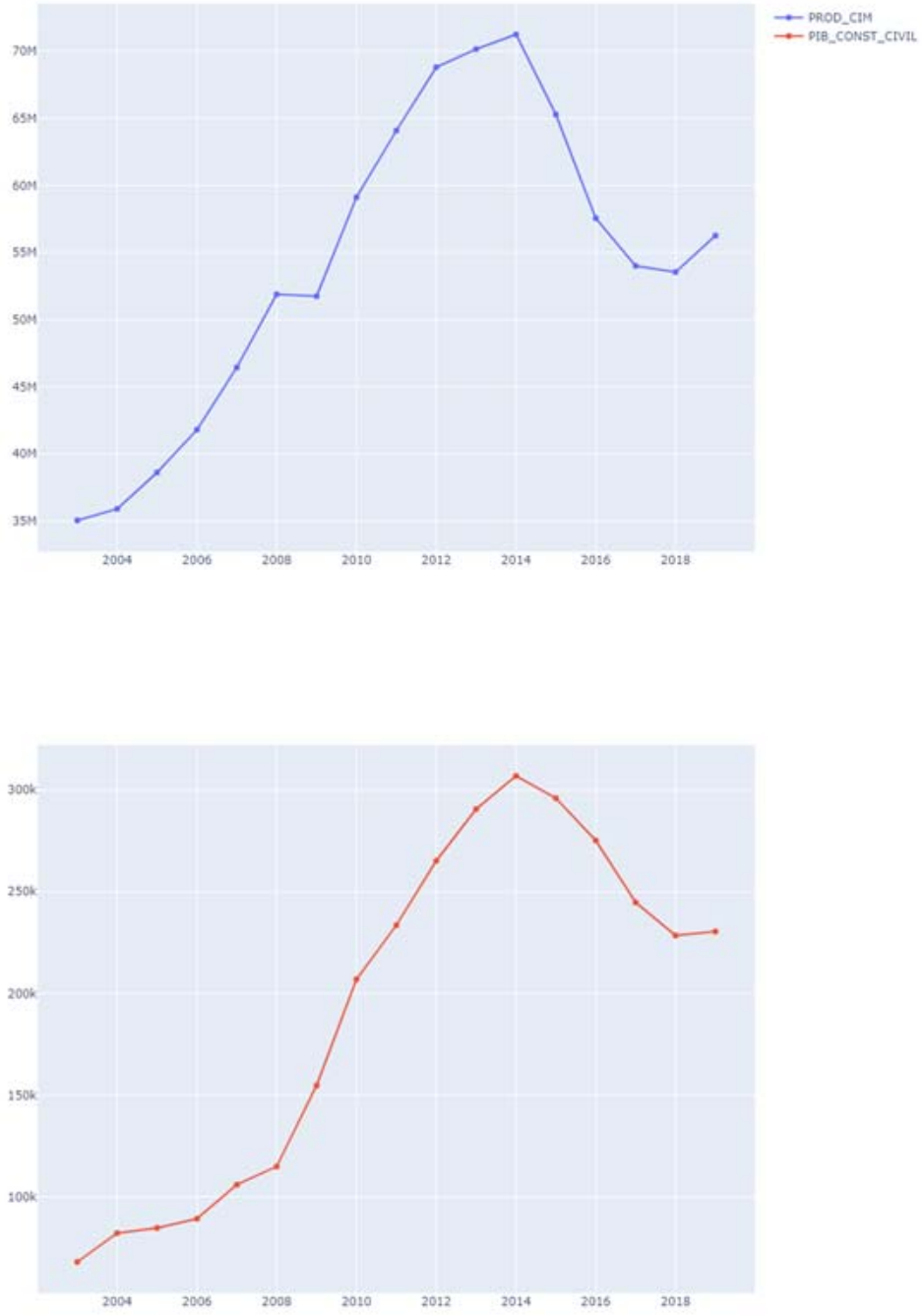
Gráfico 16:Mapa de Calor - correlação entre as ST’s



Fonte: Própria autora (2021)

O Gráfico 17 apresenta a evolução das ST’s entre o período de 2003 e 2019, a linha em azul representa a produção de cimento e linha em vermelho, representa o PIB da ICC.

Gráfico 17: Evolução das ST's da produção de cimento e do PIB da ICC



Fonte: Própria autora (2021)

Como pode-se observar no Gráfico 17, as ST's apresentam comportamentos semelhantes, crescendo e decrescendo nos mesmos períodos, demonstrando uma forte correlação positiva entre elas. Ou seja, quando há crescimento na produção de cimento há, também, crescimento no PIB da ICC e, quando há um decrescimento na produção de cimento há, também, um decrescimento no PIB da ICC.

5.6 Modelos de predição para o PIB da ICC a partir da produção de cimento

Confirmada a correlação entre as ST's, foram propostos quatro modelos de AM para tentar predizer o PIB da ICC com base na produção anual de cimento. Para a validação dos modelos, adotou-se uma acurácia mínima, ou seja, os modelos só serão validados se seu coeficiente de determinação (R^2) for maior que 80%. Quanto maior o R^2, mais explicativo é o modelo, isto é, melhor ele se ajusta aos dados. Para a construção dos modelos utilizou-se a linguagem de programação *Python.* Os valores obtidos para R^2 e o RMSE, para cada modelo, são apresentados na Tabela 7.

Tabela 7: Comparação da acurácia dos modelos de AM

Modelo	R^2	**RMSE**
Regressão Linear	84%	36.374,85
Árvore de Decisão	85%	35.107,20
Random Forest	98%	11.499,92
Gradient Boosting	88%	31.476,17

Fonte: Própria autora (2021)

Conforme os valores apresentados na Tabela 7, os quatro modelos utilizados obtiveram a acurácia mínima pretendida, isto é, $R^2 > 80\%$. O modelo *Random Forest* apresentou o melhor desempenho, obtendo um $R^2 = 98\%$. Isto significa que 98% da variabilidade do PIB da ICC é explicada pelo modelo. Em relação ao RMSE, o *Random Forest* também obteve o melhor resultado, 11.499,92. Em segundo lugar ficou o *Gradient Boosting*, com $R^2 = 88\%$ e RMSE = 31.467,17. Com base nos resultados obtidos, conclui-se que os modelos de *Ensemble Learning* melhor se adaptaram aos conjuntos de dados, gerando as previsões mais próximas dos valores reais.

Com base nos resultados obtidos neste trabalho, constatou-se a existência de uma forte correlação positiva entre a produção de cimento e o PIB da ICC no Brasil. A hipótese foi comprovada pelos valores obtidos para os coeficientes de correlação de Pearson e Spearman, 0,96 para ambos. Esta comprovação permitiu a utilização de quatro modelos de predição para relacionar as duas ST's. Dentre os modelos utilizados, o *Random Forest* foi o que mais se adaptou aos conjuntos de dados analisados, obtendo os melhores valores para R^2 e RMSE, isto é, gerou as previsões mais próximas dos valores reais. Portanto, pode-se concluir que é possível fazer uma extrapolação para períodos futuros, estimando com elevada acurácia o PIB da ICC com base na produção de cimento.

CAPÍTULO 6. CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo analisar o comportamento da produção de cimento no Brasil, identificando suas características e propondo modelos de AM para fazer previsões para períodos futuros, e verificar sua relação com o PIB da ICC. Foram, também, utilizados modelos de AM para estimar o PIB da ICC com base na produção de cimento. Para as análises, foram coletados, em duas bases, dados dos anos de 2003 a 2019. A relevância do trabalho se dá pelo fato da ICC possuir grande valor no cenário econômico do país, sendo um dos setores que mais emprega e que, quando encontra-se em ritmo acelerado, movimenta economicamente diversos outros setores. A ICC possui significativa participação no PIB brasileiro, porém necessita aumentar os investimentos em tecnologias, o que tornará seu planejamento estratégico mais assertivo.

Através da análise da ST da produção de cimento foi possível identificar uma forte tendência de crescimento na produção do insumo de 2003 a 2015. A partir de 2016 a produção passou a apresentar queda, o que ocorre até os dias atuais. Foi possível identificar, também, sazonalidade entre maio e novembro, meses em que há um aumento significativo na produção. Utilizando-se o teste de Dickey-Fuller, foi comprovado que a ST não é estacionária. Três modelos de AM foram propostos para tentar prever a produção de cimento para os anos futuros. Dentre eles, destacou-se o PROPHET, que apresentou os menores erros de previsão (RMSE). Este fato deve-se ao modelo conseguir se ajustar melhor que os demais a conjuntos de dados que apresentam comportamento não estacionário.

Na segunda etapa deste trabalho, primeiramente, analisou-se o grau de correlação entre as ST's da produção de cimento e do PIB da ICC. Para isso, foram calculados os coeficientes de correlação de Pearson e Spearman. Os valores obtidos, 0,96 para ambos, indicaram uma forte correlação positiva entre as ST's, ou seja, quando há crescimento na produção de cimento, o PIB da ICC também apresenta crescimento. Confirmada a correlação, foram utilizados quatro modelos de AM com o intuito de prever o PIB da ICC, baseando-se na produção de cimento. Todos os modelos utilizados obtiveram a acurácia mínima pretendida (R^2 > 80%). Os melhores resultados foram obtidos pelo *Random Forest*, isto é, R^2 = 98%, o que significa que 98% da variabilidade do PIB da ICC é explicada pelo modelo, e RMSE = 11.499,92. Os modelos de *Ensemble Learning* apresentaram melhor desempenho, adaptando-se mais facilmente às ST's e gerando

previsões mais próximas dos valores reais. Pôde-se, então, concluir que é possível fazer extrapolações para períodos futuros, estimando com elevada acurácia o PIB da ICC com base na produção de cimento.

As previsões realizadas pelos modelos podem auxiliar os gestores da ICC a implementar estratégias, assegurando, dessa forma, vantagem competitiva no mercado. Com base na análise do cenário futuro é possível às corporações estabelecer planejamentos mais assertivos, bem como auxiliar a administração pública na criação de políticas que promovam o aquecimento do setor. Observou-se nos estudos feitos que quando há políticas advindas do governo incentivando a construção e melhorias de infraestrutura, nota-se um aumento na produção de cimento, o que leva ao crescimento do PIB da ICC e, consequentemente, da economia.

Para trabalhos futuros sugere-se considerar outros tipos de insumos utilizados na ICC, tais como, aço e materiais cerâmicos. Isso tornará o modelo mais completo e próximo da realidade. Para isso, será necessária a utilização de outros modelos de predição, como, por exemplo, a Regressão Linear Múltipla.

REFERENCIAL TEÓRICO

ASSOCIAÇÃO BRASILEIRA DE CIMENTO PORTLAND. *Guia básico de utilização do cimento portland*, São Paulo, 7° ed. 2002, p. 28.

ANTUNES, J. L. F.; CARDOSO, M. R. A. *Uso da análise de séries temporais em estudos epidemiológicos*. Epidemiologia e Serviços de Saúde, v. 24, p. 565–576, 2015.3.

ARAUJO, G. J. F. D. *O coprocessamento na indústria de cimento: definição, oportunidades e vantagem competitiva*. Revista Nacional de Gerenciamento de Cidades, v. 8, n. 57, p. 52–61, Março 2020.

ASSAD, D. *Previsão da evolução dos casos de COVID-19 no município do Rio de Janeiro para o período de 14/abril a 04/maio.* LEGOS, 2020. Disponível em: http://www.legos.uerj.br/nota-tecnica/previsao-da-evolucao-dos-casos-de-sindrome-respiratoria-aguda-grave-srag-e-da-ocupacao-hospitalar-na-rede-publica-sus-no-municipio-do-rio-de-janeiro-para-o-periodo-de-14-abril-a-04-maio/. Acesso em: 20 Out 2020.

BEZERRA, Aguinaldo, SILVA, Ivanovitch; GUEDES, Luiz Affonso; SILVA, Diego; LEITÃO, Gustavo; SAITO, Kaku. *Extracting Value from Industrial Alarms and Events:*A Data-Driven Approach Based on Exploratory Data. Sensors, p. 1-21, 20 Junho 2019.

BIECEK, P. *Dalex: Explainers for complex predictive models in R*. Journal of Machine Learning Research, v. 19, p. 1-5, Novembro 2018.

BILAL, Muhammad; OYEDELE, Lukumon O.; QADIR, Junaid; MUNIR, Kamran; AJAYI, Saheed O.; AKINADE, Olugbenga O.; OWOLABI, Hakeem A.; ALAKA, Hafiz A.; PASHA, Maruf. *Big Data in the construction industry: A review of present status, opportunities, and future trends*. Advanced Engineering Informatics, v. 30, n. 3, p. 500-521, Agosto 2016.

BOX, George E. P; JENKINS, Gwilym M.; REINSEL, Gregory C.; LJUNG, Greta M. *Time Series Analysis:* Forecasting and Control. 5. ed. Nova Jersey: John Wiley & Sons, 2015.

BOURGARD, Bruno; GOMES, Carlos F. S. *As variáveis econômicas no Brasil e o PIB: uma análise em períodos de crises financeiras através da correlação de Pearson.* Almanaque multidiscilpinar de pesquisa, v. 1, n. 2, p. 76-98, 2017

CAI, L.; ZHU, Y. *The challenges of data quality and data quality assessment in the big data era.* Data Science Journal, 2015.

CAMILO, C. O.; SILVA, J. C. D. *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. Universidade Federal de Goiás. [S.l.], p. 29. 2009.

CAO, L. *Data science and analytics: a new era*. Int J Data Sci Anal, 2016. Disponível em: https://link.springer.com/article/10.1007/s41060-016-0006-1.

_____. Data science: A comprehensive overview. ACM Computing Surveys, v. 50, n. 3, 2017.

_____. Data science: challenges and directions, 60, n. 8, 2017. 59–68. Disponivel em: https://cacm.acm.org/magazines/2017/8/219605-data-science/fulltext.. Acesso em: 19 Set. 2020.

CASTRO, L. N. D.; FERRARI, D. G. *Introdução à mineração de dados*. 1. ed. São Paulo: Saraiva, 2016.

CBIC - CÂMARA BRASILEIRA DA INDÚSTRIA DA CONSTRUÇÃO . *Banco de Dados CBIC*, 2020. Disponivel em: <http://www.cbicdados.com.br/glossario/v/>. Acesso em: 21 Março 2020.

CBIC- CÂMARA BRASILEIRA DA INDÚSTRIA DA CONSTRUÇÃO. *Banco de Dados CBIC,* 2020. Disponivel em: http://www.cbicdados.com.br/menu/pib-e-investimento/pib-brasil-e-construcao-civil. Acesso em: 07 Set. 2020

CIMENTO.ORG. *Cimento no Mundo*. Cimento.org, 2014. Disponível em: https://cimento.org/cimento-no-mundo-2013/. Acesso em: 20 Out. 2020.

CHATFIELD,C. *The analysis of time series:* An Introduction 6. ed. Chapman & Hall/CRC, 2013

CNI - CONFEDERAÇÃO NACIONAL DA INDÚSTRIA *A indústria brasileira de cimento base para a construção do desenvolvimento*. Confederação Nacional da Indústria. Associação Brasileira de Cimento Portland, p. 60, 2017.

_____. *Fato Econômico:* Razões e condições da crise à recuperação do setor de construção. [S.l.]: [s.n.]. 2019. p. 1–3.

COWPERTWAIT, P. S. P.; METCALFE, A. V. *Introductory Times Series with R.* 1. ed. Nova Iorque: Springer, 2009.

CROARKIN, C.; TOBIAS, P. *NIST/SEMATECH e-Handbook of Statistical Methods*. . 2012. Disponível em: http://www.itl.nist.gov/div898/handbook . Acesso em: 28 de set de 2020.

CUNHA, G. D. C. *Importância do setor de Construção Civil para o desenvolvimento da Economia Brasileira e as alternativas complementares para o Funding do Crédito Imobiliário no Brasil.* 79 f. Trabalho de Conclusão de Curso (Instituto de Economia) Universidade Federal do Rio de Janeiro. Rio de Janeiro. 2012.

DIETTERICH, T. G. *Ensemble Methods in Machine Learning***.** In: Multiple Classifier Systems. Berlim: Springer, v. 1857, 2000.

DOANE, David P.; SEWARD, Lori E.. *Estatística aplicada à administração e economia*. 4. ed. Porto Alegre: AMGH, 2014. p. 847.

DUA, S.; DU, X. *Data Mining and Machine Learning in Cybersecurity***.** 3. ed. Nova Iorque: Imprensa CRC, 2016.

DUNHAM, M. H. *Data Mining:* Introductory and Advanced Topics. 1. ed. Pearson, 2003.

EHLERS, R. S. *Análise de séries temporais*. Universidade Federal do Paraná. Curitiba, 2009. p. 90.

ESPÍNDOLA, André M. S. de; ROTH, Leonardo; CAMARGO, Maria Emilia; FACHINELLI, Ana Cristina. Big Data e Inteligência Estratégica: Um Estudo de Caso Sobre a Mineração de Dados como Alternativa. Revista Espacios, v. 37, p. 16-35, 2016.

FARAH, M. F. S. *Processo de Trabalho na Construção Habitacional:* Tradicional e Mudança. 1. ed. São Paulo: Annablume, 1996.

FÁVERO, L. P.; BELFIORE, P. *Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®*. 1. ed. Rio de Janeiro: Elsevier Brasil, 2017.

FERNANDES, F. T.; CHIAVEGATTO FILHO, A. D. P. *Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho*. Revista Brasileira de Saúde Ocupacional, v. 44, p. 1–12, 2019.

FU, T. C. *A review on time series data mining*. Engineering Applications of Artificial Intelligence, v. 24, p. 164–181, 2011.

FRACARO, Nelize. *Estacionariedade das séries temporais do modelo matemático ARIMAX de propulsores eletromecânicos.* 89 f. Dissertação (Mestrado em Modelagem Matemática) - DCEEng - Departamento de Ciências Exatas e Engenharias. Universidade Regional do Noroeste do Estado do Rio Grande do Sul. Ijuí, 2018.

FREITAS, C. A. de; SÁFADI, Thelma. *Volatilidade dos Retornos de Commodities Agropecuárias Brasileiras: um teste utilizando o modelo APARCH.* Revista Economia e Sociologia Rural. v.53, n.2, Brasília, Abr -Jun 2015

GASPAR, I. D. A.; GONÇALVES, M. R.; MATIAS, I. D. O. *Time Series Prediction: Case study using artificial neural network techiniques for forecasting National Petroleum Production*. Interdisciplinary Scientific Journal., v. 5, n. 1, p. 138-152, Jan-Mar 2018.

GUJARATI, D. N.; PORTER, D. C. *Econometria*. 5. ed. Porto Alegre: AMGH Editora, 2011.

GUPTA, C. B.; GUTTMAN, I. *Estatística e probabilidade com aplicações para engenheiros e cientistas.* 1. ed. Rio de Janeiro: LTC , 2018.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining:* Concepts and Techniques. 3. ed. Waltham: Elsevier, 2011.

ISHIZAKI, Mauricio Yoiti. Reconhecimento automático de palavras. 2018. 43 f. Trabalho de Conclusão de curso (Graduação em Engenharia de Controle e Automação) - Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2018.

JACOBS, W.; ZANINI, R. R.; COSTA, M. *Estudo comparativo de séries temporais para a previsão de vendas de um produto*. Revista Iberoamericana de Engenharia Industrial, Florianopolis, v. 6, n. 12, p. 112-133, 2014.

KANTARDZIC, M. *Data Mining:* Concepts, Models, Methods, and Algorithms. 2. ed. Nova Jersey: John Wiley & Sons, 2011.

KIRCHGÄSSNER, Gebhard, WOLTERS , Jürgen. *Introduction to Modern Time Series Analysis.* 1. ed. Nova Iorque: Springer, 2007

KRISMA, A.; AZHARI, M.; WIDAGDO, P. P. *Perbandingan Metode Double Exponential Smoothing Dan Triple Exponential Smoothing Dalam Parameter Tingkat Error Mean Absolute Percentage Error (MAPE) dan Means Absolute Deviation (MAD)*. Prosiding Seminar Nasional Ilmu Komputer dan Teknologi Informasi, v. 4, p. 81-87, Setembro 2019.

KUMAR, T. S. *Introduction to Data Mining*. 5. ed. Londres: Pearson Education Limited, 2014.

KWIATKOWSKI, D.; PHILLIPS, P.C.B.; SCHMIDT P.; SHIN, Y. - *Testing the null hypothesis of stationarity against the alternative of a unit root:* How sure are we that economic time series have a unit root? Journal of Econometrics, 54(1) p. 159-178, 1992.

LAROSE, D. T.; LAROSE, C. D. *Discovering Knowledge in Data:* An Introduction to Data Mining. 2. ed. Nova Jersey: John Wiley & Sons, 2014.

LEÃO, André Luiz M.de S.; FERREIRA, Bruno Rafael T.; GOMES, Victor Pessôa de M.. *Um "elefante branco" nas dunas de Natal? Uma análise pós-desenvolvimentista dos discursos acerca da construção da Arena das Dunas.* Revista de Administração Pública, Rio de Janeiro, v.50, n.4, jul-ago. 2016.

MAYRINK, V. T. D. M. *Avaliação do algoritmo Gradient Boosting em aplicações de previsão de carga elétrica a curto prazo.* 91f. Dissertação (Mestrado em Modelagem Computacional) – ICE/Engenharia.Universidade Federeal de Juiz de Fora. Juiz de Fora. 2016.

MCCUE, C. *Data Mining and Predictive Analysis:* Intelligence Gathering and Crime Analysis. 1. ed. Grã-Bretanha: Elsevier, 2007.

MONTEIRO, Adriana Roseno; VERAS, Antonio Tolrino de Rezende. *A Questão habitacional no Brasil.* Mercator, Fortaleza, v.16, jul. 2017. Disponível em: https://www.scielo.br/pdf/mercator/v16/1984-2201-mercator-16-e16015.pdf. Acesso em: 20 Mar. 2021.

MONTGOMERY, D. C.; JENNINGS, C. L.; KULAHCI, Murat. *Introduction to Time Series Analysis and Forecasting.* 1. ed. New Jersey: John Wiley & Sons, 2008.

MONTGOMERY, D. C.; PECK, E. A.; GEOFFREY, G. *Introduction to linear regression analysis*. 5. ed. New Jersey: John Wiley & Sons, 2012.

MONTGOMERY, D. C.; RUNGER, G. C. *Estatística aplicada e probabilidade para engenheiros.* 6. ed. Rio de Janeiro: Livros Técnicos e Científicos, 2018.

MORETTIN, P. A. C.; TOLOI, C. M. *Análise de séries temporais:* modelos lineares univariados. São Paulo: Blucher, v. 3, 2018.

MOYER, B.; DUNN, A. *Medindo o Produto Interno Bruto (PIB):* The Ultimate Data Science Project. Harvard Data Science Review [Internet]., 31 Janeiro 2020. Disponivel em: https://hdsr.mitpress.mit.edu/pub/5pkkan15. Acesso em: 20 Set. 2020.

MUSARDO, I. Modelo de consultas e relatórios AD-HOC para Sistemas de BI, 2008. Disponivel em: <https://musardos.com/modelo-de-consultas-e-relatorios-ad-hoc-para-sistemas-de-bi/>. Acesso em: 04 Outubro 2020.

NABAVI-PELESARAEI, Ashkan; RAFIEE, Shahin; HOSSEINI-FASHAMI, Fatemeh; CHAU, Kwok-wing. *Predictive Modelling for Energy Management and Power Systems Engineering.* In: Artificial neural networks and adaptive neuro-fuzzy inference system in energy modeling of agricultural products. Editora: Elsevier, 2021. Cap.11. p. 299-334.

NUNES, H. *Análise do sistema construtivo de edifícios de múltiplos pavimentos no Brasil em lajes lisas com cordoalhas engraxadas.* 157 f. Tese (Doutorado em Estruturas e Construção Civil) - Universidade Federal de São Carlos. São Carlos, 2019.

OLIVEIRA, B. *Características das Séries Temporais*. Oper Data, 2019. Disponível em: https://operdata.com.br/blog/caracteristicas-das-series-temporais/. Acesso em: 20 Out. 2020.

OLIVEIRA, J. D. S. *Desenvolvimento e treinamento de Redes Neurais Artificiais para processamento de dados de radiação solar.* Universidade Federal de Santa Catarina. Araraguá, p. 37. 2017.

OLSON, D. L.; DELEN, D. *Advanced data mining techniques.* 1. ed. Nova Iorque: Springer Science & Business Media, 2008.

ORIGUELA, Letícia Aparecida. *Estudo da influência de eventos sobre a estrutura do mercado brasileiro de ações a partir de redes ponderadas por correlações de Pearson, Spearman e Kendall.* 85 f. Dissertação (Mestrado em Administração de Organizações) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, Ribeirão Preto. 2018

PARMEZAN, A. R. S. *Predição de série temporais por similaridade.* 219 f. Dissertação (Mestrado em Ciências – Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICM/USP) São Carlos, 2016.

PT NA CÂMARA. *Nos governos petistas, Minha Casa, Minha Vida garantiu moradia digna a 6,8 milhões de brasileiros,* 2020. Disponível em: https://ptnacamara.org.br/portal/2020/06/25/nos-governos-petistas-minha-casa-minha-vida-garantiu-moradia-digna-a-68-milhoes-de-brasileiros/. Acesso em: 13 Mar. 2021.

PYTHON. Sobre *Python*. Disponível em: https://www.python.org/. Acesso em: 03 nov.2020.

RASCHKA, S.; MIRJALILI, V. *Python Machine Learning*. 2. ed. Packt, 2017.

REFFAT, R. M.; GERO, J. S.; PENG, W. *Using data mining on building maintenance during the building life cycle.* International Conference of Architectural Science Association ANZAScA. Launceston: 38. 2017. p. 91–97.

REIS, M. M. *Análise de séries temporais*. p.55, Disponível em: https://www.inf.ufsc.br/~marcelo.menezes.reis/Cap4.pdf . Acesso em: 07 Set. 2020.

REZENDE, D. A.; ABREU, A. F. D. *Tecnologia da informação aplicada a sistemas de informação empresariais:* o papel estratégico da informação e dos sistemas de informação nas empresas. 9. ed. São Paulo: Atlas, 2013.

RIBEIRO, Francielle; TELEGINSKI, Jaqueline; SOUZA, Jodson; GUGELMIN, Renata. *A evolução do produto interno bruto brasileiro entre 1993 e 2009.* Vitrine da Conjuntura, Curitiba, v. 3, Julho 2010.

ROSA, C. D. O. C. S.; CHRISTO, E. D. S.; COSTA, K. A. *Análise de viabilidade de modelos SARIMA para previsão de vazões do rio Paraíba do Sul.* Coletânea nacional sobre engenharia de produção 5: pesquisa, Curitiba, p. 325, 2017.

RUSSOM, P. *Introduction to Big Data Analytics*. TDWI best practices report, v. 19, n. 4, p. 1-34, 2011.

SAMMUT, C.; WEBB, G. I. *Encyclopedia of Machine.* 1. ed. Boston: Springer, 2011.

SANTOVENA, A. Z. *Big Data : Evolution , Components , Challenges and Opportunities*. Escola de Gestão Mit Sloan. [S.l.], p. 126. 2013.

SILVA, Ermes Medeiros da; SILVA, Elio Medeiros da; GONÇALVES, Valter; MUROLO, Afrânio Carlos. *Estatística*. 5.ed. São Paulo: Atlas, 2018.

SILVEIRA, A., DE MATTOS, V., KONRATH, A. *Avaliação da estacionariedade e teste de cointegração em séries temporais:* O caso da demanda de energia elétrica no Brasil. Revista de Tecnologias, Ourinho, 9, fev. 2017. Disponível em: https://www.fatecourinhos.edu.br/retec/index.php/retec/article/view/267/163. Acesso em: 03 Nov. 2020.

SLACK, N.; BRANDON-JONES, A.; JOHNSTON, R. *Administração da produção*. 8. ed. Rio de Janeiro.: Atlas, 2018.

SNIC- SINDICATO NACIONAL DA INDÚSTRIA DE CIMENTO. *ROADMAP tecnológico do cimento:* potencial de redução das emissões de carbono da indústria do cimento brasileira até 2050. Rio de Janeiro, 2019. 64 p.

_____. *Vendas de cimento encerram 2018 em queda de 1,2%.* 2018. Disponível em: http://snic.org.br/assets/pdf/resultados-preliminares/1547058910.pdf. Acesso em: 23 mai. 2021.

SOUSA, Áurea. *Coeficiente de Correlação de Pearson e Coeficiente de correlação de Spearman. O que medem e em que situações devem ser utilizados?*. Correio Açores: Matemática, 21, março, 2019. p. 19. Disponível em: https://repositorio.uac.pt/bitstream/10400.3/5365/1/Sousa_CA_21%20Mar%c3%a7o%202019.pdf. Acesso em: 03 nov. 2020.

SOUZA, Bruno Almeida; OLIVEIRA, Camilla Araújo Coelho; SANTANA, Júlio Carlos Oliveira De; NETO, Luis Antônio da C. V; SANTOS, Débora de Gois. *Análise dos indicadores pib nacional e pib da indústria da construção civil*. Revista de Desenvolvimento Econômico, Salvador, v. 17, n. 31, p. 140-150, jan- jun 2015.

TAN, Pang-Ning; STEINBACH, Michael; KARPATNE, Anuj; KUMAR, Vipin. *Introduction to Data Mining*. 2. ed. [S.l.]: Pearson, 2019.

TAYLOR, S. J.; LETHAM, B. *Forecasting at Scale*. PeerJ Preprints, Setembo 2017.

TEIXEIRA, L. P.; CARVALHO, F. M. A. *A construção civil como instrumento do desenvolvimento da economia brasileira.* Revista Paranaense de Desenvolvimento, Curitiba, v. 109, p. 9-26, Julho - Dezembro 2005.

TOMAR, Geetam S; CHAUDHARI, Narendra S; BHADORIA, Robin Singh; DEKA, Ganesh Chandra. *The Human element of big data:* Issues, analytics, and performance. 1. ed. Nova Iorque: Impresa CRC, 2016.

TUKEY, John. W. Exploratory Data Analysis Biometrics, 1977.

TURING, A. M. Computing Machinery and Intelligence. Mind *v. 49: p.* 433-460, 1950.

UNIVERSIDADE FEDERAL DO RECÔNCAVO BAIANO. *Você sabe o que consiste um petabyte?*.Coordenadoria de Tecnologia da Informação. Disponíevel em: https://www.ufrb.edu.br/cotec/. Acesso em: 21 Out. 2020.

VICARIO, G.; COLEMAN, S. *Uma revisão da ciência de dados nos negócios e na indústria e uma visão futura*. Appl Stochastic Models Bus Ind, v. 36, p. 6 - 18, 2020.

ZHANG, Y.; HAGHANI, A. *A gradient boosting method to improve travel time prediction.* Transportation Research Part C, v. 58, p. 308-324, 2015.