Trabalho de Conclusão de Curso

Previsão de Empregos na Construção Civil em Alagoas Utilizando Séries Temporais

Lucas Ribeiro Raggi
lrr@ic.ufal.br

Orientadores:

Prof. Dr. Thiago Damasceno Cordeiro
Prof. Dr. Baldoino Fonseca dos Santos Neto

Maceió, Julho de 2023

Lucas Ribeiro Raggi

Previsão de Empregos na Construção Civil em Alagoas Utilizando Séries Temporais

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientadores:

Prof. Dr. Thiago Damasceno Cordeiro

Prof. Dr. Baldoino Fonseca dos Santos Neto

Maceió, Julho de 2023

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

R142p Raggi, Lucas Ribeiro.
Previsão de empregos na construção civil em Alagoas utilizando séries temporais / Lucas Ribeiro Raggi. – 2023.
45 f. : il. color.

Orientador: Thiago Damasceno Cordeiro.
Coorientador: Baldoino Fonseca dos Santos Neto.
Monografia (Trabalho de Conclusão de Curso em Ciência da Computação) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2023.

Bibliografia: f. 43-45.

1. Indústria de construção civil. 2. Emprego – Construção civil – Alagoas. 3. Previsão de séries temporais. I. Título.

CDU: 004.8 : 331 (813.5)

Agradecimentos

Agradeço profundamente a todos aqueles que contribuíram para a realização deste trabalho.

À minha mãe, Andreia Raggi, por todo o amor e apoio incondicional em todos os momentos da minha vida. Ao meu falecido pai, Alexandre Raggi, cujo legado continua inspirando minha jornada.

À minha namorada, Milenna, pelo constante apoio e auxílio durante todo o processo de desenvolvimento deste trabalho.

Aos meus orientadores, Thiago Cordeiro e Baldoino Fonseca, pela inestimável contribuição para a elaboração do projeto e formação acadêmica.

Agradeço aos professores Aydano Machado, Evandro Barros e Rafael Amorim, que abriram as portas dos seus laboratórios, proporcionando valiosas oportunidades de contribuição e aprendizado.

Aos meus amigos da universidade, Daniel Vassalo, França Mac Dowell, Gabriel Barbosa, Lucas Amorim, Nelson Gomes Neto, Wagner Fontes, pelo companheirismo, inspiração e pela amizade que tornou a jornada universitária mais divertida e enriquecedora.

Aos meus amigos mais próximos, Daniel Melo, Thiago Tenorio e Victor Meneghini, pela amizade e suporte constantes.

Por fim, agradeço aos meus colegas de trabalho, cuja colaboração diária enriquece minha experiência profissional e pessoal.

"Planejar é trazer o futuro para o presente para que você possa fazer algo sobre isso agora."

– Alan Lakein

Resumo

A indústria da construção civil tem um papel fundamental na economia brasileira, contribuindo significativamente para o desenvolvimento socioeconômico e a empregabilidade. No estado de Alagoas, onde o setor é uma das principais forças econômicas, a empregabilidade vem sendo afetada por uma serie de fatores, incluindo mudanças na política econômica, crises financeiras e a pandemia de COVID-19.

Este estudo propõe a previsão do saldo empregos na construção civil em Alagoas utilizando técnicas de previsão de séries temporais, a fim de ajudar instituições de formação técnica e profissional a alinharem seus programas de capacitação às demandas do setor e os formuladores de políticas no desenvolvimento de estratégias de emprego e educação mais eficazes.

Os dados de emprego no setor foram obtidos a partir das bases RAIS e CAGED, ferramentas governamentais que coletam informações sobre admissões, demissões e empregos formais no Brasil. Diversos modelos de previsão, incluindo Suavização Exponencial Holt Winters, ARIMA, SARIMA, TBATS e Prophet, foram aplicados e comparados, com otimização de hiperparâmetros e validação cruzada realizadas para melhorar a precisão das previsões e avaliar o desempenho dos modelos.

Os resultados destacam a Suavização Exponencial Holt Winters como a abordagem mais eficaz, alcançando um erro RMSE de 825.26, uma redução de 43.5% em relação ao método Prophet, amplamente utilizado na indústria. sugerindo que essa técnica clássica continua sendo uma ferramenta robusta e eficaz para prever a demanda de empregos no setor da construção civil. O estudo reconhece as limitações inerentes à abordagem univariada, que não leva em consideração variáveis exógenas, e sugere que pesquisas futuras explorem técnicas de modelagem que considerem tais variáveis.

Em suma, esta pesquisa fornece um método confiável e eficaz para prever a demanda de empregos na construção civil, oferecendo orientações relevantes para empresas de construção, instituições de formação técnica e profissional, e formuladores de políticas.

Palavras-chave: Previsão de Séries Temporais, Demanda de Empregos, Construção Civil, Alagoas.

Abstract

The construction industry plays a crucial role in the Brazilian economy, significantly contributing to socioeconomic development and employability. In the state of Alagoas, where the sector is one of the main economic forces, employability has been affected by a series of factors, including changes in economic policy, financial crises, and the COVID-19 pandemic.

This study proposes forecasting the balance of jobs in the construction industry in Alagoas using time series forecasting techniques, aiming to assist technical and professional training institutions in aligning their training programs with the sector's demands and policymakers in developing more effective employment and education strategies.

Employment data in the sector were obtained from the RAIS and CAGED databases, government tools that collect information about admissions, layoffs, and formal jobs in Brazil. Various forecasting models, including Holt Winters Exponential Smoothing, ARIMA, SARIMA, TBATS, and Prophet, were applied and compared, with hyperparameter optimization and cross-validation performed to improve forecast accuracy and evaluate model performance.

The results highlight Holt Winters Exponential Smoothing as the most effective approach, achieving an RMSE error of 825.26, a reduction of 43.5% compared to the widely used Prophet method, suggesting that this classic technique remains a robust and effective tool for predicting job demand in the construction sector. The study acknowledges the inherent limitations of the univariate approach, which does not take into account exogenous variables, and suggests that future research explore modeling techniques that consider such variables.

In summary, this research provides a reliable and effective method for predicting job demand in the construction industry, offering relevant guidance for construction companies, technical and professional training institutions, and policymakers.

Key-words: Time Series Forecasting, Job Demand, Construction Industry, Alagoas.

Lista de Figuras

Lista de Tabelas

Conteúdo

1

Introdução

Com um papel fundamental para o desenvolvimento socioeconômico do Brasil, a construção civil tem enfrentado vários desafios, incluindo questões de empregabilidade, especialmente no estado de Alagoas. Projeções recentes apontam para um crescimento do de 2,5% para o PIB de construção civil em 2023, influenciado pela recuperação econômica do Brasil, desaceleração dos custos dos materiais e o ânimo pós-pandêmico (2). No entanto, é também um setor caracterizado por sua volatilidade, com flutuações na demanda de empregos, muitas vezes em resposta a mudanças nas condições econômicas e políticas (3).

No estado de Alagoas, a construção civil é uma das principais indústrias, contribuindo significativamente para a economia local (4). No entanto, a indústria enfrentou desafios nos últimos anos, com variações na demanda de empregos que podem ser atribuídas a uma variedade de fatores, incluindo mudanças na política econômica, crises financeiras e o impacto da pandemia de COVID-19 (5).

Nesse contexto, a previsão da demanda de empregos na construção civil torna-se uma ferramenta valiosa. Previsões precisas podem ajudar as empresas a planejar melhor suas necessidades de contratação, e os formuladores de políticas a desenvolver estratégias de emprego e educação mais eficazes.

Este trabalho visa prever a demanda de empregos na construção civil em Alagoas, utilizando técnicas de previsão de séries temporais, que são séries caracterizadas como uma forma útil de modelar problemas de previsão, pois permitem que os padrões subjacentes nos dados - como tendências e sazonalidade - sejam capturados e utilizados para fazer previsões futuras (6).

Para realizar esta análise, foram coletados dados da bases RAIS e CAGED. O CAGED é um registro administrativo mantido pelo Ministério do Trabalho do Brasil, que coleta informações sobre admissões e demissões de trabalhadores sob o regime da Consolidação das Leis do Trabalho (CLT)(4). A RAIS é um sistema de coleta de dados administrativos, também mantido pelo Ministério do Trabalho do Brasil que coleta dados anuais sobre todos os estabelecimentos e empregos formais no Brasil(7). Com bases nestes dados, foram aplicados vários modelos

de previsão, incluindo Suavização Exponencial Holt Winters, ARIMA, SARIMA, TBATS e Prophet. A otimização de hiperparâmetros foi realizada para melhorar a precisão das previsões, e a validação cruzada foi utilizada para avaliar o desempenho dos modelos.

1.1 Justificativa

Dada a relevância econômica da indústria da construção civil, especialmente no contexto de Alagoas, e os desafios impostos pelas condições atuais, a previsão precisa da demanda de empregos nesse setor torna-se uma necessidade crucial. Do ponto de vista prático, tais previsões poderiam beneficiar uma série de entidades e instituições. Empresas do setor de construção civil poderiam aprimorar sua gestão de recursos humanos, planejar com mais eficiência suas necessidades de contratação e ajustar suas estratégias de negócios de acordo com as previsões futuras (8). Instituições de formação técnica e profissional, como SENAI e SESC, poderiam também utilizar essas informações para direcionar seus programas de formação e capacitação, ajustando a oferta de cursos de acordo com a demanda projetada para certos perfis profissionais na indústria da construção civil (9).

Além disso, os formuladores de políticas poderiam se beneficiar dessas previsões para desenvolver e implementar políticas públicas mais eficientes no que diz respeito à geração de empregos, formação de mão de obra e desenvolvimento econômico (10). Essas previsões também poderiam ajudar as agências governamentais a alocar melhor os recursos destinados ao setor e mitigar o impacto de possíveis flutuações na demanda de empregos.

Do ponto de vista acadêmico, a contribuição deste trabalho reside no desenvolvimento e comparação de diferentes modelos de previsão de séries temporais aplicados a um setor específico da economia brasileira. Além disso, a incorporação de técnicas de otimização de hiperparâmetros contribui para a literatura na área de ciência da computação e análise de séries temporais, fornecendo informações sobre a eficácia dessas técnicas na melhoria da precisão das previsões de séries temporais (6).

1.2 Objetivos gerais e específicos

1.2.1 Objetivo geral

O objetivo deste trabalho é desenvolver um modelo de previsão da demanda de empregos no setor da construção civil em Alagoas, utilizando algoritmos de séries temporais para fazer analise exploratória em dados históricos de fontes públicas do governo com informações sobre pessoas empregadas neste setor.

1.2.2 Objetivos específicos

Para alcançar o objetivo geral desse trabalho, os seguintes objetivos específicos foram definidos:

- Coletar dados mensais de empregados na construção civil em Alagoas, utilizando fontes das bases RAIS e CAGED;
- Analisar a decomposição e sazonalidade das séries temporais destas bases no contexto do estado de Alagoas;
- Aplicar e comparar diversos algoritmos de previsão, incluindo Suavização Exponencial Holt Winters, ARIMA, SARIMA, TBATS e Prophet após a otimização dos hiperparâmetros;
- Avaliar o desempenho dos modelos de previsão utilizando técnicas de validação cruzada e métricas de desempenho baseadas no erro entre os dados reais e os dados preditos;
- Identificar o modelo de previsão com melhor desempenho para a demanda de empregos na construção civil em Alagoas;

1.3 Estrutura do trabalho

A organização deste trabalho é a seguinte: no capítulo 2 será apresentada uma fundamentação teórica sobre séries temporais, a indústria da construção civil em Alagoas, e os modelos de previsão de séries temporais utilizados neste estudo. No capítulo 3 será descrita a metodologia empregada, incluindo a coleta e análise de dados, a aplicação e otimização dos modelos de previsão, e a avaliação do desempenho dos modelos. No capítulo 4 serão apresentados os resultados obtidos e uma discussão sobre eles. Finalmente, no capítulo 5 serão apresentadas as contribuições do estudo, suas limitações e as sugestões para trabalhos futuros.

2

Fundamentação Teórica

Neste capítulo serão descritas as bases de dados públicas utilizadas neste trabalho. Uma revisão sobre séries temporais também está contida neste capítulo abordando estratégias de previsão. Os modelos utilizados para previsão de séries temporais serão brevemente descritos, bem como a avaliação de desempenho destes modelos. Por fim, será discutido processo de otimização de hiperparâmetros destes modelos e a técnica de validação cruzada para garantir melhores resultados eliminando a possibilidade de *overfitting*.

2.1 Bases de Dados de Empregos

A análise de dados de empregos é uma área de estudo crítica, e três principais fontes de dados no Brasil são o CAGED, o RAIS e a PNAD. Estas bases de dados contêm informações vitais sobre o mercado de trabalho brasileiro.

2.1.1 CAGED - Cadastro Geral de Empregados e Desempregados

Os dados do CAGED são coletados mensalmente de empresas em todo o país, e são usados para monitorar as flutuações no emprego formal. As informações disponíveis no CAGED são essenciais para a compreensão das tendências do mercado de trabalho, incluindo setores em crescimento ou declínio, e a demanda por diferentes ocupações (4).

Em 2020, o CAGED passou por uma reformulação, dando origem ao que é agora conhecido como "Novo CAGED". Esta nova versão da base de dados incorpora mudanças significativas, incluindo a expansão do escopo para incluir todos os contratos de trabalho, não apenas aqueles sob a CLT. Além disso, o Novo CAGED agora inclui informações sobre contratos de trabalho temporários e intermitentes, proporcionando uma visão mais completa do mercado de trabalho brasileiro (11).

2.1.2 RAIS - Relação Anual de Informações Sociais

A RAIS é uma fonte de dados valiosa para o estudo do mercado de trabalho brasileiro, pois oferece uma visão detalhada das características dos trabalhadores e empregadores, incluindo idade, gênero, escolaridade, salário e setor econômico, permitindo coletar a quantidade de pessoas empregadas no ano de referência. (7).

Para um estudo compreensivo do mercado de trabalho, é essencial levar em conta todas estas bases de dados. O CAGED e o Novo CAGED trazem informações atualizadas acerca das dinâmicas de contratação e demissão, enquanto a RAIS oferece uma visão anual mais ampla e detalhada do emprego formal.

2.1.3 PNAD - Pesquisa Nacional por Amostra de Domicílios

A PNAD é uma pesquisa domiciliar realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) que coleta dados sobre uma ampla gama de indicadores sociais, econômicos e demográficos, incluindo o emprego. A pesquisa é realizada em uma amostra representativa de domicílios em todo o país, e fornece informações detalhadas sobre a população em idade ativa, incluindo a taxa de participação na força de trabalho, a taxa de desemprego, a duração do desemprego, e a natureza do emprego (por exemplo, emprego formal versus informal) (12).

Além disso, a PNAD fornece informações sobre uma variedade de outros indicadores sociais e econômicos, como educação, renda, habitação e saúde, permitindo uma compreensão mais completa do contexto social e econômico em que o emprego ocorre. Entretanto, a PNAD é uma pesquisa realizada manualmente e divulgada apenas uma vez por ano, ao final de cada ano, o que gera uma base de dados com certo atraso nas informações. Por este motivo, esta base não será utilizada neste trabalho.

2.2 Previsão utilizando series temporais

Séries temporais são sequências de dados observados ao longo do tempo, onde cada observação corresponde a um intervalo de tempo específico, como minutos, horas, dias ou anos (13). As séries temporais são comuns em diversos campos, como economia, finanças, meteorologia e engenharia, e possuem aplicações variadas, como previsão de demanda, análise de tendências e detecção de anomalias. As características das séries temporais incluem:

- **Tendência**: a direção geral da série temporal ao longo do tempo (crescente, decrescente ou estável). A tendência pode ser linear ou não linear e pode variar com o tempo (13).
- **Sazonalidade**: os padrões que se repetem em intervalos fixos e conhecidos, como diariamente, mensalmente ou anualmente (13). A sazonalidade pode ser aditiva, onde as flutuações sazonais são constantes, ou multiplicativa, onde as flutuações sazonais são proporcionais ao nível da série temporal.

- **Ciclicidade**: os padrões que se repetem em intervalos irregulares e desconhecidos, geralmente associados a ciclos econômicos ou eventos externos (13).
- **Aleatoriedade (ruído)**: a variação não explicada na série temporal, que pode ser considerada como um processo estocástico ou ruído branco (13).

A análise de séries temporais busca entender essas características e usar esse conhecimento para fazer previsões, controlar processos e tomar decisões informadas. Além disso, a análise dessas séries busca entender e extrapolar padrões existentes nos dados para fazer previsões futuras (14). Essas previsões são fundamentais para uma ampla variedade de aplicações, incluindo a previsão da demanda de empregos na construção civil em Alagoas.

Os métodos de previsão de séries temporais são baseados na ideia de que o passado contém informações importantes sobre o futuro, e a análise desses dados pode identificar padrões e tendências que se repetem ao longo do tempo (14). Os métodos quantitativos de previsão de séries temporais são projetados para identificar e separar os padrões subjacentes nos dados das variações aleatórias e, em seguida, usar esses padrões para fazer previsões futuras.

2.3 Modelos Utilizados para Previsão

Nesta seção, detalhamos os principais modelos de previsão de séries temporais utilizados neste trabalho.

Começamos com o modelo ARIMA (*AutoRegressive Integrated Moving Average*), uma técnica amplamente adotada em diversas aplicações de previsão de séries temporais devido à sua eficácia em lidar com dependências temporais.

Prosseguimos para o modelo SARIMA (*Seasonal AutoRegressive Integrated Moving Average*), que é uma extensão do ARIMA que adiciona componentes sazonais, tornando-o ideal para lidar com séries temporais que apresentam padrões sazonais.

Em seguida, abordamos o método de Suavização Exponencial Holt Winters, um método robusto de previsão de séries temporais, capaz de capturar componentes de tendência e sazonalidade em séries temporais, além de ser notavelmente simples de implementar e interpretar.

Posteriormente, discutimos o modelo TBATS (*Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components*), um modelo mais sofisticado que pode lidar com várias sazonalidades, tendências não lineares, e transformações complexas, oferecendo flexibilidade para se ajustar a uma ampla gama de séries temporais.

Por fim, falamos sobre o modelo Prophet, uma técnica moderna para previsão de séries temporais que foi projetada para lidar com a sazonalidade em diferentes escalas de tempo e incorporar efeitos de feriados e eventos especiais.

Nesta seção, também detalhamos como a performance de cada um desses modelos foi avaliada, os critérios utilizados para selecionar o modelo mais apropriado para nossos dados, como

os hiperparâmetros de cada modelo foram otimizados e como a validação cruzada foi utilizada para prevenir o *overfitting* e garantir a robustez dos resultados da previsão.

2.3.1 ARIMA

O ARIMA (*AutoRegressive Integrated Moving Average*) é um modelo versátil para prever séries temporais, particularmente útil para dados que mostram uma tendência ao longo do tempo ou que não são estacionários. A estacionariedade é uma propriedade importante nas séries temporais e se refere à consistência das propriedades estatísticas ao longo do tempo. Quando uma série temporal não é estacionária, pode-se aplicar a técnica de diferenciação uma ou mais vezes para torná-la estacionária. (15).

O modelo ARIMA é comumente denotado por ARIMA(p,d,q), onde os parâmetros p,d,q são números inteiros não negativos que indicam a ordem dos componentes autorregressivos, a ordem da diferenciação e a ordem dos componentes de média móvel, respectivamente. A descrição desses parâmetros é a seguinte:

- p é a ordem do componente *AutoRegressive* (AR). Um modelo AR é onde o valor atual de uma série é uma combinação linear dos valores anteriores mais um erro branco.
- d é o número de diferenciações requeridas para tornar a série temporal estacionária. A diferenciação é o processo de subtrair a observação atual da observação anterior.
- q é a ordem do componente *Moving Average* (MA). Um modelo MA é onde o valor atual de uma série é uma combinação linear dos erros brancos passados.

A equação do modelo ARIMA pode ser escrita como:

$$(1-\sum_{i=1}^{p}\phi_i L^i)(1-L)^d Y_t = (1+\sum_{i=1}^{q}\theta_i L^i)\varepsilon_t \tag{2.1}$$

onde L é o operador de defasagem, Y_t é a observação no tempo t, ϕ_i são os parâmetros do modelo AR, θ_i são os parâmetros do modelo MA, d é a ordem de diferenciação, e ε_t é o erro no tempo t (16). No contexto deste trabalho, o modelo ARIMA foi aplicado na previsão de demanda de empregos na construção civil em Alagoas.

2.3.2 SARIMA

O modelo SARIMA (Seasonal Autoregressive Integrated Moving Average) é uma extensão do ARIMA, incorporando elementos sazonais na análise. Este modelo é particularmente útil quando padrões ou tendências repetem em intervalos fixos dentro de uma série temporal (15).

O SARIMA é denotado como SARIMA$(p,d,q;P,D,Q,s)$, onde os parâmetros adicionais P,D,Q e s capturam a sazonalidade da série:

- P é a ordem do componente sazonal autoregressivo.
- D é a ordem da diferenciação sazonal.
- Q é a ordem do componente sazonal de média móvel.
- s é o período da sazonalidade.

Podemos expressar a equação do modelo SARIMA da seguinte maneira:

$$(1-\sum_{i=1}^{p}\phi_i L^i)(1-\sum_{i=1}^{P}\Phi_i L^{i\cdot s})(1-L)^d(1-L^s)^D Y_t = (1+\sum_{i=1}^{q}\theta_i L^i)(1+\sum_{i=1}^{Q}\Theta_i L^{i\cdot s})\varepsilon_t, \tag{2.2}$$

onde L é o operador de defasagem, Y_t a observação no tempo t, ϕ_i, Φ_i, θ_i, e Θ_i são os parâmetros dos modelos AR e MA, e seus correspondentes sazonais, d e D são as ordens das diferenciações não sazonal e sazonal, e ε_t é o erro no tempo t (16). No âmbito deste trabalho, o modelo SARIMA é utilizado para prever a demanda por empregos na indústria da construção civil em Alagoas.

2.3.3 Suavização Exponencial Holt Winters

A Suavização Exponencial Holt Winters é um método robusto e amplamente utilizado de previsão de séries temporais, capaz de capturar componentes de nível, tendência e sazonalidade em séries temporais (14). O método é uma extensão da suavização exponencial de Holt, permitindo acomodar a sazonalidade presente nos dados.

Existem duas variantes principais do método Holt Winters: o modelo aditivo e o modelo multiplicativo. O modelo aditivo é mais adequado quando as variações sazonais são aproximadamente constantes ao longo da série, enquanto o modelo multiplicativo é apropriado quando as variações sazonais estão mudando proporcionalmente ao nível da série.

Os parâmetros de suavização do modelo Holt Winters são geralmente expressos como (α, β, γ), representando o nível, a tendência e a sazonalidade, respectivamente. Estes parâmetros são selecionados de maneira a minimizar o erro de previsão.

As equações para o modelo aditivo de Suavização Exponencial Holt Winters são:

$$\ell_t = \alpha(y_t - s_{t-L}) + (1-\alpha)(\ell_{t-1} + b_{t-1}), \tag{2.3}$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}, \tag{2.4}$$

$$s_t = \gamma(y_t - \ell_t) + (1-\gamma)s_{t-L}, \tag{2.5}$$

$$\hat{y}_{t+m} = \ell_t + mb_t + s_{t-L+1+(m-1)modL}, \tag{2.6}$$

E para o modelo multiplicativo, as equações são:

$$\ell_t = \alpha \frac{y_t}{s_{t-L}} + (1-\alpha)(\ell_{t-1} + b_{t-1}), \tag{2.7}$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}, \tag{2.8}$$

$$s_t = \gamma \frac{y_t}{\ell_t} + (1-\gamma)s_{t-L}, \tag{2.9}$$

$$\hat{y}_{t+m} = (\ell_t + mb_t)s_{t-L+1+(m-1)modL}, \tag{2.10}$$

Onde, y_t é a observação no tempo t, ℓ_t é o componente de nível no tempo t, b_t é o componente de tendência no tempo t, s_t é o componente de sazonalidade no tempo t, L é o período da sazonalidade, e $\hat{y}_{t+m}$ é a previsão m passos à frente (14).

Neste trabalho, tanto o modelo aditivo quanto o multiplicativo da Suavização Exponencial Holt Winters foram explorados para prever a demanda de empregos na construção civil em Alagoas, e o melhor modelo foi selecionado com base na validação cruzada.

2.3.4 TBATS

TBATS é um modelo de previsão de séries temporais desenvolvido por De Livera, Hyndman e Snyder em 2011. O acrônimo TBATS é uma referência às principais características do modelo: *Trigonometric seasonality*, *Box-Cox transformation*, *ARMA errors*, *Trend* e *Seasonal components* (17).

TBATS é um modelo projetado para lidar com a complexidade das séries temporais que exibem múltiplos padrões sazonais e não necessariamente inteiros. O modelo utiliza uma transformação de Box-Cox para estabilizar a variância, um componente de tendência autorregressivo para capturar a tendência não sazonal, e componentes sazonais autorregressivos, usando funções trigonométricas para representar a sazonalidade, permitindo que a duração do período sazonal seja um número real, não necessariamente inteiro (17).

O método TBATS se destaca em comparação com outros métodos de previsão de séries temporais por sua capacidade de lidar com várias formas de sazonalidade, mesmo quando o período sazonal é longo. Isso o torna especialmente útil para séries temporais com padrões sazonais complexos que não podem ser capturados adequadamente por métodos de previsão tradicionais (14). Para um modelo TBATS, a formulação do modelo é dada por:

$$y_t^{(\lambda)} = l_{t-1} + \phi b_{t-1} + s_t + d_t, \tag{2.11}$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t, \tag{2.12}$$

$$b_t = \phi b_{t-1} + \beta d_t, \tag{2.13}$$

$$s_t = s_{t-m} - \gamma d_t. \tag{2.14}$$

onde $y_t^{(\lambda)}$ é a série temporal transformada pelo Box-Cox, l_t é o componente de nível, b_t é o

componente de tendência, s_t é o componente sazonal, d_t é o erro que segue um modelo ARMA e α, β, ϕ, e γ são parâmetros a serem estimados (17).

Devido à sua capacidade de modelar complexas estruturas de sazonalidade, o TBATS tem se mostrado eficaz na previsão de séries temporais em uma variedade de campos, desde a demanda de varejo até a utilização de recursos de saúde (14).

2.4 Prophet

Prophet é uma ferramenta de previsão open-source publicada pela equipe de ciência de dados do Facebook e está disponível em Python (18) e R (19). O Prophet foi desenvolvido para lidar com problemas típicos do Facebook, como prever atividades de usuários. Isso torna o método Prophet útil para prever sazonalidades, eventos especiais, dados com feriados, dados mostrando outliers e dados com tendência variável.

O método Prophet utiliza um framework chamado "*Analyst-in-the-Loop*", como mostrado na Figura 2.1.

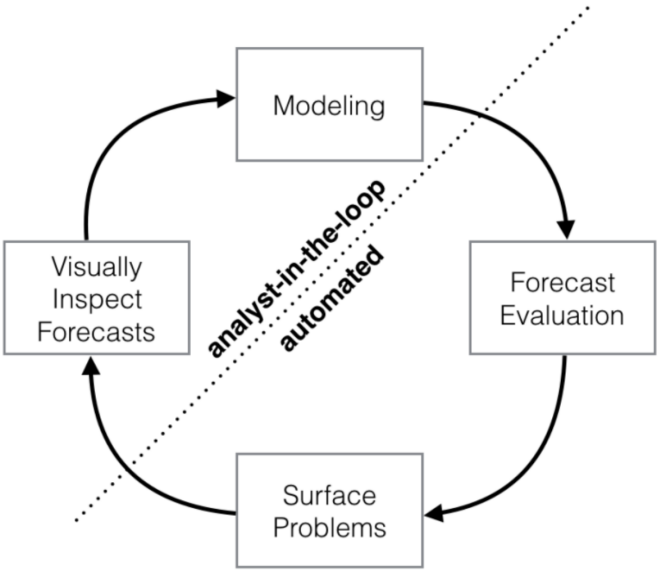

Figura 2.1: Modelagem Analyst-in-the-Loop(1)

O framework é bidirecional, onde, de um lado, o ajuste do modelo é automatizado, assumindo que o usuário não tem conhecimento estatístico, enquanto, do outro lado, o framework permite que o mesmo usuário insira informações com base em seu conhecimento de domínio ou indústria.

O método Prophet é um método de regressão aditiva que pertence à família de Modelos Aditivos Generalizados (GAM) com os seguintes componentes e forma funcional:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{2.15}$$

Na equação acima, $g(t)$ captura a tendência na série temporal, $s(t)$ captura a sazonalidade da série temporal, $h(t)$ captura feriados ou eventos especiais na série temporal e ε_t é um termo de erro irredutível. Em qualquer instância do método Prophet, apenas ε_t está sempre presente, os três termos restantes podem não estar sempre presentes, pois precisam ser fornecidos pelo usuário. Agora, vamos explicar cada componente em mais detalhes:

- $g(t)$ modela alterações não periódicas (tendência) na série temporal. A biblioteca Prophet implementa o modelo de tendência linear por partes, onde a taxa de crescimento permanece constante, e o modelo de tendência não linear, onde a taxa de crescimento diminui com o tempo t. Na implementação atual do Prophet, a taxa de crescimento não pode aumentar com o tempo.

- $s(t)$ Sazonalidade representa mudanças periódicas (sazonalidade diária/semanal/mensal/anual) na série temporal.

- $h(t)$ O componente de feriados contribui com informações sobre feriados e eventos, conforme fornecido pelo usuário.

O método Prophet usa uma técnica de ajuste de curva para o ajuste da série temporal. O método é ajustado automaticamente usando o código Stan (20) que leva em consideração sazonalidade, tendências e feriados. A robustez do Prophet, a facilidade de configuração e a rapidez de ajuste atraem usuários não especialistas e com conhecimento estatístico limitado para implantar o Prophet em suas organizações.

2.5 Avaliação de Desempenho de Modelos de Previsão

A avaliação do desempenho de modelos de previsão é uma etapa crucial em qualquer projeto de previsão. Existem várias métricas que podem ser usadas para avaliar o desempenho de um modelo de previsão. As utilizadas neste trabalho incluem a raiz do erro quadrático médio (RMSE), o erro percentual absoluto médio (MAPE) e Erro Médio Simétrico Absoluto Percentual (sMAPE)

2.5.1 Raiz do Erro Quadrático Médio (RMSE)

O RMSE é a raiz quadrada do MSE. O Erro Quadrático Médio (MSE) é a média dos quadrados dos erros. A Métrica RMSE tem a vantagem de ter as mesmas unidades que a variável de interesse, o que pode facilitar a interpretação dos resultados. É calculado da seguinte maneira:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.16}$$

2.5.2 Erro Percentual Absoluto Médio (MAPE)

O MAPE é a média dos valores absolutos dos erros percentuais. Esta métrica é expressa como uma porcentagem, e é particularmente útil quando você quer expressar o erro de previsão em termos relativos, em vez de absolutos. É calculado da seguinte maneira:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{2.17}$$

Cada uma destas métricas tem suas próprias vantagens e desvantagens, e a escolha da métrica de avaliação deve ser baseada no problema específico de previsão e nos objetivos do projeto (21).

2.5.3 Erro Médio Simétrico Absoluto Percentual (sMAPE)

O Erro Médio Simétrico Absoluto Percentual (sMAPE) é uma extensão do MAPE que corrige um de seus principais problemas: a assimetria. O MAPE penaliza mais os erros de previsão que são superestimados do que aqueles que são subestimados. O sMAPE, por outro lado, trata esses erros de maneira mais equitativa. É calculado da seguinte maneira (22):

$$sMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{\frac{1}{2}(|y_i| + |\hat{y}_i|)} \tag{2.18}$$

onde y_i é o valor real, $\hat{y}_i$ é o valor previsto e n é o número total de pontos de dados.

O sMAPE possui a vantagem de manter uma interpretação em porcentagem semelhante ao MAPE, porém evita a assimetria deste.

2.6 Otimização de Hiperparâmetros

A otimização de hiperparâmetros é uma etapa crucial na construção de modelos de aprendizado de máquina e séries temporais, pois a eficácia e a precisão do modelo resultante podem depender fortemente da seleção adequada desses hiperparâmetros. Os hiperparâmetros são, em essência, as configurações de um algoritmo de aprendizado de máquina que são definidas previamente antes do início do processo de treinamento (6).

A técnica de busca em grade, ou *Grid Search*, é uma das abordagens mais comumente utilizadas para a otimização de hiperparâmetros. Este método envolve a definição de uma grade de valores possíveis para cada hiperparâmetro, seguida pelo treinamento e avaliação do modelo para todas as combinações desses valores (23).

Em termos mais concretos, se tivermos dois hiperparâmetros e decidirmos avaliar cada um em 10 valores diferentes, a busca em grade avaliará o modelo com todas as 100 combinações possíveis desses hiperparâmetros. A avaliação do modelo é tipicamente baseada em uma medida de erro (como o erro quadrático médio para problemas de regressão) ou uma medida de

acurácia (como a acurácia ou a área sob a curva ROC para problemas de classificação) (6).

Embora a busca em grade seja uma técnica exaustiva, pois avalia todas as combinações possíveis de hiperparâmetros, ela pode ser computacionalmente cara, especialmente quando o número de hiperparâmetros e os possíveis valores para cada um é grande. Alternativas à busca em grade, como a busca aleatória ou a otimização bayesiana, podem às vezes ser mais eficientes, embora possam não garantir que a solução ótima seja encontrada (6; 23).

Na otimização de hiperparâmetros, é importante ter cuidado para evitar o *overfitting* - quando o modelo se ajusta muito bem aos dados de treinamento, mas tem um desempenho ruim em dados não vistos. Para isso, é comum usar técnicas de validação cruzada durante o processo de otimização, de modo a garantir que a seleção dos hiperparâmetros não seja excessivamente otimizada para os dados de treinamento (24).

Desta forma, a otimização de hiperparâmetros por meio do *Grid Search* é uma estratégia eficaz e amplamente aplicada para a melhoria do desempenho dos modelos de séries temporais e aprendizado de máquina.

2.7 Validação Cruzada em Séries Temporais

A validação cruzada é um método crucial na avaliação de desempenho de modelos de previsão ou classificação na aprendizagem de máquinas e estatística. Esta técnica envolve a divisão do conjunto de dados em diversos subconjuntos. O modelo é treinado em um subconjunto (o conjunto de treinamento) e testado em outro subconjunto (o conjunto de teste). Este procedimento é repetido várias vezes, cada vez com um conjunto de teste diferente, para proporcionar uma avaliação robusta do desempenho do modelo (25).

No entanto, no contexto das séries temporais, a aplicação da validação cruzada pode ser desafiadora devido à dependência temporal inerente aos dados. Em séries temporais, as observações são frequentemente interdependentes de uma forma que não se encontra em outros tipos de dados. Assim, os métodos padrão de validação cruzada, que pressupõem a independência entre as observações, podem não ser adequados.

Uma solução para esse problema é a validação cruzada do tipo *walk-forward*. Esta técnica permite que o modelo seja treinado inicialmente com um subconjunto dos dados, depois o modelo faz previsões para o próximo ponto no tempo. Este ponto é então incluído no conjunto de treinamento, e a previsão é feita para o próximo ponto. Esse processo é repetido até que todas as observações na série temporal tenham sido previstas (26). A Figura 2.2 ilustra este processo.

A validação cruzada "walk-forward"é especialmente útil para séries temporais porque considera a dependência temporal dos dados e a possível mudança nas propriedades da série temporal ao longo do tempo, fenômeno conhecido como não estacionariedade. Ao fazer previsões um passo à frente com base nos dados disponíveis até esse ponto, a técnica fornece uma avaliação

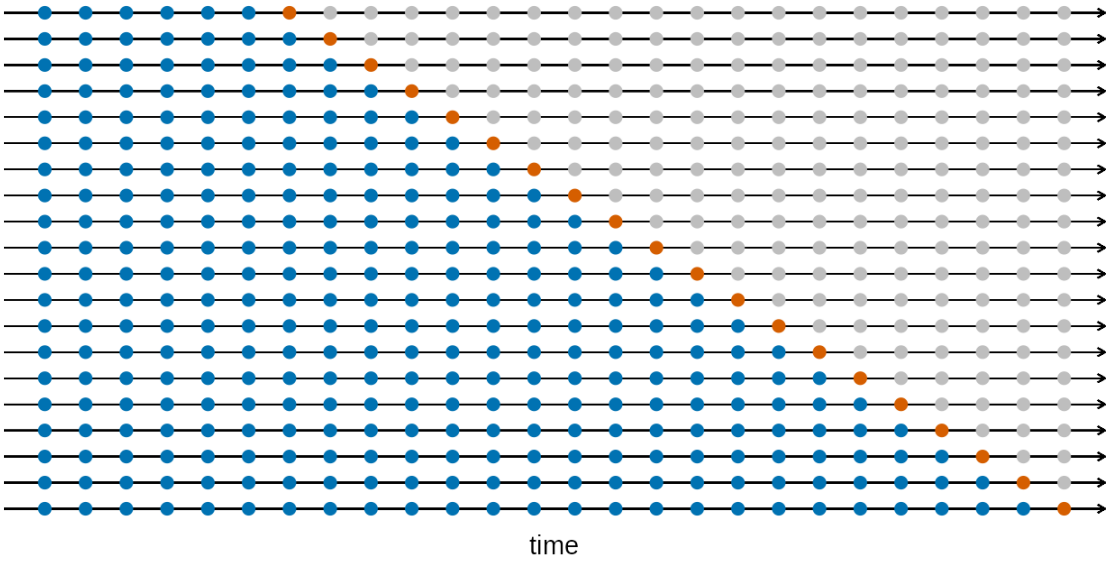

Figura 2.2: Ilustração do processo de validação cruzada walk-forward. O modelo é inicialmente treinado em uma parte da série temporal, faz previsões para o próximo ponto, inclui este ponto no conjunto de treinamento e repete o procedimento até todas as observações serem previstas.

robusta do desempenho do modelo de previsão em condições "reais"(14).

Importante destacar que a implementação correta da validação cruzada em séries temporais é essencial para obter uma estimativa precisa do desempenho do modelo de previsão e para evitar o fenômeno de *overfitting*, que ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, mas falha em generalizar para novos dados (27).

3

Metodologia

Neste capítulo, abordamos a metodologia completa adotada neste estudo, desde a coleta inicial dos dados até a aplicação e otimização dos modelos de previsão. Detalhamos o processo de análise exploratória de dados, decomposição da série temporal e análise de sazonalidade, e explicamos como aplicamos e otimizamos vários modelos de previsão de séries temporais, incluindo Suavização Exponencial Holt Winters, ARIMA, SARIMA, TBATS e Prophet. Por fim, descrevemos a técnica de validação cruzada com cortes temporais específicos usada para avaliar o desempenho dos modelos e garantir sua eficácia em diferentes conjuntos de dados.

A análise foi conduzida no ambiente Python (18), utilizando as bibliotecas pandas (28), numpy (29), matplotlib (30), statsmodels (31), scikit-learn (32), plotly (33) e fbprophet (1). A análise foi realizada no Jupyter Notebook (34).

3.1 Coleta de dados

O primeiro e fundamental passo em qualquer análise estatística é a coleta de dados. No presente estudo, os dados da RAIS[1], CAGED[2] e Novo CAGED[3] foram adquiridos das bases de microdados disponíveis nos endereços *ftp* do Programa de Disseminação das Estatísticas do Trabalho (PDET) administrados pelo Ministério do Trabalho e Emprego do Brasil. O PDET, uma iniciativa do Ministério da Economia, tem como objetivo disponibilizar informações cada vez mais abrangentes sobre o mercado de trabalho, utilizando diferentes tipos de mídia e, assim, alcançando diferentes grupos de usuários.

Para o objetivo deste estudo, inicialmente foi obtido o saldo de pessoas empregadas no setor de construção civil em Alagoas em 31 de dezembro de 2006, através da RAIS. A partir dessa data, foram aplicados mensalmente os saldos de admissões e demissões fornecidos pelo

[1] ftp://ftp.mtps.gov.br/pdet/microdados/RAIS
[2] ftp://ftp.mtps.gov.br/pdet/microdados/CAGED
[3] ftp://ftp.mtps.gov.br/pdet/microdados/NOVO CAGED

CAGED. Tal metodologia foi adotada até dezembro de 2019, quando a fonte dos dados foi alterada para o Novo CAGED, que iniciou o fornecimento de informações em janeiro de 2020.

Portanto, a série temporal compilada para este estudo combina dados do CAGED de janeiro de 2007 a dezembro de 2019 com dados do Novo CAGED de janeiro de 2020 a março de 2023. Cada ponto de dados na série temporal representa o saldo de trabalhadores empregados no setor de construção civil em Alagoas em um determinado mês.

Antes de proceder com qualquer análise, garantimos que a qualidade e consistência dos dados fossem verificadas. Essa verificação envolveu a identificação e o tratamento de dados ausentes, valores discrepantes e quaisquer outros problemas que pudessem interferir na subsequente análise dos dados.

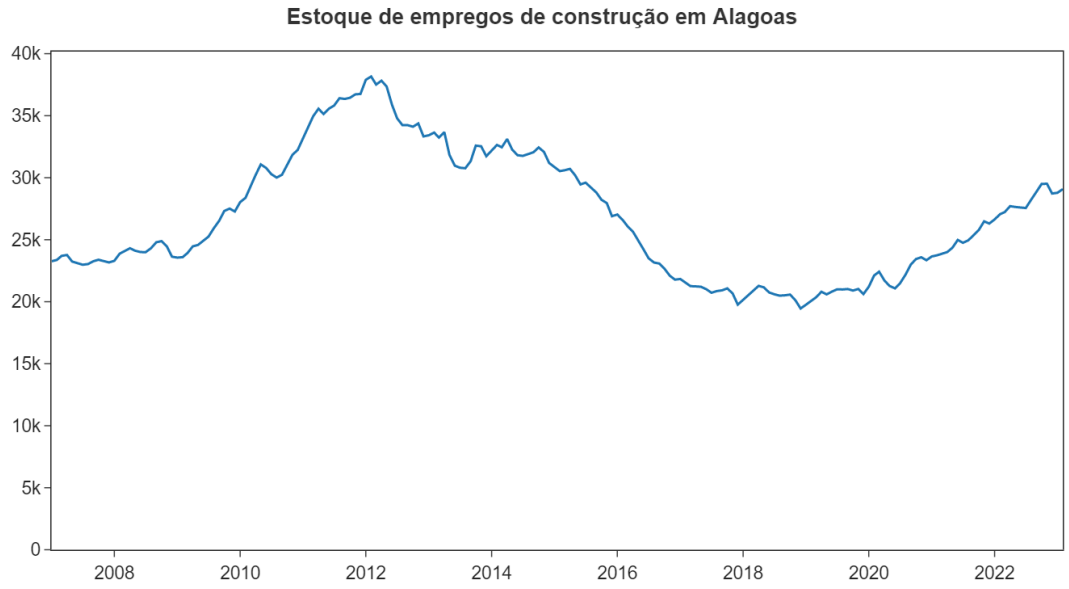

Figura 3.1: Gráfico da série temporal de empregos na construção civil em Alagoas

3.2 Análise Exploratória

O processo de análise é conduzido em três fases principais: Decomposição da Série Temporal e Análise de Sazonalidade. Cada fase emprega métodos específicos para examinar diferentes aspectos da série temporal de empregos na construção civil em Alagoas.

A Decomposição da Série Temporal, que é um passo essencial para entender os padrões e comportamentos subjacentes da série. Utilizamos o método STL (Seasonal-Trend decomposition using LOESS) para decompor a série em componentes de tendência, sazonalidade e resíduos.

Por fim, na Análise de Sazonalidade, buscamos identificar flutuações regulares na série temporal de empregos na construção civil. Para isso, utilizamos gráficos sazonais que representam

cada ciclo sazonal (ano) como uma linha separada, ajudando a revelar tendências e padrões ocultos.

É importante ressaltar que as análises detalhadas e as respectivas figuras serão discutidas no Capítulo 4. Este capítulo foca na metodologia utilizada, e os resultados das análises são apresentados no capítulo subsequente para facilitar a compreensão.

3.3 Métrica de Avaliação

Na realização desta pesquisa, escolhemos o Erro Quadrático Médio Raiz (RMSE) como a principal métrica de avaliação. Entretanto, também consideramos duas outras métricas para garantir uma boa comparação entre diferentes métodos, o Erro Percentual Absoluto Médio (MAPE) e o Erro Percentual Absoluto Médio Simétrico (sMAPE). A inclusão destas métricas se deve à sensibilidade do RMSE a *outliers*(35).

As três métricas apresentam suas próprias peculiaridades:

1. **RMSE:** Esta métrica, embora seja sensível a *outliers*, tem a vantagem de punir mais severamente erros grandes, o que é especialmente relevante em nosso estudo. O RMSE age principalmente como uma métrica de erro de previsão intra-modelo, e secundariamente como uma métrica de erro de previsão entre-modelos (35).

2. **MAPE:** O MAPE coloca uma penalidade maior em erros positivos do que em erros negativos . Esta métrica não é definida como a medição ótima de precisão de previsão, mas serve como uma métrica adicional de comparação entre-modelos (36).

3. **sMAPE:** Apesar do nome, o sMAPE não é simétrico, o que significa que não trata igualmente os erros de superestimação e subestimação . No entanto, assim como o MAPE, é útil como métrica de erro de previsão entre-modelos (22).

Embora o RMSE seja a nossa principal métrica de avaliação, o uso de três métricas diferentes para medir os erros de previsão entre os modelos deve oferecer uma orientação mais abrangente do que usar apenas uma. O MAPE e o sMAPE servem principalmente como métricas de erro de previsão entre modelos para comparação de modelos, e secundariamente para esclarecer a comparação geral em conjunto com o RMSE (22).

3.4 Abordagem para construção dos modelos de previsão

Este trabalho segue um processo de várias etapas para aplicar, otimizar e validar modelos de previsão de séries temporais, que envolve Suavização Exponencial Holt Winters, ARIMA, SARIMA, TBATS e Prophet.

1. **Aplicação dos Modelos** - Após uma análise exploratória aprofundada dos dados, os modelos de previsão de séries temporais são aplicados. Cada modelo é implementado conforme os detalhes fornecidos na seção 2.3.

2. **Otimização de Hiperparâmetros** - Uma vez aplicados, os modelos são otimizados para melhorar a precisão de suas previsões. A otimização de hiperparâmetros é realizada através da técnica de *Grid Search*, que sistematicamente varia os hiperparâmetros e avalia o desempenho do modelo em cada configuração. O objetivo da otimização é minimizar a Raiz do Erro Quadrático Médio (RMSE) nas previsões. Os parâmetros otimizados e escolhidos para cada modelo podem ser encontrados na seção 4.2.

3. **Validação Cruzada** - Para evitar *overfitting* e garantir a generalização dos modelos, é aplicada a validação cruzada durante o processo de otimização de hiperparâmetros. Neste trabalho, utilizamos o método de validação cruzada com cortes temporais específicos, Os pontos de corte foram definidos começando em Dezembro de 2010 avançando a cada 2 anos até Dezembro de 2022. Para cada ponto, o modelo é treinado com os dados até o ponto de corte e são feitas previsões para os próximos 12 meses. As previsões são então comparadas com os dados reais para calcular o erro da previsão. Este processo é repetido para cada ponto de corte, com o conjunto de hiperparâmetros que resulta no menor valor médio de RMSE sendo selecionado para o modelo final.

4. **Avaliação do Desempenho dos Modelos** - Por fim, o desempenho de cada modelo é avaliado. Isso é feito calculando métricas de erro como a Raiz do Erro Quadrático Médio (RMSE), Erro Percentual Absoluto Médio (MAPE), e o Erro Médio Simétrico Absoluto Percentual (sMAPE) a partir dos erros de previsão obtidos no processo de validação cruzada.

   O resultado deste processo estruturado de aplicação, otimização e validação é um conjunto de modelos de previsão de séries temporais altamente ajustados e eficazes, prontos para a análise de desempenho e comparação descrita no Capítulo 4.

4

Resultados e Discussões

4.1 Análise dos Dados

Nesta seção, apresentamos os resultados da análise exploratória dos dados, incluindo os gráficos produzidos a partir das análises de decomposição da série temporal e sazonalidade

4.1.1 Decomposição da Série Temporal

A análise da decomposição da série temporal é um passo fundamental em nossa investigação, que nos permite entender os principais componentes e padrões subjacentes à série. A série temporal é decomposta em componentes distintos: tendência, sazonalidade e resíduos, que juntos compõem a estrutura dos dados.

Neste estudo, a decomposição STL (Decomposição de Tendência Sazonal usando LOESS) é aplicada à nossa série temporal. A decomposição STL é um método robusto que nos permite separar a série em seus componentes de tendência, sazonalidade e residuais. A tendência revela a evolução de longo prazo da série. A sazonalidade destaca os padrões que se repetem em intervalos regulares de tempo. Os resíduos são o que resta após a remoção da tendência e sazonalidade (37).

A Figura 4.1 mostra a decomposição da série temporal dos nossos dados.

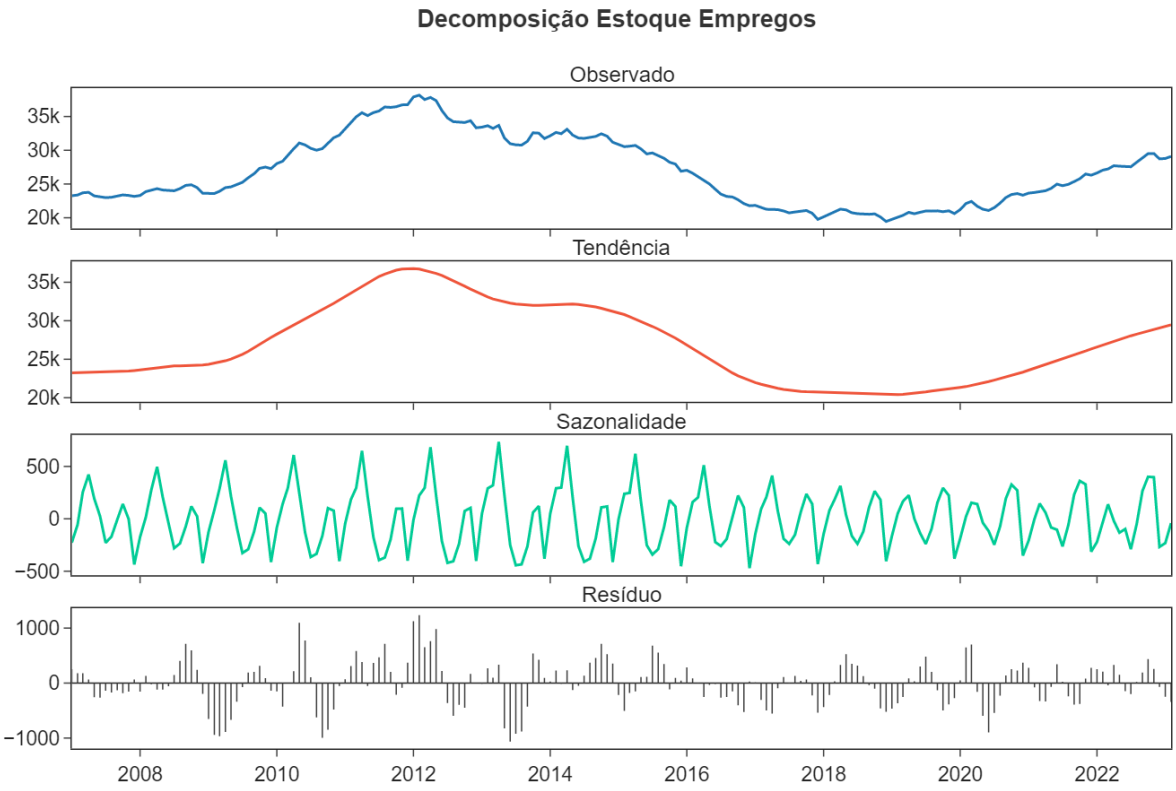

Figura 4.1: Decomposição da série temporal utilizando o método STL.

A partir da análise da decomposição da série temporal, notamos que a série apresenta uma tendência marcante de subida a partir de 2009, seguida por uma descida a partir de 2012. O período entre 2018 e 2020 é caracterizado por uma relativa estabilidade, e a série volta a subir a partir de 2020. No entanto, a falta de dados não permite a observação de um comportamento cíclico.

No que diz respeito à sazonalidade, até 2016, a série exibe um padrão de um pico maior no início do ano e um pico menor no final do ano. Esse padrão muda a partir de 2018, com picos de tamanho similar no início e no final do ano.

A análise dos resíduos indica que há uma maior concentração destes na primeira metade dos dados, entre 2008 e 2014. Isso pode indicar que a série apresentou mais ruído ou variabilidade nesse período em comparação com os anos subsequentes.

4.1.2 Análise de Sazonalidade

Na análise de séries temporais, especialmente para dados econômicos ou comerciais, a sazonalidade é um elemento crucial. É uma característica que expressa a recorrência de padrões em intervalos regulares de tempo.

Para investigar a sazonalidade na série temporal de empregos na construção civil em Alagoas, utilizamos gráficos sazonais. Esses gráficos apresentam cada ciclo sazonal (neste caso,

anual) como uma linha distinta, com o eixo x indicando os meses e o eixo y representando o valor da série temporal.

A Figura 4.2 ilustra o gráfico sazonal da nossa série temporal, expondo os padrões de sazonalidade presentes.

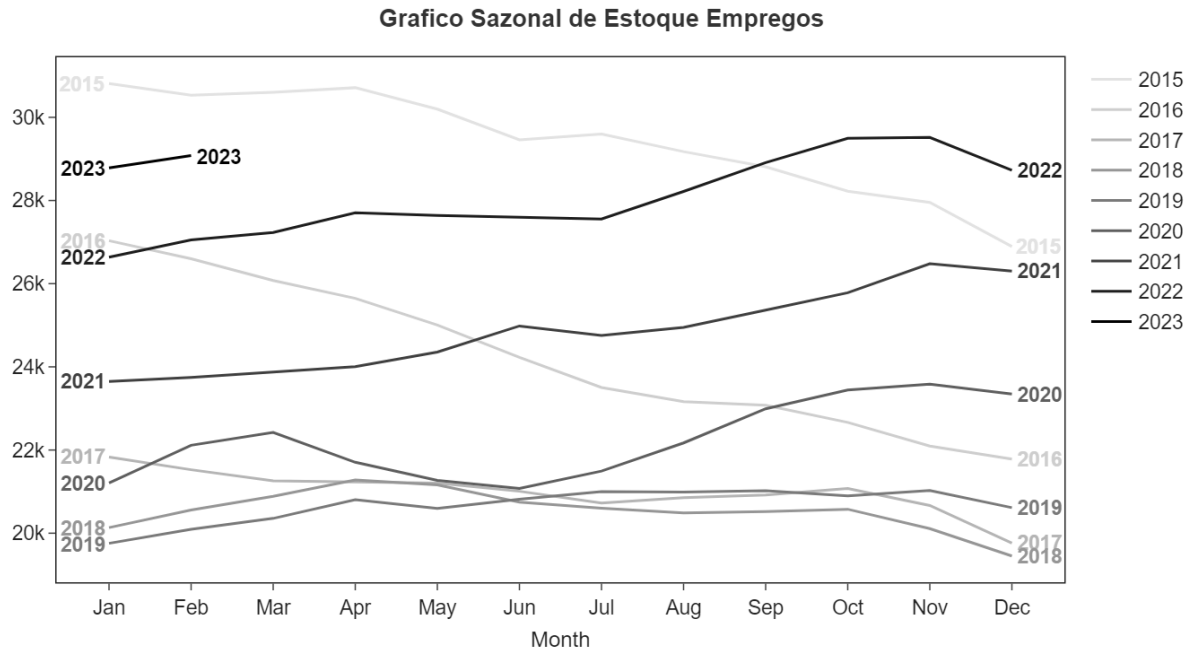

Figura 4.2: Gráfico sazonal da série temporal.

Através do gráfico sazonal, é possível revelar tendências e padrões que podem não ser imediatamente evidentes na série temporal original.

Nos anos mais recentes, a partir de 2020, nota-se uma sazonalidade de pico maior de contratações no final do ano (outubro e novembro) e início do ano (fevereiro até maio, dependendo do ano). Em contrapartida, há uma estabilização nos meses intermediários do ano (junho e julho). No período de 2015 a 2019, o gráfico exibe uma forte tendência decrescente, o que dificulta a visualização de um padrão sazonal robusto.

Essas observações evidenciam a importância de levar em conta a sazonalidade ao modelar e prever a série temporal de empregos na construção civil em Alagoas.

4.2 Desempenho dos Modelos

Neste estudo, analisamos e comparados diferentes métodos de modelagem de séries temporais para prever o número de empregos na indústria da construção civil em Alagoas, Brasil. Os modelos avaliados foram ARIMA, SARIMA, Suavização Exponencial Holt Winters e Prophet.

Cada um dos modelos tem suas próprias forças e fraquezas, e escolher o melhor modelo depende muito das características da série temporal em estudo. O ARIMA e SARIMA são

famosos por sua flexibilidade e capacidade de modelar várias formas diferentes de dados de séries temporais, enquanto o modelo de Suavização Exponencial Holt Winters é conhecido por ser especialmente útil para séries temporais com tendências claras e componentes sazonais. O Prophet, por outro lado, é uma abordagem relativamente nova que oferece um equilíbrio eficaz entre simplicidade e poder de previsão, além de ser fácil de usar.

Cada modelo foi otimizado através da escolha dos parâmetros que minimizaram a raiz do erro quadrado médio (RMSE) na validação cruzada. Além disso, a validade dos modelos foi avaliada com base em duas outras métricas de erro, a média percentual do erro absoluto (MAPE) e o erro percentual absoluto simétrico médio (sMAPE).

Os resultados dessas análises de desempenho são discutidos em detalhes nas subseções a seguir.

4.2.1 ARIMA

Nesta fase do estudo, aplicamos e avaliamos o modelo ARIMA (AutoRegressive Integrated Moving Average). Este modelo, extensivamente usado em previsões de séries temporais, permite a consideração de tendências, sazonalidades e outros componentes ao modelar os dados.

Parâmetros	Métricas de Erro		
Ordem(p,d,q)	RMSE	MAPE (%)	sMAPE (%)
(3,0,3)	1134.54	3.20	3.25

Tabela 4.1: Desempenho do melhor modelo ARIMA

Apos a etapa de otimização de hiperparâmetro o melhor conjunto dos parâmetros do modelo ARIMA como visto na Tabela 4.2.1 foi o com 3 *lags* passados(p), 0 ordem de diferenciações(d) e 3 *white noise* passados(q) obtendo um bom RMSE de 1134

Em uma análise visual da figura 4.3 as primeiras três previsões do modelo ARIMA para as datas 2010-12-01, 2012-12-01 e 2014-12-01 que possuíam menos dados de treinamento não se ajustaram tão bem quanto as previsões subsequentes. No entanto, em geral, o modelo conseguiu gerar previsões que estavam muito próximas dos valores reais, demonstrando a eficácia do ARIMA para esta série temporal.

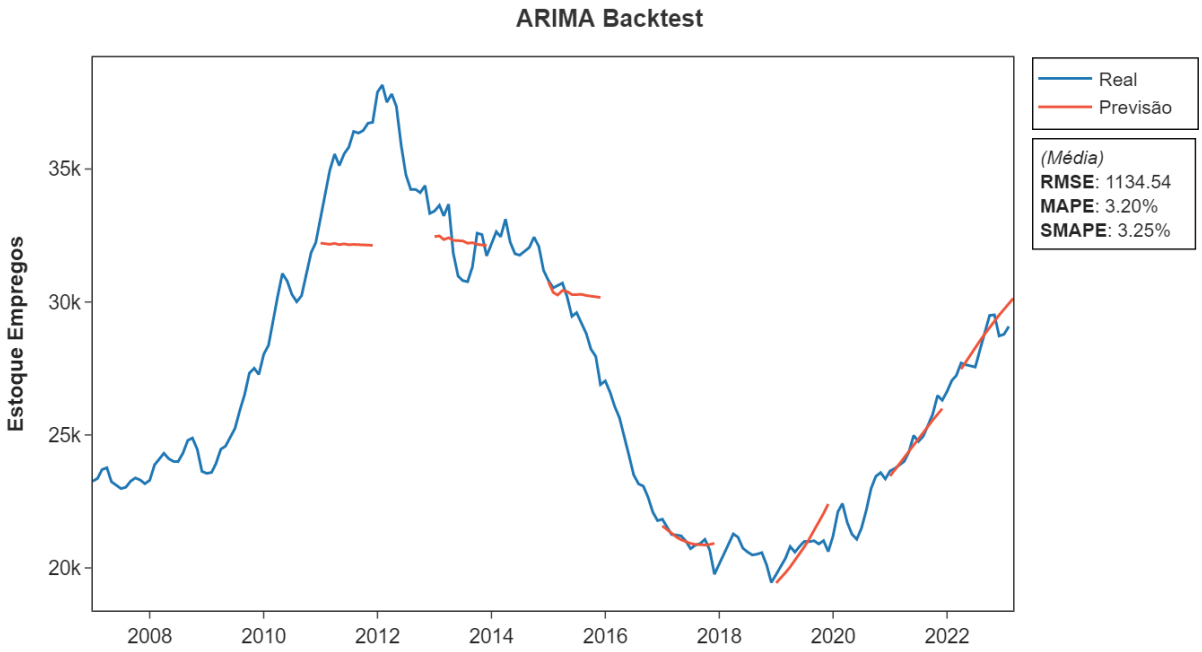

Figura 4.3: Gráfico do resultado da validação cruzada do modelo ARIMA com a média das métricas de erro para a previsão da série temporal de empregos na construção civil em Alagoas.

4.2.2 SARIMA

Em seguida, foi aplicado e avaliado o modelo SARIMA (Seasonal AutoRegressive Integrated Moving Average). Este modelo é uma extensão do ARIMA que adiciona a capacidade de modelar componentes sazonais na série temporal.

Os parâmetros não sazonais (p,d,q) e sazonais (P,D,Q,S) do modelo SARIMA de acordo com a otimização de parâmetros. O melhor conjunto de hiperparâmetros encontrado foi $(p=2,d=0,q=2)$ para os parâmetros não sazonais e $(P=0,D=1,Q=0,S=12)$ para os parâmetros sazonais, conforme demonstrado na Tabela 4.2.2.

Parâmetros		Métricas de Erro		
Ordem (p,d,q)	Ordem Sazonal (P,D,Q,S)	RMSE	MAPE (%)	sMAPE (%)
(2, 0, 2)	(0, 1, 0, 12)	967.85	3.03	2.99

Tabela 4.2: Desempenho do melhor modelo SARIMA

O parâmetro sazonal (P,D,Q,S) indica a presença de uma tendência sazonal nos dados, sendo 'D=1' uma indicação clara dessa tendência. Isso significa que o modelo identificou um padrão que se repete a cada 12 meses, consistente com um ciclo anual.

A análise visual do gráfico da Figura 4.4 mostra um ajuste visualmente muito bom para a maioria dos pontos de corte da validação cruzada. Isso indica que o modelo SARIMA foi capaz

de capturar adequadamente a estrutura subjacente da série temporal, levando em conta tanto os componentes não sazonais quanto sazonais.

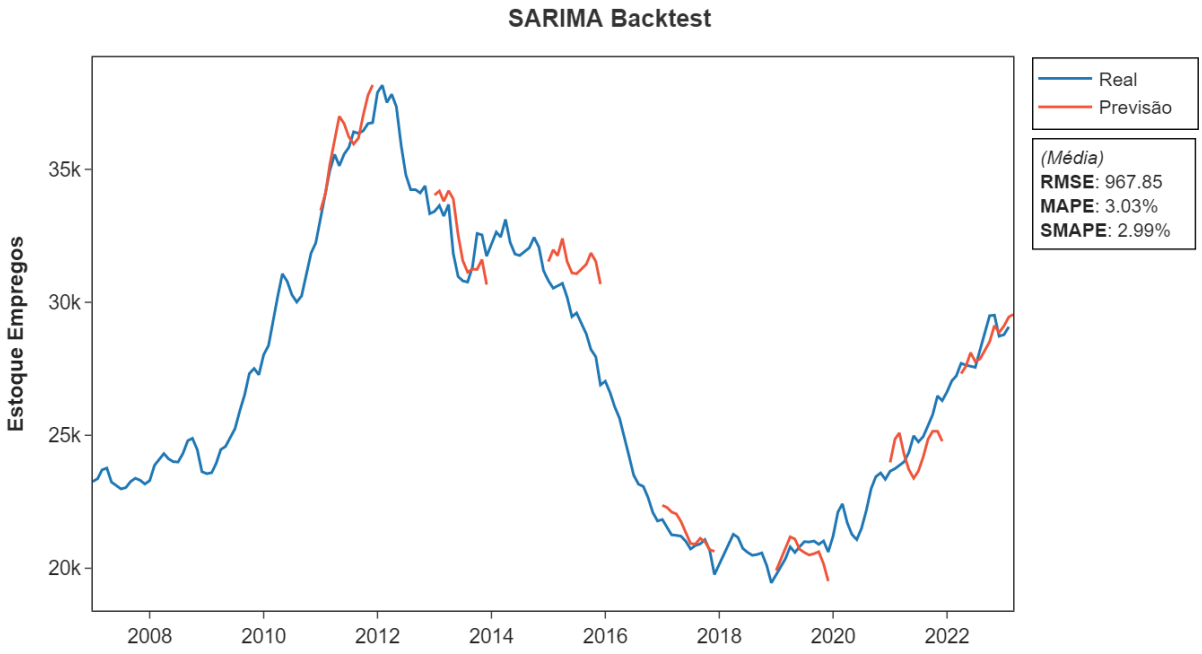

Figura 4.4: Gráfico do resultado da validação cruzada do modelo SARIMA com a media das métricas de erro para a previsão da série temporal de empregos na construção civil em Alagoas.

4.2.3 Suavização Exponencial Holt Winters

Na sequencia, aplicamos e avaliamos o modelo de Suavização Exponencial Holt Winters, tanto na forma aditiva quanto multiplicativa. Este modelo é uma extensão do método de suavização exponencial que leva em consideração tanto a tendência quanto a sazonalidade dos dados.

Os parâmetros α (nível), β (tendência) e γ (sazonalidade)) do modelo Holt Winters foram ajustados por meio de otimização de parâmetros que testou todos os valores de 0 a 1 com incrementos de 0.1 para cada parâmetro.

Depois de testar ambas as formas de suavização exponencial Holt Winters, descobrimos que a forma aditiva produziu resultados superiores para nosso conjunto de dados. O modelo aditivo de Holt-Winters que apresentou o menor RMSE (825.26) usou um α de 0.3, um β de 0.9 e um γ de 0.7, como mostrado na Tabela 4.2.3.

Parâmetros			Métricas de Erro		
α	β	γ	RMSE	MAPE (%)	sMAPE (%)
0.30	0.90	0.70	825.26	2.64	2.62

Tabela 4.3: Desempenho do melhor modelo Holt Winters Aditivo

Os parâmetros deste modelo fornecem informações importantes sobre a série temporal.

O valor de α indica a velocidade de adaptação do modelo às alterações no nível da série, ou seja, seu valor médio. Um α de 0.3 sugere que o modelo dá um peso moderado aos dados mais recentes ao atualizar o nível da série.

O valor de β demonstra a rapidez com que o modelo responde a alterações na tendência da série. Com um valor de β de 0.9, o modelo está altamente sensível a mudanças na tendência, indicando que a tendência dos dados é um fator significativo que o modelo precisa considerar ao fazer previsões.

Finalmente, o γ reflete a importância da sazonalidade na série temporal. O valor relativamente alto de γ (0.7) sugere que a sazonalidade tem um impacto notável na série temporal.

No geral, os valores desses parâmetros indicam que a tendência e a sazonalidade são fatores particularmente importantes nesta série temporal, e que o modelo aditivo de Suavização Exponencial Holt Winters foi bem-sucedido ao capturar essas características.

A análise visual da figura 4.5 indica que o ajuste do modelo Holt Winters Aditivo foi excelente na grande maioria dos pontos de corte, acompanhando muito bem a trajetória dos dados reais, demonstrando a eficácia deste método para esta série temporal.

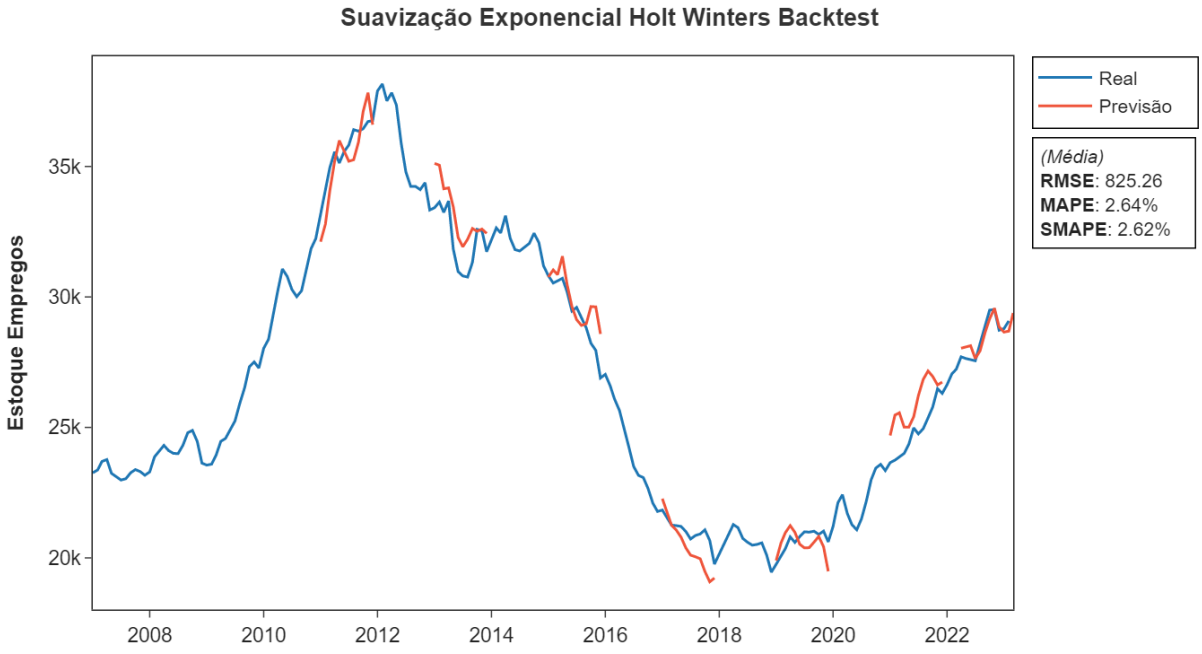

Figura 4.5: Gráfico do resultado da validação cruzada do modelo Holt Winters com a média das métricas de erro para a previsão da série temporal de empregos na construção civil em Alagoas.

4.2.4 TBATS

Por fim, aplicamos e avaliamos o modelo TBATS (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components). Este é um modelo de previsão para

séries temporais que possuem componentes complicados de sazonalidade.

Um aspecto distintivo do modelo TBATS é que ele estima os parâmetros em seu próprio método, portanto, ao contrário dos modelos anteriores, não realizamos otimização de parâmetros neste caso. Para cada iteração da validação cruzada, os parâmetros foram estimados de acordo com a série temporal, resultando em um conjunto de parâmetros para cada iteração da validação cruzada.

O único parâmetro definido foi '*seasona_periods=[12]*', que indica a presença de uma sazonalidade anual na série temporal. O desempenho do modelo TBATS, de acordo com várias métricas de erro, é mostrado na Tabela 4.2.4.

Parâmetros	Métricas de Erro		
Seasonal Periods	RMSE	MAPE (%)	sMAPE (%)
12	1280.15	4.22	4.29

Tabela 4.4: Desempenho do melhor modelo TBATS

A análise visual da série temporal revelou que o modelo teve resultados ruins apenas no final de uma forte tendência, como em dezembro de 2016 e 2018. No restante das iterações, o modelo apresentou bons resultados, indicando que o modelo TBATS foi capaz de capturar adequadamente a estrutura subjacente da série temporal na maioria dos casos.

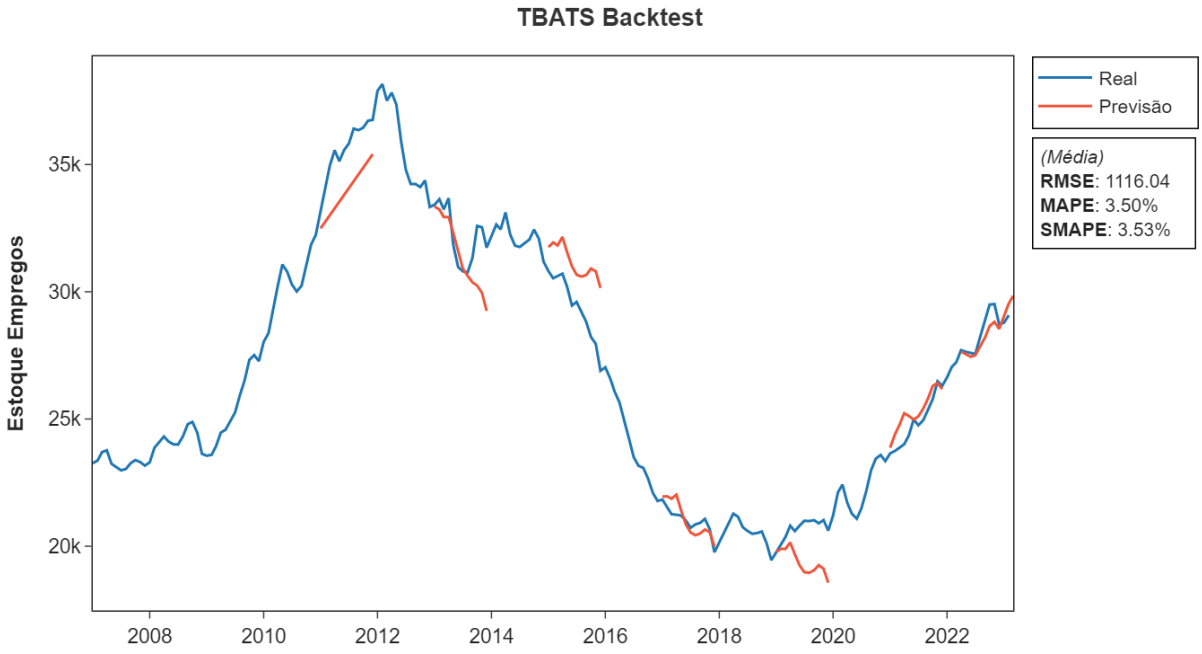

Figura 4.6: Gráfico do resultado da validação cruzada do modelo TBATS com a media das métricas de erro para a previsão da série temporal de empregos na construção civil em Alagoas.

A Figura 4.6 apresenta o gráfico da validação cruzada do modelo TBATS, ilustrando os

resultados da previsão. Pode-se observar que, apesar de alguns desvios em momentos de forte tendência, o modelo conseguiu capturar a sazonalidade e a tendência dos dados de maneira satisfatória na maior parte do tempo.

4.2.5 Prophet

A próxima etapa foi a aplicação e avaliação do modelo Prophet, que é uma abordagem de previsão baseada em modelo para séries temporais univariadas, desenvolvida pelo Facebook. Este modelo é capaz de capturar tendências diárias, semanais e anuais, bem como efeitos de feriados.

Os parâmetros do modelo Prophet de acordo com a otimização de parâmetros . O melhor conjunto de hiperparâmetros encontrado, que resultou na menor RMSE, incluiu um '*changepoint_prior_scale*' de 0.5, um '*seasonality_prior_scale*' de 2.0 e um '*changepoint_range*' de 0.95, conforme mostrado na Tabela 4.2.5.

Parâmetros			Métricas de Erro		
Changepoint P. S.	Seasonality P. S.	Changepoint R.	RMSE	MAPE (%)	sMAPE (%)
0.5	2.0	0.95	1460.75	4.66	4.69

Tabela 4.5: Desempenho do melhor modelo Prophet

O parâmetro '*changepoint_prior_scale*' controla a flexibilidade do modelo em relação às mudanças na tendência. Um valor de 0.5 indica que o modelo é relativamente flexível, permitindo que a tendência mude significativamente. '*Seasonality_prior_scale*' controla a flexibilidade do componente de sazonalidade, sendo que um valor de 2.0 sugere que o modelo tem uma flexibilidade moderada para capturar a sazonalidade. O '*changepoint_range*' controla a proporção da série temporal onde são permitidos pontos de mudança potenciais na tendência. Um valor de 0.95 indica que os pontos de mudança são permitidos em 95% da série temporal.

A análise visual do gráfico da Figura 4.7 mostra que o modelo obteve melhores resultados nos dados mais recentes. Contudo, parece que a importância alta dada à tendência fez com que o modelo errasse bastante na previsão de 12 meses a partir de dezembro de 2010 e dezembro de 2016, onde o modelo assumiu erroneamente que a tendência abrupta de subida ou queda continuaria.

4.3 Comparação entre resultados

A Tabela 4.3 ilustra a comparação entre as métricas de erro de validação cruzada dos melhores conjuntos de hiperparâmetros de cada modelo analisado. As métricas de erro apresentadas são a Raiz do Erro Médio Quadrático (RMSE), a Porcentagem Absoluta Média do Erro (MAPE) e o Erro Percentual Absoluto Simétrico Médio (sMAPE)

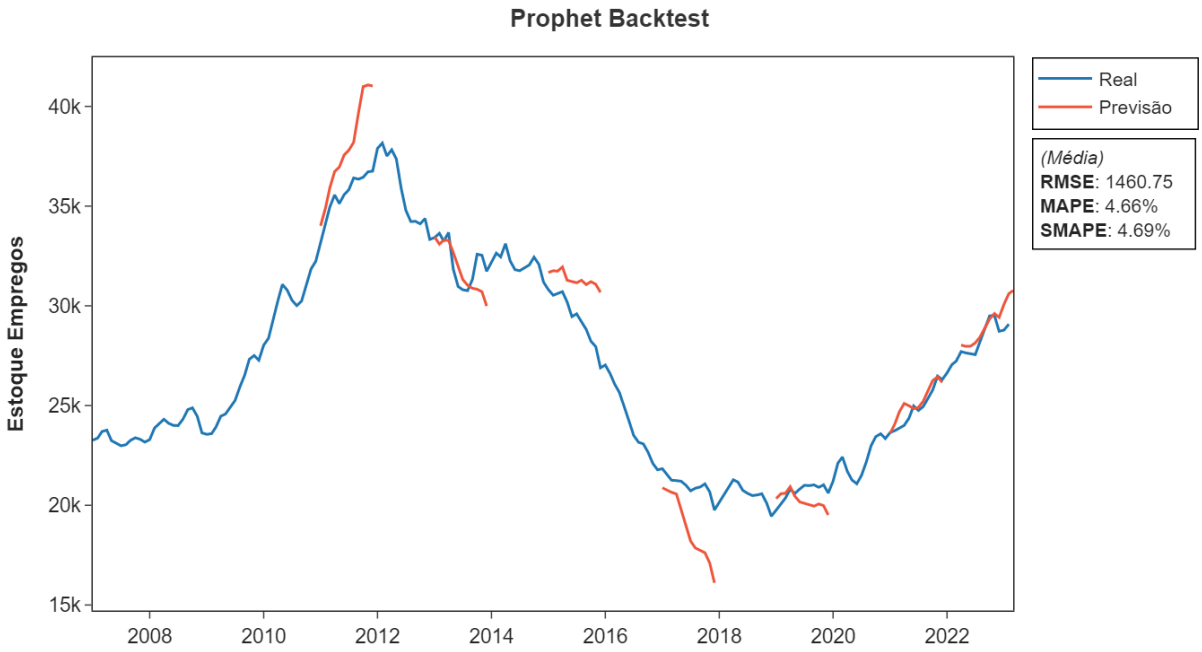

Figura 4.7: Gráfico do resultado da validação cruzada do modelo Prophet com a media das métricas de erro para a previsão da série temporal de empregos na construção civil em Alagoas.

	Métricas de Erro		
Método	RMSE	MAPE (%)	sMAPE (%)
SE Holt Winters	**825.26**	**2.64**	**2.62**
SARIMA	967.85	3.03	2.99
ARIMA	1134.54	3.20	3.25
TBATS	1280.15	4.22	4.29
Prophet	1460.75	4.66	4.69

Tabela 4.6: Comparação das métricas de erro dos modelos

A partir da análise desses resultados, é possível observar que o método de Suavização Exponencial Holt Winters (SE Holt Winters) apresentou os menores valores para a todas as métricas de erro. Com um RMSE de 825.26, MAPE de 2.64% e sMAPE de 2.62%, o SE Holt Winters se destacou em relação aos outros modelos analisados, indicando que este foi o modelo que melhor se adequou aos dados e, consequentemente, obteve a melhor capacidade de previsão.

O SARIMA, embora não tenha apresentado os menores valores de erro, obteve resultados próximos ao do melhor modelo, mostrando-se também como uma alternativa viável para a previsão dos dados analisados.

Já os modelos ARIMA, TBATS e Prophet obtiveram resultados inferiores em comparação com o SE Holt Winters e SARIMA, sendo o modelo Prophet o que apresentou o maior erro em todas as métricas analisadas.

Dessa forma, a partir da comparação dos resultados, pode-se concluir que o método de Suavização Exponencial Holt Winters apresentou o melhor desempenho na previsão dos dados

analisados, conforme demonstrado pelas menores métricas de erro.

4.4 Modelo final e previsões

Com base na análise detalhada das métricas de erro discutida na Seção 4.4, a Suavização Exponencial Holt-Winters foi selecionada como nosso modelo final para previsão. Este modelo demonstrou o melhor desempenho entre os cinco modelos avaliados, alcançando a menor taxa de erro em todas as métricas utilizadas - RMSE, MAPE e sMAPE.

O modelo de Holt-Winters escolhido para as previsões finais usou um valor de α de 0.3, β de 0.9 e γ de 0.7. Estes são os coeficientes de suavização para o nível, tendência, sazonalidade e fator de amortecimento dos dados, respectivamente. Esses parâmetros foram determinados por meio do procedimento de validação cruzada e otimização de hiperparâmetros, que minimizou o erro quadrático médio (RMSE) do modelo, resultando em um RMSE de 825.26.

Com o modelo final configurado, passamos à fase de previsão. A seguir apresentamos as previsões para os próximos 12 meses.

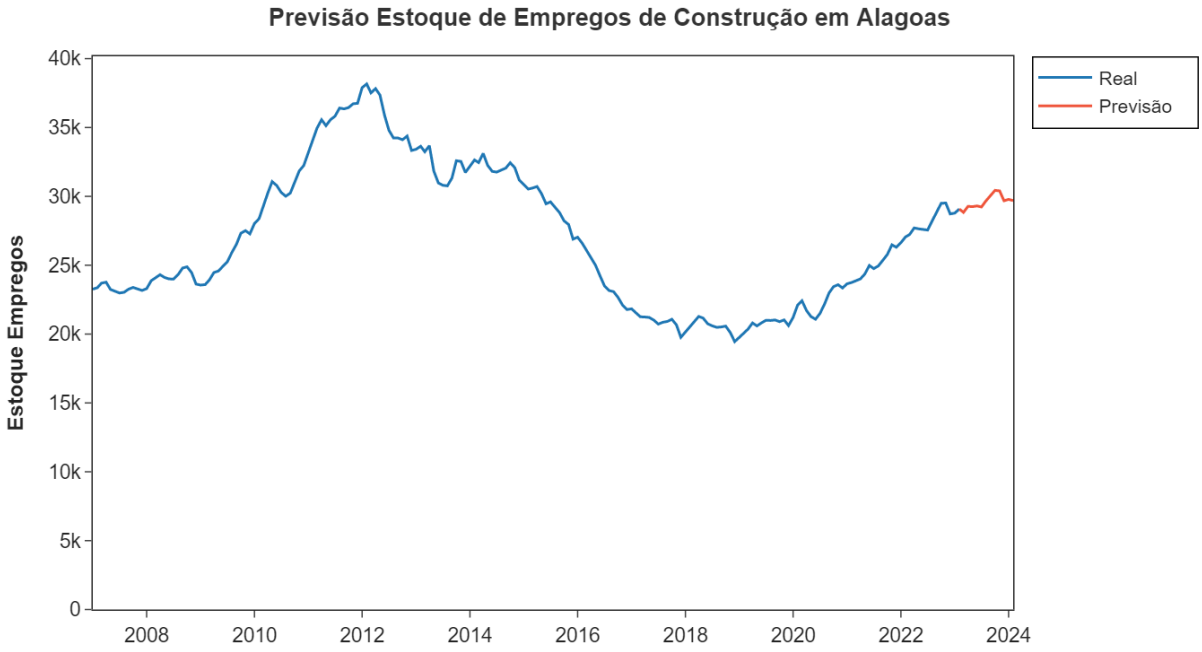

Figura 4.8: Previsão de 12 meses para o estoque de empregos usando a Suavização Exponencial Holt-Winters

Como pode ser observado na figura 4.8, o modelo capturou bem a tendência crescente e a sazonalidade dos dados. No entanto, vale lembrar que, como em qualquer modelo de previsão, há incertezas associadas a essas previsões.

As previsões apresentadas são as melhores estimativas que temos com base nos dados históricos e na configuração atual do modelo. No entanto, mudanças futuras nos padrões subjacentes

que não são refletidas nos dados históricos podem afetar a precisão das previsões.

Este modelo, portanto, pode ser uma ferramenta valiosa para orientar decisões que dependem de uma visão de futuro das métricas em questão. No entanto, é importante reavaliar o desempenho do modelo regularmente e ajustar os parâmetros conforme necessário, especialmente se forem observadas mudanças significativas nas tendências ou na sazonalidade dos dados.

5

Conclusão

Este trabalho dedicou-se à investigação de diversas técnicas de modelagem de séries temporais para a previsão do estoque de empregos no setor da construção civil, com foco em Alagoas. Após uma rigorosa avaliação de desempenho usando modelos de series temporais, a Suavização Exponencial Holt-Winters, um método clássico que tem resistido ao teste do tempo, emergiu como o modelo mais adequado. Surpreendentemente, com uma extensiva otimização de hiperparâmetros e validação cruzada, este modelo foi capaz de superar técnicas amplamente utilizadas na indústria, como o Prophet, devido à sua flexibilidade e capacidade de lidar efetivamente com séries univariadas.

As previsões produzidas por este modelo capturaram com eficácia a tendência e a sazonalidade dos dados. As previsões resultantes deste estudo têm o potencial de servir como uma ferramenta valiosa para a tomada de decisões em diversos setores, desde empresas do setor de construção civil, até instituições de formação técnica e profissional e formuladores de políticas. No entanto, é importante notar que este trabalho se limitou a usar apenas métodos de previsão univariada, focando-se unicamente na própria série temporal para prever o futuro. Esta abordagem, embora eficaz, não incorpora o impacto de possíveis variáveis exógenas, o que pode limitar a precisão das previsões. Além disso, embora modelos mais complexos como TBATS e Prophet sejam mais fáceis de aplicar e inicialmente tenham apresentado resultados promissores, a limitada possibilidade de customização desses modelos resultou em um desempenho inferior ao da Suavização Exponencial Holt-Winters após a calibração de parâmetros.

Para futuras pesquisas, recomendamos a exploração de outras técnicas de modelagem, especialmente aquelas que consideram a inclusão de variáveis exógenas. Essas técnicas têm o potencial de melhorar ainda mais a precisão das previsões e oferecer orientações adicionais sobre a dinâmica do mercado de trabalho. Por fim, este trabalho demonstrou que a Suavização Exponencial Holt-Winters pode ser uma ferramenta eficaz para prever o estoque de empregos utilizando dados públicos no estado de Alagoas. Ainda que haja limitações, as previsões geradas têm potencial para beneficiar uma gama de entidades e instituições.

Referências bibliográficas

[1] Sean J Taylor and Benjamin Letham. Prophet: forecasting at scale. *PeerJ Preprints*, 5:e3190v2, 2017.

[2] Câmara Brasileira da Indústria da Construção. Panorama da construção, dezembro 2022, 2022.

[3] FIRJAN. Construção civil: Desafios 2020, 2020.

[4] CAGED. Cadastro geral de empregados e desempregados - caged, 2023.

[5] Câmara Brasileira da Indústria da Construção CBIC. Impactos jurídicos da covid-19 na construção civil, 2020.

[6] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[7] Relação anual de informações sociais - rais, 2023.

[8] M. Loosemore, A. Dainty, and H. Lingard. *Human Resource Management in Construction Projects: Strategic and Operational Approaches*. Spon Press, London, 2003.

[9] SENAI. Senai: Formação profissional, 2021.

[10] IBGE. Ibge: Indicadores sociais e de desenvolvimento, 2021.

[11] Novo cadastro geral de empregados e desempregados - novo caged, 2023.

[12] Pesquisa nacional por amostra de domicílios - pnad, 2023.

[13] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer, 3 edition, 2016.

[14] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, Melbourne, Australia, 3 edition, 2021.

[15] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 5 edition, 2015.

[16] James Douglas Hamilton. *Time series analysis*, volume 2. Princeton University Press, 1994.

[17] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011.

[18] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*, 2013.

[20] Bob Carpenter, Andrew Gelman, Matt D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[21] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

[22] Spyros Makridakis, Steven C Wheelwright, and Rob J Hyndman. *Forecasting methods and applications*. Wiley New York, 1998.

[23] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

[24] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, pages 1137–1143, 1995.

[25] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4(1):40–79, 2010.

[26] Christoph Bergmeir, Rob J Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018.

[27] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of machine learning research*, 11:2079–2107, 2010.

[28] Wes Mckinney. Data structures for statistical computing in python. In *9th Python in Science Conference*, pages 51–56, 2010.

[29] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[30] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[31] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[33] Plotly Technologies Inc. Plotly: Modern visualization for the data era. https://plotly.com, 2015.

[34] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In Fernando Loizides and Birgit Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90. IOS Press, 2016.

[35] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

[36] J Scott Armstrong and Fred Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, 8(1):69–80, 1992.

[37] Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990.