Trabalho de Conclusão de Curso

Aplicação de Filtro de Kalman para predição do alunado em escolas públicas do Estado de Alagoas e da região Nordeste do Brasil

de **José Ferreira Leite Neto**

orientado por
Prof. Dr. Bruno Almeida Pimentel
Prof. Dr. Ícaro Bezerra Queiroz de Araújo

Universidade Federal de Alagoas
Instituto de Computação
Maceió, Alagoas
05 de Julho de 2024

UNIVERSIDADE FEDERAL DE ALAGOAS
Instituto de Computação

APLICAÇÃO DE FILTRO DE KALMAN PARA PREDIÇÃO DO ALUNADO EM ESCOLAS PÚBLICAS DO ESTADO DE ALAGOAS E DA REGIÃO NORDESTE DO BRASIL

Trabalho de Conclusão de Curso submetido ao Instituto de Computação da Universidade Federal de Alagoas como requisito parcial para a obtenção do grau de Engenheiro de Computação.

José Ferreira Leite Neto

Orientador: Prof. Dr. Bruno Almeida Pimentel
Coorientador: Prof. Dr. Ícaro Bezerra Queiroz de Araújo

Banca Avaliadora:

Dalgoberto Miquilino Pinho Júnior	Prof. Dr., UFAL
Glauber Rodrigues Leite	Prof. Dr., UFAL

Maceió, Alagoas
05 de Julho de 2024

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Girlaine da Silva Santos – CRB-4 – 1127

L533a Leite Neto, José Ferreira.
Aplicação de filtro de kalman para predição do alunado em escolas públicas do Estado de Alagoas e da região nordeste do Brasil / José Ferreira Leite Neto. – 2024.
80 f. : il. color.

Orientador: Bruno Almeida Pimentel.
Coorientador: Ícaro Bezerra Queiroz de Araújo.
Monografia (Trabalho de Conclusão de Curso em Engenharia de Computação) - Universidade Federal de Alagoas. Instituto de Computação, Maceió, 2024.

Bibliografia: f. 68-70.
Apêndices: f.71- 80.

1. Kalman, Filtragem de. 2. Predição de alunado. 3. Programa Nacional do Livro Didático. 4. Analise de séries temporais. 5. Mineração de dados (computação). I. Título.

CDU: 004.62:371.671

Dedicatória

À mulher mais importante e batalhadora da minha vida, que mesmo caminhando com os pés desnudos em brasas, me levou em seu colo e fez com que eu nunca precisasse sequer tirar os sapatos. Que me deu o sopro da existência e inflou minhas asas para que pudesse alçar voos cada vez mais altos e me tornasse quem sou hoje. Agradeço por estar comigo em todos os momentos, tanto nas vitórias quanto nas derrotas, por toda a paciência e amor.

Adriana Gomes Ferreira Leite

Continue a nadar, para achar a solução

— **Dory**

Resumo

Este trabalho investigou a aplicação do Filtro de Kalman para a predição do número de alunos em escolas públicas do estado de Alagoas e da região Nordeste do Brasil. A pesquisa foi motivada pela observação da necessidade de predições precisas do número de alunos para a logística do Programa Nacional do Livro Didático (PNLD). Foram coletados e processados dados do Censo Escolar de 2007 a 2023, e o Filtro de Kalman foi implementado e comparado com baselines como média e regressão linear. Os resultados mostraram que o Filtro de Kalman proporcionou predições mais precisas, adaptando-se melhor às variações nos dados. Como futuras investigações, sugere-se utilizar variações anuais e analisar a evolução de turmas específicas ao longo do tempo para refinar ainda mais a precisão das previsões.

***Palavras-chave*: *Predição do Alunado*; *Programa Nacional do Livro Didático*; *Ciência de Dados*; *Séries temporais*; *Filtro de Kalman*.**

Abstract

This study investigated the application of the Kalman Filter for forecast the number of students in public schools in the state of Alagoas and the Northeast region of Brazil. The research was motivated by the observed need for accurate student number predictions for the logistics of the National Textbook Program (PNLD). Data from the School Census from 2007 to 2023 were collected and processed, and the Kalman Filter was implemented and compared with baselines such as mean and linear regression. The results showed that the Kalman Filter provided more accurate predictions, better adapting to variations in the data. For future investigations, it is suggested to use annual variations and analyze the evolution of specific classes over time to further refine the accuracy of predictions.

***Keywords*: *Student Enrollment Prediction*; *Programa Nacional do Livro Didático*; *Data Science*; *Time Series*; *Kalman Filter*.**

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AM Aprendizado de Máquina

CD Ciência de Dados

ECT Empresa Brasileira de Correios e Telégrafos

FNDE Fundo Nacional de Desenvolvimento da Educação

IA Inteligência Artificial

INEP Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

INL Instituto Nacional do Livro

MAE Mean Absolute Error (Erro Médio Absoluto)

MAPE Mean Absolute Percentage Error (Erro Percentual Absoluto Médio)

MEC Ministério da Educação

MPF Ministério Público Federal

MSE Mean Squared Error (Erro Médio Quadrático)

NEES Núcleo de Excelência em Tecnologias Educacionais

PNBE Programa Nacional Biblioteca da Escola

PNLD Programa Nacional do Livro e do Material Didático

RMSE Root Mean Squared Error (Raiz do Erro Médio Quadrático)

STD Standard Deviation (Desvio Padrão)

TCU Tribunal de Contas da União

TED Termo de Execução Descentralizada

UFAL Universidade Federal de Alagoas

Sumário

Capítulo 1

Introdução

1.1 Motivação

A educação é um dos pilares estruturais no desenvolvimento de qualquer nação, fornecendo subsídios para o progresso social, cultural e econômico [UNESCO, 2015]. A Constituição da República Federativa do Brasil de 1988 estabelece um importante marco ao garantir a educação como direito inalienável de todos, tendo destaque a responsabilidade do Estado ao garantir o acesso universal e igualitário a esta, através de um ensino de qualidade, como se destaca nos seguintes artigos:

> **Art. 6º** São direitos sociais a **educação**, a saúde, a alimentação, o trabalho, a moradia, o transporte, o lazer, a segurança, a previdência social, a proteção à maternidade e à infância, a assistência aos desamparados, na forma desta Constituição. [BRASIL, 1988]

> **Art. 205** A educação, **direito de todos e dever do Estado** e da família, será promovida e incentivada com a colaboração da sociedade, visando ao pleno desenvolvimento da pessoa, seu preparo para o exercício da cidadania e sua qualificação para o trabalho. [BRASIL, 1988]

Esse marco legal reforça o compromisso do país com a promoção da transformação social e inclusão através do acesso à educação. Ao garantir que todos os cidadãos tenham a oportunidade de se educar, o país reconhece a educação como pilar fundamental e alicerce para o desenvolvimento humano e econômico da nação. A educação de qualidade não apenas transmite conhecimento, mas também incentiva o desenvolvimento do pensamento crítico, a capacidade de análise e a habilidade de questionar e compreender o mundo ao seu redor.

Cidadãos bem-educados são capazes de participar ativamente nos processos democráticos, contribuir para debates informados e tomar decisões fundamentadas que afetam a sociedade como um todo. Portanto, ao investir na educação e implementar políticas

inclusivas, o país está construindo uma base sólida para um futuro mais justo, equitativo e próspero. Segundo a [OECD, 2010], a educação desempenha um papel crucial na promoção da coesão social e na formação de cidadãos informados e engajados.

Antes mesmo da promulgação da Constituição Federal de 1988, nascia em 1929 o que hoje é a política pública de maior longevidade da história do Brasil: o **PNLD: Programa Nacional do Livro e do Material Didático**. Criado sob a denominação de INL, Instituto Nacional do Livro, o programa passou por diversas mudanças ao longo dos anos para se tornar o que hoje é um dos maiores e mais bem estruturados programas de distribuição de material didático em todo o mundo, referência absoluta quando se trata do assunto.

Com sua mudança mais recente através do Decreto Nº 9.099, de 18 de Julho de 2017, que uniu o PNLD e o PNBE, Programa Nacional Biblioteca da Escola, em uma só política, o PNLD passou a ter por definição oficial:

> **Art. 1º** O Programa Nacional do Livro e do Material Didático - PNLD, executado no âmbito do Ministério da Educação, será destinado a avaliar e a disponibilizar obras didáticas, pedagógicas e literárias, entre outros materiais de apoio à prática educativa, de forma sistemática, regular e gratuita, às escolas públicas de educação básica das redes federal, estaduais, municipais e distrital e às instituições comunitárias, confessionais ou filantrópicas sem fins lucrativos e conveniadas com o Poder Público. [BRASIL, 2017]

Segundo o [Fundo Nacional de Desenvolvimento da Educação, s d], o PNLD é fundamental para assegurar a equidade no acesso aos materiais didáticos e promover a melhoria da qualidade da educação no Brasil. Ao disponibilizar livros didáticos gratuitos a milhões de alunos, o programa tem um papel crucial na democratização do acesso ao conhecimento, na padronização da qualidade dos materiais educativos e no estímulo à leitura e ao aprendizado autônomo. A importância do PNLD transcende a distribuição de livros, refletindo um esforço contínuo para garantir que todos os estudantes brasileiros, independentemente de sua localização geográfica ou condição socioeconômica, tenham as mesmas oportunidades de aprender e crescer.

Este programa é um exemplo do comprometimento do Brasil com a educação como um direito de todos, e sua evolução ao longo dos anos ilustra o papel dinâmico que as políticas públicas desempenham na adaptação às necessidades educacionais de uma nação em constante mudança. O programa é administrado pelo Ministério da Educação (MEC) e executado em colaboração com o Fundo Nacional de Desenvolvimento da Educação (FNDE).

Em números recentes do PNLD, destacam-se os mais de **194 milhões** de livros distribuídos para o ano letivo de 2024, atendendo cerca de **31 milhões** de alunos nos mais de 5 mil municípios do país e movimentando mais de **R$ 2 bilhões** [Agência Brasil, 2024].

Apesar de todo o sucesso em sua implementação e funcionamento, uma das informações mais relevantes para o PNLD, que é o quantitativo de alunos em cada escola e turma do país, traz alguns desafios. Considerando um país de dimensões continentais como o Brasil, nota-se que a logística de coleta de dados educacionais, tais como o quantitativo de matrículas, é complexa, especialmente em áreas remotas e com acesso limitado a tecnologias e infraestrutura. Esta coleta de dados é feita através do Censo Escolar, que será melhor abordado no item 2.1.2.

A variação sazonal de turmas, as migrações internas e a falta de informações atualizadas são apenas alguns exemplos de fatores que podem dificultar a obtenção de números precisos a respeito da quantidade exata de alunos em cada escola e turma do país. Além disso, a diversidade socioeconômica e regional contribui para disparidades no acesso à educação, complicando ainda mais a tarefa de determinar com precisão o número de livros didáticos e outros recursos educacionais que devem ser adquiridos.

Diante desses desafios, como podemos desenvolver e implementar estratégias eficazes para garantir que o maior número possível de alunos, independentemente de sua localização ou condição socioeconômica, tenha acesso aos materiais educacionais necessários? Como a ciência de dados pode contribuir para esta missão?

1.2 Justificativa

Vencida a etapa de coleta de dados, que por si só já traz seus desafios, como citado, o Censo Escolar é consolidado pelo INEP com as informações de um ano letivo. A partir daí, faz-se necessário mapear a quantidade de alunos em cada escola do país para que a informação possa ser utilizada pelo FNDE para suportar a tomada de decisões em relação ao PNLD.

Ao fazer esse mapeamento, a variabilidade natural do número de alunos nas escolas ao longo dos anos se evidencia. Uma escola pode receber novos alunos ou mesmo ter alguns de seus estudantes mudando de instituição por conta de fatores diversos. Essa quantidade de alunos afeta diretamente a negociação e compra do material didático, que já tem o desafio de ser feita tendo dados defasados em dois anos em relação ao início do período de utilização do material, conforme evidenciado na subseção 2.1.4.

Para ter dados que façam sentido no contexto de compra, se faz necessário analisar e processar os dados existentes, o que pode ser feito manualmente ou de forma automatizada. A análise manual envolve custos, tanto de tempo quanto de dinheiro, além de ser suscetível a erros humanos e de baixa viabilidade devido ao volume de dados. Diante dessa realidade, surgem abordagens mais modernas e eficientes como a Inteligência Artificial (IA), Aprendizado de Máquina (AM) e a Ciência de Dados (CD).

Quando a quantidade de livros comprada é imprecisa, podem ocorrer dois cenários: sobra de livros, o que traz um prejuízo financeiro por conta do desperdício de recursos

públicos, ou falta de livros, que traz um prejuízo pedagógico, uma vez que impacta o processo de aprendizagem e pode afetar de forma permanente os alunos atingidos.

Observando o contexto, o uso da Ciência de Dados emerge como uma opção de grande potencial para auxiliar no enfrentamento da problemática de predição do número de alunos nas escolas públicas de todo o Brasil.

De acordo com [Skiena, 2017], podemos definir Ciência de Dados como a interseção entre Ciência da Computação, Estatística e Domínios de aplicação. Da primeira, vem o poder de fogo e a tecnologia para fazer frente aos grandes volumes de dados e à complexidade das operações; da segunda vem toda uma tradição em análise de dados, testes de hipóteses e visualização. Por fim, do último se originam os desafios que serão enfrentados, além dos padrões de avaliação para saber se os objetivos foram atingidos.

A aplicação de modelos preditivos desempenha um papel fundamental quando se deseja transformar um grande volume de dados em previsões acuradas para o futuro. Diversas técnicas podem ser utilizadas e adaptadas para atender à problemática da predição do alunado, levando em consideração tanto tendências históricas quanto sazonalidade nos dados, o que traz resultados confiáveis mesmo com a variabilidade inerente ao domínio. A aplicação bem-sucedida dessas técnicas pode ajudar a vencer os desafios atuais de determinação do número preciso de alunos em um determinado período-alvo, ou simplesmente, predição do alunado, permitindo a distribuição dos recursos do PNLD de maneira mais eficiente, garantindo a entrega do material didático nas escolas no momento certo e na quantidade adequada, sem desperdício e sem prejuízo pedagógico.

Tendo essas tendências em vista, o próprio FNDE utiliza há alguns anos métodos para análise dos dados, buscando predizer com maior acurácia a quantidade de alunos. Embora tenham sido feitos aprimoramentos no processo, ainda há oportunidades para melhorias adicionais, seja em métodos mais eficientes (em acurácia ou tempo), seja em uma ferramenta para auxiliar a tomada de decisão.

Nessa toada, o presente trabalho investiga novas metodologias para processar os dados do censo escolar e predizer, no planejamento e compra de livros didáticos, a quantidade esperada de alunos para o próximo ano letivo. A metodologia central foi o Filtro de Kalman, uma abordagem inovadora e ainda pouco explorada nesse contexto, em que se pode apontar como exemplos os estudos de [Lazăr and Lazăr, 2015], que utilizou métodos tradicionais, como regressão linear e suavização exponencial, além de [Parkhi et al., 2023], que empregou aprendizado de máquina, exemplificado por árvores de decisão.

O Filtro de Kalman, melhor definido na Seção 2.3, é um algoritmo recursivo que utiliza uma série de medições ao longo do tempo, contendo ruído estatístico e outras imprecisões, para produzir estimativas mais precisas dos estados desconhecidos de um sistema dinâmico. Sua capacidade de lidar com incertezas e ajustar previsões com base em novas informações torna-o especialmente adequado para cenários de predição com dados temporais variáveis, como o caso das matrículas escolares.

1.3 Formulação de Hipóteses

Este trabalho se propõe a investigar a aplicação do Filtro de Kalman na predição do número de alunos em escolas públicas, com foco no estado de Alagoas. A motivação para essa pesquisa é baseada nos desafios logísticos e na necessidade de precisão na predição do alunado para programas como o PNLD. Diante disso, foram formuladas as seguintes hipóteses:

- **Hipótese 1:** A aplicação do Filtro de Kalman pode melhorar a acurácia das predições de matrículas escolares em comparação com métodos tradicionais, como média ou regressão linear.
- **Hipótese 2:** O método de predição utilizando Filtro de Kalman pode ser adaptado para outras regiões além do estado de Alagoas.

1.4 Objetivos

O objetivo geral do trabalho foi: **Explorar o potencial do Filtro de Kalman como um estimador linear estatístico para predizer o número de alunos para anos letivos futuros em escolas públicas do estado de Alagoas e da Região Nordeste do Brasil**.

Dentre os objetivos específicos, podem ser destacados:

- Coletar dados do censo escolar do ano de 2023 referentes às escolas públicas do estado de Alagoas e da região Nordeste;
- Pré-processar os dados de modo a manter apenas informações relacionadas às escolas e turmas alvo de pesquisa, seguindo os critérios estabelecidos;
- Fazer uma análise estatística dos dados das escolas e turmas que foram mantidas;
- Determinar baselines que servirão como base de comparação para os resultados obtidos na predição do alunado, incluindo a análise de seus erros em relação aos dados reais de matrículas no ano predito;
- Utilizar a abordagem do Filtro de Kalman para fazer a predição do alunado esperado para o próximo ano letivo e para o posterior (2 anos à frente) baseando-se em uma série temporal contendo a quantidade de alunos dos anos anteriores, com o objetivo de aprimorar as estimativas de matrículas futuras.
- Avaliar a performance do Filtro de Kalman comparando os resultados obtidos com as baselines determinadas, utilizando métricas de avaliação coerentes com essa análise.

1.5 Estrutura do trabalho

Este trabalho está organizado em cinco capítulos principais, cada um dedicado a uma etapa para compreensão da pesquisa realizada e apresentação dos resultados obtidos. A seguir, estão detalhadas a estrutura e o conteúdo de cada capítulo:

1. **Introdução**: Apresenta um panorama geral do Programa Nacional do Livro e do Material Didático, com enfoque especial na predição do alunado e sua relevância no contexto da política pública. Esta seção é dividida em subseções que discutem motivação, justificativa, objetivos e hipóteses do trabalho, além desta descrição da estrutura;

2. **Fundamentação Teórica**: Expõe os conceitos teóricos fundamentais que suportam o estudo, relacionados ao Programa Nacional do Livro Didático, séries temporais, Filtro de Kalman e análise de desempenho;

3. **Metodologia**: Aborda em detalhes a metodologia adotada para a análise estatística dos dados, determinação de baselines e implementação do Filtro de Kalman para a predição do alunado;

4. **Resultados**: Apresenta os resultados obtidos a partir da aplicação da metodologia proposta, incluindo a análise estatística e a comparação com os baselines pr é-definidos, com a finalidade de determinar a eficácia do Filtro de Kalman na predição do número de alunos;

5. **Conclusão**: Reflete sobre as descobertas do estudo, avaliando o desempenho do Filtro de Kalman e sua aplicabilidade no contexto das escolas públicas de Alagoas e do Nordeste. Aborda as limitações do estudo e sugere direções para pesquisas futuras que possam expandir ou aprofundar o entendimento do tema.

Capítulo 2

Fundamentação Teórica

2.1 Programa Nacional do Livro Didático

O Programa Nacional do Livro Didático (PNLD), como já citado, é uma iniciativa do Ministério da Educação do Brasil (MEC) que tem como objetivo fornecer livros didáticos e outros materiais pedagógicos aos alunos da educação básica nas escolas públicas. Atende diversas etapas de ensino, a saber [Abrelivros, s d]:

- Educação infantil.
- Anos iniciais do ensino fundamental (1º ao 5º ano).
- Anos finais do ensino fundamental (6º ao 9º ano).
- Ensino médio.
- Educação de Jovens e Adultos (EJA).

A seguir, são abordados apectos técnicos do programa e relacionados ao processo de escolha e aquisição do material didático e outros conceitos que dão suporte à compreensão deste trabalho.

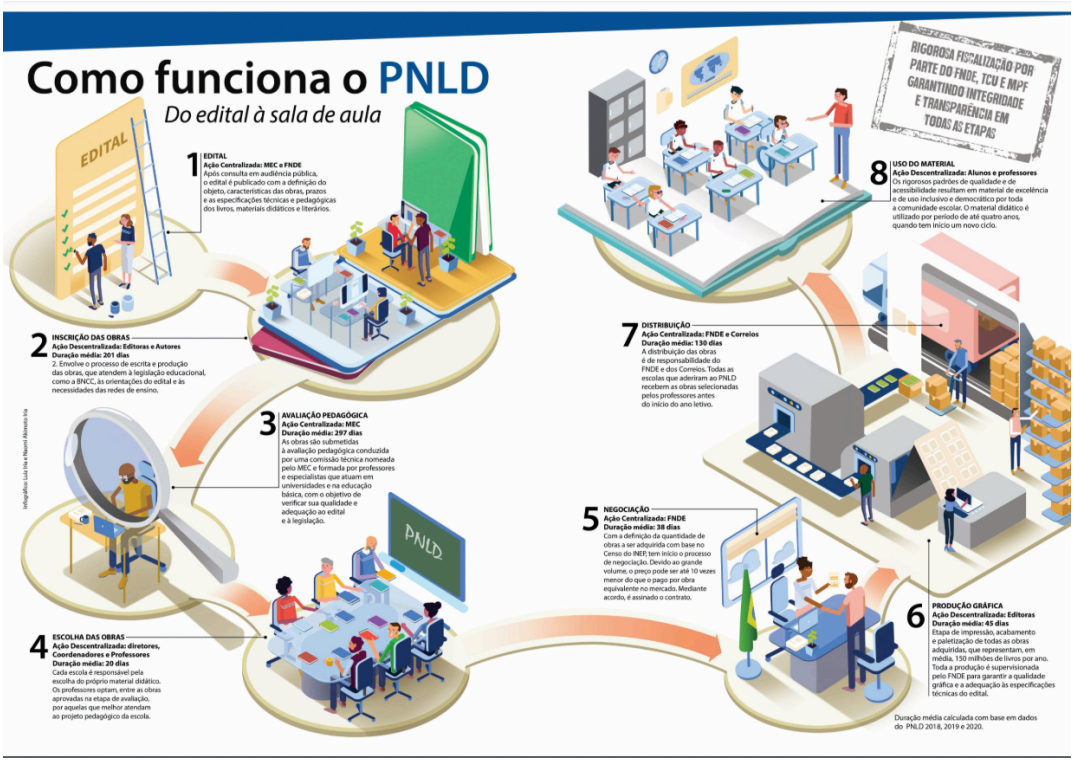
2.1.1 Processo de Escolha e Aquisição do Material Didático

O PNLD possui uma estrutura complexa para o processo de seleção e distribuição de obras, que envolve diversas etapas técnicas que vão da seleção dos livros até a distribuição nas escolas. As obras passam por diversas etapas de avaliação em que são avaliadas por especialistas em educação e, uma vez aprovadas, são incluídas no catálogo do programa. Cada parte do processo é acompanhada e fiscalizada de perto por órgãos como o Fundo Nacional de Desenvolvimento da Educação (FNDE), o Tribunal de Contas da União (TCU) e o Ministério Público Federal (MPF).

A seleção dos livros é baseada em critérios rigorosos que consideram a qualidade pedagógica, a adequação ao currículo nacional e a conformidade com os princípios éticos e de direitos humanos. Além disso, o programa também prevê a inclusão de livros em formato acessível para alunos com deficiência visual e auditiva.

O infográfico da Figura 2.1 [Abrelivros, 2020] apresenta as diversas etapas que acontecem em um processo de seleção de material didático:

Figura 2.1: Fluxo de seleção do PNLD.



Fonte: Abrelivros.

Tudo começa com o lançamento do edital, onde são definidas as especificações técnicas e pedagógicas que cada obra deve atender para participar do certame, além dos prazos de cada etapa. Em seguida, as obras são produzidas de forma a atender às diretrizes propostas e inscritas para seleção por parte das editoras e autores. Desde a etapa de inscrição, existe uma triagem das obras inscritas, que só seguem no processo após passar por uma minuciosa avaliação por analistas certificados.

Após a inscrição, as obras passam pelo processo de avaliação pedagógica, onde uma comissão técnica nomeada pelo Ministério da Educação e formada por professores e especialistas nas mais diversas áreas do conhecimento se certifica de que cada página do material atende aos requisitos e ao edital.

Em seguida, concluída a avaliação pedagógica, cada escola tem autonomia para escolher, dentre uma lista de obras aprovadas, qual atende melhor às suas necessidades e ao seu projeto pedagógico. Esse é um processo que conta com a participação de diretores,

coordenadores e professores em cada unidade de ensino.

Sabendo quais obras cada escola escolheu, segue-se o processo de negociação e compra por parte do FNDE com as editoras e autores. Devido às altíssimas quantidades, o valor de um livro pode ser até dez vezes menor que o praticado no mercado, garantindo economia aos cofres públicos. Na definição dessas quantidades, é fundamental saber com precisão quantos alunos serão atendidos em cada escola e etapa de ensino, a problemática central estudada no presente trabalho.

Por fim, com os contratos assinados, as editoras procedem com a produção gráfica das obras adquiridas, com supervisão do FNDE. Após a produção, as obras são distribuídas em parceria com a Empresa Brasileira de Correios e Telégrafos (ECT) e finalmente podem ser utilizadas pelos alunos e professores.

2.1.2 Censo Escolar

O Censo Escolar é a principal pesquisa estatística sobre a educação básica no Brasil, realizada anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) em parceria com secretarias de educação, dele participam tanto escolas públicas quanto privadas. Este levantamento é fundamental para a formulação, monitoramento e avaliação das políticas públicas educacionais no país [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, s d].

Os dados coletados pela pesquisa falam sobre as escolas, turmas, alunos, docentes e outros profissionais da educação. As informações coletadas incluem infraestrutura das escolas, recursos pedagógicos, rendimento e movimento escolar, entre outros. Esses dados são essenciais para:

- Planejamento e gestão educacional em nível municipal, estadual e federal;
- Definição de políticas públicas e programas educacionais, como o Programa Nacional do Livro Didático (PNLD);
- Monitoramento da evolução dos indicadores educacionais e identificação de desigualdades e necessidades específicas;
- Distribuição de recursos financeiros

2.1.3 Predição do Alunado

Os dados do Censo Escolar são a base para a predição do alunado em escolas públicas, objeto de estudo deste trabalho, uma vez que fornecem uma base histórica anual e detalhada da quantidade de alunos em cada turma de cada escola do país que registrou dados no Censo Escolar. A aplicação de métodos de previsão, como o Filtro de Kalman, pode

utilizar essas informações como base para gerar estimativas mais precisas do número de alunos em um horizonte futuro.

2.1.4 A necessidade de um modelo preditivo

Devido às dimensões do país e à abrangência do PNLD, a logística de compra e distribuição dos livros didáticos é uma operação complexa e que requer informações que apoiem a tomada de decisão, dentre elas o número de alunos que serão atendidos, para que seja determinada a quantidade a ser comprada.

A predição do alunado é fundamental para garantir que todos os alunos recebam seus materiais no início do ano letivo, uma vez que existe uma defasagem de dois anos entre o momento representado pelo dado real, que é fornecido pelo Censo Escolar, e o momento em que os livros serão utilizados, uma vez que, por conta da logística envolvida, a compra ocorre cerca de um ano antes do início do próximo período letivo.

Por exemplo: no início do ano de 2024, é efetivada a compra de material didático que será usado em 2025. Contudo, a informação mais recente é do Censo Escolar de 2023, de forma que se faz necessário um método para preencher esse espaço e determinar, da forma mais precisa possível, a quantidade de alunos que deverá receber o material no ano de 2025.

2.1.5 Terminologia

Ao longo do texto, alguns conceitos específicos do domínio são mencionados, e a compreensão destes conceitos é fundamental para a perfeita compreensão do trabalho. São eles:

- **Entidade**: Uma entidade consiste em uma escola, que pode atender várias etapas de ensino e registra a quantidade de alunos para cada etapa no censo escolar;
- **Etapa de Ensino**: Para além das etapas de ensino já mencionadas (Educação infantil, Ensino médio, entre outros), ao analisar os dados, nota-se a existência de uma etapa de ensino para cada ano/série de ensino. Isso significa que o primeiro ano e o segundo ano do ensino fundamental tem a sua quantidade de alunos em uma determinada escola registrada de maneira individualizada, sob um código específico de etapa de ensino, mas pertencendo à mesma entidade (nesse casso);
- **Turma**: Ao longo deste texto, o conceito de turma será tratado de forma mais abrangente: cada turma é uma **etapa de ensino** dentro de uma **entidade**, ou seja, todas as turmas do quinto ano do ensino fundamental da escola ”A”correspondem a apenas uma quantidade de alunos a cada ano dentro dos dados trabalhados. Esta abstração simplifica as operações, pois o aluno que pertence a uma turma ou outra dentro da mesma etapa de ensino não interfere na predição realizada.

2.2 Séries Temporais

De acordo com [Morettin, 2006], séries temporais se referem a qualquer conjunto de observações coletadas de maneira sequencial ao longo do tempo. Normalmente, essas observações são espaçadas de maneira uniforme, como segundos, minutos, horas, dias, meses ou anos. A análise de séries temporais envolve métodos para analisar dados temporais a fim de extrair características importantes e fazer previsões futuras.

Diversos campos do conhecimento se beneficiam do conhecimento da análise de séries temporais, como economia, meteorologia, medicina, engenharia e outros. A principal característica das séries temporais, como o próprio nome indica, é a dependência temporal, ou seja, os dados coletados em um determinado tempo t são, em geral, correlacionados com os dados de tempos anteriores. Esta correlação temporal pode ser explorada para modelar o comportamento da série e inferir valores futuros.

A maioria dos métodos de previsão se ampara na ideia de que valores passados são referenciais confiáveis de informação sobre o comportamento futuro de uma série temporal. O principal objetivo desses métodos é identificar padrões importantes entre o ruído presente nas observações, a fim de utilizar tais padrões na previsão dos valores futuros da série. Uma ampla variedade de modelos de previsão busca abstrair as flutuações, empregando técnicas de suavização como média móvel e suavização exponencial. Estas técnicas partem do pressuposto de que os valores extremos na série são aleatórios e que, ao suavizar esses extremos, se identifica o padrão real da série temporal [Morettin, 2006].

Além da previsão, a análise de séries temporais pode ser usada para detectar padrões sazonais, identificar mudanças estruturais, filtrar ruídos e suavizar séries de dados.

De acordo com [Box et al., 2015], a análise de séries temporais é essencial para compreender o comportamento dos sistemas dinâmicos e para o desenvolvimento de modelos de previsão robustos que podem ser aplicados em diversas áreas do conhecimento humano. Esses modelos ajudam na tomada de decisões, planejamento e controle de processos.

2.2.1 Regressão Linear

A regressão linear uma técnica amplamente utilizada para modelar séries temporais, especialmente quando se busca entender a relação entre variáveis independentes e dependentes. Na análise de séries temporais, a regressão linear pode ser usada para modelar a tendência linear nos dados. A fórmula básica da regressão linear pode ser definida segundo [Morettin, 2006] de acordo com a Equação 2.1:

$$Z_t = \alpha + \beta t + \alpha_t, \tag{2.1}$$

onde:

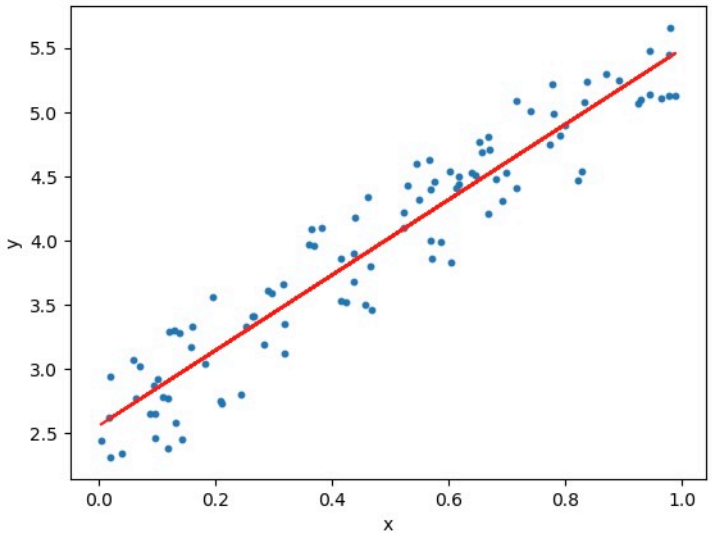
- Z_t é a variável dependente (valor da série temporal no tempo t),

- t é a variável independente (tempo t ou outra variável explicativa, t = 1, ..., N),
- α é o intercepto,
- β é o coeficiente de inclinação,
- α_t é o termo de erro.

A regressão linear pode ser estendida para incluir múltiplas variáveis explicativas ou para modelar componentes não lineares através de transformações e termos polinomiais.

Na figura 2.2 [Rathore, 2023], é possível observar um exemplo de regressão linear aplicado a um conjunto qualquer de dados:

Figura 2.2: Exemplo de reta de regressão linear.



Fonte: Medium.

2.2.2 Suavização Exponencial

A suavização exponencial é uma técnica comum para modelar séries temporais que lida melhor com a presença de ruído nos dados do que a regressão linear, fornecendo previsões baseadas em valores passados. Existem várias variantes da suavização exponencial, mas considerando a suavização exponencial simples, usada quando os dados não apresentam tendência ou sazonalidade significativas, temos, de acordo com a Equação 2.2 [Morettin, 2006]:

$$\hat{Z}_t(h) = \hat{Z}_{t-1}(h+1) + \frac{Z_t - Z_{t-r}}{r}, \quad \forall h > 0. \tag{2.2}$$

onde:

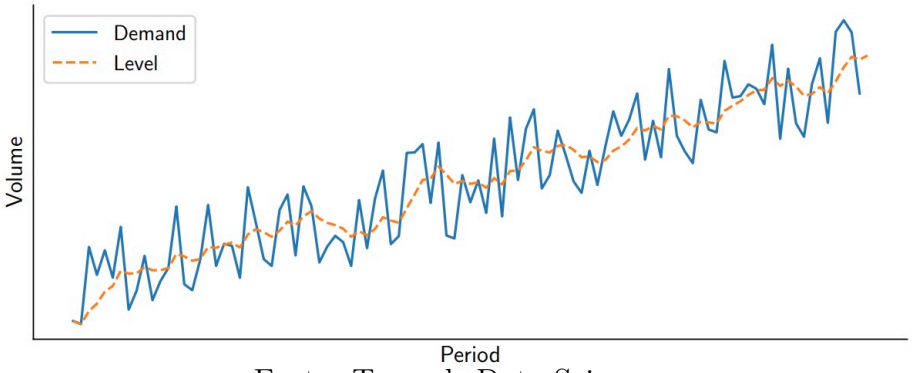
- $\hat{Z}_t(h)$ é a previsão da série temporal no tempo t para um horizonte h,

- $\hat{Z}_{t-1}(h+1)$ é a previsão anterior da série temporal no tempo $t-1$ para o horizonte $h+1$,
- Z_t é o valor observado da série temporal no tempo t,
- Z_{t-r} é o valor observado da série temporal no tempo $t-r$,
- r é um parâmetro que determina o intervalo de tempo considerado para a correção da previsão.

A equação representa a política de atualização de previsão. Essa fórmula ajusta a previsão com base em novas observações, tornando-a mais precisa à medida que novos dados se tornam disponíveis. Cada novo valor observado corrige a estimativa anterior, levando em conta a diferença entre o valor atual e um valor passado, normalizados por r.

Na figura 2.3 [Vandeput, 2019], é possível observar um exemplo de suavização exponencial aplicado ao nível ('level') de um conjunto qualquer de dados:

Figura 2.3: Exemplo de curva de suavização exponencial.



Fonte: Towards Data Science.

2.3 Filtro de Kalman

O Filtro de Kalman é um algoritmo recursivo que fornece estimativas de estados de sistemas dinâmicos lineares a partir de medições ruidosas. Desenvolvido por Rudolf E. Kálmán em 1960, este filtro é amplamente utilizado em controle e sistemas de estimativa, navegação e tracking, processamento de sinais e econometria [Kalman, 1960].

O Filtro de Kalman foi projetado para operar em um processo de dois passos: *previsão* e *atualização* [Welch and Bishop, 1995]:

- Durante a fase de previsão, o filtro usa o modelo do sistema para prever o estado atual e sua incerteza.

- Na fase de atualização, o filtro ajusta essa previsão com base em novas medições. O algoritmo é eficiente computacionalmente, pois utiliza apenas os estados anteriores e a nova medição para calcular o novo estado estimado [Welch and Bishop, 1995].

O Filtro de Kalman assume que o sistema pode ser descrito por um modelo linear com ruído Gaussiano. A Equação 2.3 representa o modelo de estado [Welch and Bishop, 1995]:

$$x_k = Ax_{k-1} + Bu_k + w_k, \tag{2.3}$$

onde:

- x_k é o vetor de estado no instante k,
- A é a matriz de transição de estado,
- B é a matriz de controle (se houver controle no sistema),
- u_k é o vetor de controle no instante k,
- w_k é o ruído do processo, assumido como Gaussiano com média zero ($E\{w_k\} = 0$) e covariância Q ($E\{w_k^T w_k\} = Q$).

A equação de observação é representada por 2.4 [Welch and Bishop, 1995]:

$$z_k = Hx_k + v_k, \tag{2.4}$$

onde:

- z_k é o vetor de medição no instante k,
- H é a matriz de observação,
- v_k é o ruído da medição, assumido como Gaussiano com média zero ($E\{v_k\} = 0$) e covariância R ($E\{v_k^T v_k\} = R$),
- v_k e w_k (ruído do processo) são assumidos independentes, ou seja, $E\{w_k^T v_k\} = 0$.

2.3.1 Filtro de Kalman Simples

O filtro de Kalman simples se comporta da seguinte forma nas fases de *previsão* e *atualização*:

Fase de Previsão

Na fase de previsão, o filtro estima o estado a priori e a covariância de erro a priori de acordo com as Equações 2.5 e 2.6:

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} + Bu_k, \tag{2.5}$$

$$P_{k|k-1} = AP_{k-1|k-1}A^T + Q, \tag{2.6}$$

onde:

- $\hat{x}_{k|k-1}$ é a estimativa a priori do estado no instante k,
- $P_{k|k-1}$ é a covariância do erro de previsão,
- $P_{k-1|k-1}$ é a covariância do erro de estimativa do estado anterior.

Fase de Atualização

Na fase de atualização, a estimativa do estado e a covariância do erro são corrigidas com base na nova medição, de acordo com as Equações 2.7, 2.8 e 2.9:

$$K_k = P_{k|k-1}H^T(HP_{k|k-1}H^T + R)^{-1}, \tag{2.7}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(z_k - H\hat{x}_{k|k-1}), \tag{2.8}$$

$$P_{k|k} = (I - K_kH)P_{k|k-1}, \tag{2.9}$$

onde:

- K_k é o ganho de Kalman, calculado para minimizar a covariância do erro de estimativa. Ele é considerado **ótimo**, pois utiliza as características estatísticas dos ruídos de processo w_k e medição v_k, com base no critério de mínimos quadrados.
- $\hat{x}_{k|k}$ é a estimativa **a posteriori** do estado no instante k,
- $P_{k|k}$ é a covariância do erro de estimativa **a posteriori**.z
- $(z_k - H\hat{x}_{k|k-1})$ é denominado **erro de inovação**, que representa a discrepância entre a medição real e a predição feita com base no modelo do sistema. Esse termo é essencial na atualização da estimativa.

2.3.2 Matrizes Envolvidas

As matrizes fundamentais no Filtro de Kalman são:

- A: Matriz de transição de estado, que descreve a dinâmica do sistema.

- B: Matriz de controle, que relaciona o vetor de controle u_k com o estado do sistema (se houver controle).
- H: Matriz de observação, que relaciona o estado do sistema com as medições.
- Q: Matriz de covariância do ruído do processo, que captura a incerteza no modelo do sistema.
- R: Matriz de covariância do ruído de medição, que captura a incerteza nas medições.
- $P_{k|k-1}$: Matriz de covariância do erro de previsão, que captura a incerteza na previsão do estado.
- $P_{k|k}$: Matriz de covariância do erro de estimativa, que captura a incerteza na estimativa corrigida do estado.

2.3.3 Algoritmo de Estimação EM

O algoritmo Expectation-Maximization (EM) é uma abordagem iterativa usada para estimar os parâmetros do modelo em situações onde os dados estão incompletos ou possuem ruído. No contexto do Filtro de Kalman, o algoritmo EM pode ser utilizado para estimar as matrizes de covariância Q e R.

O algoritmo EM consiste em duas etapas principais [Dempster et al., 1977]:

Etapa de Expectation (E)

Na etapa de Expectation, calcula-se a expectativa do logaritmo da função de verossimilhança completa dos dados, usando as estimativas atuais dos parâmetros.

Etapa de Maximization (M)

Na etapa de Maximization, encontram-se os parâmetros que maximizam a expectativa calculada na etapa anterior.

O algoritmo alterna entre essas duas etapas até que a convergência seja atingida, ou seja, até que as mudanças nos parâmetros estimados sejam pequenas o suficiente.

2.4 Estimação, Predição e Forecasting

No contexto da análise de séries temporais e modelagem estatística, os termos ”estimação”, ”predição”e ”forecasting”são frequentemente usados de maneira intercambiável. No entanto, cada termo possui um significado específico e é importante distinguir entre eles para aplicar corretamente os métodos e técnicas em análises. Esta seção explora as diferenças e inter-relações entre estes conceitos.

2.4.1 Estimação

A estimação refere-se ao processo de inferir os valores dos parâmetros de um modelo com base em dados observados. No contexto estatístico, isso geralmente envolve a utilização de métodos como Máxima Verossimilhança (MLE) ou Mínimos Quadrados (OLS) para calcular os parâmetros que melhor explicam os dados observados.

Definição: Estimação é o processo de inferir ou calcular os valores dos parâmetros desconhecidos de um modelo estatístico a partir de dados observados [Greene, 2012].

2.4.2 Predição

Predição é o processo de usar um modelo estatístico preparado para prever os valores de uma variável dependente com base em novas observações das variáveis independentes. Este processo se concentra em prever valores dentro do intervalo dos dados observados.

Definição: Predição é o processo de usar um modelo estatístico ou de machine learning para prever valores de uma variável dependente com base em novas observações das variáveis independentes [Hastie et al., 2009].

2.4.3 Forecasting

Forecasting, ou previsão, refere-se ao processo de prever valores futuros de uma série temporal com base em observações passadas. Ao contrário da predição, que pode ser aplicada a qualquer tipo de dado, forecasting é especificamente utilizado em dados temporais.

Definição: Forecasting é o processo de usar modelos estatísticos ou de machine learning para prever valores futuros de uma série temporal com base em observações passadas [Hyndman and Athanasopoulos, 2018].

2.4.4 Diferenças e Inter-relações

Apesar das semelhanças, há distinções importantes entre estimação, predição e forecasting:

- **Contexto**: A estimação é geralmente focada em inferir parâmetros do modelo com base em dados observados. Predição e Forecasting têm como finalidade para prever valores futuros em uma série temporal.
- **Método**: A estimação utiliza métodos estatísticos. A predição pode usar modelos estatísticos ou de machine learning. O forecasting é especializado em modelos de séries temporais.
- **Objetivo**: O objetivo da estimação é determinar os melhores parâmetros do modelo. O objetivo da predição é prever valores futuros com base em variáveis independentes.

O objetivo do forecasting é prever valores futuros de uma série temporal com base em dados passados.

Estas distinções são cruciais para a correta aplicação dos métodos e para a interpretação adequada dos resultados em análises estatísticas e de séries temporais.

2.5 Validação e Análise de Desempenho

A validação e análise de desempenho são etapas cruciais na avaliação de modelos de previsão de séries temporais. Para esta tarefa, pode-se lançar mão de várias métricas para quantificar a precisão e a eficácia dos modelos. A seguir estão detalhadas as métricas de avaliação utilizadas para este trabalho, além do Teste de Wilcoxon, utilizado como teste para validação estatística dos resultados encontrados.

2.5.1 Métricas de Avaliação

As métricas de avaliação utilizadas para comparar a performance entre diferentes modelos de previsão neste trabalho foram: MAE (Mean Absolute Error), STD (Standard Deviation), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R^2 (R-Squared -Coeficiente de Determinação) e MAPE (Mean Absolute Percentage Error).

Detalhando e caracterizando as equações relacionadas a cada uma das métricas, temos:

MAE (Mean Absolute Error)

O MAE, em português: Erro Absoluto Médio, é uma medida da precisão de um modelo de previsão, calculada como a média das diferenças absolutas entre os valores previstos e os valores observados. Podemos calcular utilizando a Equação 2.10 [Hyndman and Athanasopoulos, 2018]:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.10}$$

onde:

- n é o número total de observações,
- y_i é o valor observado,
- $\hat{y}_i$ é o valor previsto.

A principal vantagem do MAE é que ele fornece uma medida simples e intuitiva da precisão da previsão: quanto menor, menor é o erro. No entanto, não penaliza grandes erros tanto quanto outras métricas como o MSE ou o RMSE.

STD (Standard Deviation)

O STD, em português: Desvio Padrão, é uma medida que quantifica a quantidade de variação ou dispersão dos dados. No contexto de previsões, é útil para avaliar a consistência dos erros de previsão. O STD é definido como a raiz quadrada da variância, que é a média dos quadrados das diferenças entre cada valor e a média dos valores. Sua equação é a 2.11 [Montgomery et al., 2015]:

$$\text{STD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2.11}$$

onde:

- n é o número total de observações,
- y_i é o valor observado,
- $\bar{y}$ é a média dos valores observados.

Quanto maior o desvio padrão, maior é a distância média entre o dado previsto e o dado real, indicando que os valores de erro são mais expressivos. O STD é uma ótima métrica para verificar se o modelo tem um desempenho consistente.

MSE (Mean Squared Error)

O MSE, Erro Médio Quadrado em português, é uma métrica que calcula a média dos quadrados dos erros (diferenças entre os valores previstos e observados). Sua equação, muito semelhante à do MAE, é dada pela Equação 2.12 [Hyndman and Athanasopoulos, 2018]:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.12}$$

O MSE é particularmente útil porque penaliza erros grandes mais do que erros pequenos, o que pode ser importante dependendo do contexto da previsão. No entanto, essa métrica, bem como todas as que envolvem médias, tem maior sensibilidade a outliers.

RMSE (Root Mean Squared Error)

O RMSE, Raiz do Erro Médio Quadrado, como o próprio nome sugere, consiste na raiz quadrada do MSE, sendo uma medida de precisão que fornece uma interpretação mais intuitiva dos erros, uma vez que está na mesma escala que os valores observados. Sua equação é a do MSE com a operação de raiz quadrada, como mostrado em 2.13 [Hyndman and Athanasopoulos, 2018]:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{2.13}$$

O RMSE busca conciliar as vantagens do MAE e do MSE, de maneira que a métrica é sensível a grandes erros, mas tem uma interpretação mais direta por manter a escala dos dados.

R^2 (Coeficiente de Determinação)

O R^2, coeficiente de determinação, mede a proporção da variância total dos valores observados que é explicada pelos valores previstos. Sua equação é dada por 2.14 [Montgomery et al., 2015]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2.14}$$

onde:

- y_i é o valor observado,
- $\hat{y}_i$ é o valor previsto,
- $\bar{y}$ é a média dos valores observados.

O R^2 varia entre 0 e 1, de maneira que valores próximos de 1 significam que a maior parte da variabilidade dos dados observados é explicada pelo modelo de previsão utilizado. Valores menores indicam um ajuste ruim, ou seja, uma representação imprecisa da variabilidade dos dados.

MAPE (Mean Absolute Percentage Error)

O MAPE, em português: Erro Médio Percetual Absoluto, é uma métrica que calcula a média das diferenças absolutas entre os valores previstos e observados, expressas como uma porcentagem dos valores observados. A equação para sua determinação é a 2.15 [Hyndman and Athanasopoulos, 2018]:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%. \tag{2.15}$$

Onde:

- n é o número total de observações,
- y_i é o valor observado no tempo i,
- $\hat{y}_i$ é o valor previsto no tempo i.

O MAPE é particularmente útil porque fornece uma medida intuitiva da precisão das previsões em termos percentuais. No entanto, ele pode ser influenciado por valores observados muito pequenos, resultando em valores de MAPE muito altos. Portanto, é importante considerar essa limitação ao interpretar os resultados.

2.5.2 Teste de Wilcoxon

O teste de Wilcoxon é um teste estatístico não paramétrico usado para comparar duas amostras emparelhadas para avaliar se suas populações de origem diferem. Ele é particularmente útil quando não se pode assumir que os dados seguem uma distribuição normal. Nele, se verifica a mediana das diferenças entre pares de observações. A hipótese nula do teste é que as diferenças entre os pares de observações têm uma mediana zero. A fórmula básica para calcular o valor de W no teste estatístico é apresentada na Equação 2.16:

$$W = \sum_{i=1}^{n} R_i, \tag{2.16}$$

onde:

- R_i é a classificação (ou rank) da diferença absoluta $|D_i|$ entre os pares de observações,
- $D_i = X_i - Y_i$ é a diferença entre as observações emparelhadas X_i e Y_i,
- n é o número de pares de observações.

Os passos para determinar o resultado do teste de Wilcoxon são os seguintes:

1. Calcular as diferenças $D_i = X_i - Y_i$.
2. Tomar os valores absolutos das diferenças $|D_i|$.
3. Classificar os valores absolutos $|D_i|$ em ordem crescente.
4. Atribuir aos ranks os sinais das diferenças D_i.
5. Calcular o estatístico de teste W, que é a soma dos ranks com o mesmo sinal (ou a soma dos ranks positivos, dependendo da convenção).

Aplicação do Teste

O teste de Wilcoxon é robusto a distribuições não normais e outliers, sendo uma escolha adequada quando essas condições são suspeitas. No contexto de validação de modelos de séries temporais, o teste pode ser usado para comparar as predições de dois modelos diferentes, determinando se há uma diferença estatisticamente significativa entre elas [Conover, 1999].

Capítulo 3

Metodologia

Nesta seção está detalhada a metodologia de cada etapa da pesquisa, de forma a permitir sua replicabilidade e facilitar contribuições futuras para melhorias nos resultados obtidos.

É apresentado o Organograma da Pesquisa e na sequência são descritos os passos de: Coleta de Dados, Limpeza e Pré-processamento dos Dados, Análise Exploratória, Determinação de Baselines, determinação de métricas de avaliação e Desenvolvimento do Filtro de Kalman.

Os códigos utilizados podem ser consultados nos apêndices deste trabalho e no repositório hospedado no GitHub. [1]

3.1 Organograma

A figura 3.1 a seguir apresenta o organograma da pesquisa, detalhando todas as suas etapas. Este organograma suplementa a compreensão da metodologia adotada, proporcionando uma visão clara e estruturada do fluxo de trabalho e das interações entre as diferentes fases da pesquisa.

O organograma da pesquisa está dividido em três partes, representadas por caixas na cor preta: Pesquisa Inicial, Determinação da Metodologia, e Execução dos Experimentos. Cada uma dessas partes compreende diversas etapas, representadas na cor cinza e detalhamentos, como subprocessos ou definições, que estão nos textos conectados à etapa.

A fase de pesquisa inicial foi orientada por duas etapas principais: Levantamento de Metodologias e Coleta de Dados. A fase de determinação da metodologia se deu após a escolha do método de predição a ser estudado, estando separada em: Limpeza e Pré-processamento, Análise Exploratória, Determinação de Baselines, Determinação de Métricas de Avaliação e Desenvolvimento do Filtro de Kalman.

Por fim, na fase de execução dos experimentos, foram cumpridas as etapas de Aplicação da Metodologia, Compilação dos Resultados e Comparação de Resultados.

[1] https://github.com/jflnetobr/kalman-filter/blob/main/codigo.ipynb

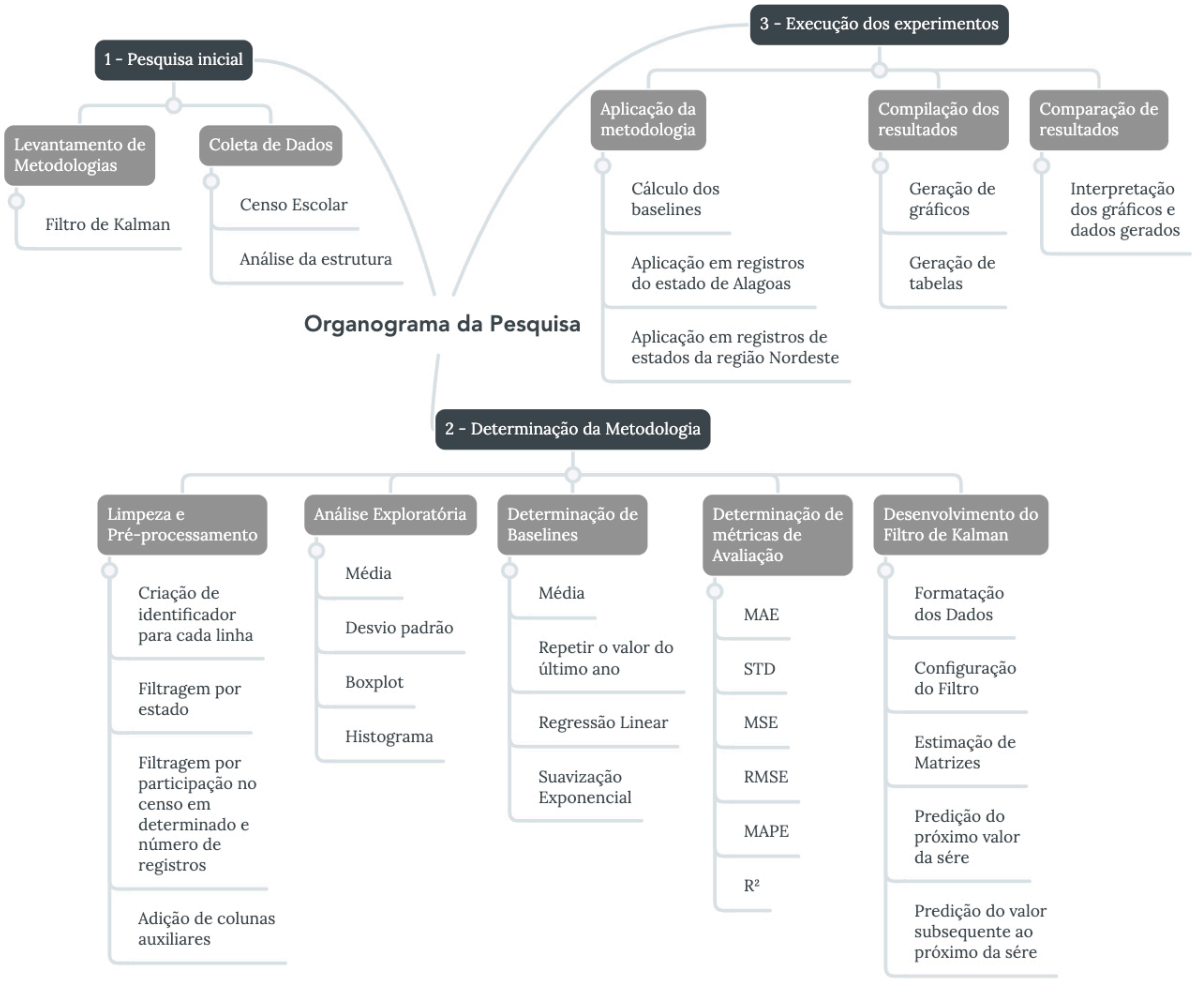


Figura 3.1: Organograma detalhado da pesquisa.

3.2 Coleta de Dados

A partir do Censo Escolar de 2023, as informações relativas ao número de alunos por etapa de ensino se tornaram públicas, podendo ser acessadas livremente através da política de dados abertos do Governo Federal.

Em relação aos anos de 2007 a 2022, os dados utilizados tiveram origem na base de dados interna do FNDE, que teve a tabela que detalha a quantidade de alunos matriculados a cada ano por etapa de ensino e entidade (escola) exportada. O autor, por fazer parte de projetos conduzidos pelo laboratório NEES/UFAL [2], teve acesso a estes dados e autorização do FNDE.

A condução da pesquisa foi autorizada por escrito por membro competente do FNDE e a privacidade dos dados obtidos foi resguardada, tendo sido manipulados apenas pelo próprio autor. Quaisquer dados apresentados neste trabalho foram devidamente descaracterizados e utilizados apenas como exemplo.

Outrossim, o experimento pode ser replicado utilizando qualquer série temporal que

[2]Termos de Execução Descentralizada (TEDs) 11668 e 12244

indique a variação da quantidade de alunos de uma determinada turma ao longo dos anos ou mesmo o comportamento de outro tipo de entidade que apresente variação semelhante.

Para fins de compreensão das etapas da metodologia, convém citar as principais colunas que compõem a estrutura do arquivo CSV utilizado como base para a pesquisa:

- **nu_ano_censo**: ano do registro em questão;
- **sg_uf**: estado (Unidade da Federação) em que se localiza a entidade em questão;
- **co_entidade**: código da entidade (escola) em questão;
- **co_etapa_ensino**: código da etapa de ensino em questão;
- **qt_censo**: quantidade de alunos informada no censo escolar para o ano, entidade e etapa de ensino relacionados.

3.3 Limpeza e Pré-processamento dos Dados

Para simplificar a manipulação dos dados, considerando que a tupla **co_entidade** e **co_etapa_ensino** é um valor que não se repete no mesmo ano, foi criada a coluna **co_turma**, composta pela combinação dos dois valores citados anteriormente e seguindo o padrão [**co_entidade**]-[**co_etapa_ensino**]. Esta coluna configurou um identificador que facilitou o acompanhamento da evolução da quantidade de alunos daquela etapa de ensino naquela entidade ao longo dos anos.

A base de dados original conta com mais de 8.760.000 linhas, que correspondem a mais de 1.630.000 turmas em cerca de 240.000 escolas. Neste trabalho, foram consideradas apenas entidades/etapas de ensino que atendessem aos seguintes critérios:

- **Participaram do censo escolar de 2023**: apenas os valores de **co_turma** que tivessem um registro com **nu_ano_censo** igual a 2023. Com esse conjunto de identificadores, o conjunto de dados foi filtrado para conter apenas valores cujo valor de **co_turma** estivesse presente no conjunto de identificadores citado.;
- **Cenário 1 - Dados do estado de Alagoas**
    1. **Localizadas no estado de Alagoas**: apenas os registros em que **sg_uf** fosse igual a "AL";
    2. **Contivessem ao menos dez registros ao longo dos anos (não necessariamente em sequência)**: os dados do item anterior foram agrupados pela coluna **co_turma** e foi adicionada a coluna **quantitativo**. O conjunto de dados resultante foi ordenado em ordem decrescente por esse quantitativo e foram excluídos itens com **quantitativo** menor que 10. A partir desses dados

filtrados foi obtido um conjunto de identificadores **co_turma**. Semelhante ao item anterior, o conjunto de dados foi filtrado para conter apenas valores presentes entre esses identificadores.

- **Cenário 2 - Dados da região Nordeste**

  1. **Localizadas na região Nordeste**: apenas os registros em que **sg_uf** fosse igual a um dos valores: "AL", "BA", "CE", "MA", "PB", "PE", "PI", "RN", "SE";
  2. **Contivessem ao menos treze registros ao longo dos anos (não necessariamente em sequência)**: os dados do item anterior foram agrupados pela coluna **co_turma** e foi adicionada a coluna **quantitativo**. O conjunto de dados resultante foi ordenado em ordem decrescente por esse quantitativo e foram excluídos itens com **quantitativo** menor que 13. A partir desses dados filtrados foi obtido um conjunto de identificadores **co_turma**. Semelhante ao item anterior, o conjunto de dados foi filtrado para conter apenas valores presentes entre esses identificadores.

Ao final do processo, restaram dois grupos de dados atendendo aos critérios fixados:

1. Apenas entidades/etapas de ensino do estado de Alagoas, que participaram do censo escolar de 2023 e com ao menos dez registros ao longo dos anos.

2. Apenas entidades/etapas de ensino da região Nordeste, que participaram do censo escolar de 2023 e com ao menos treze registros ao longo dos anos.

Para simplificar o processamento da série temporal foi gerado um novo conjunto de dados, com o objetivo de reunir em apenas uma linha toda a série temporal de uma entidade/etapa de ensino (representada por um **co_turma**). Para isso, distribuiu-se os valores de **qt_censo** conforme o **nu_ano_censo** para cada **co_turma**. Caso um registro para aquele ano não fosse encontrado, o valor zero foi utilizado.

Este novo conjunto de dados, contendo apenas o **co_turma** e os quantitativos de alunos para cada ano compondo a série temporal, foi utilizado para construção dos resultados de acordo com os métodos descritos a seguir.

Com o objetivo de facilitar a compreensão e replicabilidade, o código utilizado na etapa de pré-processamento está disponível na íntegra no Apêndice A.

3.4 Análise Exploratória

Para melhor compreensão dos dados, foram feitas algumas análises básicas a fim de determinar a dinâmica de variação do número de alunos e possíveis padrões que venham a ocorrer ao longo dos anos.

A princípio, calculou-se as seguintes estatísticas das séries temporais relativas a cada entidade/etapa de ensino, desconsiderando as interrupções (valores nulos):

1. Média ao longo dos anos
2. Mediana (média entre turmas)
3. Primeiro Quartil (média entre turmas)
4. Terceiro Quartil (média entre turmas)
5. Desvio Padrão médio por turma
6. Mínimo Geral
7. Máximo Geral

A partir desses valores, foi determinada a estatística geral para o cunjunto de dados, seguindo a metodologia adequada (média de todas as linhas, media por ano, média por turma, etc).

Para análise gráfica, os valores zero também foram desconsiderados e o processamento foi feito de acordo com o código referente à etapa de análise exploratória no Apêndice B.

Foram gerados um gráfico do tipo boxplot da quantidade de alunos por ano e um histograma da quantidade de alunos. Ambos os gráficos não consideraram o ano de 2023, uma vez que este foi utilizado apenas para cálculo de métricas e, para todos os efeitos, não era conhecido por outra parte do código além desta.

3.5 Determinação de Baselines

De forma a mensurar a eficácia das predições feitas pelo modelo de predição do alunado a ser criado, foram usados baselines como referências comparativas. Esses baselines proporcionam uma base para avaliar o desempenho do modelo proposto, permitindo identificar se as novas abordagens realmente oferecem melhorias significativas em relação aos métodos mais elementares de predição.

Neste trabalho, foram definidos quatro baselines principais para cada turma a ser analisada:

1. **Média dos Anos Anteriores:**
    - Utiliza a média das quantidades de alunos para o intervalo entre 2007 e 2022 (ou 2007 e 2021 para predições de 2 anos à frente), excluindo os anos em que o valor foi zero. Este método fornece uma estimativa que suaviza as variações anuais e remove prováveis ruídos, proporcionando uma visão geral da tendência histórica.

2. **Último Ano:**

    - Utiliza a quantidade de alunos informada no penúltimo ano disponível (2022) ou antepenúltimo (2021) para predições de 2 anos à frente como suposição para a quantidade de alunos do próximo ano (2023). Este método assume que a quantidade de alunos permanecerá constante e é o mais imediato quando se pensa em determinar próximos valores de séries temporais.

3. **Regressão Linear:**

    - Aplica um modelo de regressão linear aos dados históricos, ajustando uma linha reta que melhor represente a tendência ao longo do tempo. O valor predito é obtido extrapolando esta linha para o próximo ano ou para os próximos dois anos. Este método captura a tendência linear subjacente nos dados, sendo útil para séries temporais com tendências claras de crescimento ou declínio.

4. **Suavização Exponencial:**

    - Utiliza o método de suavização exponencial, que dá pesos maiores aos dados mais recentes, permitindo que a previsão seja mais sensível às mudanças recentes na série temporal. Este método é particularmente útil para capturar padrões sazonais e recentes tendências nos dados.

O código implementando os itens descritos está detalhado no Apêndice C. Os baselines definidos aqui servem como uma base para validar uma potencial melhoria trazida pelo uso do Filtro de Kalman na predição do número de alunos, oferecendo uma referência para medir o sucesso da abordagem proposta.

3.6 Métricas de Avaliação

Para avaliar a performance do modelo de predição utilizado neste trabalho, foram implementadas diversas métricas de avaliação que permitem quantificar a precisão e a robustez das predições em relação aos dados reais. As métricas escolhidas fornecem uma visão abrangente do desempenho do modelo sob diferentes perspectivas, desde a magnitude dos erros absolutos até a proporção dos erros relativos e a variabilidade dos erros quadráticos.As métricas escolhidas foram:

1. **Mean Absolute Error (MAE):**

    - O MAE é uma métrica que calcula a média das diferenças absolutas entre os valores preditos e os valores reais. Esta métrica fornece uma medida clara da magnitude média dos erros de predição, independentemente da direção dos erros (subestimação ou superestimação).

2. **Standard Deviation (STD):**

    - O desvio padrão dos erros de predição mede a variabilidade desses erros em torno da média. Um desvio padrão baixo indica que os erros são consistentemente próximos da média, enquanto um desvio padrão alto sugere maior dispersão dos erros.

3. **Mean Squared Error (MSE):**

    - O MSE calcula a média dos quadrados das diferenças entre os valores preditos e os valores reais. Esta métrica dá maior peso aos grandes erros, penalizando predições que estão muito distantes dos valores reais. O MSE é útil para destacar a presença de outliers.

4. **Root Mean Squared Error (RMSE):**

    - O RMSE é a raiz quadrada do MSE, mantendo as mesmas unidades dos valores originais. Esta métrica também penaliza erros maiores, mas é mais interpretável que o MSE por estar na mesma escala que os dados.

5. **Mean Absolute Percentage Error (MAPE):**

    - O MAPE expressa o erro médio absoluto como uma porcentagem dos valores reais, fornecendo uma perspectiva relativa dos erros. Esta métrica é particularmente útil para comparar a precisão das predições entre diferentes escalas ou contextos.

6. **R-squared (R^2):**

    - O R^2, ou coeficiente de determinação, mede a proporção da variância nos valores reais que é explicada pelas predições do modelo. Um R^2 próximo de 1 indica que o modelo explica bem a variabilidade dos dados, enquanto um R^2 próximo de 0 sugere que o modelo não capta bem a variabilidade dos dados.

Para cada predição gerada, foram calculadas as métricas acima utilizando como referência os valores reais do número de alunos no ano de 2023. As métricas foram calculadas tanto para os baselines definidos quanto para o Filtro de Kalman proposto, permitindo uma comparação direta de desempenho. O procedimento segue os seguintes passos:

1. **Seleção das Colunas de Interesse:** As colunas que contêm os valores reais (ex: 2023) e as predições (ex: baseline_media, baseline_last_year, baseline_linear_regression, baseline_exponential_smoothing, etc) foram selecionadas para avaliação.

2. **Cálculo das Métricas:** Utilizando a função `evaluate_metrics`, foram calculadas as métricas para cada combinação de valor real e predição, armazenando os resultados para análise posterior.

3. **Análise Comparativa:** Com base nos resultados das métricas, foi possível comparar o desempenho do Filtro de Kalman em relação aos baselines, identificando pontos fortes e fracos de cada abordagem.

Além disso, o teste de Wilcoxon foi utilizado para comparar as predições do modelo com diferentes baselines, avaliando se há uma diferença estatisticamente significativa entre as distribuições dos erros das predições e dos baselines. Este teste não paramétrico é ideal para amostras pareadas e para avaliar se a mediana das diferenças entre pares é significativamente diferente de zero, fornecendo uma validação da superioridade ou equivalência das predições do modelo em comparação com os métodos tradicionais.

Esta abordagem metodológica permite uma avaliação sob diversos aspectos do desempenho do modelo de predição proposto, além de fornecer direcionamentos para melhorias futuras e ajustes no modelo. As métricas de avaliação são fundamentais para garantir que as predições são realmente precisas e úteis no contexto do planejamento educacional e alocação de recursos do PNLD.

O código utilizado para determinação de métricas está disponível no Apêndice D.

3.7 Desenvolvimento do Filtro de Kalman

Como já mencionado, o objeto de estudo do trabalho foi o Filtro de Kalman simples para realizar a predição do número de alunos nas escolas públicas, neste caso, do estado de Alagoas e da Região Nordeste.

3.7.1 Formatação dos Dados

O primeiro passo na implementação do Filtro de Kalman foi a formatação dos dados. O objetivo era preparar as séries temporais de forma a ser eficientemente utilizadas. Esta etapa contou com a remoção de zeros iniciais, preenchimento de valores faltantes e a transformação dos dados para o formato apropriado. Esse processo foi realizado conforme detalhado nas linhas 6 a 11 do código presente no Apêndice E.

3.7.2 Configuração do Filtro de Kalman

Após o pré-processamento, o filtro foi configurado, definindo as matrizes de transição e observação. No código presente no Apêndice E, na linha 14, está indicada a configuração das matrizes utilizando o parâmetro `em_vars`, que as inicializa automaticamente.

3.7.3 Execução do Algoritmo EM

Para estimar os parâmetros do filtro, foi utilizado o algoritmo de Expectation-Maximization (EM). Este algoritmo ajusta iterativamente as matrizes de covariância de transição e observação para maximizar a probabilidade dos dados observados. Os detalhes da execução do algoritmo EM estão disponíveis no Apêndice E, nas linhas 17 a 19.

3.7.4 Predição do Próximo Valor

Com o filtro configurado e os parâmetros estimados, utilizou-se o estado suavizado para predizer o próximo valor da série. A predição é realizada aplicando a matriz de transição ao último estado suavizado e, em seguida, a matriz de observação ao novo estado estimado.

Para a predição de dois anos à frente, o mesmo procedimento é aplicado iterativamente. Inicialmente, prediz-se o valor para o próximo ano como descrito anteriormente. Esse valor predito é então usado como entrada para uma segunda iteração do filtro, aplicando novamente a matriz de transição ao novo estado suavizado e, em seguida, a matriz de observação para estimar o valor do ano subsequente.

O processo completo de predição, incluindo a manipulação dos estados e matrizes, está descrito no Apêndice E nas linhas 22 a 27.

Capítulo 4

Resultados

O capítulo de resultados está estruturado de forma a apresentar de maneira detalhada e sistemática os achados da pesquisa, com foco na análise dos dados das escolas do estado de Alagoas e também da região Nordeste. Neste capítulo, discutimos a eficácia dos métodos de predição aplicados e a validação das hipóteses levantadas no início do estudo. Os resultados são organizados em duas grandes seções: uma dedicada exclusivamente ao estado de Alagoas e outra abrangendo a região Nordeste, que passaram rigorosamente pelos mesmos experimentos, descritos nos parágrafos a seguir.

Dentro de cada seção, a subseção **Limpeza e Pré-processamento dos Dados** detalha o processo de tratamento dos dados brutos, descrevendo as quantidades de escolas e turmas presentes antes e após filtrar os dados de acordo com os critérios pré-estabelecidos. Na subseção **Análise Exploratória**, alguns dados relevantes são apresentados, além da análise do histograma e boxplot dos dados.

Em **Predizendo Valores para o Ano Posterior**, inicialmente são determinados os baselines e, em seguida, aplicado o Filtro de Kalman para estimar o número de alunos no ano seguinte. A subseção **Predizendo Valores para o Ano Subsequentemente ao Posterior** estende o mesmo tipo de análise para predizer o número de alunos dois anos à frente, uma vez que o desafio principal do problema é ser o mais preciso possível com uma defasagem de dois anos. Por fim, em **Analisando os Resultados**, são consolidadas as informações apresentadas, discutidos os achados principais e suas implicações. Esta seção sintetiza os resultados das predições e comenta sobre a performance da abordagem sugerida.

A tabela 4.1 traz um exemplo com algumas linhas do conjunto de dados base, onde foram aplicados todos os procedimentos. Alguns valores foram substituídos por * para manter o anonimato das informações, além de não serem relevantes para o trabalho executado.

Tabela 4.1: Dados de Exemplo de Escolas no arquivo original

nu_ano_censo	co_uf	sg_uf	co_municipio	no_municipio	co_entidade
2022	*	UF	*	*	*
2023	*	UF	*	*	*
2015	*	UF	*	*	*
2012	*	UF	*	*	*
2018	*	UF	*	*	*

no_entidade	co_tp_localizacao	ds_tp_localizacao	co_etapa_ensino
*	1	Urbana	0
*	1	Urbana	0
*	1	Urbana	25
*	1	Urbana	25
*	1	Urbana	25

no_etapa_ensino	qtd_alunos	tp_dependencia
Não Se Aplica	13	Estadual
Não Se Aplica	5	Estadual
Ensino Médio - 1ª Série	92	
Ensino Médio - 1ª Série	154	
Ensino Médio - 1ª Série	114	

Fonte: Elaborado pelo autor.

4.1 Analisando dados de escolas do estado de Alagoas

A seção dedicada à análise dos dados das escolas do estado de Alagoas tem como objetivo apresentar os processos e resultados das predições realizadas, estando subdividida da forma que já foi descrita anteriormente. A razão da escolha, além de ser o estado onde está localizada a UFAL e que merece receber destaque, sendo o primeiro a receber os testes da nova metodologia, reside no fato de que Alagoas é um dos estados com desafios educacionais significativos no país, de forma que uma predição precisa do número de alunos pode contribuir para a alocação eficiente de recursos educacionais.

4.1.1 Limpeza e Pré-processamento dos Dados

Inicialmente, foi realizada uma limpeza e pré-processamento dos dados brutos no formato apresentado na Tabela 4.1, de forma a manter apenas turmas de escolas que participaram do censo escolar do ano de 2023, necessário para determinação das métricas, e com 10 ou mais registros ao longo dos anos (que passaram a compor sua série temporal), de forma a garantir uma quantidade de dados relevante e ao mesmo tempo manter uma amostra significativa.

A Tabela 4.2 traz os quantitativos a cada etapa, em que podemos destacar o total de turmas participantes do censo escolar 2023, cerca de 17.300, e a quantidade de turmas

que foram consideradas para as demais etapas, pouco mais de 3.100, que corresponde a aproximadamente um quinto do total possível.

Tabela 4.2: Análise do arquivo original com dados relativos à quantidade de alunos das escolas do estado de Alagoas.

Descrição	**Quantidade**
Total de registros no arquivo original	141273
Total de escolas no arquivo original	2895
Total de registros das turmas que participaram do censo de 2023	104534
Total de turmas que participaram do censo de 2023	17319
Total de registros das turmas que participaram do censo de 2023 e têm 10 ou mais registros	33502
Total de turmas que participaram do censo de 2023 e têm 10 ou mais registros	3134

Fonte: Elaborado pelo autor.

Após a filtragem, os dados foram reorganizados da forma descrita no capítulo relativo à metodologia de maneira a criar uma série temporal em cada linha, correspondendo a uma turma, além de acrescentar um contador e uma variável indicando se existem interrupções ou não na série. Um exemplo do formato em que os dados foram trabalhados deste ponto em diante pode ser visto na Tabela 4.3. O código da turma foi substituído por letras que indicam seu formato (E para dígitos relativos à escola e T para dígitos relativos à turma) para manter a confidencialidade dos dados:

Tabela 4.3: Série temporal resultante do pré-processamento

co_turma	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
EEEEEEEE-TT	0.0	178.0	0.0	0.0	0.0	0.0	238.0	0.0	202.0	187.0
EEEEEEEE-TT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	48.0	50.0	37.0
EEEEEEEE-TT	170.0	0.0	0.0	0.0	57.0	61.0	0.0	0.0	84.0	0.0
EEEEEEEE-TT	83.0	0.0	52.0	0.0	0.0	0.0	0.0	0.0	0.0	77.0
EEEEEEEE-TT	0.0	190.0	0.0	96.0	0.0	0.0	0.0	0.0	68.0	0.0

2017	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**nu_valores**	**has_interruption**
0.0	137.0	77.0	82.0	81.0	79.0	119.0	10	True
42.0	43.0	48.0	40.0	42.0	68.0	31.0	10	False
0.0	45.0	53.0	48.0	49.0	50.0	50.0	10	True
50.0	68.0	48.0	49.0	56.0	42.0	46.0	10	True
84.0	78.0	77.0	81.0	78.0	66.0	93.0	10	True

Fonte: Elaborado pelo autor.

4.1.2 Análise Exploratória

Nesta seção, apresentamos uma visão inicial dos dados após o tratamento, permitindo a identificação de padrões e outliers que possam influenciar as etapas subsequentes ou explicar itens delas. Um dos primeiros passos foi contar o número de escolas que possuem diferentes quantidades de registros em suas séries temporais. Esse levantamento é fundamental para entender a distribuição dos dados e a presença de dados das escolas ao longo dos anos. A Tabela 4.4 mostra a contagem de escolas com determinada quantidade de registros na série temporal das escolas do estado de Alagoas.

Tabela 4.4: Contagem de escolas com determinada quantidade de registros na série temporal das escolas do estado de Alagoas.

Quantidade	**Número de Escolas**
10	1670
11	946
12	364
13	133
14	17
15	3
16	1

Fonte: Elaborado pelo autor.

A maioria das escolas tem exatamente 10 registros na série temporal, totalizando 1670 escolas. À medida que o número de registros aumenta, a quantidade de escolas diminui drasticamente, com apenas uma escola possuindo 16 registros.

Além da contagem de registros, outra análise importante é a das estatísticas dos dados. A Tabela 4.5 apresenta as principais estatísticas em relação às turmas analisadas.

Tabela 4.5: Estatísticas dos Dados das escolas do estado de Alagoas (média entre os anos).

Estatística	**Valor**
Média ao longo dos anos	75.48
Mediana (média entre turmas)	50.38
Primeiro Quartil (média entre turmas)	25.75
Terceiro Quartil (média entre turmas)	93.70
Desvio Padrão médio por turma	24.09
Mínimo Geral	1.0
Máximo Geral	2384.0

Fonte: Elaborado pelo autor.

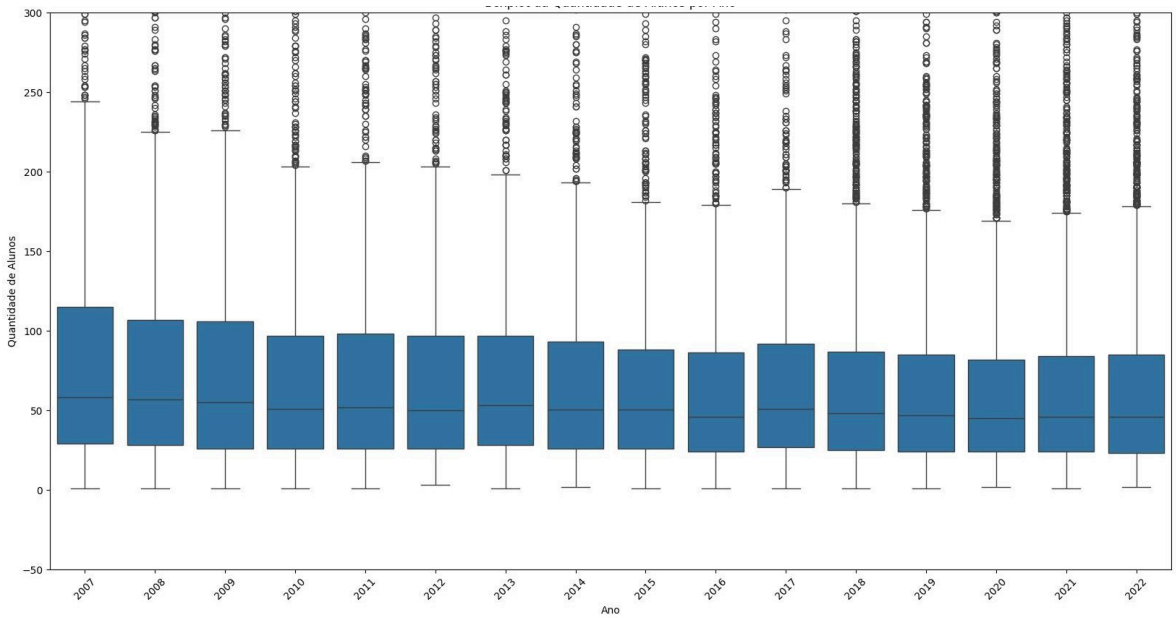
A média de alunos por turma é de cerca de 75 alunos, enquanto o desvio padrão médio por turma é de cerca de 18, o que sugere uma variação relevante no número de alunos de uma turma entre um ano e outro. A mediana gira em torno de 50 alunos por turma,

indicando que, em média, metade das turmas tem essa quantidade ou menos alunos ao longo dos anos.

A menor turma observada foi de 1 aluno e a maior contava com 2384 estudantes, o que sugere uma agregação de dados de várias turmas em uma só, tendo em vista a inviabilidade de turmas desse tamanho na prática. Essa informação também indica variações significativas no tamanho das turmas, o que traz impacto no desempenho dos métodos preditivos.

O boxplot apresentado na Figura 4.1 exibe a distribuição da quantidade de alunos por turma ao longo dos anos para as escolas do estado de Alagoas.

Figura 4.1: Boxplot dos dados do estado de Alagoas.



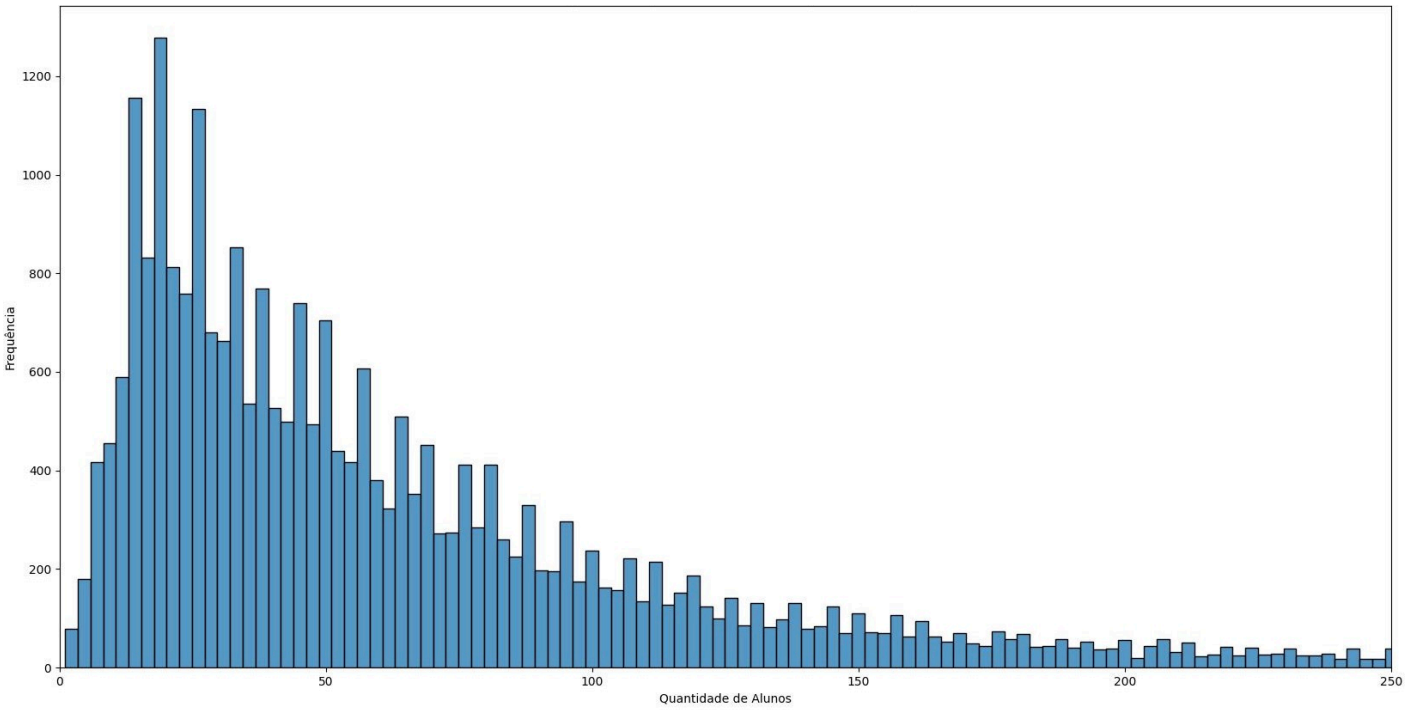
Fonte: Elaborado pelo autor.

Através de sua análise, podemos notar medianas semelhantes, assim como o intervalo interquartil ao longo dos anos, o que sugere que a quantidade de alunos por turma se mantém razoavelmente estável durante o período observado, mas, conforme o desvio padrão apresentado na Tabela 4.5, com grande variabilidade durante os anos. Além disso, em todos os anos, o número de outliers é substancial, com muitos pontos de dados acima dos limites superiores do boxplot.

A semelhança entre medianas é um aspecto positivo que pode facilitar a tarefa de predição. No entanto, a presença significativa de outliers e a variação dos dados ao longo dos anos representam desafios importantes. Estes fatores destacam a necessidade de técnicas robustas para lidar com os dados e fornecer resultados verdadeiramente relevantes, garantindo que as predições sejam tanto precisas quanto confiáveis.

O histograma apresentado na Figura 4.2 exibe a distribuição da quantidade de alunos por turma ao longo dos anos para as escolas do estado de Alagoas selecionadas:

Figura 4.2: Histograma dos dados do estado de Alagoas.



Fonte: Elaborado pelo autor.

O histograma se apresenta de forma unimodal, com um único pico na faixa de turmas entre 20 e 30 alunos, indicando ser a mais comum entre as escolas analisadas. À medida que a quantidade de alunos por turma aumenta, a frequência das turmas diminui drasticamente, como é possível observar na cauda.

A distribuição é assimétrica à direita (skewed right), com uma longa cauda que se estende para a direita. Isso sugere que há algumas turmas com números muito elevados de alunos.

4.1.3 Predizendo valores para 1 ano à frente

Determinação de Baselines

Para garantir que as predições obtidas com o modelo alvo do estudo são realmente relavantes e confiáveis, foram estabelecidos os baselines: Média, Repetir o Último Ano, Regressão Linear e Suavização Exponencial, seguindo o descrito na metodologia. A tabela 4.6 resume as métricas de avaliação desses baselines para as escolas do estado de Alagoas, considerando uma predição para 1 ano à frente.

A partir dos dados apresentados na tabela, observa-se que o baseline de Suavização Exponencial obteve os melhores resultados gerais, com menor MAE, STD, MSE, RMSE e MAPE, além do maior valor de R^2. Estes resultados sugerem que este método é o

Tabela 4.6: Métricas dos Baselines para escolas do estado de Alagoas predizendo 1 ano à frente

Baseline	**MAE**	**STD**	**MSE**	**RMSE**	**MAPE**	R^2
Média	21.85	40.14	1664.46	40.80	49.59	0.6210
Repetir o último Ano	15.27	34.65	1204.54	34.71	30.28	0.7257
Regressão Linear	17.95	33.85	1152.56	33.95	35.76	0.7376
Suavização Exponencial	14.30	28.31	805.01	28.37	29.72	0.8167

Fonte: Elaborado pelo autor.

mais adequado para capturar a variabilidade e as tendências presentes nos dados, proporcionando predições mais precisas para o ano seguinte, de forma que apresentar um desempenho superior acaba sendo a ”meta”do Filtro de Kalman.

Comparativamente, o baseline de Repetir o Último Ano também apresenta um desempenho considerável, especialmente no valor do MAE, indicando que em muitos casos a quantidade de alunos permanece relativamente constante de um ano para o outro. No entanto, para capturar variações e tendências mais complexas, a Suavização Exponencial demonstra ser mais eficiente.

Execução dos experimentos

Buscando o melhor resultado possível, foram realizados experimentos variando a quantidade de iterações na execução do Filtro de Kalman. Com estes resultados, é possível visa entender como o número de iterações influencia nas métricas avaliadas.

A Tabela 4.7 resume os resultados dos experimentos para diferentes quantidades de iterações:

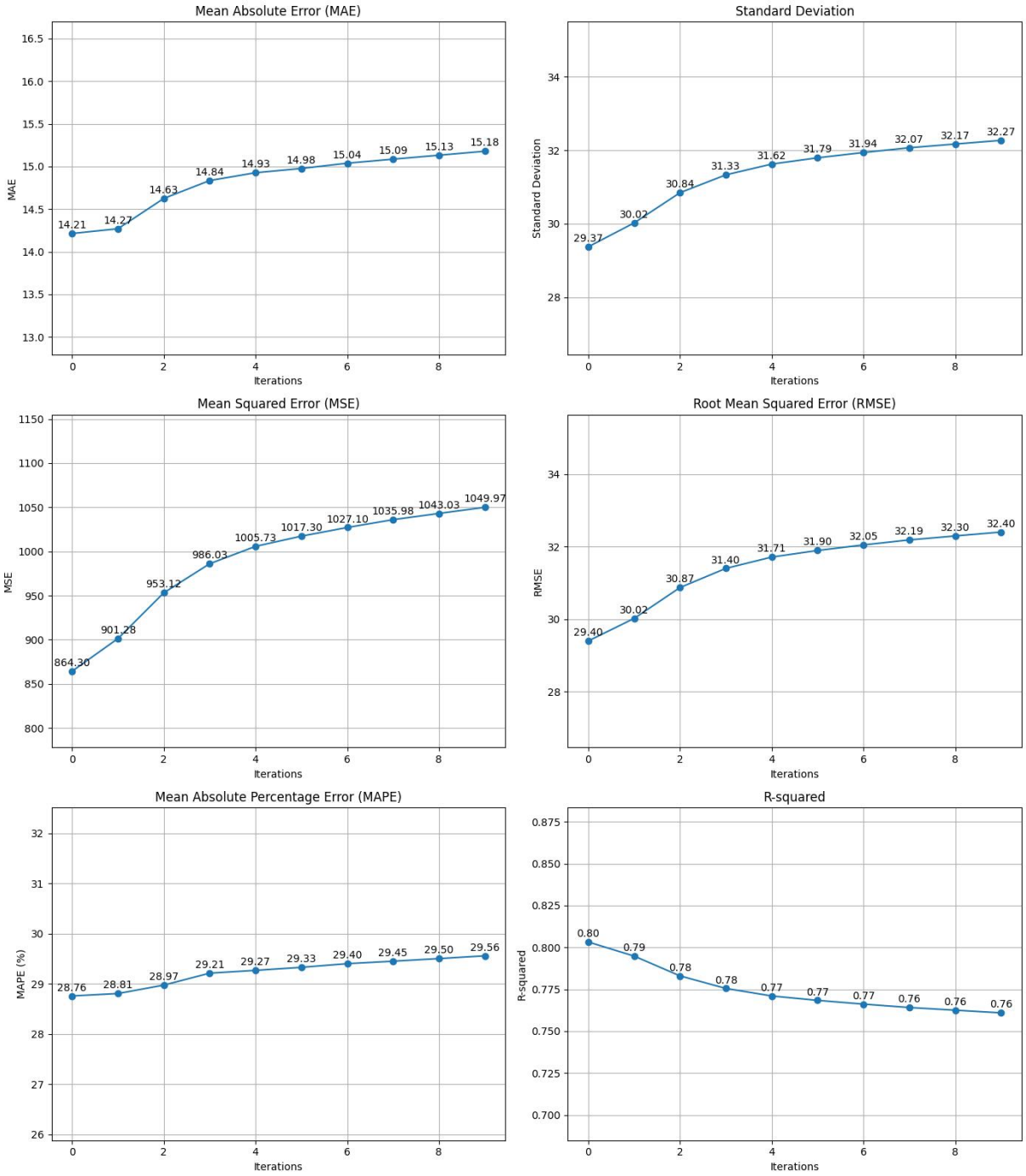
Tabela 4.7: Resultados do Experimento de predição para 1 ano à frente em escolas do estado de Alagoas com diferentes quantidades de iterações no Filtro de Kalman

Métrica	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
MAE	14.21	14.27	14.63	14.84	14.93	14.98	15.04	15.09	15.13	15.18
STD	29.37	30.02	30.84	31.33	31.62	31.79	31.94	32.07	32.17	32.27
MSE	864.30	901.28	953.12	986.03	1005.73	1017.30	1027.10	1035.98	1043.03	1049.97
RMSE	29.40	30.02	30.87	31.40	31.71	31.90	32.05	32.19	32.30	32.40
MAPE	28.76	28.81	28.97	29.21	29.27	29.33	29.40	29.45	29.50	29.56
R^2	0.8032	0.7948	0.7830	0.7755	0.7710	0.7684	0.7661	0.7641	0.7625	0.7609

Fonte: Elaborado pelo autor.

A Figura 4.3 traz uma visualização gráfica das informações contidas na Tabela 4.7.

Figura 4.3: Gráficos apresentando as métricas das execuções do Filtro de Kalman nos dados do estado de Alagoas com diferentes quantidades de iterações no Filtro de Kalman predizendo 1 ano à frente.



Fonte: Elaborado pelo autor.

Os resultados dos experimentos mostram uma tendência clara em relação ao impacto das iterações na performance do Filtro de Kalman:

MAE (Mean Absolute Error):

- Observa-se que o MAE aumenta de forma gradual conforme o número de iterações aumenta. Este comportamento indica que, embora o Filtro de Kalman melhore a suavização dos dados com mais iterações, ele também tende a se ajustar demais aos dados da série fornecida, reduzindo a capacidade de generalizar para novos dados.

STD (Standard Deviation):

- O desvio padrão também aumenta com o número de iterações. Isso sugere que os erros das predições se tornam mais dispersos conforme o modelo passa por mais iterações, refletindo uma menor consistência nos resultados preditivos.

MSE (Mean Squared Error) e RMSE (Root Mean Squared Error):

- Tanto o MSE quanto o RMSE seguem um padrão de aumento similar ao do MAE e STD. A penalização mais severa de erros maiores pelo MSE e RMSE indica que as predições tendem a apresentar desvios mais significativos com o aumento das iterações.

MAPE (Mean Absolute Percentage Error):

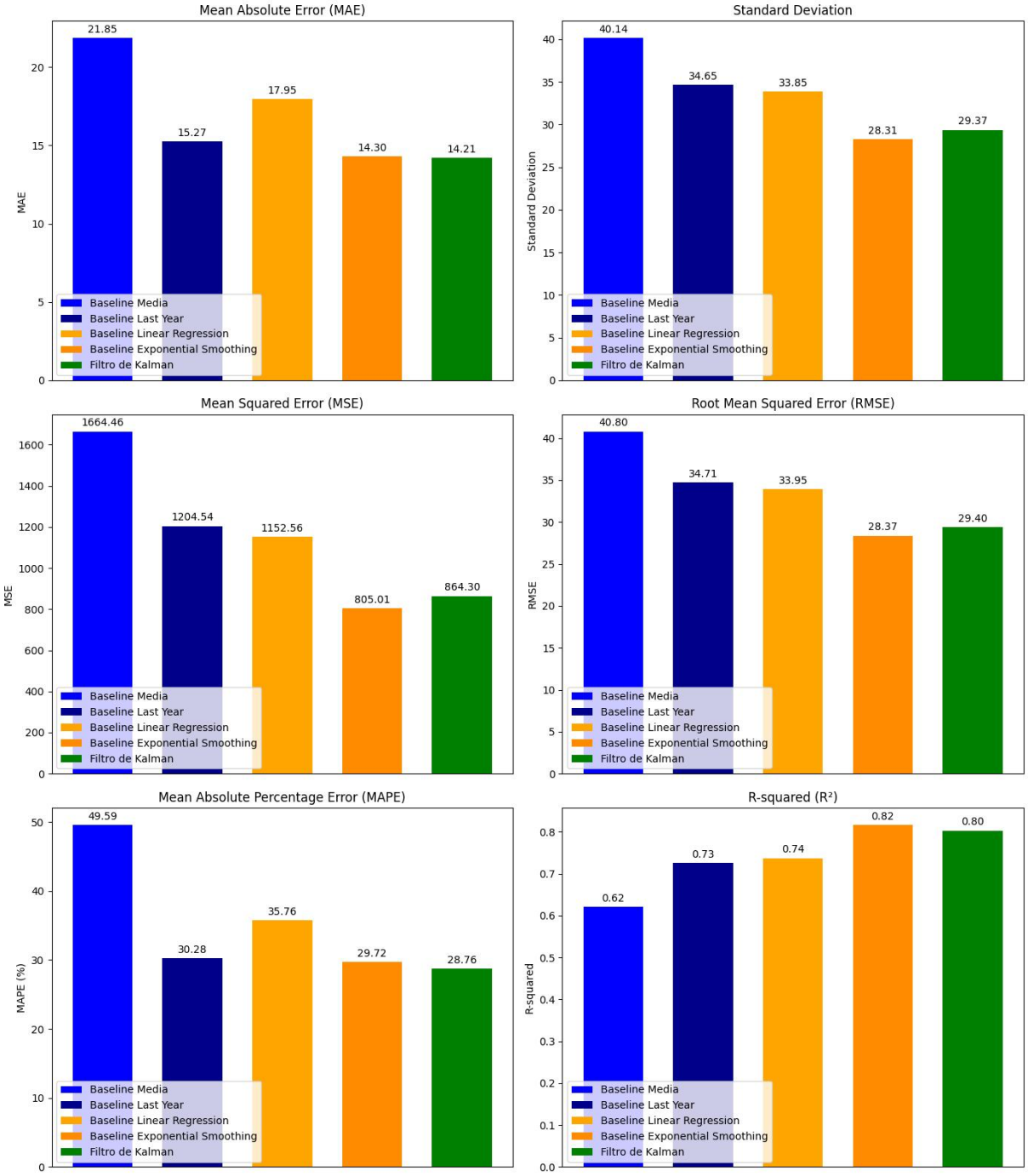
- O MAPE apresenta um leve aumento com o número de iterações, mostrando que a precisão percentual das predições em relação aos valores reais se degrada lentamente conforme o modelo se ajusta mais aos dados de treinamento.

R^2 (R-squared):

- O valor do R^2 diminui gradualmente, o que indica uma redução na capacidade do modelo em explicar a variabilidade dos dados. Este comportamento reflete um ajuste excessivo aos dados da série fornecida como entrada.

O gráfico apresentado na Figura 4.4 mostra uma comparação métrica a métrica entre os baselines e o Filtro de Kalman (considerando o melhor resultado do MAE).

Figura 4.4: Gráficos apresentando a comparação entre diferentes métricas dos baselines e do Filtro de Kalman (melhor resultado) com dados do estado de Alagoas predizendo 1 ano à frente.



Fonte: Elaborado pelo autor.

Comparando o desempenho do Filtro de Kalman em relação aos baselines, podemos fazer uma análise sob a perspectiva de cada métrica:

MAE (Mean Absolute Error):

- Observa-se que a suavização exponencial e o Filtro de Kalman apresentaram os menores MAEs, com valores de 14.30 e 14.21 respectivamente, destacando-se como as técnicas mais precisas. Em contrapartida, a média dos anos anteriores resultou no maior MAE (21.85), demonstrando ser a menos eficiente entre os baselines considerados.

STD (Standard Deviation):

- A suavização exponencial e o Filtro de Kalman mantiveram os menores desvios padrões, 28.31 e 29.37 respectivamente, indicando uma maior consistência nas predições em comparação com as outras técnicas. O maior desvio padrão foi observado para a média dos anos anteriores (40.14), reforçando sua falta de precisão.

MSE (Mean Squared Error) e RMSE (Root Mean Squared Error):

- Tanto o MSE quanto o RMSE seguiram padrões semelhantes ao MAE e STD, onde a suavização exponencial (805.01 e 28.37) e o Filtro de Kalman (864.30 e 29.40) apresentaram os melhores resultados. A média dos anos anteriores teve os piores desempenhos com valores de 1664.46 para MSE e 40.80 para RMSE.

MAPE (Mean Absolute Percentage Error):

- O Filtro de Kalman e a suavização exponencial se destacaram novamente com os menores valores de MAPE, 28.76% e 29.72%, respectivamente. A média dos anos anteriores apresentou o maior MAPE (49.59%), confirmando sua baixa eficiência.

R^2 (R-squared):

- O Filtro de Kalman apresentou o segundo maior valor de R^2 (0.80), enquanto a suavização exponencial teve o melhor desempenho (0.82), indicando que ambos têm uma boa capacidade de explicação da variabilidade dos dados. A média dos anos anteriores teve o menor R^2 (0.62), refletindo uma menor capacidade preditiva.

Em suma, a comparação entre as diferentes técnicas mostra que o Filtro de Kalman e a suavização exponencial são os métodos mais eficazes para predição, proporcionando maior precisão e consistência, mas sem unanimidade entre as métricas para esse primeiro experimento. A média dos anos anteriores se mostrou ineficaz, com maiores erros e menor capacidade de explicação dos dados.

4.1.4 Predizendo valores para 2 anos à frente

Determinação de Baselines

Aqui são apresentadas as métricas de avaliação dos diferentes baselines utilizados para predizer os valores para 2 anos à frente, no caso das escolas do estado de Alagoas. A Tabela 4.8 resume as métricas de desempenho para cada baseline:

Tabela 4.8: Métricas dos Baselines para escolas do estado de Alagoas predizendo 2 anos à frente

Baseline	**MAE**	**STD**	**MSE**	**RMSE**	**MAPE**	R^2
Média	23.56	43.72	1974.10	44.43	53.34	0.5505
Repetir o último Ano	18.07	34.59	1197.43	34.60	35.43	0.7273
Regressão Linear	21.28	39.62	1585.71	39.82	42.01	0.6389
Suavização Exponencial	16.53	31.34	985.25	31.39	34.49	0.7757

Fonte: Elaborado pelo autor.

Os resultados obtidos seguem um padrão semelhante ao observado na predição para um ano à frente. O baseline de suavização exponencial e o de Repetir o último ano apresentaram os melhores desempenhos, com menores valores de MAE, STD, MSE, RMSE e MAPE, além de valores mais altos de R^2. Em contrapartida, a média dos anos anteriores continua demonstrando um desempenho inferior, com maiores erros absolutos e quadráticos.

Execução dos experimentos

A Tabela 4.9 apresenta os resultados dos experimentos para escolas do estado de Alagoas ao predizer 2 anos à frente, considerando diferentes quantidades de iterações do Filtro de Kalman:

Tabela 4.9: Resultados do Experimento de predição para 2 anos à frente em escolas do estado de Alagoas com diferentes quantidades de iterações no Filtro de Kalman

Métrica	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
MAE	16.34	16.18	16.62	16.87	17.00	17.10	17.17	17.24	17.31	17.38
STD	31.27	31.02	31.30	31.49	31.61	31.70	31.78	31.85	31.92	31.99
MSE	977.68	963.21	994.46	1015.62	1027.06	1034.63	1041.61	1047.74	1053.85	1059.52
RMSE	31.27	31.04	31.54	31.87	32.05	32.17	32.27	32.37	32.46	32.55
MAPE	32.81	32.37	32.42	32.52	32.61	32.75	32.83	32.94	33.02	33.11
R^2	0.7774	0.7807	0.7736	0.7687	0.7661	0.7644	0.7628	0.7614	0.7600	0.7587

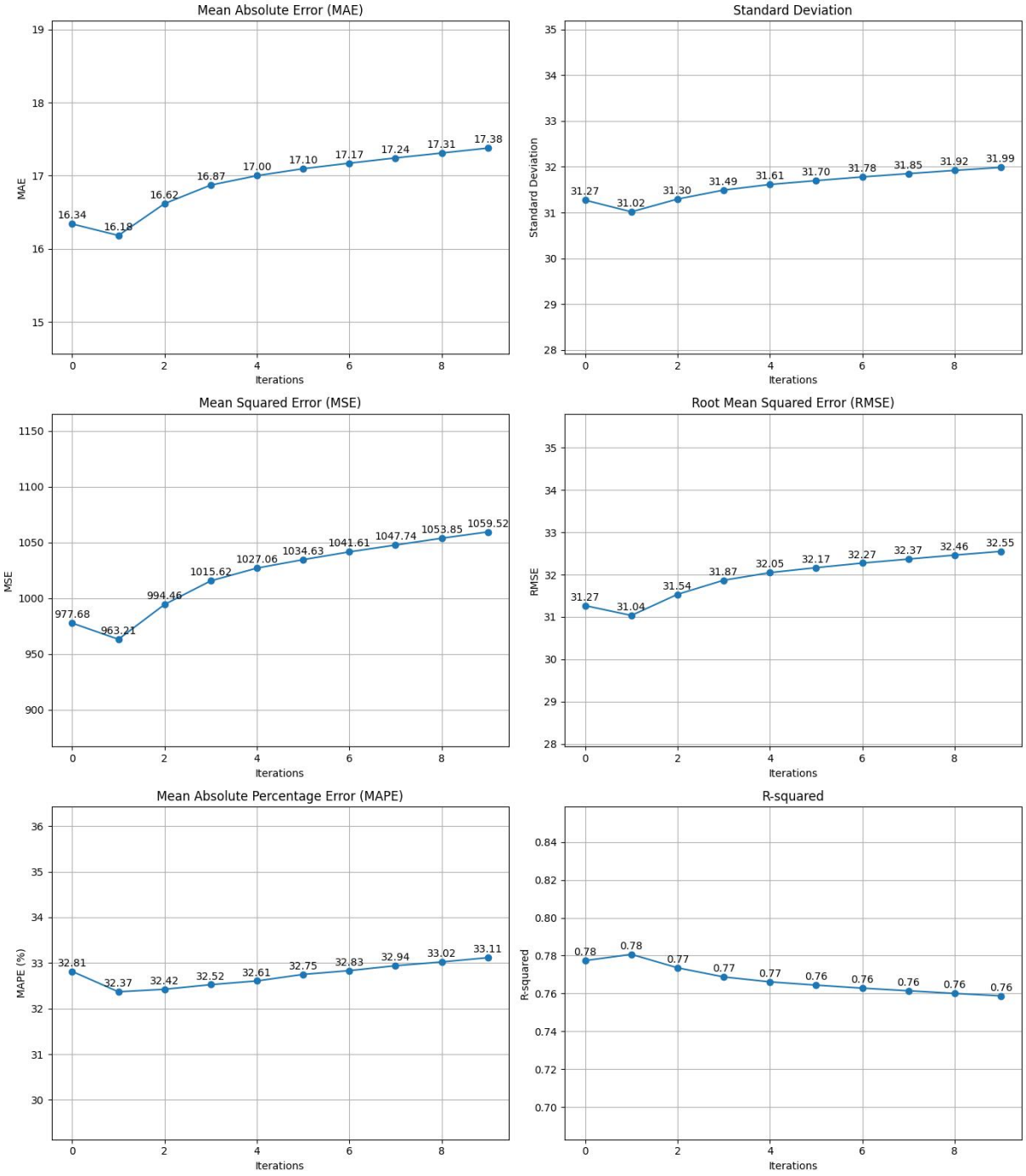
Fonte: Elaborado pelo autor.

Diferentemente do observado na predição para um ano à frente, neste caso, não foi a execução com zero iterações que obteve o melhor resultado. Além disso, os valores das métricas de erro foram consistentemente maiores do que os encontrados na predição para um ano, refletindo a maior incerteza associada a predições realizadas para um período mais longo.

A Figura 4.5 traz uma visualização gráfica das informações contidas na Tabela 4.9.

Observa-se que, ao longo das iterações, as métricas tendem a piorar gradualmente, assim como na predição para um ano à frente, indicando um aumento nos erros das predições. No entanto, um ponto interessante é que, ao utilizar apenas uma iteração, há uma ligeira melhoria nas métricas em comparação com a execução sem iterações. Isso sugere que uma única iteração pode proporcionar um ajuste inicial aos dados que melhora a qualidade das predições, mas iterações adicionais podem levar a um ajuste excessivo aos dados de treinamento, reduzindo a capacidade do modelo de generalizar para novos dados.

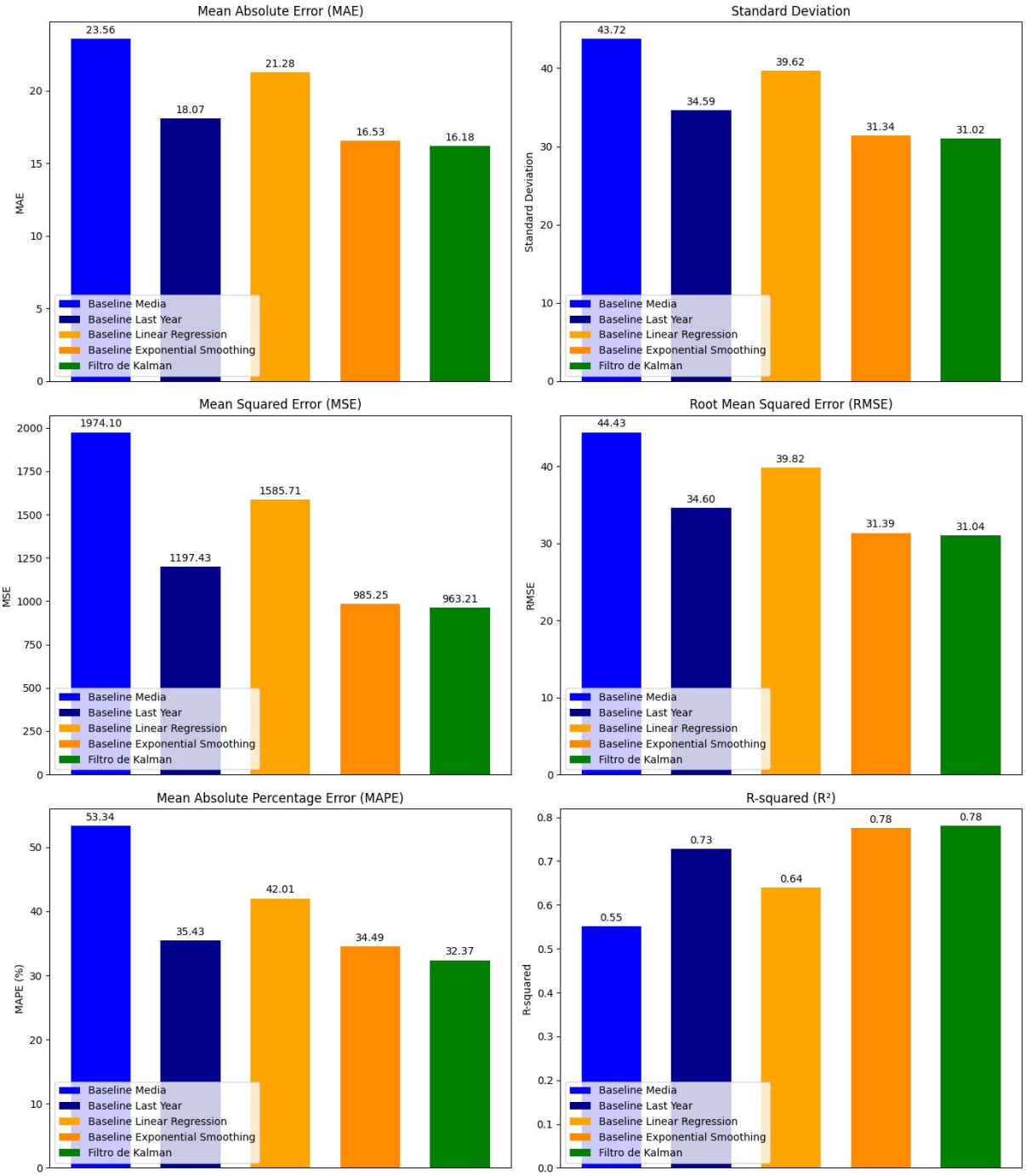
Figura 4.5: Gráficos apresentando as métricas das execuções do Filtro de Kalman nos dados do estado de Alagoas com diferentes quantidades de iterações predizendo 2 anos à frente.



Fonte: Elaborado pelo autor.

O gráfico apresentado na Figura 4.6 mostra uma comparação métrica a métrica entre os baselines e o Filtro de Kalman (considerando o melhor resultado do mae).

Figura 4.6: Gráficos apresentando a comparação entre diferentes métricas dos baselines e do Filtro de Kalman (melhor resultado) com dados do estado de Alagoas predizendo 2 anos à frente.



Fonte: Elaborado pelo autor.

Observa-se que o Filtro de Kalman, neste caso, apresentou o melhor desempenho em basicamente todas as métricas. A métrica MAE mostra que o Filtro de Kalman tem um erro absoluto médio mais baixo comparado aos baselines,sendo cerca de 0.35 menor que

a Suavização Exponencial, indicando predições mais precisas. O desvio padrão (STD) é o mesmo e o erro médio quadrático (MSE) é menor para o Filtro de Kalman.

4.1.5 Analisando os resultados

Nos itens a seguir, os resultados obtidos serão discutidos, tendo seu impacto analisado na dimensão do PNLD no Estado de Alagoas. Também serão feitas considerações sobre o tempo de execução.

Discussão e impacto dos resultados

A análise dos resultados demonstra que o Filtro de Kalman, quando configurado com um número excessivo de iterações, tende a sobreajustar aos dados de treinamento. Isso resulta em predições menos precisas e menos confiáveis, como evidenciado pelo aumento das métricas de erro e a redução do R^2.

Para alcançar um equilíbrio entre a precisão e a capacidade de generalização, é crucial selecionar um número adequado de iterações que permita ao modelo capturar as tendências nos dados sem se ajustar demais aos ruídos e anomalias. Pelos experimentos, pode-se notar que esse número de iterações ideal é normalmente 0 ou 1.

Ainda que a diferença tenha sido pequena, comparando uma situação em que seria usada a Suavização Exponencial ao invés do Filtro de Kalman para predizer dois anos à frente (que é a condição em que a predição é executada nos casos reais), temos uma diferença no MAE de 0,35.

Em um cenário hipotético, considerando apenas o estado de Alagoas e que os valores de erro médio absoluto se mantenham, para o total de mais de 17 mil turmas presentes no censo escolar de 2023, o erro na predição se refletiria em uma diferença de cerca de **6 mil alunos**.

Ainda hipoteticamente, considerando que o erro resultasse sempre em alunos a mais e que um aluno precise utilizar livros em 6 disciplinas, com um custo médio de 5 reais por livro, teríamos uma diferença de aproximadamente **36 mil livros** ou **R$ 180.000,00**, somente mudando o método de predição utilizado, conforme demonstrado na simulação abaixo:

$$\begin{aligned}
\text{Erro por turma} &= 0.35 \\
\text{Total de turmas} &= 17\,000 \\
\text{Quantidade de alunos} &= 17\,000 \times 0.35 = 6\,000 \\
\text{Quantidade de livros} &= 6\,000 \times 6 \text{ (livros por aluno)} = 36\,000 \\
\text{Custo total} &= 36\,000 \times 5 \text{ (valor unitário por livro)} = \mathbf{R\$\ 180\,000}
\end{aligned}$$

Considerando por fim o cenário oposto, em que o erro fosse sempre para menos, prevendo um número de alunos menor que o real, seriam **36 mil** livros a menos, que deixariam de atender mais de **6 mil alunos** no estado.

Este exercício, ainda que considerando premissas muito específicas é interessante para entender a dimensão do Programa Nacional do Livro Didático e o impacto que qualquer melhoria na predição, ainda que pequena, tem sobre a educação mesmo em um estado apenas.

Na Tabela 4.10, estão os resultados do teste de Wilcoxon, aplicados aos resultados do experimento com o Filtro de Kalman em relação a cada baseline, para predições 1 ano e 2 anos à frente.

Tabela 4.10: Resultados do Teste de Wilcoxon em relação aos dados das escolas de Alagoas

Período	Baseline	W-Statistic	P-Value
1 ano à frente	Média	1456274.5	1.2494e-62
1 ano à frente	Último ano	1919948.0	0.7361
1 ano à frente	Regressão Linear	1499765.5	8.6419e-45
1 ano à frente	Suavização exponencial	961299.0	7.2514e-39
2 anos à frente	Média	1094596.5	2.7333e-133
2 anos à frente	Último ano	1563630.5	3.9306e-35
2 anos à frente	Regressão Linear	1766591.0	7.5313e-17
2 anos à frente	Suavização exponencial	221611.0	0.0

Fonte: Elaborado pelo autor.

Interpretando os resultados da tabela, podemos notar:

1 ano à frente

- **Baseline Média:** O p-valor extremamente baixo, $p = 1.2494e-62$, e o valor de $W = 1456274.5$ indicam que há uma diferença estatisticamente significativa entre os resultados do Filtro de Kalman e o baseline da média. Isso sugere que o Filtro de Kalman tem um desempenho diferente da média histórica ao prever o número de alunos.

- **Baseline Último ano:** O p-valor alto, $p = 0.7361$, e o valor de $W = 1919948.0$ indicam que não há diferença estatisticamente significativa entre os resultados do Filtro de Kalman e o baseline do último ano. Isso sugere que o Filtro de Kalman tem um desempenho semelhante ao repetir o valor do último ano para a predição.

- **Baseline Regressão Linear:** O p-valor extremamente baixo, $p = 8.6419e-45$,e o valor de $W = 1499765.5$ indicam uma diferença estatisticamente significativa entre os resultados do Filtro de Kalman e a regressão linear. Isso sugere que o Filtro de

Kalman tem um desempenho diferente ao prever o número de alunos comparado à regressão linear.

- **Baseline Suavização Exponencial:** O p-valor muito baixo, $p = 7.2514e-39$, e o valor de $W = 961299.0$ indicam uma diferença significativa entre o Filtro de Kalman e a suavização exponencial, sugerindo que o desempenho do Filtro de Kalman é diferente neste caso.

2 anos à frente

- **Baseline Média:** O p-valor extremamente baixo, $p = 2.7333e-133$, e o valor de $W = 1094596.5$ sugerem uma diferença significativa entre os resultados do Filtro de Kalman e o baseline da média, indicando que o Filtro de Kalman tem um desempenho diferente da média histórica para predições de dois anos à frente.

- **Baseline Último ano:** O p-valor muito baixo, $p = 3.9306e-35$, e o valor de $W = 1563630.5$ indicam uma diferença significativa entre o Filtro de Kalman e o baseline do último ano, sugerindo que o Filtro de Kalman tem um desempenho diferente ao prever para dois anos à frente.

- **Baseline Regressão Linear:** O p-valor baixo, $p = 7.5313e-17$, e o valor de $W = 1766591.0$ indicam uma diferença significativa entre o Filtro de Kalman e a regressão linear para predições de dois anos, sugerindo que o Filtro de Kalman apresenta resultados diferentes.

- **Baseline Suavização Exponencial:** O p-valor zero, $p = 0.0$, e o valor de $W = 221611.0$ indicam uma diferença altamente significativa entre o Filtro de Kalman e a suavização exponencial, sugerindo que os dois métodos têm desempenhos muito diferentes para predições de dois anos à frente.

Os resultados do teste de Wilcoxon mostram que, para a maioria dos baselines, há uma diferença estatisticamente significativa entre os resultados do Filtro de Kalman e os diferentes baselines, tanto para predições de um ano quanto de dois anos à frente. Isso indica que o Filtro de Kalman tem um desempenho diferente dos métodos tradicionais de baseline. A exceção é o baseline do último ano para a predição de um ano à frente, onde os resultados não mostraram diferença significativa, indicando um desempenho semelhante entre os dois métodos.

Nos casos em que o Filtro de Kalman foi superior, como evidenciado pelos menores valores de MAE, MSE e RMSE, o teste de Wilcoxon comprova que a melhoria é estatisticamente significativa. Isso significa que o Filtro de Kalman não apenas apresentou melhores métricas de desempenho, mas esses resultados também são robustos e confiáveis. Em outras palavras, o Filtro de Kalman oferece predições mais precisas e eficazes

em comparação com os baselines definidos, validando sua superioridade na predição do número de alunos.

Vale ressaltar que o Filtro de Kalman utilizado nesta análise é o mais simples. Outros tipos de Filtros de Kalman, como o Filtro de Kalman Estendido (EKF) e o Filtro de Kalman Unscented (UKF), ou ajustes na forma de aplicar o método, podem apresentar resultados ainda mais significativos. A utilização dessas variações mais sofisticadas pode potencialmente melhorar ainda mais a precisão e a robustez das predições.

Tempo de execução

Os tempos de execução apresentados na Tabela 4.11 mostram a eficiência computacional do Filtro de Kalman na predição do número de alunos, na máquina utilizada para os experimentos [1]:

Tabela 4.11: Tempos de Execução para Predição de 1 e 2 Anos à Frente no estado de Alagoas

Descrição	**Tempo (s)**
Predizendo 1 ano à frente	
Execução com 0 iterações	7.26
Execuções de 0 a 9 iterações	246.39
Predizendo 2 anos à frente	
Execução com 0 iterações	7.55
Execuções de 0 a 9 iterações	236.90

Fonte: Elaborado pelo autor.

Observa-se que, para predições de um ano à frente, a execução com 0 iterações leva aproximadamente 7,26 segundos, enquanto a execução completa, variando de 0 a 9 iterações, requer cerca de 246,39 segundos. A predição para dois anos à frente traz resultados similares, isso considerando as cerca de 3 mil turmas processadas no experimento.

Quando escalonado para um conjunto de dados maior, por exemplo, as 950 mil turmas presentes no censo escolar de 2023, o tempo de execução para a predição com 0 iterações seria de cerca de quarenta minutos na máquina utilizada, que é um computador doméstico, o que representa um grande benefício em termos de eficiência e viabilidade prática.

Além disso, a diferença de tempo entre a execução com 0 iterações e a execução completa (0 a 9 iterações) destaca a flexibilidade do método. Usuários com restrições de tempo ou recursos podem optar por realizar predições com menos iterações para obter resultados mais rapidamente, embora com uma possível redução na precisão. Por outro lado, para cenários onde a precisão é crucial, a execução completa de iterações fornece uma solução mais robusta, embora com um tempo de execução maior.

[1] MacBook Pro - Apple, CPU ARM M3 Pro, de 11 núcleos e 18 GB de memória RAM

Por fim, podemos apontar que o Filtro de Kalman é uma ferramenta eficaz e eficiente para a predição do número de alunos, oferecendo um equilíbrio entre precisão e tempo de execução que pode ser ajustado conforme as necessidades e recursos disponíveis.

4.2 Analisando dados de escolas da região Nordeste

A seção dedicada à análise dos dados das escolas dos estados da região Nordeste do Brasil, assim como a anterior, tem como objetivo apresentar os processos e resultados das predições realizadas, estando subdividida de forma análoga àquela referente ao estado de Alagoas. A região foi escolhida para validar a eficácia do Filtro de Kalman considerando dados mais diversos, mas ainda em uma região que possui desafios educacionais e que pode ser potencialmente beneficiada por uma predição mais precisa do número de alunos.

4.2.1 Limpeza e Pré-processamento dos Dados

O mesmo pré-processamento feito nos dados do estado de Alagoas foi feito naqueles referentes à região Nordeste, mantendo apenas turmas de escolas que participaram do censo escolar do ano de 2023, porém, com 13 ou mais registros ao longo dos anos. A escolha de um limite inferior de registros mais elevado na série temporal teve a finalidade de trazer turmas com dados mais completos, mantendo um tamanho de amostra semelhante ao caso anterior.

A Tabela 4.12 traz os quantitativos a cada etapa do pré-processamento. Pode-se destacar um total de turmas participantes do censo escolar 2023 mais significativo, de cerca de 60.300 turmas, sendo que a quantidade de turmas que foram consideradas para o experimento, que atendiam os critérios após o pré-processamento, foi de 3.095 turmas, neste caso correspondente a cerca de 5% do total possível.

Tabela 4.12: Análise do arquivo original com dados relativos à quantidade de alunos das escolas da região Nordeste.

Descrição	**Quantidade**
Total de registros no arquivo original	2985045
Total de escolas no arquivo original	60380
Total de registros das turmas que participaram do censo de 2023	1973420
Total de turmas que participaram do censo de 2023	322901
Total de registros das turmas que participaram do censo de 2023 e têm 13 ou mais registros	40928
Total de turmas que participaram do censo de 2023 e têm 13 ou mais registros	3095

Fonte: Elaborado pelo autor.

Após a filtragem, os dados também foram reorganizados de maneira a criar uma série temporal em cada linha, correspondendo a uma turma, além de acrescentar um contador e uma variável indicando se existem interrupções ou não na série.

4.2.2 Análise Exploratória

Analisando os dados após o tratamento, buscando a identificação de padrões e dados discrepantes que possam influenciar as etapas subsequentes, iniciou-se pela contagem do número de escolas que possuem diferentes quantidades de registros em suas séries temporais. A Tabela 4.13 mostra a contagem de escolas com determinada quantidade de registros na série temporal das escolas da região Nordeste.

Tabela 4.13: Contagem de escolas com determinada quantidade de registros na série temporal das escolas da região Nordeste.

Quantidade	**Número de Escolas**
13	2495
14	516
15	75
16	9

Fonte: Elaborado pelo autor.

Observa-se que a maioria das escolas tem exatamente 13 registros na série temporal, totalizando 2495 escolas. À medida que o número de registros aumenta, a quantidade de escolas diminui, com nove escolas possuindo 16 registros (máximo no intervalo estudado). Mantendo a tendência das escolas do estado de Alagoas, temos uma concentração de dados com séries temporais mais curtas, mas ainda com um tamanho significativo.

Analisando a média e desvio padrão dos dados, que consta na Tabela 4.14, temos uma visão geral da variabilidade e do comportamento central dos dados.

Tabela 4.14: Estatísticas dos Dados das escolas da região Nordeste (média entre os anos).

Estatística	**Valor**
Média ao longo dos anos	61.06
Mediana (média entre turmas)	39.75
Primeiro Quartil (média entre turmas)	21.38
Terceiro Quartil (média entre turmas)	75.83
Desvio Padrão médio por turma	18.29
Mínimo Geral	1.0
Máximo Geral	1137.0

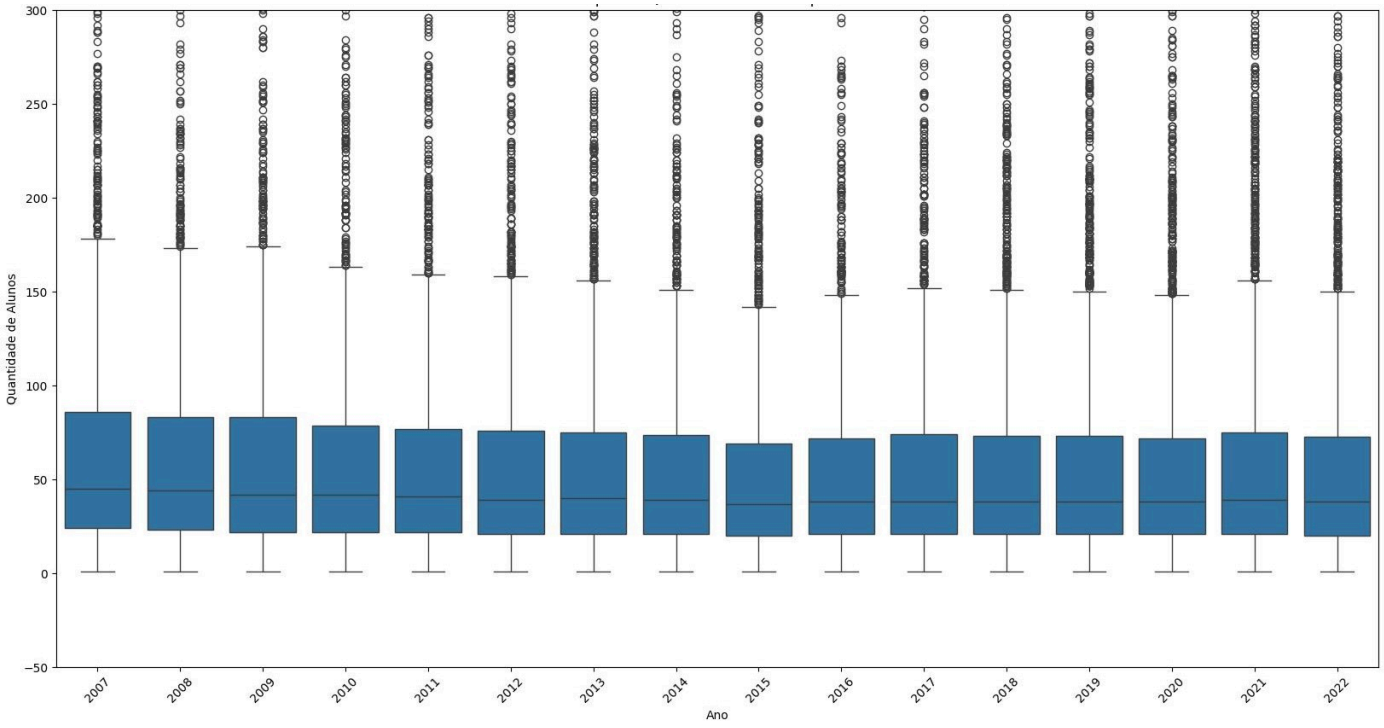
Fonte: Elaborado pelo autor.

A média de alunos por turma é de cerca de 61 alunos, enquanto o desvio padrão médio é de cerca de 18, mantendo a tendência já apresentada de variação relevante no número de alunos de uma turma entre diferentes anos.

A mediana gira em torno de 40 alunos por turma, ou seja, metade das turmas (em média entre os anos) tem essa quantidade ou menos alunos, sendo um mínimo de 1 aluno e máximo de 1137 na amostra analisada, também sinalizando agregação de dados de diferentes turmas em uma só.

O boxplot apresentado na Figura 4.7 exibe a distribuição da quantidade de alunos por turma ao longo dos anos para as escolas da região Nordeste.

Figura 4.7: Boxplot dos dados da região Nordeste.



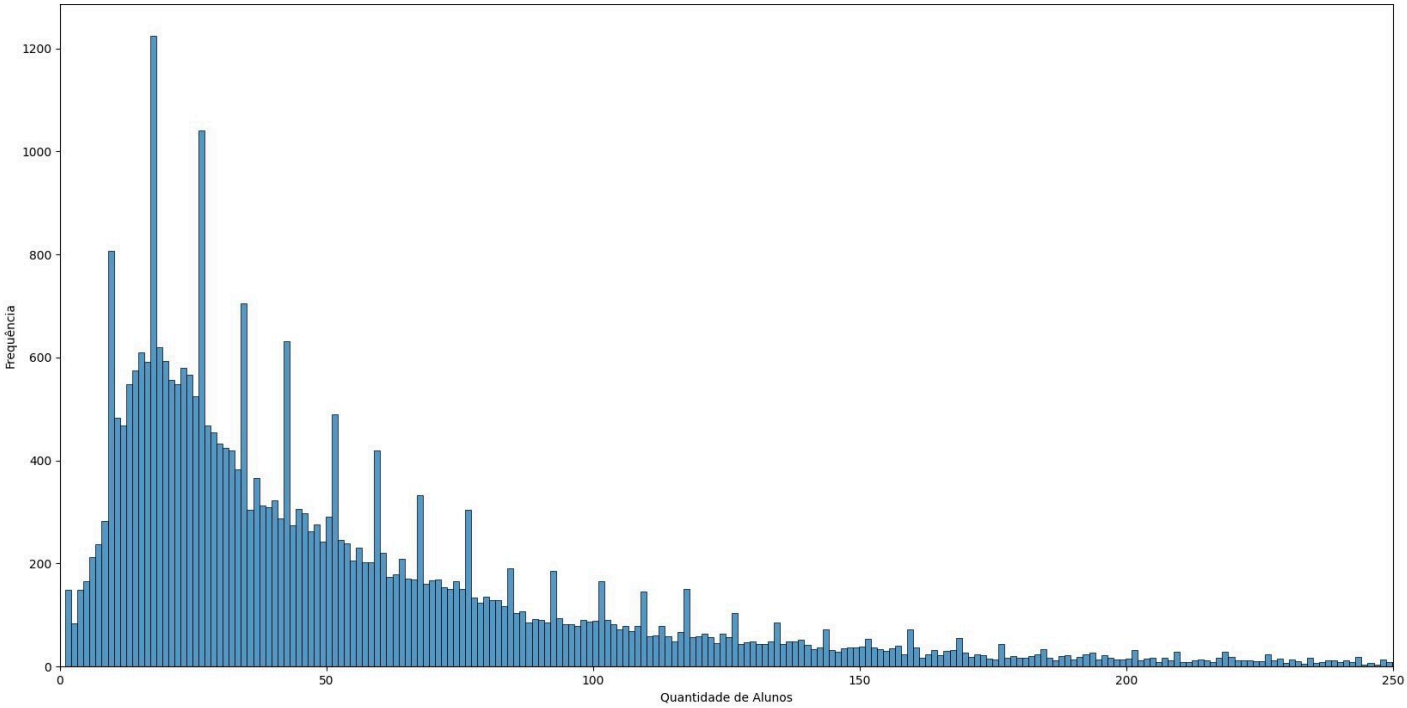
Fonte: Elaborado pelo autor.

A análise aponta medianas semelhantes ao longo dos anos, em torno do valor de 40 já apontado na Tabela 4.14 tal como o intervalo interquartil, que se mantém ainda mais homogêneo ao longo dos anos do que o que foi visto no caso das escolas do estado de Alagoas. Além disso, em todos os anos, estão presentes muitos outliers.

Aspectos como semelhança entre medianas, tamanho dos intervalos interquartis ao longo dos anos e presença de outliers contribuem de forma positiva, para os dois primeiros itens, e negativa, para o último, replicando o que foi observado na seção anterior.

O histograma apresentado na Figura 4.8 exibe a distribuição da quantidade de alunos por turma ao longo dos anos para as escolas da região Nordeste estudadas:

Figura 4.8: Histograma dos dados da região Nordeste.



Fonte: Elaborado pelo autor.

De maneira muito semelhante ao apresentado na seção anterior, o histograma se apresenta de forma unimodal com distribuição assimétrica à direita e um pico de frequência de turmas entre 20 e 30 alunos, indicando ser a faixa mais comum entre as escolas analisadas. Se observam outliers, assim como no boxplot, e à medida que a quantidade de alunos por turma aumenta, a frequência das turmas diminui drasticamente, o que é sugerido pela distribuição assimétrica.

4.2.3 Predizendo valores para 1 ano à frente

Determinação de Baselines

Os mesmos baselines da seção anterior foram calculados para os dados das escolas da região Nordeste: Média, Repetir o Último Ano, Regressão Linear e Suavização Exponencial. A tabela 4.15 resume as métricas de avaliação desses baselines, considerando uma predição para 1 ano à frente.

Tabela 4.15: Métricas de dos Baselines para escolas da região Nordeste predizendo 1 ano à frente

Baseline	**MAE**	**STD**	**MSE**	**RMSE**	**MAPE**	R^2
Média	16.97	29.99	920.09	30.33	40.23	0.7299
Repetir o último Ano	11.39	19.92	398.10	19.95	26.91	0.8832
Regressão Linear	14.07	25.61	655.73	25.61	30.93	0.8075
Suavização Exponencial	11.18	20.04	404.91	20.12	26.24	0.8812

Fonte: Elaborado pelo autor.

Observando a tabela se nota que o baseline de Suavização Exponencial segue obtendo os melhores resultados gerais, com menor MAE, STD, MSE, RMSE e MAPE, além do maior valor de R^2, seguido pelo baseline de Repetir o Último Ano.

Execução dos experimentos

De forma análoga à seção anterior, os experimentos foram executados variando a quantidade de iterações na execução do Filtro de Kalman, buscando compreender como o número de iterações influencia no desempenho.

A Tabela 4.16 resume os resultados dos experimentos:

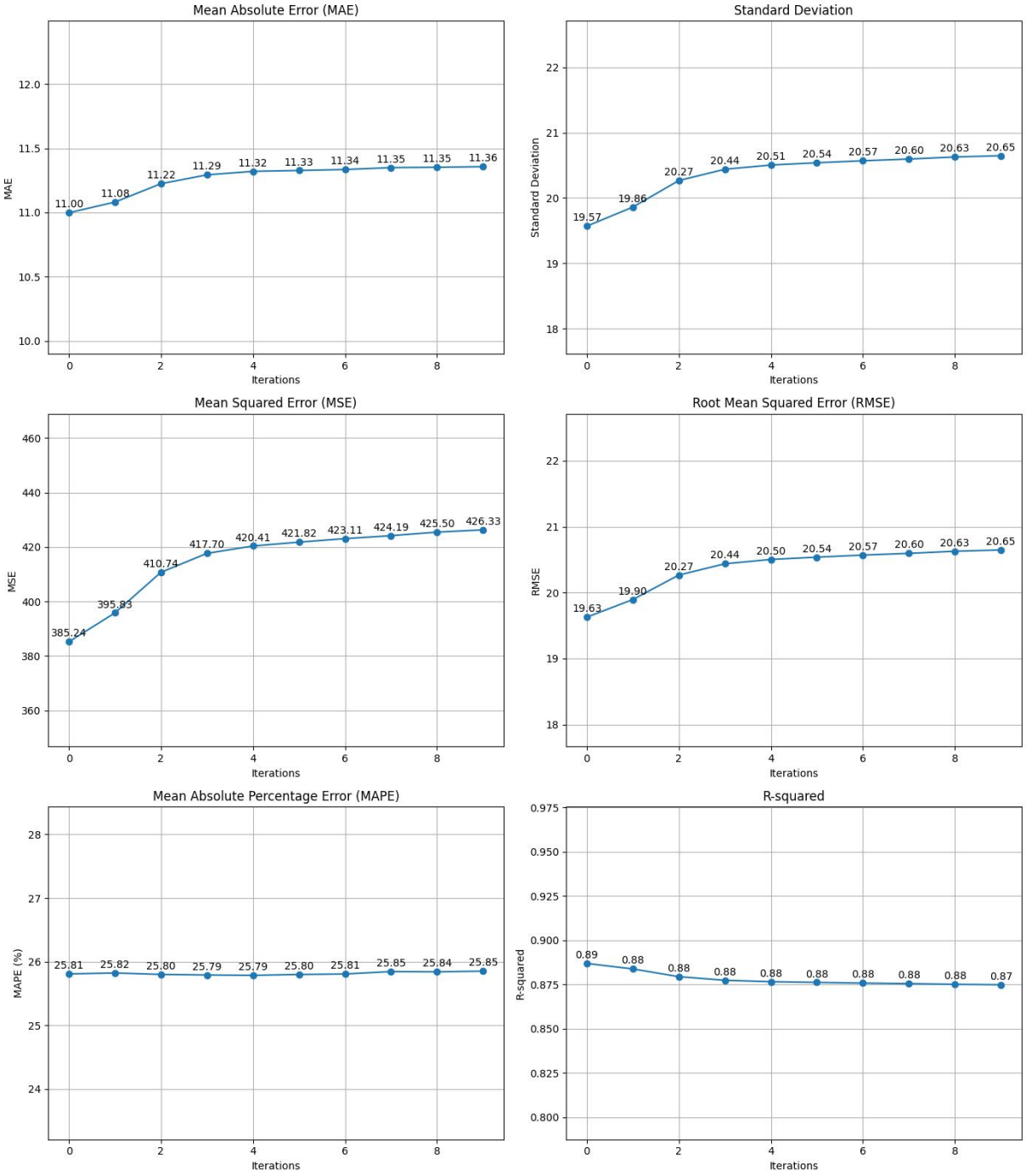
Tabela 4.16: Resultados do Experimento de predição para 1 ano à frente em escolas da região Nordeste com diferentes quantidades de iterações no Filtro de Kalman

Métrica	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
MAE	11.00	11.08	11.22	11.29	11.32	11.33	11.34	11.35	11.35	11.36
STD	19.57	19.86	20.27	20.44	20.51	20.54	20.57	20.60	20.63	20.65
MSE	385.24	395.83	410.74	417.70	420.41	421.82	423.11	424.19	425.50	426.33
RMSE	19.63	19.90	20.27	20.44	20.50	20.54	20.57	20.60	20.63	20.65
MAPE	25.81	25.82	25.80	25.79	25.79	25.80	25.81	25.85	25.84	25.85
R^2	0.8869	0.8838	0.8794	0.8774	0.8766	0.8762	0.8758	0.8755	0.8751	0.8749

Fonte: Elaborado pelo autor.

A Figura 4.9 traz uma visualização gráfica das informações contidas na Tabela 4.16:

Figura 4.9: Gráficos apresentando as métricas das execuções do Filtro de Kalman nos dados da região Nordeste com diferentes quantidades de iterações no Filtro de Kalman predizendo 1 ano à frente.



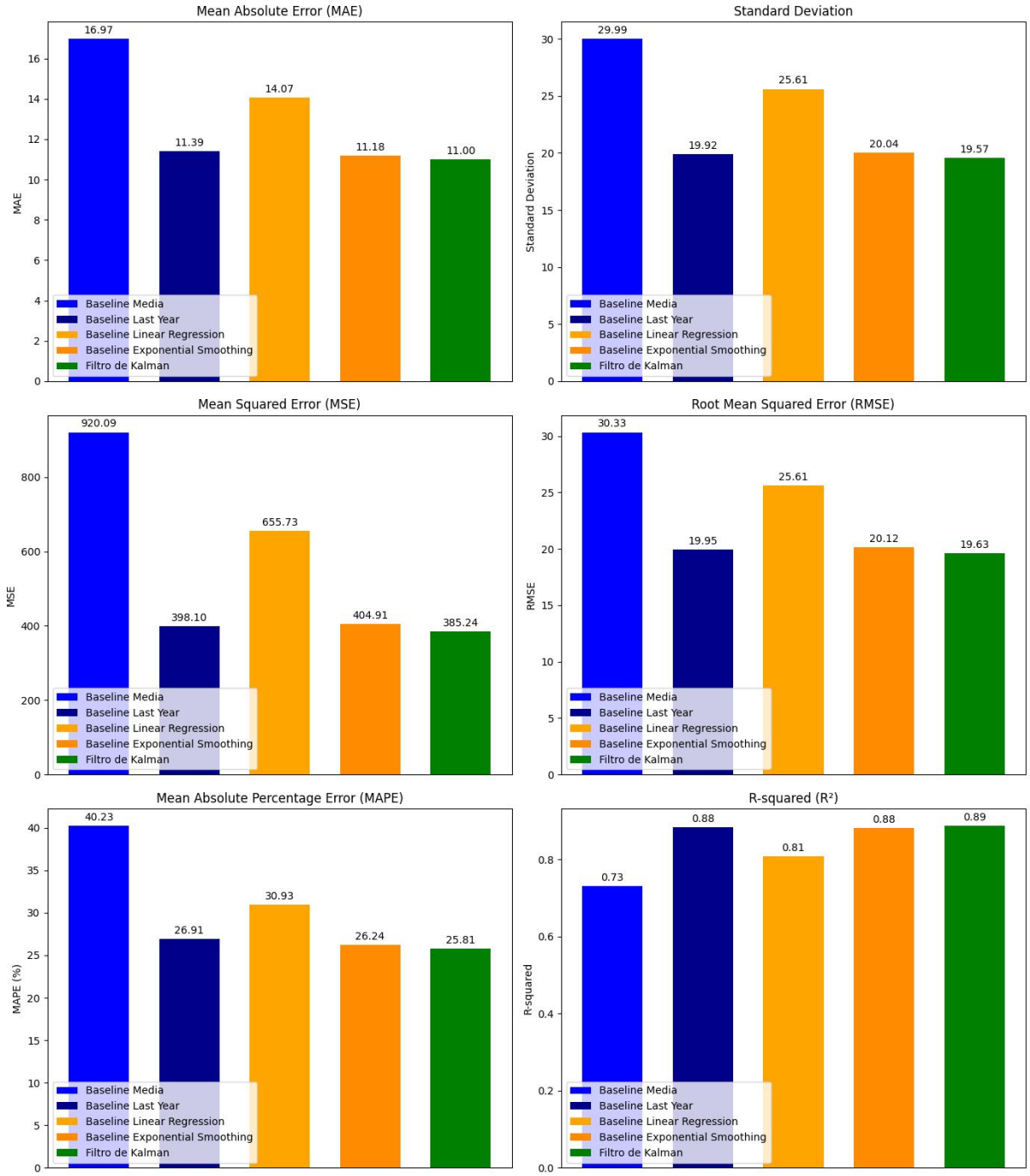
Fonte: Elaborado pelo autor.

Os resultados dos experimentos reforçam a tendência já observada em relação ao impacto das iterações na performance do Filtro de Kalman. O resultado das métricas: MAE, STD, MSE e RMSE aumenta gradualmente conforme o número de iterações aumenta. No

caso da MAPE, se observa uma certa estabilidade, enquanto o R^2 diminui até também chegar a um valor estável.

O gráfico apresentado na Figura 4.10 mostra uma comparação métrica a métrica entre os baselines e o Filtro de Kalman (considerando o melhor resultado do mae).

Figura 4.10: Gráficos apresentando a comparação entre diferentes métricas dos baselines e do Filtro de Kalman (melhor resultado) com dados da região Nordeste predizendo 1 ano à frente.



Fonte: Elaborado pelo autor.

Comparando o desempenho do Filtro de Kalman em relação aos baselines em relação a cada métrica, pode-se observar:

MAE (Mean Absolute Error):

- A suavização exponencial e o Filtro de Kalman apresentaram os menores MAEs, com valores de 11.18 e 11.00 respectivamente, destacando-se como as técnicas mais precisas, seguidas pela repetição do último ano.

STD (Standard Deviation):

- Para este caso, a Repetição do último ano e o Filtro de Kalman mantiveram os menores desvios padrões, 19.92 e 19.57 respectivamente. O maior desvio padrão foi observado para a média dos anos anteriores (29.99), demonstrando uma falta de precisão, conforme já observado.

MSE (Mean Squared Error) e RMSE (Root Mean Squared Error):

- Tanto o MSE quanto o RMSE seguiram um padrão semelhante ao STD, onde a repetição do último ano (398.10 e 19.95) e o Filtro de Kalman (385.24 e 19.63) apresentaram os melhores resultados.

MAPE (Mean Absolute Percentage Error):

- Para esta métrica, o Filtro de Kalman e a suavização exponencial se destacaram com os menores valores de MAPE, 26.24% e 25.81%, respectivamente.

R^2 (R-squared):

- Também nesta métrica o Filtro de Kalman apresentou o melhor valor, ficando 0,01 a frente da suavização exponencial, indicando que ambos têm uma boa capacidade de explicação da variabilidade dos dados.

Em suma, a comparação entre as diferentes técnicas mostra uma unanimidade entre todas as métricas sinalizando que o o Filtro de Kalman é o método mais eficaz para predição neste conjunto de dados. A média dos anos anteriores se mostrou como a mais ineficaz, com as piores métricas entre os métodos abordados, assim como aconteceu com os dados das escolas do estado de Alagoas.

4.2.4 Predizendo valores para 2 anos à frente

Determinação de Baselines

A Tabela 4.17 resume as métricas de desempenho para cada baseline utilizado para predizer os valores para 2 anos à frente, no caso das escolas da região Nordeste:

Tabela 4.17: Métricas dos Baselines para escolas da região Nordeste predizendo 2 anos à frente

Baseline	**MAE**	**STD**	**MSE**	**RMSE**	**MAPE**	R^2
Média	17.99	32.00	1047.40	32.36	42.51	0.6926
Repetir o último Ano	15.38	30.01	908.11	30.13	33.91	0.7335
Regressão Linear	16.61	30.89	955.03	30.90	35.96	0.7197
Suavização Exponencial	13.65	26.01	682.65	26.13	30.99	0.7996

Fonte: Elaborado pelo autor.

Os resultados obtidos seguem um padrão semelhante ao observado em experimentos anteriores. O baseline de suavização exponencial e o de Repetir o último ano apresentaram os melhores desempenhos, com menores valores de MAE, STD, MSE, RMSE e MAPE, além de valores mais altos de R^2. Em contrapartida, a média dos anos anteriores, neste último experimento, se consolida como o pior desempenho entre os baselines utilizados.

Execução dos experimentos

A Tabela 4.18 apresenta os resultados dos experimentos para escolas da região Nordeste ao predizer 2 anos à frente, considerando diferentes quantidades de iterações do Filtro de Kalman:

Tabela 4.18: Resultados do Experimento de predição para 2 anos à frente em escolas da região Nordeste com diferentes quantidades de iterações no Filtro de Kalman

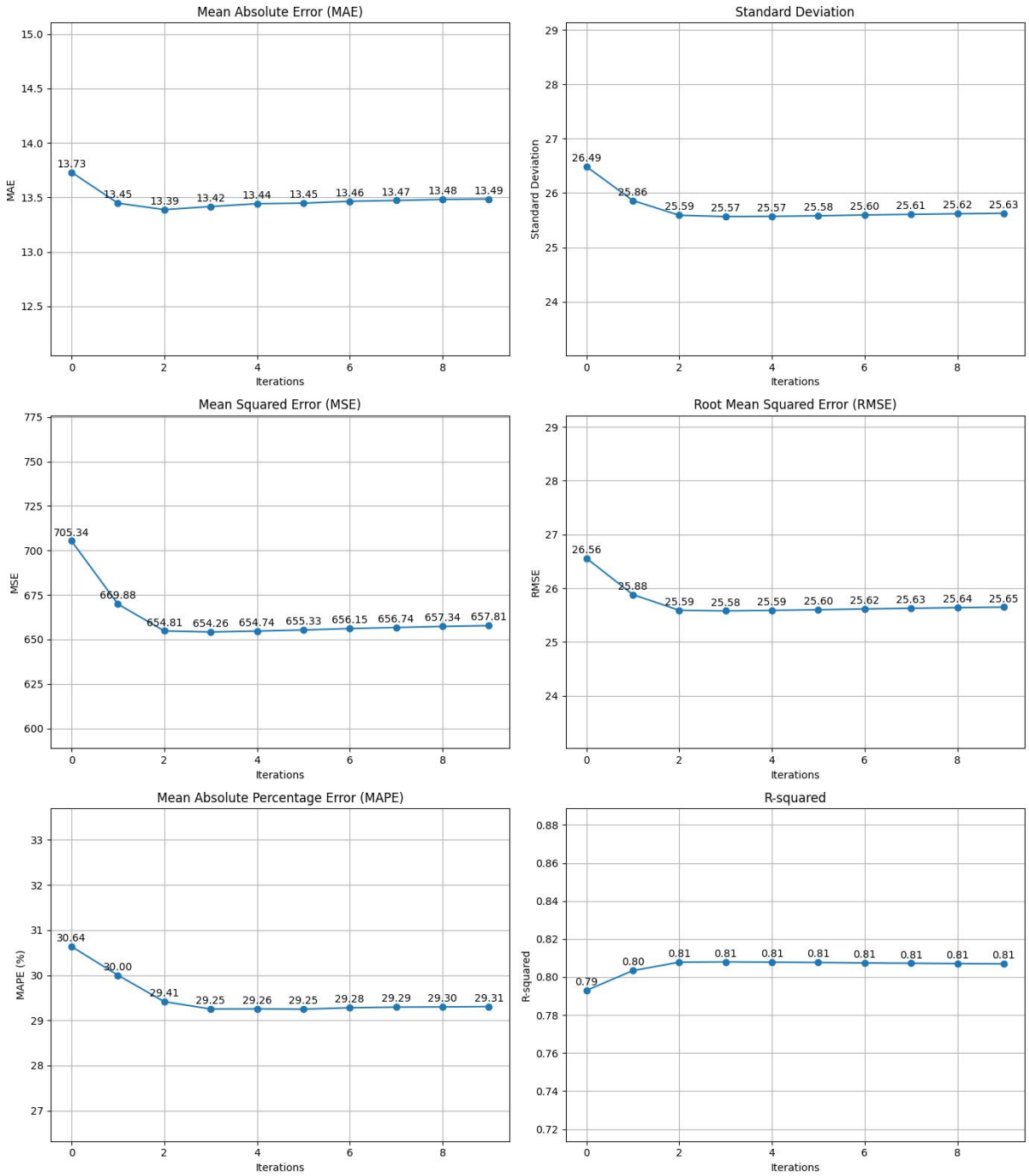
Métrica	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
MAE	13.73	13.45	13.39	13.42	13.44	13.45	13.46	13.47	13.48	13.49
STD	26.49	25.86	25.59	25.57	25.57	25.58	25.60	25.61	25.62	25.63
MSE	705.34	669.88	654.81	654.26	654.74	655.33	656.15	656.74	657.34	657.81
RMSE	26.56	25.88	25.59	25.58	25.59	25.60	25.62	25.63	25.64	25.65
MAPE	30.64	30.00	29.41	29.25	29.26	29.25	29.28	29.29	29.30	29.31
R^2	0.7930	0.8034	0.8078	0.8080	0.8078	0.8077	0.8074	0.8072	0.8071	0.8069

Fonte: Elaborado pelo autor.

Assim como ocorreu na predição para 2 anos à frente no estado de Alagoas, não foi a execução com zero iterações que obteve o melhor resultado, mas sim com duas iterações (considerando o MAE como métrica principal). Além disso, os valores das métricas de erro foram consistentemente maiores do que os encontrados na predição para 1 ano à frente, o que é natural devido à incerteza associada a predições realizadas para um intervalo maior de tempo.

A Figura 4.11 traz uma visualização gráfica das informações contidas na Tabela 4.18.

Figura 4.11: Gráficos apresentando as métricas das execuções do Filtro de Kalman nos dados da região Nordeste com diferentes quantidades de iterações predizendo 2 anos à frente.



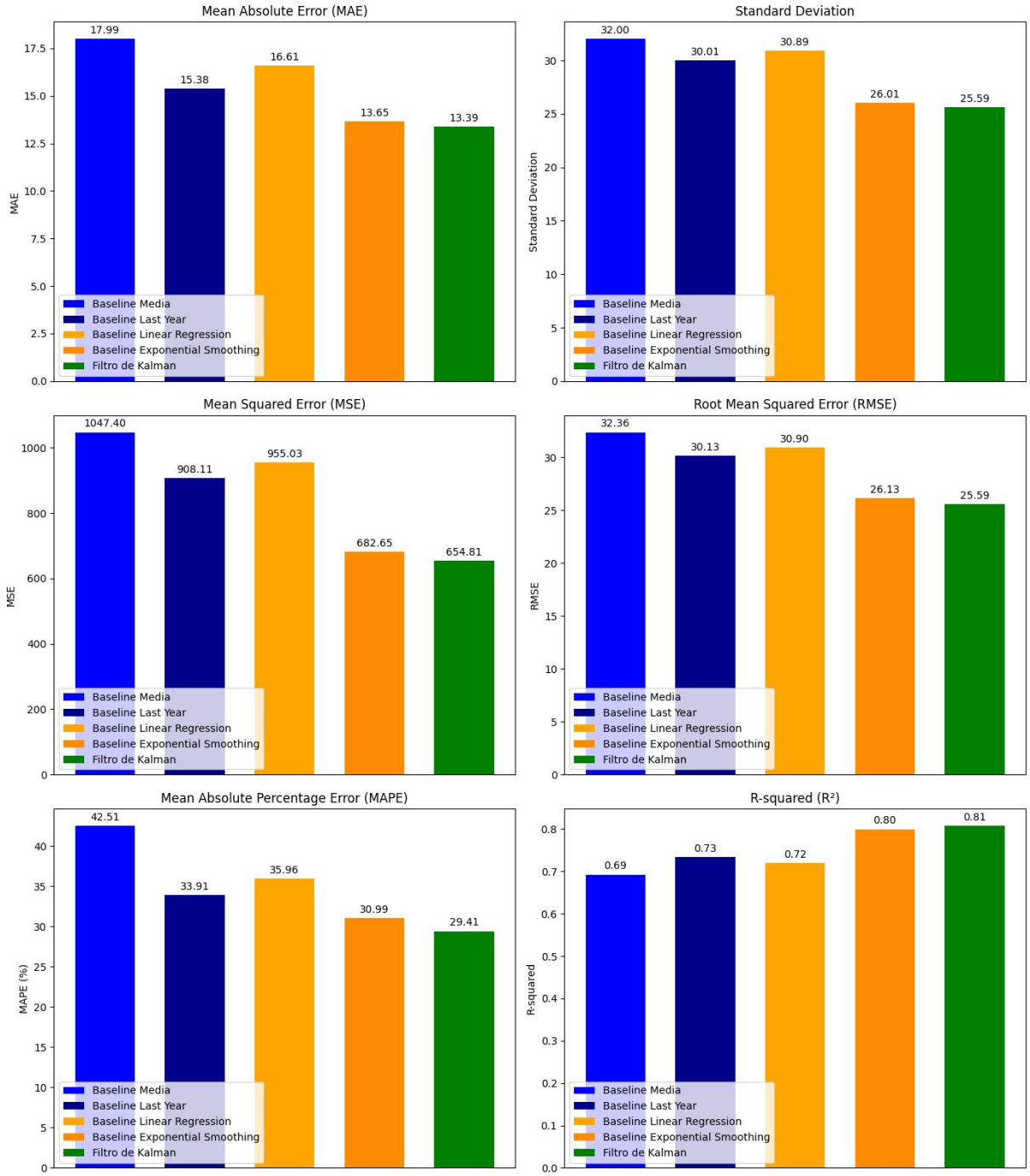
Fonte: Elaborado pelo autor.

Observa-se que, ao longo das primeiras iterações, as métricas apresentam melhorias graduais, estabilizando em seguida, o que não ocorreu na predição para 1 ano à frente. Isso sugere que um maior número de iterações, neste caso, proporciona um melhor ajuste

aos dados e consequente melhora no desempenho das predições.

O gráfico apresentado na Figura 4.12 mostra uma comparação métrica a métrica entre os baselines e o Filtro de Kalman (considerando o melhor resultado do mae).

Figura 4.12: Gráficos apresentando a comparação entre diferentes métricas dos baselines e do Filtro de Kalman (melhor resultado) com dados da região Nordeste predizendo 2 anos à frente.



Fonte: Elaborado pelo autor.

Observa-se que o Filtro de Kalman apresentou resultados ainda melhores em relação

à predição para 1 ano à frente, sendo superior aos baselines em todas as métricas e com maiores margens. O MAE mostra que o Filtro de Kalman teve o menor erro absoluto médio,sendo 0.26 menor que a Suavização Exponencial, indicando predições mais precisas, ainda que por uma pequena margem. O desvio padrão (STD) também foi ligeiramente inferior, indicando maior consistência e menores desvios nos resultados preditos.

4.2.5 Analisando os resultados

De maneira análoga à seção anterior, os resultados obtidos serão discutidos a seguir, sob os aspectos de impacto no PNLD na região Nordeste e tempo de execução.

Discussão e impacto dos resultados

Os resultados desse segundo bloco de experimentos reforçam que o Filtro de Kalman, quando configurado com um número excessivo de iterações, tende a sobreajustar aos dados de treinamento, resultando em predições menos confiáveis, demonstrado pelo aumento das métricas de erro e a redução do R^2.

O equilíbrio entre a precisão e a capacidade de generalização foi percebido ao utilizar 0 a 2 iterações no Filtro de Kalman. Ainda que a diferença tenha sido pequena, comparando uma situação em que seria usada a Suavização Exponencial ao invés do Filtro de Kalman para predizer dois anos à frente, a diferença no MAE foi de 0,26.

Refazendo o exercício da seção anterior para a diferença no MAE obtida e considerando toda a região Nordeste, com a premissa hipotética de que os valores de erro médio absoluto se mantenham, para o total de cerca de 323 mil turmas presentes no censo escolar de 2023, o erro na predição se refletiria em uma diferença de cerca de **84 mil alunos**.

Ainda hipoteticamente, considerando que o erro resultasse sempre em alunos a mais e que um aluno precise utilizar livros em 6 disciplinas, com um custo médio de 5 reais por livro, teríamos uma diferença de aproximadamente **504 mil livros** ou **R$ 2.520.000,00**, somente mudando o método de predição utilizado, conforme demonstrado na simulação abaixo:

$$\begin{aligned}
\text{Erro por turma} &= 0.26 \\
\text{Total de turmas} &= 323\,000 \\
\text{Quantidade de alunos} &= 323\,000 \times 0.26 = 84\,000 \\
\text{Quantidade de livros} &= 84\,000 \times 6 \text{ (livros por aluno)} = 504\,000 \\
\text{Custo total} &= 504\,000 \times 5 \text{ (valor unitário por livro)} = \mathbf{R\$\ 2.520\,000}
\end{aligned}$$

Considerando por fim o cenário oposto, em que o erro fosse sempre para menos, predizendo um número de alunos menor que o real, seriam **504 mil** livros a menos, que

deixariam de atender quase **84 mil alunos** na região.

Através deste exercício, percebe-se o grande impacto que qualquer melhoria na predição, ainda que singela, sobre a educação na região Nordeste do país.

Na Tabela 4.19, estão os resultados do teste de Wilcoxon, aplicados aos resultados do experimento com o Filtro de Kalman em relação a cada baseline, para predições 1 ano e 2 anos à frente.

Tabela 4.19: Resultados do Teste de Wilcoxon em relação aos dados das escolas do Nordeste

Período	Baseline	W-Statistic	P-Value
1 ano à frente	Média	1701892.0	6.1388e-25
1 ano à frente	Último ano	1687722.0	7.8059e-05
1 ano à frente	Regressão Linear	1642815.5	2.8088e-24
1 ano à frente	Suavização exponencial	903699.0	3.8152e-17
2 anos à frente	Média	1260994.0	9.1330e-85
2 anos à frente	Último ano	1364553.0	6.5242e-57
2 anos à frente	Regressão Linear	1974132.0	6.5674e-02
2 anos à frente	Suavização exponencial	266417.5	1.9414e-292

Fonte: Elaborado pelo autor.

Interpretando os resultados da tabela, podemos notar:

1 ano à frente

- **Baseline Média:** O p-valor extremamente baixo, $p = 6.1388e-25$, e o valor de $W = 1701892.0$ indicam que há uma diferença estatisticamente significativa entre os resultados do Filtro de Kalman e o baseline da média. Isso sugere que o Filtro de Kalman tem um desempenho diferente da média histórica ao prever o número de alunos.

- **Baseline Último ano:** O p-valor baixo, $p = 7.8059e-05$, e o valor de $W = 1687722.0$ indicam que há uma diferença estatisticamente significativa entre os resultados do Filtro de Kalman e o baseline do último ano. Isso sugere que o Filtro de Kalman tem um desempenho diferente de repetir o valor do último ano para a predição.

- **Baseline Regressão Linear:** O p-valor extremamente baixo, $p = 2.8088e-24$, e o valor de $W = 1642815.5$ indicam uma diferença estatisticamente significativa entre os resultados do Filtro de Kalman e a regressão linear. Isso sugere que o Filtro de Kalman tem um desempenho diferente ao prever o número de alunos comparado à regressão linear.

- **Baseline Suavização Exponencial:** O p-valor muito baixo, $p = 3.8152e-17$, e o valor de $W = 903699.0$ indicam uma diferença significativa entre o Filtro de Kalman e a suavização exponencial, sugerindo que o desempenho do Filtro de Kalman é diferente neste caso.

2 anos à frente

- **Baseline Média:** O p-valor extremamente baixo, $p = 9.1330e-85$, e o valor de $W = 1260994.0$ sugerem uma diferença significativa entre os resultados do Filtro de Kalman e o baseline da média, indicando que o Filtro de Kalman tem um desempenho diferente da média histórica para predições de dois anos à frente.

- **Baseline Último ano:** O p-valor muito baixo, $p = 6.5242e-57$, e o valor de $W = 1364553.0$ indicam uma diferença significativa entre o Filtro de Kalman e o baseline do último ano, sugerindo que o Filtro de Kalman tem um desempenho diferente ao prever para dois anos à frente.

- **Baseline Regressão Linear:** O p-valor moderado, $p = 6.5674e-02$, e o valor de $W = 1974132.0$ indicam que não há diferença estatisticamente significativa entre o Filtro de Kalman e a regressão linear para predições de dois anos. Isso sugere que o Filtro de Kalman tem um desempenho semelhante à regressão linear nesse caso.

- **Baseline Suavização Exponencial:** O p-valor extremamente baixo, $p = 1.9414e-292$, e o valor de $W = 266417.5$ indicam uma diferença altamente significativa entre o Filtro de Kalman e a suavização exponencial, sugerindo que os dois métodos têm desempenhos muito diferentes para predições de dois anos à frente.

Os resultados do teste de Wilcoxon mostram que, para a maioria dos baselines, há uma diferença estatisticamente significativa entre os resultados do Filtro de Kalman e os diferentes baselines, tanto para predições de um ano quanto de dois anos à frente. Isso indica que o Filtro de Kalman tem um desempenho diferente dos métodos tradicionais de baseline. A exceção é o baseline de regressão linear para a predição de dois anos à frente, onde os resultados não mostraram diferença significativa, indicando um desempenho semelhante entre os dois métodos.

Nos casos em que o Filtro de Kalman foi superior, como evidenciado pelos menores valores de MAE, MSE e RMSE, o teste de Wilcoxon comprova mais uma vez que a melhoria é estatisticamente significativa. Isso significa que o Filtro de Kalman não apenas apresentou melhores métricas de desempenho, mas esses resultados também são robustos e confiáveis.

Tempo de execução

Os tempos de execução apresentados na Tabela 4.20 mostram a eficiência computacional do Filtro de Kalman na predição do número de alunos. A máquina utilizada para execução dos experimentos foi a mesma dos experimentos anteriores[2]:

Tabela 4.20: Tempos de Execução para Predição de 1 e 2 Anos à Frente nos estados do Nordeste

Descrição	**Tempo (s)**
Predizendo 1 ano à frente	
Execução com 0 iterações	7.67
Execuções de 0 a 9 iterações	261.04
Predizendo 2 anos à frente	
Execução com 0 iterações	7.21
Execuções de 0 a 9 iterações	249.49

Fonte: Elaborado pelo autor.

Observa-se que, para predições de um ano à frente, a execução com 0 iterações leva aproximadamente 7,67 segundos, enquanto a execução completa, variando de 0 a 9 iterações, requer cerca de 261,04 segundos. A predição para dois anos à frente traz resultados similares, isso considerando as cerca de 3 mil turmas processadas no experimento.

Quando escalonado para um conjunto de dados maior, por exemplo, as 950 mil turmas presentes no censo escolar de 2023, o tempo de execução para a predição com 0 iterações continuaria na casa dos quarenta minutos na máquina utilizada, que é um computador doméstico, o que representa um grande benefício em termos de eficiência e viabilidade prática.

[2]MacBook Pro - Apple, CPU ARM M3 Pro, de 11 núcleos e 18 GB de memória RAM

Capítulo 5

Conclusão

Neste trabalho, investigou-se a aplicação do Filtro de Kalman para a predição do número de alunos em escolas públicas do estado de Alagoas e da região Nordeste do Brasil, no contexto do Programa Nacional do Livro Didático (PNLD). O objetivo foi explorar o potencial do Filtro de Kalman como um estimador linear estatístico para aprimorar a precisão das previsões de matrículas escolares, abordando os desafios logísticos e de coleta de dados em um país de dimensões continentais como o Brasil.

Inicialmente, foi realizada uma revisão da literatura sobre métodos de predição de séries temporais e suas aplicações na educação. A importância do PNLD e a necessidade de previsões precisas para otimizar a distribuição de materiais didáticos foram alguns dos temas abordados, destacando os desafios enfrentados devido à variabilidade das matrículas escolares e às limitações dos métodos tradicionais de previsão.

A metodologia incluiu a coleta e o pré-processamento dos dados do Censo Escolar, seguidos pela análise estatística segmentada por etapa de ensino. Foram estabelecidos baselines utilizando métodos tradicionais como média e regressão linear para comparação com os resultados obtidos pelo Filtro de Kalman.

Os resultados demonstraram que o Filtro de Kalman apresentou melhorias na acurácia das predições de matrículas em comparação com os métodos testados. Do ponto de vista prático, o uso do Filtro de Kalman poderia melhorar a eficiência na alocação de recursos educacionais, traduzindo-se em economias substanciais e melhor distribuição de materiais didáticos, com potencial para beneficiar diretamente milhares de alunos e professores.

Além disso, a flexibilidade do Filtro de Kalman permite sua adaptação para diferentes regiões e contextos educacionais, ampliando o impacto de sua aplicação. Apesar de utilizar a versão mais simples do filtro, os resultados obtidos são promissores, indicando que variações do método podem trazer ganhos de desempenho.

Por fim, todos os objetivos da pesquisa foram alcançados com sucesso, validando a metodologia proposta e fornecendo uma base sólida para estudos futuros. Este trabalho destaca o potencial do Filtro de Kalman como uma ferramenta para solução de problemas no contexto educacional brasileiro.

5.1 Trabalhos Futuros

Com base nos resultados obtidos, várias direções para trabalhos futuros emergem, visando aprimorar ainda mais a precisão e a aplicabilidade do Filtro de Kalman na predição do alunado. Três das principais são:

5.1.1 Avaliar a variação ao invés do valor bruto

Uma abordagem potencialmente promissora é testar o modelo utilizando as variações anuais no número de alunos, ao invés dos valores brutos. Isso envolve um pré-processamento dos dados para calcular as variações de um ano para o outro (por exemplo, se uma turma tinha 50 alunos em 2020 e 47 em 2021, a variação é -3).

Em seguida, seria aplicado o Filtro de Kalman sobre esta série de variações e, posteriormente, transformado novamente em quantidades de alunos (somando a variação predita ao último valor conhecido). Esta abordagem pode melhorar a capacidade do modelo de capturar tendências e flutuações.

5.1.2 Evolução de turmas específicas ao longo do tempo

Outra abordagem interessante seria avaliar os dados das turmas não por turma e ano isoladamente, mas sim acompanhando a evolução de um mesmo conjunto de alunos ao longo do tempo. Por exemplo, ao predizer a matrícula de 2021 para o segundo ano, que é justamente a turma do primeiro ano em 2020, utilizar a série temporal dos alunos que compunham o primeiro ano em 2020.

Aplicar o Filtro de Kalman nesta série temporal pode proporcionar uma análise com maior taxa de sucesso, uma vez que é o mesmo conjunto de pessoas ao longo dos anos, flutuando apenas nos casos de entrada ou saída de alunos. A forma que foi testada neste trabalho calcula predições com base em séries temporais em que cada valor corresponde a um conjunto diferente de pessoas, o que traz

Um elemento central para essa abordagem é a matriz de transição A, que no Filtro de Kalman é responsável por modelar a evolução do estado ao longo do tempo. Neste caso, A descreve como o número de alunos de um ano é projetado para o ano seguinte, considerando as alterações esperadas devido às entradas e saídas de alunos. Trabalhar com o mesmo conjunto de alunos ao longo do tempo permite que a matriz A capture de forma mais fiel a dinâmica da turma, reduzindo incertezas no modelo e mitigando o impacto de ruídos ou valores discrepantes.

A abordagem sugerida pode enfrentar limitações devido à quantidade insuficiente de dados para algumas turmas. Outrossim, caso validada, essa abordagem só poderá ser aplicada a turmas específicas com histórico consolidado, o que restringe sua aplicabilidade.

5.1.3 Uso de Machine Learning

Uma abordagem baseada em Aprendizado de Máquina apresenta um grande potencial para aprimorar as predições relacionadas ao número de alunos em anos futuros, especialmente devido à vasta quantidade de dados disponibilizados anualmente nos microdados do Censo Escolar, que incluem mais de 300 colunas. Esses dados abrangem informações demográficas, estruturais, socioeconômicas e de desempenho das escolas, oferecendo uma rica base de conhecimento para explorar padrões complexos e não lineares.

Ao utilizar algoritmos tradicionais, como regressão, árvores de decisão e redes neurais, é possível modelar relações subjacentes nos dados, identificando tendências e fatores que impactam as matrículas. Além disso, o uso de heurísticas pode ser incorporado para priorizar variáveis relevantes e reduzir a dimensionalidade, simplificando o treinamento e aumentando a interpretabilidade do modelo. Técnicas como seleção de características (feature selection) e análise de importância de variáveis podem ser aplicadas para garantir que o modelo seja eficiente e focado nos aspectos mais significativos. O uso desse tipo de abordagem pode trazer resultados muito interessantes para a predição do alunado em anos futuros.

5.2 Considerações Finais

Este trabalho demonstrou que a aplicação do Filtro de Kalman pode trazer avanços significativos na predição de matrículas escolares, contribuindo para uma gestão mais eficiente dos recursos educacionais. A confirmação das hipóteses levantadas reforça que, respeitadas as premissas de linearidade e distribuição gaussiana do ruído, o filtro pode reduzir a incerteza associada às estimativas e produzir resultados mais precisos do que métodos tradicionais.

A continuidade deste estudo, através das propostas para trabalhos futuros, promete expandir ainda mais as aplicações do Filtro de Kalman, ou mesmo de outras técnicas. A garantia de otimalidade e eficiência computacional do Filtro são grandes atrativos para continuar evoluindo modelos que o utilizam, tendo em mente a incerteza inerente a estimativas que utilizam a matriz de covariância P.

Por fim, ao unir tecnologia computacional com políticas públicas bem estruturadas, pavimenta-se o caminho para um sistema educacional mais adaptável e resiliente. Esse modelo, que combina precisão técnica com consciência de incertezas e planejamento estratégico, tem o potencial de não apenas transformar a educação no Brasil, mas também de servir como referência para sistemas educacionais em diferentes contextos ao redor do mundo.

Bibliografia

[Abrelivros, 2020] Abrelivros (2020). Como funciona o pnld - do edital à sala de aula. Disponível em: https://abrelivros.org.br/site/wp-content/uploads/2020/10/2020_Como_funciona_o_PNLD.pdf. Acesso em: 4 jul. 2024.

[Abrelivros, s d] Abrelivros (s. d.). Pnld - programa nacional do livro e do material didático. Disponível em: https://abrelivros.org.br/site/pnld/. Acesso em: 4 jul. 2024.

[Agência Brasil, 2024] Agência Brasil (2024). Brasil celebra o dia nacional do livro didático nesta terça-feira. Disponível em: https://agenciagov.ebc.com.br/noticias/202402/brasil-celebra-o-dia-nacional-do-livro-didatico-nesta-terca-feira. Acesso em: 4 jul. 2024.

[Box et al., 2015] Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2015). *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken.

[BRASIL, 1988] BRASIL (1988). Constituição da república federativa do brasil de 1988. Brasília, DF: Presidente da República. Art. 6º, Art. 205. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 4 jul. 2024.

[BRASIL, 2017] BRASIL (2017). Decreto nº 9.099, de 18 de julho de 2017. Dispõe sobre o Programa Nacional do Livro e do Material Didático. Diário Oficial da União. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/decreto/D9099.htm. Acesso em: 4 jul. 2024.

[Conover, 1999] Conover, W. J. (1999). *Practical Nonparametric Statistics*. John Wiley & Sons.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

[Fundo Nacional de Desenvolvimento da Educação, s d] Fundo Nacional de Desenvolvimento da Educação (s. d.). Programas do livro. Disponível em: https://www.gov.br/fnde/pt-br/acesso-a-informacao/acoes-e-programas/programas/programas-do-livro. Acesso em: 4 jul. 2024.

[Greene, 2012] Greene, W. H. (2012). *Econometric Analysis*. Pearson.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[Hyndman and Athanasopoulos, 2018] Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, s d] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (s. d.). Censo escolar. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar. Acesso em: 4 jul. 2024.

[Kalman, 1960] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

[Lazăr and Lazăr, 2015] Lazăr, C. and Lazăr, M. (2015). Forecasting methods of the enrolled students' number. *Economic Insights – Trends and Challenges*, IV(LXVII)(2):41–51.

[Montgomery et al., 2015] Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2015). *Introduction to Time Series Analysis and Forecasting*. John Wiley & Sons.

[Morettin, 2006] Morettin, P. A. (2006). *Análise de Séries Temporais*. Egard Blucher, São Paulo.

[OECD, 2010] OECD (2010). Education and social cohesion. in *Education at a Glance 2010: OECD Indicators*. Disponível em: https://www.oecd.org/education/skills-beyond-school/educationataglance2010oecdindicators.htm. Acesso em: 4 jul. 2024.

[Parkhi et al., 2023] Parkhi, P. N., Patel, A., Solanki, D., Ganwani, H., and Anandani, M. (2023). Machine learning based prediction model for college admission. In *2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET-SIP)*. IEEE.

[Rathore, 2023] Rathore, B. S. (2023). Linear regression - day 27. Disponível em: https://medium.com/@bsinghrathore32/linear-regression-day-27-f7ecb96db2c5. Acesso em: 4 jul. 2024.

[Skiena, 2017] Skiena, S. S. (2017). *The Data Science Design Manual*. Springer, Stony Brook, Nova Iorque.

[UNESCO, 2015] UNESCO (2015). *Education for All 2000-2015: Achievements and Challenges*. UNESCO, Paris. Disponível em: http://unesdoc.unesco.org/images/0023/002322/232205e.pdf. Acesso em: 4 jul. 2024.

[Vandeput, 2019] Vandeput, N. (2019). Simple exponential smoothing. Disponível em: https://towardsdatascience.com/simple-exponential-smoothing-749fc5631bed. Acesso em: 4 jul. 2024.

[Welch and Bishop, 1995] Welch, G. and Bishop, G. (1995). An introduction to the kalman filter. University of North Carolina at Chapel Hill.

Apêndice A

Limpeza e Pré-processamento

```
import pandas as pd
import numpy as np

# Carregando o dataframe
df = pd.read_csv('<caminho_dados>.csv', sep=',', low_memory=False)

# Criando uma nova coluna 'co_turma' composta por 'co_entidade' e 'co_etapa_ensino'
# Esse código deve aparecer uma vez a cada ano que a turma foi registrada no censo escolar
df['co_turma'] = df['co_entidade'].astype(str) + '-' + df['co_etapa_ensino'].astype(str)

# Levantando o código das turmas que aparecem no censo de 2023
turmas_com_2023 = df[df['nu_ano_censo'] == 2023]['co_turma']

# Filtrando o dataframe apenas pelas turmas que aparecem no censo de 2023
df_filtrado_2023 = df[df['co_turma'].isin(turmas_com_2023)]

# Agrupando por turma e adicionando a quantidade de registros
df_filtrado_2023_agrupado = df_filtrado_2023.groupby(['co_turma']).size().reset_index(name='quantitativo')
df_filtrado_2023_agrupado = df_filtrado_2023_agrupado.sort_values(by='quantitativo', ascending=False)

# Removendo linhas onde o quantitativo de registros é menor que 10 e pegando o código das turmas
df_filtrado_2023_agrupado_maior_igual_10 = df_filtrado_2023_agrupado[df_filtrado_2023_agrupado['quantitativo'] >= 10]
turmas_2023_maior_igual_10 = df_filtrado_2023_agrupado_maior_igual_10['co_turma'].unique()

# Filtrando o dataframe apenas pelas turmas que aparecem no censo de 2023 e tem 10 ou mais registros
```

```
df_filtrado_2023_maior_igual_10 = df[df['co_turma'].isin(
    turmas_2023_maior_igual_10)]

# Copiando o df e convertendo os valores de ano para string para poder usar
    como nome de coluna
df_filtrado_2023_maior_igual_10_copia = df_filtrado_2023_maior_igual_10.
    copy()
df_filtrado_2023_maior_igual_10_copia['nu_ano_censo'] =
    df_filtrado_2023_maior_igual_10_copia['nu_ano_censo'].astype(str)

# Usando pivot_table para montar o df da série temporal
# Os valores de 'qtd_alunos' serão distribuídos conforme 'nu_ano_censo',
    para cada 'co_turma'
# Valores NaN serão preenchidos com 0 e o name do index será removido
df_serie_temporal = df_filtrado_2023_maior_igual_10_copia.pivot_table(
        index='co_turma',columns='nu_ano_censo', values='qtd_alunos',
    aggfunc="sum"
    ).reset_index().fillna(0)
df_serie_temporal.columns.name = None

# Adicionando uma coluna com a quantidade de registros no censo para cada
    turma
anos = [str(ano) for ano in range(2007, 2024)]
df_serie_temporal['nu_valores'] = df_serie_temporal[anos].apply(lambda row:
    (row != 0).sum(), axis=1)

# Verificando se há interrupção na série temporal
def verifica_interrupcao(row):
    serie_anos = row[anos]
    inicio_atividade = serie_anos[serie_anos != 0].first_valid_index()
    fim_atividade = serie_anos[serie_anos != 0].last_valid_index()
    return (serie_anos.loc[inicio_atividade:fim_atividade] == 0).any()

df_serie_temporal['has_interruption'] = df_serie_temporal.apply(
    verifica_interrupcao, axis=1)

# Salvando o df da série temporal
df_serie_temporal.to_csv('<caminho_dados_preprocessados>.csv', index=False)
```

Apêndice B

Análise Exploratória

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Transformando para formato longo
df_long = pd.melt(df_serie_temporal, id_vars=['co_turma'], var_name='
    nu_ano_censo', value_name='qt_censo',
                  value_vars=[str(year) for year in range(2007, 2023)])
df_long = df_long[df_long['qt_censo'] > 0]

# Função para calcular estatísticas de um ano específico
def calcular_estatisticas(df, ano):
    dados_ano = df[df['nu_ano_censo'] == str(ano)]['qt_censo']
    return {
        'Média ao longo dos anos': dados_ano.mean(),
        'Mediana (média entre turmas)': dados_ano.median(),
        'Primeiro Quartil (média entre turmas)': dados_ano.quantile(0.25),
        'Terceiro Quartil (média entre turmas)': dados_ano.quantile(0.75),
    }

# Calculando estatísticas para cada ano
anos = range(2007, 2023)
estatisticas_anuais = {ano: calcular_estatisticas(df_long, ano) for ano in
    anos}

# Convertendo para DataFrame e calculando a média das estatísticas
df_estatisticas_anuais = pd.DataFrame(estatisticas_anuais).T
estatisticas_media = df_estatisticas_anuais.mean().to_dict()

# Calculando o desvio padrão
df_serie_temporal['desvio_padrao'] = df_serie_temporal[[str(year) for year
    in anos]].replace(0, np.nan).std(axis=1, ddof=0)
desvio_padrao_medio = df_serie_temporal['desvio_padrao'].mean()
```

```
32
33 # Calculando o mínimo e máximo
34 minimo_geral = df_long['qt_censo'].min()
35 maximo_geral = df_long['qt_censo'].max()
36
37 # Adicionando os valores ao dicionário de estatísticas médias
38 estatisticas_media.update({
39     'Desvio Padrão por turma': desvio_padrao_medio,
40     'Mínimo Geral': df_long['qt_censo'].min(),
41     'Máximo Geral': df_long['qt_censo'].max(),
42 })
43
44 # Criando o boxplot para todos os anos
45 plt.figure(figsize=(20, 10))
46 plt.ylim(-50, 300)
47
48 plt.title('Boxplot da Quantidade de Alunos por Ano')
49 plt.xlabel('Ano')
50 plt.ylabel('Quantidade de Alunos')
51 plt.xticks(rotation=45)
52
53 sns.boxplot(x='nu_ano_censo', y='qt_censo', data=df_long)
54
55 plt.show()
56
57 # Geração do histograma combinado para todos os anos
58 plt.figure(figsize=(20, 10))
59 plt.xlim(0, 250)
60
61 plt.title('Histograma da Quantidade de Alunos (Todos os Anos)')
62 plt.xlabel('Quantidade de Alunos')
63 plt.ylabel('Frequência')
64
65 sns.histplot(data=df_long, x='qt_censo', bins=1000)
66
67 plt.show()
```

Apêndice C

Determinação de Baselines

```
import numpy as np
from sklearn.linear_model import LinearRegression

def linear_regression_baseline(row, anos):
    y = row[anos].values
    X = np.array([int(year) for year in anos]).reshape(-1, 1)
    mask = y != 0
    X, y = X[mask], y[mask]

    if len(y) > 0:
        model = LinearRegression().fit(X, y)
        return model.predict(np.array([[2023]]))[0]
    else:
        return np.nan

def exponential_smoothing(series, alpha):
    result = [series[0]]
    for n in range(1, len(series)):
        result.append(alpha * series[n] + (1 - alpha) * result[n-1])
    return result

def exponential_smoothing_baseline(row, anos, alpha=0.5):
    y = row[anos].values
    mask = y != 0
    y = y[mask]

    if len(y) > 0:
        smoothed_values = exponential_smoothing(pd.Series(y), alpha)
        return smoothed_values[-1]
    else:
        return np.nan

# Definindo baselines para cada turma (média e repetir o último valor)
anos = [str(ano) for ano in range(2007, 2023)]
```

```
35
36 # Calculando a média para os anos entre 2010 e 2022 apenas com valores
       diferentes de zero
37 df_serie_temporal['baseline_media'] = df_serie_temporal[anos].replace(0, np
       .NaN).mean(axis=1).round(0)
38
39 # Repetindo o valor do ano de 2022 como baseline
40 df_serie_temporal['baseline_last_year'] = df_serie_temporal['2022']
41
42 # Calculando predição através de regressão linear
43 df_serie_temporal['baseline_linear_regression'] = df_serie_temporal.apply(
       linear_regression_baseline, axis=1, anos=anos).round(0)
44
45 # Calculando predição através de suavização exponencial
46 df_serie_temporal['baseline_exponential_smoothing'] = df_serie_temporal.
       apply(exponential_smoothing_baseline, axis=1, anos=anos, alpha=0.5).
       round(0)
```

Apêndice D

Métricas de Avaliação

```python
import numpy as np
from scipy.stats import wilcoxon

def evaluate_metrics(dataframe, column_base, column_predicted):
    # Calculando o Mean Absolute Error (MAE)
    mae = np.abs(dataframe[column_base] - dataframe[column_predicted]).mean()

    # Calculando o Standard Deviation (STD)
    std = (dataframe[column_base] - dataframe[column_predicted]).std()

    # Calculando o Mean Squared Error (MSE)
    mse = ((dataframe[column_base] - dataframe[column_predicted]) ** 2).mean()

    # Calculando o Root Mean Squared Error (RMSE)
    rmse = np.sqrt(mse)

    # Calculando o Mean Absolute Percentage Error (MAPE)
    mape = (np.abs((dataframe[column_base] - dataframe[column_predicted]) / dataframe[column_base])).mean() * 100

    # Calculando o R-squared (R²)
    ss_res = ((dataframe[column_base] - dataframe[column_predicted]) ** 2).sum()
    ss_tot = ((dataframe[column_base] - dataframe[column_base].mean()) ** 2).sum()
    r_squared = 1 - (ss_res / ss_tot)

    return {
        "mae": mae,
        "std": std,
        "mse": mse,
        "rmse": rmse,
```

```
30          "mape": mape,
31          "r_squared": r_squared
32      }
33
34  def apply_wilcoxon_test(df, column1, column2):
35      stat, p_value = wilcoxon(df[column1], df[column2])
36      return stat, p_value
37
38  def test_wilcoxon(df):
39      results = {
40          'Baseline': [],
41          'W-Statistic': [],
42          'P-Value': []
43      }
44
45      baselines = ['baseline_media', 'baseline_last_year', '
        baseline_linear_regression', 'baseline_exponential_smoothing']
46
47      for baseline in baselines:
48          w_stat, p_value = apply_wilcoxon_test(df, baseline, 'predicao_kf')
49          results['Baseline'].append(baseline)
50          results['W-Statistic'].append(w_stat)
51          results['P-Value'].append(p_value)
52
53      return pd.DataFrame(results)
```

Apêndice E

Desenvolvimento do Filtro de Kalman

```
1  from pykalman import KalmanFilter
2  import numpy as np
3  import pandas as pd
4
5  def run_kalman_filter(data, em_iterations, year_gap):
6      # Criando DataFrame temporário para pré-processamento
7      temp_df = pd.DataFrame(data)
8      temp_df = temp_df[temp_df.cumsum() != 0] # Removendo os zeros iniciais
       das séries temporais
9      temp_df = temp_df.ffill()                # Preenchendo os valores zero
       no meio da série temporal pelo último valor válido
10     temp_df = temp_df.dropna()               # Removendo os NaNs
       resultantes
11     series = temp_df.values.reshape(-1, 1)   # Extraindo os valores e
       redimensionando
12
13     # Definindo filtro
14     kalman_filter = KalmanFilter(em_vars=['transition_covariance', '
       observation_covariance'])
15
16     # Rodando Filtro de Kalman
17     kalman_filter = kalman_filter.em(series, n_iter=em_iterations) #
       Executando o algoritmo EM para estimativa de parâmetros
18     _, _ = kalman_filter.filter(series)                            #
       Filtrando os dados
19     smoothed_state_means, _ = kalman_filter.smooth(series)         #
       Suavizando os dados
20
21     # Predizendo o próximo valor da série
22     last_smoothed_state = smoothed_state_means[-1]
       # Obtendo o último estado suavizado
23     next_state = kalman_filter.transition_matrices.dot(last_smoothed_state)
       # Aplicando a matriz de transição para prever o próximo estado
24     if year_gap == 2:
```

```
25          next_state = kalman_filter.transition_matrices.dot(next_state)
        # Aplicando novamente a matriz de transição para casos em que o gap é
        de 2 anos
26
27      next_observation = kalman_filter.observation_matrices.dot(next_state)
        # Aplicando a matriz de observação para prever a próxima observação
28
29      # Retornar o valor previsto
30      return np.round(next_observation[0])
```