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

José Ferreira Leite Neto[1], **Glauber Rodrigues Leite**[1], **Ícaro Bezerra Queiroz de Araújo**[1]
Bruno Almeida Pimentel[1]

[1] Instituto de Computação
Universidade Federal de Alagoas (UFAL)
Maceió – AL – Brazil

{jfln,glauber,icaro,brunopimentel}@ic.ufal.br

Abstract. *Brazil faces logistical challenges that require advance planning for textbook acquisition under the National Textbook and Teaching Materials Program (Programa Nacional do Livro e do Material Didático, PNLD). Moreover, planning relies on past School Census data, leading administrators to estimate student numbers with at least a two-year delay. Traditional forecasting models often fail to accurately predict student enrollment, motivating the search for alternative approaches. In this study, we propose using the Kalman Filter to estimate enrollment in public schools in Brazil's Northeast, comparing its performance against traditional time-series baselines. We analyzed 11,971 classes and observed a reduction in the average error of 1.47 students per class, indicating potential improvements.*

Resumo. *O Brasil traz desafios logísticos que demandam planejamento na compra de livros para os próximos anos no Programa Nacional do Livro e do Material Didático (PNLD). Além disso, os dados para o planejamento são obtidos a partir de Censos Escolares anteriores, o que leva os gestores a estimar o quantitativo de alunos com uma defasagem de, pelo menos, dois anos. Modelos clássicos de previsão frequentemente não predizem com precisão futuras matrículas, o que motiva a busca por outras alternativas. Neste trabalho, propõe-se o uso do Filtro de Kalman para estimar matrículas em escolas públicas do Nordeste brasileiro, comparando seu desempenho com baselines tradicionais de séries temporais. Foram analisadas 11.971 turmas, e houve redução no erro médio em 1,47 aluno/turma, indicando potencial de melhorias.*

1. Introdução

De acordo com [OECD 2024], sistemas educacionais de qualidade são catalisadores de oportunidades, embora o maior desafio seja garantir equidade. Indo nesse sentido, o Programa Nacional do Livro e do Material Didático (PNLD), criado em 1937, é a política pública mais longeva do país e distribui obras didáticas, pedagógicas e literárias para toda a rede básica. Em 2024, o programa investiu mais de R$ 2 bi e beneficiou 31 milhões de alunos, com números semelhantes nos anos anteriores [Agência Gov 2024][FNDE 2025]. Esse porte traz diversos desafios, sendo um deles: "Como determinar quantos livros comprar para cada turma, em cada escola e município do Brasil?"

A região Nordeste, com taxa de analfabetismo de 11,2% em 2023 — quase quatro vezes a do Sudeste e Sul — ilustra as graves desigualdades educacionais no país

DOI: 10.5753/sbie.2025.12614

[Instituto Brasileiro de Geografia e Estatística 2023]. Nesse contexto, um modelo preditivo de matrículas mais preciso pode otimizar a logística do PNLD, reduzir custos e contribuir para a melhoria da qualidade de ensino e equidade regional.

O processo de seleção e distribuição de obras do PNLD é complexo e composto de diversas etapas, durando cerca de 24 meses. Participam diversos especialistas em educação, atributos editoriais e acessibilidade, de forma que, uma vez aprovada, a obra fica disponível para ser escolhida pelas escolas para uso no próximo ciclo [Ministério da Educação 2024]. Durante a seleção, são considerados rigorosos critérios de qualidade pedagógica, adequação ao currículo nacional proposto e conformidade com os direitos humanos e princípios éticos.

Na fase de negociação, é imprescindível definir com precisão quantos livros deverão ser adquiridos, valendo-se para isso dos dados do Censo Escolar da Educação Básica [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira 2023]. Porém, como esses dados só ficam disponíveis semanas após o fechamento do ano letivo, ocorre a seguinte situação: considere a compra de livros para 2026, feita em 2025 - o último censo disponível, 2024 é usado para planejar as aquisições. Isso gera um descompasso de dois anos entre a informação disponível e o momento de uso. Esse intervalo temporal evidencia a necessidade de um modelo preditivo capaz de corrigir essa defasagem e apoiar decisões mais precisas na compra de material didático.

Atendendo milhões de alunos, o PNLD não comporta estimativas manuais, que demandam tempo, elevam custos e ficam sujeitas a variações metodológicas e erros humanos. Por isso, recorre-se ao tratamento computacional dessas informações por meio da Ciência de Dados – interseção entre Ciência da Computação, Estatística e Domínios de Aplicação [Skiena 2017]. Como informação base para o processamento, pode-se utilizar o quantitativo de alunos em anos anteriores e formar séries temporais de matrículas com base no Censo Escolar, divulgado anualmente pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira 2023].

Para o processamento dos dados, podem ser empregados métodos como a Regressão Linear, a Suavização Exponencial e o algoritmo recursivo desenvolvido por Rudolf E. Kálmán em 1960, e que leva seu nome: Filtro de Kalman, amplamente utilizado nas áreas de controle, navegação e sistemas de estimativa [Kalman 1960], que se mostra como uma solução ainda pouco explorada para a predição do alunado através das séries temporais obtidas por meio do quantitativo de matrículas do Censo Escolar.

A proposta do presente estudo é a utilização do Filtro de Kalman para resolver o problema de predição do número de alunos em escolas da região Nordeste do Brasil; para isso,serão comparados os resultados obtidos entre diferentes métodos, aqui tratados como baselines, e os obtidos utilizando a metodologia proposta, avaliando a performance utilizando métricas estatisticamente reconhecidas em séries temporais, como MAE, R^2 e MAPE, cujas propriedades estão descritas em [Kotu and Deshpande 2018], [Montgomery et al. 2015] e [Hyndman and Athanasopoulos 2018].

Prever com precisão o número de estudantes matriculados em instituições educacionais motiva pesquisadores a desenvolver diversos modelos preditivos baseados em séries temporais. [Yang et al. 2020] propuseram uma abordagem utilizando al-

goritmos de otimização para aprimorar previsões de matrícula em Taiwan, enquanto [Lavilles and Arcilla 2012] estabeleceram metodologias usando análise de regressão para sistemas de gestão escolar. [Silva et al. 2025] evidenciaram a importância de métodos preditivos precisos para programas governamentais como o PNLD, motivando a exploração de técnicas como o Filtro de Kalman, que apresenta características recursivas e adaptativas que podem contribuir em contextos como o do PNLD.

Este trabalho divide-se em outras quatro seções, além da Introdução: Fundamentação Teórica, em que se abordam as definições de séries temporais, Filtro de Kalman e Métricas de Avaliação; Metodologia, que detalha as etapas de coleta e tratamento dos dados, cálculo de baselines e aplicação do Filtro de Kalman; Resultados, que apresenta os resultados dos experimentos realizados e os analisa; e Conclusão, que discute os principais achados, suas implicações e sugere trabalhos futuros.

2. Fundamentação Teórica

Esta seção abordará as técnicas utilizadas durante o desenvolvimento e validação do estudo. Serão apresentadas definições, aplicações e formalização de conceitos relacionados a séries temporais, Filtro de Kalman e métricas de avaliação.

2.1. Séries Temporais

Séries temporais, de acordo com [Pimentel 2025], podem ser caracterizadas como uma sequência de valores numéricos coletados em intervalos regulares de tempo discreto (podendo ser em segundos, minutos, horas, dias, meses, anos, entre outros) e que possuem observações vizinhas dependentes. Diversos campos do conhecimento podem se beneficiar de sua análise, tais como economia, medicina, meteorologia e uma infinidade de outros.

Um das técnicas muito utilizadas na modelagem de séries temporais é a regressão linear, que, como o próprio nome indica, visa simplificar o comportamento da série temporal a uma linha de tendência histórica, baseando-se na ideia de que valores anteriores podem ser um referencial para a tendência futura. De acordo com [Morettin 2006], podemos definir a fórmula básica da regressão linear como mostrado na Equação 1:

$$Z_t = \alpha + \beta t + \alpha_t, \tag{1}$$

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.

Já a suavização exponencial é uma técnica que lida melhor com possíveis ruídos presentes nos dados. Sua política de atualização de previsão é ajustada de acordo com as novas observações e os resultados são mais precisos conforme mais dados são fornecidos. Considerando o caso em que não existe uma tendência ou sazonalidade significativa, podemos definir, de acordo com [Morettin 2006], a equação 2:

$$\hat{Z}_t(h) = \hat{Z}_{t-1}(h+1) + \frac{Z_t - Z_{t-r}}{r}, \quad \forall h > 0. \tag{2}$$

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 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 determina o intervalo de tempo para a correção da previsão.

2.2. Filtro de Kalman

O Filtro de Kalman é um algoritmo recursivo utilizado para estimar estados em sistemas dinâmicos lineares, tendo como base medições ruidosas. É eficiente computacionalmente, utilizando apenas os estados anteriores e a nova medição para estimar o novo estado. Foi projetado para operar em dois passos: previsão e atualização [Welch and Bishop 1995]. Neste trabalho, foi considerado o Filtro de Kalman sem variável exógena.

Durante a fase de previsão, o filtro usa o modelo do sistema para prever o estado atual e sua incerteza. São estimados o estado *a priori* e a covariância de erro *a priori*, considerando que a incerteza segue uma distribuição gaussiana com média zero e covariância determinada pela matriz de covariância Q, de acordo com as Equações 3 e 4.

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1}, \tag{3}$$

$$P_{k|k-1} = AP_{k-1|k-1}A^T + Q, \tag{4}$$

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.
- A é a matriz que representa o modelo de transição do sistema.

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 5, 6 e 7:

$$K_k = P_{k|k-1}H^T(HP_{k|k-1}H^T + R)^{-1}, \tag{5}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(z_k - H\hat{x}_{k|k-1}), \tag{6}$$

$$P_{k|k} = (I - K_kH)P_{k|k-1}, \tag{7}$$

onde:

- K_k é o ganho de Kalman ótimo, obtido via minimização da covariância do erro com base nas estatísticas dos ruídos de processo w_k e de medição v_k.
- $\hat{x}_{k|k}$ é a estimativa *a posteriori* do estado no instante k,
- $P_{k|k}$ é a covariância do erro de estimativa *a posteriori*.
- H é a matriz de observação.
- $z_k - H\hat{x}_{k|k-1}$ é o erro de inovação, ou seja, a discrepância entre a medição real e a predição usada na atualização.

2.3. Métricas de Avaliação

Nesta seção, descrevem-se brevemente as métricas adotadas para comparar os modelos de predição de matrículas. Dentre o conjunto de métricas possíveis, optou-se pelos seguintes itens:

- **MAE (Mean Absolute Error)**: média das diferenças absolutas entre valores previstos e observados [Kotu and Deshpande 2018].

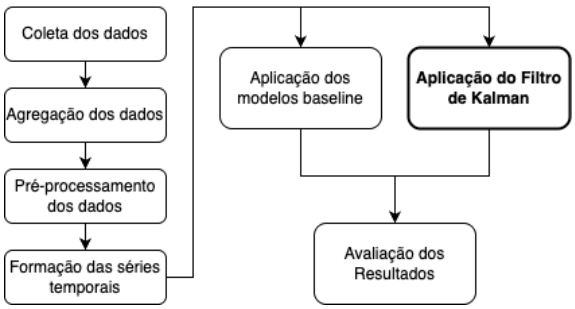
- **STD (Standard Deviation)**: desvio padrão dos dados, quantificando sua dispersão. É definido como a raiz quadrada da variância [Montgomery et al. 2015].
- **MSE (Mean Squared Error)**: média dos quadrados dos erros [Hyndman and Athanasopoulos 2018].
- **RMSE (Root Mean Squared Error)**: raiz quadrada do MSE, sendo mais intuitiva por manter a escala original [Hyndman and Athanasopoulos 2018].
- $\mathbf{R^2}$ **(Coeficiente de Determinação)**: proporção da variância total dos valores observados explicada pelos valores previstos [Montgomery et al. 2015].
- **MAPE (Mean Absolute Percentage Error)**: média percentual dos erros absolutos [Hyndman and Athanasopoulos 2018].

Conforme sugerido em [Makridakis et al. 1998], a combinação de métricas pode trazer uma visão mais abrangente do desempenho de cada modelo avaliado, não devendo se restringir a uma métrica isolada. A escolha desse conjunto de métricas deve levar em conta o contexto e a natureza dos dados, como afirmado em [Hyndman and Athanasopoulos 2018].

3. Metodologia

A Figura 1 apresenta o fluxo metodológico adotado neste estudo, detalhando as etapas do processamento dos dados e da modelagem preditiva.

Figura 1. Fluxo da metodologia.



Fonte: Elaborado pelos autores.

Foram utilizados os dados do Censo Escolar, que são públicos e disponibilizados pelo INEP em página própria [1], de acordo com a política nacional de dados abertos; os números de matrículas foram agrupados de acordo com a coluna **CO_ENTIDADE** (código INEP da escola) e o ano/série, obtido através das colunas **QT_MAT_X**, onde X é o ano/série desejado, por exemplo, **QT_MAT_FUND_AI_1**, para o Ensino Fundamental, Anos Iniciais - 1ª série, e colhidos ao longo dos anos de 2007 a 2024 (respectivamente, o primeiro ano com dados de matrículas e o último ano com dados disponíveis). Foi criado o identificador **CO_TURMA** para identificar um ano/série em uma determinada entidade, composto por: **[CO_ENTIDADE]-[CO_ETAPA_ENSINO]**, sendo este último de acordo com os valores definidos pelo INEP para cada um dos anos/séries.

Para delimitar os dados a serem usados neste trabalho, foram selecionadas apenas linhas de estados da região nordeste (em que a coluna **SG_UF** fosse um dos valores:

[1] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar

”AL”, ”BA”, ”CE”, ”MA”, ”PB”, ”PE”, ”PI”, ”RN”, ”SE”) e totalizando ao menos treze registros ao longo dos anos (podendo ter interrupções), de forma a melhorar a qualidade dos dados fornecidos aos modelos.

Para formar a série temporal, os dados foram agrupados por **CO_TURMA**, de forma que cada coluna representou a quantidade de matrículas em um ano entre 2007 e 2024. O valor zero foi utilizado caso um registro para aquele ano não fosse encontrado. Os experimentos foram repetidos para os anos de 2020 a 2024.

Como *baselines*, para comparação com os resultados obtidos pelo Filtro de Kalman, foram utilizados os seguintes métodos:

- **Média dos Anos Anteriores**: utiliza a média das quantidades de alunos para o intervalo entre 2007 e Y-2 como estimativa para o valor no ano Y (ex: 2007 a 2022 estimando 2024) e excluindo os anos em que o valor foi zero.
- **Último ano**: utiliza a quantidade de alunos informada no antepenúltimo ano disponível (Y-2) como estimativa para a quantidade de alunos no ano Y (ex: 2022 como estimativa de 2024).
- **Regressão Linear**: aplica um modelo de regressão linear aos dados de 2007 a Y-2, ajustando uma reta que melhor represente a tendência ao longo do tempo. O valor predito para o ano Y é obtido extrapolando esta linha para os próximos dois anos.
- **Suavização Exponencial**: utiliza nos dados de 2007 a Y-2 o método que dá pesos maiores aos dados mais recentes, permitindo que a previsão para o ano Y seja mais sensível às últimas mudanças de tendência da série temporal.

Para o desenvolvimento do Filtro de Kalman, foram seguidas as seguintes etapas de forma sequencial:

- **Formatação dos dados**: foi necessário remover os zeros iniciais, preencher valores intermediários vazios com o último valor informado e remover valores, NaNs.
- **Configuração do Filtro**: Foi utilizada a biblioteca *pykalman*[2], que implementa o Filtro de Kalman. O parâmetro **em_vars**, que inicializa automaticamente as matrizes de transição e observação, foi utilizado.
- **Estimação de parâmeteros**: Foi utilizado o algoritmo de Expectation-Maximization (EM), que faz um ajuste iterativo das matrizes de transição A, covariância Q e observação H. O parâmetro *n_iter*, que varia o número de iterações de sua execução, foi passado variando entre 0 e 4, permitindo a escolha do melhor resultado entre 5 execuções.
- **Predição do próximo valor**: Foi utilizado o estado suavizado para predizer o próximo valor da série, aplicando a matriz de transição a ele, seguido da aplicação da matriz de observação ao novo estado estimado. Este valor foi utilizado como entrada para uma segunda iteração para obter o valor de dois anos à frente.

Como métricas de avaliação, foram utilizadas as seis já descritas na seção 2.3, Métricas de Avaliação: MAE (Mean Absolute Error), STD (Standard Deviation), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R^2 (Coeficiente de Determinação) e MAPE (Mean Absolute Percentage Error).

[2]https://pypi.org/project/pykalman/

As métricas foram obtidas considerando o valor gerado pela predição (tanto para os baselines quanto para o Filtro de Kalman) e o valor real de matrículas do censo para aquele **CO_TURMA** para o ano-alvo Y. Foram consideradas apenas linhas com **CO_TURMA** que tivessem um registro no censo do ano-alvo, uma vez que este era o valor a ser utilizado como referência para o cálculo das métricas. Com base nos valores calculados, foram elaboradas tabelas comparativas, tendo a métrica MAE como principal indicador de desempenho.

4. Resultados

Ao manter apenas anos/séries de escolas com mais de 13 registros na sua série temporal, foram consideradas para o experimento um total de 11.971 turmas, cerca de 1,72% do total possível de 695.363 turmas na região Nordeste. Esse total de turmas variou para cada ano-alvo entre os pretendidos (2020 a 2024) de maneira a sempre ter um valor para cálculo de métricas.

Na sequência, os dados foram organizados em séries temporais por **CO_TURMA**. Ao executar os algoritmos baseline com o objetivo de predizer o ano-alvo de 2024, dois anos à frente dos dados fornecidos, foram obtidas as métricas da tabela 1:

Tabela 1. Métricas dos Baselines para a região Nordeste em 2024

Baseline	**MAE**	**STD**	**MSE**	**RMSE**	**MAPE**	R^2
Média	16.19	30.71	969.00	31.13	48.3	0.7157
Repetir o último ano	12.94	27.79	777.00	27.87	37.1	0.7721
Regressão Linear	13.94	28.04	786.00	28.04	37.3	0.7693
Suavização Exponencial	11.98	24.86	625.00	24.99	34.5	0.8167

Fonte: Elaborado pelos autores.

Pode-se observar que os baselines ”Repetir o último ano”e ”Suavização Exponencial”apresentaram os melhores resultados, com menores valores de MAE, STD, MSE, RMSE e MAPE, além de valores mais altos de R^2. Já a média dos anos anteriores teve o pior desempenho. O resultado dos testes de execução do Filtro de Kalman para a predição com o ano-alvo de 2024 resultou em valores de MAE de: 12,07 (0 iterações), 11,86 (1 iteração), 11,79 (2 iterações), 11,80 (3 iterações) e 11,80 (4 iterações)

Notou-se nesta execução que uma maior quantidade de iterações no algoritmo EM tendeu a sobreajustar levemente o filtro para os dados de entrada, reduzindo a qualidade do resultado (considerando a métrica MAE como referência). Contudo, o intervalo de iterações entre 2 e 4 trouxe melhores métricas. A execução com duas iterações obteve a melhor performance e este resultado foi utilizado para este ano na comparação com os baselines. A mesma regra foi aplicada aos demais anos, de forma que os resultados obtidos foram: 11,90 (2020, 3 iterações), 12,46 (2021, 2 iterações), 10,57 (2022, 1 iteração), 10,07 (2023, 0 iterações) e 11,79 (2024, 2 iterações).

Ao repetir o experimento para os demais anos (2020 a 2023) e calcular a média entre as métricas obtidas, tanto para os baselines quanto para o Filtro de Kalman, obteve-se os resultados da Tabela 2, que traz o comparativo entre esses resultados em valores absolutos e percentuais, de forma que, em todas as métricas analisadas, houve redução do erro e, no caso da métrica R^2, aumento do valor, indicando performance superior:

Tabela 2. Métricas do Filtro de Kalman x baselines (média entre 2020 a 2024)

Modelo	**MAE**	**STD**	**MSE**	**RMSE**	**MAPE**	R^2
Filtro de Kalman	11.36	22.91	536.37	23.07	29.25	0.8584
Média	16.82	32.03	1043.6	32.29	46.40	0.7256
Variação (%)	**-32.47%**	**-28.49%**	**-48.60%**	**-28.58%**	**-36.95%**	**18.30%**
Repetir o último ano	13.65	27.15	743.8	27.20	36.34	0.8039
Variação (%)	**-16.79%**	**-15.64%**	**-27.89%**	**-15.19%**	**-19.50%**	**6.78%**
Regressão Linear	18.01	34.47	1224.2	34.63	44.74	0.6804
Variação (%)	**-36.94%**	**-33.55%**	**-56.19%**	**-33.40%**	**-34.61%**	**26.17%**
Suavização Exponencial	12.83	25.26	642.4	25.32	34.62	0.8311
Variação (%)	**-11.52%**	**-9.30%**	**-16.50%**	**-8.89%**	**-15.50%**	**3.29%**

Fonte: Elaborado pelos autores.

A média da diferença entre o MAE do Filtro de Kalman e o da Suavização Exponencial (melhor baseline), de 1,47 alunos por turma, aplicada à escala do PNLD, permite entender o impacto que melhorias nesse processo podem ter na educação e no orçamento público.

Considerando a quantidade de 385.833 turmas da região Nordeste registradas no censo escolar de 2024 e um erro médio de 1,47 aluno por turma, caso se mantenha para todas as turmas da região, pode-se estimar que 3.403.047 livros (considerando uma compra de 6 unidades por aluno) não seriam entregues, caso o modelo usado não apontasse a existência dos 567.175 alunos a quem estes livros atenderiam, ou ainda, um impacto orçamentário de aproximadamente R$ 27.224.376,00 (R$ 8,00 por livro), caso o modelo usado informasse mais alunos que a realidade. Esse valor ilustra o potencial ganho financeiro e pedagógico de aprimorar a precisão na previsão de matrículas.

5. Conclusão

Este trabalho investigou a aplicação do Filtro de Kalman Simples como método preditivo do número de alunos em escolas do Nordeste brasileiro no contexto do Programa Nacional do Livro e do Material Didático (PNLD), comparando seu desempenho com baselines tradicionais de séries temporais. Aplicado a 11.971 turmas, o filtro reduziu em 1,47 aluno por turma o erro médio absoluto, o que, extrapolado às 385.833 turmas da região, representa uma economia potencial de mais de R$ 27 milhões e maior eficiência na alocação de material didático.

O Filtro de Kalman superou os métodos testados, considerando a amostra analisada, aprimorando a precisão das previsões e podendo beneficiar também outros programas governamentais — como o PNAE, que hoje se apoia em dados defasados do Censo Escolar para estimar recursos de alimentação [FNDE 2024].

Como trabalhos futuros, propõe-se a expansão do estudo para escolas de todo o Brasil; a criação de séries temporais considerando o mesmo conjunto de alunos, isto é: ao invés de avaliar o quarto ano do ensino fundamental da mesma escola ao longo dos anos, por exemplo, fazer a predição para o quarto ano do ensino fundamental de 2026 considerando o histórico da turma que foi primeiro ano do ensino fundamental em 2022 ao longo dos anos, composta pelos mesmos alunos; abordagem de outras técnicas e incorporação da dinâmica de transição de estados para representar características como migração e reprovação, permitindo melhoria das estimativas.

6. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Referências

Agência Gov (2024). Decreto de lula regulamenta política nacional de leitura e escrita. Disponível em: https://agenciagov.ebc.com.br/noticias/202409/lula-assina-decreto-que-regulamenta-politica-nacional-de-leitura-e-escrita. Acesso em: 21 mar. 2025.

FNDE (2024). Pnae - programa nacional de alimentação escolar. Disponível em: https://www.gov.br/fnde/pt-br/acesso-a-informacao/acoes-e-programas/programas/pnae. Acesso em: 21 mar. 2025.

FNDE (2025). Dados estatísticos do programa nacional do livro e do material didático (pnld). Disponível em: https://www.gov.br/fnde/pt-br/acesso-a-informacao/acoes-e-programas/programas/programas-do-livro/pnld/dados-estatisticos. Acesso em: 21 mar. 2025.

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

Instituto Brasileiro de Geografia e Estatística (2023). Pesquisa nacional por amostra de domicílios contínua de 2023. Disponível em: https://www.ibge.gov.br. Acesso em: 21 mar. 2025.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2023). Censo escolar. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar. Acesso em: 21 mar. 2025.

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

Kotu, V. and Deshpande, B. (2018). *Data Science: Concepts and Practice*. Elsevier, 2 edition.

Lavilles, R. Q. and Arcilla, M. J. B. (2012). Enrollment forecasting for school management system. *International Journal of Modeling and Optimization*, 2(5):563–566.

Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. (1998). *Forecasting: Methods and Applications*. Wiley, New York, 3 edition.

Ministério da Educação (2024). Como funciona o programa nacional do livro e do material didático (pnld). Disponível em: https://www.gov.br/mec/pt-br/pnld/como-funciona. Acesso em: 21 mar. 2025.

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

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

OECD (2024). *Education at a Glance 2024: OECD Indicators*. OECD Publishing, Paris.

Pimentel, B. (2025). *Ciência de dados da teoria à prática*. Ed. do Autor, 1 edition.

Silva, L. C., Cabral, L. S., Santos Junior, J. J., Santos, L. L. P., Oliveira, T. T. M., Costa, B. J. D., and Pimentel, B. A. (2025). Forecasting student enrollments in brazilian schools for equitable and efficient education resource allocation. In *26th Annual International Conference on Digital Government Research*, pages 1–15.

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

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

Yang, S., Chen, H. C., Chen, W. C., and Yang, C. H. (2020). Student enrollment and teacher statistics forecasting based on time-series analysis. *Computational Intelligence and Neuroscience*, page 1246920.