Universidade de Brasília

Faculdade de Administração, Contabilidade, Economia e Gestão de Políticas Públicas - FACE/UnB

Programa de Pós-Graduação em Economia (PPGECO)

Previsão da Despesa Primária do Governo Central: Uma Análise Comparativa entre Técnicas Estatísticas, Aprendizado de Máquina, Aprendizado Profundo e Combinação de Previsões

Autor: Eduardo Jacomo Seraphim Nogueira

Orientador: Prof. Dr. Daniel Oliveira Cajueiro

Brasília, DF

2025

Eduardo Jacomo Seraphim Nogueira

Previsão da Despesa Primária do Governo Central: Uma Análise Comparativa entre Técnicas Estatísticas, Aprendizado de Máquina, Aprendizado Profundo e Combinação de Previsões

Dissertação apresentada ao curso de Mestrado Acadêmico do Programa de Pós-Graduação em Economia da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre em Economia.

Universidade de Brasília

Orientador: Prof. Dr. Daniel Oliveira Cajueiro

Brasília, DF
2025

Eduardo Jacomo Seraphim Nogueira

Previsão da Despesa Primária do Governo Central: Uma Análise Comparativa entre Técnicas Estatísticas, Aprendizado de Máquina, Aprendizado Profundo e Combinação de Previsões

Dissertação apresentada ao curso de Mestrado Acadêmico do Programa de Pós-Graduação em Economia da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre em Economia.

Trabalho aprovado. Brasília, DF, 21 de agosto de 2025:

Prof. Dr. Daniel Oliveira Cajueiro
Orientador

Prof. Dr. João Frois Caldeira
Membro

Prof. Dr. João Gabriel de Moraes Souza
Membro

Brasília, DF
2025

Aos meus pais, Maria e Hélio; à minha esposa Débora;
e aos meus filhos, Isadora, Teco e Bizu.

Agradecimentos

Agradeço, primeiramente, a Deus, pela força e pela luz em todos os momentos desta jornada. À minha família, Maria, Hélio, Débora e Isadora; pelo apoio incondicional, pela paciência e pelo incentivo constante. Ao meu orientador, Prof. Dr. Daniel Cajueiro, pela orientação atenta, pela confiança depositada e pelos ensinamentos que foram fundamentais para a realização deste trabalho. Aos membros da banca, Prof. Dr. João Caldeira e Prof. Dr. João Souza, pela disponibilidade, pelas críticas construtivas e pelas contribuições valiosas. Estendo, ainda, meus agradecimentos a todos os alunos e professores do Programa de Pós-Graduação em Economia da UnB, que compartilharam seus conhecimentos e contribuíram de forma decisiva para a minha formação acadêmica e pessoal.

Resumo

A previsão do comportamento das despesas públicas é essencial para o planejamento e a condução da política fiscal, bem como para a sustentabilidade das contas públicas. No entanto, muitos países, especialmente os de baixa e média renda, ainda realizam previsões fiscais por meio de métodos simples, subjetivos ou de meras extrapolações em planilhas eletrônicas. No cenário internacional, apesar da consolidação do uso de métodos estatísticos tradicionais, o avanço da aplicação de técnicas de aprendizado de máquina e de aprendizado profundo permanece limitado e concentrado, em grande parte, na previsão de receitas públicas. Alguns estudos apontam ganhos de precisão com o uso de algoritmos de aprendizado de máquina e de aprendizado profundo, sobretudo pela capacidade teórica de lidar com não linearidades e padrões complexos; entretanto, outros evidenciam dificuldades desses modelos em lidar com séries temporais curtas e ruidosas, comumente observadas em dados fiscais de países emergentes. No contexto brasileiro, não encontramos pesquisas voltadas à previsão desagregada das despesas primárias federais nem à comparação sistemática entre diferentes classes de modelos. Neste trabalho, investigamos empiricamente o desempenho preditivo de modelos estatísticos, de aprendizado de máquina, de aprendizado profundo e de combinações de modelos na previsão de séries temporais de despesas primárias federais brasileiras. Para tanto, combinamos modelos estatísticos, algoritmos supervisionados e otimização automática de hiperparâmetros para obterm previsões pontuais robustas. Utilizamos validação cruzada temporal como procedimento de seleção de modelos e previsão conformal para gerar intervalos de confiança calibrados. Empregamos diversas métricas de erro para comparar o desempenho das diferentes abordagens. De modo geral, os modelos estatísticos apresentam desempenho altamente competitivo, superando algoritmos de aprendizado de máquina e de aprendizado profundo em horizontes mais longos. Os modelos de aprendizado profundo são mais competitivos em horizontes mais curtos, enquanto a combinação de previsões apresenta desempenho equilibrado em diferentes contextos. Partindo de dados oficiais do Governo Federal, demonstramos que o uso de técnicas avançadas de previsão de séries temporais é ferramenta relevante para subsidiar o trabalho dos formuladores da política fiscal. Pesquisas futuras podem explorar variáveis exógenas e modelos estruturais ou semi-estruturais, de forma a ampliar a acurácia, a robustez e a interpretabilidade econômica dos modelos preditivos.

Palavras-chaves: previsão. séries temporais. política fiscal.

Abstract

Forecasting the behavior of public expenditures is essential for fiscal policy planning and implementation, as well as for the sustainability of public finances. However, many countries, especially low- and middle-income ones, still produce fiscal forecasts using simple, subjective methods or mere extrapolations in spreadsheets. Internationally, despite the consolidation of traditional statistical methods, the application of machine learning and deep learning techniques remains limited and is largely concentrated on revenue forecasting. Some studies point to accuracy gains from the use of machine learning and deep learning algorithms, mainly due to their theoretical ability to capture nonlinearities and complex patterns; however, other studies highlight the difficulties these models face in handling short and noisy time series, which are commonly observed in fiscal data from emerging economies. In the Brazilian context, we do not find research dedicated to disaggregated forecasting of federal primary expenditures nor to systematic comparisons across different classes of models. In this study, we empirically investigate the predictive performance of statistical models, machine learning models, deep learning models, and forecast combinations in predicting time series of Brazilian federal primary expenditures. To this end, we combine statistical models, supervised algorithms, and automated hyperparameter optimization to obtain robust point forecasts. We use temporal cross-validation as a model selection procedure and conformal prediction to generate calibrated confidence intervals. We employ several error metrics to compare the performance of the different approaches. Overall, statistical models display highly competitive performance, outperforming machine learning and deep learning algorithms at longer horizons. Deep learning models prove more competitive at shorter horizons, while forecast combinations achieve balanced performance across different contexts. Using official data from the Federal Government, we demonstrate that the application of advanced time series forecasting techniques constitutes a relevant tool to support the work of fiscal policy makers. Future research may explore exogenous variables and structural or semi-structural models in order to enhance the accuracy, robustness, and economic interpretability of predictive models.

Key-words: forecasting. time series. fiscal policy.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
BCB	Banco Central do Brasil
BCE	Banco Central Europeu
BiTCN	Bidirectional Temporal Convolutional Networks
BPC	Benefício de Prestação Continuada
CatBoost	Categorical Boosting
CBO	Congressional Budget Office
CES	Complex Exponential Smoothing
CSR	Complete Subset Regression
DeepAR	Deep Autoregressive
DeepsNPTS	Deep Non-Parametric Time Series
DilatedRNN	Dilated Recurrent Neural Network
DL	Deep Learning
Dlinear	Decomposition Linear Model
EnbPI	Ensemble Batch Prediction Intervals
ES-RNN	Exponential Smoothing Recurrent Neural Network
ETS	Error, Trend, and Seasonality
FEDformer	Frequency Enhanced Decomposed Transformer
FMI	Fundo Monetário Internacional
FNN	Feedforward Neural Network
FUNDEB	Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação
GRU	Gated Recurrent Units

HW	Holt-Winters
IFI	Instituição Fiscal Independente
IFS	Institute for Fiscal Studies
INSS	Instituto Nacional do Seguro Social
IPCA	Índice Nacional de Preços ao Consumidor Amplo
KAN	Kolmogorov-Arnold Network
KNN	K-Nearest Neighbors
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MFLES	Median, Fourier seasonality, Linear trend, and Exponential Smoothing
MiDaS	Mixed-data sampling
ML	Machine Learning
MLP	Multilayer Perceptron
MLP-Multivariate	Multivariate Multilayer Perceptron
MSSE	Mean Squared Scaled Error
MSE	Mean Squared Error
N-BEATS	Neural Basis Expansion Analysis for Time Series
NBEATS	Neural Basis Expansion Analysis for Time Series
NHITS	Neural Hierarchical Interpolation for Time Series
Nlinear	Nonlinear Forecasting Model
OBR	Office for Budget Responsibility
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
PatchTST	Patch-based Time Series Transformer
PIB	Produto Interno Bruto

RFB	Secretaria da Receita Federal do Brasil
RNN	Recurrent Neural Networks
RGPS	Regime Geral de Previdência Social
RMSE	Root Mean Squared Error
RMSSE	Root Mean Squared Scaled Error
RTN	Boletim Resultado do Tesouro Nacional
SARIMA	Seasonal Autoregressive Integrated Moving Average
SIAFI	Sistema Integrado de Administração Financeira do Governo Federal
SJP	Sentenças Judiciais e Precatórios
SMAPE	Symmetric Mean Absolute Percentage Error
SOFTS	State-Of-The-Forecast Time Series
StemGNN	Spectral-Temporal Graph Neural Network
STN	Secretaria do Tesouro Nacional
SVM	Support Vector Machine
TBATS	Trigonometric, ARMA errors, Box-Cox transformation, Trend, and Seasonality
TCN	Temporal Convolutional Networks
TFT	Temporal Fusion Transformer
TiDE	Time-series Dense Encoder
TimesNet	TimesNet Convolutional Architecture
TPE	Tree-structured Parzen Estimator
VAR	Vector Autoregression
XGBoost	Extreme Gradient Boosting
TSMixer	Time Series Token Mixing

Sumário

1 Introdução

1.1 Motivação

A previsão das despesas públicas é elemento central para o planejamento e a condução da política fiscal, influenciando diretamente a credibilidade das contas públicas, o cumprimento das regras fiscais e a alocação eficiente de recursos. Apesar disso, em muitos países, especialmente os de baixa e média renda, ainda prevalecem métodos simples e subjetivos, frequentemente insuficientes para capturar a complexidade e a volatilidade dos gastos públicos. No caso brasileiro, a tarefa é dificultada por fatores como a rigidez institucional de parcelas significativas do orçamento, a exposição a choques exógenos, a ocorrência de quebras estruturais e a instabilidade da condução da política fiscal.

1.2 Contexto

O presente trabalho se diferencia da literatura em três aspectos principais. Primeiro, em relação à experiência internacional, observamos que, embora os países desenvolvidos procurem utilizar métodos estruturais ou semi-estruturais, conforme os trabalhos de Arnold (2018) e Ciccarelli et al. (2024), com foco em despesas desagregadas, como nos estudos de Stern et al. (2023), OBR (2011a) e OBR (2011b), nos países em desenvolvimento ainda predominam modelos simples e subjetivos, conforme destacado por ECLAC (2015) e Kyobe e Danninger (2005). Nosso trabalho está alinhado às melhores práticas internacionais, pois utilizamos dados desagregados e critérios objetivos, mas procuramos explorar modelos de previsão de séries temporais, visto que estes modelos são muito competitivos e conseguem inclusive superar modelos estruturais e semi-estruturais em tarefas de previsão fiscal, conforme destacado por Favero e Marcellino (2005) há mais de duas décadas.

Segundo, em relação à experiência nacional, realizamos uma revisão de literatura que não identificou trabalhos dedicados à investigação aprofundada da modelagem das despesas públicas federais no Brasil. A literatura nacional concentra-se majoritariamente na previsão de receitas ou de saldos fiscais agregados, ou investiga erros de previsão, como, por exemplo, nos trabalhos de Medeiros, Aragón e Besarria (2022); Gadelha, Lima e Polli (2020); Carneiro e Costa (2021); Deus e Mendonça (2017). Não encontramos estudos que realizem previsão desagregada de despesa federal nem comparação entre o desempenho de diferentes famílias de modelos. Assim, o presente estudo preenche uma lacuna relevante na literatura e oferece contribuição valiosa ao servir de base para futuras pesquisas e aplicações práticas.

Terceiro, em relação à comparação sistemática de modelos, embora existam trabalhos que avaliem o desempenho de modelos de previsões estatísticas, de aprendizado de máquina e de aprendizado profundo, como, por exemplo, Godahewa et al. (2021) e Makridakis, Spiliotis e Assimakopoulos (2020), nosso estudo se destaca por ser direcionado para a previsão fiscal, englobar diversas métricas de desempenho distintas e estabelecer um processo de avaliação e seleção de modelos de forma dinâmica. Assim, ao contrário das tradicionais competições de desempenho de modelos, que avaliam bancos de dados estáticos, propomos um processo que possui aplicação prática, pois fornece previsões consistentes com base nos dados disponíveis, que são automaticamente atualizados.

1.3 Objetivos

O objetivo principal deste estudo é investigar e avaliar o desempenho de modelos estatísticos, de aprendizado de máquina, de aprendizado profundo e de combinação de previsões na previsão de séries temporais de despesas públicas. Buscamos aprimorar a precisão e a confiabilidade das projeções fiscais realizadas no Brasil e demonstrar a viabilidade de métodos estatísticos na prática. Outros objetivos incluem comparar o desempenho preditivo de diversos modelos, de forma individual e combinada, e em diferentes horizontes de projeção fiscal, avaliando sua acurácia e robustez por meio de métricas de erro absoluto (MAE e MASE), quadrático (MSE, RMSE, MSSE e RMSSE) e percentual (MAPE e SMAPE).

Adicionalmente, analisamos a sensibilidade e a robustez dos modelos em situações de eventos exógenos, identificando como diferentes métodos lidam com alterações estruturais e rupturas nos padrões históricos das séries temporais. Por fim, aplicamos um processo de seleção automática de modelos de previsão fiscal com base nas características específicas das séries temporais e no horizonte de previsão. Com isso, buscamos contribuir para a prática institucional da projeção fiscal, oferecendo recomendações para a adoção de métodos mais avançados e fomentando o debate sobre as contas públicas nacionais.

Empregamos técnicas de análise, previsão e avaliação de séries temporais, incluindo separação dos dados entre treino e teste, otimização automática de hiperparâmetros, validação cruzada, comparação com modelos de referência, análise de resíduos e cálculo de métricas de erro. Também testamos combinações de previsões entre classes de modelos, por meio de estratégias de *ensemble*.

Os resultados obtidos contribuem para o debate sobre boas práticas em previsão fiscal, pois fornecem evidência sistemática sobre o desempenho relativo de diferentes abordagens metodológicas em um contexto real e institucionalmente relevante. Mais especificamente, observamos que o desempenho relativo dos modelos é sensível ao horizonte de previsão e à existência de choques e mudanças significativas na condução da política

pública.

Nossos resultados mostram que não há um método dominante em todos os horizontes: modelos estatísticos superam as alternativas em horizontes mais longos, enquanto arquiteturas de aprendizado profundo apresentam vantagem em horizontes curtos. Combinações de previsões oferecem desempenho equilibrado, embora não liderem em termos de menor erro em nenhum cenário. Esses achados reforçam a importância de estratégias adaptadas ao horizonte e à natureza das séries, com potencial aplicação prática na formulação da política fiscal.

1.4 Organização do trabalho

Este trabalho está organizado da seguinte forma: a Seção 2 revisa a literatura relevante sobre previsão fiscal e séries temporais financeiras, apresentando o estado da arte nacional e internacional, o que serve de base para a metodologia proposta. A Seção 3 descreve os dados e os métodos empíricos, incluindo modelos, procedimentos de validação e métricas de avaliação. A Seção 4 apresenta e a Seção 5 discute os resultados empíricos. Por fim, a Seção 6 conclui o trabalho, ressaltando as principais implicações para a prática de previsão fiscal das despesas primárias federais brasileiras.

2 Revisão da Literatura

2.1 Fundamentos e desafios da previsão de despesa pública

A elaboração de previsões fiscais confiáveis constitui atividade central para o funcionamento eficiente do setor público, tanto no âmbito da formulação de políticas quanto à sustentabilidade das finanças públicas no médio e longo prazo (ECLAC, 2015, p. 1). A qualidade dessas projeções afeta diretamente a credibilidade dos governos, o cumprimento das regras fiscais e a alocação eficiente dos recursos públicos. Embora a literatura sobre previsão fiscal tenha tradicionalmente concentrado esforços na modelagem da receita, observa-se consenso crescente quanto à necessidade de que a despesa pública seja objeto de análise metodologicamente rigorosa (ECLAC, 2015, p. 2).

Estudos como o de Kyobe e Danninger (2005, p. 4) investigam de forma aprofundada as práticas de previsão de receita em países em desenvolvimento, evidenciando a escassez de pesquisas sobre seus determinantes. Os autores observam que, em países de baixa renda, as projeções de receita são predominantemente realizadas de forma agregada, ao passo que, em países de renda mais elevada, utilizam-se dados mais desagregados. Além disso, os autores destacam que, embora métodos estatísticos mais sofisticados sejam empregados em determinados contextos, prevalece o uso de avaliações subjetivas e técnicas simples de extrapolação como prática dominante na maioria dos países de baixa renda para derivação das projeções de receita orçamentária (KYOBE; DANNINGER, 2005, p. 15).

Como observado por Kyobe e Danninger (2005, p. 15), a maioria dos países em desenvolvimento ainda utiliza métodos muito simples ou subjetivos, ou meras extrapolações para a realização de projeções de receitas, o que não difere da realidade observada no Brasil, tanto para a previsão de receitas quanto para a projeção de despesas públicas. Por consequência, os gestores da política fiscal dependem fortemente de estruturas baseadas em planilhas, julgamento e projeções das autoridades nacionais, sujeitas a ajustes discricionários mais fáceis de manipular e difíceis de detectar, em detrimento de modelos econométricos formais (KYOBE; DANNINGER, 2005, p. 16).

Em contraste, a projeção das despesas públicas é tradicionalmente abordada com uma metodologia que, embora fundamental para a gestão orçamentária, frequentemente carece da mesma profundidade analítica baseada em modelos estatísticos preditivos. O conceito predominante para a projeção de despesas reside na elaboração de "linhas de base" (*baselines*) ou "cenários de políticas inalteradas" (*no-policy change*), nos quais se estima o custo futuro dos serviços públicos assumindo a continuidade das políticas e

estruturas existentes (RAHIM; WENDLING; PEDASTSAAR, 2022, p. 1).

Essas linhas de base são construídas a partir dos custos de insumos (mão de obra, custos operacionais, equipamentos), cujos fatores (preço e volume) são ajustados por parâmetros macroeconômicos, como inflação ou crescimento populacional (RAHIM; WENDLING; PEDASTSAAR, 2022, p. 3). Segundo os autores, embora tal abordagem seja vital para o planejamento e a disciplina fiscal, ela se concentra em custear as políticas existentes em vez de prever o comportamento futuro das despesas com base em relações estatísticas complexas e dinâmicas econômicas e sociais subjacentes.

Diversos organismos internacionais enfatizam a importância estratégica da previsão da despesa. A Comissão Econômica para a América Latina e o Caribe (CEPAL) destaca que, especialmente em países latino-americanos, a despesa pública apresenta características estruturais que a tornam desafiadora de prever, como rigidez orçamentária, vinculações legais e exposição a choques exógenos (ECLAC, 2015, p. 2). O Fundo Monetário Internacional (FMI) reforça que previsões de despesa são determinantes para a construção de linhas de base orçamentárias e para a realização de análises de sustentabilidade fiscal (RAHIM; WENDLING; PEDASTSAAR, 2022, p. 5), sendo particularmente críticas em contextos de consolidação fiscal ou reformas estruturais (IEO, 2014, p. 17).

A Organização para a Cooperação e Desenvolvimento Econômico (OCDE) argumenta que previsões de despesa bem fundamentadas são essenciais para a atuação eficaz de Instituições Fiscais Independentes[1] (IFIs), uma vez que permitem a detecção precoce de riscos fiscais e o aprimoramento da transparência orçamentária (SHAW, 2017, p. 17). Além disso, Cameron (2022, p. 3) recomenda que as previsões sejam avaliadas sistematicamente por meio de processos de revisão *ex-post*, com ênfase na aprendizagem institucional, e não apenas na precisão pontual.

Apesar do reconhecimento da importância do tema, a previsão da despesa pública enfrenta uma série de desafios práticos. Entre os mais relevantes destacam-se: (i) a baixa frequência e o reduzido número de observações disponíveis nas séries históricas; (ii) a presença de quebras estruturais resultantes de mudanças institucionais, alterações legais ou reclassificações contábeis; (iii) a coexistência de componentes altamente rígidos (como aposentadorias e pensões, pessoal e transferências obrigatórias) com parcelas discricionárias sujeitas a instabilidade política e contingenciamentos (ECLAC, 2015, p. 2); e (iv) a dificuldade de antecipar o comportamento de agentes públicos na execução do gasto, especialmente em anos eleitorais (HADZI-VASKOV et al., 2021, p. 6).

[1] De acordo com a OCDE, Instituições fiscais independentes (IFIs) são instituições públicas independentes com o mandato de avaliar criticamente e, em alguns casos, fornecer aconselhamento imparcial sobre política e desempenho fiscal. As IFIs visam promover uma política fiscal sólida e finanças públicas sustentáveis, ajudando a promover maior transparência sobre as contas públicas. Disponível em: https://www.oecd.org/en/topics/parliamentary-budget-offices-and-independent-fiscal-institutions.html.

Do ponto de vista metodológico, a literatura aponta que, tradicionalmente, as previsões de despesa baseiam-se em métodos determinísticos, planilhas estruturadas por elasticidades e modelos estatísticos univariados ou multivariados, como regressões lineares, modelos autorregressivos integrados de médias móveis e modelos de suavização exponencial (ECLAC, 2015, p. 13). Os modelos estruturais são mais empregados em análises de médio e longo prazo, muitas vezes incorporando projeções macroeconômicas exógenas para variáveis como PIB, inflação e demografia (ANDO; KIM, 2022, p. 19).

Nos últimos anos, observa-se um crescimento do interesse pelo uso de métodos de aprendizado de máquina, ou *machine learning* (ML), e de aprendizado profundo, ou *deep learning* (DL), em previsão fiscal. Estudos empíricos recentes publicados pelo FMI exploram essa agenda em diversas frentes e sugerem que modelos de aprendizado de máquina podem superar métodos tradicionais em determinadas tarefas preditivas, especialmente quando há grandes volumes de dados, múltiplas variáveis correlacionadas e padrões não lineares complexos (JUNG; PATNAM; TER-MARTIROSYAN, 2018, p. 8); (BOLHUIS; RAYNER, 2020, p. 12). Além disso, Chen e Ranciere (2016, p. 8) demonstram que informações financeiras de alta frequência, como *spreads* soberanos e taxas de juros de mercado, podem antecipar o comportamento de variáveis fiscais, indicando potencial valor preditivo complementar.

No entanto, também se reconhece que esses métodos apresentam limitações importantes, como baixa interpretabilidade, sensibilidade a sobreajuste e necessidade de calibração cuidadosa (BOLHUIS; RAYNER, 2020, p. 12). Destaca-se ainda que a maior parte das aplicações de ML e DL no campo fiscal concentra-se na previsão da arrecadação ou do crescimento econômico, havendo escassez de evidência empírica sobre seu desempenho na previsão de despesas públicas (EICHER et al., 2018, p. 6). Ademais, são raros os trabalhos que realizam comparações sistemáticas entre diferentes famílias de modelos, avaliando o desempenho preditivo com base em múltiplas métricas, intervalos de confiança e validação fora da amostra (ANDO; KIM, 2022, p. 10).

2.2 A prática internacional de previsão fiscal

A literatura internacional sobre previsão fiscal tem avançado significativamente nas últimas décadas, tanto sob a ótica metodológica quanto institucional. Esse desenvolvimento está intrinsecamente relacionado à crescente demanda por regras fiscais críveis, à necessidade de coordenação intertemporal das políticas públicas e à centralidade da transparência na governança orçamentária moderna. Nesse contexto, instituições como o Congressional Budget Office[2] (CBO), o Office for Budget Responsibility[3] (OBR), o Banco

[2] O CBO desempenha o papel de IFI nos Estados Unidos. Para mais informações, ver: https://www.cbo.gov/.

[3] O OBR desempenha o papel de IFI no Reino Unido. Para mais informações, ver: https://www.obr.uk/.

Central Europeu (BCE) e o Institute for Fiscal Studies[4] (IFS) desempenham papéis fundamentais no desenvolvimento e na disseminação de metodologias para projeção fiscal com elevado nível de rigor técnico.

No caso norte-americano, o CBO adota abordagem detalhada e estruturada para a elaboração de suas projeções macroeconômicas e fiscais. Em seu relatório metodológico, o órgão descreve o processo de formulação das previsões de longo prazo com base em modelos estruturais e análise de julgamento técnico. O modelo principal do CBO incorpora uma estrutura de equilíbrio geral com rigidez nominal e real, integrando expectativas racionais dos agentes, hipóteses sobre produtividade, taxas de juros e participação da força de trabalho (ARNOLD, 2018, p. 3). O órgão reconhece que essas projeções são altamente sensíveis a suposições sobre tendências demográficas e crescimento de produtividade, o que implica revisão constante dos parâmetros e validação periódica com dados históricos (ARNOLD, 2018, p. 12).

Além disso, o relatório sobre a elaboração da linha de base orçamentária detalha a forma como o CBO constrói suas previsões de despesa pública a partir de informações institucionais e projeções desagregadas por subfunções orçamentárias (STERN et al., 2023, p. 5). A previsão de cada subcomponente da despesa leva em conta regras legais, tendências históricas, pressões demográficas e parâmetros macroeconômicos projetados, assegurando a consistência entre os blocos orçamentários (STERN et al., 2023, p. 6). Essa estratégia de construção *bottom-up* serve de referência para a abordagem empírica adotada neste estudo.

Na experiência do Reino Unido, destaca-se a atuação do OBR, tanto pela sofisticação técnica quanto pelo compromisso com a transparência. Como exposto em OBR (2011a, p. 3), as previsões macroeconômicas e fiscais do OBR combinam modelos econométricos de demanda agregada com julgamento institucional, de modo a incorporar informações qualitativas de ministérios, agências e especialistas do setor público. Essa integração entre modelos formais e conhecimento tácito é fundamental para capturar aspectos institucionais frequentemente ausentes em abordagens puramente quantitativas.

No âmbito fiscal, o OBR emprega metodologia desagregada, projetando separadamente as categorias de despesa obrigatória, discricionária e encargos com juros (OBR, 2011b, p. 6). Os métodos utilizados vão desde extrapolações por tendência até projeções parametrizadas por regras legais, além de ajustes discricionários com base em eventos recentes. Segundo OBR (2011b, p. 10), esse tipo de ajuste é indispensável diante de mudanças legislativas ou reorganizações administrativas, ressaltando a importância do julgamento técnico no processo preditivo.

Outro aspecto relevante na atuação do OBR é a ênfase no monitoramento da exe-

[4] O IFS é o principal instituto independente de pesquisa econômica do Reino Unido. Para mais informações, ver: https://www.ifs.org.uk/.

cução fiscal ao longo do ano, com revisões periódicas das previsões à medida que dados de execução orçamentária se tornam disponíveis. Conforme descrito em OBR (2018, p. 7), as atualizações intra-anuais são realizadas com base em relatórios mensais e dados administrativos, permitindo revisar previsões de acordo com sinais de desvio no comportamento orçamentário. Esse processo torna-se especialmente importante em contextos de alta volatilidade política ou econômica, nos quais a rigidez das previsões estáticas compromete sua utilidade para a política fiscal.

No campo da comunicação da incerteza, o OBR desenvolveu práticas pioneiras de representação probabilística das projeções fiscais. O órgão utiliza gráficos de leque (*fan charts*) para expressar intervalos de confiança em torno das projeções do déficit e de outras variáveis fiscais, baseando-se na distribuição empírica dos erros de previsão passados (OBR, 2012, p. 3). Segundo OBR (2012, p. 4), tais gráficos são úteis para comunicar a amplitude de incerteza associada às projeções, ainda que não capturem eventos extremos com precisão. Essas práticas estão diretamente relacionadas com os objetivos do presente estudo, ao mensurar e comparar o desempenho preditivo de diferentes modelos atualizados automaticamente e com base não apenas em projeções pontuais, mas também em métricas intervalares, utilizando previsão conformal.

Ampliando a perspectiva institucional, o estudo de Leal et al. (2008), publicado pelo Institute for Fiscal Studies, oferece uma visão crítica e abrangente dos desafios inerentes às previsões fiscais. Os autores destacam o problema recorrente de vieses otimistas nas previsões oficiais, especialmente em anos eleitorais, e argumentam que tais distorções reduzem a credibilidade da política fiscal e minam a sustentabilidade das contas públicas (LEAL et al., 2008, p. 10). A tensão entre transparência e sofisticação técnica também é abordada, sendo apontado que modelos mais complexos, embora potencialmente mais acurados, tendem a ser menos compreensíveis para o público e mais difíceis de auditar (LEAL et al., 2008, p. 14). Os autores destacam ainda que a previsão da despesa pública é mais difícil que a da receita, devido à rigidez institucional de muitos gastos e à imprevisibilidade das políticas discricionárias. A literatura europeia aponta que, mesmo com regras fiscais, os erros de previsão de despesa persistem e se concentram em itens sujeitos à volatilidade política ou contábil (LEAL et al., 2008, p. 9).

As metodologias de previsão empregadas por essas instituições são diversas e frequentemente combinam diferentes abordagens em uma "suite de modelos". O Banco Central Europeu também adota essa estratégia, que busca equilibrar complexidade e simplicidade, ajuste empírico e solidez teórica. Os modelos macroeconométricos podem ser categorizados, por exemplo, em Modelos Dinâmicos Estocásticos de Equilíbrio Geral (DSGE), Modelos de Vetores Autorregressivos (VAR) e modelos semi-estruturais (CICCARELLI et al., 2024, p. 24). No campo fiscal, modelos DSGE "satélites" também são utilizados para analisar multiplicadores fiscais sob política monetária restrita, com orientação futura

e flexibilização quantitativa, frequentemente incluindo uma gama relativamente ampla de instrumentos de política fiscal (CICCARELLI et al., 2024, p. 87).

Outros autores, como Cimadomo, Giannone e Lenza (2017), propõem o uso de técnicas de *nowcasting* baseadas em modelos bayesianos com dados de frequência mista para prever saldos fiscais. O modelo utiliza dados de fluxo de caixa de maior frequência (mensal) para antecipar variações no saldo anual, contendo poucas variáveis e nenhum julgamento, demonstrando que abordagens mais parcimoniosas[5] podem superar previsões oficiais baseadas em centenas de variáveis (CIMADOMO; GIANNONE; LENZA, 2017, p. 10). A discussão crítica de Foroni (2017) sugere aprimoramentos como o uso de modelos de amostragem de dados de frequência mista (MiDaS) ou a decomposição do saldo em receitas e despesas, proposta alinhada ao escopo deste estudo.

O artigo de Asimakopoulos, Paredes e Warmedinger (2013) apresenta abordagem empírica para previsão de componentes da receita e da despesa com base em dados de alta frequência. Os autores utilizam modelos MiDaS para combinar dados trimestrais com metas anuais, e mostram que previsões desagregadas (*bottom-up*), atualizadas sempre que possível, geram ganhos preditivos expressivos, especialmente para o lado da despesa (ASIMAKOPOULOS; PAREDES; WARMEDINGER, 2013, p. 11). Essa conclusão fornece respaldo à estratégia empírica adotada neste estudo, que estima individualmente séries específicas da despesa pública e atualiza os modelos a cada nova informação disponível.

2.3 A prática nacional de previsão fiscal

A literatura nacional sobre previsões fiscais ainda é relativamente escassa e concentra-se majoritariamente em variáveis agregadas, como arrecadação tributária e resultado primário, havendo poucos estudos sobre despesa pública propriamente dita. De modo geral, os trabalhos brasileiros abordam três frentes: previsões macroeconômicas, previsões de arrecadação federal e estadual e análises de erros de execução orçamentária. A ausência de estudos focados na previsão desagregada da despesa federal reflete uma lacuna que este trabalho busca suprir.

No campo das previsões macroeconômicas, Kava (2022, p. 4) aplica métodos de aprendizado de máquina para prever séries brasileiras de inflação, atividade econômica e taxa de juros, comparando o desempenho de algoritmos como *Random Forest*, redes neurais e *Gradient Boosting* com modelos tradicionais, como ARMA e VAR. Os resultados mostram ganhos consistentes para modelos de *machine learning* em horizontes curtos, embora o autor ressalte a necessidade de procedimentos de validação cuidadosa e ajuste de

[5] Diversos autores argumentam em favor da parcimônia, uma versão da "Navalha de Occam", princípio que sustenta que modelos mais simples, com menos parâmetros, são geralmente preferíveis, devido ao *trade-off* entre viés e variância. Ver, por exemplo, Stoffi, Cevolani e Gnecco (2022) e Goldblum et al. (2024).

hiperparâmetros. Essa experiência, embora centrada em séries macroeconômicas, revela a viabilidade do uso de técnicas de aprendizado de máquina em dados brasileiros e introduz metodologias de interpretação agnóstica (KAVA, 2022, p. 6).

No âmbito da arrecadação tributária federal, diversos estudos exploraram a previsão e a combinação de modelos. Medeiros, Aragón e Besarria (2022, p. 8) compararam diferentes algoritmos de aprendizado supervisionado, como *Elastic Net*, *Complete Subset Regression* (CSR) e técnicas de *bagging*, para prever a arrecadação mensal de tributos federais entre 2002 e 2021. Os autores constataram que o *Elastic Net* apresentou melhor desempenho em horizontes curtos, seguido pela CSR, enquanto métodos de *bagging* apresentaram desempenho inferior, sobretudo para amostras menores ou horizontes mais longos. Além disso, *benchmarks* simples, como a média ou a mediana das previsões individuais, mostraram-se competitivos, corroborando a literatura internacional de combinação de previsões.

De modo similar, Gadelha, Lima e Polli (2020, p. 12) aplicaram técnicas de combinação simples e ótima de Bates e Granger (1969) para projetar a arrecadação de nove tributos federais, concluindo que a combinação de previsões supera consistentemente modelos individuais como SARIMA e o método de suavização exponencial tripla de Holt-Winters (HW). Resultados semelhantes foram obtidos por (MENDONçA; MEDRANO, 2016, p. 15), que mostraram ganhos de acurácia ao empregar modelos fatoriais dinâmicos associados a combinações lineares ponderadas para reduzir viés e erro quadrático médio. Tais trabalhos destacam a relevância de abordagens *ensemble* no contexto fiscal brasileiro, mesmo em cenários de baixa frequência e alta volatilidade institucional.

No que concerne à despesa, as evidências empíricas disponíveis são mais restritas e focam principalmente na análise de erros de projeção. Carneiro e Costa (2021, p. 7) analisaram os determinantes do erro de previsão de despesa nos municípios brasileiros, mostrando que categorias rígidas, como pessoal e encargos, apresentam maior acurácia, enquanto despesas discricionárias, como investimentos, tendem a ser sistematicamente subestimadas. Os autores identificaram que restos a pagar e práticas de incrementalismo orçamentário são fatores estruturais que perpetuam os erros, enquanto variáveis políticas, como anos eleitorais, tiveram menor impacto do que se supunha. Além disso, municípios com maior autonomia financeira e melhor qualidade de gestão fiscal apresentaram previsões mais precisas (CARNEIRO; COSTA, 2021, p. 9). Esses achados sugerem que, também no nível federal, a rigidez institucional pode facilitar a previsão de certas despesas, enquanto categorias discricionárias permanecem mais voláteis.

Deus e Mendonça (2017, p. 11) apresentam um avanço adicional ao analisar a qualidade das previsões fiscais agregadas no Brasil entre 2003 e 2013. O estudo revela a existência de viés otimista persistente, especialmente em anos eleitorais, bem como a influência dos erros de previsão do Produto Interno Bruto (PIB) sobre o resultado fiscal.

Os autores argumentam que a baixa eficiência das previsões está relacionada não apenas ao ciclo econômico, mas também à fragilidade institucional, que reduz os incentivos para projeções realistas. Esse diagnóstico dialoga com a literatura internacional ao demonstrar que erros fiscais em economias emergentes não são aleatórios, mas estruturais, refletindo tanto limitações técnicas quanto incentivos políticos (DEUS; MENDONçA, 2017, p. 14).

Apesar dessas contribuições, nota-se que a maior parte da literatura nacional concentra-se em receitas e saldos agregados, deixando de lado a previsão detalhada da despesa pública federal. Ademais, quase todos os estudos limitam-se a previsões pontuais, sem avaliação probabilística ou intervalar, e não realizam comparações sistemáticas entre famílias de modelos estatísticos, de aprendizado de máquina e de aprendizado profundo. Por fim, não encontramos estudos que adotem uma abordagem *bottom-up* para despesas federais, desagregando por categorias ou programas específicos, o que reforça a originalidade e a relevância deste estudo.

2.4 Previsão de séries temporais financeiras

A literatura recente sobre previsão de séries temporais tem avançado no desenvolvimento de abordagens capazes de capturar padrões complexos e melhorar a acurácia preditiva em diferentes domínios. Além dos métodos econométricos tradicionais, técnicas de aprendizado de máquina e de aprendizado profundo vêm sendo amplamente exploradas para lidar com não linearidades, múltiplos preditores e séries de maior complexidade. Além disso, à medida que cresce a disponibilidade de dados e o poder computacional, observa-se tendência para o uso de modelos mais flexíveis e intensivos em dados. Entretanto, ainda não há consenso quanto à superioridade de uma única família de modelos, sobretudo em séries curtas e de baixa frequência, como as séries mensais de despesa pública federal, foco deste estudo. Dada a escassez de trabalhos específicos sobre previsão fiscal, o conhecimento dos resultados de pesquisas empíricas e teóricas sobre séries temporais financeiras contribui para a compreensão do trabalho realizado e dos resultados obtidos.

Kaushik e Giri (2020, p. 5) realizaram análise comparativa multivariada entre um modelo tradicional (VAR), uma técnica de aprendizado supervisionado (SVM) e uma arquitetura de aprendizado profundo (LSTM) para prever a taxa de câmbio USD/INR (Dólar americano/Rupia indiana) com dados mensais relativamente curtos. Os autores mostraram que, embora as redes LSTM tenham apresentado melhor desempenho geral, os ganhos em relação ao SVM foram modestos e condicionados a um pré-processamento cuidadoso dos dados. Para séries curtas e ruidosas, o VAR permaneceu como *benchmark* competitivo, ainda que menos capaz de capturar não linearidades. Essa evidência reforça a importância de comparar diferentes famílias de modelos em contextos de dados limitados.

O papel central do pré-processamento no desempenho preditivo é enfatizado por Larson e Overton (2024, p. 8), que mostraram que transformações como normalização logarítmica, remoção de tendência e ajuste de sazonalidade podem ter impacto mais significativo na acurácia das previsões do que a escolha do modelo em si. Em estudo sobre previsão de impostos sobre vendas em cidades do Texas, os autores constataram que modelos de ML (XGBoost, kNN) só superaram ARIMA e ETS após um pipeline de pré-processamento robusto, evidenciando que o preparo adequado dos dados é determinante para o sucesso de métodos avançados.

Para compreender o desempenho relativo de diferentes famílias de modelos em múltiplos domínios e frequências, Godahewa et al. (2021, p. 12) organizaram o *Monash Time Series Forecasting Archive*, coleção de 25 conjuntos de dados, incluindo séries mensais curtas, e avaliaram métodos tradicionais (ETS, ARIMA, Theta), algoritmos globais de ML (CatBoost, FNN) e arquiteturas de DL (DeepAR, N-BEATS, Transformers). Os resultados mostraram que, para séries mensais de baixa frequência e tamanho reduzido, modelos teoricamente mais simples como Theta e ETS continuaram competitivos, enquanto modelos DL só se destacaram em séries longas e estáveis. Além disso, modelos globais que treinam múltiplas séries relacionadas conjuntamente apresentaram ganhos consideráveis quando havia similaridade estrutural entre as séries. Essa evidência é relevante para o presente estudo, que adota abordagem *bottom-up* por categoria de despesa e utiliza modelos treinados de forma global.

Achados semelhantes foram relatados pela competição M4, uma das maiores comparações empíricas já realizadas em previsão de séries temporais, envolvendo mais de 100 mil séries e 61 métodos distintos. (MAKRIDAKIS; SPILIOTIS; ASSIMAKOPOULOS, 2020, p. 17) mostraram que, para as séries mensais (quase metade do conjunto), métodos tradicionais como Theta, ETS e ARIMA permaneceram entre os melhores, enquanto DL puros apresentaram ganhos marginais em dados curtos e ruidosos. Os maiores avanços vieram de abordagens híbridas, como ES-RNN, e de combinações de previsões, que consistentemente superaram métodos individuais. Essa evidência é consistente com os achados de Bates e Granger (1969, p. 9), que provaram teoricamente que a combinação linear de previsões reduz o erro quadrático médio sempre que os erros individuais não são perfeitamente correlacionados.

A robustez empírica dessa conclusão foi confirmada na revisão abrangente de Clemen (1989, p. 11), que sintetizou 20 anos de estudos sobre combinação de previsões. O autor mostrou que combinações simples, como a média aritmética ou a mediana das previsões, tendem a apresentar desempenho tão bom quanto combinações ótimas baseadas em pesos estimados, especialmente em séries curtas nas quais a estimativa de covariâncias de erros é instável. Esses resultados justificam o uso de *ensembles* simples neste estudo como estratégia para aumentar a robustez das previsões de despesas públicas.

Além das previsões pontuais, a literatura recente destaca a importância de quantificar de forma confiável a incerteza associada às projeções. Xu e Xie (2021, p. 3) propuseram o uso de *conformal prediction* para séries temporais, apresentando o método *Ensemble Batch Prediction Intervals* (EnbPI), que fornece intervalos de previsão com garantias formais de cobertura, sem depender de suposições paramétricas. A utilização de intervalos de previsão conformais é especialmente relevante para o presente estudo, pois permite gerar intervalos calibrados para qualquer modelo de base, oferecendo avaliação mais completa da incerteza preditiva.

Por fim, a escolha e calibração dos modelos demandam procedimentos de seleção que evitem tanto o sobreajuste quanto a perda de informação. Nesse sentido, Bergmeir, Hyndman e Koo (2018, p. 6) demonstraram que a validação cruzada tradicional é válida para modelos autorregressivos sempre que os resíduos finais forem não correlacionados, e que esse procedimento fornece estimativas mais estáveis de erro preditivo do que divisões simples de treino e teste, especialmente em séries curtas. Tal conclusão legitima o uso de validação cruzada temporal no conjunto de treino, adotado neste estudo para selecionar hiperparâmetros e escolher os melhores modelos de previsão para o conjunto de teste.

Em conjunto, esses estudos mostram que, para séries mensais curtas e heterogêneas, como as despesas federais brasileiras, *benchmarks* estatísticos continuam relevantes, mas podem ser complementados por métodos de ML e DL bem ajustados. A combinação de previsões se apresenta como estratégia robusta para reduzir a variância dos erros, enquanto a previsão conformal oferece ferramenta avançada para quantificar a incerteza dos intervalos. Ademais, o uso de validação cruzada temporal fornece maior confiabilidade na escolha dos modelos em um cenário de dados limitados.

3 Metodologia

3.1 Base de dados

Adotamos o Boletim Resultado do Tesouro Nacional (RTN) como fonte dos dados de despesas públicas federais utilizadas no estudo. O RTN, publicação mensal da Secretaria do Tesouro Nacional (STN), é referência oficial para a mensuração do resultado primário do Governo Central brasileiro desde 1995. O boletim consolida, de forma transparente, as estatísticas fiscais referentes ao desempenho das receitas e despesas, e é reconhecido como o principal instrumento de acompanhamento da execução orçamentária federal e de análise da política fiscal corrente (STN, 2016, p. 3).

O RTN tem abrangência correspondente ao Governo Central, englobando administração pública federal direta, autarquias, fundações, fundos, empresas públicas dependentes, Banco Central do Brasil (BCB) e Regime Geral de Previdência Social (RGPS). Assim, as estatísticas fiscais do RTN refletem o resultado primário do Governo Central, diferentemente de demonstrativos que abrangem o setor público consolidado ou todos os entes federados. Os resultados fiscais consolidados pelo RTN incluem ainda receitas e despesas previdenciárias e contas do Tesouro Nacional, incorporando transferências constitucionais e legais para estados e municípios (STN, 2016, p. 6).

A apuração do RTN baseia-se em fluxos de caixa registrados no Sistema Integrado de Administração Financeira do Governo Federal (SIAFI), complementados por informações da Secretaria da Receita Federal do Brasil (RFB), do BCB e do Instituto Nacional do Seguro Social (INSS). O SIAFI registra, acompanha e controla a execução orçamentária, financeira e patrimonial do governo federal. O fluxo de apuração inicia-se com a coleta e o processamento dos dados primários, seguido de rotinas de consolidação, conferência e ajustes, culminando na elaboração dos demonstrativos do RTN (STN, 2016, p. 4).

Informações externas são usadas para confirmação e ajuste dos registros, assegurando a qualidade e integridade dos dados fiscais. A apuração do resultado primário ocorre pela diferença entre receitas primárias líquidas e despesa primária total, sendo o resultado positivo (superávit) ou negativo (déficit), e calculado em valores mensais e acumulados no ano. O RTN divulga comparações entre diferentes períodos e valores corrigidos pelo Índice Nacional de Preços ao Consumidor Amplo (IPCA), para análise em termos reais (STN, 2016, p. 5).

Distinguem-se, nas estatísticas fiscais brasileiras, dois conceitos fundamentais: resultado nominal e resultado primário. O resultado nominal representa a diferença entre receitas totais (incluindo financeiras) e despesas totais (incluindo juros nominais), refle-

tindo a necessidade total de financiamento do setor público. O resultado primário, por sua vez, corresponde ao resultado nominal excluindo as despesas com juros incidentes sobre a dívida líquida, captando o esforço fiscal do governo sem considerar o serviço da dívida. No Brasil, o RTN apura e divulga prioritariamente o resultado primário do Governo Central, indicador central para formulação e monitoramento das metas fiscais (STN, 2016, p. 6).

Metodologicamente, distinguem-se as abordagens "acima da linha" e "abaixo da linha" para a mensuração do resultado fiscal. O RTN adota a metodologia acima da linha, apurando o resultado fiscal pela diferença entre fluxos de receitas e despesas, o que fornece retrato detalhado da execução orçamentária e financeira. Essa abordagem permite identificar os determinantes do resultado fiscal corrente e realizar o acompanhamento da política fiscal de forma transparente. Em contraste, a metodologia abaixo da linha, utilizada pelo Banco Central, baseia-se na variação do saldo da dívida líquida do setor público, evidenciando fontes de financiamento e necessidades líquidas do setor público. No Brasil, o RTN apura o resultado primário do Governo Central pelo critério "acima da linha" (STN, 2016, p. 7).

Outro ponto fundamental refere-se aos regimes de reconhecimento das receitas e despesas: caixa e competência. O regime de caixa registra os ingressos e desembolsos efetivamente realizados no período; o de competência reconhece receitas e despesas conforme o fato gerador, independentemente do fluxo financeiro. O RTN utiliza, para apuração do resultado primário, o critério acima da linha no regime caixa, considerando receitas e despesas na data do efetivo recebimento ou pagamento, conforme a prática consolidada das estatísticas fiscais brasileiras (STN, 2016, p. 8). Para o resultado nominal apurado pelo Banco Central, adota-se o regime de competência para despesas financeiras líquidas, resultando em abordagem híbrida no cálculo das necessidades de financiamento do setor público.

O RTN apresenta estrutura detalhada de receitas e despesas primárias, segmentando as receitas em categorias administradas pela Receita Federal, receitas previdenciárias, receitas próprias do Tesouro Nacional, entre outras; e as despesas em benefícios previdenciários, pessoal e encargos sociais, outras despesas obrigatórias, despesas discricionárias e transferências. Essas classificações permitem análises desagregadas e facilitam o acompanhamento da execução orçamentária de cada componente, servindo como base para estudos empíricos que investigam padrões de comportamento ou avaliem impactos de políticas públicas específicas (STN, 2016, p. 17).

A Secretaria do Tesouro Nacional adota práticas de transparência ativa, disponibilizando não apenas os boletins mensais do RTN, mas também as séries históricas completas, metadados detalhados, bases em formato aberto e informações metodológicas em seu portal institucional. O acesso amplo aos dados[1] e a publicação dos critérios

[1] A atualização dessas informações é automatizada. A planilha mais atualizada pode ser encontrada

de apuração contribuem para a legitimidade, o controle social e a qualidade dos estudos acadêmicos e institucionais que utilizam essas informações (STN, 2016, p. 19). Para fins deste estudo, utilizamos a série histórica das despesas públicas federais, em valores reais, apurada pelo RTN, considerando que o resultado primário do Governo Central é calculado pelo critério "acima da linha" no regime caixa (STN, 2016, p. 8). Tal escolha fornece aderência metodológica, comparabilidade e solidez empírica às análises desenvolvidas.

3.2 Escolha das variáveis

O RTN apresenta 159 linhas de variáveis, abrangendo receitas (57), despesas (92) e resultados fiscais (10). O escopo do presente estudo concentra-se exclusivamente nas variáveis associadas às despesas públicas primárias. Para a construção da base de dados, definimos como data de corte o início da série histórica analisada. Embora o RTN disponha de dados desde 1997 para despesas agregadas, o nível de detalhamento necessário, que permite identificação e análise individualizada de linhas específicas de despesa, só está disponível a partir de janeiro de 2010. Limitamos a série temporal a este ponto de partida, assegurando comparabilidade e homogeneidade do detalhamento ao longo do período analisado, evitando problemas de dados faltantes, quebras de série ou inconsistências classificatórias.

Quanto ao nível de granularidade das previsões, inicialmente selecionamos 41 linhas de despesa, constituindo o menor grupo que identifica unicamente cada despesa do RTN, sem redundâncias ou combinações lineares triviais. Demais linhas disponíveis correspondem a detalhamentos mais específicos, sem interesse preditivo, ou resultam de agregações das variáveis já selecionadas, não acrescentando informação ao modelo. Durante análise exploratória, observamos muitos meses com valores iguais a zero em algumas variáveis do conjunto original com 41 linhas de despesa.

Identificamos ainda que 18 dessas variáveis apresentavam valores individualmente irrelevantes sob a ótica da materialidade orçamentária, cada uma representando fração muito pequena do total de despesas federais. Como critério objetivo, consideramos materialmente irrelevantes as variáveis cuja soma representa menos de 5% do total de despesas no período analisado. Para evitar superajuste em variáveis pouco informativas, agregamos em duas novas variáveis: uma com a soma de 11 despesas obrigatórias e outra com a soma de 7 despesas discricionárias. Adicionalmente, concatenamos em uma única variável as duas linhas de despesas com sentenças judiciais e precatórios de benefícios previdenciários (urbano e rural).

Questão relevante refere-se à presença de valores faltantes ou nulos nas séries. Observamos grande número de meses com valores iguais a zero em algumas variáveis do

em: http://sisweb.tesouro.gov.br/apex/cosis/thot/link/rtn/serie-historica?conteudo=cdn.

conjunto original com 41 linhas de despesa. Destaca-se, contudo, que tais registros não correspondem a dados faltantes ou omissões, mas a situações reais de ausência de despesa na respectiva linha e período. Ou seja, os zeros refletem a ausência efetiva de execução orçamentária, não problemas de qualidade da informação. Portanto, não realizamos imputação de dados para substituição dos zeros estruturais.

O agrupamento de variáveis reduziu o número de séries analisadas de 41 para 24 e eliminou 696 pontos de dados zerados. As novas variáveis agregadas não apresentam valores zero ao longo do período. Apenas 4 das 24 séries finais exibem algum valor zero em determinado mês[2]. Ressalte-se que essa baixa proporção de zeros não caracteriza cenário típico de "inflação de zeros" (*zero-inflated models*), que exigiria técnicas específicas de análise, conforme literatura desenvolvida inicialmente por Croston (1972). Tal abordagem parcimoniosa contribui para a robustez da análise e evita distorções pela inclusão de variáveis pouco informativas ou tratamento desnecessário dos dados. A Tabela 16 (Apêndice A) descreve as variáveis do RTN utilizadas.

A Figura 1 apresenta a evolução, e a Figura 2, a sazonalidade das despesas primárias, ambas em milhões de R$ corrigidas pelo IPCA até junho de 2025 (última informação disponível) e para o período de janeiro de 2010 a dezembro de 2022, definido como conjunto de treino. De acordo com Hyndman et al. (2025), o conjunto de teste normalmente representa cerca de 20% da amostra, embora esse valor dependa do tamanho da amostra e do horizonte de previsão. Embora a base vá até junho de 2025, toda a análise exploratória de dados é realizada sobre o período de treino (13 anos ou 156 pontos), sendo o restante dos dados usado como teste (30 pontos), evitando o *data leakage*[3].

[2] Quais sejam: i) Abono Salarial e Lei Kandir, com 31 contagens cada (16,8% dos pontos), ii) FUNDEB, com 8 contagens (4,3%), iii) sentenças judiciais e precatórios do BPC LOAS/RMV, com 7 contagens (3,8%).

[3] Data leakage ocorre quando informações do conjunto de teste (ou de dados futuros) são, direta ou indiretamente, usadas no treinamento do modelo, gerando avaliações artificialmente superiores. Para aprofundamento, ver (APICELLA; ISGRò; PREVETE, 2025).

Figura 1 – Evolução das despesas primárias em termos reais (milhões de R$)

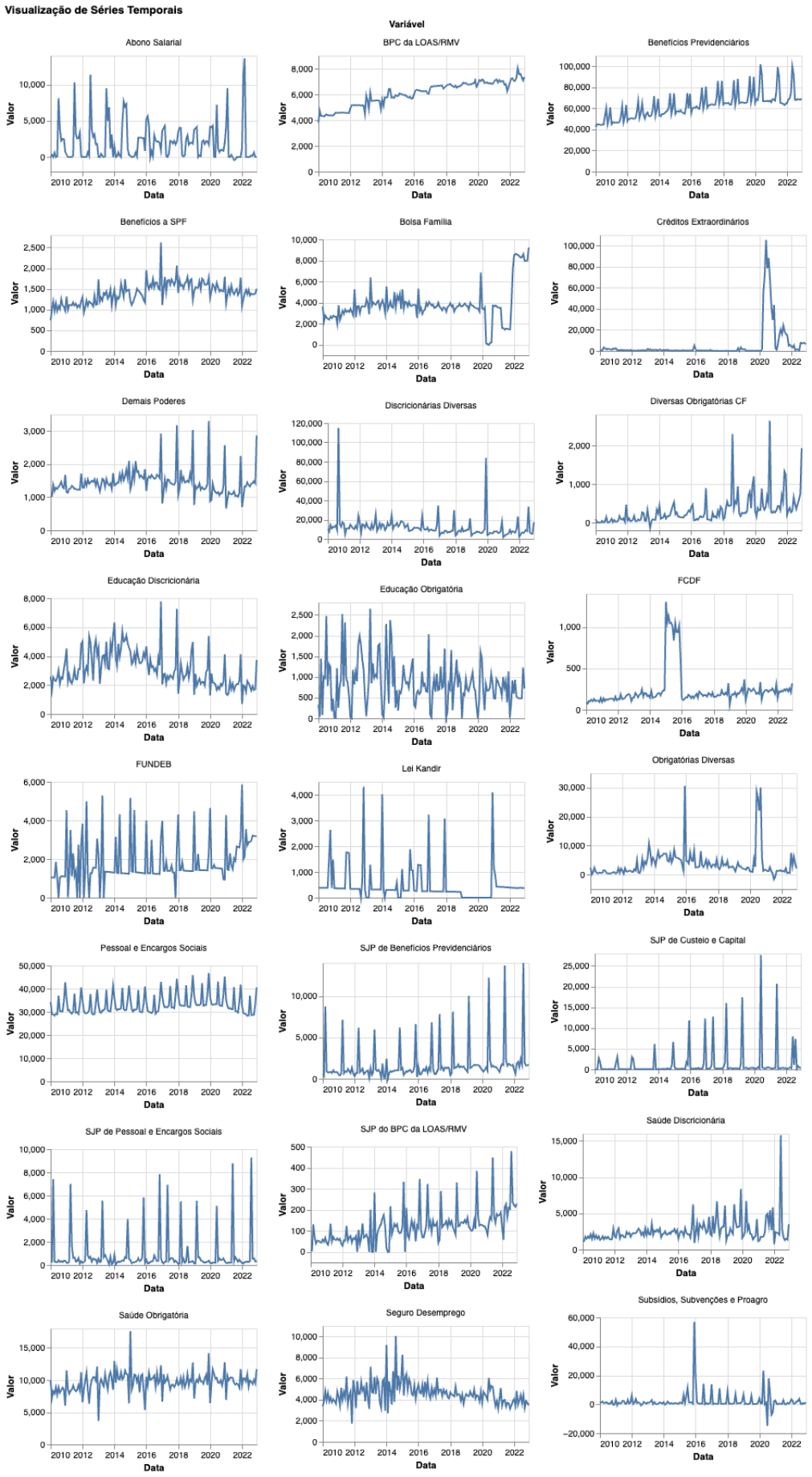
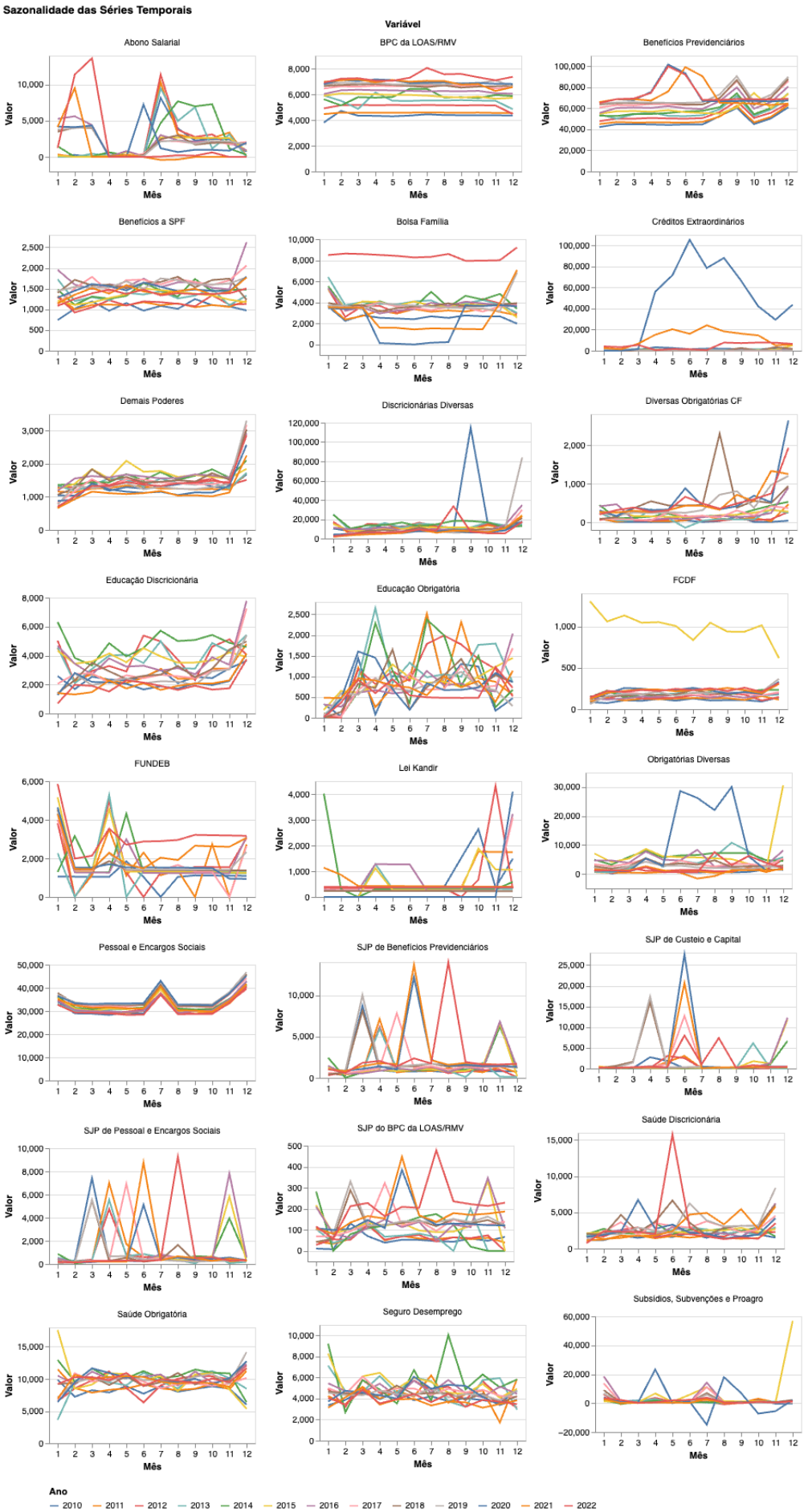


Figura 2 – Sazonalidade das despesas primárias em termos reais (milhões de R$)



Com base nas figuras, identificamos que as séries apresentam comportamento bastante diverso entre si, o que indica que dificilmente um único modelo representa bem todas as séries. Tal fato reforça a escolha da abordagem *bottom-up*. Quanto à tendência, algumas séries, como Benefícios Previdenciários e Benefícios de Prestação Continuada (BPC) da LOAS/RMV, apresentam crescimento persistente, refletindo fatores demográficos e mudanças normativas. Outras séries, como Educação Discricionária e Obrigatória, apresentam decréscimo, enquanto Saúde e Pessoal e Encargos Sociais mostram estabilidade relativa, sugerindo rigidez institucional.

Quanto à sazonalidade, algumas categorias exibem padrões sistemáticos ao longo dos meses do ano (ex: Pessoal e Encargos Sociais e Benefícios Previdenciários). Sazonalidade pronunciada tende a facilitar a previsão, desde que capturada adequadamente pelos modelos e que seja constante ao longo do tempo, o que não é o caso para várias despesas analisadas. Por outro lado, diversas despesas apresentam comportamento irregular, com picos em meses distintos, como Sentenças Judiciais e Precatórios (SJP), decorrente da natureza discricionária do pagamento de precatórios. O componente cíclico, entendido como movimentos de longo prazo associados a flutuações econômicas ou institucionais, é mais difícil de isolar, mas pode ser sugerido por choques ou mudanças de nível, sobretudo em períodos de crise. Um caso notório é a substituição do Bolsa Família pelo Auxílio Emergencial na pandemia de Covid-19.

Por fim, notamos que a irregularidade é marcante em séries como Abono Salarial, FUNDEB e Sentenças Judiciais e Precatórios, com valores atípicos e rupturas de tendência. Comportamentos imprevisíveis, associados a fatores externos ou decisões discricionárias, dificultam a previsão, exigindo modelos flexíveis capazes de lidar com não linearidades e rupturas estruturais. Essa diversidade reforça a importância de múltiplos métodos e validações rigorosas.

3.3 Processo de avaliação e seleção de modelos

Estruturamos o processo de avaliação e seleção dos modelos preditivos em um *pipeline* modular, alinhado às melhores práticas em previsão de séries temporais, descritas nos Capítulos 5 e 13 do livro de Hyndman et al. (2025). O objetivo é selecionar o modelo mais adequado para cada série temporal, dadas as suas características individuais, sem desprezar as informações que podemos obter ao utilizar modelos multivariados[4].

[4] De acordo com Hyndman et al. (2025), um modelo multivariado modela explicitamente as interações entre múltiplas séries temporais em um conjunto de dados e fornece previsões para múltiplas séries temporais simultaneamente. Em contraste, um modelo univariado treinado em múltiplas séries temporais modela implicitamente as interações entre múltiplas séries temporais e fornece previsões para séries temporais únicas simultaneamente. Modelos multivariados são tipicamente custosos em termos computacionais e, empiricamente, não oferecem necessariamente melhor desempenho de previsão em comparação com o uso de um modelo univariado.

No âmbito da previsão e análise de séries temporais, uma das tarefas mais complexas é identificar o modelo mais adequado para um grupo específico de séries. Muitas vezes, esse processo de seleção depende fortemente da intuição, que pode não necessariamente se alinhar à realidade empírica do nosso conjunto de dados. Fazemos isso por meio de uma técnica de validação cruzada com base fixa (*"expanding window"*). A validação cruzada para séries temporais avalia como um modelo teria se comportado no passado, usando uma janela deslizante nos dados históricos para prever períodos futuros. Conforme destacam Hyndman et al. (2025), os erros de medida dos resíduos por validação cruzada são maiores, visto que as previsões correspondentes são baseadas em um modelo ajustado a uma parte do conjunto de dados. Entretanto, esse processo se aproxima mais da geração de erros de previsões verdadeiras, por simular o processo natural em que as novas informações são agregadas à base de dados com o passar do tempo.

Uma boa maneira de escolher o melhor modelo de previsão é encontrar aquele com o menor MAE ou RMSE calculado usando validação cruzada de séries temporais. Conforme explicam Hyndman et al. (2025), ao comparar métodos de previsão aplicados a uma única série temporal ou a várias séries com as mesmas unidades, utilizamos o MAE, por ser mais popular e fácil de entender e calcular. Um método de previsão que minimize o MAE leva a previsões da mediana, enquanto a minimização do RMSE leva a previsões da média. Consequentemente, o RMSE também é amplamente utilizado, apesar de ser mais difícil de interpretar. Treinamos uma variedade de modelos de vários paradigmas de previsão: modelos de referência, técnicas estatísticas, aprendizado de máquina e aprendizado profundo. Ao final, utilizamos a técnica de combinação de previsões. Essa abordagem ajuda a selecionar aquele que apresenta o melhor desempenho para cada série temporal dentre cada paradigma e a estimar o desempenho preditivo de nossos modelos, com base na combinação de previsões. Aplicamos o procedimento individualmente a cada série de despesa e seguimos as seguintes etapas:

Divisão dos dados em treino e teste

Dividimos os dados em dois subconjuntos disjuntos: treino ($\mathcal{T}$) e teste ($\mathcal{S}$), sempre mantendo a ordem temporal das observações. Utilizamos o conjunto de treino para ajuste, escolha e validação dos modelos, enquanto o conjunto de teste serve para mensurar o desempenho preditivo fora da amostra, simulando a tarefa real de previsão futura.

Instanciação e ajuste dos modelos

Para cada série i, tempo t e modelo M_k, ajustamos os parâmetros a partir dos dados de treino ($\mathcal{T}$), considerando periodicidade mensal ($s = 12$) e seleção automática de hiperparâmetros:

$$\hat{\theta}_{i,k} = \arg\min_{\theta} L(y_t^{(i)}, \hat{y}_{t,k}^{(i)}(\theta)), \quad t \in \mathcal{T},$$

onde:

- i: índice da série temporal de despesa ($i = 1, 2, ..., n_{series}$);
- t: índice temporal da observação no conjunto de treino;
- M_k: modelo preditivo de índice k;
- θ: vetor de parâmetros do modelo;
- L: função de perda (exemplo: MAE ou RMSE);
- $y_t^{(i)}$: valor observado da série i no tempo t;
- $\hat{y}_{t,k}^{(i)}$: valor previsto pelo modelo M_k com parâmetros ajustados;

Diagnóstico dos resíduos

Analisamos os resíduos $e_{t,k}^{(i)} = y_t^{(i)} - \hat{y}_{t,k}^{(i)}$ nos dados de treino quanto à ausência de autocorrelação (teste de Ljung-Box), homocedasticidade (teste ARCH de Engle), média zero e normalidade. De acordo com Hyndman et al. (2025), um bom método de previsão produz resíduos com duas propriedades principais: (i) os resíduos não devem ser correlacionados; e (ii) devem ter média zero. Se houver correlação entre os resíduos, então há informações restantes que devem ser capturadas pelo modelo. Se apresentarem média diferente de zero, então as previsões são tendenciosas. Segundo os autores, qualquer método de previsão que não satisfaça essas propriedades pode ser melhorado. No entanto, isso não significa que os métodos que as satisfaçam não possam ser aperfeiçoados.

Além dessas propriedades essenciais, é útil, mas não necessário, que os resíduos também tenham outras duas características: (iii) variância constante ("homocedasticidade"); e (iv) distribuição normal. Essas duas propriedades facilitam o cálculo dos intervalos de previsão. No entanto, conforme destacam Hyndman et al. (2025), um método que não as satisfaça não necessariamente pode ser melhorado. Às vezes, aplicar uma transformação pode ajudar, mas geralmente há pouco que possamos fazer para garantir que os resíduos tenham variância constante e distribuição normal. Em vez disso, os autores recomendam a adoção de uma abordagem alternativa para obtenção de intervalos de predição.

Intervalos de previsão conformais

De acordo com Hyndman et al. (2025), enquanto intervalos de previsão tradicionais dependem de hipóteses sobre a distribuição das previsões, tipicamente assumindo normalidade, a previsão conformal oferece uma alternativa flexível e agnóstica ao modelo, que não exige pressupostos fortes. Em vez de estimar intervalos com base em um modelo probabilístico predefinido, a previsão conformal constrói intervalos utilizando os erros passados, garantindo que a cobertura empírica esteja próxima do nível de confiança desejado. Assim, os intervalos de previsão conformais são geralmente construídos usando resíduos passados para calibrar a incerteza futura. Esses resíduos são calculados a partir de um conjunto de calibração, que consiste nos erros de h-passos à frente, definidos por $e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$, que servem de base para estimar a incerteza das previsões futuras.

Uma suposição fundamental na previsão conformal, conforme destacam Hyndman et al. (2025), é a de permutabilidade (*exchangeability*), isto é, a distribuição dos resíduos passados deve ser representativa dos erros futuros. Essa hipótese permite estimar a distribuição dos erros de previsão utilizando seus valores absolutos, assegurando que os intervalos construídos tenham cobertura empírica válida, mesmo quando a distribuição subjacente é desconhecida. Embora essa hipótese seja a base da abordagem conformal clássica, pesquisas recentes, como as de Barber et al. (2023), vêm explorando métodos que vão além dessa suposição, possibilitando aplicações em contextos de séries temporais. Essa flexibilização da permutabilidade é fundamental na construção de intervalos de confiança em séries temporais, dado que existe clara dependência temporal entre os dados.

Hyndman et al. (2025) propõem um método simples de *split conformal* para construir um intervalo de previsão $(1 - \alpha)$ no horizonte h:

- primeiro ajustamos um modelo de previsão aos dados históricos e calculamos os resíduos das observações passadas; e
- em seguida, estimamos o intervalo de previsão usando os quantis empíricos dos valores absolutos desses resíduos.

Para um nível de confiança $(1 - \alpha)$, o intervalo conformal é dado por:

$$\hat{y}_{T+h|T} \pm Q_{1-\alpha}(|e_{t+h|t}|),$$

onde $Q_{1-\alpha}(|e_{t+h|t}|)$ é o quantil $(1 - \alpha)$ dos resíduos absolutos passados para o erro de h-passos à frente.

Stankevičiūtė, Alaa e Schaar (2021) destacam que essa abordagem garante que os intervalos tenham cobertura empírica próxima a $(1 - \alpha)$, mesmo quando a distribuição subjacente seja desconhecida ou não siga uma normal.

Validação cruzada temporal no conjunto de treino

Para cada série e modelo, realizamos a validação cruzada temporal do tipo *expanding window* apenas dentro do conjunto de treino ($\mathcal{T}$). Em cada rodada j, treinamos o modelo em uma janela crescente de dados e avaliamos as próximas h observações:

$$\hat{y}_{t,k,j}^{(i)} = M_k(\mathcal{T}_j), \quad t \in \mathcal{V}_j \subset \mathcal{T},$$

onde:

- $\mathcal{T}$: conjunto de treino;
- j: rodada de validação cruzada temporal ($j = 1, 2, ..., J$);
- $\hat{y}_{t,k,j}^{(i)}$: valor previsto para a série i, no tempo t, pelo modelo M_k, na j-ésima rodada de validação cruzada;
- $M_k(\mathcal{T}_j)$: operador que representa o ajuste do modelo k com os dados de treino da janela $\mathcal{T}_j$ e a posterior geração de previsões para os tempos em $\mathcal{V}_j$;
- $\mathcal{T}_j$: subconjunto de treino para a rodada j (janela expansiva de dados históricos até o tempo t_j); e
- $\mathcal{V}_j$: janela de validação para a rodada j, contendo os h períodos imediatamente posteriores a $\mathcal{T}_j$, todos ainda contidos no conjunto de treino $\mathcal{T}$.

Em resumo, em cada rodada j da validação cruzada, ajustamos o modelo M_k aos dados históricos de $\mathcal{T}_j$ e o utilizamos para prever os próximos h valores da série i em $\mathcal{V}_j$, gerando as previsões $\hat{y}_{t,k,j}^{(i)}$.

Avaliação das métricas de erro na validação cruzada

Para cada série i, modelo M_k, e rodada j, calculamos as métricas de erro nas predições da validação cruzada:

- $MAE_k^{(i)} = \frac{1}{N_{val}} \sum_{j=1}^{J} \sum_{t \in \mathcal{V}_j} |y_t^{(i)} - \hat{y}_{t,k,j}^{(i)}|$;
- $RMSE_k^{(i)} = \sqrt{\frac{1}{N_{val}} \sum_{j=1}^{J} \sum_{t \in \mathcal{V}_j} (y_t^{(i)} - \hat{y}_{t,k,j}^{(i)})^2}$.

onde:

- $e_{t,k,j}^{(i)} = y_t^{(i)} - \hat{y}_{t,k,j}^{(i)}$: resíduo na validação cruzada; e
- N_{val}: número total de pontos na validação cruzada.

Seleção do melhor modelo no conjunto de treino

Após a validação cruzada, para cada série i, selecionamos o modelo M_i^* que apresentou a menor métrica de erro médio no conjunto de treino:

$$M_i^* = \arg\min_{M_k \in \mathcal{M}} MAE_k^{(i)},$$

onde:

- $\hat{y}_{t,M_i^*}^{(i)}$: valor previsto pelo modelo escolhido, no teste; e
- o vetor $M_i^* = (M_1^*, M_2^*, ..., M_{n_{series}}^*)$ define o modelo escolhido para cada série.

Empregamos como critério principal de seleção o menor erro absoluto médio (MAE), mas consideramos o RMSE como métrica complementar para capturar potenciais distorções associadas a valores extremos.

Previsão e avaliação no conjunto de teste

Treinamos cada modelo selecionado, por série, em todos os dados do conjunto de treino ($\mathcal{T}$) e utilizamos o modelo "fitado" para gerar previsões no conjunto de teste ($\mathcal{S}$), fora da amostra. Para cada tempo $t' \in \mathcal{S}$, a previsão para a série i é:

$$\hat{y}_{t',M_i^*}^{(i)} = M_i^*(\mathcal{T}; t'),$$

onde:

- M_i^*: modelo selecionado para a série i com base na validação cruzada do treino;
- $\mathcal{T}$: conjunto de treino usado para ajuste (*fit*);
- $t' \in \mathcal{S}$: índice temporal do conjunto de teste;
- $\hat{y}_{t',M_i^*}^{(i)}$: valor previsto para a série i, no tempo t', usando o modelo M_i^* ajustado em $\mathcal{T}$.

Calculamos novamente as métricas de erro neste conjunto, permitindo avaliar a capacidade preditiva dos modelos em dados não utilizados para calibração ou seleção. Para acrescentar robustez ao modelo, avaliamos o ajuste fora da amostra por meio de métricas de erro absoluto (MAE e MASE), quadrático (MSE, RMSE, MSSE e RMSSE) e percentual (MAPE e SMAPE).

Combinação de previsões

Para melhorar a acurácia preditiva das séries temporais, implementamos a estratégia de combinação de previsões (*forecast combination*), conforme recomendado por Hyndman et al. (2025). A literatura demonstra de forma consistente que a combinação de previsões oriundas de múltiplos modelos frequentemente resulta em desempenho superior à utilização isolada de qualquer modelo individual. A abordagem mais simples e, surpreendentemente, difícil de superar em termos de desempenho consiste na média aritmética simples das previsões individuais, conforme destacado por Wang et al. (2023).

Sejam $\hat{y}_t^{(\ell)}$ as previsões de L diferentes paradigmas de modelos para a série i no tempo t, a previsão combinada (*ensemble*) é definida como:

$$\hat{y}_t^{(\text{ensemble})} = \frac{1}{L} \sum_{\ell=1}^{L} \hat{y}_t^{(\ell)},$$

onde L é o número de paradigmas de modelos combinados (neste trabalho, $L = 3$: estatístico, aprendizado de máquina e aprendizado profundo).

Como a combinação de previsões não resulta diretamente de um modelo estatístico tradicional, os intervalos de previsão não são obtidos via distribuição analítica dos erros, mas sim por métodos empíricos. Utilizamos a abordagem de intervalos de previsão conformais, que estima a incerteza da previsão combinada a partir dos resíduos absolutos observados em uma amostra de calibração ou validação, conforme proposto por Hyndman et al. (2025). Os resíduos da combinação são dados por:

$$e_t = y_t - \hat{y}_t^{(\text{ensemble})},$$

e, para um nível de confiança $(1 - \alpha)$, os limites inferior e superior do intervalo conformal para a previsão combinada são construídos usando o quantil empírico $Q_{1-\alpha}$ dos resíduos absolutos:

$$\hat{y}_t^{(\text{ensemble})} \pm Q_{1-\alpha}(|e_t|)$$

onde $Q_{1-\alpha}(|e_t|)$ é o quantil empírico de ordem $(1 - \alpha)$ dos valores absolutos dos resíduos da combinação.

3.4 Modelos

A fundamentação teórica desta pesquisa apoia-se em obras de referência que cobrem diferentes paradigmas de modelagem preditiva. Hamilton (1994) oferece uma visão abrangente sobre modelos estatísticos lineares e não lineares de séries temporais, discutindo tanto os fundamentos teóricos quanto as dificuldades práticas de análise e interpretação de dados do mundo real, além de integrar conceitos da teoria econômica à modelagem temporal.

Complementarmente, Hastie, Tibshirani e Friedman (2009) constitui uma das referências mais influentes em estatística aplicada, aprendizado de máquina e modelagem preditiva, apresentando de forma rigorosa a teoria por trás de diversos algoritmos supervisionados e não supervisionados, como métodos de regressão, árvores de decisão, técnicas de ensemble e regularização, que embasam a compreensão e a interpretação dos modelos de aprendizado de máquina empregados neste estudo.

No campo do aprendizado profundo, Goodfellow, Bengio e Courville (2016) é considerada a principal referência acadêmica, sistematizando desde os fundamentos matemáticos e estatísticos até arquiteturas modernas de redes neurais, fornecendo o suporte teórico necessário para entendermos o funcionamento e as propriedades dos modelos de deep learning aplicados à previsão das despesas públicas analisadas.

Treinamos 46 modelos para cada uma das 24 despesas primárias avaliadas no presente estudo. Inicialmente, utilizamos 6 modelos de base, abarcando média histórica (*Historic Average*), modelos ingênuos (*Naive*, *Seasonal Naive* e *Random Walk With Drift*) e médias móveis simples e sazonais (*Window Average* e *Seasonal Window Average*). Esses modelos de base serviram como *benchmark* para avaliar o desempenho dos demais modelos. Em seguida, utilizamos 6 modelos tradicionais de previsão de séries temporais: ARIMA, ETS, Theta, *Complex Exponential Smoothing* (CES), *Median, Fourier seasonality, Linear trend, and Exponential Smoothing* (MFLES) e *Trigonometric, ARMA errors, Box-Cox transformation, Trend, and Seasonality* (TABTS). Também avaliamos 8 modelos de aprendizado de máquina: *Random Forest*, *Elastic Net*, *Lasso*, *Ridge*, *Linear Regressor*, *CatBoost*, *XGBoost* e *Light GBM*.

Exploramos ainda a eficiência de 26 modelos de aprendizado profundo, separados em 5 classes de arquiteturas. Avaliamos 7 modelos da classe de Redes Neurais Recorrentes: *Recurrent Neural Networks* (RNN), LSTM, *Gated Recurrent Units* (GRU), *Temporal Convolutional Networks* (TCN), *Deep Autoregressive* (DeepAR), *Dilated Recurrent Neural Network* (DilatedRNN) e *Bidirectional Temporal Convolutional Networks* (BiTCN). Também testamos 7 modelos da classe de Perceptron Multicamadas: *Multilayer Perceptron* (MLP), *Neural Basis Expansion Analysis for Time Series* (NBEATS), *Neural Hierarchical Interpolation for Time Series* (NHITS), *Decomposition Linear Model* (DLinear),

Nonlinear Forecasting Model (NLinear), *Time-series Dense Encoder* (TiDE) e *Deep Non-Parametric Time Series* (DeepNPTS).

Testamos ainda 6 modelos baseados em transformers: *Temporal Fusion Transformer* (TFT), *Vanilla Transformer*, *Informer*, *Autoformer*, *Frequency Enhanced Decomposed Transformer* (FEDformer) e *Patch-based Time Series Transformer* (PatchTST). Por fim, avaliamos 4 modelos multivariativos: *Spectral-Temporal Graph Neural Network* (StemGNN), *Time Series Token Mixing* (TSMixer), *Multivariate Multilayer Perceptron* (MLP-Multivariate) e *State-Of-The-Forecast Time Series* (SOFTS), além de 2 modelos com arquiteturas diversas: *Kolmogorov-Arnold Network* (KAN) e *TimesNet Convolutional Architecture* (TimesNet). Ao final, também analisamos os melhores modelos independentemente das classes designadas anteriormente.

Para testar essa grande quantidade de modelos, recorremos ao uso de otimização automática de hiperparâmetros, conforme os trabalhos de Hyndman e Khandakar (2008), Akiba et al. (2019), Garza et al. (2022) e Olivares et al. (2022). A otimização automática de hiperparâmetros desempenha papel central na melhoria do desempenho de modelos de aprendizado de máquina, especialmente em contextos em que a busca manual é inviável ou ineficiente. Nesse sentido, Akiba et al. (2019) propõem uma abordagem de última geração baseada em dois princípios fundamentais: a definição dinâmica do espaço de busca (*define-by-run*) e a amostragem eficiente com técnicas de otimização Bayesiana. O mecanismo *define-by-run* permite que o espaço de hiperparâmetros seja construído de forma programática durante a execução da função objetivo, conferindo maior flexibilidade e expressividade à definição condicional de parâmetros. A amostragem é realizada predominantemente por meio do algoritmo *Tree-structured Parzen Estimator* (TPE), que modela separadamente a distribuição de boas e más configurações, priorizando regiões mais promissoras do espaço de busca.

Além disso, Akiba et al. (2019) incorporam técnicas de *pruning* para interromper precocemente avaliações com baixo potencial de desempenho, o que reduz significativamente o custo computacional da otimização. Essa estratégia se baseia em monitoramento contínuo de métricas parciais durante o treinamento e é especialmente útil em tarefas de alto custo, como o ajuste de redes neurais profundas. Os experimentos apresentados pelos autores demonstram que o modelo proposto supera outros *frameworks* populares, tanto no tempo de convergência quanto na qualidade das soluções encontradas, mesmo sob restrições de orçamento computacional. De acordo com os autores, a arquitetura leve e o suporte à execução paralela e distribuída compatível com múltiplas bibliotecas tornam a ferramenta robusta e altamente adaptável para aplicações reais de modelagem preditiva.

4 Resultados

Nesta seção, apresentamos e analisamos os resultados obtidos a partir dos modelos estatísticos, de aprendizado de máquina e de aprendizado profundo aplicados às séries de despesas primárias federais. Realizamos a avaliação em três etapas: (i) desempenho na validação cruzada dentro do conjunto de treino; (ii) desempenho no conjunto de teste, fora da amostra; e (iii) análise do desempenho para investigar a ocorrência de *overfitting*, *underfitting* ou padrões de generalização. Também discutimos a comparação entre os conjuntos e realizamos um teste de robustez, alterando o horizonte de previsão.

4.1 Modelos Estatísticos

Desempenho no Conjunto de Treino (Validação Cruzada)

A Tabela 1 resume as médias, dentre todas as despesas, do erro absoluto médio (MAE) e do erro quadrático médio (RMSE) obtidos na validação cruzada (*expanding window*) para cada modelo avaliado. Entre os modelos de referência, o *Seasonal Naive* apresenta o menor MAE médio (2.259,48), seguido do *Historic Average* (2.406,35). No grupo dos modelos estatísticos, o destaque é o *MFLES*, com o menor MAE médio (2.576,11), seguido pelo *ARIMA* (2.852,98).

Tabela 1 – Desempenho dos modelos *benchmark* e estatísticos no conjunto de treino

Modelo	**MAE**	**RMSE**
Seasonal Naive	2.259,48	4.418,89
Historic Average	2.406,35	4.032,77
Naive	6.059,89	7.991,80
Random Walk With Drift	6.918,17	9.118,06
Window Average	7.498,67	9.026,54
MFLES	2.576,11	4.210,70
ARIMA	2.852,98	4.521,43
TBATS	3.561,68	5.521,35
ETS	3.801,51	5.645,45
Theta	4.199,90	6.140,13
CES	13.696,61	24.196,63

Nota: Resultados ordenados de forma crescente pelo MAE.

A Tabela 2 mostra a frequência em que cada modelo foi selecionado como o melhor, ou seja, com o menor erro, em cada uma das séries de despesa.

Tabela 2 – Frequência de seleção como melhor modelo *benchmark* e estatístico no conjunto de treino

Modelo	**MAE**	**RMSE**
Seasonal Naive	12	11
Historic Average	8	10
Window Average	2	1
Random Walk With Drift	1	1
Naive	1	1
MFLES	6	6
TBATS	6	2
ARIMA	5	5
ETS	3	4
CES	2	4
Theta	2	3

Nota: Resultados ordenados de forma decrescente pelo MAE.

Observamos que, entre os modelos de referência, o *Seasonal Naive* e o *Historic Average* concentram a maioria das seleções como melhores modelos para as séries avaliadas, especialmente no critério MAE. Já no grupo dos modelos estatísticos, verificamos maior diversidade, com destaque para MFLES, TBATS e ARIMA, mas nenhum modelo domina todas as séries, indicando que o ajuste personalizado é fundamental para maximizar o desempenho preditivo nesse contexto.

Desempenho no Conjunto de Teste (Fora da Amostra)

A Tabela 3 sintetiza as métricas de erro fora da amostra (conjunto de teste) para as previsões geradas pelos melhores modelos selecionados para cada linha de despesa. Destacamos reduções expressivas nos erros médios em comparação ao *benchmark*.

Tabela 3 – Desempenho dos modelos estatísticos comparados ao *benchmark* no conjunto de teste

Métrica	**Benchmark**	**Estatístico**	**Redução do erro (%)**
MAE	2.257,30	1.731,89	23,28%
MSE	22.907.035,21	15.686.055,51	31,52%
RMSE	3.255,79	2.646,59	18,71%
MAPE	1.057,16	370,73	64,93%
SMAPE	32,97	22,93	30,45%
MASE	2,04	1,57	23,04%
MSSE	3,99	2,61	34,59%
RMSSE	1,55	1,30	16,13%

Nota: A coluna “Redução do erro” mostra a melhoria percentual dos modelos estatísticos em relação ao benchmark para cada métrica.

Análise de Desempenho

A Tabela 4 resume a comparação entre os erros obtidos na validação cruzada (treino) e no teste. Observamos que os modelos estatísticos não apenas superam os *benchmarks* em todos os critérios, mas também apresentam erro médio menor no teste do que no treino. Esse resultado, embora à primeira vista possa parecer contraintuitivo, pode ser explicado por três fatores: (i) amostras de teste menos voláteis, com períodos mais regulares; (ii) amostragem robusta, sem vazamento de dados e com boa segmentação das séries; e (iii) tamanho menor da amostra de teste, que pode, por acaso, ser menos complexa do que o conjunto de treino.

Tabela 4 – Comparação dos erros médios dos modelos estatísticos e *benchmarks* nos conjuntos de treino e teste

Métrica	**Benchmark**		**Estatístico**	
	Treino	**Teste**	**Treino**	**Teste**
MAE	1.880,44	2.257,30	1.952,87	1.731,89
RMSE	3.621,43	3.255,79	3.609,39	2.646,59

De modo geral, a ausência de aumento dos erros no conjunto de teste indica que os modelos ajustados conseguem capturar padrões estáveis e generalizáveis nas séries avaliadas, sem evidências de *overfitting*. O ganho absoluto e relativo dos modelos estatísticos em relação ao *benchmark* demonstra a importância de adotar abordagens mais sofisticadas, mesmo em contextos de séries curtas e com rigidez orçamentária. Em resumo, os resultados evidenciam que: (i) os modelos estatísticos, especialmente MFLES, ARIMA e TBATS, superam os modelos de referência em todos os critérios de avaliação; (ii) não há indícios de sobreajuste, já que os erros de teste permanecem iguais ou inferiores aos do treino; e (iii) a metodologia adotada assegura robustez e confiabilidade às previsões de despesas federais, reforçando o papel de modelos estatísticos bem calibrados como instrumentos relevantes para a política fiscal.

Adicionalmente, não observamos sinais de subajuste (*underfitting*). Em situações de *underfitting*, seria esperado que tanto os erros no treino quanto no teste permanecessem elevados, sugerindo que o modelo seria incapaz de capturar padrões estruturais relevantes das séries. No entanto, como os modelos estatísticos apresentam desempenho substancialmente superior aos *benchmarks* em ambas as amostras, constatamos que a modelagem adotada extrai informações relevantes dos dados, sem limitar-se a reproduzir apenas tendências triviais. Em seguida, avaliamos o desempenho dos modelos de *machine learning* e suas nuances frente ao contexto fiscal brasileiro.

4.2 Modelos de Machine Learning

Desempenho no Conjunto de Treino (Validação Cruzada)

A Tabela 5 apresenta as médias do erro absoluto médio (MAE) e do erro quadrático médio (RMSE) para os principais algoritmos de ML na validação cruzada. Observamos que o LightGBM apresenta o menor MAE médio, enquanto o CatBoost obtém o menor RMSE médio.

Tabela 5 – Desempenho médio dos modelos de *machine learning* no conjunto de treino

Modelo	**MAE**	**RMSE**
LightGBM	1.928,42	3.769,79
Random Forest	1.986,33	3.719,98
Ridge	2.090,85	3.681,96
XGBoost	2.128,99	3.623,39
CatBoost	2.192,85	3.585,35
Lasso	2.193,80	3.819,97
Linear Regression	2.205,38	3.713,62
Elastic Net	2.304,26	3.772,46

Nota: Resultados ordenados de forma crescente pelo MAE.

A Tabela 6 apresenta a frequência em que cada modelo é selecionado como o melhor para cada série (menor MAE e RMSE). Essa diversidade de seleções evidencia que nenhuma abordagem é universalmente superior para todas as séries, ressaltando a importância da escolha específica para cada contexto orçamentário e a relevância da estratégia *bottom-up* na previsão de despesas públicas.

Tabela 6 – Frequência de seleção como melhor modelo de *machine learning* no conjunto de treino

Modelo	**MAE**	**RMSE**
LightGBM	11	6
CatBoost	4	5
Ridge	4	2
Random Forest	3	2
Linear Regression	1	3
Lasso	1	2
XGBoost	–	4
Elastic Net	–	–

Nota: Resultados ordenados de forma decrescente pelo MAE.

Desempenho no Conjunto de Teste (Fora da Amostra)

A Tabela 7 resume as métricas de erro dos melhores modelos de ML no conjunto de teste, permitindo comparação com o *benchmark*.

Tabela 7 – Desempenho dos modelos de *machine learning* comparados ao *benchmark* no conjunto de teste

Métrica	**Benchmark**	**Machine Learning**	**Redução do erro (%)**
MAE	2.257,30	2.196,71	2,68%
MSE	22.907.035,21	25.898.316,48	-13,07%
RMSE	3.255,79	3.322,90	-2,06%
MAPE	1.057,16	658,48	37,69%
SMAPE	32,97	24,24	26,48%
MASE	2,04	1,94	4,90%
MSSE	3,99	4,32	-8,27%
RMSSE	1,55	1,61	-3,87%

Nota: A coluna "Redução do erro" mostra a melhoria percentual dos modelos de ML em relação ao benchmark para cada métrica. Valores negativos indicam piora em relação ao benchmark.

A análise dos resultados fora da amostra mostra que, embora os modelos de *machine learning* apresentem desempenho levemente superior no critério MAE e avanços consideráveis nas métricas percentuais (MAPE e SMAPE), não identificamos ganhos sistemáticos em todas as métricas avaliadas. Em particular, tanto o RMSE quanto o MSE e suas variantes padronizadas (MSSE, RMSSE) ficam ligeiramente acima dos valores observados para o *benchmark*, indicando que os modelos de ML, apesar de eficientes para prever a mediana dos desvios absolutos, são mais sensíveis a grandes erros em algumas séries ou eventos atípicos. Por outro lado, esse desempenho menos consistente evidencia a importância de ajustar expectativas quanto ao uso dessas técnicas em ambientes caracterizados por volatilidade fiscal, mudanças institucionais frequentes e séries curtas.

Assim, nossos resultados sugerem que o melhor desempenho dos modelos estatísticos em horizontes mais longos decorre de sua estrutura parcimoniosa e do viés indutivo embutido nas técnicas clássicas de estimação, que atuam como regularizadores naturais contra o sobreajuste. Esses modelos preservam memória histórica de maneira eficiente e projetam tendências e sazonalidades de forma estável, enquanto arquiteturas de aprendizado de máquina e de aprendizado profundo, embora potentes para capturar padrões locais em horizontes curtos, tendem a sofrer com a propagação de erros e a alta variância quando estendidas a horizontes longos. Tal evidência é consistente com a literatura em competições de previsão discutida em Makridakis, Spiliotis e Assimakopoulos (2020) e Godahewa et al. (2021), reforçando que a escolha metodológica deve considerar não apenas o tipo de série, mas também o horizonte de interesse.

Análise de Desempenho

A Tabela 8 sintetiza os resultados comparativos dos melhores modelos de ML entre treino e teste, considerando as principais métricas.

Tabela 8 – Comparação dos erros médios dos modelos de *machine learning* e *benchmark* nos conjuntos de treino e teste

Métrica	**Benchmark**		**Machine Learning**	
	Treino	**Teste**	**Treino**	**Teste**
MAE	1.880,44	2.257,30	1.761,53	2.196,71
RMSE	3.621,43	3.255,79	3.338,61	3.322,90

Observamos que, para o MAE, os modelos de ML apresentam desempenho superior ao *benchmark* no treino, mas sofrem pequena elevação no teste, o que é natural quando avaliamos a capacidade preditiva fora da amostra. Ainda assim, o MAE no teste dos modelos de ML permanece competitivo e menor que o *benchmark*, demonstrando boa generalização. No caso do RMSE, a diferença entre treino e teste permanece mínima, sugerindo estabilidade na dispersão dos erros, mesmo em cenários de grandes desvios.

Não identificamos indícios de *overfitting*, pois a diferença dos erros entre treino e teste é pequena e não há explosão do erro fora da amostra. Da mesma forma, não observamos sinal de *underfitting*, visto que os modelos conseguem capturar padrões relevantes das séries e superam o *benchmark* em ambos os conjuntos.

Entre os algoritmos de aprendizado de máquina avaliados, o modelo LightGBM se destaca por sua recorrente seleção como melhor modelo em múltiplas séries. Ainda assim, o modelo composto pelos melhores algoritmos em cada série apresenta desempenho superior, alinhando-se à literatura sobre o potencial da técnica de previsão *bottom-up* para lidar com a heterogeneidade estrutural típica das séries de despesas públicas.

4.3 Modelos de Deep Learning

Desempenho no Conjunto de Treino (Validação Cruzada)

A Tabela 9 apresenta as médias do erro absoluto médio (MAE) e do erro quadrático médio (RMSE) para os principais modelos de *deep learning* avaliados na validação cruzada do conjunto de treino. Observamos que TS-Mixer, StemGNN, TCN e DilatedRNN apresentam os menores valores de MAE, enquanto NBEATS, DeepNPTS e TS-Mixer se destacam nos menores valores de RMSE. A ampla dispersão entre arquiteturas indica forte dependência da estrutura da série para a eficácia de cada abordagem.

Tabela 9 – Desempenho médio dos modelos de *deep learning* no conjunto de treino

Modelo	**MAE**	**RMSE**
TS-Mixer	2.111,43	3.548,56
StemGNN	2.200,92	3.753,66
TCN	2.220,60	3.734,76
DilatedRNN	2.220,75	3.746,72
LSTM	2.230,13	3.752,41
GRU	2.236,10	3.739,45
NBEATS	2.238,47	3.481,22
RNN	2.253,07	3.747,12
SOFTS	2.267,67	3.580,12
Vanilla Transformer	2.290,53	3.844,24
TiDE	2.327,32	3.772,22
KAN	2.351,86	3.731,07
FEDformer	2.353,89	3.753,21
DLinear	2.360,65	3.739,22
NHITS	2.402,37	3.827,10
Informer	2.409,95	3.881,29
Autoformer	2.414,23	3.788,67
MLP	2.454,65	3.895,57
TFT	2.487,47	3.819,84
Bi-TCN	2.511,75	3.966,84
MLP Multivariate	2.516,32	3.994,98
DeepNPTS	2.524,49	3.535,31
DeepAR	2.529,63	4.003,66
NLinear	2.534,48	3.636,27
CNN	2.712,42	3.737,28
PatchTST	2.852,34	3.681,96

Nota: Resultados ordenados de forma crescente pelo MAE.

A Tabela 10 apresenta a frequência em que cada modelo é selecionado como o melhor para cada série (menor MAE e RMSE). Observamos maior dispersão entre os modelos de *deep learning*, sugerindo que o melhor desempenho é bastante sensível ao tipo de arquitetura, série e ajuste de hiperparâmetros.

Tabela 10 – Frequência de seleção como melhor modelo de *deep learning* no conjunto de treino

Modelo	**MAE**	**RMSE**
DLinear	3	5
SOFTS	3	2
TFT	3	1
Autoformer	2	2
TS-Mixer	2	1
DilatedRNN	2	1
PatchTST	1	4
NBEATS	1	1
CNN	1	1
Bi-TCN	1	1
MLP Multivariate	1	1
TiDE	1	1
KAN	1	–
StemGNN	1	–
Vanilla Transformer	1	–
NLinear	–	2
DeepNPTS	–	1

Nota: Resultados ordenados de forma decrescente pelo MAE.

De forma geral, os resultados evidenciam que, embora modelos como TS-Mixer, StemGNN, TCN e DLinear liderem em termos de erro absoluto e quadrático médio, não há domínio absoluto de uma arquitetura sobre todas as séries. Isso sugere que a seleção individualizada por série, acompanhada de ajuste automático de hiperparâmetros, é fundamental para extrair o melhor desempenho dos modelos de *deep learning* no contexto das despesas federais.

Desempenho no Conjunto de Teste (Fora da Amostra)

A Tabela 11 resume as métricas de erro dos melhores modelos de DL no conjunto de teste, permitindo comparação com o *benchmark*. Os resultados do conjunto de teste indicam que os modelos de *deep learning* não apresentam ganhos robustos sobre o *benchmark*, especialmente nas métricas mais sensíveis a grandes desvios. Embora o MAE, MAPE e SMAPE mostrem reduções de erro em relação ao *benchmark*, sugerindo uma leve vantagem dos modelos de DL na previsão dos desvios médios e percentuais, as métricas baseadas em quadrados dos resíduos (MSE, RMSE, MSSE, RMSSE) apresentam desempenho inferior ao dos modelos de referência.

Tabela 11 – Desempenho dos modelos de *deep learning* comparados ao *benchmark* no conjunto de teste

Métrica	**Benchmark**	**Deep Learning**	**Redução do erro (%)**
MAE	2.257,30	2.130,28	5,63%
MSE	22.907.035,21	28.205.277,21	-23,11%
RMSE	3.255,79	3.340,65	-2,61%
MAPE	1.057,16	670,49	36,54%
SMAPE	32,97	24,12	26,81%
MASE	2,04	2,03	0,49%
MSSE	3,99	4,82	-20,80%
RMSSE	1,55	1,71	-10,32%

Nota: A coluna "Redução do erro" mostra a melhoria percentual dos modelos de deep learning em relação ao benchmark para cada métrica. Valores negativos indicam piora em relação ao benchmark.

Esses resultados indicam que, apesar de os modelos de *deep learning* conseguirem capturar padrões médios ou recorrentes das séries de despesas, eles são mais vulneráveis à ocorrência de grandes erros em pontos específicos, especialmente em contextos marcados por choques, sazonalidades atípicas ou mudanças abruptas de regime fiscal. A leve redução do MAE e MASE, acompanhada do aumento do RMSE e métricas relacionadas, reforça a hipótese de que esses modelos podem ser sensíveis a *outliers* ou eventos incomuns, especialmente em amostras de teste relativamente curtas e heterogêneas.

Análise de Desempenho

A Tabela 12 sintetiza os resultados comparativos dos melhores modelos de DL entre treino e teste, considerando as principais métricas.

Tabela 12 – Comparação dos erros médios dos modelos de *deep learning* e *benchmark* nos conjuntos de treino e teste

Métrica	**Benchmark**		**Deep Learning**	
	Treino	**Teste**	**Treino**	**Teste**
MAE	1.880,44	2.257,30	1.889,77	2.130,28
RMSE	3.621,43	3.255,79	3.244,67	3.340,65

No conjunto de treino, os modelos de *deep learning* apresentam desempenho semelhante ao *benchmark* em termos de MAE, com valores praticamente iguais, e desempenho superior em termos de RMSE. No conjunto de teste, observamos que o MAE do modelo de *deep learning* permanece inferior ao *benchmark*, indicando maior precisão preditiva fora da amostra. O RMSE dos modelos de *deep learning* fica levemente acima do *benchmark*,

sugerindo que, apesar de menor MAE, houve alguma ocorrência de desvios maiores na previsão de determinadas séries.

A diferença entre treino e teste para os modelos de *deep learning* é modesta. O aumento do MAE é esperado em predições fora da amostra e está em linha com o observado para modelos estatísticos e de *machine learning*. Isso sugere que não há sinais claros de sobreajuste (*overfitting*), já que o erro não cresce de forma abrupta, e o modelo mantém desempenho competitivo em dados não vistos. Tampouco observamos evidências de subajuste (*underfitting*), uma vez que os erros não permanecem elevados em ambos os conjuntos. Em síntese, observamos que os modelos de *deep learning* demonstram capacidade de generalização, conseguindo superar o *benchmark* em termos de precisão absoluta no teste, ainda que apresentem ligeira desvantagem no RMSE, um reflexo natural da maior sensibilidade desse indicador a valores extremos.

Por outro lado, o desempenho menos robusto dos modelos de *machine learning* e *deep learning* no teste reforça que, em séries curtas e altamente heterogêneas, métodos clássicos bem calibrados podem manter vantagem relevante sobre alternativas modernas, cuja eficácia plena depende de amostras maiores ou menos sujeitas a choques.

4.4 Combinação de Previsões

A Tabela 13 apresenta o desempenho da combinação de previsões no conjunto de teste, comparando-a ao *benchmark* ingênuo.

Tabela 13 – Desempenho da combinação de previsões comparado ao *benchmark* no conjunto de teste

Métrica	**Benchmark**	**Ensemble**	**Redução do erro (%)**
MAE	2.257,30	1.932,98	14,37%
MSE	22.907.035,21	21.635.445,46	5,55%
RMSE	3.255,79	3.002,49	7,78%
MAPE	1.057,16	563,50	46,70%
SMAPE	32,97	22,65	31,30%
MASE	2,04	1,78	12,75%
MSSE	3,99	3,67	8,02%
RMSSE	1,55	1,50	3,23%

Nota: A coluna "Redução do erro" mostra a melhoria percentual da combinação de previsões em relação ao benchmark para cada métrica.

Observamos que a estratégia de combinação de previsões supera o *benchmark* em todas as métricas avaliadas, com ganhos particularmente expressivos no MAPE (46,7%) e no SMAPE (31,3%), além de reduções relevantes no MAE (14,4%) e no RMSE (7,8%).

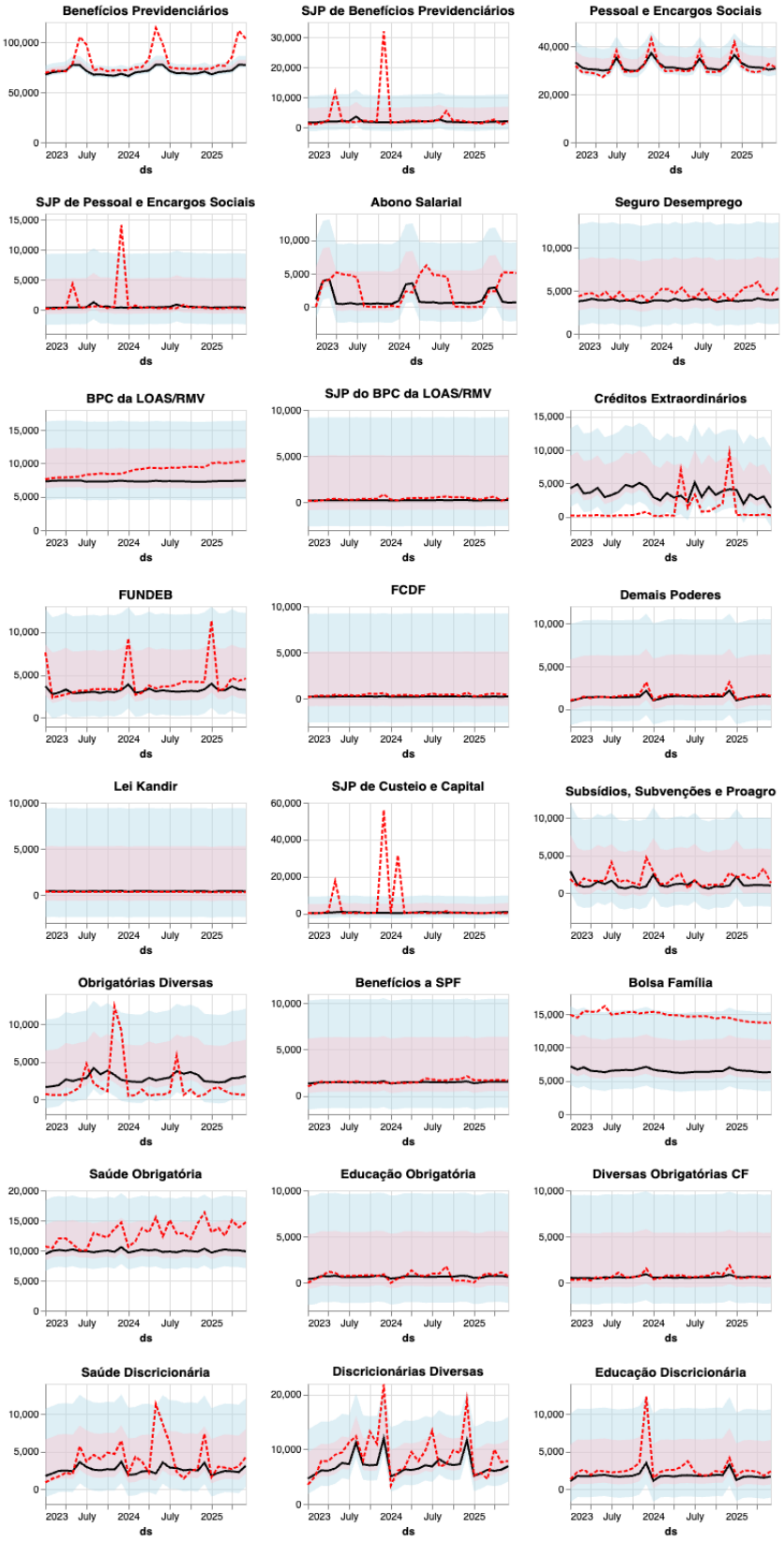
Esses ganhos confirmam o potencial da combinação para promover robustez e estabilidade às previsões, diluindo erros específicos de modelos individuais.

Ao compararmos o desempenho da combinação com os melhores modelos estatísticos, verificamos que o *ensemble* supera os estatísticos apenas no SMAPE, e por pequena margem. Em todas as demais métricas, os modelos estatísticos mantêm desempenho superior, o que reflete a forte aderência ao padrão das séries avaliadas. Portanto, embora a combinação de previsões seja recomendada como estratégia de robustez, especialmente em cenários de elevada heterogeneidade, sua eficácia pode ser limitada quando uma família de modelos já se mostra claramente dominante. Ainda assim, a combinação agrega valor ao evitar dependência de um único modelo e ao aumentar a resiliência do sistema preditivo frente a cenários incertos.

Nossos resultados também sugerem que esse padrão observado, com modelos estatísticos se destacando em horizontes longos, redes neurais em horizontes curtos e a combinação apresentando desempenho mais estável, reflete não apenas a heterogeneidade das séries, mas também o papel da"memória" e do viés indutivo embutidos nas técnicas clássicas de estimação. Modelos estatísticos tendem a errar de forma mais parcimoniosa, o que preserva sua vantagem em horizontes longos. Já a combinação funciona como mecanismo de diversificação, equilibrando viés e variância: raramente lidera em desempenho absoluto, mas quase nunca é a pior opção, assegurando previsões consistentes mesmo diante de choques.

A Figura 3 apresenta a combinação de previsões para as 24 despesas primárias em termos reais (milhões de R$) para o horizonte de 30 meses, incluindo os intervalos de confiança de 80% e 95%, além dos valores realizados.

Figura 3 – Combinação de previsões (horizonte = 30 meses)



Nota: Linha vermelha: realizado; linha preta: previsto; área azul: IC 95%; e área vermelha: IC 80%.

4.5 Testes de Robustez

Para avaliarmos a robustez dos resultados em diferentes configurações, realizamos um teste alternativo reduzindo o horizonte de previsão. Estendemos o conjunto de treino até dezembro de 2023 e realizamos previsões para o período de janeiro de 2024 a junho de 2025, totalizando 18 meses (em vez dos 30 meses do cenário principal).

A Tabela 14 apresenta as métricas de erro para cada classe de modelo avaliada no novo horizonte de previsão. Os valores referem-se ao desempenho médio das melhores previsões para cada linha de despesa.

Tabela 14 – Desempenho dos modelos no teste de robustez (horizonte de 18 meses)

Métrica	**Baseline**	**Estatístico**	**ML**	**DL**	**Ensemble**
MAE	2.052,63	1.728,99	1.786,51	1.389,87	1.499,48
MSE	27.297.285,45	16.384.217,14	17.031.842,42	12.984.599,96	12.460.859,76
RMSE	3.171,17	2.390,75	2.457,77	2.119,89	2.140,12
MAPE	349,48	222,41	263,72	180,66	214,43
SMAPE	23,78	21,82	23,01	18,68	21,59
MASE	1,72	1,36	1,52	1,18	1,25
MSSE	2,72	1,38	2,02	1,54	1,34
RMSSE	1,28	0,95	1,11	0,98	0,93

Nota: Resultados para o conjunto de teste de janeiro de 2024 a junho de 2025.

A Tabela 15 apresenta a redução percentual dos erros de cada modelo em relação ao baseline.

Tabela 15 – Redução percentual dos erros em relação ao baseline (teste de robustez)

Métrica	**Estatístico**	**ML**	**DL**	**Ensemble**
MAE	15,76%	12,96%	32,32%	26,93%
MSE	39,96%	37,57%	52,43%	54,35%
RMSE	24,59%	22,53%	33,16%	32,51%
MAPE	36,36%	24,53%	48,29%	38,65%
SMAPE	8,24%	3,24%	21,44%	9,19%
MASE	20,93%	11,63%	31,40%	27,33%
MSSE	49,26%	25,74%	43,38%	50,74%
RMSSE	25,78%	13,28%	23,44%	27,34%

Nota: Cálculo: 1 − (erro modelo/erro baseline).

Os resultados do teste de robustez reforçam a superioridade dos modelos avançados (estatísticos, ML, DL e ensemble) em relação ao *baseline* simples, mesmo quando reduzimos o horizonte de previsão de 30 para 18 meses. Observamos que todos os modelos mantêm desempenho superior, com destaque para os de *deep learning* e *ensemble*, que apresentam as maiores reduções relativas de erro em praticamente todas as métricas

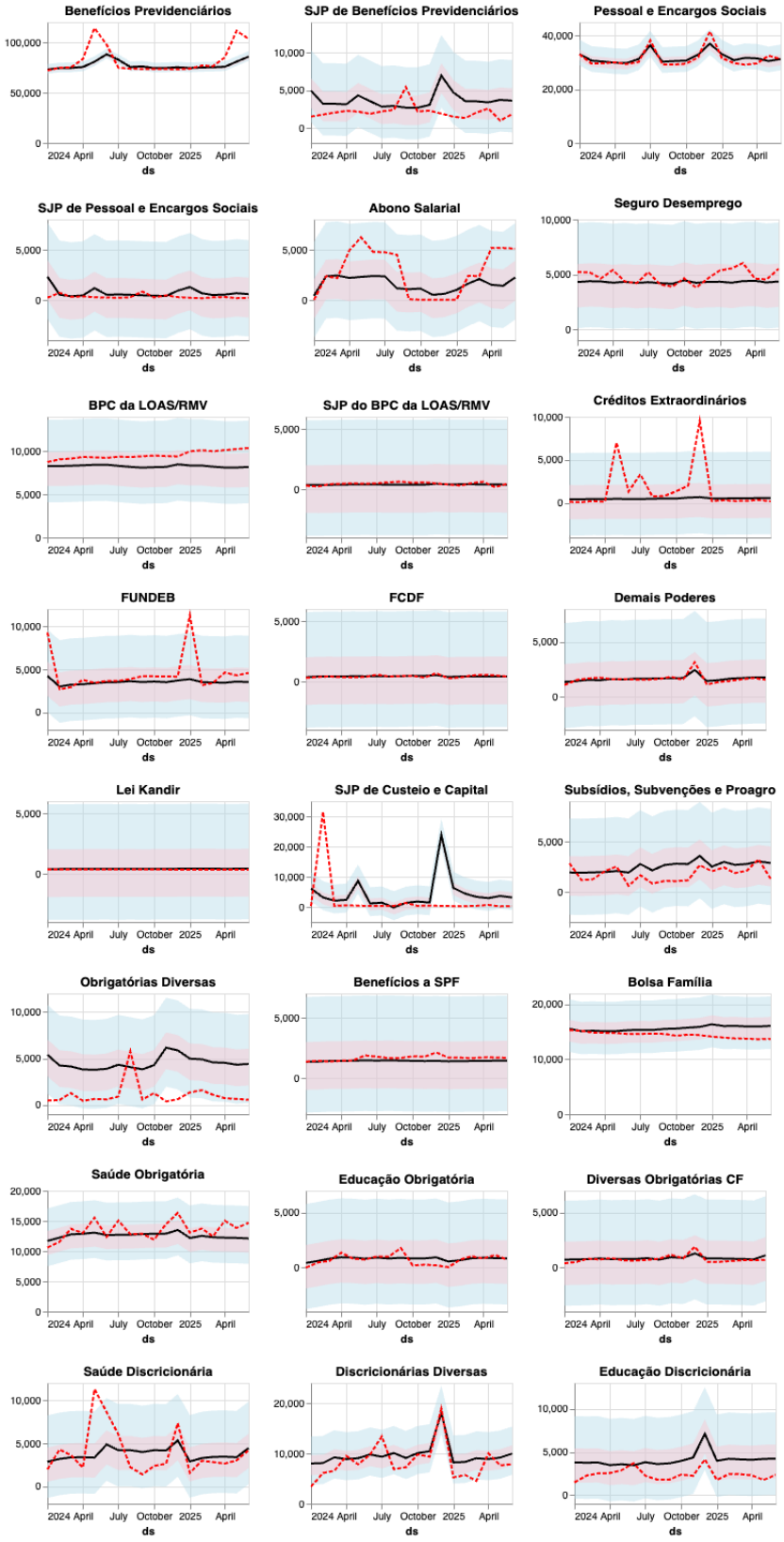
analisadas. Os modelos estatísticos e de ML preservam consistência e robustez, enquanto a combinação de previsões mostra-se especialmente eficaz na redução de MSE, MSSE e RMSSE. Notamos ainda que os ganhos do DL em relação ao *baseline* são notáveis, sugerindo melhor adaptação a choques recentes ou mudanças de padrão.

Também observamos que, em horizontes mais curtos, os ganhos dos modelos sofisticados sobre o *benchmark* tendem a ser mais expressivos em métricas absolutas, embora as diferenças entre as classes (estatísticos, ML, DL, *ensemble*) fiquem menos acentuadas. Isso indica que, em cenários de maior previsibilidade, as vantagens relativas dos métodos avançados persistem, mas o *benchmark* também se beneficia de menor incerteza.

Em síntese, os resultados confirmam que o uso criterioso de modelos estatísticos, de *machine learning*, *deep learning* e de suas combinações representa uma estratégia promissora para aprimorarmos a previsão fiscal brasileira, desde que respeitemos as especificidades do contexto, as limitações dos dados e as exigências de validação robusta. Os métodos clássicos continuam altamente competitivos, mas a combinação de paradigmas garante maior resiliência frente à incerteza e à heterogeneidade inerentes à gestão orçamentária pública.

A Figura 4 apresenta a combinação de previsões para as 24 despesas primárias em termos reais (milhões de R$) para o horizonte de 18 meses, incluindo os intervalos de confiança de 80% e 95%, além dos valores realizados.

Figura 4 – Combinação de previsões (horizonte = 18 meses)



Nota: Linha vermelha: realizado; linha preta: previsto; área azul: IC 95%; e área vermelha: IC 80%.

5 Discussão

Os resultados obtidos neste estudo reafirmam a complexidade da previsão de despesas públicas em contextos marcados por elevada incerteza, mudanças de regime e choques exógenos relevantes, como a pandemia de Covid-19 e a transição de governo federal ocorrida em 2023. Esses eventos impactam tanto o comportamento das séries históricas quanto a eficácia dos diferentes paradigmas de modelagem testados.

Constatamos que os modelos estatísticos clássicos superam, de forma consistente, as alternativas de *machine learning* e *deep learning* no horizonte de 30 meses. Isso evidencia que métodos bem calibrados e adaptados à estrutura dos dados mantêm desempenho robusto mesmo em cenários de elevada volatilidade e tamanho amostral reduzido. Esse resultado pode estar associado à maior capacidade dos modelos estatísticos de capturar padrões sazonais e tendências estruturais persistentes, além de seu menor risco de sobreajuste em séries curtas e ruidosas.

Por outro lado, quando reduzimos o horizonte de previsão para 18 meses, concentrando a análise em um período mais recente e posterior ao choque da pandemia, observamos reversão parcial dessa tendência. Nessa configuração, os modelos de *deep learning* apresentam desempenho superior em várias métricas relevantes, enquanto os estatísticos mantêm robustez, mas sem a mesma vantagem observada em horizontes mais longos. Esse resultado confirma o potencial do *deep learning* para capturar padrões complexos e dinâmicos em séries mais curtas, desde que o contexto de previsão seja menos afetado por choques atípicos e as arquiteturas sejam ajustadas de forma adequada aos dados disponíveis.

A análise comparativa mostra que modelos de *machine learning*, embora competitivos no conjunto de treino, tendem a perder desempenho fora da amostra, especialmente diante de eventos extremos ou mudanças bruscas no regime fiscal. Essa limitação ressalta o desafio de generalização dessas abordagens em ambientes heterogêneos e voláteis, reforçando a necessidade de validação rigorosa e de seleção criteriosa de hiperparâmetros.

A variação dos resultados conforme o horizonte e a janela histórica confirma que o desempenho relativo de cada classe de modelos depende fortemente do contexto institucional, da presença de choques estruturais e do volume de dados disponíveis para ajuste. A pandemia de Covid-19, ao gerar descontinuidades e saltos nas séries de despesa, e a mudança de governo, ao alterar prioridades e dinâmicas das políticas públicas, aumentam a incerteza e dificultam a modelagem de tendências, exigindo maior adaptabilidade dos métodos preditivos.

Diante desse cenário, verificamos que a estratégia de combinação de previsões se

mostra particularmente relevante. Como sugerem Hyndman et al. (2025), a combinação permite balancear as limitações e vantagens específicas de cada abordagem, promovendo previsões mais estáveis, resilientes a choques e menos suscetíveis a vieses de modelagem. Embora, neste estudo, o *ensemble* não supere de forma consistente o melhor modelo individual em nenhum horizonte de previsão (com exceção do SMAPE no horizonte de 30 meses e do MSE, MSSE e RMSSE no horizonte de 18 meses), constatamos que seu desempenho equilibrado em diferentes configurações e janelas temporais indica que a diversificação metodológica constitui resposta prudente à incerteza fiscal.

Concluímos, portanto, que não há solução única ou universal para previsão de despesas públicas em ambientes sujeitos a mudanças frequentes e choques relevantes. A escolha do paradigma ótimo deve considerar não apenas o desempenho absoluto em métricas tradicionais, mas também a robustez em cenários de instabilidade, a capacidade de adaptação a mudanças de padrão e a viabilidade de implementação prática em ambientes institucionais. A integração de métodos, associada a processos de validação contínua, configura-se como a principal recomendação para gestores e pesquisadores que buscam aprimorar a qualidade das previsões fiscais no Brasil.

Adicionalmente, reconhecemos algumas limitações que merecem registro. A principal refere-se à disponibilidade e à granularidade das séries de despesa pública: embora tenhamos trabalhado com dados mensais desagregados por categoria, a ausência de informações mais detalhadas sobre programas e subfunções, bem como a indisponibilidade de séries históricas mais longas, pode ter restringido o potencial de ajuste de alguns modelos, especialmente os mais intensivos em dados. Também não incorporamos variáveis explicativas, dados de alta frequência ou indicadores antecedentes que poderiam enriquecer a modelagem. Essas limitações abrem espaço para pesquisas futuras. Investigações subsequentes podem explorar a incorporação de variáveis exógenas macroeconômicas e setoriais, a utilização de bases de dados administrativas de maior frequência e granularidade, e a integração de modelos estruturais e semi-estruturais com técnicas de *machine learning* e *deep learning* em configurações híbridas.

Do ponto de vista de políticas públicas, nossos achados sugerem que a diversificação metodológica, com uso combinado de abordagens estatísticas, de *machine learning* e de *deep learning*, aumenta a resiliência das previsões fiscais diante de choques e mudanças de regime. A adoção de processos de validação contínua e de seleção dinâmica de modelos contribui para reduzir vieses, melhorar a transparência e fortalecer a credibilidade das contas públicas.

6 Conclusão

Os resultados deste estudo evidenciam que o desempenho relativo dos diferentes paradigmas de modelagem preditiva é sensível ao horizonte de previsão e à ocorrência de choques e mudanças institucionais relevantes, como a pandemia de Covid-19 e as alterações na condução da política fiscal. Observamos que, em horizontes de previsão mais curtos e em períodos mais recentes, os modelos de *deep learning* ganham destaque, enquanto os modelos estatísticos tradicionais permanecem superiores em contextos mais amplos e com maior variabilidade histórica. Essa variabilidade de desempenho reforça a importância da adoção de técnicas de combinação de previsões, que promovem maior robustez e equilíbrio, reduzem o risco de dependência de um único modelo e aumentam a resiliência das projeções fiscais frente a diferentes cenários.

Concluímos que a escolha da abordagem preditiva mais adequada deve considerar as características do problema, o contexto temporal e a possibilidade de eventos disruptivos. Independentemente do cenário, constatamos que os modelos avançados de previsão são ferramentas valiosas para lidar com a incerteza inerente à previsão de despesas públicas, oferecendo previsões confiáveis para subsidiar o planejamento da política fiscal. Sugerimos que pesquisas futuras explorem variáveis exógenas e modelos estruturais ou semi-estruturais, de modo a ampliar a acurácia, a robustez e a interpretabilidade dos modelos preditivos.

Referências

AKIBA, T. et al. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2019. Citado na página 40.

ANDO, S.; KIM, T. Systematizing macroframework forecasting: High-dimensional conditional forecasting with accounting identities. *IMF Working Paper*, 2022. Disponível em: <https://doi.org/10.5089/9798400211683.001>. Citado na página 18.

APICELLA, A.; ISGRò, F.; PREVETE, R. *Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning*. 2025. Disponível em: <https://arxiv.org/abs/2401.13796>. Citado na página 29.

ARNOLD, R. W. How cbo produces its 10-year economic forecast. *CBO Working Paper*, 2018. Disponível em: <https://www.cbo.gov/publication/53537>. Citado 2 vezes nas páginas 13 e 19.

ASIMAKOPOULOS, I.; PAREDES, J.; WARMEDINGER, T. *Forecasting Fiscal Time Series Using Mixed Frequency Data*. [S.l.], 2013. (ECB Working Paper No. 1553). Disponível em: <https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1550.pdf>. Citado na página 21.

BARBER, R. F. et al. Conformal prediction beyond exchangeability. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 51, n. 2, p. 816 – 845, 2023. Disponível em: <https://doi.org/10.1214/23-AOS2276>. Citado na página 35.

BATES, J. M.; GRANGER, C. W. J. The combination of forecasts. *Operational Research Quarterly*, v. 20, n. 4, p. 451–468, 1969. Disponível em: <https://doi.org/10.2307/3008764>. Citado 2 vezes nas páginas 22 e 24.

BERGMEIR, C.; HYNDMAN, R. J.; KOO, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, v. 120, p. 70–83, 2018. Disponível em: <https://doi.org/10.1016/j.csda.2017.11.003>. Citado na página 25.

BOLHUIS, M. A.; RAYNER, B. Deus ex machina: A framework for macro forecasting with machine learning. *IMF Working Paper*, 2020. Disponível em: <https://doi.org/10.5089/9781513531724.001>. Citado na página 18.

CAMERON, S. How can independent fiscal institutions make the most of assessing past economic forecasts? *OECD Journal on Budgeting*, v. 22/2, p. 104–113, 2022. Disponível em: <https://doi.org/10.1787/5aa7765a-en>. Citado na página 17.

CARNEIRO, L. M.; COSTA, M. C. Fatores associados ao erro de previsão de despesa orçamentária nos municípios brasileiros. *Cadernos de Finanças Públicas*, v. 21, n. 2, p. 1–41, 2021. Disponível em: <https://doi.org/10.55532/1806-8944.2021.121>. Citado 2 vezes nas páginas 13 e 22.

CHEN, S.; RANCIERE, R. Financial information and macroeconomic forecasts. *IMF Working Paper*, 2016. Disponível em: <https://doi.org/10.5089/9781475563177.001>. Citado na página 18.

CICCARELLI, M. et al. Ecb macroeconometric models for forecasting and policy analysis. *ECB Occasional Paper*, n. 344, 2024. Disponível em: <http://dx.doi.org/10.2139/ssrn.4761328>. Citado 3 vezes nas páginas 13, 20 e 21.

CIMADOMO, J.; GIANNONE, D.; LENZA, M. *Fiscal Nowcasting*. [S.l.], 2017. Disponível em: <https://www.ecb.europa.eu/press/conferences/shared/pdf/20170929_advances_in_short_term_forecasting/Paper_2_Cimadono_Giannone_Lenza.pdf>. Citado na página 21.

CLEMEN, R. T. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, v. 5, n. 4, p. 559–583, 1989. Disponível em: <https://doi.org/10.1016/0169-2070(89)90012-5>. Citado na página 24.

CROSTON, J. D. Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, v. 23, p. 289–303, 1972. Citado na página 29.

DEUS, J. D. B. V. d.; MENDONçA, H. F. d. Fiscal forecasting performance in an emerging economy: An empirical assessment of brazil. *Economic Systems*, v. 41, n. 3, p. 408–419, 2017. Disponível em: <http://dx.doi.org/10.1016/j.ecosys.2016.10.004>. Citado 3 vezes nas páginas 13, 22 e 23.

ECONOMIC COMMISSION FOR LATIN AMERICA AND THE CARIBBEAN. *Revenue and Expenditure Forecasting Methods for a PER Spending*. Santiago, Chile, 2015. Disponível em: <https://www.cepal.org/en/projects/strengthening-technical-capacity-public-finance-managers-select-caribbean-small-island>. Citado 4 vezes nas páginas 13, 16, 17 e 18.

EICHER, T. S. et al. Forecasting in times of crises. *IMF Working Paper*, 2018. Disponível em: <https://doi.org/10.5089/9781484345436.001>. Citado na página 18.

FAVERO, C. A.; MARCELLINO, M. Modelling and forecasting fiscal variables for the euro area. *IGIER Working Paper*, n. 298, 2005. Disponível em: <http://dx.doi.org/10.2139/ssrn.812985>. Citado na página 13.

FORONI, C. Discussion of "fiscal nowcasting". Conference presentation. 2017. Disponível em: <https://www.ecb.europa.eu/press/conferences/shared/pdf/20170929_advances_in_short_term_forecasting/Paper_2_Foroni.pdf>. Citado na página 21.

GADELHA, S. R. d. B.; LIMA, A. F. R.; POLLI, D. A. Uso da metodologia de combinação de previsões para projeções da arrecadação de receitas brutas primárias de tributos federais. *Revista Cadernos de Finanças Públicas*, v. 1, n. 1, p. 1–70, 2020. Disponível em: <https://doi.org/10.55532/1806-8944.2020.77>. Citado 2 vezes nas páginas 13 e 22.

GARZA, A. et al. *StatsForecast: Lightning fast forecasting with statistical and econometric models*. 2022. PyCon Salt Lake City, Utah, US 2022. Disponível em: <https://github.com/Nixtla/statsforecast>. Citado na página 40.

GODAHEWA, R. et al. Monash time series forecasting archive. In: *Neural Information Processing Systems Track on Datasets and Benchmarks*. [s.n.], 2021. Disponível em: <https://forecastingdata.org/>. Citado 3 vezes nas páginas 14, 24 e 45.

GOLDBLUM, M. et al. *The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning*. 2024. Disponível em: <https://arxiv.org/abs/2304.05366>. Citado na página 21.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA: MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>. Citado na página 39.

HADZI-VASKOV, M. et al. Authorities' fiscal forecasts in latin america: Are they optimistic? *IMF Working Paper*, 2021. Disponível em: <https://doi.org/10.5089/9781513573403.001>. Citado na página 17.

HAMILTON, J. D. *Time Series Analysis*. Princeton, NJ: Princeton University Press, 1994. Disponível em: <https://press.princeton.edu/books/hardcover/9780691042893/time-series-analysis>. Citado na página 39.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009. Disponível em: <https://hastie.su.domains/ElemStatLearn/>. Citado na página 39.

HYNDMAN, R. J. et al. *Forecasting: Principles and Practice, the Pythonic Way*. Melbourne, Australia: OTexts, 2025. Citado 7 vezes nas páginas 29, 32, 33, 34, 35, 38 e 57.

HYNDMAN, R. J.; KHANDAKAR, Y. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, v. 27, n. 3, p. 1–22, 2008. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v027i03>. Citado na página 40.

INDEPENDENT EVALUATION OFFICE OF THE INTERNATIONAL MONETARY FUND. *IMF Forecasts: Process, Quality, and Country Perspectives*. Washington, D.C., 2014. Disponível em: <https://doi.org/10.5089/9781475599510.017>. Citado na página 17.

JUNG, J.-K.; PATNAM, M.; TER-MARTIROSYAN, A. An algorithmic crystal ball: Forecasts based on machine learning. *IMF Working Paper*, 2018. Disponível em: <https://doi.org/10.5089/9781484380635.001>. Citado na página 18.

KAUSHIK, M.; GIRI, A. K. *Forecasting Foreign Exchange Rate: A Multivariate Comparative Analysis between Traditional Econometric, Contemporary Machine Learning & Deep Learning Techniques*. 2020. Disponível em: <https://doi.org/10.48550/arXiv.2002.10247>. Citado na página 23.

KAVA, L. E. *Além da Caixa Preta: Aprendizagem de Máquina Interpretável para Previsão de Séries Temporais Macroeconômicas Brasileiras*. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 2022. Disponível em: <https://repositorio.ufsc.br/handle/123456789/234659>. Citado 2 vezes nas páginas 21 e 22.

KYOBE, A. J.; DANNINGER, S. Revenue forecasting—how is it done? results from a survey of low-income countries. *IMF Working Paper*, 2005. Disponível em: <https://doi.org/10.5089/9781451860436.001>. Citado 2 vezes nas páginas 13 e 16.

LARSON, S. E.; OVERTON, M. Modeling approach matters, but not as much as preprocessing: Comparison of machine learning and traditional revenue forecasting techniques. *Public Finance Journal*, v. 1, n. 1, p. 29–48, 2024. Disponível em: <https://doi.org/10.59469/pfj.2024.8>. Citado na página 24.

LEAL, T. et al. Fiscal forecasting: Lessons from the literature and challenges. *Fiscal Studies*, v. 29, n. 3, p. 347–386, 2008. Disponível em: <https://ifs.org.uk/journals/fiscal-forecasting-lessons-literature-and-challenges>. Citado na página 20.

MAKRIDAKIS, S.; SPILIOTIS, E.; ASSIMAKOPOULOS, V. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, v. 36, n. 1, p. 54–74, 2020. Disponível em: <https://doi.org/10.1016/j.ijforecast.2019.04.014>. Citado 3 vezes nas páginas 14, 24 e 45.

MEDEIROS, R. K. d.; ARAGóN, E. K. d. S. B.; BESARRIA, C. d. N. Estratégias de previsão fiscal: um estudo empírico para a economia brasileira. In: ASSOCIAçãO NACIONAL DOS CENTROS DE PóS-GRADUAçãO EM ECONOMIA. *Anais do 50º Encontro Nacional de Economia*. 2022. Disponível em: <https://www.anpec.org.br/encontro/2022/submissao/files_I/i4-ff06fa7bb786e5f1039adaef6ed4559f.pdf>. Citado 2 vezes nas páginas 13 e 22.

MENDONçA, M. J.; MEDRANO, L. A. Um modelo de combinação de previsões para arrecadação de receita tributária no brasil. *Texto para Discussão*, n. 2186, 2016. Disponível em: <https://repositorio.ipea.gov.br/bitstream/11058/6610/1/td_2186.pdf>. Citado na página 22.

OFFICE FOR BUDGET RESPONSIBILITY. *Forecasting the Economy*. [S.l.], 2011. Disponível em: <https://obr.uk/docs/dlm_uploads/Forecasting-the-economy.pdf>. Citado 2 vezes nas páginas 13 e 19.

OFFICE FOR BUDGET RESPONSIBILITY. *Forecasting the Public Finances*. [S.l.], 2011. Disponível em: <https://obr.uk/docs/dlm_uploads/obr_briefing1.pdf>. Citado 2 vezes nas páginas 13 e 19.

OFFICE FOR BUDGET RESPONSIBILITY. *How We Present Uncertainty*. [S.l.], 2012. Disponível em: <https://obr.uk/docs/dlm_uploads/Briefing-paper-No4-How-we-present-uncertainty.pdf>. Citado na página 20.

OFFICE FOR BUDGET RESPONSIBILITY. *In-year Fiscal Forecasting and Monitoring*. [S.l.], 2018. Disponível em: <https://obr.uk/docs/dlm_uploads/In-year_fiscal_forecasting_and_monitoring_2.pdf>. Citado na página 20.

OLIVARES, K. G. et al. *NeuralForecast: User friendly state-of-the-art neural forecasting models*. 2022. PyCon Salt Lake City, Utah, US 2022. Disponível em: <https://github.com/Nixtla/neuralforecast>. Citado na página 40.

RAHIM, F. S.; WENDLING, C.; PEDASTSAAR, E. How to prepare expenditure baselines. *IMF How To Notes*, 2022. Disponível em: <https://doi.org/10.5089/9798400210129.061>. Citado na página 17.

SECRETARIA DO TESOURO NACIONAL. *Manual de Estatísticas Fiscais do Boletim Resultado do Tesouro Nacional*. Brasília, 2016. Disponível em: <https://www.tesourotransparente.gov.br/publicacoes/manual-de-estatisticas-fiscais-do-resultado-do-tesouro-nacional>. Citado 3 vezes nas páginas 26, 27 e 28.

SHAW, T. Long-term fiscal sustainability analysis: Benchmarks for independent fiscal institutions. *OECD Journal on Budgeting*, v. 17/1, p. 125–152, 2017. Disponível em: <https://doi.org/10.1787/budget-v17-1-en>. Citado na página 17.

STANKEVIčIūTė, K.; ALAA, A. M.; SCHAAR, M. van der. Conformal time-series forecasting. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2021. (NIPS '21). Citado na página 36.

STERN, E. et al. Cbo explains how it develops the budget baseline. *CBO Report*, 2023. Disponível em: <https://www.cbo.gov/publication/59085>. Citado 2 vezes nas páginas 13 e 19.

STOFFI, F. B.; CEVOLANI, G.; GNECCO, G. Simple models in complex worlds: Occam's razor and statistical learning theory. *Minds and Machines*, v. 32, 03 2022. Disponível em: <https://doi.org/10.1007/s11023-022-09592-z>. Citado na página 21.

WANG, X. et al. Forecast combinations: An over 50-year review. *International Journal of Forecasting*, v. 39, n. 4, p. 1518–1547, 2023. Disponível em: <https://ideas.repec.org/a/eee/intfor/v39y2023i4p1518-1547.html>. Citado na página 38.

XU, C.; XIE, Y. Conformal prediction for time series. In: *Proceedings of the 38th International Conference on Machine Learning*. [s.n.], 2021. Disponível em: <https://doi.org/10.48550/arXiv.2010.09107>. Citado na página 25.

Apêndices

APÊNDICE A – Descrição das variáveis

Tabela 16 – Variáveis utilizadas nos modelos

Despesa	Códigos no RTN
Benefícios Previdenciários	4.1
SJP de Benefícios Previdenciários	4.1.1.1 + 4.1.2.1
Pessoal e Encargos Sociais	4.2
SJP de Pessoal e Encargos Sociais	4.2.1
Abono Salarial	4.3.01.1
Seguro Desemprego	4.3.01.2
BPC da LOAS/RMV	4.3.05
SJP do BPC da LOAS/RMV	4.3.05.1
Créditos Extraordinários	4.3.07
FUNDEB	4.3.10
FCDF	4.3.11
Demais Poderes	4.3.12
Lei Kandir	4.3.13
SJP de Custeio e Capital	4.3.14
Subsídios, Subvenções e Proagro	4.3.15
Obrigatórias Diversas	4.3.02 + 4.3.03 + 4.3.04 + 4.3.06 + 4.3.08 + 4.3.09 + 4.3.16 + 4.3.17 + 4.3.18 + 4.3.19 + 4.3.20
Benefícios a SPF	4.4.1.1
Bolsa Família	4.4.1.2
Saúde Obrigatória	4.4.1.3
Educação Obrigatória	4.4.1.4
Diversas Obrigatórias CF	4.4.1.5
Saúde Discricionária	4.4.2.1
Educação Discricionária	4.4.2.2
Discricionárias Diversas	4.4.2.3 + 4.4.2.4 + 4.4.2.5 + 4.4.2.6 + 4.4.2.7 + 4.4.2.8 + 4.4.2.9

Nota: Elaborado com base na Tabela 1.2-A Resultado Primário do Governo Central do RTN.