30º Prêmio Tesouro de Finanças Públicas

Revista
Cadernos de Finanças Públicas

2026

Edição Especial

TESOURONACIONAL

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

Eduardo Jacomo Seraphim Nogueira

Universidade de Brasília - UnB

RESUMO

A previsão das despesas públicas é fundamental para o planejamento fiscal, mas em muitos países ainda se utiliza métodos simples e pouco robustos. Embora o uso de técnicas estatísticas seja consolidado, a aplicação de aprendizado de máquina e profundo ainda é limitada, sobretudo em despesas. Este trabalho investiga o desempenho de diferentes classes de modelos estatísticos, de aprendizado de máquina, profundo e suas combinações na previsão de séries de despesas primárias federais brasileiras. O estudo utiliza dados oficiais, otimização automática de parâmetros, validação cruzada temporal e previsão conformal para construir previsões e intervalos de confiança. Os resultados mostram que modelos estatísticos permanecem altamente competitivos, superando algoritmos mais complexos em horizontes longos, enquanto modelos profundos se destacam em horizontes curtos. A combinação de previsões, por sua vez, apresenta desempenho equilibrado. Conclui-se que técnicas avançadas de previsão são ferramentas úteis para subsidiar a política fiscal no Brasil.

Palavras-Chave: Previsão. Séries temporais. Política fiscal.
JEL: C53, H68, E62.

SUMÁRIO

1. INTRODUÇÃO

A elaboração de previsões fiscais confiáveis constitui atividade central para o funcionamento eficiente do setor público, tanto no âmbito da formulação de políticas, quanto na sustentabilidade das finanças públicas no médio e longo prazo (CEPAL, 2015). 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 (CEPAL, 2015).

Estudos como o de Kyobe & Danninger (2005) 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 (Kyobe & Danninger, 2005).

Como observado por Kyobe & Danninger (2005), 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).

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 inalterados (*no-policy change*), nos quais se estima o custo futuro dos serviços públicos assumindo a continuidade das políticas e estruturas existentes (Rahim et al., 2022).

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 et al., 2022). 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 (CEPAL, 2015). 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 et al., 2022), sendo particularmente críticas em contextos de consolidação fiscal ou reformas estruturais (FMI, 2014).

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). Além disso, Cameron (2022) 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 (CEPAL, 2015); e (iv) a dificuldade de antecipar o comportamento de agentes públicos na execução do gasto, especialmente em anos eleitorais (Hadzi-Vaskov et al., 2021).

Do ponto de vista metodológico, a literatura aponta que, tradicionalmente, as previsões

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. Acesso em: 21/9/2025.

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 (CEPAL, 2015). 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).

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 et al., 2018); (Bolhuis & Rayner, 2020). Além disso, Chen & Ranciere (2016) 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). Destaca-se ainda que a maioria 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). 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).

2. REVISÃO DE LITERATURA

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 Of-*

fice[2] (CBO), o *Office for Budget Responsibility*[3] (OBR), o Banco 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). 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).

Além disso, o relatório sobre a elaboração da linha de base orçamentária detalha como o CBO constrói suas previsões de despesa pública a partir de informações institucionais e projeções desagregadas por subfunções orçamentárias (Stern et al., 2023). 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). 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 (2011-a), 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, 2011-b). Os métodos utilizados vão desde extrapolações por tendência até projeções parametrizadas por

2 O CBO desempenha o papel de IFI nos Estados Unidos. Para mais informações, ver: https://www.cbo.gov/. Acesso em: 21/9/2025.

3 O OBR desempenha o papel de IFI no Reino Unido. Para mais informações, ver: https://www.obr.uk/. Acesso em: 21/9/2025.

4 O IFS é o principal instituto independente de pesquisa econômica do Reino Unido. Para mais informações, ver: https://www.ifs.org.uk/. Acesso em: 21/9/2025.

regras legais, além de ajustes discricionários com base em eventos recentes. Segundo OBR (2011-b), 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 execução 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), 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). Segundo OBR (2012), 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 IFS, 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). 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). 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).

As metodologias de previsão empregadas por essas instituições são diversas e frequentemente combinam diferentes abordagens em uma suíte 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 semiestruturais (Ciccarelli et al., 2024). 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).

Outros autores, como Cimadomo et al. (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 et al., 2017). 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 et al. (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 et al., 2013). 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.

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) 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*

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, Bargagli Stoffi et al. (2022) e Goldblum et al. (2024).

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 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).

No âmbito da arrecadação tributária federal, diversos estudos exploraram a previsão e a combinação de modelos. Medeiros et al. (2022) 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 et al. (2020) aplicaram técnicas de combinação simples e ótima proposta por Bates & 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), 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 & Costa (2021) 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).

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 & Mendonça (2017) 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).

Apesar dessas contribuições, nota-se que a maioria da literatura nacional se concentra 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.

3. METODOLOGIA

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).

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 sobreajuste 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 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, e 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 de todo o período. Apenas 4 das 24 séries originais exibem algum valor zero em determinado mês[6]. 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

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

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 1 descreve as variáveis do RTN utilizadas.

Tabela 1 – Variáveis utilizadas nos modelos[7]

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

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 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*[8].

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

8 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 et al., 2025).

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

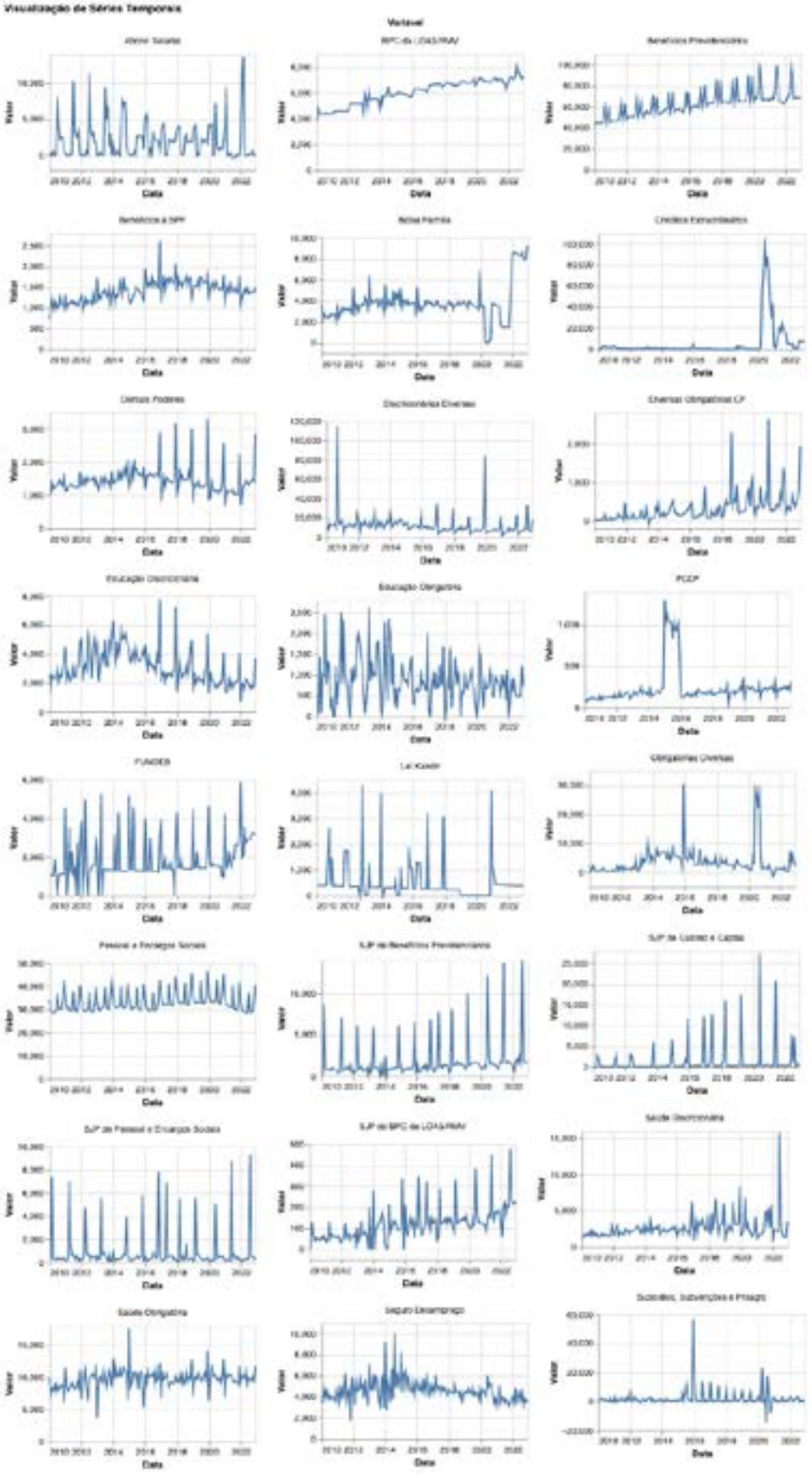
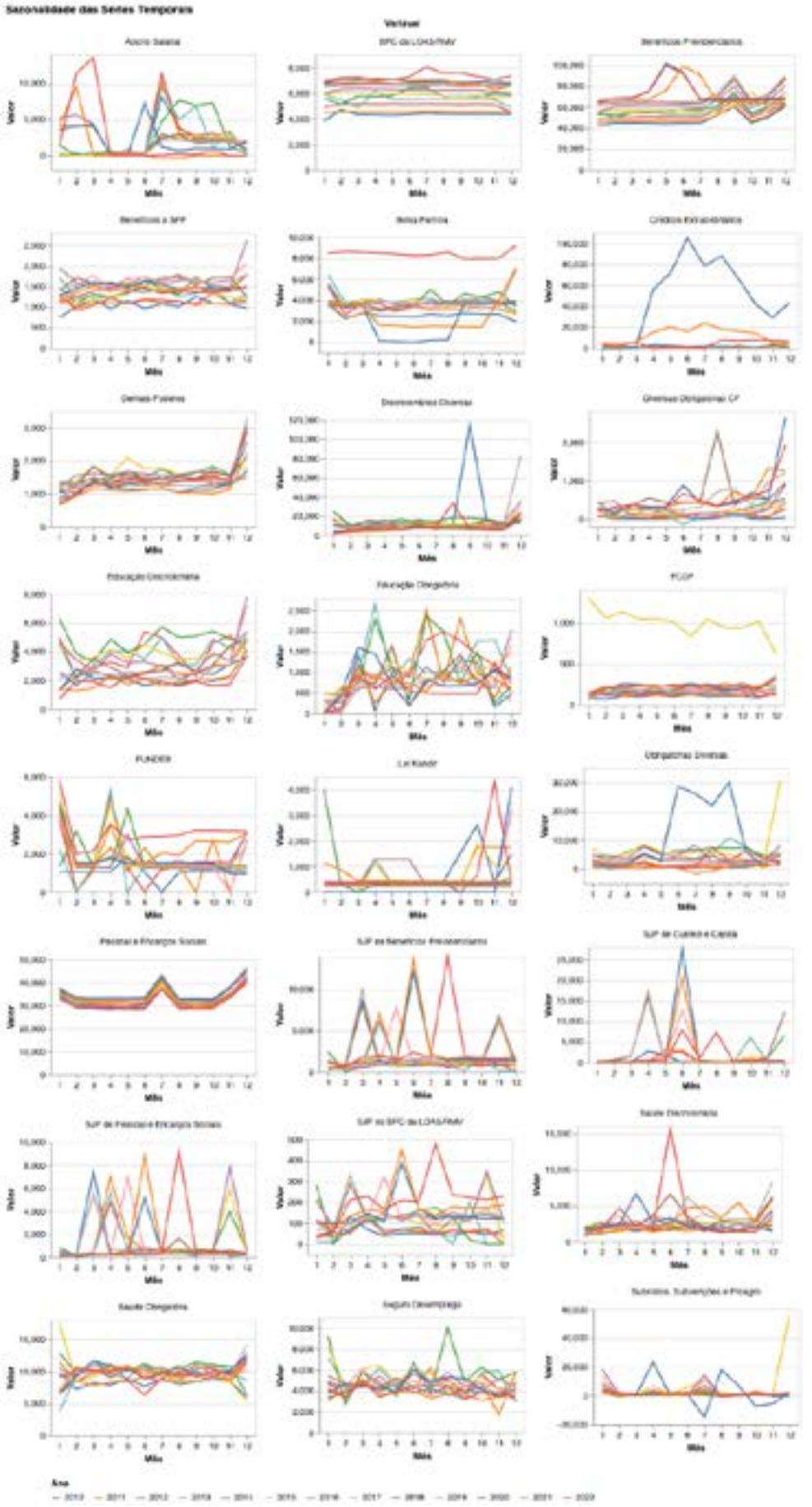


Figura 2 – Sazonalidade das despesas primárias em termos reais (milhões de R$)



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.

Estruturamos o processo de avaliação e seleção dos modelos preditivos em um *pipeline* modular, alinhado às melhores práticas em previsão de séries temporais, conforme 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[9]. A ideia central é identificar o modelo mais adequado para cada série de despesa pública, levando em conta suas características específicas, mantendo a robustez que pode ser obtida ao considerar diferentes abordagens. Como a escolha de modelos baseada ape-

9 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.

nas em intuição muitas vezes não se confirma nos dados, adotamos um procedimento de validação cruzada temporal com janela expansiva, que simula como os modelos se comportariam ao prever períodos futuros. Esse método, embora produza erros maiores do que o ajuste simples, aproxima-se mais da situação real de previsão, em que novas informações são incorporadas ao longo do tempo.

A comparação entre os modelos foi feita a partir de métricas de erro, principalmente o erro absoluto médio (MAE) e a raiz do erro quadrático médio (RMSE). O MAE é privilegiado por sua simplicidade de interpretação, enquanto o RMSE ajuda a capturar distorções associadas a valores extremos. O *pipeline* inicia-se com a divisão dos dados em conjuntos de treino e teste, preservando a ordem temporal. O conjunto de treino é usado para ajustar e validar os modelos, enquanto o teste serve para avaliar sua capacidade preditiva fora da amostra. Cada modelo é ajustado automaticamente a partir dos dados de treino, explorando diferentes paradigmas estatísticos, de aprendizado de máquina e de aprendizado profundo.

Após o ajuste, realizamos o diagnóstico dos resíduos, isto é, a análise dos erros gerados pelas previsões. Bons modelos devem produzir resíduos não correlacionados e sem viés, o que indica previsões mais confiáveis. Quando possível, também se avalia a constância da variância e a normalidade, ainda que essas propriedades não sejam indispensáveis. Para construir intervalos de confiança das previsões, utilizamos a abordagem conformal, que, em vez de depender de suposições fortes sobre distribuições probabilísticas, recorre ao histórico dos erros para estimar a incerteza futura. Essa técnica assegura que os intervalos de previsão reflitam de forma realista o grau de incerteza do modelo.

A etapa seguinte é a validação cruzada temporal no conjunto de treino, que consiste em treinar e avaliar os modelos em janelas crescentes de dados, testando sua estabilidade e capacidade de generalização. O modelo escolhido para cada série é aquele que apresenta o menor erro médio nesse processo, usando o MAE como critério principal. Esse modelo é então ajustado em todo o conjunto de treino e aplicado ao conjunto de teste, permitindo medir sua real capacidade de previsão em dados novos. Por fim, para aumentar a robustez, as previsões finais são obtidas a partir da combinação de diferentes paradigmas de modelos. A literatura mostra que, de forma consistente, a média simples das previsões costuma superar o desempenho de modelos individuais. A incerteza das previsões combinadas também é estimada de maneira empírica, utilizando novamente os erros passados.

Em síntese, esse processo busca garantir que a seleção de modelos seja feita de maneira sistemática e transparente, equilibrando rigor estatístico e flexibilidade prática. Em vez de

depender de escolhas subjetivas, compara-se o desempenho de várias abordagens, avaliam-se seus erros de forma realista e combinam-se previsões para produzir resultados mais confiáveis.

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 & 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.

A Tabela 2 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 2 – Desempenho dos modelos *benchmark* e estatísticos no conjunto de treino[10]

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

A Tabela 3 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 3 – Frequência de seleção como melhor modelo *benchmark* e estatístico no conjunto de treino[11]

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

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,

10 Nota: Resultados ordenados de forma crescente pelo MAE.
11 Nota: Resultados ordenados de forma decrescente pelo MAE.

indicando que o ajuste personalizado é fundamental para maximizar o desempenho preditivo nesse contexto.

A Tabela 4 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 4 – Desempenho dos modelos estatísticos comparados ao *benchmark* no conjunto de teste[12]

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%

A Tabela 5 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 5 – 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

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

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 se limitar a reproduzir apenas tendências triviais. Em seguida, avaliamos o desempenho dos modelos de *Machine Learning* e suas nuances frente ao contexto fiscal brasileiro.

A Tabela 6 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 6 – Desempenho médio dos modelos de *Machine Learning* no conjunto de treino[13]

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

A Tabela 7 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

13 Nota: Resultados ordenados de forma crescente pelo MAE.

de despesas públicas.

Tabela 7 – Frequência de seleção como melhor modelo de *Machine Learning* no conjunto de treino[14]

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	–	–

A Tabela 8 resume as métricas de erro dos melhores modelos de ML no conjunto de teste, permitindo comparação com o *benchmark*.

Tabela 8 – Desempenho dos modelos de *Machine Learning* comparados ao *benchmark* no conjunto de teste[15]

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%

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

14 Nota: Resultados ordenados de forma decrescente pelo MAE.

15 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.

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 et al. (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.

A Tabela 9 sintetiza os resultados comparativos dos melhores modelos de ML entre treino e teste, considerando as principais métricas.

Tabela 9 – 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.

A Tabela 10 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 10 – Desempenho médio dos modelos de *Deep Learning* no conjunto de treino[16]

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

A Tabela 11 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.

16 Nota: Resultados ordenados de forma crescente pelo MAE.

Tabela 11 – Frequência de seleção como melhor modelo de *Deep Learning* no conjunto de treino[17]

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

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.

A Tabela 12 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.

17 Nota: Resultados ordenados de forma decrescente pelo MAE.

Tabela 12 – Desempenho dos modelos de *Deep Learning* comparados ao *benchmark* no conjunto de teste[18]

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%

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. A Tabela 13 sintetiza os resultados comparativos dos melhores modelos de DL entre treino e teste, considerando as principais métricas.

Tabela 13 – 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.

18 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.

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. A Tabela 14 apresenta o desempenho da combinação de previsões no conjunto de teste, comparando-a ao *benchmark* ingênuo.

Tabela 14 – Desempenho da combinação de previsões comparado ao *benchmark* no conjunto de teste[19]

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%

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 a combinação supera os estatísticos apenas no SMAPE, e por pequena mar-

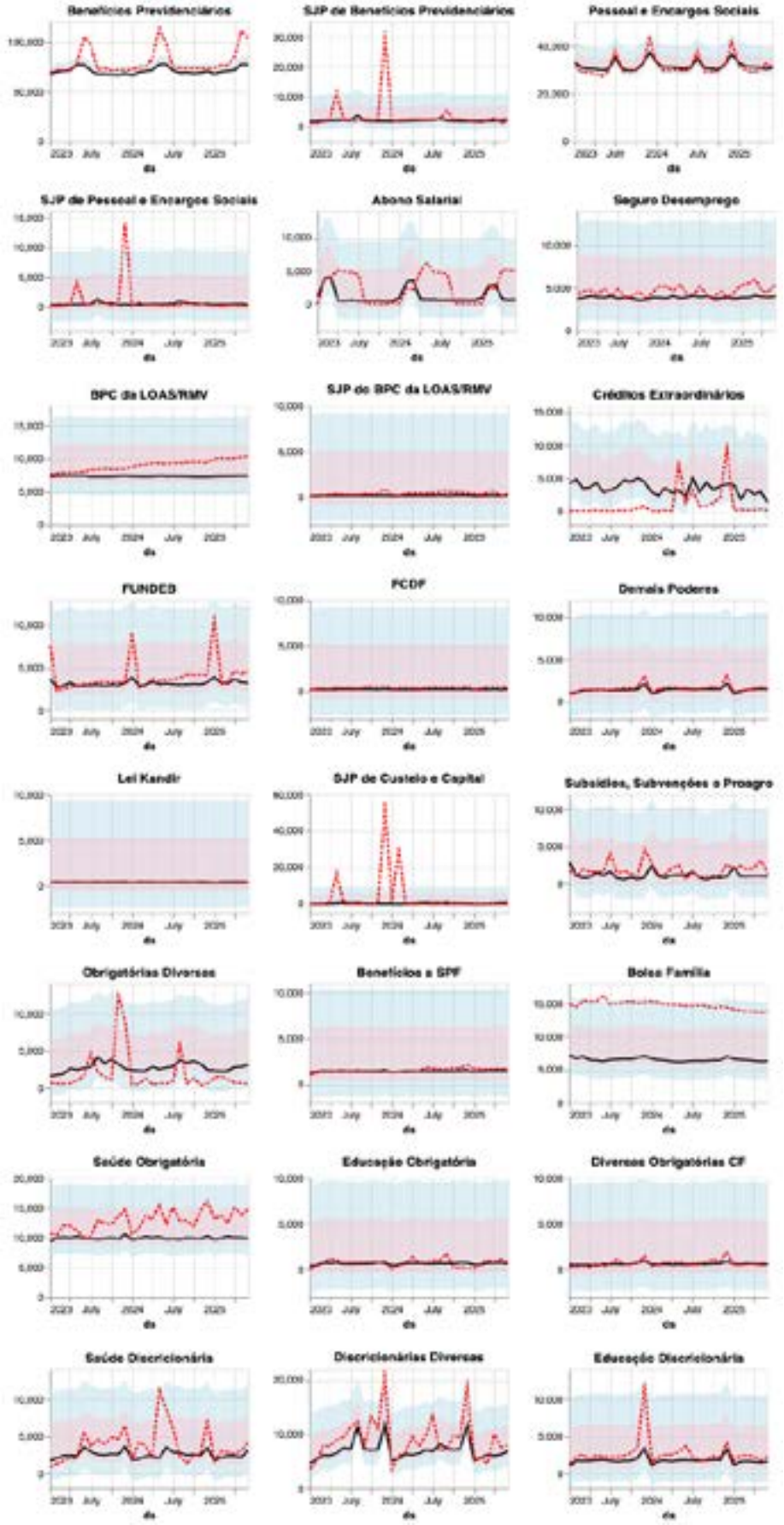
19 Nota: A coluna redução do erro mostra a melhoria percentual da combinação de previsões em relação ao benchmark para cada métrica.

gem. Nas 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 é raramente 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)[20]



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

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 15 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 15 – Desempenho dos modelos no teste de robustez (horizonte de 18 meses)[21]

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	0,96	1,11	0,98	0,93

A Tabela 16 apresenta a redução percentual dos erros de cada modelo em relação ao *baseline*.

Tabela 16 – Redução percentual dos erros em relação ao *baseline* (teste de robustez)[22]

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	8,21%	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%

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 estatísti-

21 Nota: Resultados para o conjunto de teste de janeiro de 2024 a junho de 2025.
22 Nota: Cálculo: 1 - (erro do modelo / erro baseline).

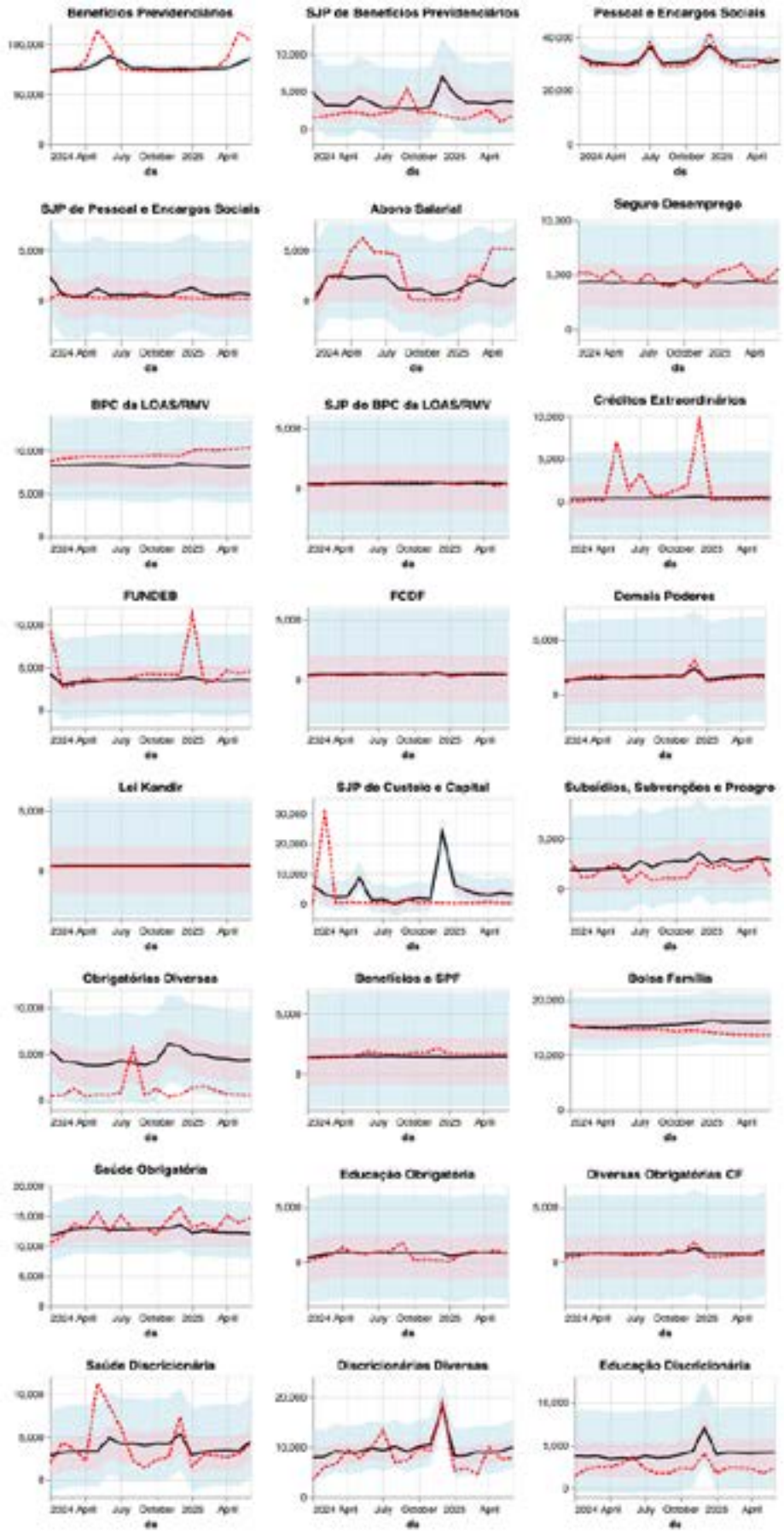
cos e de ML preservam consistência e robustez, enquanto a combinação de previsões mostra-se especialmente eficaz na redução de MSE, MSSE e RMSSE. Notamos ainda que os ganhos do DL em relação ao *baseline* são notáveis, sugerindo melhor adaptação a choques recentes ou mudanças de padrão.

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)[23]



23 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 confirmam 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 mos-

tra 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 quase 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 semiestruturais 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 semiestruturais, de modo a ampliar a acurácia, a robustez e a interpretabilidade dos modelos preditivos.

REFERÊNCIAS BIBLIOGRÁFICAS

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.

ANDO, S.; KIM, T. **Systematizing macroframework forecasting: High-dimensional conditional forecasting with accounting identities**. IMF Working Paper, 2022.

APICELLA, A.; ISGRò, F.; PREVETE, R. **Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning**. 2025

ARNOLD, R. W. **How cbo produces its 10-year economic forecast**. CBO Working Paper, 2018.

ASIMAKOPOULOS, I.; PAREDES, J.; WARMEDINGER, T. **Forecasting Fiscal Time Series Using Mixed Frequency Data**, 2013. ECB Working Paper No. 1553.

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.

BATES, J. M.; GRANGER, C. W. J. **The combination of forecasts**. Operational Research Quarterly, v. 20, n. 4, p. 451–468, 1969.

BERGMEIR, C.; HYNDMAN, R. J.; KOO, B. **A note on the validity of crossvalidation for evaluating autoregressive time series prediction.** Computational Statistics & Data Analysis, v. 120, p. 70–83, 2018.

BOLHUIS, M. A.; RAYNER, B. **Deus ex machina: A framework for macroforecasting with machine learning**. IMF Working Paper, 2020.

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.

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.

CHEN, S.; RANCIERE, R. **Financial information and macroeconomic forecasts**. IMF Working Paper, 2016.

CICCARELLI, M. et al. **Ecb macroeconometric models for forecasting and policy analysis**. ECB Occasional Paper, n. 344, 2024.

CIMADOMO, J.; GIANNONE, D.; LENZA, M. **Fiscal Nowcasting**. [S.l.], 2017.

CLEMEN, R. T. **Combining forecasts: A review and annotated bibliography**. International Journal of Forecasting, v. 5, n. 4, p. 559–583, 1989.

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

DEUS, J. D. B. V. d.; MENDONçA, H. F. d. **Fiscal forecasting performance in na emerging economy: An empirical assessment of Brazil**. Economic Systems, v. 41, n. 3, p. 408–419, 2017.

CEPAL. **Revenue and Expenditure Forecasting Methods for a PER Spending**. Santiago, Chile, 2015.

EICHER, T. S. et al. **Forecasting in times of crises**. IMF Working Paper, 2018.

FAVERO, C. A.; MARCELLINO, M. **Modelling and forecasting fiscal variables for the euro area.** IGIER Working Paper, n. 298, 2005.

FORONI, C. **Discussion of "fiscal nowcasting".** Conference presentation. 2017.

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.

GARZA, A. et al. **StatsForecast: Lightning fast forecasting with statistical and econometric models.** 2022. PyCon Salt Lake City, Utah, US 2022.

GODAHEWA, R. et al. **Monash time series forecasting archive**. In: Neural Information Processing Systems Track on Datasets and Benchmarks. [s.n.], 2021.

GOLDBLUM, M. et al. **The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning**. 2024.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016.

HADZI-VASKOV, M. et al. **Authorities fiscal forecasts in latin america: Are they optimistic?** IMF Working Paper, 2021.

HAMILTON, J. D. **Time Series Analysis.** Princeton, NJ: Princeton University Press, 1994.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009.

HYNDMAN, R. J. et al. **Forecasting: Principles and Practice, the Pythonic Way**. Melbourne, Australia: OTexts, 2025.

HYNDMAN, R. J.; KHANDAKAR, Y. A**utomatic time series forecasting: The forecast package for R.** Journal of Statistical Software, v. 27, n. 3, p. 1–22, 2008.

FMI. **IMF Forecasts: Process, Quality, and Country Perspectives**. Washington, D.C., 2014.

JUNG, J.-K.; PATNAM, M.; TER-MARTIROSYAN, A. **An algorithmic crystal ball: Forecasts based on machine learning.** IMF Working Paper, 2018.

KAUSHIK, M.; GIRI, A. K. **Forecasting Foreign Exchange Rate: A Multivariate Comparative Analysis between Traditional Econometric, Contemporary Machine Learning & Deep Learning Techniques.** 2020.

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.

KYOBE, A. J.; DANNINGER, S. **Revenue forecasting - how is it done? Results from a survey of low-income countries.** IMF Working Paper, 2005.

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.

LEAL, T. et al. **Fiscal forecasting: Lessons from the literature and challenges.** Fiscal Studies, v. 29, n. 3, p. 347–386, 2008.

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.

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: ANPEC. Anais do 50º Encontro Nacional de Economia. 2022.

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.

OBR. **Forecasting the Economy.** [S.l.], 2011-a.

OBR. **Forecasting the Public Finances.** [S.l.], 2011-b.

OBR. **How We Present Uncertainty.** [S.l.], 2012.

OBR. **In-year Fiscal Forecasting and Monitoring.** [S.l.], 2018.

OLIVARES, K. G. et al. **NeuralForecast: User friendly state-of-the-art neural forecasting models.** 2022. PyCon Salt Lake City, Utah, US 2022.

RAHIM, F. S.; WENDLING, C.; PEDASTSAAR, E. **How to prepare expenditure baselines.** IMF How To Notes, 2022.

STN. **Manual de Estatísticas Fiscais do Boletim Resultado do Tesouro Nacional.** Brasília, 2016.

SHAW, T. **Long-term fiscal sustainability analysis: Benchmarks for independent fiscal institutions.** OECD Journal on Budgeting, v. 17/1, p. 125–152, 2017.

STANKEVIčIuTe, 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).

STERN, E. et al. **CBO explains how it develops the budget baseline.** CBO Report, 2023.

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.

WANG, X. et al. **Forecast combinations: An over 50-year review.** International Journal of Forecasting, v. 39, n. 4, p. 1518–1547, 2023.

XU, C.; XIE, Y. **Conformal prediction for time series.** In: Proceedings of the 38th International Conference on Machine Learning. [s.n.], 2021.