30º Prêmio Tesouro
de Finanças Públicas

Revista
Cadernos de Finanças Públicas

2026

Edição Especial

TESOURO NACIONAL

Forecasting Central Government Primary Expenditure: A Comparative Analysis of Statistical Techniques, Machine Learning, Deep Learning, and Combination of Forecasts

Eduardo Jacomo Seraphim Nogueira

Universidade de Brasília - UnB

ABSTRACT

Forecasting public expenditures is essential for fiscal planning, but in many countries simple and less robust methods are still used. Although the use of statistical techniques is well established, the application of Machine Learning and Deep Learning remains limited, especially in expenditure forecasting. This paper investigates the performance of different classes of statistical models, Machine Learning, Deep Learning, and their combinations in forecasting Brazilian federal primary expenditure series. The study employs official data, automatic parameter optimization, temporal cross-validation, and conformal prediction to build forecasts and confidence intervals. The results show that statistical models remain highly competitive, outperforming more complex algorithms in long horizons, while deep models perform better in short horizons. Forecast combinations, in turn, deliver, balanced performance. It is concluded that advanced forecasting techniques are useful tools to support fiscal policy in Brazil.

Keywords: forecasting; time series; fiscal policy.
JEL: C53, H68, E62.

SUMMARY

1. INTRODUCTION

The preparation of reliable fiscal forecasts is a central activity for the efficient functioning of the public sector, both in terms of policy formulation and the sustainability of public finances in the medium and long term (ECLAC, 2015). The quality of these projections directly affects the credibility of governments, compliance with fiscal rules, and the efficient allocation of public resources. Although the literature on fiscal forecasting has traditionally focused on revenue modeling, there is growing consensus on the need for public expenditure to be subject to rigorous methodological analysis (ECLAC, 2015).

Studies such as that by Kyobe & Danninger (2005) investigate in depth revenue forecasting practices in developing countries, highlighting the scarcity of research on their determinants. The authors note that in low-income countries, revenue projections are predominantly made in aggregate form, whereas in higher-income countries, more disaggregated data are used. In addition, the authors point out that, although more sophisticated statistical methods are used in certain contexts, the use of subjective assessments and simple extrapolation techniques prevails as the dominant practice in most low-income countries for deriving revenue projections (Kyobe & Danninger, 2005).

As noted by Kyobe & Danninger (2005), most developing countries still use very simple or subjective methods, or mere extrapolations, to make revenue projections, which is no different from the reality observed in Brazil, both for revenue forecasting and public expenditure projections. As a result, fiscal policy managers rely heavily on *spreads*heet-based structures, judgment, and projections by national authorities, which are subject to discretionary adjustments that are easier to manipulate and difficult to detect, to the detriment of formal econometric models (Kyobe & Danninger, 2005).

In contrast, public expenditure projections are traditionally approached with a methodology that, although fundamental to budget management, often lacks the same analytical depth based on predictive statistical models. The predominant concept for expenditure projection lies in the development of baselines or no-policy-change scenarios, in which the future cost of public services is estimated assuming the continuity of existing policies and structures (Rahim et al., 2022).

These baselines are constructed from input costs (labor, operating costs, equipment), whose factors (price and volume) are adjusted by macroeconomic parameters such as inflation or population growth (Rahim et al., 2022). According to the authors, although this approach is vital for planning and fiscal discipline, it focuses on funding existing policies rather than predicting future expenditure behavior based on complex statistical relationships and underlying economic and social dynamics.

Several international organizations emphasize the strategic importance of expenditure forecasting. The Economic Commission for Latin America and the Caribbean (ECLAC) points out that, especially in Latin American countries, public spending has structural characteristics that make it challenging to forecast, such as budgetary rigidity, legal constraints, and exposure to exogenous shocks (ECLAC, 2015). The International Monetary Fund (IMF) reinforces that expenditure forecasts are crucial for constructing budget baselines and conducting fiscal sustainability analyses (Rahim et al., 2022), and are particularly critical in contexts of fiscal consolidation or structural reforms (IMF, 2014).

The Organization for Economic Cooperation and Development (OECD) argues that well--founded expenditure forecasts are essential for the effective performance of Independent Fiscal Institutions (IFIs)[1], as they enable the early detection of fiscal risks and the improvement of budgetary transparency (Shaw, 2017). In addition, Cameron (2022) recommends that forecasts be systematically evaluated through *ex-post* review processes, with an emphasis on institutional learning, rather than just on-point accuracy.

Despite recognition of the importance of the issue, public expenditure forecasting faces a number of practical challenges. Among the most relevant are: (i) the low frequency and small number of observations available in historical series; (ii) the presence of structural breaks resulting from institutional changes, legal changes, or accounting reclassifications; (iii) the coexistence of highly rigid components (such as retirement and pensions, personnel, and mandatory transfers) with discretionary portions subject to political instability and contingencies (ECLAC, 2015); and (iv) the difficulty of anticipating the behavior of public agents in the execution of spending, especially in election years (Hadzi-Vaskov et al., 2021).

From a methodological point of view, the literature points out that, traditionally, expenditure forecasts are based on deterministic methods, *spreads*heets structured by elasticities, and univariate or multivariate statistical models, such as linear regressions, integrated autoregressive moving average models, and exponential smoothing models (ECLAC, 2015). Structural models are more commonly used in medium- and long-term analyses, often incorporating

[1] According to the OECD, Independent Fiscal Institutions (IFIs) are independent public institutions with a mandate to critically assess and, in some cases, provide impartial advice on fiscal policy and performance. IFIs aim to promote sound fiscal policy and sustainable public finances, helping to promote greater transparency in public accounts. Available at: https://www.oecd.org/en/topics/parliamentary-budget-offices-and-independent-fiscal-institutions.html. Accessed on: 9/21/2025.

exogenous macroeconomic projections for variables such as GDP, inflation, and demographics (Ando & Kim, 2022).

In recent years, there has been growing interest in the use of *machine learning* (ML) and *deep learning* (DL) methods in fiscal forecasting. Recent empirical studies published by the IMF explore this agenda on several fronts and suggest that *machine learning* models can outperform traditional methods in certain predictive tasks, especially when there are large volumes of data, multiple correlated variables, and complex nonlinear patterns (Jung et al., 2018); (Bolhuis & Rayner, 2020). In addition, Chen & Ranciere (2016) demonstrate that high-frequency financial information, such as sovereign *spreads* and market interest rates, can anticipate the behavior of fiscal variables, indicating potential complementary predictive value.

However, it is also recognized that these methods have important limitations, such as low interpretability, sensitivity to *overfitting*, and the need for careful calibration (Bolhuis & Rayner, 2020). It should also be noted that most ML and DL applications in the fiscal field focus on forecasting tax revenue or economic growth, with a scarcity of empirical evidence on their performance in forecasting public expenditure (Eicher et al., 2018). Furthermore, there are few studies that systematically compare different model families, evaluating predictive performance based on multiple metrics, confidence intervals, and out-of-sample validation (Ando & Kim, 2022).

2. LITERATURE REVIEW

International literature on fiscal forecasting has advanced significantly in recent decades, both from a methodological and institutional perspective. This development is intrinsically linked to the growing demand for credible fiscal rules, the need for intertemporal coordination of public policies, and the centrality of transparency in modern budgetary governance. In this context, institutions such as the Congressional Budget Office[2] (CBO), the *Office for Budget Responsibility*[3] (OBR), the European Central Bank (ECB), and the *Institute for Fiscal Studies*[4] (IFS) play key roles in developing and disseminating methodologies for fiscal projections with a high level of technical rigor.

[2] The CBO plays the role of IFI in the United States. For more information, see: https://www.cbo.gov/. Accessed on: 9/21/2025.

[3] The OBR plays the role of an IFI in the United Kingdom. For more information, see: https://www.obr.uk/. Accessed on: 9/21/2025.

[4] The IFS is the UK's leading independent economic research institute. For more information, see: https://www.ifs.org.uk/. Accessed on: 9/21/2025.

In the case of the United States, the CBO adopts a detailed and structured approach to preparing its macroeconomic and fiscal projections. In its methodological report, the agency describes the process of formulating long-term forecasts based on structural models and technical judgment analysis. The CBO's main model incorporates a general equilibrium structure with nominal and real rigidity, integrating rational expectations of agents, assumptions about productivity, interest rates, and labor force participation (Arnold, 2018). The agency acknowledges that these projections are highly sensitive to assumptions about demographic trends and productivity growth, which implies constant revision of parameters and periodic validation with historical data (Arnold, 2018).

In addition, the report on the preparation of the budget baseline details how the CBO constructs its public expenditure forecasts based on institutional information and projections disaggregated by budget subfunctions (Stern et al., 2023). The forecast for each subcomponent of expenditure takes into account legal rules, historical trends, demographic pressures, and projected macroeconomic parameters, ensuring consistency between budget blocks (Stern et al., 2023). This *bottom-up* construction strategy serves as a reference for the empirical approach adopted in this study.

In the UK experience, the OBR stands out for both its technical sophistication and its commitment to transparency. As stated in OBR (2011-a), the OBR's macroeconomic and fiscal forecasts combine econometric models of aggregate demand with institutional judgment in order to incorporate qualitative information from ministries, agencies, and public sector experts. This integration between formal models and tacit knowledge is essential to capture institutional aspects that are often absent in purely quantitative approaches.

In the fiscal sphere, the OBR employs a disaggregated methodology, projecting mandatory and discretionary expenditure categories and interest charges separately (OBR, 2011-b). The methods used range from trend extrapolations to projections parameterized by legal rules, as well as discretionary adjustments based on recent events. According to OBR (2011-b), this type of adjustment is indispensable in the face of legislative changes or administrative reorganizations, highlighting the importance of technical judgment in the predictive process.

Another relevant aspect of the OBR's work is its emphasis on monitoring fiscal execution throughout the year, with periodic revisions of forecasts as budget execution data becomes available. As described in OBR (2018), intra-annual updates are made based on monthly reports and administrative data, allowing forecasts to be revised according to signs of deviation in budgetary behavior. This process becomes especially important in contexts of high political or economic volatility, in which the rigidity of static forecasts compromises their usefulness for fiscal policy.

In the field of uncertainty communication, the OBR has developed pioneering practices for the probabilistic representation of fiscal projections. The agency uses fan charts to express confidence intervals around deficit projections and other fiscal variables, based on the empirical distribution of past forecast errors (OBR, 2012). According to OBR (2012), such charts are useful for communicating the range of uncertainty associated with projections, even if they do not accurately capture extreme events. These practices are directly related to the objectives of this study, which measures and compares the predictive performance of different models that are automatically updated and based not only on point projections but also on interval metrics, using conformal forecasting.

Broadening the institutional perspective, the study by Leal et al. (2008), published by the IFS, offers a critical and comprehensive view of the challenges inherent in fiscal forecasts. The authors highlight the recurring problem of optimistic biases in official forecasts, especially in election years, and argue that such distortions reduce the credibility of fiscal policy and undermine the sustainability of public accounts (Leal et al., 2008). The tension between transparency and technical sophistication is also addressed, pointing out that more complex models, although potentially more accurate, tend to be less understandable to the public and more difficult to audit (Leal et al., 2008). The authors also highlight that forecasting public expenditure is more difficult than forecasting revenue, due to the institutional rigidity of many expenditures and the unpredictability of discretionary policies. European literature points out that, even with fiscal rules, expenditure forecasting errors persist and are concentrated in items subject to political or accounting volatility (Leal et al., 2008).

The forecasting methodologies employed by these institutions are diverse and often combine different approaches in a suite of models. The European Central Bank also adopts this strategy, which seeks to balance complexity and simplicity, empirical adjustment, and theoretical soundness. Macroeconometric models can be categorized, for example, into Dynamic Stochastic General Equilibrium (DSGE) models, Vector Autoregression (VAR) models, and semi-structural models (Ciccarelli et al., 2024). In the fiscal field, satellite DSGE models are also used to analyze fiscal multipliers under tight monetary policy, with forward guidance and quantitative easing, often including a relatively wide range of fiscal policy instruments (Ciccarelli et al., 2024).

Other authors, such as Cimadomo et al. (2017), propose the use of *nowcasting* techniques based on Bayesian models with mixed-frequency data to forecast fiscal balances. The model uses higher frequency (monthly) cash flow data to anticipate changes in the annual balance, containing few variables and no judgment, demonstrating that more parsimonious approaches [5]can outperform official forecasts based on hundreds of variables (Cimadomo et al., 2017). Foroni's (2017) critical discussion suggests improvements such as the use of mixed-frequency data sampling (MiDaS) models or the decomposition of the balance into revenues and expenditures, a proposal in line with the scope of this study.

The article by Asimakopoulos et al. (2013) presents an empirical approach to forecasting revenue and expenditure components based on high-frequency data. The authors use MiDaS models to combine quarterly data with annual targets and show that disaggregated (*bottom-up*) forecasts, updated whenever possible, generate significant predictive gains, especially on the expenditure side (Asimakopoulos et al., 2013). This conclusion supports the empirical strategy adopted in this study, which estimates specific public expenditure series individually and updates the models with each new piece of information available.

The national literature on fiscal forecasts is still relatively scarce and focuses mainly on aggregate variables, such as tax revenue and primary results, with few studies on public expenditure itself. In general, Brazilian studies address three fronts: macroeconomic forecasts, federal and state revenue forecasts, and analyses of budget execution errors. The absence of studies focused on disaggregated federal expenditure forecasts reflects a gap that this study seeks to fill.

In the field of macroeconomic forecasts, Kava (2022) applies *machine learning* methods to forecast Brazilian inflation, economic activity, and interest rate series, comparing the performance of algorithms such as Random Forest, neural networks, and *Gradient Boosting* with traditional models such as ARMA and VAR. The results show consistent gains for *machine learning* models in the short term, although the author emphasizes the need for careful validation procedures and hyperparameter tuning. This experiment, although focused on macroeconomic series, reveals the feasibility of using *machine learning* techniques on Brazilian data and introduces agnostic interpretation methodologies (Kava, 2022).

In the context of federal tax collection, several studies have explored forecasting and model combination. Medeiros et al. (2022) compared different supervised learning algorithms, such as *Elastic Net, Complete Subset Regression* (CSR), and *bagging* techniques, to predict monthly federal tax collection between 2002 and 2021. The authors found that *Elastic Net*

[5] Several authors argue in favor of parsimony, a version of Occam's Razor, a principle that holds that simpler models with fewer parameters are generally preferable due to the trade-off between bias and variance. See, for example, Bargagli Stoffi et al. (2022) and Goldblum et al. (2024).

performed best in the short term, followed by CSR, while *bagging* methods performed worse, especially for smaller samples or longer horizons. In addition, simple *benchmarks*, such as the mean or median of individual forecasts, proved competitive, corroborating the international literature on forecast combination.

Similarly, Gadelha et al. (2020) applied simple and optimal combination techniques proposed by Bates & Granger (1969) to project the collection of nine federal taxes, concluding that the combination of forecasts consistently outperforms individual models such as SARIMA and the Holt-Winters (HW) triple exponential smoothing method. Similar results were obtained by (Mendonça & Medrano, 2016), who showed gains in accuracy when employing dynamic factorial models associated with weighted linear combinations to reduce bias and mean square error. These studies highlight the relevance of *ensemble* approaches in the Brazilian fiscal context, even in scenarios of low frequency and high institutional volatility.

With regard to expenditure, the available empirical evidence is more limited and focuses mainly on the analysis of projection errors. Carneiro & Costa (2021) analyzed the determinants of expenditure forecast error in Brazilian municipalities, showing that rigid categories, such as personnel and charges, are more accurate, while discretionary expenditures, such as investments, tend to be systematically underestimated. The authors identified that outstanding payments and budgetary incrementalism practices are structural factors that perpetuate errors, while political variables, such as election years, had less impact than expected. In addition, municipalities with greater financial autonomy and better fiscal management quality had more accurate forecasts (Carneiro & Costa, 2021). These findings suggest that, at the federal level as well, institutional rigidity can facilitate the forecasting of certain expenditures, while discretionary categories remain more volatile.

Deus & Mendonça (2017) present a further advance by analyzing the quality of aggregate fiscal forecasts in Brazil between 2003 and 2013. The study reveals the existence of persistent optimistic bias, especially in election years, as well as the influence of gross domestic product (GDP) forecast errors on fiscal outcomes. The authors argue that the low efficiency of forecasts is related not only to the economic cycle but also to institutional fragility, which reduces incentives for realistic projections. This diagnosis is in line with the international literature in demonstrating that fiscal errors in emerging economies are not random but structural, reflecting both technical limitations and political incentives (Deus & Mendonça, 2017).

Despite these contributions, it should be noted that most of the national literature focuses on aggregate revenues and balances, leaving aside detailed forecasts of federal public expenditure. Furthermore, almost all studies are limited to specific forecasts, without probabilistic or interval assessment, and do not make systematic comparisons between families of statistical, *machine learning*, and *deep learning* models. Finally, we found no studies that adopt a *bottom--up* approach to federal expenditures, disaggregating by specific categories or programs, which reinforces the originality and relevance of this study.

3. METHODOLOGY

We used the National Treasury Results Bulletin (RTN) as the source of federal public expenditure data used in the study. The RTN, a monthly publication of the National Treasury Secretariat (STN), has been the official reference for measuring the primary results of the Brazilian Central Government since 1995. The bulletin transparently consolidates fiscal statistics on revenue and expenditure performance and is recognized as the main instrument for monitoring federal budget execution and analyzing current fiscal policy (STN, 2016).

The RTN presents 159 lines of variables, covering revenues (57), expenditures (92), and fiscal results (10). The scope of this study focuses exclusively on variables associated with primary public expenditures. To construct the database, we defined the start of the historical series analyzed as the cut-off date. Although the RTN has data since 1997 for aggregate expenditures, the level of detail required to identify and analyze specific lines of expenditure individually is only available from January 2010 onwards. We limited the time series to this starting point, ensuring comparability and homogeneity of detail throughout the period analyzed, avoiding problems of missing data, breaks in the series, or classification inconsistencies.

Regarding the level of granularity of the forecasts, we initially selected 41 expenditure lines, constituting the smallest group that uniquely identifies each RTN expenditure, without redundancies or trivial linear combinations. Other available lines correspond to more specific details, without predictive interest, or result from aggregations of the variables already selected, adding no information to the model. During exploratory analysis, we observed many months with zero values in some variables of the original set with 41 expense lines.

We also identified that 18 of these variables had individually irrelevant values from a budgetary materiality perspective, each representing a very small fraction of total federal expenditures. As an objective criterion, we consider variables whose sum represents less than 5% of total expenditures in the analyzed period to be materially irrelevant. To avoid *overfitting* in uninformative variables, we aggregated them into two new variables: one with the sum of

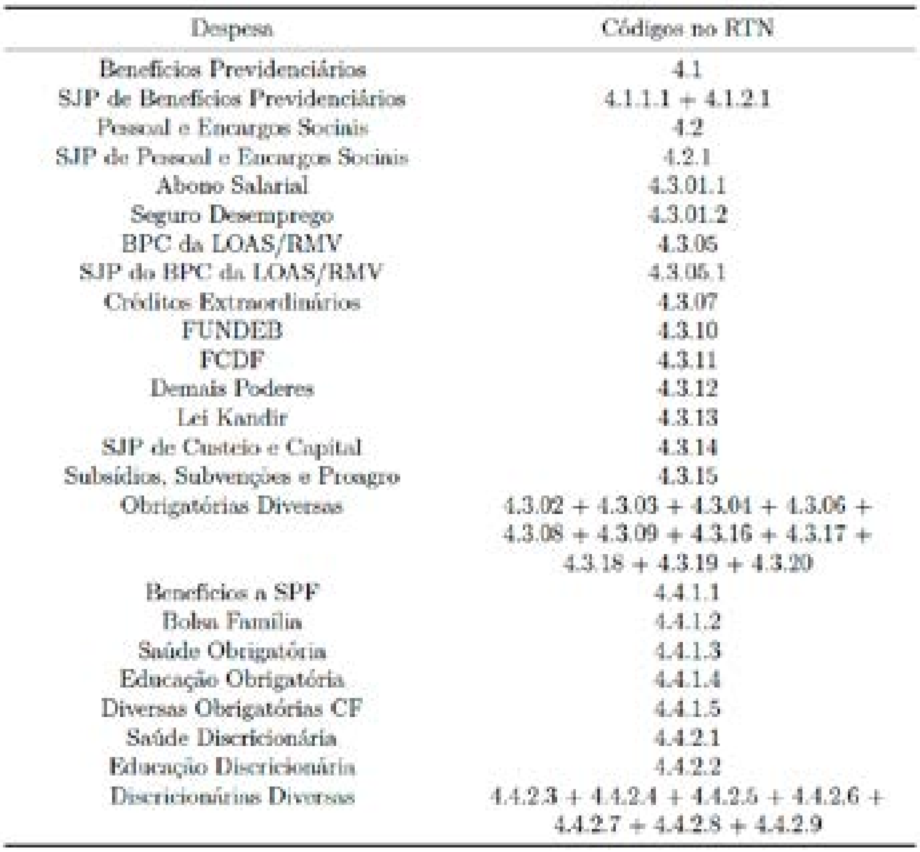
11 mandatory expenditures and another with the sum of 7 discretionary expenditures. Additionally, we concatenated the two lines of expenditures with court rulings and social security benefit warrants (urban and rural) into a single variable.

A relevant issue concerns the presence of missing or zero values in the series. We observed a large number of months with values equal to zero in some variables of the original set with 41 expenditure lines. It should be noted, however, that these records do not correspond to missing data or omissions, but to actual situations of absence of expenditure in the respective line and period. In other words, the zeros reflect the actual absence of budget execution, not problems with the quality of the information. Therefore, we did not impute data to replace the structural zeros.

The grouping of variables reduced the number of series analyzed from 41 to 24 and eliminated 696 zero data points. The new aggregate variables do not have zero values throughout the entire period. Only 4 of the 24 original series show any zero values in a given month[6]. It should be noted that this low proportion of zeros does not characterize a typical scenario of zero-inflated models, which would require specific analysis techniques, according to the literature initially developed by Croston (1972). This parsimonious approach contributes to the robustness of the analysis and avoids distortions due to the inclusion of uninformative variables or unnecessary data processing. Table 1 describes the RTN variables used.

[6] Namely: (i) Salary Bonus and Kandir Law, with 31 counts each (16.8% of data points), (ii) FUNDEB, with 8 counts (4.3%), (iii) court rulings and BPC LOAS/RMV court orders, with 7 counts (3.8%).

Table 1 – Variables used in the models[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

Figure 1 shows the evolution, and Figure 2 shows the seasonality of primary expenditures, both in millions of R$ adjusted by the IPCA until June 2025 and for the period from January 2010 to December 2022, defined as the training set. According to Hyndman et al. (2025), the test set typically represents about 20% of the sample, although this value depends on the sample size and the forecast horizon. Although the base runs until June 2025, all exploratory data analysis is performed on the training period (13 years or 156 points), with the remaining data used as a test (30 points), avoiding *data leakage*[8].

[7] Note: Prepared based on Table 1.2-A Primary Result of the Central Government of the RTN.

[8] *Data leakage* occurs when information from the test set (or future data) is directly or indirectly used in model training, generating artificially superior evaluations. For further information, see (Apicella et al., 2025).

Figure 1 – Evolution of primary expenditures in real terms (millions of R$)

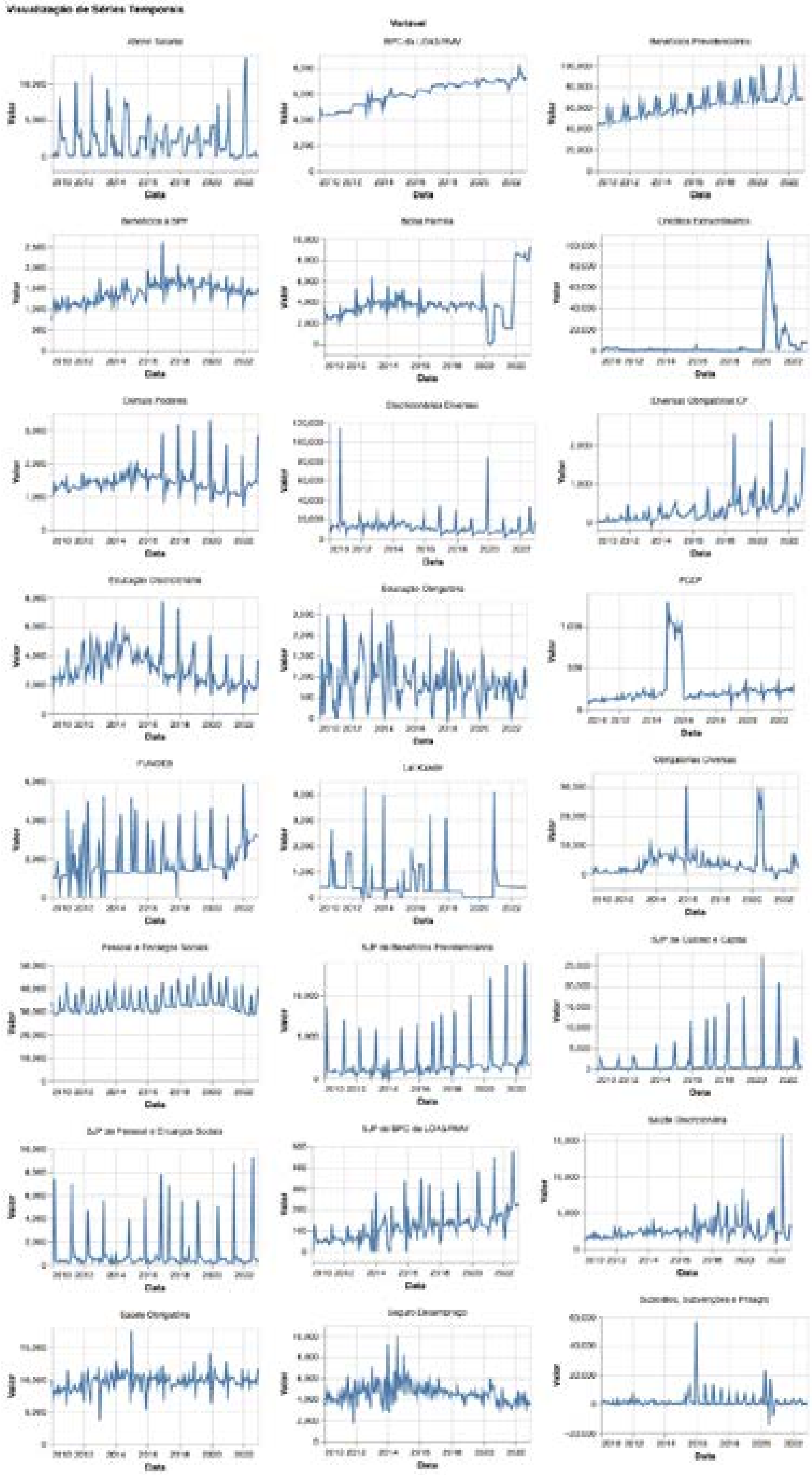
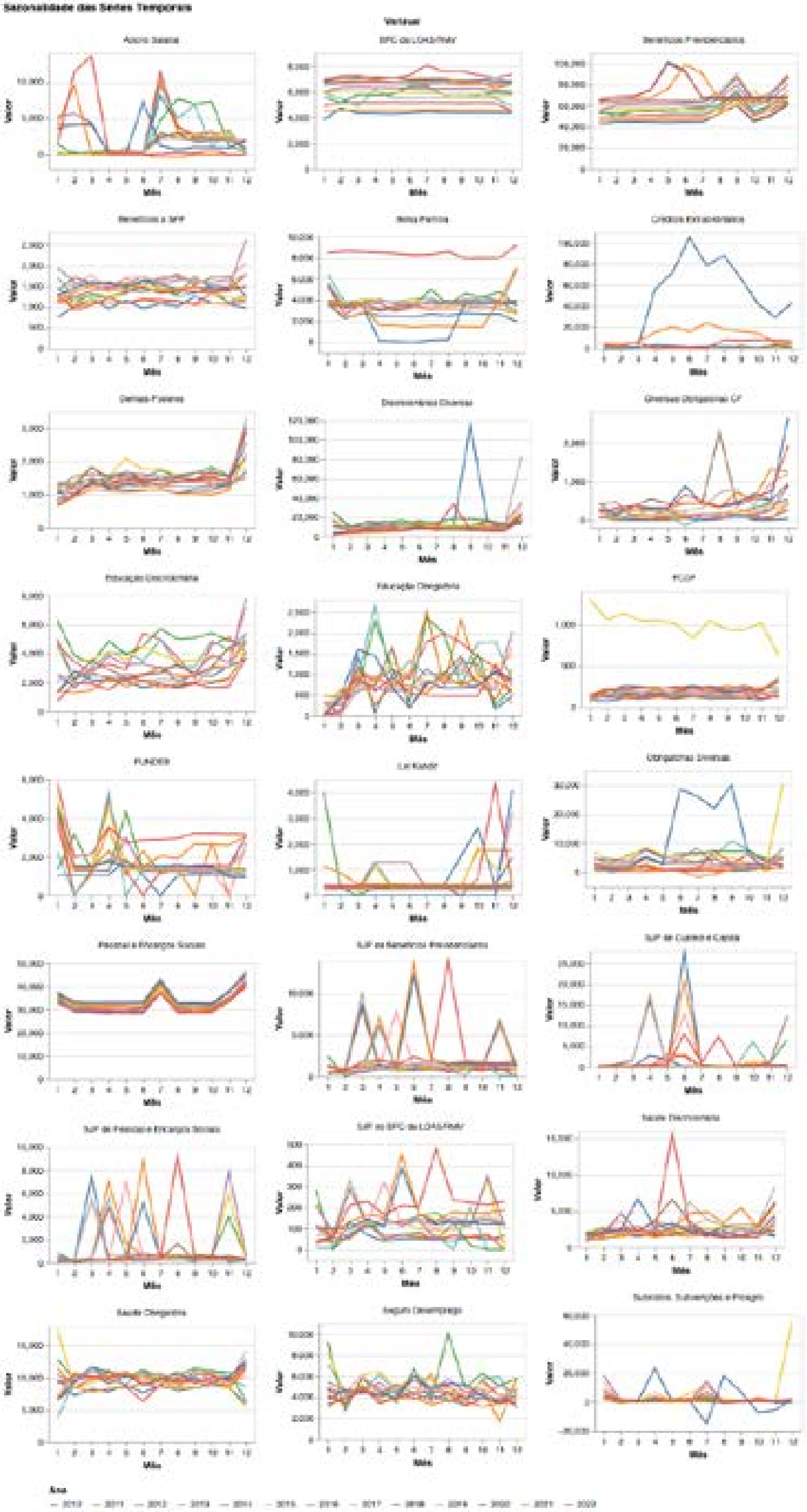


Figure 2 – Seasonality of primary expenditures in real terms (millions of R$)



Based on the figures, we identified that the series exhibit very different behavior, indicating that it is difficult for a single model to represent all series well. This fact reinforces the choice of the *bottom-up* approach. In terms of trends, some series, such as Social Security Benefits and Continuous Cash Benefit (BPC) from LOAS/RMV, show persistent growth, reflecting demographic factors and regulatory changes. Other series, such as Discretionary and Compulsory Education, show a decline, while Health and Personnel and Social Charges show relative stability, suggesting institutional rigidity.

In terms of seasonality, some categories exhibit systematic patterns throughout the months of the year (e.g., Personnel and Social Charges and Social Security Benefits). Pronounced seasonality tends to facilitate forecasting, provided that it is adequately captured by the models and remains constant over time, which is not the case for several of the expenses analyzed. On the other hand, several expenditures exhibit irregular behavior, with peaks in different months, such as Court Judgments and Writs of Payment (SJP), due to the discretionary nature of the payment of writs of payment. The cyclical component, understood as long-term movements associated with economic or institutional fluctuations, is more difficult to isolate, but can be suggested by shocks or changes in level, especially in times of crisis. A notable case is the replacement of Bolsa Família by Emergency Aid during the Covid-19 pandemic.

Finally, we note that irregularity is striking in series such as Salary Bonus, FUNDEB, and Court Rulings and Judicial Payment Orders, with atypical values and trend breaks. Unpredictable behavior, associated with external factors or discretionary decisions, makes forecasting difficult, requiring flexible models capable of dealing with non-linearities and structural breaks. This diversity reinforces the importance of multiple methods and rigorous validations.

We structured the process of evaluating and selecting predictive models in a modular *pipeline*, aligned with best practices in time series forecasting, according to Hyndman et al. (2025). The goal is to select the most appropriate model for each time series, given its individual characteristics, without disregarding the information we can obtain by using multivariate models[9]. The central idea is to identify the most appropriate model for each public expenditure series, taking into account its specific characteristics, while maintaining the robustness that can be obtained by considering different approaches. Since choosing models based solely on intuition is often not confirmed by the data, we adopted a cross-validation procedure with an expan-

[9] According to Hyndman et al. (2025), a multivariate model explicitly models the interactions between multiple time series in a dataset and provides forecasts for multiple time series simultaneously. In contrast, a univariate model trained on multiple time series implicitly models the interactions between multiple time series and provides forecasts for single time series simultaneously. Multivariate models are typically computationally costly and, empirically, do not necessarily offer better forecasting performance compared to using a univariate model.

sive window, which simulates how the models would behave when predicting future periods. Although this method produces larger errors than simple adjustment, it is closer to the actual forecasting situation, in which new information is incorporated over time.

The comparison between models was made using error metrics, mainly the mean absolute error (MAE) and the root mean square error (RMSE). MAE is preferred for its simplicity of interpretation, while RMSE helps capture distortions associated with extreme values. The *pipeline* begins by dividing the data into training and test sets, preserving the temporal order. The training set is used to adjust and validate the models, while the test set is used to evaluate their predictive capacity outside the sample. Each model is automatically adjusted based on the training data, exploring different statistical, *machine learning*, and *deep learning* paradigms.

After fitting, we perform residual diagnosis, that is, the analysis of errors generated by the predictions. Good models should produce uncorrelated and unbiased residuals, which indicates more reliable predictions. When possible, we also evaluate the constancy of variance and normality, although these properties are not indispensable. To construct confidence intervals for the forecasts, we use the conformal approach, which, instead of relying on strong assumptions about probability distributions, uses the history of errors to estimate future uncertainty. This technique ensures that the forecast intervals realistically reflect the degree of uncertainty in the model.

The next step is temporal cross-validation on the training set, which consists of training and evaluating the models on increasing data windows, testing their stability and generalization ability. The model chosen for each series is the one with the lowest average error in this process, using MAE as the main criterion. This model is then adjusted across the entire training set and applied to the test set, allowing its actual predictive power on new data to be measured. Finally, to increase robustness, the final predictions are obtained from the combination of different model paradigms. The literature consistently shows that the simple average of predictions usually outperforms individual models. The uncertainty of the combined predictions is also estimated empirically, again using past errors.

In summary, this process seeks to ensure that model selection is done in a systematic and transparent manner, balancing statistical rigor and practical flexibility. Instead of relying on subjective choices, the performance of various approaches is compared, their errors are evaluated realistically, and forecasts are combined to produce more reliable results.

We trained 46 models for each of the 24 primary expenditures evaluated in this study. Initially, we used six base models, covering historical *average*, naive models (*Naive, Seaso-*

nal Naive, and Random Walk With Drift), and simple and seasonal moving averages (*Window Average and Seasonal Window Average*). These base models served as *benchmarks* for evaluating the performance of the other models. Next, we used six traditional time series forecasting models: *ARIMA, ETS, Theta, Complex Exponential Smoothing (CES), Median, Fourier seasonality, Linear trend, and Exponential Smoothing* (MFLES), and *Trigonometric*, *ARMA errors, Box-Cox transformation, Trend,* and *Seasonality* (TABTS). We also evaluated eight *machine learning* models: *Random Forest, Elastic Net, Lasso, Ridge, Linear Regressor, CatBoost, XGBoost,* and *Light GBM*.

We also explored the efficiency of 26 *deep learning* models, separated into 5 classes of architectures. We evaluated 7 models from the Recurrent Neural Networks class: *Recurrent Neural Networks* (RNN), LSTM, *Gated Recurrent Units* (GRU), *Temporal Convolutional Networks* (TCN), *Deep Autoregressive* (DeepAR), *Dilated Recurrent Neural Network* (DilatedRNN), and *Bidirectional Temporal Convolutional Networks* (BiTCN). We also tested seven models from the Multilayer *Perceptron* class: *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), and *Deep Non-Parametric Time Series* (DeepNPTS).

We also tested six transformer-based models: *Temporal Fusion Transformer* (TFT), *Vanilla Transformer, Informer, Autoformer, Frequency Enhanced Decomposed Transformer* (FEDformer), and *Patch-based Time Series Transformer* (PatchTST). Finally, we evaluated four multivariate models: *Spectral-Temporal Graph Neural Network* (StemGNN), *Time Series Token Mixing* (TSMixer), *Multivariate Multilayer Perceptron* (MLP-Multivariate), and *State-Of-The-Forecast Time Series* (SOFTS), as well as two models with different architectures: *Kolmogorov-Arnold Network* (KAN) and *TimesNet Convolutional Architecture* (TimesNet). Finally, we also analyzed the best models regardless of the previously designated classes.

To test this large number of models, we resorted to the use of automatic hyperparameter optimization, according to the works of Hyndman & Khandakar (2008), Akiba et al. (2019), Garza et al. (2022), and Olivares et al. (2022). Automatic hyperparameter optimization plays a central role in improving the performance of *machine learning* models, especially in contexts where manual search is unfeasible or inefficient. In this sense, Akiba et al. (2019) propose a state-of-the-art approach based on two fundamental principles: dynamic definition of the search space (*define-by-run*) and efficient sampling with Bayesian optimization techniques. The *define-by-run* mechanism allows the hyperparameter space to be constructed programmatically during the execution of the objective function, giving greater flexibility and expressiveness to the conditional definition of parameters. Sampling is performed predominantly using the *Tree-structured Parzen Estimator* (TPE) algorithm, which separately models the distribution of good and bad configurations, prioritizing more promising regions of the search space.

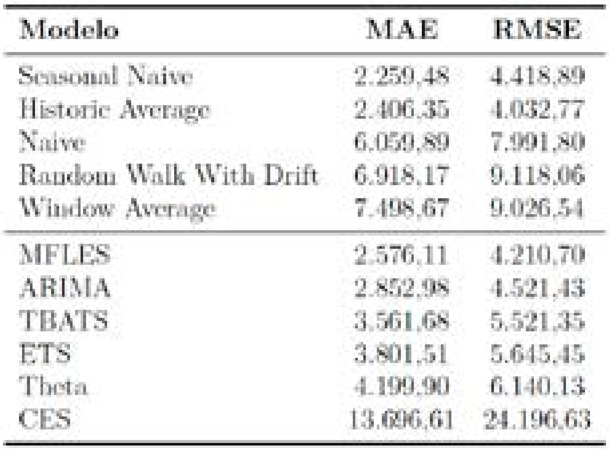
In addition, Akiba et al. (2019) incorporate *pruning* techniques to terminate evaluations with low performance potential early, which significantly reduces the computational cost of optimization. This strategy is based on continuous monitoring of partial metrics during training and is especially useful in high-cost tasks, such as deep neural network tuning. The experiments presented by the authors demonstrate that the proposed model outperforms other popular *frameworks*, both in convergence time and in the quality of the solutions found, even under computational budget constraints. According to the authors, the lightweight architecture and support for parallel and distributed execution compatible with multiple libraries make the tool robust and highly adaptable for real-world predictive modeling applications.

4. RESULTS

In this section, we present and analyze the results obtained from statistical, *machine learning*, and *deep learning* models applied to federal primary expenditure series. We performed the evaluation in three stages: (i) cross-validation performance within the training set; (ii) performance on the test set, outside the sample; and (iii) performance analysis to investigate the occurrence of *overfitting*, *underfitting*, or generalization patterns. We also discuss the comparison between the sets and perform a robustness test by changing the forecast horizon.

Table 2 summarizes the means, among all expenses, of the mean absolute error (MAE) and root mean square error (RMSE) obtained in cross-validation (*expanding window*) for each model evaluated. Among the reference models, *Seasonal Naive* has the lowest average MAE (2,259.48), followed by *Historic Average* (2,406.35). In the group of statistical models, MFLES stands out with the lowest average MAE (2,576.11), followed by ARIMA (2,852.98).

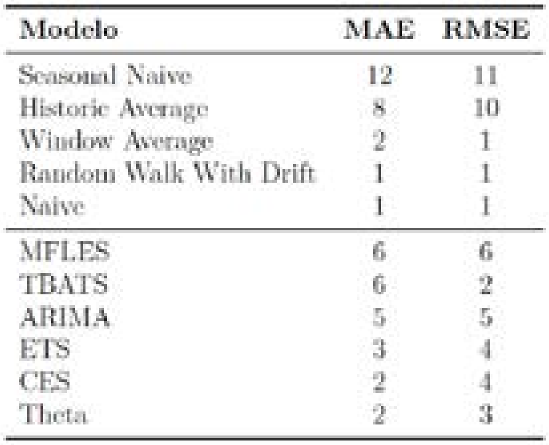
Table 2 – Performance of *benchmark* and statistical models in the training set[10]



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

Table 3 shows the frequency with which each model was selected as the best, i.e., with the lowest error, in each of the expenditure series.

Table 3 – Frequency of selection as the best *benchmark* and statistical model in the training set[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

We observed that, among the reference models, *Seasonal Naive* and *Historic Average* account for most of the selections as the best models for the series evaluated, especially in the MAE criterion. In the group of statistical models, we found greater diversity, with MFLES, TBATS, and ARIMA standing out, but no model dominates all series, indicating that custom adjustment is essential to maximize predictive performance in this context.

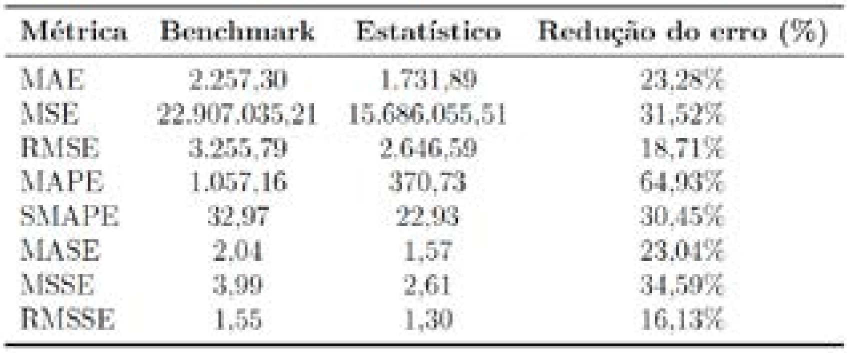
Table 4 summarizes the out-of-sample error metrics (test set) for the forecasts generated by the best models selected for each expense line. We highlight significant reductions in avera-

[10] Note: Results sorted in ascending order by MAE.

[11] Note: Results sorted in descending order by MAE.

ge errors compared to the *benchmark*.

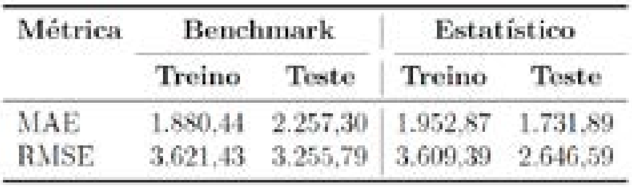
Table 4 – Performance of statistical models compared to the *benchmark* in the test set[12]



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%

Table 5 summarizes the comparison between the errors obtained in cross-validation (training) and in the test. We observe that the statistical models not only outperform the *benchmarks* in all criteria, but also have a lower average error in the test than in the training. This result, although at first glance it may seem counterintuitive, can be explained by three factors: (i) less volatile test samples, with more regular periods; (ii) robust sampling, without *data leakage* and with good series segmentation; and (iii) smaller test sample size, which may, by chance, be less complex than the training set.

Table 5 – Comparison of average errors of statistical models and *benchmarks* in training and test sets



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

In general, the absence of increased errors in the test set indicates that the adjusted models are able to capture stable and generalizable patterns in the evaluated series, with no evidence of *overfitting*. The absolute and relative gain of statistical models over the *benchmark* demonstrates the importance of adopting more sophisticated approaches, even in contexts of short series and budget constraints. In summary, the results show that: (i) statistical models, especially MFLES, ARIMA, and TBATS, outperform reference models in all evaluation criteria; (ii) there is no evidence of *overfitting*, since the test errors remain equal to or lower than those of the

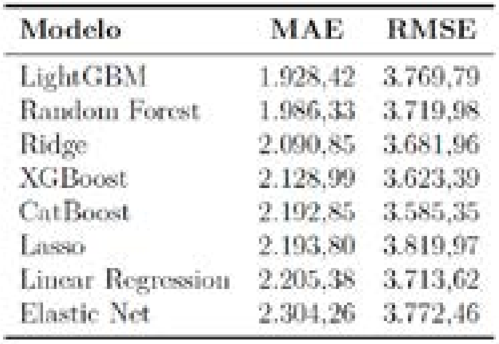
[12] Note: The error reduction column shows the percentage improvement of statistical models relative to the *benchmark* for each metric.

training; and (iii) the methodology adopted ensures robustness and reliability in federal expenditure forecasts, reinforcing the role of well-calibrated statistical models as relevant instruments for fiscal policy.

Additionally, we observed no signs of *underfitting*. In situations of *underfitting*, it would be expected that both training and test errors would remain high, suggesting that the model would be unable to capture relevant structural patterns in the series. However, as the statistical models perform substantially better than the *benchmarks* in both samples, we find that the adopted modeling extracts relevant information from the data, without limiting itself to reproducing only trivial trends. Next, we evaluate the performance of *Machine Learning* models and their nuances in the Brazilian fiscal context.

Table 6 shows the means of the mean absolute error (MAE) and mean square error (RMSE) for the main ML algorithms in cross-validation. We observed that *LightGBM* has the lowest mean MAE, while *CatBoost* has the lowest mean RMSE.

Table 6 – Average performance of *machine learning* models in the training set[13]

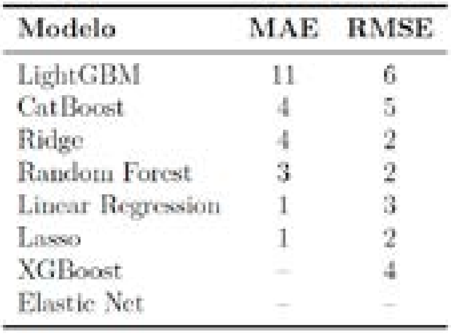


Modelo	MAE	RMSE
LightGBM	1.928,42	3.769,79
Random Forest	1.986,33	3.719,98
Ridge	2.090,85	3.681,96
XGBoost	2.128,99	3.623,39
CatBoost	2.192,85	3.585,35
Lasso	2.193,80	3.819,97
Linear Regression	2.205,38	3.713,62
Elastic Net	2.304,26	3.772,46

Table 7 shows the frequency with which each model is selected as the best for each series (lowest MAE and RMSE). This diversity of selections shows that no approach is universally superior for all series, highlighting the importance of specific choices for each budgetary context and the relevance of the *bottom-up* strategy in forecasting public expenditures.

[13] Note: Results sorted in ascending order by MAE.

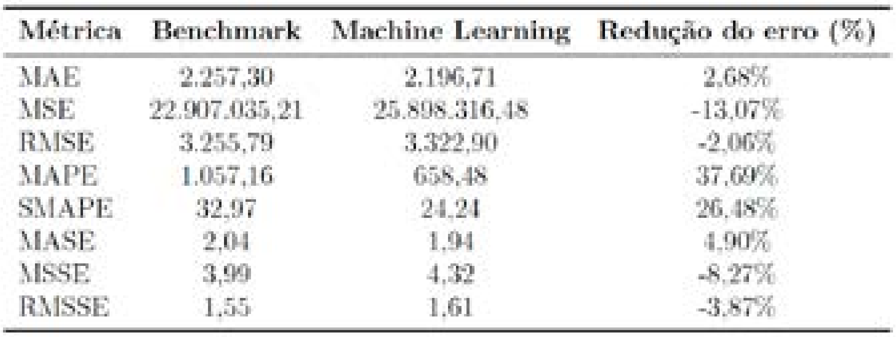
Table 7 – Frequency of selection as the best *Machine Learning* model in the training set[14]



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	–	–

Table 8 summarizes the error metrics of the best ML models in the test set, allowing comparison with the *benchmark*.

Table 8 – Performance of *Machine Learning* models compared to the *benchmark* in the test set[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%

Analysis of the out-of-sample results shows that, although *Machine Learning* models perform slightly better on the MAE criterion and show considerable advances in the percentage metrics (MAPE and SMAPE), we did not identify systematic gains in all the metrics evaluated. In particular, both RMSE and MSE and their standardized variants (MSSE, RMSSE) are slightly above the values observed for the *benchmark*, indicating that ML models, although efficient in predicting the median of absolute deviations, are more sensitive to large errors in some series or atypical events. On the other hand, this less consistent performance highlights the importance of adjusting expectations regarding the use of these techniques in environments characterized by fiscal volatility, frequent institutional changes, and short series.

Thus, our results suggest that the better performance of statistical models over longer

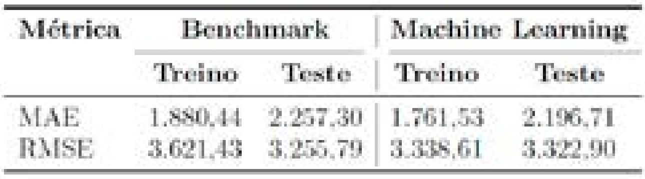
[14] Note: Results sorted in descending order by MAE.

[15] Note: The error reduction column shows the percentage improvement of ML models relative to the *benchmark* for each metric. Negative values indicate a deterioration relative to the *benchmark*.

horizons stems from their parsimonious structure and the inductive bias embedded in classical estimation techniques, which act as natural regularizers against *overfitting*. These models efficiently preserve historical memory and project trends and seasonality in a stable manner, while *machine learning* and *deep learning* architectures, although powerful for capturing local patterns in short horizons, tend to suffer from error propagation and high variance when extended to long horizons. This evidence is consistent with the literature on forecasting competitions discussed in Makridakis et al. (2020) and Godahewa et al. (2021), reinforcing that the methodological choice should consider not only the type of series but also the horizon of interest.

Table 9 summarizes the comparative results of the best ML models between training and testing, considering the main metrics.

Table 9 – Comparison of the average errors of *Machine Learning* models and *benchmarks* in training and test sets



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

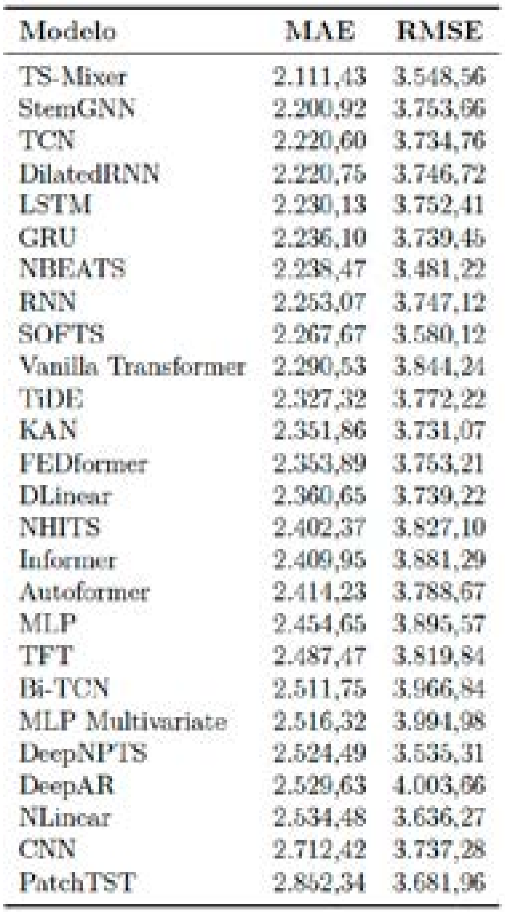
We observe that, for MAE, ML models outperform the *benchmark* in training but suffer a slight increase in testing, which is natural when evaluating predictive ability outside the sample. Even so, the MAE in testing ML models remains competitive and lower than the *benchmark*, demonstrating good generalization. In the case of RMSE, the difference between training and testing remains minimal, suggesting stability in error dispersion, even in scenarios with large deviations.

We did not identify any signs of *overfitting*, as the difference in errors between training and testing is small and there is no explosion of error outside the sample. Similarly, we did not observe any signs of *underfitting*, as the models are able to capture relevant patterns in the series and outperform the *benchmark* in both sets.

Among the *machine learning* algorithms evaluated, the *LightGBM* model stands out for its recurring selection as the best model in multiple series. Still, the model composed of the best algorithms in each series performs better, aligning with the literature on the potential of the *bottom-up* forecasting technique to deal with the structural heterogeneity typical of public expenditure series.

Table 10 presents the means of the mean absolute error (MAE) and mean square error (RMSE) for the main *Deep Learning* models evaluated in the cross-validation of the training set. We observe that TS-Mixer, StemGNN, TCN, and DilatedRNN have the lowest MAE values, while NBEATS, DeepNPTS, and TS-Mixer stand out with the lowest RMSE values. The wide dispersion among architectures indicates a strong dependence on the series structure for the effectiveness of each approach.

Table 10 – Average performance of *Deep Learning* models in the training set[16]

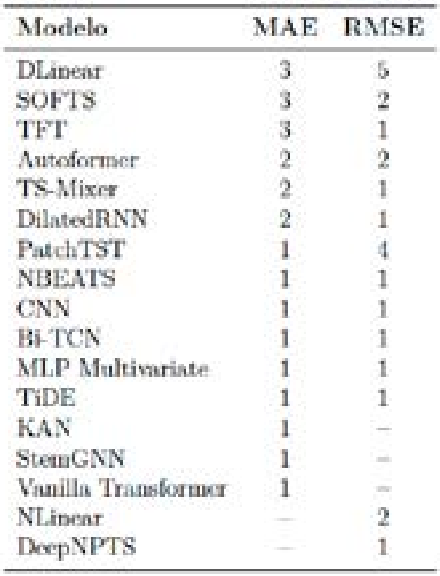


Modelo	MAE	RMSE
TS-Mixer	2.111,43	3.548,56
StemGNN	2.200,92	3.753,66
TCN	2.220,60	3.734,76
DilatedRNN	2.220,75	3.746,72
LSTM	2.230,13	3.752,41
GRU	2.236,10	3.739,45
NBEATS	2.238,47	3.481,22
RNN	2.253,07	3.747,12
SOFTS	2.267,67	3.580,12
Vanilla Transformer	2.290,53	3.844,24
TiDE	2.327,32	3.772,22
KAN	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

Table 11 shows the frequency with which each model is selected as the best for each series (lowest MAE and RMSE). We observed greater dispersion among *Deep Learning* models, suggesting that the best performance is quite sensitive to the type of architecture, series, and hyperparameter tuning.

[16] Note: Results sorted in ascending order by MAE.

Table 11 – Frequency of selection as the best *Deep Learning* model in the training set[17]



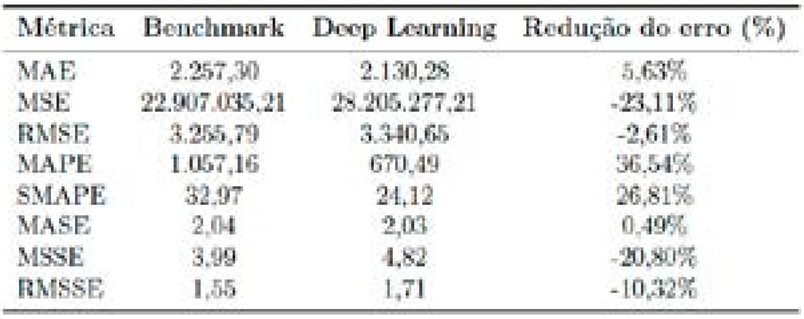
Modelo	MAE	RMSE
DLinear	3	5
SOFTS	3	2
TFT	3	1
Autoformer	2	2
TS-Mixer	2	1
DilatedRNN	2	1
PatchTST	1	4
NBEATS	1	1
CNN	1	1
Bi-TCN	1	1
MLP Multivariate	1	1
TiDE	1	1
KAN	1	–
StemGNN	1	–
Vanilla Transformer	1	–
NLinear	–	2
DeepNPTS	–	1

Overall, the results show that, although models such as TS-Mixer, StemGNN, TCN, and DLinear lead in terms of absolute and mean square error, no single architecture dominates across all series. This suggests that individualized selection by series, accompanied by automatic hyperparameter tuning, is essential to extract the best performance from *Deep Learning* models in the context of federal expenditures.

Table 12 summarizes the error metrics of the best DL models in the test set, allowing comparison with the *benchmark*. The results of the test set indicate that *Deep Learning* models do not show robust gains over the *benchmark*, especially in metrics that are more sensitive to large deviations. Although MAE, MAPE, and SMAPE show error reductions compared to the *benchmark*, suggesting a slight advantage of DL models in predicting average and percentage deviations, metrics based on residual squares (MSE, RMSE, MSSE, RMSSE) perform worse than the reference models.

[17] Note: Results sorted in descending order by MAE.

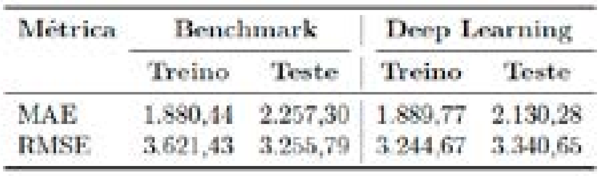
Table 12 – Performance of *Deep Learning* models compared to the *benchmark* in the test set[18]



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%

These results indicate that, although *Deep Learning* models are able to capture average or recurring patterns in the expenditure series, they are more vulnerable to large errors at specific points, especially in contexts marked by shocks, atypical seasonality, or abrupt changes in the fiscal regime. The slight reduction in MAE and MASE, accompanied by an increase in RMSE and related metrics, reinforces the hypothesis that these models may be sensitive to *outliers* or unusual events, especially in relatively short and heterogeneous test samples. Table 13 summarizes the comparative results of the best DL models between training and testing, considering the main metrics.

Table 13 – Comparison of the average errors of *Deep Learning* and *benchmark* models in the training and test sets



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

In the training set, *Deep Learning* models perform similarly to the *benchmark* in terms of MAE, with virtually equal values, and superior performance in terms of RMSE. In the test set, we observe that the MAE of the *Deep Learning* model remains lower than the *benchmark*, indicating greater predictive accuracy outside the sample. The RMSE of *Deep Learning* models is slightly above the *benchmark*, suggesting that, despite lower MAE, there were some larger deviations in the prediction of certain series.

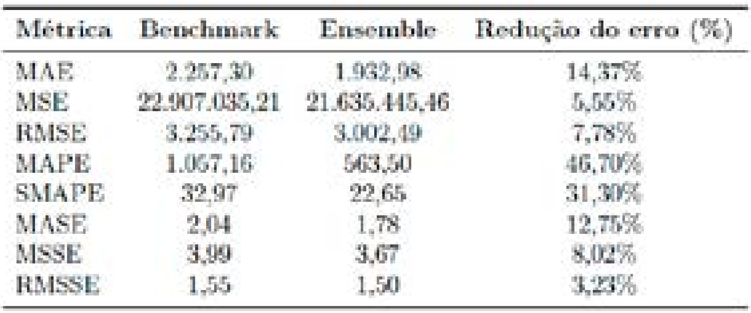
The difference between training and testing for *Deep Learning* models is modest. The

[18] Note: The error reduction column shows the percentage improvement of *Deep Learning* models relative to the *benchmark* for each metric. Negative values indicate a deterioration relative to the *benchmark*.

increase in MAE is expected in out-of-sample predictions and is in line with that observed for statistical and *Machine Learning* models. This suggests that there are no clear signs of *overfitting*, since the error does not grow abruptly, and the model maintains competitive performance on unseen data. Nor do we observe evidence of *underfitting*, since the errors do not remain high in both sets. In summary, we observe that *Deep Learning* models demonstrate generalization capacity, managing to exceed the *benchmark* in terms of absolute accuracy in the test, even though they present a slight disadvantage in RMSE, a natural reflection of this indicator's greater sensitivity to extreme values.

On the other hand, the less robust performance of *Machine Learning* and *Deep Learning* models in the test reinforces that, in short and highly heterogeneous series, well-calibrated classical methods can maintain a relevant advantage over modern alternatives, whose full effectiveness depends on larger samples or those less subject to shocks. Table 14 shows the performance of the combination of forecasts in the test set, comparing it to the naive *benchmark*.

Table 14 – Performance of the combination of forecasts compared to the *benchmark* in the test set[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%

We observe that the forecast combination strategy outperforms the *benchmark* in all metrics evaluated, with particularly significant gains in MAPE (46.7%) and SMAPE (31.3%), in addition to relevant reductions in MAE (14.4%) and RMSE (7.8%). These gains confirm the potential of the combination to promote robustness and stability in forecasts, diluting specific errors in individual models.

When comparing the performance of the combination with the best statistical models, we find that the combination outperforms the statistics only in SMAPE, and by a small margin. In the other metrics, the statistical models maintain superior performance, reflecting strong adherence to the pattern of the series evaluated. Therefore, although forecast combination is recom-

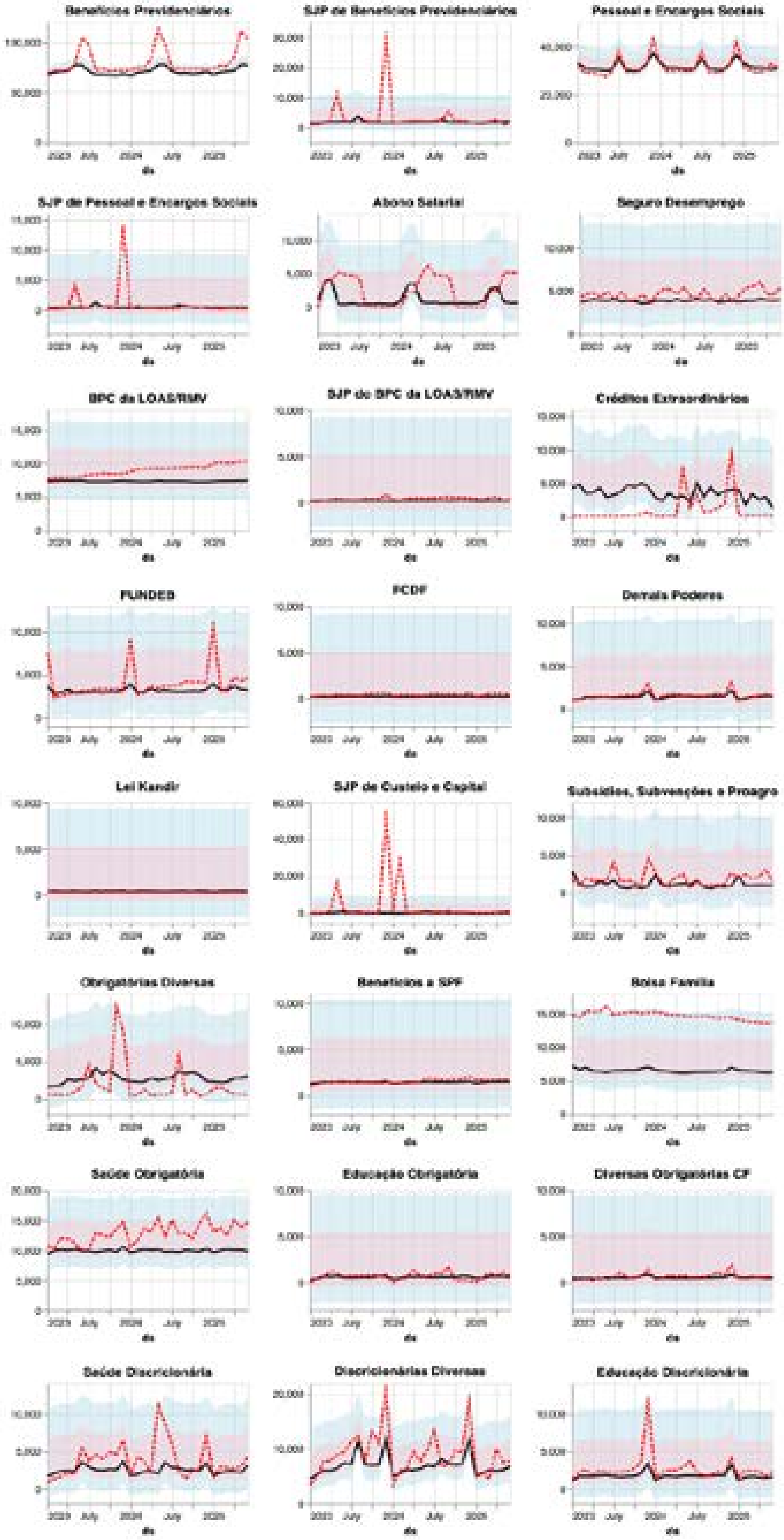
[19] Note: The error reduction column shows the percentage improvement of the combination of predictions relative to the *benchmark* for each metric.

mended as a robustness strategy, especially in highly heterogeneous scenarios, its effectiveness may be limited when a family of models is already clearly dominant. Still, combination adds value by avoiding dependence on a single model and increasing the resilience of the predictive system in the face of uncertain scenarios.

Our results also suggest that this observed pattern, with statistical models excelling in long horizons, neural networks in short horizons, and the combination showing more stable performance, reflects not only the heterogeneity of the series but also the role of memory and inductive bias embedded in classical estimation techniques. Statistical models tend to err on the side of caution, which preserves their advantage over long horizons. The combination, on the other hand, acts as a diversification mechanism, balancing bias and variance: it rarely leads in absolute performance, but it is rarely the worst option, ensuring consistent forecasts even in the face of shocks.

Figure 3 shows the combination of forecasts for the 24 primary expenditures in real terms (R$ million) for the 30-month horizon, including the 80% and 95% confidence intervals, in addition to the actual values.

Figure 3 – Combination of forecasts (horizon = 30 months)[20]

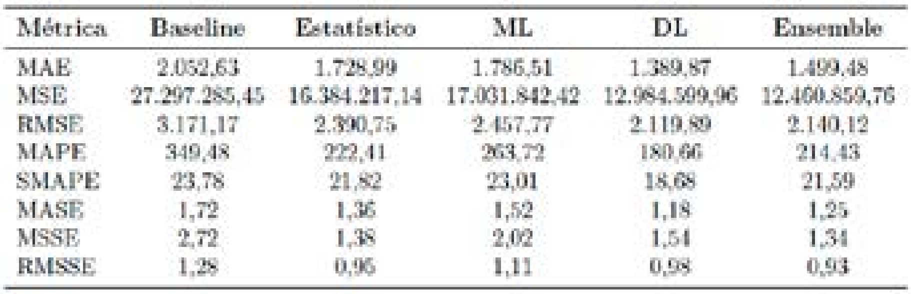


[20] Note: Red line: actual; black line: forecast; blue area: 95% CI; and red area: 80% CI.

To assess the robustness of the results in different configurations, we performed an alternative test by reducing the forecast horizon. We extended the training set to December 2023 and made forecasts for the period from January 2024 to June 2025, totaling 18 months (instead of the 30 months in the main scenario).

Table 15 presents the error metrics for each model class evaluated in the new forecast horizon. The values refer to the average performance of the best forecasts for each expenditure line.

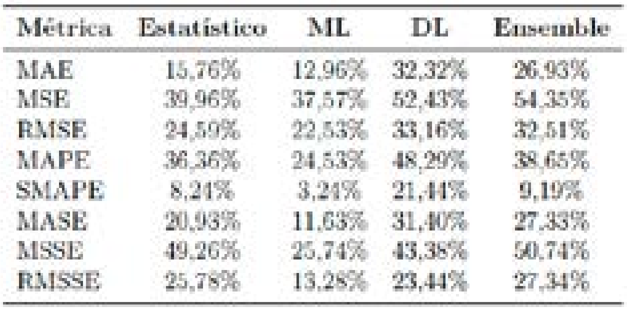
Table 15 – Model performance in the robustness test (18-month horizon)[21]



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

Table 16 shows the percentage reduction in errors for each model relative to the baseline.

Table 16 – Percentage reduction in errors relative to the baseline (robustness test)[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	8,24%	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%

The results of the robustness test reinforce the superiority of advanced models (statistical, ML, DL, and *ensemble*) in relation to the simple baseline, even when we reduce the forecast horizon from 30 to 18 months. We observe that all models maintain superior performance, with emphasis on *Deep Learning* and *ensemble* models, which present the largest relative reductions in error in practically all metrics analyzed. The statistical and ML models preserve consistency and robustness, while the combination of forecasts proves particularly effective in reducing

[21] Note: Results for the test set from January 2024 to June 2025.

[22] Note: Calculation: 1 - (model error / baseline error).

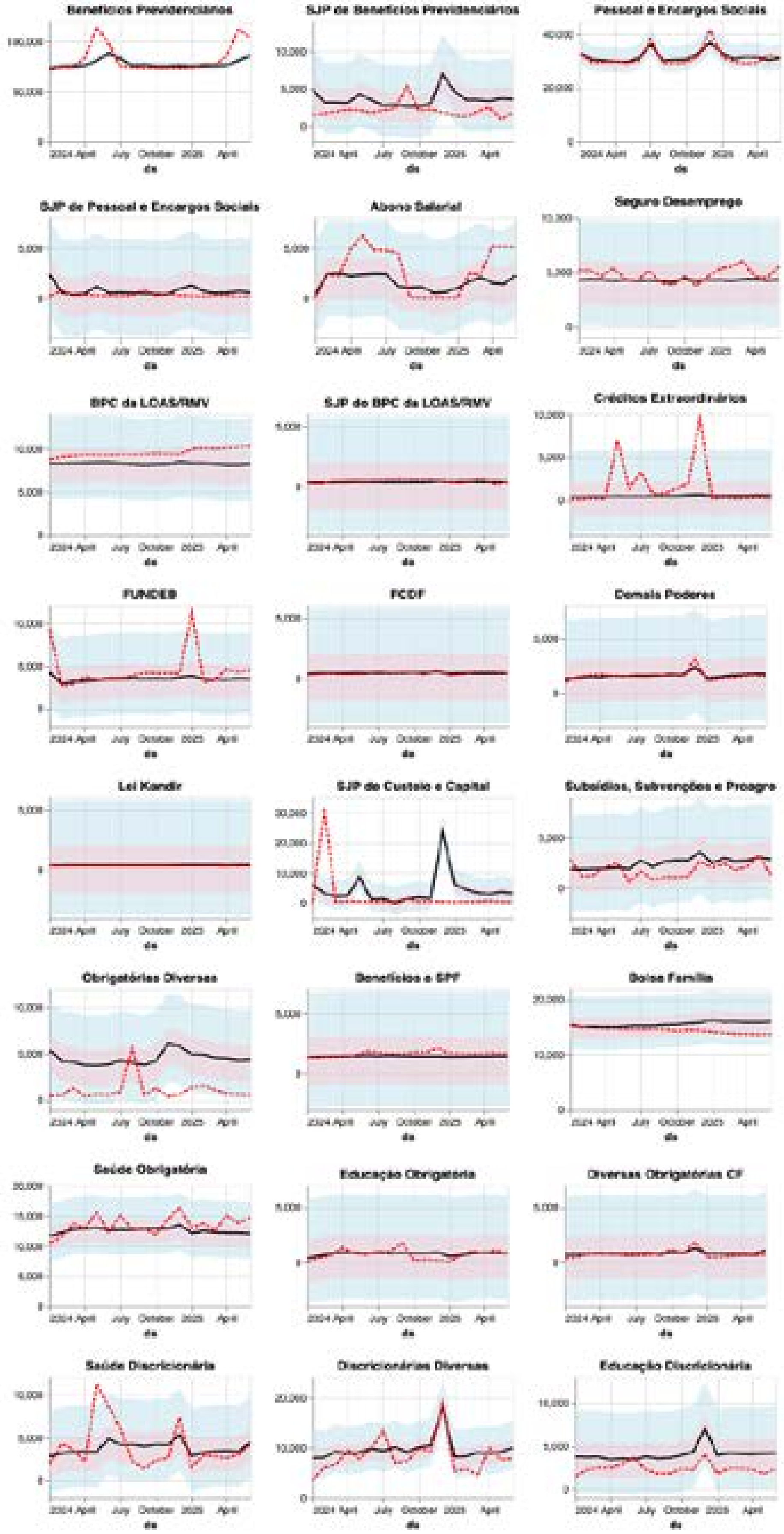
MSE, MSSE, and RMSSE. We also note that the gains of DL over the baseline are remarkable, suggesting better adaptation to recent shocks or pattern changes.

We also observe that, in shorter horizons, the gains of sophisticated models over the *benchmark* tend to be more significant in absolute metrics, although the differences between classes (statistical, ML, DL, *ensemble*) are less pronounced. This indicates that, in more predictable scenarios, the relative advantages of advanced methods persist, but the *benchmark* also benefits from lower uncertainty.

In summary, the results confirm that the judicious use of statistical, *machine learning*, and *deep learning* models and their combinations represents a promising strategy for improving Brazilian fiscal forecasting, provided that we respect the specificities of the context, the limitations of the data, and the requirements for robust validation. Classic methods remain highly competitive, but the combination of paradigms ensures greater resilience in the face of the uncertainty and heterogeneity inherent in public budget management.

Figure 4 shows the combination of forecasts for the 24 primary expenditures in real terms (millions of R$) for the 18-month horizon, including the 80% and 95% confidence intervals, in addition to the actual values.

Figure 4 – Combination of forecasts (horizon = 18 months)[23]



[23] Note: Red line: actual; black line: predicted; blue area: 95% CI; and red area: 80% CI.

5. DISCUSSION

The results obtained in this study reaffirm the complexity of forecasting public expenditures in contexts marked by high uncertainty, regime changes, and relevant exogenous shocks, such as the Covid-19 pandemic and the federal government transition that occurred in 2023. These events impact both the behavior of historical series and the effectiveness of the different modeling paradigms tested.

We found that classical statistical models consistently outperform *Machine Learning* and *Deep Learning* alternatives over a 30-month horizon. This shows that well-calibrated methods adapted to the data structure maintain robust performance even in scenarios of high volatility and small sample size. This result may be associated with the greater ability of statistical models to capture seasonal patterns and persistent structural trends, in addition to their lower risk of *overfitting* in short and noisy series.

On the other hand, when we reduce the forecast horizon to 18 months, focusing the analysis on a more recent period after the pandemic shock, we observe a partial reversal of this trend. In this configuration, *Deep Learning* models perform better on several relevant metrics, while statistical models remain robust, but without the same advantage observed in longer horizons. This result confirms the potential of *Deep Learning* to capture complex and dynamic patterns in shorter series, provided that the forecasting context is less affected by atypical shocks and the architectures are properly adjusted to the available data.

The comparative analysis shows that *machine learning* models, although competitive in the training set, tend to lose performance outside the sample, especially in the face of extreme events or sudden changes in the fiscal regime. This limitation highlights the challenge of generalizing these approaches in heterogeneous and volatile environments, reinforcing the need for rigorous validation and careful selection of hyperparameters.

The variation in results according to the horizon and historical window confirms that the relative performance of each class of models depends heavily on the institutional context, the presence of structural shocks, and the volume of data available for adjustment. The Covid-19 pandemic, by generating discontinuities and jumps in expenditure series, and the change of government, by altering priorities and dynamics of public policies, increase uncertainty and make it difficult to model trends, requiring greater adaptability of predictive methods.

Given this scenario, we find that the strategy of combining forecasts is particularly relevant. As suggested by Hyndman et al. (2025), combining forecasts allows for balancing the specific limitations and advantages of each approach, promoting more stable forecasts that are resilient to shocks and less susceptible to modeling biases. Although, in this study, the *ensemble* does not consistently outperform the best individual model in almost any forecast horizon (with the exception of SMAPE in the 30-month horizon and MSE, MSSE, and RMSSE in the 18-month horizon), we found that its balanced performance in different configurations and time windows indicates that methodological diversification is a prudent response to fiscal uncertainty.

We therefore conclude that there is no single or universal solution for forecasting public expenditures in environments subject to frequent changes and significant shocks. The choice of the optimal paradigm should consider not only absolute performance in traditional metrics, but also robustness in unstable scenarios, the ability to adapt to pattern changes, and the feasibility of practical implementation in institutional environments. The integration of methods, associated with continuous validation processes, is the main recommendation for managers and researchers seeking to improve the quality of fiscal forecasts in Brazil.

Additionally, we recognize some limitations that are worth noting. The main one concerns the availability and granularity of public expenditure series: although we worked with monthly data disaggregated by category, the absence of more detailed information on programs and sub-functions, as well as the unavailability of longer historical series, may have restricted the adjustment potential of some models, especially the most data-intensive ones. We also did not incorporate explanatory variables, high-frequency data, or leading indicators that could enrich the modeling. These limitations open up space for future research. Subsequent investigations may explore the incorporation of exogenous macroeconomic and sectoral variables, the use of administrative databases with higher frequency and granularity, and the integration of structural and semi-structural models with *Machine Learning* and *Deep Learning* techniques in hybrid configurations.

From a public policy perspective, our findings suggest that methodological diversification, with the combined use of statistical, *machine learning*, and *deep learning* approaches, increases the resilience of fiscal forecasts in the face of shocks and regime changes. The adoption of continuous validation and dynamic model selection processes contributes to reducing biases, improving transparency, and strengthening the credibility of public accounts.

6. CONCLUSION

The results of this study show that the relative performance of different predictive modeling paradigms is sensitive to the forecast horizon and the occurrence of relevant shocks and institutional changes, such as the Covid-19 pandemic and changes in the conduct of fiscal policy. We observe that, in shorter forecast horizons and more recent periods, *Deep Learning* models stand out, while traditional statistical models remain superior in broader contexts with greater historical variability. This variability in performance reinforces the importance of adopting forecast combination techniques, which promote greater robustness and balance, reduce the risk of dependence on a single model, and increase the resilience of fiscal projections in different scenarios.

We conclude that the choice of the most appropriate predictive approach should consider the characteristics of the problem, the temporal context, and the possibility of disruptive events. Regardless of the scenario, we find that advanced forecasting models are valuable tools for dealing with the uncertainty inherent in forecasting public expenditures, offering reliable forecasts to support fiscal policy planning. We suggest that future research explore exogenous variables and structural or semi-structural models in order to increase the accuracy, robustness, and interpretability of predictive models.

BIBLIOGRAPHICAL REFERENCES

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.