CIRANO
Knowledge into action

PANEL MACHINE LEARNING WITH MIXED-FREQUENCY DATA: MONITORING STATE-LEVEL FISCAL VARIABLES

PHILIPPE GOULET COULOMBE
MASSIMILIANO MARCELLINO
DALIBOR STEVANOVIC

C
2025s-15
WORKING PAPER

The purpose of the **Working Papers** is to disseminate the results of research conducted by CIRANO research members in order to solicit exchanges and comments. These reports are written in the style of scientific publications. The ideas and opinions expressed in these documents are solely those of the authors.

Les cahiers de la série scientifique visent à rendre accessibles les résultats des recherches effectuées par des chercheurs membres du CIRANO afin de susciter échanges et commentaires. Ces cahiers sont rédigés dans le style des publications scientifiques et n'engagent que leurs auteurs.

CIRANO is a private non-profit organization incorporated under the Quebec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the government of Quebec, and grants and research mandates obtained by its research teams.

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du gouvernement du Québec, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO Partners – *Les partenaires du CIRANO*

Corporate Partners – Partenaires Corporatifs	**Governmental partners - Partenaires gouvernementaux**	**University Partners – Partenaires universitaires**
Autorité des marchés financiers	Ministère des Finances du Québec	École de technologie supérieure
Banque de développement du Canada	Ministère de l'Économie, de l'Innovation et de l'Énergie	École nationale d'administration publique
Banque du Canada	Innovation, Sciences et Développement Économique Canada	de Montréal
Banque Nationale du Canada	Ville de Montréal	HEC Montreal
Bell Canada		Institut national de la recherche scientifique
BMO Groupe financier		Polytechnique Montréal
Caisse de dépôt et placement du Québec		Université Concordia
Énergir		Université de Montréal
Hydro-Québec		Université de Sherbrooke
Intact Corporation Financière		Université du Québec
Investissements PSP		Université du Québec à Montréal
Manuvie		Université Laval
Mouvement Desjardins		Université McGill
Power Corporation du Canada		
Pratt & Whitney Canada		
VIA Rail Canada		

CIRANO collaborates with many centers and university research chairs; list available on its website. *Le CIRANO collabore avec de nombreux centres et chaires de recherche universitaires dont on peut consulter la liste sur son site web.*

ISSN 2292-0838 (online version)

Panel Machine Learning with Mixed-Frequency Data: Monitoring State-Level Fiscal Variables

Philippe Goulet Coulombe[*], *Massimiliano Marcellino*[†], *Dalibor Stevanovic*[‡]

Abstract/Résumé

We study the nowcasting of U.S. state-level fiscal variables using machine learning (ML) models and mixed-frequency predictors within a panel framework. Neural networks with continuous and categorical embeddings consistently outperform both linear and nonlinear alternatives, especially when combined with pooled panel structures. These architectures flexibly capture differences across states while benefiting from shared patterns in the panel structure. Forecast gains are especially large for volatile variables like expenditures and deficits. Pooling enhances forecast stability, and ML models are better suited to handle cross-sectional nonlinearities. Results show that predictive improvements are broad-based and that even a few high frequency state indicators contribute substantially to forecast accuracy. Our findings highlight the complementarity between flexible modeling and cross-sectional pooling, making panel neural networks a powerful tool for timely and accurate fiscal monitoring in heterogeneous settings.

Nous étudions le nowcasting des variables budgétaires des États américains à l'aide de modèles d'apprentissage automatique (machine learning) et de prédicteurs à fréquence mixte, dans un cadre en panel. Les réseaux de neurones intégrant des variables continues et des identifiants catégoriels surpassent systématiquement les alternatives linéaires, en particulier lorsqu'ils sont combinés à des structures en panel mutualisé. Ces architectures permettent de capter les différences entre les États tout en tirant parti des régularités partagées. Les gains de prévision sont particulièrement importants pour les variables volatiles comme les dépenses et les déficits. Le regroupement des données améliore la stabilité des prévisions, et les modèles d'apprentissage automatique sont mieux adaptés pour traiter les non-linéarités transversales. Les résultats montrent que les améliorations prédictives sont généralisées et que même quelques indicateurs infranuels spécifiques aux États contribuent de manière significative à la précision des prévisions. Nos résultats soulignent la complémentarité entre la modélisation flexible et le regroupement transversal, faisant des réseaux de neurones en panel un outil puissant pour un suivi budgétaire rapide et précis dans des contextes hétérogènes

Keywords/Mots-clés: Machine learning, Nowcasting, Panel, Mixed-frequency, Fiscal indicators / Apprentissage automatique, Nowcasting, Panel, Fréquences mixtes, Indicateurs budgétaires.

[*] Université du Québec à Montréal
[†] Bocconi University, IGIER, Baffi and CEPR
[‡] Université du Québec à Montréal and CIRANO

JEL Codes/Codes JEL: C53, C55, E37, H72.

Pour citer ce document / To quote this document
Goulet Coulombe, P., Marcellino, M. , & Stevanovic, D. (2025). Panel Machine Learning with Mixed-Frequency Data: Monitoring State-Level Fiscal Variables (2025s-15, Cahiers scientifiques, CIRANO.) https://doi.org/10.54932/QGJA3449

1 Introduction

Fiscal forecasts play a critical role in macroeconomic policymaking, particularly following the expansive fiscal measures implemented during the Covid pandemic, which led to a massive increase in public deficits and debt at both the federal and state levels in the United States and many other countries. A variety of econometric methods are available for fiscal forecasting, ranging from univariate and multivariate time series models to semi-structural models incorporating accounting identities and fully-fledged structural models; see, for example, Favero and Marcellino (2005) and the survey in Leal et al. (2008). A common feature across much of the literature is the focus on national or federal fiscal variables, whereas regional and state public finances are also critical, especially when key spending and taxation decisions—such as those related to health and education—are decentralized.

This paper contributes to the fiscal forecasting literature along three main dimensions. First, we focus on U.S. state-level fiscal variables, such as revenues, expenditures, and deficits (see, e.g., Ghysels et al. (2022)). Second, we exploit information available at a higher frequency than the fiscal variables themselves (see, e.g., Onorante et al. (2010), Asimakopoulos et al. (2020)). Third, and most importantly, we employ novel machine learning (ML) techniques that have been shown to often outperform standard econometric models, particularly during crises and in the presence of large information sets (see, e.g., Goulet Coulombe et al. (2021b, 2022); Hauzenberger et al. (2024)). While previous work has considered state-level and mixed-frequency data separately (Ghysels et al., 2022), to the best of our knowledge, integrating them into a unified ML-based framework that includes nonlinear models and categorical embeddings is new.

More specifically, we forecast state revenues, expenditures and deficits using data from the Annual Survey of State & Local Government Finances conducted by the U.S. Census Bureau. For high-frequency predictors, we use a combination of a large number of national indicators and three state-specific economic indicators as in Ghysels et al. (2022): the growth rates of unemployment, personal income, and the coincident economic activity index. We conduct a comprehensive out-of-sample evaluation from 2000 to 2020, forecasting four fiscal variables across 48 U.S. states. Several classes of models are considered: benchmark unrestricted MIDAS

regressions, linear ML models (Ridge, Lasso, Sparse-Group Lasso), and nonlinear ML methods (Random Forests, Boosted Trees, Neural Networks).

Another novelty is the combination of panel and ML methods, resulting in what can be labeled a panel-ML approach, to assess whether explicitly accounting for the panel structure of the state-level fiscal data enhances the forecasting performance of ML models. For each model, we estimate variants using different panel structures, including no pooling, full pooling, and clustered pooling schemes based on geography or economic similarity. Previously, Khalaf et al. (2021) and Fosten and Greenaway-McGrevy (2022) introduced mixed-frequency data into a panel framework, while Babii et al. (2024) addressed the nowcasting problem using penalized linear regressions in a mixed-frequency context. We extend these approaches by incorporating nonlinear machine learning techniques.

The results show that combining machine learning with appropriate panel structures yields substantial predictive gains. Among all approaches, neural networks with continuous and categorical embeddings consistently perform best. These models are particularly well suited to capturing fixed effects and unobserved heterogeneity across states. When paired with global or hierarchical pooling schemes, they dominate both UMIDAS regressions and linear ML models in terms of forecast accuracy. Their advantage is marked for volatile fiscal variables like expenditures and deficits, where nonlinearities and cross-sectional interactions matter most.

In particular, results show that the main drivers of predictive performance are model flexibility and the ability to pool information efficiently. Neural networks with embeddings offer a structured way to incorporate categorical panel information while flexibly modeling nonlinear dynamics. Panel pooling further enhances predictive performance by reducing estimation variance and exploiting common patterns across states. These two features—nonlinear learning and pooling—are complementary, not substitutes, and their joint use explains the superior performance of panel neural networks.

Further detailed analysis provides evidence that the improvements in forecast quality are broad-based and not concentrated in a few states. The analysis also shows that the limited number of state-level indicators carries substantial predictive content—particularly for expen-

ditures—when combined with powerful ML techniques and panel structures. In contrast, purely individual-level models or those ignoring the panel nature of the data tend to suffer from instability and misspecification.

Overall, our findings support the use of panel neural networks with embeddings as a powerful tool for monitoring and nowcasting subnational fiscal outcomes. By integrating mixed-frequency information, flexible nonlinear modeling, and panel pooling, the panel-ML framework offers a scalable and effective approach for fiscal surveillance and short-term policy analysis at the state level.

The paper is structured as follows. Section 2 presents the data, forecasting models, and evaluation procedures. Section 3 reports the aggregate forecasting results, including by fiscal variable and model type, and presents a disaggregated analysis by state. Section 4 analyzes the sources of predictive performance, quantifying the marginal and joint contributions of model specification and panel structure. Section 5 concludes.

2 Data and Forecasting Framework

In this section we discuss the data we use in the empirical analysis, how we handle mixed frequency indicators, the forecasting models, and the set-up of the forecast evaluation exercise.

2.1 Data

State revenue and expenditure data are from the Annual Survey of State & Local Government Finances conducted by the U.S. Census Bureau. This survey provides detailed information on revenue and expenditure sources for each state and its local government. Data have been collected annually since 1957 and cover all 50 state governments in the United States. A census is conducted every five years (years ending in '2' and '7'), while in other years a subsample of state and local governments is used. To ensure the reliability and completeness of our analysis, we focus our attention on data collected from 1958 to 2020. As Ghysels et al. (2022) pointed out, we chose this time-frame because there are substantial missing observations for state-year

combinations prior to 1958. Due to limited fiscal data availability during the sample period, we had to exclude Alaska and Hawaii from our analysis, which leaves comprehensive data for 48 states. To make the data comparable across states and over time, we adjust total revenues and expenditures based on the population of each state, as documented by the U.S. Census Bureau. Additionally, we take into account changes in the price level by deflating each variable using the consumer price index (CPI) developed by the U.S. Bureau of Labor Statistics. Hence, in the end, our target variables are the revenues and expenditures year-over-year real per capita growths, the deficit per capita, which is obtained as the difference between revenues and expenditures, divided by population, as well as the deficit in percentage of GDP.

Regarding the high(er) frequency (HF) indicators used for monitoring the infra-annual evolution of the fiscal variables, we drew inspiration from the work of Ghysels et al. (2022). Our HF predictors include three state-specific economic indicators available at the quarterly frequency: the unemployment growth rate, personal income, and the coincident economic activity index. These indicators provide insights into the economic conditions at the state level. We consider the ten quarterly US series that have been used in their analysis: 3-Month Treasury Bill: Secondary Market Rate, 10-Year Treasury Constant Maturity Rate, Effective Federal Funds Rate, S&P500 Returns, Default Spread (Moodys BAA - AAA yields), Spot Crude Oil Price (WTI), Industrial Production Index, CPI, Federal Government Budget Surplus and Real GDP.

Moreover, in our analysis, we incorporate many more quarterly US macroeconomic and financial data, which can provide additional information for intra-annual monitoring the state fiscal conditions. The data source is the FRED-QD dataset, which can be accessed from the Federal Reserve of St. Louis' website. This dataset includes 243 macroeconomic and financial indicators, providing a comprehensive view of the United States' economic landscape. Many of these indicators exhibit high persistence or non-stationarity, so we adopt the approach outlined by McCracken and Ng (2020) to transform them and ensure stationarity.

2.2 Handling Mixed-Frequency

Let us define $t = 1, ..., T$ as the low frequency (LF) time unit and $t_m = 1, ..., T_m$ as the high frequency (HF) time unit. The HF time unit is observed m times in the LF time unit. Let us assume LF is annual and HF is quarterly, hence $m = 4$. In addition, L indicates the lag operator at t_m frequency, while L^m is the lag operator at t frequency. Let us then define y_t as the stationary low frequency target variable and x_t as the high frequency stationary exogenous predictor, so that x is observable for every period t_m, while y is observable only every m periods. Using this notation, as in Foroni et al. (2019), the models take the following general form:

$$y_{t_m} = \rho(L^m) y_{t_m - h_m} + \delta(L) x_{t_m - h_m + w} + u_{t_m}, \tag{1}$$

where $t_m = m, 2m, 3m, \ldots, T_m$, h_m is the forecast horizon, w is the number of quarters with which x is leading y, and the error term u_{t_m} is white noise with $\mathrm{E}(u_{t_m}) = 0$ and $\mathrm{E}(u_{t_m}^2) = \sigma_u^2 < \infty$. Thus, with $m = 4$, we have $t_m = 4, 8, 12 \ldots$

The model in (1) is known as unrestricted mixed data sampling regression (UMIDAS, see Foroni et al. (2015)). The restricted version of the UMIDAS in (1), MIDAS, see Ghysels et al. (2006) , is obtained by imposing a particular structure (e.g. Almon polynomial) on the distributed lag polynomial $\delta(L)$:

$$y_{t_m} = \rho(L^m) y_{t_m - h_m} + \beta B(L, \theta) x_{t_m - h_m + w} + u_{t_m}. \tag{2}$$

MIDAS is more parsimonious than UMIDAS in terms of the number of parameters, which often helps forecasting. However, it is nonlinear and must be estimated via nonlinear least squares (NLS), while ordinary least squares (OLS) suffices for UMIDAS.

The MIDAS model in (2) typically includes a single predictor x, as using multiple indicators complicates NLS estimation. Forecasts are then averaged to exploit the full information set. In contrast, UMIDAS can easily accommodate multiple predictors via OLS. But when the frequency mismatch (m) is large, the number of parameters grows quickly. This can be mitigated by using regularized estimators, as discussed below in the machine learning section.

2.3 Panel Machine Learning in Mixed-Frequency

To avoid introducing MIDAS-type nonlinearities, especially when dealing with a large number of predictors, and given the modest frequency mismatch in our application, we follow the UMIDAS approach for machine learning models.[1] Let X_{i,t_m} be an N_X-dimensional vector of high-frequency predictors, where each variable leads y with w periods, and let Z_{i,t_m} be the N_Z-dimensional vector collecting lags of the target variable y_{i,t_m} and lags of X_{i,t_m} aligned via the UMIDAS transformation. The model is then:

$$y_{i,t_m} = g(Z_{i,t_m-h_m+w};\, \theta_i) + u_{i,t_m}, \tag{3}$$

where $g(\cdot)$ is a flexible function approximator parameterized by θ_i—which may vary by unit depending on the degree of pooling—and u_{i,t_m} is an error term. The function g may correspond to penalized linear regression, tree-based models, or neural networks, as detailed below.

The function g is trained to approximate the conditional mean of y_{i,t_m} given the inputs Z_{i,t_m-h_m+w}, using regularization and tuning procedures suited to high-dimensional and nonlinear settings.[2] In the purely individual case, θ_i is estimated separately for each i; in pooled models, it is shared or partially shared across units.

Our application differs from standard macroeconomic forecasting in two key respects. First, it uses mixed-frequency predictors, requiring specific transformations (via UMIDAS) to align quarterly and annual information. Second, the target variable is a panel, y_{i,t_m}, which varies across cross-sectional units $i = 1, \ldots, N_y$. Exploiting cross-sectional structure is crucial, as correlations across units may contain predictive content. To this end, we consider several panel strategies: estimation by unit (no pooling), full pooling, and clustered pooling across economically meaningful groups. These choices define the degree of heterogeneity allowed in the

[1] Carriero et al. (2015) adopt the UMIDAS approach for a similar reason, though they use Bayesian priors as a shrinkage method in the presence of many predictors. Mogliani and Simoni (2021) and Hauzenberger et al. (2024) develop instead Bayesian MIDAS models for nowcasting with large information sets.

[2] The predictive importance of flexible nonlinear approximators g has been extensively studied in univariate forecasting applications with macroeconomic data. In particular, Goulet Coulombe et al. (2021a, 2022) show that nonlinear models can improve forecasting performance by more than 20% compared to linear benchmarks.

model's parameters θ_i, as detailed next.[3]

NO PANEL DATA. We can ignore completely the panel structure, so the model remains

$$y_{i,t_m} = g(\tilde{Z}_{i,t_m-h_m+w}; \theta_i) + u_{i,t_m},$$

where $\tilde{Z}_{i,t_m} \subset Z_{i,t_m}$ includes only state-specific predictors and national indicators. This specification is estimated separately for each i, without borrowing information across states.

INDIVIDUAL. We can acknowledge the existence of the panel structure, but assume complete heterogeneity across units, such that the nowcasting model becomes

$$y_{i,t_m} = g(Z_{i,t_m-h_m+w}; \theta_i) + u_{i,t_m},$$

where now Z_{i,t_m} contains covariates from all states. The model is specified and estimated separately for each state i.

POOLED. We can embrace the panel structure and assume homogeneity across units, which results in the pooled version

$$y_{i,t_m} = g(Z_{i,t_m-h_m+w}; \theta) + u_{i,t_m},$$

where Z_{i,t_m} contains all covariates and fixed effects to account for unobserved state-specific heterogeneity. The model is specified and estimated by pooling all states.

CLUSTERING. Finally, we can allow for some heterogeneity conditional on a prior clustering

$$y^l_{i,t_m} = g(Z^l_{i,t_m-h_m+w}; \theta^l) + u^l_{i,t_m},$$

where units i are grouped in the cluster $l = 1, \dots L$, for a total of L clusters. Given the nature of our predictive problem, we divided the 48 US states in the following (alternative) clusters (presented in Figure 14 in the Appendix).

[3] Theoretical insights on the relative forecasting performance of pooled and heterogeneous panel estimators in linear settings have been recently explored by Pesaran et al. (2024).

- REGIONAL: The most obvious regional clustering is to pool states over 4 regions: the Northeast, the Midwest, the South and the West.

- POLITICAL: The states are pooled together according to their political allegiance through time which results into 2 clusters: republican and democrat. These clusters are updated every 4 years after the election.

- GDP: States are grouped according to their GDP per capita: poorest (those belonging to less than 0.25 percentile), middle (those between 0.25 and 0.75) and rich (>0.75). These clusters are updated every year.

- HIERARCHICAL: States are grouped using a hierarchical clustering algorithm (Ward Jr, 1963), an unsupervised learning approach. Clustering is performed on the target variables across states used in the supervised forecasting task. Clusters are updated every year.

2.4 Forecasting Models

We construct forecasting models by means of the direct approach (Marcellino et al., 2006), which requires to specify a model for each forecast horizon of interest but does not need forecasts for the exogenous regressors, which is a big plus in our context with possibly hundreds of regressors, and the use of simulation methods to compute multiple step ahead forecasts for the nonlinear models, which leads to a major reduction in computing time. For simplicity, we omit the subscript i but, in every model, the target variable and predictors are observable across the panel dimension.

2.4.1 Linear Models

BENCHMARK. Like in Ghysels et al. (2022), we consider a random walk model as benchmark, which, though simple, often forecasts well.

UMIDAS. The UMIDAS models are specified with one high-frequency indicator indicator at a time, to avoid the curse of dimensionality. As in Ghysels et al. (2022), the set of predictors for UMIDAS models is composed of three state-level variables and ten national indicators

described above. The final forecast is then obtained by averaging the predictions across all single-indicator models.

PENALIZED REGRESSIONS. These models extend the linear UMIDAS framework to a high-dimensional setting. Regularization permits their estimation even when the number of regressors is larger than that of observations. Penalized regressions shrink coefficient estimates toward zero, which can introduce some bias in the estimators and associated forecasts but also reduce estimation uncertainty, which in turn can improve out-of-sample forecast accuracy. The general form of the penalized estimator is:

$$\hat{\beta} = \arg\min_{\beta} \sum_{t_m=1}^{T_m} \left(y_{t_m} - Z_{t_m-h_m+w}\beta\right)^2 + \lambda \sum_{j=1}^{N_Z} |\beta_j|^{\eta}, \quad \eta > 0,$$

where $Z_{t_m-h_m+w}$ denotes the available predictors at time t_m, λ is a tuning parameter controlling the degree of regularization, and η determines the type of penalty. We consider two widely used special cases. **Ridge regression** ($\eta = 2$), imposes an ℓ_2 penalty to shrink all coefficients, and it is typically useful when predictors are highly collinear. **Lasso** ($\eta = 1$), uses an ℓ_1 penalty to perform variable selection, which often leads to setting some coefficients exactly to zero.

To incorporate known structure among predictors—such as lags of the same variable—we also consider the **Sparse-Group Lasso (sg-Lasso)**, introduced by Simon et al. (2013). This estimator combines ℓ_1 and group-level ℓ_2 penalties, allowing for both sparsity across variables and within groups of coefficients. Recent work by Babii et al. (2022, 2023, 2024) adapts this approach to time series forecasting and demonstrates predictive gains in mixed-frequency contexts. The sg-Lasso estimator solves:

$$\hat{\beta} = \arg\min_{\beta} \sum_{t_m=1}^{T_m} \left(y_{t_m} - Z_{t_m-h_m+w}\beta\right)^2 + \lambda \left[\alpha \sum_{g=1}^{G} w_g |\beta_g| 2 + (1-\alpha) \sum j = 1^{N_Z} |\beta_j|\right],$$

where β_g denotes the subvector of coefficients for group g, and w_g is an optional weight (e.g., $w_g = \sqrt{|g|}$). The mixing parameter $\alpha \in [0,1]$ interpolates between Lasso ($\alpha = 0$) and Group Lasso ($\alpha = 1$), with intermediate values yielding the Sparse-Group Lasso.

As in Babii et al. (2022), we define each group as the full set of high-frequency lags associated

with a given predictor. The hyperparameters λ and α are selected via cross-validation.

2.4.2 Nonlinear Models

RANDOM FORESTS. This algorithm, besides being capable of handling large datasets, provides a means of approximating nonlinear functions by combining regression trees. Hence, it allows to go beyond the linear specification for the conditional mean of the UMIDAS type regressions. Each regression tree partitions the feature space defined by $Z_{t_m-h_m+w}$ into distinct regions and uses the region-specific mean of the target variable y_{t_m} as the forecast, i.e. for M_{RF} leaf nodes

$$\hat{y}_{t_m} = \sum_{m_{RF}=1}^{M_{RF}} c_{m_{RF}} I_{(Z_{t_m-h_m+w} \in R_{m_{RF}})},$$

where $R_1, ..., R_{M_{RF}}$ is a partition of the feature space. Since individual trees tend to overfit, Breiman (2001) introduced the Random Forest algorithm to stabilize the predictions through aggregation. The method builds multiple trees on bootstrapped subsamples of the data, and at each split only a random subset of features is considered. This feature sampling promotes diversity across trees and reduces correlation between them. The final forecast is obtained by averaging over the predictions of all trees.

BOOSTED TREES. This algorithm provides an alternative ensemble strategy based on sequential learning. Instead of averaging predictions, Boosted Trees iteratively add new trees to correct the errors of the previous ones. At each step n, a tree $f(Z_{t_m-h_m+w}; c_n)$ is trained to fit the pseudo-residuals $e_{t_m}^{(n)} = y_{t_m} - \hat{y}_{t_m}^{(n)}$, and the prediction is updated as:

$$\hat{y}_{t_m}^{(n+1)} = \hat{y}_{t_m}^{(n)} + \rho_n f(Z_{t_m-h_m+w}; c_n),$$

where (ρ_n, c_n) minimizes the squared error loss. The process continues for a pre-specified number of iterations or until convergence.

In our setup, each tree has a maximum depth of 5. We use a subsample of 75% of the training data at each iteration, and we tune both the number of trees and the learning rate $\eta_{\text{BT}} \in \{0.01, 0.005\}$ via cross-validation. This approach often yields high predictive accuracy by

efficiently correcting residual patterns in the data.

Neural Networks. We employ fully connected feed-forward neural networks with a dual-input design tailored to panel forecasting. The network processes two types of input features: (i) continuous variables, which include quarterly and annual macroeconomic predictors, and (ii) categorical variables identifying the cross-sectional unit. This dual-stream architecture enables the network to capture both temporal dynamics and cross-sectional heterogeneity.

Traditionally, fixed effects in panel models are introduced through dummy variables—i.e., one-hot encoding of the categorical unit identifiers. In a neural network, however, this approach becomes computationally inefficient and uninformative when the number of categories (e.g., states) is large. Instead, we adopt a more scalable and expressive solution: categorical embedding. An embedding layer maps each state identifier to a low-dimensional continuous vector that is jointly optimized with the rest of the network via backpropagation.

This approach offers several key advantages. First, it reduces the dimensionality of the input space by replacing sparse one-hot vectors with dense, compact representations. Second, and more importantly, it allows the network to learn latent similarities between units. States with similar fiscal behavior can be mapped to nearby locations in the embedding space, enabling parameter sharing and improving generalization. This is richer than static intercepts and supports nonlinear interactions between state identity and macroeconomic predictors.

Recent empirical evidence from Ma and Zhang (2020) shows that entity embeddings significantly outperform both one-hot and label encoding in classification tasks, particularly when the number of categories is large. Not only do embeddings mitigate the curse of dimensionality, but they also yield meaningful geometric relationships among categories, which can help improve model interpretability and predictive power. This aligns with our findings in the fiscal forecasting context, where embeddings enable the model to uncover complex fiscal dynamics across states without manually engineering groupings or hierarchies.

Formally, our neural network architecture unfolds as follows. Let Z_{t_m} denote the continuous inputs (e.g., yearly and quarterly predictors defined above) at time t_m, and C_{t_m} denote the categorical input identifying the state. The model proceeds in five steps:

$$
\begin{aligned}
\mathbf{u}_{t_m} &= f_{\text{cont}}\left(Z_{t_m}; \boldsymbol{\theta}_{\text{cont}}\right), && \text{(transformation of continuous inputs)} \\
\mathbf{v}_{t_m} &= \text{Embed}\left(C_{t_m}; \boldsymbol{\theta}_{\text{emb}}\right), && \text{(embedding of categorical inputs)} \\
\mathbf{x}_{t_m} &= \left[\mathbf{u}_{t_m};\, \mathbf{v}_{t_m}\right], && \text{(concatenation of both representations)} \\
\mathbf{h}_{t_m} &= f_{\text{ff}}\left(\mathbf{x}_{t_m}; \boldsymbol{\theta}_{\text{ff}}\right), && \text{(hidden feed-forward layers)} \\
\hat{y}_{t_m} &= \mathbf{w}_{\text{out}}^{\top} \mathbf{h}_{t_m - h_m + w} + b_{\text{out}}, && \text{(final forecast)}
\end{aligned}
$$

Here, $f_{\text{cont}}(\cdot)$ is a feed-forward network applied to the continuous inputs, $\text{Embed}(\cdot)$ maps a categorical input to a learned continuous vector, and $f_{\text{ff}}(\cdot)$ is another feed-forward network operating on the concatenated representation. The final layer outputs a point forecast for the fiscal variable of interest.

This hybrid structure, illustrated in Figure 1, allows the model to learn both shared and unit-specific predictive patterns. We use ReLU activation functions throughout: $f^{(n)}(z) = \max\{0, z\}$ for each hidden layer n. Training uses mini-batch gradient descent with the Adam optimizer (learning rate 5×10^{-5}) and early stopping after 20 epochs without improvement in validation loss. The description of the embedding hemisphere is presented in Figure 13 in the Appendix.

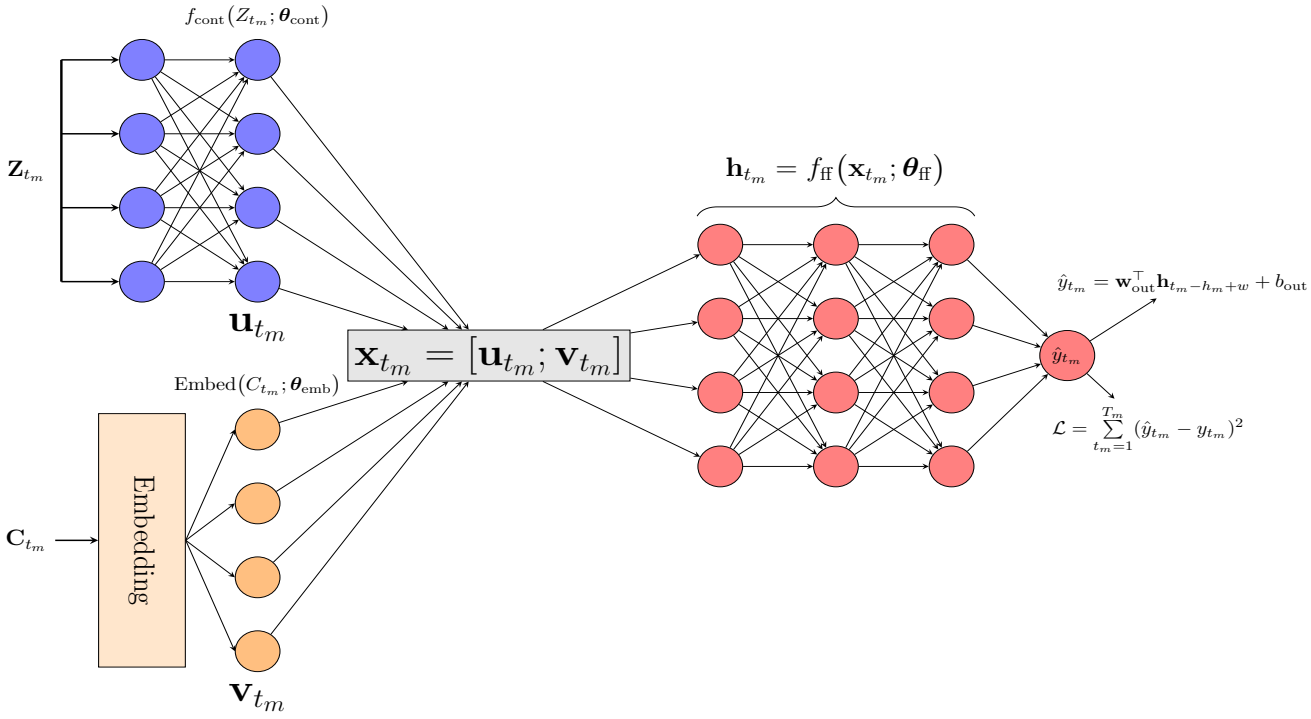
2.5 Tuning Hyperparameters

Following Kalfa et al. (2024), we explicitly document the hyperparameter choices and model configurations across all machine learning methods. This is especially important in a multidimensional setting such as ours, where the size and complexity of the hyperparameter space can substantially affect performance and interpretability.

We use two lags of the target variable and allow up to 8 lags for quarterly predictors. Table 1 summarizes the main hyperparameters and architectural choices for each model, including regularization parameters, optimization settings, and early stopping rules where applicable.

To ensure robust and generalizable performance in a panel time-series context, we implement a cross-validation procedure that respects the temporal ordering of observations—thereby eliminating look-ahead bias. In addition, we impose a balance constraint across panel units (e.g., U.S. states), ensuring that each fold contains a comparable number of observations per

Figure 1: Neural network architecture



Note: The architecture combines a dense network for continuous predictors with an embedding layer for categorical identifiers (states). The embedding captures unit-specific fixed effects in a low-dimensional representation that is learned jointly with the rest of the model via back-propagation.

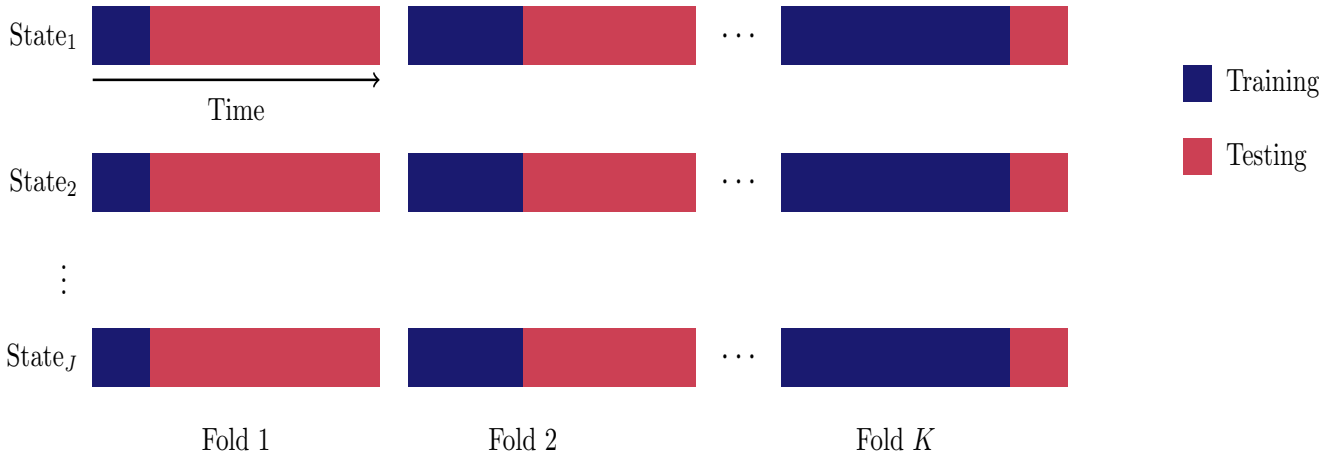
Table 1: Hyperparameters and model specifications

Model	**Hyperparameters and Settings**
Ridge	$\lambda \in [0.1, 100{,}000]$; geometric spacing
Lasso	$\lambda \in [10^{-4}, 1]$; 500 values; linear spacing; max iterations = 200,000
Sparse-Group Lasso	$\lambda \in [10^{-4}, 1]$; 500 values; $\alpha \in [0, 1]$ with 10 grid points; linear spacing
Random Forest	Max features $\in \{0.2, 0.35, 0.5\}$; number of trees = 500
Boosting Trees	Subsample = 0.75; learning rate $\in \{0.01, 0.005\}$; max depth = 5; number of trees $\in \{100, 250, 500, 750, 1000\}$
Neural Network	3 hidden layers; 400 neurons per layer; ReLU activations; Adam optimizer; learning rate = 5e-5; early stopping patience = 20; max epochs = 100; embedding dimension = $\min\{3, \sqrt{J}\}$

Note: Summary of key hyperparameters and modeling choices for each machine learning model. Grid values are selected via cross-validation, respecting the panel and time-series structure of the data.

unit. This helps avoid overfitting to overrepresented states and encourages more uniform generalization across the panel. Figure 2 illustrates this cross-validation structure, where training and testing windows are defined sequentially for each unit. Blue segments represent training periods, while red segments denote test periods.

Figure 2: Cross-validation structure over time and panel units



Note: Each row corresponds to a panel identity (e.g., State$_1$, . . . , State$_J$), and each block illustrates a train-test split for a given fold. Blue segments indicate training windows; red segments indicate testing windows.

2.6 Forecast Evaluation Framework

We evaluate the forecasting performance of all models over the period 2000 to 2020. Although the fiscal indicators to be predicted are annual, we leverage the availability of quarterly predictors to update forecasts in real time as new information becomes available. Specifically, we focus on nowcasts produced at the end of the first quarter of each year—i.e., when Q1 data is available but before official annual outcomes are released.

All models are estimated recursively using an expanding window. This setup ensures that each forecast is based only on information that would have been available at the time of forecasting (apart for the data revisions, as we use final vintage data due to the large set of indicators we consider). The expanding window approach also reflects how forecasters update their models in practice, allowing us to incorporate more data over time, potentially improving estimation precision—especially for more flexible models that benefit from larger sample sizes.

Following standard practice in the forecasting literature, we assess point forecast accuracy

using the mean squared error (MSE), computed over the evaluation sample. All results reported in the next sections refer to out-of-sample performance based on this recursive procedure.

3 Forecasting Performance Across Models and Panel Structures

We now present the results of the forecast evaluation exercise, examining how model choice, panel structure, and the nature of the target variable affect predictive performance. The analysis begins with aggregate results across states, then moves to more disaggregated and illustrative comparisons, highlighting the respective contributions of model flexibility, cross-sectional pooling, and mixed-frequency information.

3.1 Aggregate Forecasting Performance

We begin by assessing the overall predictive performance by averaging across all target variables and states. Our analysis focuses on two key aspects: the performance of different modeling approaches and the role of panel structures. To capture the full distribution of forecasting outcomes across states, variables, models, and panel structures, we present the results using violin plots. The performance metric is based on the mean squared error (MSE) computed across these dimensions. Unlike simple bar charts or box plots, violin plots simultaneously display measures of central tendency, dispersion, and the shape of the distribution, offering a more comprehensive depiction of the underlying heterogeneity. This choice is particularly relevant in our context, where cross-sectional and model-driven differences are substantial, and where capturing the entire distribution of forecast accuracy is crucial for evaluating model robustness.

3.1.1 Impact of Model Specification on Forecast Accuracy

Figure 3 presents the overall distribution of forecasting performance across models, relative to the RW, averaged over all fiscal variables and states. The results reveal substantial differences in predictive accuracy depending on the choice of $g(\cdot)$. Machine learning models clearly outperform the benchmark UMIDAS forecast, which is the mean of the forecasts resulting from

the single indicator UMIDAS models. In particular, penalized regressions and neural networks deliver the strongest overall performance, reflecting their capacity to exploit both sparse and dense predictive signals in the data. Nonlinear models provide additional gains, particularly in the tails of the distribution, suggesting their ability to handle complex interactions and heterogeneities that are not well captured by linear models. The distribution of MSEs across models also shows that, while variability exists across states and target variables, machine learning models generally deliver more stable and concentrated performance than UMIDAS, whose forecasting accuracy exhibits wider dispersion.

Figure 3: Models' average performance

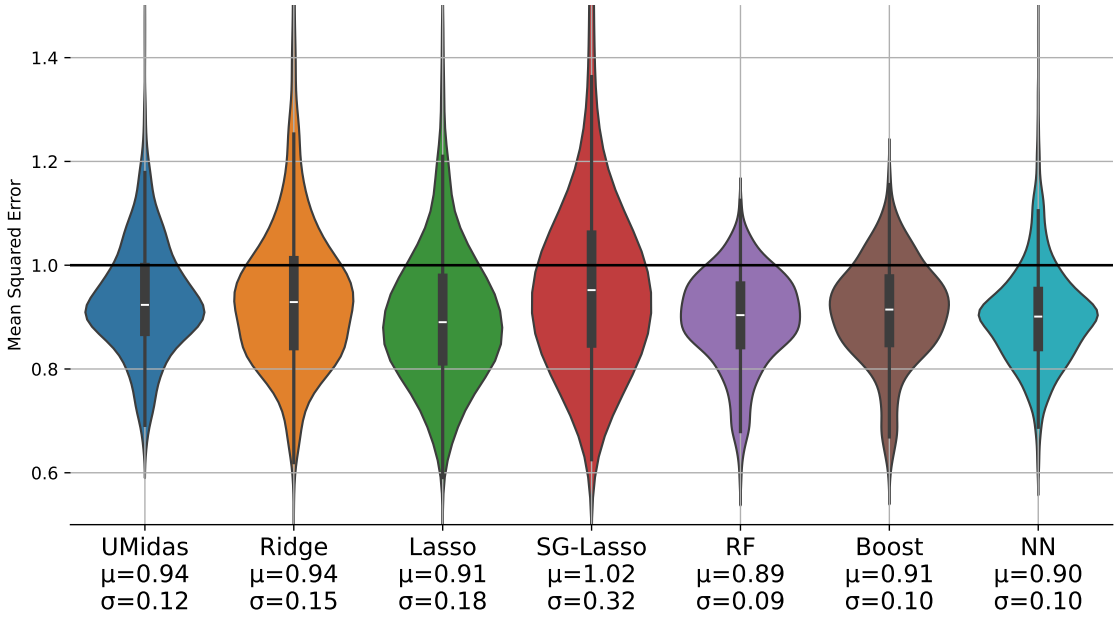
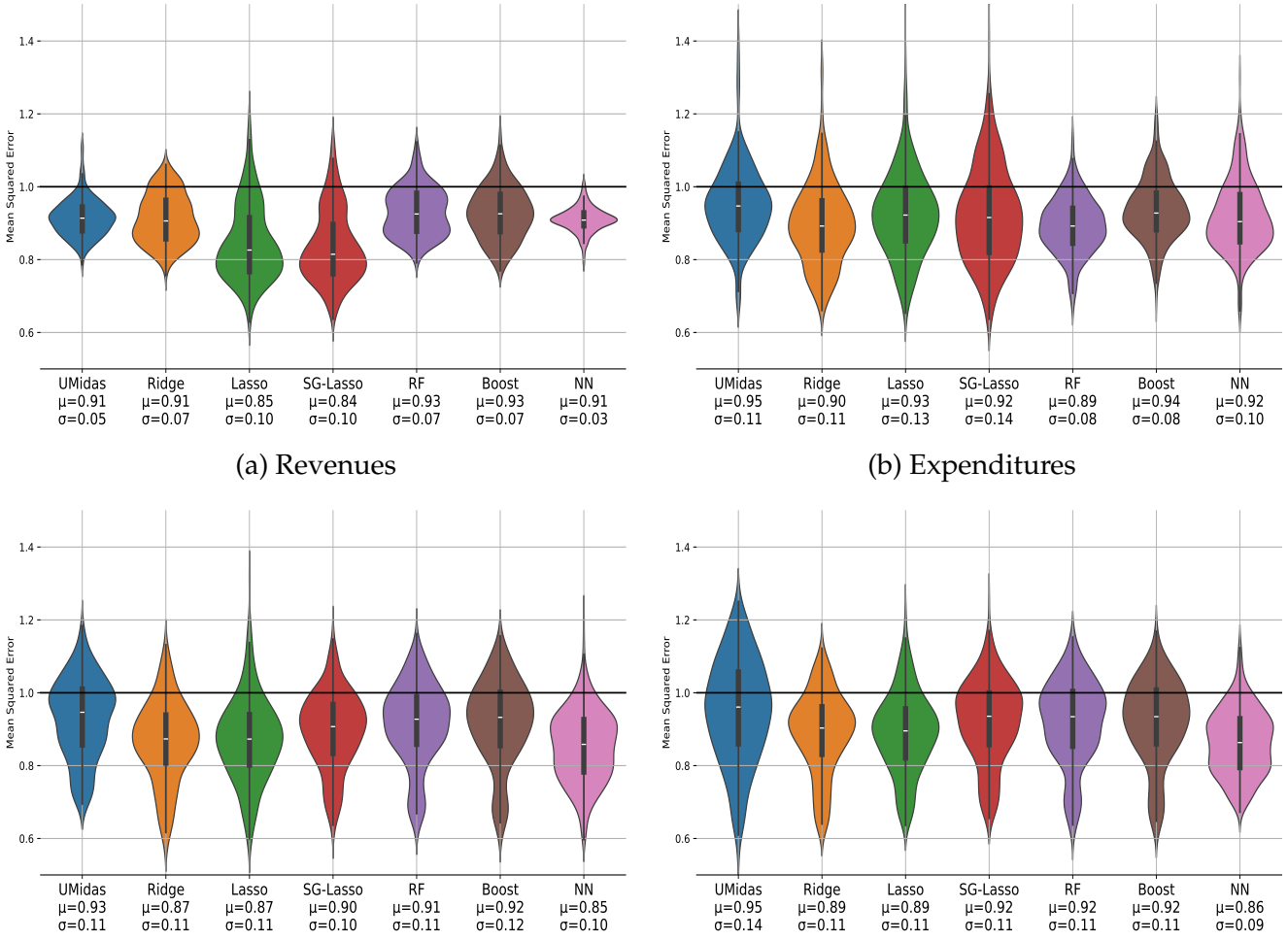


Figure 4 disaggregates the comparison by target variable, highlighting how the gains from machine learning methods vary across different fiscal outcomes. For total expenditure and deficit measures (both per capita and as a share of GDP), the improvement over UMIDAS is particularly pronounced. These results suggest that expenditures and deficits are more difficult to forecast using simple linear specifications and benefit more from the flexibility offered by machine learning techniques. In contrast, for total revenues, the relative performance gap between UMIDAS and machine learning models is smaller, indicating that revenue series may exhibit more stable or linear dynamics that are adequately captured even by simpler models.

Figure 4: Models' performance by variable



(a) Revenues

(b) Expenditures

(c) Deficit per capita

(d) Deficit in % of GDP

Overall, the disaggregated results reinforce the conclusion that the predictive gains from machine learning are not uniform across fiscal variables, with the largest benefits appearing for outcomes that are inherently more volatile or complex to model.

3.1.2 Impact of Panel Structure on Forecast Accuracy

Figure 5 examines the impact of panel structure on forecasting performance, averaged across all models and target variables. The results show that introducing some form of pooling across states systematically improves predictive accuracy compared to a no-pooling approach. Both full pooling (global structure) and clustered pooling (based on regional, political, GDP, or hierarchical groupings) outperform models estimated separately for each state. This highlights the benefit of exploiting cross-sectional information to enhance forecasts. Among pooling strate-

gies, no single structure clearly dominates, although clustered approaches generally provide slightly better median performance than global pooling. The dispersion of forecast errors is also reduced under clustered structures, suggesting that partial pooling based on economic or political similarities offers a good balance between flexibility and the benefits of aggregation.

Figure 5: Average impact of clustering

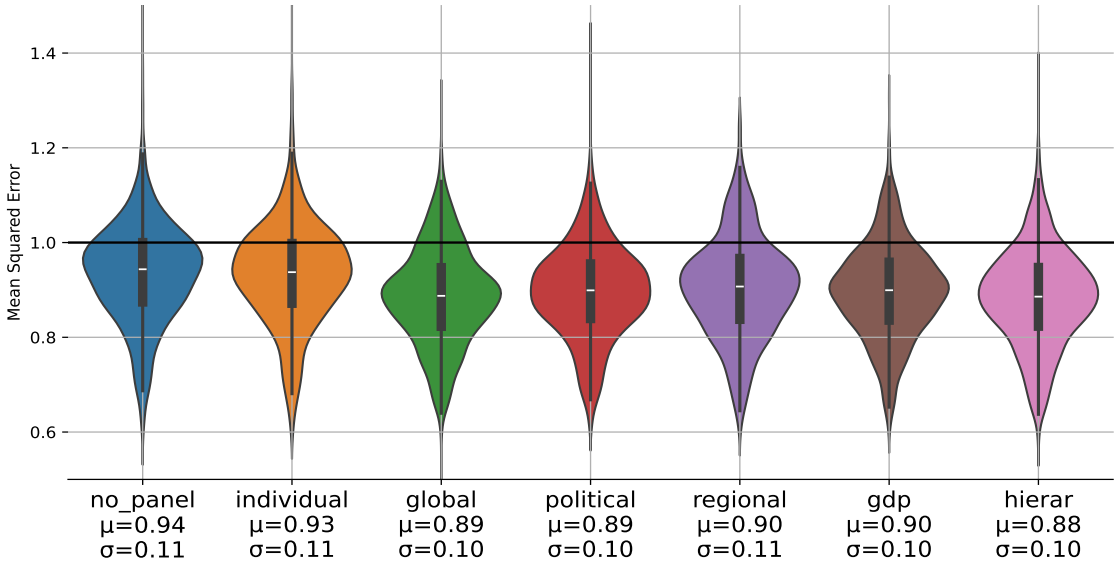
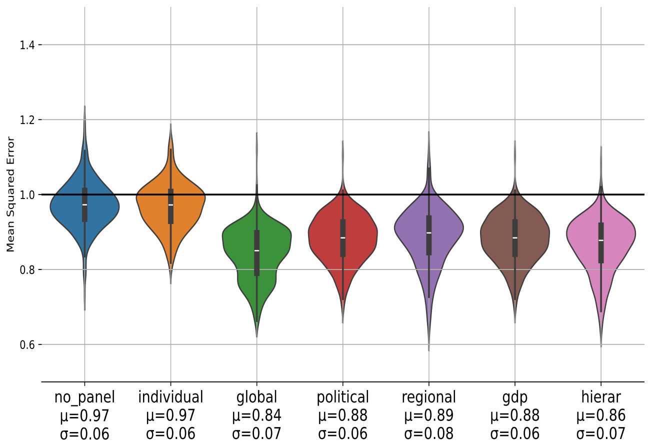
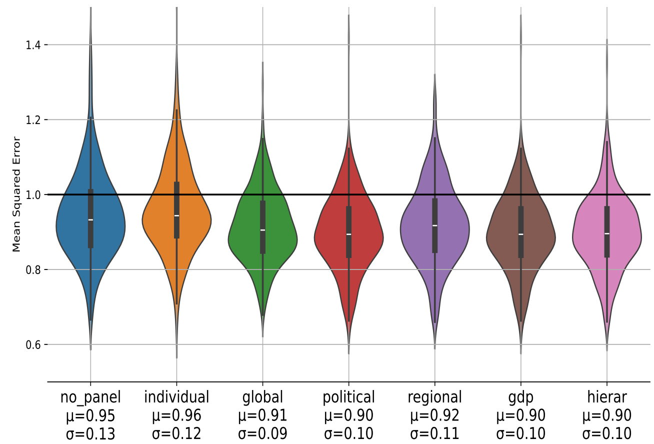


Figure 6 disaggregates the role of panel structure by target variable. The advantage of pooling appears particularly strong for deficit measures (both per capita and as a share of GDP), where clustered pooling strategies such as GDP-based or hierarchical clustering deliver noticeably better forecasts than global pooling. This suggests that deficits, which may reflect diverse fiscal behaviors and exposures across states, benefit from more targeted pooling strategies that respect underlying economic heterogeneity. For total revenues and expenditures, however, the differences between pooling structures are less pronounced, and global pooling performs almost as well as clustered alternatives. These patterns underscore the importance of adapting the pooling structure to the nature of the fiscal variable being forecasted: complex and heterogeneous outcomes require more flexible cross-sectional modeling than more stable fiscal aggregates.

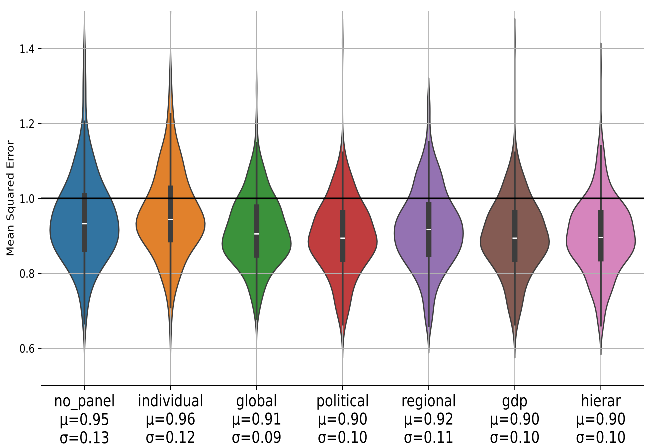
Figure 6: Clustering performance by variable



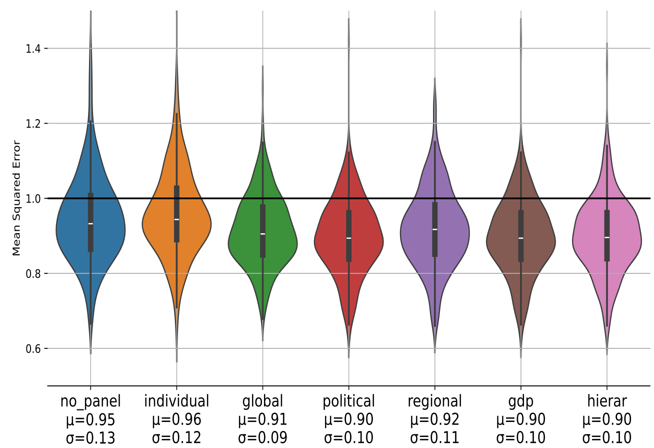
(a) Revenues



(b) Expenditures



(c) Deficit per capita



(d) Deficit in % of GDP

3.1.3 Longer forecast horizon

As a complementary exercise, we also evaluate the forecasting performance for the next fiscal year, using all information available up to the end of the current year (i.e., including predictors up to Q4 of year t). Figures15 and 16 in the Appendix present results analogous to Figures 3 and 5. Although the overall patterns remain broadly consistent—with nonlinear models and pooled structures improving forecast accuracy—some differences emerge, particularly in how model performance varies across fiscal variables.

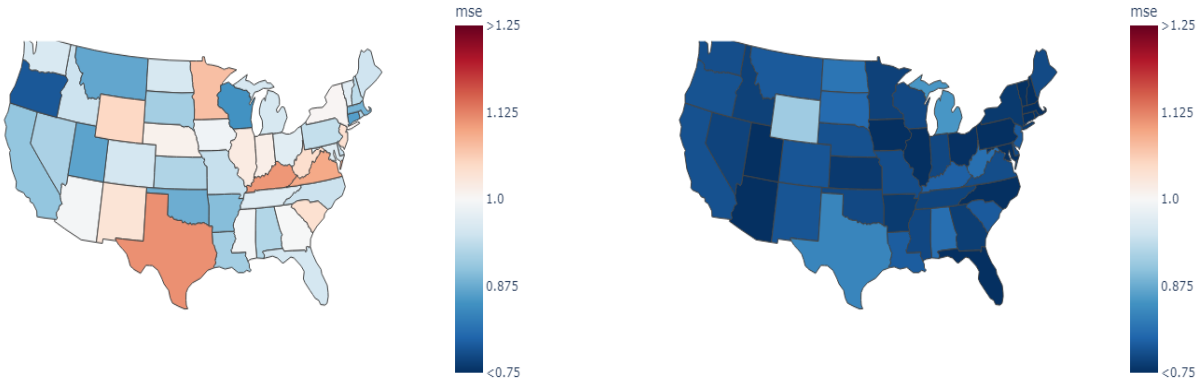
3.2 Variables, States, and Key Predictors

While the previous sections have focused on aggregate forecast performance across models, panel structures, and fiscal variables, such analyses inevitably abstract from the specific dynamics at play in individual cases. To complement the global findings, we now turn to a set of illustrative examples that provide a more granular perspective on the forecasting process.

3.2.1 Nowcasting total revenues

Figure 7 compares the distribution of forecast accuracy, relative to RW benchmark, across U.S. states for total revenues under two representative modeling strategies. The left panel shows the MSEs obtained from the benchmark UMIDAS model estimated without pooling (no panel), while the right panel displays those from the Lasso model with global pooling. The gains from adopting a machine learning model with cross-sectional pooling are clear: the Lasso × global combination delivers uniformly lower MSEs across almost all states. The improvement is particularly pronounced in smaller or more volatile states, where the UMIDAS model exhibits substantial forecast errors. This supports the earlier conclusion that pooling information across states stabilizes predictions and that flexible model specifications can better adapt to heterogeneous fiscal dynamics.

Figure 7: Nowcasting total revenue: comparison across states



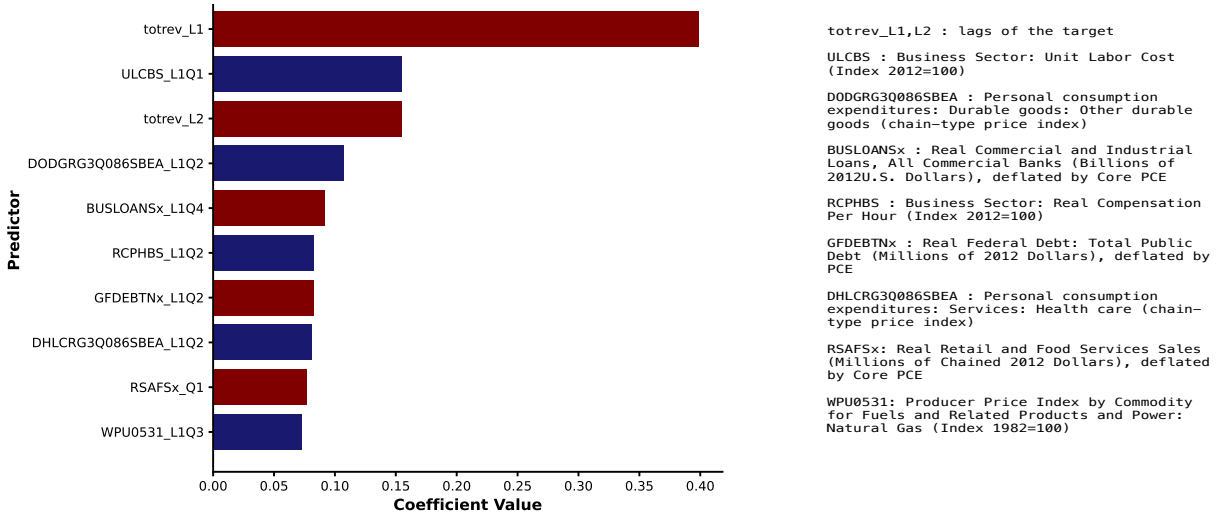
(a) UMIDAS, No panel

(b) Lasso, Global

Note: Mean squared forecast errors (MSE) computed across states for total revenue forecasts. Left panel reports results from the UMIDAS model without pooling. Right panel reports results from a Lasso model with global pooling. Lower MSEs indicate better accuracy.

Figure 8 reports the ten most influential predictors selected by the Lasso model under the global structure, along with their descriptions.[4] These predictors include both national macroeconomic indicators (e.g., GDP growth, industrial production, consumer sentiment) and high-frequency state-level variables (e.g., employment or wage growth). The prominence of macro indicators suggests that national economic conditions play a central role in shaping revenue expectations across states. At the same time, the inclusion of state-specific labor market indicators highlights the model's capacity to capture local fiscal signals. This mix of predictors confirms that the strength of the Lasso × global model lies in its ability to combine rich national information with relevant local dynamics through regularization and pooling.

Figure 8: Nowcasting total revenue: selected predictors



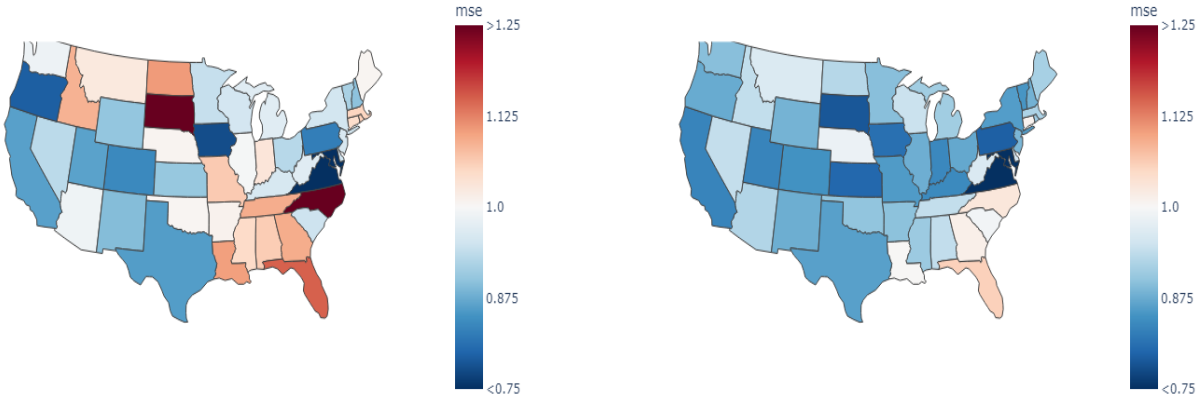
Note: Figure presents coefficient values (in absolute values) of the ten most important Lasso regressors.

3.2.2 Nowcasting total expenditures

Figure 9 compares forecast accuracy for total expenditures across states using two contrasting approaches: UMIDAS with no pooling and Random Forest with global pooling. As in the revenue case, incorporating pooling significantly improves performance relative to the no-pooling benchmark. However, the gains appear even larger when switching from UMIDAS to Random Forest, particularly for smaller and more volatile states. This highlights the advantage of non-linear models in capturing complex fiscal dynamics that simple linear approaches fail to model

[4]Since in Lasso model regressors are standardized, the size of the estimated coefficients are directly comparable.

Figure 9: Nowcasting total expenditures: comparison across states



(a) UMIDAS, No panel

(b) RF, Global

Note: Mean squared forecast errors (MSE) computed across states for total expenditures nowcasts. Left panel reports results from the UMIDAS model without pooling. Right panel shows results from a RF model with global pooling. Lower MSEs indicate better accuracy.

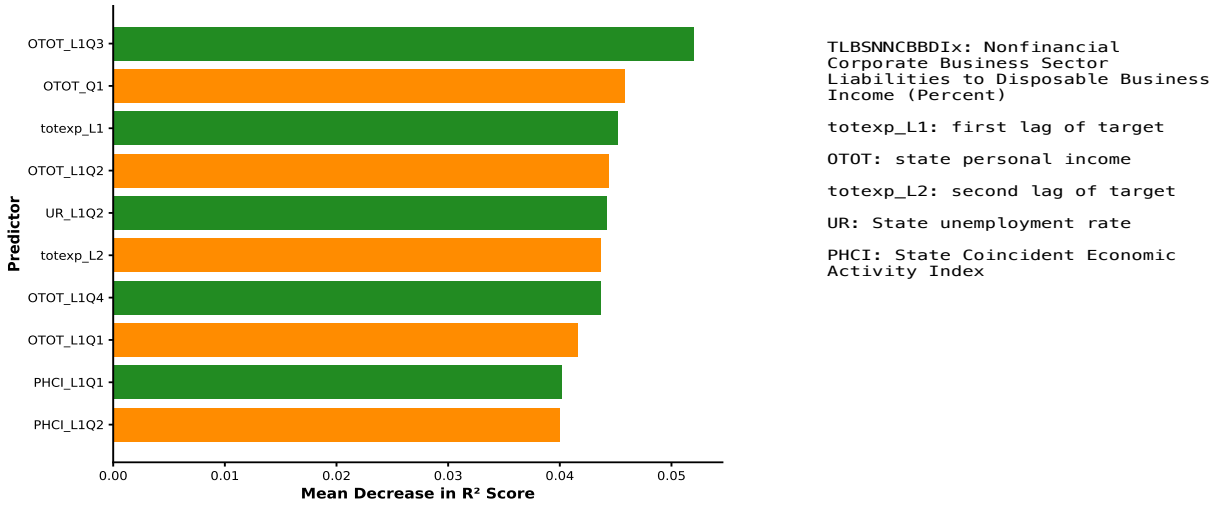
adequately. The RF model not only stabilizes forecasts across states but also substantially reduces the dispersion in prediction errors, supporting the earlier finding that nonlinearity and pooling jointly enhance forecast performance for volatile fiscal variables like expenditures.

Figure 10 displays the ten most important predictors in RF model. Unlike the coefficient estimates from penalized regressions, variable importance measures how much each predictor reduces impurity across the ensemble of decision trees. The selected predictors encompass a range of national macroeconomic indicators (e.g., industrial production growth, unemployment claims) and state-level variables. The prominence of real economic activity indicators suggests that expenditure forecasts are highly sensitive to contemporaneous macroeconomic and labor market conditions. Compared to the revenue case, the importance of financial indicators appears somewhat lower, reflecting differences in the drivers of expenditures versus revenues.

An important additional insight is that several state-specific predictors rank among the most influential variables, despite being vastly outnumbered by aggregate national indicators in the dataset. While previous work, such as Gu et al. (2020), proposed creating ex-ante interactions between state and national variables to address this imbalance, the Random Forest model is capable of capturing relevant nonlinear interactions endogenously. This ability reduces the need for explicit feature engineering. Moreover, the stronger role of state-specific variables in

forecasting expenditures relative to revenues suggests that local economic dynamics are particularly critical for modeling fiscal spending behavior across states.

Figure 10: Nowcasting total expenditures: selected predictors



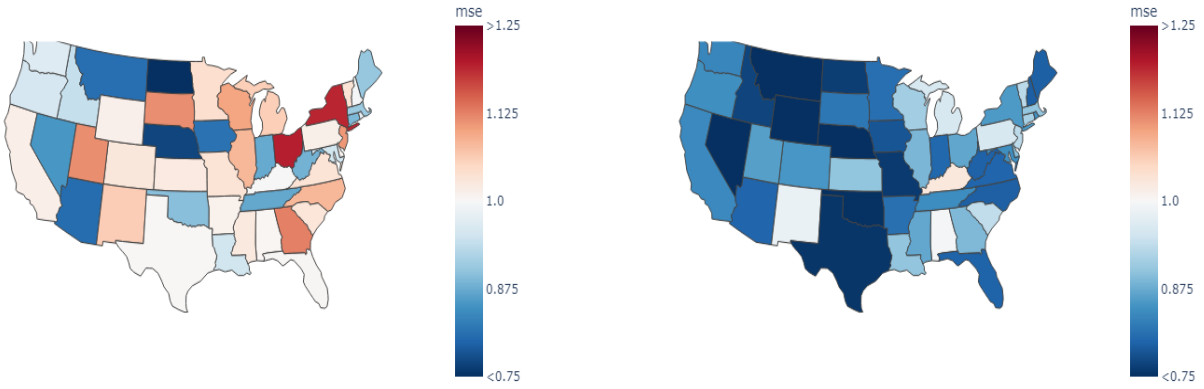
Note: Figure reports the ten most important predictors selected by Random Forests for each fiscal variable, based on mean decrease in R^2 score.

3.2.3 Nowcasting deficit as share of GDP

Figure 11 compares the state-level mean squared errors (MSEs) for forecasts of total deficits as a share of GDP under two modeling strategies: UMIDAS with no pooling and a Neural Network with GDP-based clustering. Consistent with the patterns observed for revenues and expenditures, the neural network model yields substantially lower forecast errors across almost all states. However, the relative gains are particularly striking in this case, with the UMIDAS model performing poorly for several states—especially those with more volatile fiscal environments. This suggests that deficits are more challenging to nowcast using linear, unpooled models and benefit from both nonlinearity and cross-sectional information sharing.

Taken together, the case studies presented in this section reinforce the aggregate findings by illustrating how pooling strategies and model flexibility translate into tangible improvements in forecast accuracy at the state level. Across all fiscal outcomes examined, machine learning models—especially those incorporating nonlinearity and clustering—consistently outperform traditional benchmarks. Moreover, the variable importance analyses reveal that effective forecasting relies on a combination of national macroeconomic signals and state-specific indicators.

Figure 11: Nowcasting deficit as share of GDP: across states



(a) UMIDAS, No panel

(b) NN, GDP clustering

Note: Mean squared forecast errors (MSE) computed across states for total deficit as a share of GDP nowcasts. Left panel reports results from the UMIDAS model without pooling. Right panel shows the NN model with GDP clustering case. Lower MSEs indicate better accuracy.

4 A Closer Look at the Forecasting Results

The results presented in the previous section highlight the gains achieved by combining flexible modeling approaches with appropriate panel structures. However, while aggregate performance metrics provide valuable insights, they may mask important variations across states, variables, models, and pooling strategies. We now study these more granular patterns.

4.1 Quantifying Impacts of Modeling and Panel Structure

To better understand the sources of forecasting accuracy and the statistical significance of differences across models, variables, and panel structures, we estimate a linear regression where the dependent variable is the pseudo-R^2 defined at the level of each state i, year t, model m, target variable v, and panel structure s. The pseudo-R^2 is computed as:

$$R^2_{i,t,m,v,s} = 1 - \frac{(y_{i,t,v} - \hat{y}_{i,t,v,m,s})^2}{\text{MSFE}^{\text{const}}_{i,v}},$$

where $\text{MSFE}^{\text{const}}_{i,v}$ is the out-of-sample mean squared forecast error of the RW benchmark that predicts $y_{i,t,v}$ using the historical average of $y_{i,t',v}$ for $t' < t$. [5]

[5]This approach, also known as response surface (Davidson and MacKinnon, 1993, Section 21.7), has been used in Goulet Coulombe et al. (2022). An alternative would be to adapt the method from Qu et al. (2024) to this

Since our evaluation period spans two decades, including business cycle turning points and crisis episodes, we include year fixed effects to account for common variation in forecast difficulty over time. This ensures that comparisons between models, variables, and panel structures are not confounded by macroeconomic shocks that affect all forecast targets in a given year.

We first estimate the following additive model with year fixed effects:

$$R^2_{i,t,m,v,s} = \alpha + \gamma_m + \delta_s + \theta_v + \lambda_t + \varepsilon_{i,t,m,v,s}, \tag{4}$$

where γ_m, δ_s, θ_v, and λ_t are categorical indicators for model class, panel structure, target variable, and year, respectively. All model classes, panel structures, and target variables are included in this specification, with the baseline category set to neural network for the model, pooling global for the panel structure, and revenues for the target variable. The intercept thus corresponds to the pseudo-R^2 for this reference configuration in the base year (2000).

Table 2 reports the estimated coefficients and the associated p-values where the t-stats have been produced using HAC standard errors. The stars indicate statistical significance levels (* for $p < 0.10$, ** for $p < 0.05$, and *** for $p < 0.01$). The interpretation of each coefficient is relative to the corresponding baseline. For instance, negative coefficients on alternative model classes imply a lower pseudo-R^2 relative to NN model, controlling for the other factors.

The results confirm that the neural network model consistently achieves the highest pseudo-R^2, with most alternative models performing significantly worse. The largest performance gaps are observed for UMidas, Boosting, and Random Forests, while Lasso and Ridge appear closer to the benchmark, though without statistical significance in the case of Lasso.

Panel structure also plays a significant role. The pooled version (global) remains the most favorable, as reflected by significantly negative coefficients for alternative structures such as predicting states on individual basis, without panel consideration, and regional clustering. This suggests that pooling across states contributes to forecast accuracy, likely by stabilizing estimation in the presence of limited annual data. Notably, the negative effect of the individual structure persists even after controlling for year effects, implying that its relative performance

multidimensional setup.

is not solely driven by crisis periods.

Regarding target variables, forecasting performance is significantly lower for deficits per capita and deficit as % of GDP, and expenditures compared to revenues. This aligns with the idea that fiscal expenditure and deficit dynamics may be harder to predict.

Finally, the year fixed effects highlight systematic differences in forecast difficulty across the sample period. In particular, the financial crisis period (2008–2010) is associated with sharply lower pseudo-R^2, confirming the challenge of forecasting during periods of heightened volatility. Some rebound is observed in later years, although the coefficients for most post-crisis years remain negative relative to the early 2000s. Interestingly, the impact of the 2020-COVID is much smaller than the effect of the Great Recession years (2008-09).

Overall, these results emphasize the importance of model choice, pooling strategy, and target variable in determining forecasting success. They also underline the value of controlling for time variation when comparing predictive models in the presence of possible instability.

While the additive specification provides valuable insights into the average marginal effects of model class and panel structure, it implicitly assumes that these effects operate independently. However, it is plausible that certain models may benefit more from specific pooling strategies, especially in contexts where regularization, nonlinearity, or flexibility interacts with the amount of available cross-sectional information.

To explore this possibility, we estimate an extended specification that includes interaction terms between model type and panel structure, capturing the joint effect of model choice and pooling strategy on forecasting performance. The estimated equation becomes:

$$R^2_{i,t,m,v,s} = \alpha + \theta_v + \eta_{m,s} + \lambda_t + \varepsilon_{i,t,m,v,s}, \tag{5}$$

where $\eta_{m,s}$ denotes the interaction between model and panel structure, and the additive terms γ_m and δ_s are absorbed into the interaction terms. Year and variable fixed effects remain included to control for systematic differences across time and target variables.

This interaction framework permits to test whether the relative performance of a given model depends on the chosen pooling strategy. For instance, models with strong regulariza-

Table 2: Marginal Predictive Performance: Pseudo-R^2 Regression Results

Coefficient	Estimate	p-value	Coefficient	Estimate	p-value
Constant	0.61***	0.00	year_2001	-0.36***	0.00
UMIDAS	-0.10***	0.00	year_2002	-0.39***	0.00
Ridge	-0.01	0.19	year_2003	0.12***	0.00
Lasso	-0.01	0.49	year_2004	-0.17***	0.00
SG-Lasso	-0.03**	0.01	year_2005	0.18***	0.00
RF	-0.05***	0.00	year_2006	0.22***	0.00
Boost	-0.08***	0.00	year_2007	-0.11***	0.00
No Panel	-0.09***	0.00	year_2008	-1.22***	0.00
Individual	-0.10***	0.00	year_2009	-2.72***	0.00
Political	-0.02	0.12	year_2010	-0.96***	0.00
Regional	-0.04***	0.00	year_2011	-0.22***	0.00
GDP	-0.02	0.11	year_2012	-1.35***	0.00
Hierar(2)	0.00	0.80	year_2013	0.08***	0.00
Hierar(4)	-0.03***	0.00	year_2014	0.13***	0.00
Expenditures	-0.08***	0.00	year_2015	-0.04**	0.01
Deficit/cap.	-0.05***	0.00	year_2016	-0.03	0.11
Deficit %GDP	-0.09***	0.00	year_2017	-0.18***	0.00
			year_2018	0.16***	0.00
			year_2019	0.21***	0.00
			year_2020	-0.26***	0.00

Notes: This table reports linear regression results from equation (4), estimating the marginal effects of model type, panel structure, target variable, and year dummies on pseudo-R^2 performance. Coefficients are relative to the baseline category: NN model, Global panel structure, Revenue target, and year 2000. Standard errors are HAC-robust. Asterisks denote significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

tion such as Lasso should benefit more from global pooling than from individual estimation, due to better stabilization of coefficient estimates. In contrast, flexible nonparametric models like Random Forests may be less sensitive to the choice of panel structure.

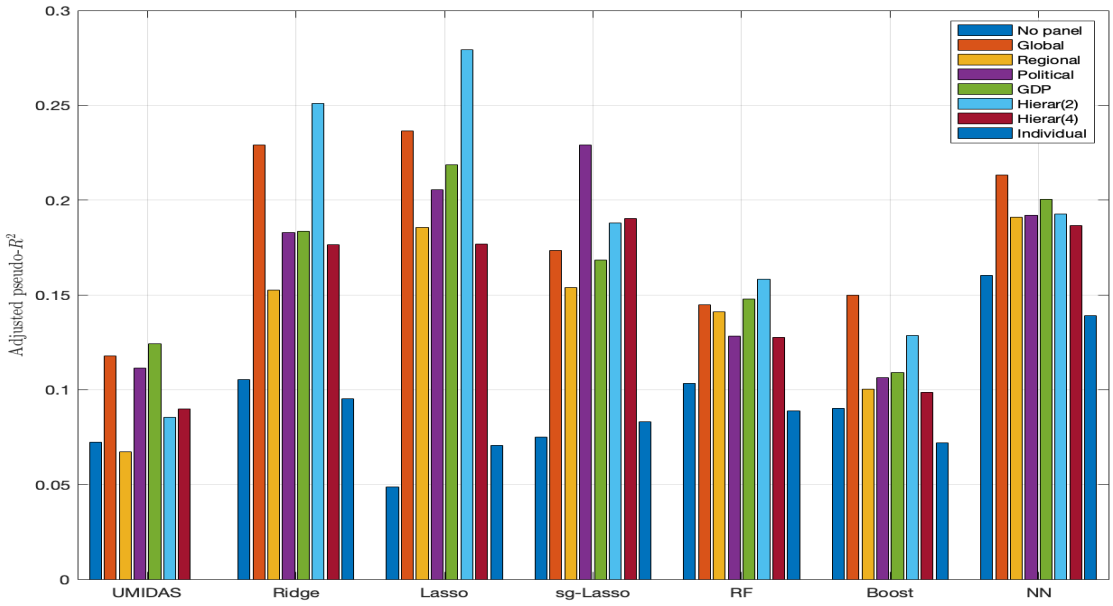
The results, illustrated in Figure 12, provide a clearer comparison of model robustness across different panel structures. Each group of bars represents a specific forecasting model, allowing us to observe how its performance changes depending on the pooling strategy. Neural networks consistently achieve high adjusted pseudo-R^2 values across most panel structures, particularly under global, GDP-based, and hierarchical pooling. This suggests that nonlinear models benefit from information sharing but are also robust to structural variation across states.

Lasso and sparse-group Lasso exhibit greater sensitivity to panel structure. Lasso reaches its peak performance under hierarchical pooling (Hierar), but its accuracy drops under individual

or no-panel settings. This pattern confirms that regularization alone is not sufficient without some form of cross-sectional pooling. UMIDAS and Ridge tend to underperform across the board, especially under more fragmented structures. Meanwhile, tree-based methods like Random Forests (RF) and Boosting exhibit stable but slightly lower predictive accuracy, indicating flexibility across structures but less overall gain from pooling.

This visualization reinforces the idea that forecasting accuracy depends jointly on model complexity and the degree of pooling. Nonlinear models—particularly neural networks—appear most robust, whereas linear regularized models require the support of appropriate panel structures to realize their full potential.

Figure 12: Pseudo-R^2 by model and panel structure



Note: Adjusted pseudo-R^2 values are obtained from regression (5). Each group of bars corresponds to a model, and colors indicate the panel structure used. The pseudo-R^2 is computed from out-of-sample forecasts and adjusted for the number of predictors.

4.2 Forecast Optimality: Evaluating Bias and Informational Efficiency

To assess whether forecasts are unbiased and well-calibrated, we implement the Mincer-Zarnowitz regression (Mincer and Zarnowitz, 1969) at the state level:

$$y_{i,t,v} = \alpha_{[m,s]} + \beta_{[m,s]}\widehat{y}_{i,t,v} + \mathrm{FE}_{v,t} + \varepsilon_{i,t,v}, \tag{6}$$

where $y_{i,t,v}$ denotes the realized value, $\widehat{y}_{i,t,v}$ is the forecast, and fixed effects $\text{FE}_{v,t}$ capture both fiscal variable identity and year effects. The regression is estimated separately for each model m and panel structure s, with HAC-corrected standard errors.

The Mincer-Zarnowitz regressions (6) test two core conditions for forecast optimality: unbiasedness (i.e., $\alpha = 0$) and informational efficiency (i.e., $\beta = 1$). The latter condition is often interpreted as verifying whether the forecast incorporates all available information in a proportional and undistorted way. However, it is important to note that the regression evaluates the response of the realized value to the forecast $\widehat{y}_{i,t,v}$—that is, to the transformation of available information through the forecasting model. A slope $\hat{\beta} < 1$ implies underreaction to predictive signals, while $\hat{\beta} > 1$ reflects overreaction. We also report the p-value from a joint Wald test of H_0: $\alpha = 0,\ \beta = 1$, offering a summary measure of overall forecast optimality.

Table 3 presents the detailed results, broken down by model and panel structure. Each model–structure pair yields estimates for $\hat{\alpha}$ and $\hat{\beta}$, with statistical significance denoted by stars. The joint p-value allows us to identify cases where forecasts are jointly biased and inefficient.

The results reveal several key patterns. First, UMIDAS, Lasso, and Sparse-Group Lasso produce forecasts that are systematically attenuated relative to the realized values, as evidenced by $\hat{\beta} \ll 1$ and strong rejections of the efficiency hypothesis across all pooling strategies. These models also tend to exhibit positive and significant bias terms, particularly under partial pooling (e.g., Political or Regional). Despite strong regularization, their predictive signals appear overly dampened, suggesting underutilization of available variation in the data.

In contrast, neural networks generally produce $\hat{\beta} > 1$, suggesting strong reactions to input signals. However, they also tend to have negative $\hat{\alpha}$, indicating an average upward bias in their forecasts. This overreaction–offset pattern hints at a different kind of inefficiency: while the NN model fully exploits variation in predictors, they may overshoot due to nonlinearities not fully aligned with the true data-generating process. Still, joint tests for optimality are not rejected in many configurations involving structured pooling (e.g., Global, Hierar(2)), underscoring their relative robustness.

The Random Forest and Boosting models show milder deviations: they often exhibit slopes

Table 3: Mincer-Zarnowitz Test Results by Cluster Grouping

	UMIDAS	Ridge	Lasso	SG-Lasso	RF	Boost	NN
No Panel							
$\hat{\alpha}$	0.02	0.01	0.02	0.02	0.01	0.01	-0.02
$\hat{\beta}$	0.49***	0.82	0.67***	0.45***	0.76*	0.79**	1.70**
p-value	0.00	0.24	0.00	0.00	0.21	0.14	0.11
Global							
$\hat{\alpha}$	0.03*	0.03*	0.02	0.03*	0.01	0.01	-0.02
$\hat{\beta}$	0.30***	0.72	0.82	0.38***	0.63***	0.70*	1.47*
p-value	0.00	0.11	0.31	0.00	0.01	0.16	0.22
Regional							
$\hat{\alpha}$	0.03*	0.02*	0.02	0.03**	0.02	0.01	-0.01
$\hat{\beta}$	0.26***	0.79	0.80	0.50***	0.49***	0.54***	1.46*
p-value	0.00	0.13	0.19	0.00	0.00	0.00	0.20
Political							
$\hat{\alpha}$	0.03*	0.03*	0.01	0.04**	0.01	0.01	-0.01
$\hat{\beta}$	0.27***	0.62**	0.64**	0.51***	0.59***	0.62**	1.50
p-value	0.00	0.04	0.14	0.00	0.00	0.05	0.23
GDP							
$\hat{\alpha}$	0.03*	0.03*	0.02	0.03*	0.02	0.01	-0.03
$\hat{\beta}$	0.29***	0.65**	0.66**	0.56***	0.54***	0.55***	1.67*
p-value	0.00	0.02	0.07	0.00	0.00	0.00	0.23
Hierar(2)							
$\hat{\alpha}$	0.03*	0.03*	0.02	0.04**	0.00	0.01	-0.00
$\hat{\beta}$	0.28***	0.72	0.74	0.40***	0.85	0.69***	1.34
p-value	0.00	0.10	0.27	0.00	0.47	0.02	0.46
Hierar(4)							
$\hat{\alpha}$	0.03*	0.03*	0.02	0.03**	0.01	0.01	-0.02
$\hat{\beta}$	0.27***	0.66**	0.58***	0.58***	0.67***	0.66***	1.40
p-value	0.00	0.03	0.00	0.00	0.00	0.00	0.27
Individual							
$\hat{\alpha}$	–	0.01	0.01	0.02	0.01	0.01	-0.02
$\hat{\beta}$	–	0.90	0.79	0.53***	0.79*	0.79	1.36*
p-value	–	0.65	0.32	0.01	0.22	0.32	0.25

Note: Each cell reports the estimate of $\hat{\alpha}$ (bias term), $\hat{\beta}$ (efficiency term), and the p-value from the joint test H_0: $\alpha = 0$, $\beta = 1$. Asterisks denote significance of individual tests: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. All standard errors are HAC-corrected.

below one, similar to the linear models, but with smaller bias terms and fewer rejections of the joint null. These tree-based models seem to balance flexibility with regularization more effectively when combined with richer pooling structures.

Finally, the individual and no-panel specifications almost always lead to significant departures from the benchmark of forecast optimality—especially for linear models—highlighting

the importance of borrowing strength across states through pooled estimation.

4.3 Testing for Conditional Heteroskedasticity in Forecast Errors

Heteroskedasticity is a frequent feature in linear models, in particular with a small number of regressors, while the added flexibility of nonlinear models and/or a large information set can reduce its extent. While the White test (White, 1980) is traditionally applied to in-sample residuals to detect conditional heteroskedasticity, we implement it here on the out-of-sample forecast errors $\hat{e}_{i,t,v} = y_{i,t,v} - \hat{y}_{i,t,v}$. This is justified by our interest in identifying whether the dispersion of prediction errors varies systematically with the magnitude of the forecast itself, and also by the fact that, under correct specification, the forecast errors should be a function of the unobservable future model errors.

Formally, we estimate the following regression separately for each combination of model m and panel structure s:

$$\hat{e}^2_{i,t,v} = c_{[m,s]} + \theta_{1,[m,s]}\hat{y}_{i,t,v} + \theta_{2,[m,s]}\hat{y}^2_{i,t,v} + \mu_{i,t,v}, \tag{7}$$

and test the null hypothesis H_0: $\theta_{1,[m,s]} = \theta_{2,[m,s]} = 0$. A rejection suggests that the variance of forecast errors is conditionally heteroskedastic, i.e., systematically related to the level of the predicted value. In each regression, we control for fixed effects at the variable × time level to absorb structural shifts across fiscal indicators and macroeconomic periods. This ensures that the test is sensitive only to variance instability conditional on the forecast itself, not to scale differences across targets or years. All standard errors are HAC-corrected.

The results, reported in Table 4, indicate that forecast errors from UMIDAS models are consistently heteroskedastic across all panel structures, as the null is strongly rejected in each case. This reflects the relative rigidity of UMIDAS specifications, which lack mechanisms to adapt the variance of their forecast errors across the range of predicted outcomes. Among linear machine learning methods, Ridge, Lasso, and sg-Lasso display mixed results: in simple panel structures such as No Panel or Global, the test is usually rejected, whereas more granular structures like

Individual or Hierarchical yield higher p-values, suggesting greater error stability. These findings indicate that pooling and regularization do help stabilize prediction dispersion, but only when the panel structure is sufficiently fine.

Tree-based models such as Random Forests and Boosting also show signs of heteroskedasticity in most configurations. This is not entirely surprising, as these methods can overfit local patterns, including in the tails of the distribution, thereby amplifying forecast variance in extreme regions. That said, some configurations involving hierarchical pooling mitigate this issue, particularly in the case of Boosting, where p-values rise under more structured panels.

The most robust model remains the neural network, for which the null of homoskedasticity is not rejected in several settings. Notably, under global pooling, the White test yields a p-value of 0.43, and values larger than 0.10 are observed under other specifications, such as Hierar(2) and Hierar(4). This suggests that neural networks, despite their high flexibility and potential for overfitting, can produce stable forecast dispersion when paired with appropriately structured pooling. These results reinforce the findings from earlier tests of forecast optimality: neural networks may overreact in terms of slope, but they produce forecast errors with more stable variance properties, especially when panel structure is well calibrated.

Table 4: White Test p-values for Conditional Heteroskedasticity in Forecast Errors

Model	No Panel	Global	Regional	Political	GDP	Hierar	Hierar4	Individual
UMIDAS	0	0	0	0	0	0	0	
Ridge	0.02	0	0	0	0	0	0	0.02
Lasso	0.25	0	0	0	0	0.03	0.01	0.03
SG-Lasso	0.11	0	0	0	0	0	0	0.10
RF	0	0	0	0	0	0.02	0	0.05
Boost	0.09	0.04	0	0	0	0.01	0	0.13
NN	0	0.43	0.16	0.14	0	0.12	0.17	0.11

Notes: Each cell reports the p-value from a White test for conditional heteroskedasticity applied to out-of-sample forecast errors. The test regresses squared forecast errors on predicted values and their squares, controlling for fixed effects at the variable × time level. The null hypothesis is that forecast error variance is not systematically related to the level of the forecast. A low p-value indicates evidence of misspecification in the conditional variance.

5 Conclusion

This paper studies the nowcasting and short-term forecasting of U.S. state-level fiscal variables by leveraging machine learning methods in a panel and mixed-frequency environment. Using a rich dataset of quarterly macroeconomic and financial indicators alongside annual fiscal outcomes, we systematically evaluated the predictive performance of several modeling strategies, focusing on the role of model flexibility, pooling structures, and the nature of fiscal targets.

Our findings suggest that combining mixed-frequency information with machine learning techniques substantially improves forecasting performance relative to traditional econometric models. Nonlinear machine learning models such as random forests, boosted trees, and neural networks consistently outperform benchmark specifications, especially for more volatile fiscal variables like expenditures and deficits. Moreover, incorporating panel structures—either via full pooling or through clustering across economically meaningful dimensions—delivers additional gains by efficiently exploiting cross-sectional information.

A key insight from our analysis is that predictive gains are heterogeneous across fiscal variables and depend on both model flexibility and the choice of pooling strategy. Forecasting expenditures and deficits particularly benefits from nonlinear methods and partial pooling approaches, while revenues are comparatively easier to predict and show smaller differences across models. Through forecast error diagnostics, we further show that nonlinear models are better at mitigating cross-sectional nonlinearities.

Finally, disaggregated case studies demonstrate that these improvements are not driven by a few states but are broadly shared across the panel, and that state-specific economic conditions play an important role in driving forecast performance. This highlights the advantage of flexible models capable of capturing both national and local dynamics.

Overall, our results suggest that mixed-frequency panel machine learning models are powerful tools for timely fiscal monitoring at the state level. They offer a promising avenue for improving the responsiveness and accuracy of fiscal surveillance frameworks, particularly in an environment characterized by economic heterogeneity and rapid information flows.

References

Asimakopoulos, S., Paredes, S., and Warmedinger, T. (2020). Real-Time Fiscal Forecasting Using Mixed-Frequency Data. *Scandinavian Journal of Economics*, 122:369–390.

Babii, A., Ball, R. T., Ghysels, E., and Striaukas, J. (2023). Machine learning panel data regressions with heavy-tailed dependent data: Theory and application. *Journal of Econometrics*, 237(2, Part C):105–115.

Babii, A., Ball, R. T., Ghysels, E., and Striaukas, J. (2024). Panel data nowcasting: The case of price–earnings ratios. *Journal of Applied Econometrics*, 39(2):292–307.

Babii, A., Ghysels, E., and Striaukas, J. (2022). Machine Learning Time Series Regressions With an Application to Nowcasting. *Journal of Business & Economic Statistics*, 40(3):1094–1106.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Carriero, A., Clark, T. E., and Marcellino, M. (2015). Realtime Nowcasting with a Bayesian Mixed Frequency Model with Stochastic Volatility. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 178(4):837–862.

Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York.

Favero, C. and Marcellino, M. (2005). Modelling and Forecasting Fiscal Variables for the Euro Area. *Oxford Bulletin of Economics and Statistics*, 67:755–783.

Foroni, C., Marcellino, M., and Schumacher, C. (2015). U-MIDAS: MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society - Series A*, 178(1):57–82.

Foroni, C., Marcellino, M., and Stevanovic, D. (2019). Mixed frequency models with MA components. *Journal of Applied Econometrics*, 34(5):688–706.

Fosten, J. and Greenaway-McGrevy, R. (2022). Panel data nowcasting. *Econometric Reviews*, 41(7):675–696.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131:59–95.

Ghysels, S., Grigoris, F., and Ozkan, N. (2022). Real-time Forecasts of State and Local Government Budgets with an Application to the COVID-19 Pandemic. *National Tax Journal*, 75(4):731–763.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2021a). Macroeconomic data transformations matter. *International Journal of Forecasting*, 37:1338–1354.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2022). How is Machine Learning Useful for Macroeconomic Forecasting? *Journal of Applied Econometrics*, 37(5):920–964.

Goulet Coulombe, P., Marcellino, M., and Stevanovic, D. (2021b). Can Machine Learning Catch the COVID-19 Recession? *National Institute Economic Review*, 256:71–109.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hauzenberger, N., Marcellino, M., Pfarrhofer, M., and Stelzer, A. (2024). Nowcasting with Mixed Frequency Data Using Gaussian Processes. *arXiv*, 2402.10574:https://arxiv.org/abs/2402.10574.

Kalfa, S. Y., Timmermann, A., and van der Zwan, T. (2024). Overhyped? Can ML Models Reliably Predict Stock Returns? Working paper, October 2024.

Khalaf, L., Kichian, M., Saunders, C. J., and Voia, M. (2021). Dynamic panels with MIDAS covariates: Nonlinearity, estimation and fit. *Journal of Econometrics*, 220(2):589–605. Annals Issue: Celebrating 40 Years of Panel Data Analysis: Past, Present and Future.

Leal, T., Pérez, J. J., Tujula, M., and Vidal, J.-P. (2008). Fiscal Forecasting: Lessons from the Literature and Challenges. *Fiscal Studies*, 29(3):347–386.

Ma, Y. and Zhang, Z. (2020). Travel mode choice prediction using deep neural networks with entity embeddings. *IEEE Access*, 8:64959–64970.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2):499–526.

McCracken, M. and Ng, S. (2020). FRED-QD: A quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research.

Mincer, J. and Zarnowitz, V. (1969). The evaluation of economic forecasts. *In: J. Mincer (ed.), Economic Forecasts and Expectations*, pages 3–46.

Mogliani, M. and Simoni, A. (2021). Bayesian MIDAS penalized regressions: Estimation, selection, and prediction. *Journal of Econometrics*, 222(1, Part C):833–860.

Onorante, L., Pedregal, D. J., Pérez, J. J., and Signorini, S. (2010). The usefulness of infra-annual government cash budgetary data for fiscal forecasting in the euro area. *Journal of Policy Modeling*, 32(1):98–119.

Pesaran, M. H., Pick, A., and Timmermann, A. (2024). Forecasting with panel data: Estimation uncertainty versus parameter heterogeneity. Cambridge Working Papers in Economics, CWPE 2219.

Qu, R., Timmermann, A., and Zhu, Y. (2024). Comparing forecasting performance with panel data. *International Journal of Forecasting*, 40(3):918–941.

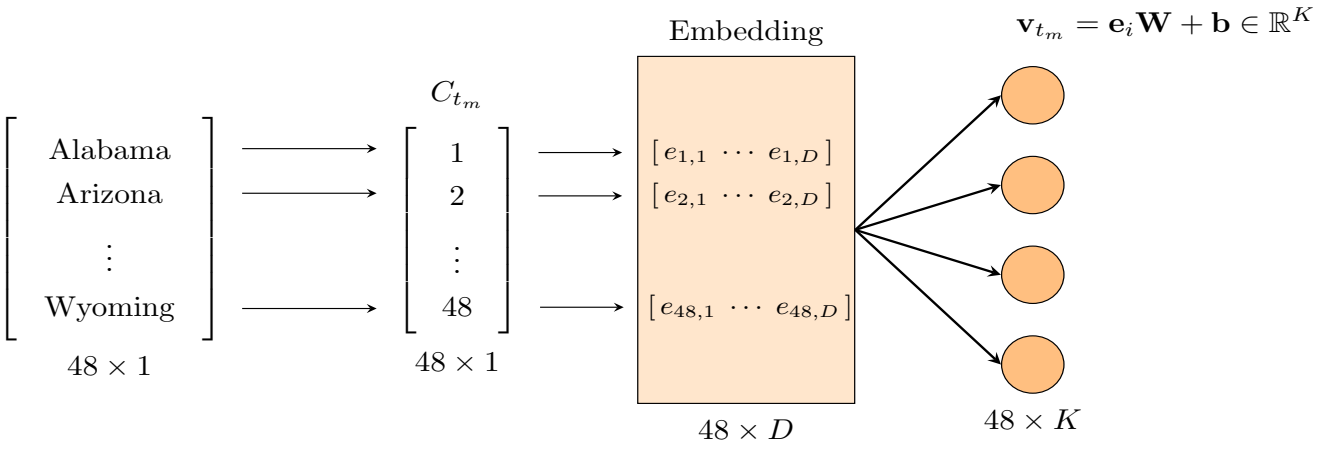
Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

A Neural Network Embedding

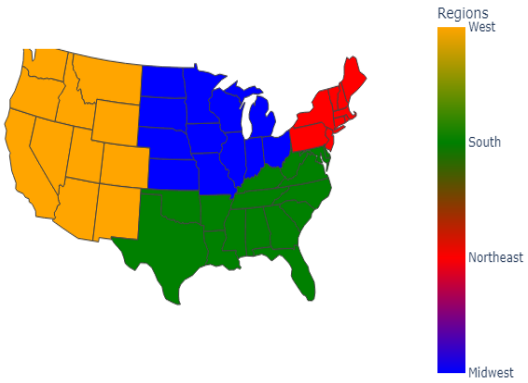
Figure 13: Embedding Layer



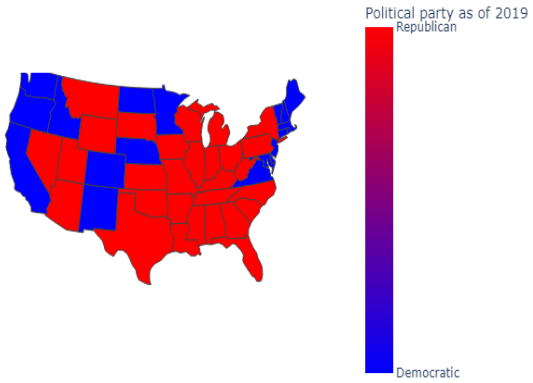
Note: This figure illustrates the embedding lookup process. Categorical input indices $i \in \{1, \ldots, 48\}$ are first mapped into dense vectors $\mathbf{e}_i \in \mathbb{R}^D$ via an embedding matrix $\mathbf{E} \in \mathbb{R}^{48 \times D}$, where D denotes the embedding dimension. In practice, the lookup is computed as a row selection or equivalently as $\mathbf{e}_i = \mathbf{1}_i^\top \mathbf{E}$, using a one-hot vector $\mathbf{1}_i \in \mathbb{R}^{48}$. The output vector is then passed through a dense layer with weights $\mathbf{W} \in \mathbb{R}^{D \times K}$ and bias $\mathbf{b} \in \mathbb{R}^K$, producing $\mathbf{v}_{t_m} = \mathbf{e}_i\mathbf{W} + \mathbf{b} \in \mathbb{R}^K$, where K is the number of output units.

B Clusters

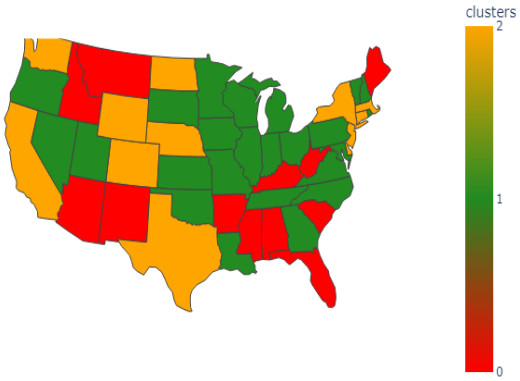
Figure 14: Clusters



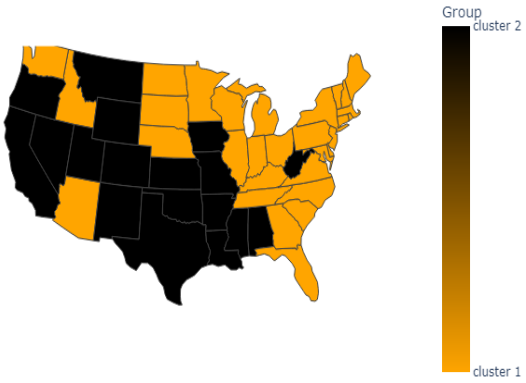
(a) Regional



(b) Political



(c) GDP



(d) Hierarchical

C Forecasting results

Figure 15: Forecasting: models' performance

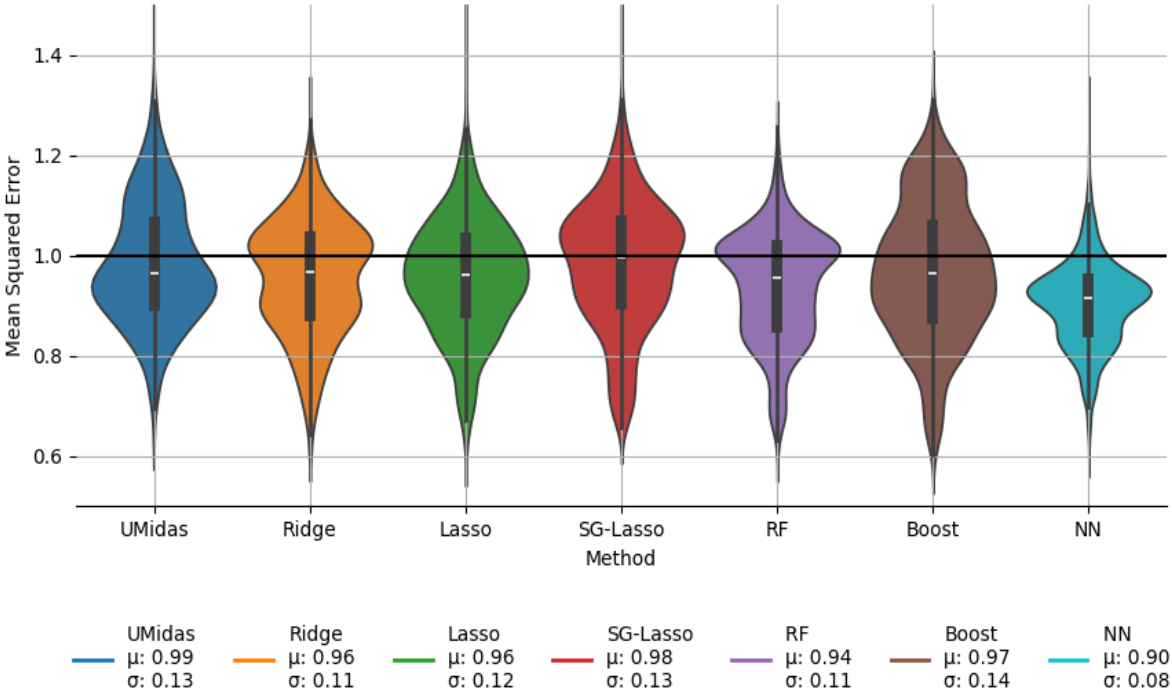


Figure 16: Forecasting: panel structure performance