Targeted Transformations for Macroeconomic Forecasting

KOF Working Papers

Targeted Transformations for Macroeconomic Forecasting

Maurizio Daniele, Philipp Kronenberg, and Tim Reinicke

No. 517, 06 / 2024

Targeted Transformations for Macroeconomic Forecasting*

Maurizio Daniele†, Philipp Kronenberg†, and Tim Reinicke†

†KOF Swiss Economic Institute, ETH Zurich

May 30, 2024

Abstract

The crisis periods of the past decades have highlighted the difficulty of forecasting economic indicators due to increased non-linearity and rapidly changing dynamics. To address this challenge, we introduce the Transform-Sparsify-Forecast (TSF) framework. The TSF framework first applies multiple transformations to each predictor to account for non-linear effects. It then dynamically applies a dimension reduction (targeting) to select the most relevant predictors from the high-dimensional dataset. Our approach is straightforward and can be seamlessly incorporated into existing forecasting frameworks. We demonstrate its versatility by applying it to five distinct classes of econometric and machine learning models. Using the FRED-MD dataset, we show that the TSF framework substantially improves the forecasting accuracy across all models, with notable gains at short forecast horizons and during periods of high economic uncertainties, attributable to the adaptive realignment of variable interactions and the ability to capture non-linearities. On short horizons, improvements are on average more than twice as high compared to standard methods without transformations. Our findings highlight the importance of the dataset composition and variable selection in effectively capturing evolving relationships and dynamics over time.

JEL Classification: C32, C38, C53, C55, E32, E37

Keywords: Data Transformations, High-dimensional Forecasting, Machine Learning, Targeted Predictors, Regularization

*The authors thank Philippe Goulet-Coulombe, Efrem Castelnuovo, the participants of the International Symposium of Forecasting (2023), and participants of the Macroeconomic Workshop at the University of Basel for useful discussions and comments.

1 Introduction

During times of elevated uncertainty, economic data often exhibit non-linear relationships and are subject to sudden changes. Forecasting with linear models might miss those dynamics, resulting in large prediction errors and misleading policy recommendations. Recent research on modern machine learning techniques, such as the factor-augmented random forest or artificial neural networks, see e.g., Coulombe et al. (2021), Medeiros et al. (2021), Hauzenberger et al. (2022) and Hauzenberger et al. (2023), highlights the superior predictive performance of non-linear models. Non-linear models excel in capturing complex relationships between predictors and target variables, which are crucial for forecasting key macroeconomic indicators and understanding the current and future state of the economy. However, implementing complex non-linear models can be challenging and computationally intensive. An alternative approach involves capturing non-linear relationships within linear model frameworks by expanding the model space through the inclusion of alternative transformations of the variables, as demonstrated e.g. in Bai and Ng (2008).

When forecasting aggregate time series, researchers often rely on theoretical economic reasoning when selecting the transformation of the data. One example is the FRED-MD dataset provided by McCracken and Ng (2016) that covers a comprehensive collection of macroeconomic and financial time series for the US with a recommended transformation for each variable. Nonetheless, relying solely on these recommended transformations may have limitations. The presumed optimal transformations may provide only a static perspective, disregarding the dynamic nature of the level of integration of the variables over time, as highlighted by McCracken and Ng (2020). Consequently, varying transformations might be required to maintain stationarity, underscoring the necessity for adaptive approaches. Moreover, the complexity of economic dynamics often entails non-linear relationships, which challenges the assumption of a single optimal transformation per variable for different regimes in time. Further, according to Koop and Potter (2000) the evolving relationship between variables, influenced by non-linearities over the business cycle, structural breaks, and outliers, challenges the assumption of a constant relationship over time.

We argue that there is no ex-ante optimal choice of the dataset. Rather, researchers should dynamically choose the relevant variables and transformations as relationships change over time. The dataset itself can then be seen as another hyperparameter to be optimized. Despite the potential benefits of using alternative transformations, determining the optimal transformations and selecting the right variables remain unresolved.

Instead of applying a single best transformation to each variable, resulting in a fixed dataset with predefined transformations, we propose to allow for multiple transformations per variable, exploiting the full potential of the data combined with a regularization approach to select the most informative transformations for prediction. We call our approach *Transform-Sparsify-Forecast (TSF)*. This paper aims to show that the TSF framework, by allowing multiple transformations per variable combined with a pre-shrinkage step on the inflated dataset before forecasting, effectively improves the forecast performance, especially of linear models, which may result from better capturing relevant dynamics, interactions, non-linearities or changing relationships.

The TSF approach unifies several components into a single framework that have individually demonstrated performance improvements, thereby enhancing forecasting accuracy beyond what each component achieves on its own. Recently, studies such as e.g., Medeiros et al. (2021) and Coulombe et al. (2021) demonstrate the benefits of including various transformations in the form of latent factors to capture different dynamics in the data. Our TSF framework goes beyond these approaches by incorporating a broader range of transformations, including levels, first and second differences, natural logarithms, first and second differences of the natural logarithm, and percentage changes. Further, Coulombe et al. (2021) highlight that the inclusion of moving averages and levels to the set of predictors helps to capture low-frequency dynamics with random forests.

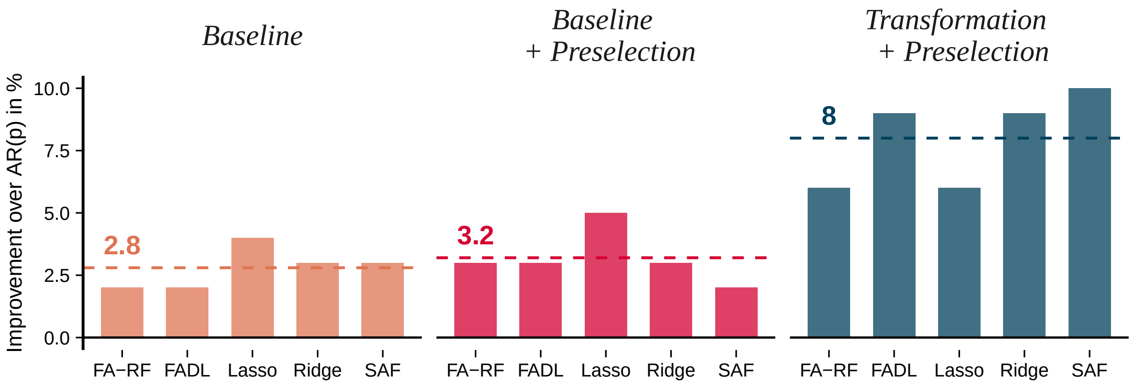
The TSF approach combines the use of multiple transformations with the concept of targeting predictors, as introduced by Bai and Ng (2008). This concept involves selecting a subset of the variables that carry the highest information and helps to mitigate issues related to overfitting and multicollinearity arising from the curse of dimensionality. They suggest that when data is excessively noisy, it might be advantageous to discard some of it to enhance the forecast performance. Their study demonstrates substantial benefits by applying targeted predictor methods for inflation, where only the most informative predictors are selected to estimate the factors. Similarly, Boivin and Ng (2006) show in the context of macroeconomic forecasting that the performance of the PCA estimator deteriorates as more uninformative series are included in the panel, emphasizing the merits of a preselection step. Giannone et al. (2021) provide evidence that macroeconomic datasets, are sparse, meaning only a subset of the whole dataset has explanatory power and a preselection of the most relevant predictors can improve the model performance. In line with this finding, various studies have shown forecast improvements with targeted predictors. For example Kotchoni et al. (2019); Kim and Swanson (2014, 2018) elaborate on the benefits of combining factor models and a lasso-based targeting, especially in turbulent times for forecasting macroeconomic key variables. Further Borup et al. (2023) applied a variable preselection on Random Forests, showing gains, especially for larger horizons. More recently, Ferrara and Simoni (2023) show that targeting predictors before ridge regression improves prediction performance using Google search data. Similarly, Borup et al. (2023) apply targeting to Google Trends data for predicting employment growth, employing methods such as random forest and bagging. Chinn et al. (2023) demonstrate the effectiveness of targeting in nowcasting world trade.

In this paper, we aim to bridge a significant gap in the literature by integrating multiple data transformations within a dynamic forecasting framework. This approach effectively allows the measurement of non-linear effects within a linear setting. Our method combines the strengths of targeting predictors with the flexibility of applying various transformations, thereby capturing complex non-linear relationships in economic data.

The TSF methodology is straightforward and can be seamlessly incorporated into existing forecasting frameworks utilized by central bankers, policymakers, and other practitioners, making it both feasible and practical. This approach addresses several critical challenges: It enhances forecasting accuracy using the full potential existing datasets, which can alleviate the necessity of collecting additional data. By increasing the dimensionality of the original dataset using multiple transformations our framework relies on a high-dimensional linear approximation of a potential non-linear true underlying model (see e.g. Belloni and Chernozhukov, 2011). Moreover, it avoids any model overfit by targeting the most relevant predictor variables. In addition, it highlights the importance of incorporating appropriate transformations, which can yield significant predictive gains. Third, it demonstrates that the optimal approach often involves applying multiple transformations simultaneously and employing a targeting strategy to identify the most effective predictor combination. By analyzing the composition of the dataset, we elaborate on the sources of the forecast improvements achieved through the TSF framework and identify circumstances in which its application may be particularly beneficial.

Furthermore, we contribute to the literature on sparse models by applying a two-stage dimension reduction procedure through the additional targeting step prior to the model estimation, which significantly enhances forecasting performance. Meinshausen (2007) and Belloni et al. (2013) show the theoretical and empirical superiority of a two-stage Lasso procedure over the regular Lasso approach. Sparse factor models, as analyzed in Luciani (2014), Kristensen (2017), Daniele et al. (2019), and Despois and Doz (2023),

Figure 1: Average Short-term Forecast Improvement over AR(p)



Notes: This figure presents the average RMSFE improvement over an AR(p) in percent, where p is optimized by BIC. The out-of-sample analysis spans from 2007 to 2022 for CPI, INDPRO, and UNRATE for one-step ahead forecasts. The x-axis categorizes the forecast models: Factor-augmented random forest (FA-RF), static factor model (FADL), Lasso, Ridge, and the sparse approximate factor (SAF) model. The Baseline utilizes the FRED-MD data; Baseline + Preselection dynamically selects 30 targeted variables via Lasso; Transformation + Preselection applies up to seven stationary transformations to each predictor and selects 60 targeted transformations via Lasso.

offer advantages in thoroughly analyzing and interpreting the underlying latent factors, however, they do not necessarily lead to more precise forecasts. In contrast, our framework shows that by incorporating alternative transformations and a predictor preselection step, the predictive accuracy of these models is markedly improved, making them competitive with other advanced machine-learning models.

To empirically assess the benefits of our TSF approach, we conduct a comprehensive pseudo-out-of-sample forecasting experiment using the FRED-MD dataset. We consider the growth rate of the Consumer Price Index (CPI), the growth rate of Industrial Production (INDPRO), and the change in the Unemployment Rate (UNRATE) as target variables and analyze the impact of the TSF framework on a large class of forecasting models that have shown to be among the best-performing model types, including factor, regularized, and machine-learning models. We compare the performance of the FRED-MD baseline dataset, with our augmented dataset featuring multiple transformations, along with a Lasso-based targeting, varying the fixed size of selected predictors to be 30 (as used by Bai and Ng (2008); Kotchoni et al. (2019)) or 60 (as used by Chinn et al.

(2023)).

The key findings of our study reveal several important aspects. First, the TSF framework substantially improves the forecasting accuracy across all model classes, with notable gains at short forecast horizons. As illustrated in Figure 1, our method shows an average root mean squared forecast error (RMSFE) improvement of eight percent over the standard AR(p) on short horizons, making it more than twice as effective as conventional targeting methods.[1] These improvements are particularly pronounced during times of economic instability, where the TSF method effectively captures dynamic changes and non-linear interactions in the data. Our results align with the findings of Goulet Coulombe et al. (2022) and Kim and Swanson (2014), who show that incorporating non-linear dynamics can enhance forecasting performance, especially for linear models during crisis periods. Second, the forecast gains can vary strongly between the target variables. Specifically, the gains are substantial for CPI and UNRATE, while the improvement for INDPRO is more moderate. Our analysis reveals that these outcomes stem from variations in dataset compositions. For target variables exhibiting substantial gains, the dataset composition shifts towards a higher proportion of variables related to the target variable. Conversely, for target variables with no significant gains, there is no notable shift in dataset composition, underscoring the importance of the variables included in the original dataset. Third, the TSF approach selects a more diverse range of variable transformations using often multiple transformations of the same variable, with reduced emphasis on log-differences and levels, and increased focus on alternative transformations.

The remainder of the paper is structured as follows. In Section 2, we introduce our forecasting framework and elaborate on the rationale behind incorporating multiple variable transformations in the set of predictors, as well as the preselection step for selecting the

[1] Figure C.1 in the Appendix provides details on the short-term forecast improvement of the TSF approach for each target variable.

most informative predictor variables through shrinkage. Section 3 provides details on our forecasting experiment, including the data description, the models used, and the forecasting setup. In Section 4, we present the forecasting results, while Section 5 provides an in-depth analysis of the factors leading to the forecast improvement. Finally, Section 6 concludes.

2 Forecasting Framework

In this section, we present our forecasting framework and provide details on each component within the forecasting process. Our methodology comprises three important parts. Initially, we elaborate on the role of data transformations in augmenting the model domain, particularly in the context of linear models, thereby improving their accuracy in predicting key macroeconomic and financial indicators. Subsequently, we discuss the necessity of incorporating shrinkage techniques for reducing the estimation complexity and circumventing a potential model over-parametrization caused by the curse of dimensionality. Lastly, we further condense the dimension of the problem by extracting the most valuable information from the data through latent factor models.

2.1 Econometric Model

We start with introducing our transform-sparsify forecasting (TSF) framework. Given a large set of N predictors X_t at period t, where $X_t = (X_{1,t}, \cdots, X_{N,t})'$, for $t = 1, \cdots, T$, we aim at predicting the target variable y_{t+h}, h-steps ahead based on the following general prediction model

$$y_{t+h} = f\left(X_t\right) + \epsilon_{t+h}, \tag{1}$$

where $f(\cdot)$ represents the functional form of a particular modeling strategy, which commonly encompasses linear or non-linear parametric formulations or a nonparametric representation.

Our TSF approach aims to enhance the predictive capabilities, especially of linear models, while preserving their inherent structure, thereby mitigating potential increases in computational complexity associated with structural modifications.

The predictive improvement is realized through several modeling steps. First, we largely expand the predictor space by incorporating diverse transformations for each of the N variables in X_t. In particular, we consider the following transformations: first and second differences, natural logarithm, first and second differences of natural logarithm, and the first difference of the percentage change. Given that the true underlying model f in (1) may exhibit a complex non-linear functional form, we rely on a high-dimensional linear approximation of the true model (see e.g., Belloni and Chernozhukov (2011), Caner and Eliaz (2024)). Our objective is to capture non-linear relationships within linear models by integrating multiple transformations of the predictors.

Second, we address potential issues of overfitting and multicollinearity that arise from the curse of dimensionality when dealing with a large set of predictors, where the number of predictors N may exceed the number of observations T. We achieve this by running a h-step ahead regression

$$y_{t+h} = X_t' \beta + \epsilon_{t+h}, \tag{2}$$

and employing the Lasso approach for estimating the model parameters β, identifying the most relevant predictors in X_t and substantially reducing the estimation error. Specifically, we solve the following optimization problem

$$\hat{\beta} = \arg\min_{\beta} \left[\sum_{t=1}^{T} (y_{t+h} - X_t'\beta)^2 + \lambda \sum_{i=1}^{N} |\beta_i| \right], \tag{3}$$

where in each out-of-sample forecasting step, we dynamically select the regularization parameter $\lambda > 0$ to obtain active set sizes close to either 30 or 60 variables. We opt for the soft-thresholding selection approach instead of hard-thresholding, as the latter, selecting predictors one at a time does not take the information from the entire set of predictors into account and may be prone to selecting similar predictors. Moreover, it is sensitive to small changes in the data due to the discrete nature of the decision rule.

Third, we utilize this enriched dataset featuring diverse transformations in a broad array of forecasting models, including modern machine learning models, regularization techniques, and dimension reduction approaches.[2]

In the following, we elaborate in more detail on the importance of each component of our TSF framework.

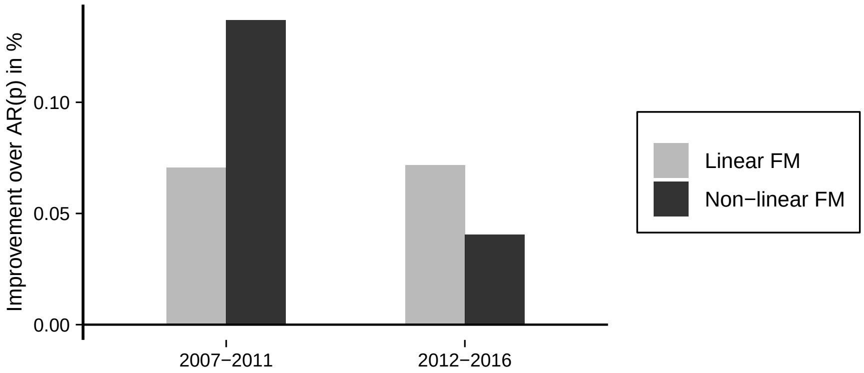
2.2 Data Transformations

By including multiple transformations of the predictors, we aim to capture non-linear relationships within a linear model, avoiding the need for significant alterations to the model structure, which potentially increases the complexity of model estimation. Specifically, in our framework we deviate from the common practice of applying one single best transformation[3] to each considered variable and instead consider multiple transformations per variable largely enhancing the space spanned by the dataset. Especially, for short-term forecasting horizons, Borup et al. (2023) shows that capturing non-linearities

[2]Appendix B provides a detailed overview of the model included in our empirical application.

[3]For the commonly used macroeconomic and financial dataset FRED-MD introduced by McCracken and Ng (2016), the authors recommend a specific transformation for each variable to which researchers usually rely on, potentially leading to a suboptimal choice of the variable transformation.

Figure 2: RMSFEs for Linear and Non-linear Factor Models



in the data is essential for the predictive accuracy.

In the following, we provide a short empirical demonstration that analyzes the merits of using non-linear transformations of the predictor variables for improving the predictive ability of traditional linear latent factor models. Specifically, we compute the one-step ahead forecast of a target variable x_i according to the following factor-augmented autoregressive process as introduced in Stock and Watson (2002)

$$x_{i,t+1} = \beta_{i,0} + \beta_i(L)x_{i,t} + \gamma_{i,1}\hat{F}_{t-1} + \epsilon_{i,t+1}, \tag{4}$$

where $\beta_{i,0}$ is an intercept, $\beta_i(L) = \beta_{i,1}L + \cdots + \beta_{i,l}L^l$ are lag polynomials corresponding to the variable x_i and $\hat{F}$ are the latent factors estimated based on principal component analysis (PCA). We select the number of lags l according to the Bayesian information criterion (BIC), where we consider a maximum of six lags, and include ten factors.

To analyze whether a polynomial transformation of a subset of the included predictor variables has a positive impact on the forecast accuracy, we consider the following non-linear factor-augmented autoregressive process

$$x_{i,t+1} = \beta_{i,0} + \beta_i(L)x_{i,t} + \gamma_{i,1}\hat{F}_{t-1} + \gamma_{i,2}\hat{F}^2_{t-1} + \epsilon_{i,t+1}, \tag{5}$$

where we extend the process in (4) by additionally incorporating the squared factors $\hat{F}^2$. Hence, the process in (5) keeps the linear relation between the target variable x_i and the predictors, however incorporates non-linearities through the squared factors.

Identical as in our empirical application presented in Section 3, we conduct a rolling window out-of-sample forecasting experiment using the FRED-MD dataset for the period January 1975 - December 2022,[4] where we aim at predicting the monthly CPI inflation rate. Figure 2 illustrates the average root mean squared forecast errors (RMSFE) of the linear and non-linear factor models relative to an autoregressive model of order p (AR(p)) used as a benchmark model, where p is selected based on the BIC. We analyze the average relative RMSFEs for two out-of-sample periods, where the first one covers the more turbulent period January 2007 - December 2011, which includes the recent Financial crisis, and the second one spans the rather calm period January 2012 - December 2016. The findings show that the linear and non-linear factor models outperform the benchmark model in both periods. Further, we see that the non-linear factor model leads to a higher forecast precision compared to the linear factor model during the period 2007-2011, which emphasizes the importance of incorporating non-linearities in the factor modeling, especially during more volatile periods. For the period 2012-2016, the linear factor model specification is slightly more precise compared to the non-linear factor model, indicating that during more calm periods including non-linearities in the factor modeling does not necessarily improve the predictive power.

Hence, the results from our real data illustration emphasize the potential benefit of including non-linear transformations in the predictor variables for improving the forecasting performance of linear models. Specifically, augmenting linear models with additional predictor transformations potentially leads to enhanced model flexibility, capturing a broader array of relationships between predictors and the response. However, the results

[4]The data description can be found in Section 3.1.

further show that it is important to choose the relevant transformation dynamically as the relations in the data can change over time.

2.3 Regularization and Dimension Reduction

While incorporating multiple transformations can indeed enhance the predictive capability of linear models, it introduces some challenges. One issue emerges from the inflated size of the dataset, where the number of variables can be much larger than the number of observations. In such a high-dimensional setting conventional estimation techniques like ordinary least squares (OLS) yield inconsistent estimates, and statistical inference suffers from a lack of degrees of freedom. Additionally, models may become over-parameterized, increasing the risk of overfitting and diminishing predictive performance. Furthermore, there is the potential risk of multicollinearity, especially prevalent in macroeconomic datasets where many variables depict similar dynamics, thus conveying redundant information. The inclusion of various transformations exacerbates this issue, as some transformations are highly relevant while others may not contribute substantially to model informativeness.

To address the challenges posed by the curse of dimensionality and multicollinearity, our framework employs a dynamic preselection to identify the most relevant predictors. We utilize the Lasso procedure for variable selection, significantly reducing the predictor space. Shrinkage techniques, like the Lasso, have the added benefit of amplifying the signal-to-noise ratio, leading to more accurate estimates and enhancing the predictive performance. It has been shown in the literature that a predictor preselection by selecting only the variables with the highest predictive power can significantly increase forecast accuracy (see e.g., Bai and Ng, 2008, Kotchoni et al., 2019, Ferrara and Simoni, 2023, Borup et al., 2023). This is especially important in datasets with variables with a high component of noise and with many highly correlated variables.

The need for preselecting the most relevant predictors is further justified asymptotically. Meinshausen (2007) shows that the convergence rate of the ordinary Lasso estimator significantly deteriorates when the number of variables N grows faster than the number of observations T.[5] Specifically, in the high-dimensional scenario when $\log(N) \sim T^{\zeta}$, where $0 < \zeta < 1$, the convergence rate of the lasso estimator can degrade to $\mathcal{O}(T^{-r})$, for $r < 1-\zeta$, contrasting the common rate of convergence of the Lasso of $\mathcal{O}(T^{-1})$. The quantity ζ determines the rate at which the logarithm of the number of predictors diverges, when T increases to infinity, lying between constant ($\zeta = 0$) and exponential ($\zeta = 1$) increase. Intuitively, this slow convergence occurs because a large regularization parameter is required to shrink the coefficients of irrelevant predictor variables towards zero. However, this also impacts the estimated coefficients of non-zero components, shrinking them towards zero as well, thereby leading to suboptimal predictions.

Meinshausen (2007) proposes a straightforward adjustment to the ordinary Lasso estimation technique to circumvent this issue. This modification involves employing the Lasso procedure twice. In the first step, the relevant predictors with non-zero coefficients are preselected using a relatively large regularization parameter. In the second step, the coefficients of the selected variables are re-estimated using a relaxed penalty. The convergence rate of the two-step lasso procedure is not affected by the presence of many irrelevant predictor variables and the coefficients of the non-zero variables can be estimated at the standard $\sqrt{T}$ rate.

Although we reduce the number of predictor variables to a large extent using our preselection procedure as previously described, the number of predictors can still be quite large and can be close to the number of observations. As part of our forecasting models used, we rely on factor modeling, to further reduce the number of predictors. Factor models are used to measure the information content of a large set of economic variables

[5]Note that Meinshausen (2007) conducts the analysis under an iid assumption.

by a much smaller number of latent factors. Thus, these models allow high-dimensional problems to be mapped into spaces of much lower dimension, reducing the number of parameters to be estimated and the corresponding estimation error.

Specifically, the static approximate factor model (AFM) as introduced in Chamberlain and Rothschild (1983) is defined as follows

$$x_{it} = \lambda_i' f_t + u_{it}, \tag{6}$$

where x_{it} denotes the data for the i-th variable at time t, for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. λ_i is a $(r \times 1)$-dimensional vector of unobserved factor loadings and f_t is the corresponding $(r \times 1)$-dimensional vector of latent factors, where r is the number of included factors. Finally, u_{it} denotes the idiosyncratic component, which can incorporate some weak temporal and cross-sectional correlations. We estimate the loadings and factors based on principal component analysis (PCA).

A common assumption in the traditional AFM is that the first r eigenvalues of the covariance matrix of the loadings $\Lambda\Lambda'$, where $\Lambda = (\lambda_1, \cdots, \lambda_N)'$ diverge with the number of variables in the system N. Intuitively, this means that all factors are strong and affect the entire set of time series. However, the strong factor assumption is usually not supported by empirical evidence. E.g., Ludvigson and Ng (2009) and Uematsu and Yamagata (2022) show that the spectral structure of macroeconomic and financial datasets is better captured by factor models, which incorporate weak factors, affecting only a subset of the considered time series.

For that reason, we extend our analysis by including the sparse approximate factor (SAF) model by Daniele et al. (2019). The SAF is a hybrid model that combines the structure of the AFM with regularization techniques. Specifically, the model incorporates sparsity via l_1-regularization on the factor loadings. In this way, the model shrinks elements of the factor loadings to zero and thus allows for measuring weak factors, reduces the number of parameters to be estimated, and therefore potentially diminishes the estimation noise.

3 Empirical Application

In this section, we provide details on the design of our empirical application, which refers to an out-of-sample forecasting experiment for predicting various key macroeconomic variables.

3.1 Data

To conduct the out-of-sample forecasting experiment, we utilize historical data from the United States, based on the FRED-MD dataset, introduced by McCracken and Ng (2016). The dataset includes observations from January 1975 to December 2022, resulting in an observation length of 576 months.[6] Moreover, it comprises 134 monthly macroeconomic and financial variables and covers a wide range of economic variables, including output, income, the labor market, housing, and price variables.

In the majority of cases, the models used for forecasting rely on a stationarity assumption on the variables under consideration, necessitating the transformation of each variable by a suitable predefined transformation to remove a potential unit root. The transformation is commonly determined through unit root tests and expert judgment. As we question the optimality of only using one predefined transformation, we transform each of the 134 variables by multiple transformations. More precisely, we consider all data transformations described in McCracken and Ng (2016) and apply each transformation to all variables in the dataset, regardless of their optimal transformation. Thus, each variable is included in level, first and second differences, natural logarithm, first and second dif-

[6] We disregarded the period before 1975, as more than 80% of the series had not yet commenced at that time.

ferences of natural logarithm, and first difference of the percentage change, resulting in 938 variables.

Before the model estimation, we remove any outliers[7] and verify the stationarity of each variable transformation, based on the augmented Dickey-Fuller test adjusted for the effective sample size and the Phillips-Perron test (Dickey and Fuller, 1979 and Phillips and Perron, 1988).[8] A variable or variable transformation is retained only if both tests reject the null hypothesis of a unit root. After removing the non-stationary transformations, the dataset comprises 652 variables. Finally, the data is standardized to have a sample mean of zero and a sample variance of one.

Our target variables are the growth rate of the Consumer Price Index (CPI), the growth rate of Industrial Production (INDPRO), and the change in the Unemployment Rate (UNRATE). Those target variables have high economic relevance and are analyzed in a large number of studies.[9]

3.2 Forecasting Setup

The sample used for the model estimation starts in January 1975 and ends in December 2006, comprising 32 years (384 months) of data. The out-of-sample evaluation is conducted as a rolling estimation window from January 2007 to December 2022, encompassing 16 years (192 months) and representing approximately one-third of the entire sample.[10] We compute direct forecasts over a $h = 1$, 6, and 12-month horizon. All

[7]Outliers are defined as values that are larger than ten times the interquartile range of the deviation to the absolute value of the median and are replaced for estimation with values generated by a cubic filter with seven lags moving average.

[8]For the Dickey-Fuller test, the optimal lag selection is chosen according to the Bayesian information criterion. For the Phillips-Perron test, the appropriate number of autocovariance lags to include in the Newey-West estimator of the long-run variance is chosen by lag $L_t = 4 * (T/100)^{1/4}$ with T being the length of the respective time series.

[9]See e.g., Stock and Watson (2002); Kotchoni et al. (2019); Coulombe et al. (2021); Borup et al. (2023) for studies analyzing the same target variables.

[10]In line with the prevailing view in the literature, highly influential observations during the COVID-19 pandemic are treated as outliers, for instance in Carriero et al. (2022), Antolín-Díaz et al. (2024), Schorfheide and Song (2021), Lenza and Primiceri (2022), or Baumeister and Hamilton (2023). Con-

models, including the entire set of hyperparameters, are re-estimated for each out-of-sample period, each target, and each forecast horizon. This approach allows us to follow a dynamic model selection strategy, taking into account the potential evolution of relationships between variables or the underlying data-generating process. The relevance of such a dynamic model selection approach has been demonstrated in several studies, including Kotchoni et al. (2019), who found that the optimal number of factors varies considerably across out-of-sample periods, target periods, and forecast horizons.

Prior to the estimation of each forecasting model, we select the most informative variables for predicting the respective target variable. As the optimal information set may vary over time, we conduct the variable selection for each OOS period. As outlined in Section 2.1, we employ the Lasso approach to identify the relevant predictors. For selecting the regularization parameter λ in (3), we follow the approach proposed by Bai and Ng (2008) and utilize a soft-thresholding rule such that a predefined fixed number of variables is obtained and the datasets are comparable. The soft-threshold selects 30 and 60 targeted predictors.[11] For the baseline FRED-MD dataset we refer to these fixed set specifications as targeted predictors, while for our large dataset encompassing all stationary transformations of each considered predictor, we denote them as targeted transformations.

In Table 1, we provide an overview of the models used in the forecasting experiment. The models are implemented according to the references and details mentioned. More details can be found in Appendix B.

To assess the predictive performance of the considered models across various periods, we evaluate our forecasts not only for the entire sample period, which spans from January 2007 to December 2022 but also within distinct economic episodes. Specifically, we evalu-

sequently, the four months from March 2020 to June 2020 have been excluded from the sample.

[11]The size of the fixed sets is chosen because 30 and 60 variables are commonly used in the literature. For example, Bai and Ng (2008) show that a soft-thresholding for the FRED-MD dataset to 30 variables results in forecast gains. Also Giannone et al. (2021) demonstrates that the optimal shrinkage for the FRED-MD dataset is approximately 30 variables. Chinn et al. (2023) use a dataset of 60 variables.

Table 1: Summary of Forecasting Models

Acronym	Model	Details
AR	Autoregressive Model	AR-lags selected by BIC
Lasso	Least Absolute Shrinkage and Selection Operator as in Tibshirani (1996)	Regularization parameter λ selected by CV
Ridge	Ridge Regression as in Hoerl and Kennard (1970)	Regularization parameter λ selected by CV
FADL	Factor-augmented Distributed Lag Model as in Stock and Watson (2002)	Number of factors chosen by the IR_a information criterion of Bai and Ng (2002) and AR-lags selected by BIC
SAF	Sparse Approximate Factor Model as in Daniele et al. (2019)	Regularization parameter μ selected by the adapted BIC type criterion
FA-RF	Factor-augmented Random Forest as in Medeiros et al. (2021)	Trees estimated via block bootstrapping First 4 PCs are included as regressors

Notes: The OOS CV is performed using 70% as training and 30% as the test set. The maximum number of AR-lags is set to six. The maximum number of factors is set to ten.

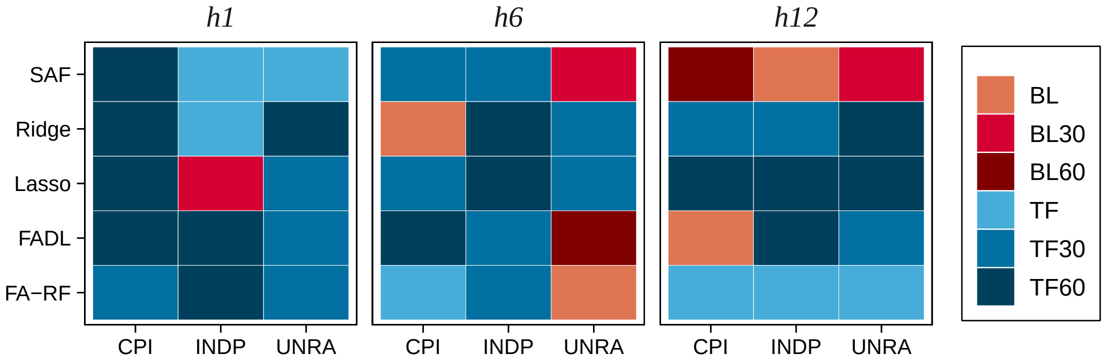
ate the models during periods of economic crisis, subsequent recovery phases, and periods excluding these events, separately. In the crisis sample, we incorporate all recessionary events identified according to the definition provided by the National Bureau of Economic Research (NBER) within our OOS period. This encompasses the Great Recession and the COVID-19 crisis. We define crisis periods as the three months preceding and the twelve months succeeding the commencement of the NBER recessionary phase.[12]

4 Forecasting Results

Figure 3 provides a graphical visualization of the strength of our forecasting framework. It uses a color-coded scheme to show the RMSFE minimizing specification - across different models, target variables, and forecast horizons. Blue highlights areas where transformation-based approaches outperform, while red indicates the superior performance of existing methods. The figure shows that the TSF approach consistently outperforms at short horizons with $h = 1$ (in 14 out of 15 cases) and maintains a significant advantage at $h = 6$ and $h = 12$ (both 11 out of 15 cases).

[12]This definition of crisis periods is in alignment with for example Goulet Coulombe et al. (2022).

Figure 3: Best-Performing Dataset Specification



Notes: The figure shows for each target variable, model, and forecast horizon the dataset specification that minimizes the root mean squared forecast error (RMSFE). Each heatmap illustrates the results for one forecast horizon (h1: one-month, h6: six-months, h12: twelve-months). The columns represent the different target variables and the rows the different models. The dataset specifications are BL: baseline, BL30: 30 targeted predictors, BL60: 60 targeted predictors, TF: transformations, TF30: 30 targeted transformations, TF60: 60 targeted transformations.

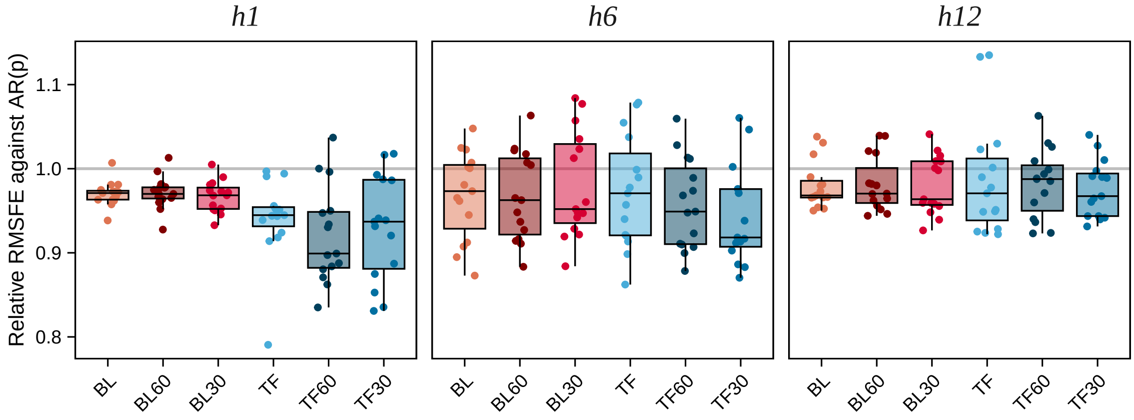
This chapter builds on the findings of Figure 3. First, we show that the most significant effects are observed at shorter forecast horizons. We then examine the strength of these effects in more detail by analyzing their distribution across different models and target variables. Finally, we show that these gains are particularly significant in times of crisis, underlining the robustness of our approach in difficult economic conditions.

4.1 Forecast Improvements on Short Horizons

To quantify the magnitude of the improvement offered by transformations and shrinkage, Figure 4 shows the relative RMSFE against an AR(p) model for different horizons across all target variables and models for each dataset specification in a box-plot.

Forecast improvements are largest for short horizons. For the shortest horizon, $h = 1$, all dataset specifications significantly outperform the AR(p) model. Notably, the transformation datasets (TF) show enhanced forecast gains; median improvements reach 5%, and increase to approximately 10% for TF60 and nearly 6% for TF30. This is consistent with the findings of Borup et al. (2023), who show that including squared and cubic transformations in the targeting step for Random Forests tends to improve

Figure 4: Boxplots of relative RMSFE across Dataset Specifications



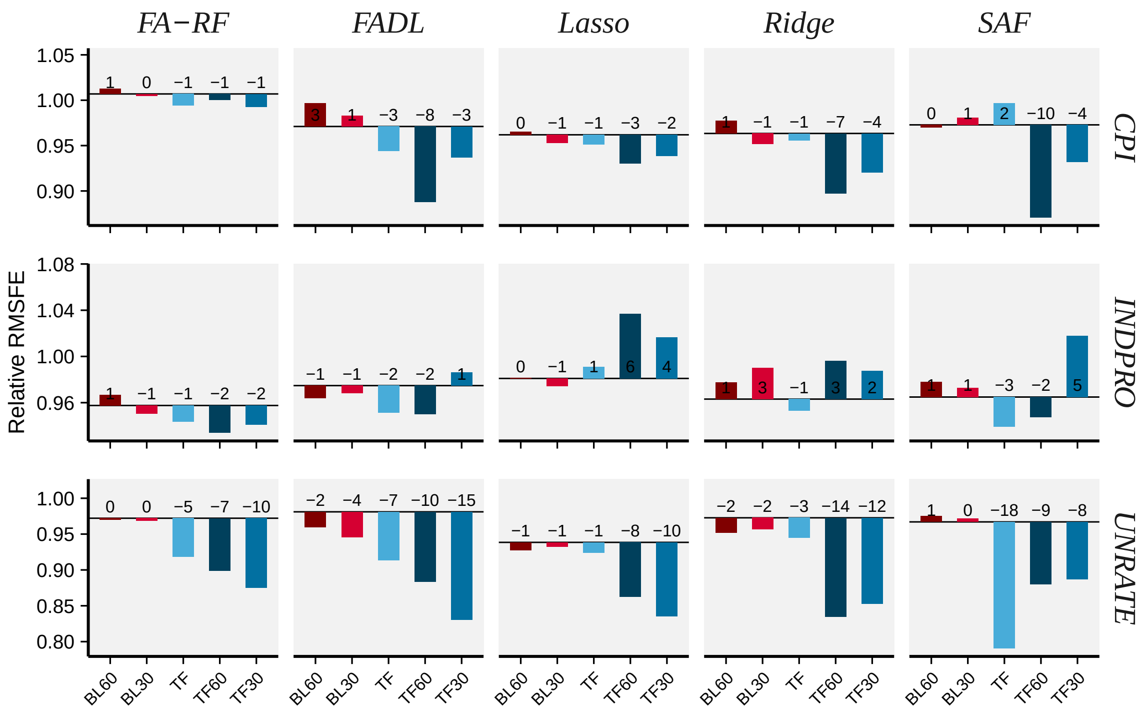
Notes: The figure shows the relative RMSFEs against an AR(p) model across models and targets for different horizons and dataset specifications. Each boxplot contains the RMSFEs of five models and three target variables resulting in 15 observations for each boxplot. The figure contains three subfigures for the different horizons (h1: one-month, h6: six-months, h12: twelve-months).

short-term forecast accuracy. Despite a slightly broader performance spread for TF30 and TF60, with a few observations underperforming or matching the AR(p) model, the majority demonstrates a superior performance.

For longer forecast horizons, the advantages of our TSF approach diminish. An increased variability in the forecast performance is notably observed for the six-month horizon as depicted in the middle boxplot of Figure 4. While TF30 and TF60 still demonstrate lower relative RMSFEs, affirming their sustained improvements over baseline models, the distinction between targeting methods lessens as we extend the horizon to twelve months.

These insights advocate for using transformations in macroeconomic forecasting. Given that the gains are most significant for short forecast horizons, and our preselection approach is optimized for this setting, as we do not incorporate longer lag structures, we concentrate on short-term forecasts with $h = 1$ in our subsequent analysis. To improve the robustness and enhance the performance at higher horizons, an extension could involve modifying the soft-thresholding to incorporate lags through grouping, as suggested by Babii et al. (2022).

Figure 5: Forecast Improvements over the Baseline Dataset across Dataset Specifications



Notes: The figure displays the relative RMSFEs of the baseline dataset against an AR(p) model in a dotted line. The bars indicate the absolute deviation in percentage points of the relative RMSFEs for the different datasets from the baseline. The forecast horizon is equal to one month.

4.2 Model Comparison

This section presents a quantitative analysis of the impact of data transformations and the dimension of the predictor set on the forecast accuracy. Figure 5 displays the relative forecast gains from various dataset specifications (BL60, BL30, TF, TF60, TF30) compared to the baseline dataset (BL) for short-term ($h = 1$) predictions. The figure also depicts the relative RMSFE of BL in comparison to the AR(p) benchmark model, represented by a horizontal line. Values below one indicate that models using the baseline dataset outperform the AR(p) model and the opposite for values above one. The bars represent the absolute differences in relative RMSFE between the dataset specifications and BL.

Selecting the right dataset is more critical than the model. In general, the relative RMSFE ratios presented in Figure 5 demonstrate the strongest forecast improvements for TF specifications. This pattern is consistent across different sizes of the preselected dataset sizes (30 and 60), highlighting the effectiveness of targeted transformations in enhancing forecast accuracy.

Selecting the right transformation and preselection size within the TF framework proves more crucial than the choice of forecasting model. Our findings reveal that, across various targets, a randomly selected model employing TF60 consistently outperforms the top-performing model using standard dataset specifications such as BL, BL60, and BL30. All models employing TF60 exhibit enhanced performance in forecasting CPI and UNRATE. However, the results for INDPRO are more dispersed, indicating that the effectiveness of the TF60 approach can vary depending on the specific economic target.

Building on these insights, Table 2 presents a detailed comparison of the model performances. It presents the relative RMSFEs of each model and dataset specification against the AR(p) benchmark and applies the Diebold-Mariano test (Diebold and Mariano (1995)) to assess the significance of differences in predictive accuracy between the AR(p) and each model. As detailed in Table 2, the TF60 specification consistently demonstrates the most significant improvements in RMSFE when compared to the AR(p) model, confirming the robustness of this approach.

This evidence highlights the crucial role of dataset selection over model selection in macroeconomic forecasting and supports a strategic emphasis on improving data quality and relevance through targeted transformations.

Linear models benefit the most from our TSF approach. The efficacy of our TSF approach varies significantly among different models. Especially, linear models benefit the most through the integration of multiple data transformations. The TSF effectively addresses the limitations of linear models in capturing non-linear data structures, thereby enhancing their predictive capabilities as detailed in Section 2.2. Sparse models such as

Table 2: Relative RMSFE against an AR(p) Model

	BL	BL30	BL60	TF	TF30	TF60
CPI						
FA-RF	1.007	1.005	1.013	0.994	**0.993**	1.000
FADL	0.971	0.983	0.997	0.944	0.937	**0.888***
Lasso	0.962	0.953	0.965	0.951	0.939	**0.930**
Ridge	0.963	0.952	0.977	0.956	0.921**	**0.897****
SAF	0.973	0.981	0.97	0.997	0.932	**0.871***
INDPRO						
FA-RF	0.958	0.950	0.967	0.944	0.941	**0.934***
FADL	0.975	0.968	0.964	0.951	0.986	**0.950**
Lasso	0.981	**0.974**	0.982	0.991	1.017	1.037
Ridge	0.963	0.990	0.977	**0.953**	0.987	0.996
SAF	0.965	0.973	0.978	**0.939**	1.018	0.947
UNRATE						
FA-RF	0.972	0.968	0.970	0.918**	0.875***	0.899**
FADL	0.981	0.946	0.960	0.914	**0.831****	0.884*
Lasso	0.938	0.933	0.928*	0.924**	**0.835****	0.862***
Ridge	0.973	0.957	0.952	0.945*	0.853***	**0.835****
SAF	0.967	0.972	0.975	**0.791***	0.887**	0.880**

Notes: This table presents the relative RMSFE against an AR(p) model for the one-month ahead forecast horizon. Significance levels indicate differences in forecast performance according to the one-sided Diebold-Mariano test against the AR(p) benchmark and are denoted by asterisks: *** (1% level), ** (5% level), and * (10% level).

the SAF, and the remaining shrinkage methods like Lasso and Ridge, significantly improve through the "double sparsity" obtained via the predictor preselection. The preselection improves the model performance by effectively reducing noise motivated by Meinshausen (2007). While the improvements for random forest models are generally modest, the FA-RF model still benefits from our TSF approach, especially for CPI. This is largely due to the noise reduction in the PCA factors used as predictors, following Medeiros et al. (2021), and an enhanced signal-to-noise ratio resulting from more robust trees through our preselection process, as indicated by Borup et al. (2023).

Improvements vary significantly across different target variables. The application of the TSF approach yields highly significant forecast improvements for UNRATE across all specifications with transformations. This stands in contrast to the BL specifications with targeted predictors, which show mostly minor non-significant gains. CPI also benefits substantially from the TSF approach, particularly with the TF60 and TF30 specifications where ridge regression and factor models demonstrate significant improvements in predictive performance. For instance, the SAF with TF60 records a gain of 9.1 percentage points over the optimal baseline model, underscoring the effectiveness of targeted transformations in forecasting CPI. Conversely, INDPRO presents a contrast to the pattern observed in UNRATE and CPI. The improvements for INDPRO are modest and not consistently significant, with some models even showing a deterioration in performance when using the TSF approach.

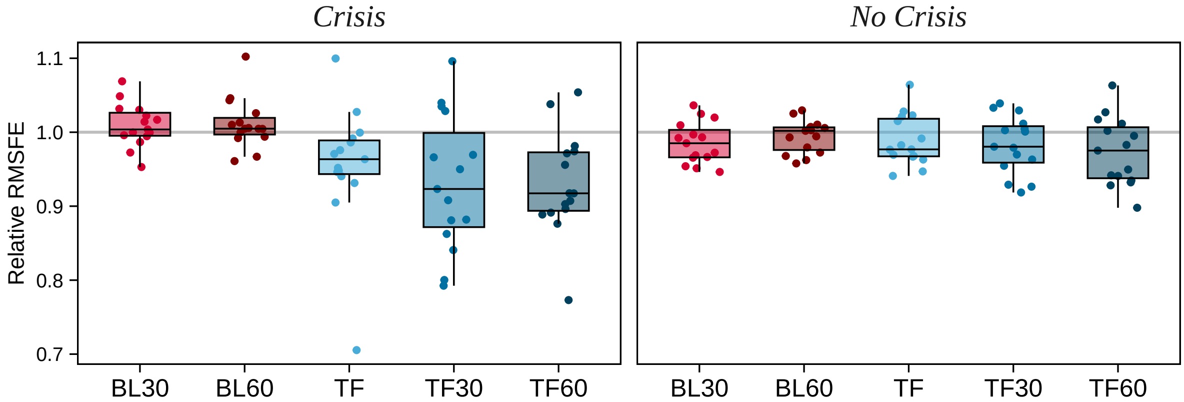
4.3 Timing - When do Transformations Matter?

As the forecast performance can be significantly influenced by specific periods, particularly those characterized by economic downturns and subsequent recovery, it is essential to analyze these periods separately. Consequently, results are presented for the crisis and non-crisis periods. The crisis periods are defined as the three months preceding and the twelve months following the official start of the recession period, as determined by the National Bureau of Economic Research (NBER). In our sample, this encompasses the period of the Great Recession, as well as the Covid-19 pandemic.

Figure 6 shows the relative RMSFEs for each model and target variable in comparison to its baseline specification, presented as a boxplot for the different dataset specifications, split into crisis and non-crisis periods. Consequently, each boxplot combines the forecasts from five different models across three target variables.

Forecast improvements are particularly substantial in turbulent times. Figure 6 illustrates that the BL30 and BL60 specifications perform similarly well as the BL specification, with a median around one and a relatively small dispersion, exhibiting a slight upward skew. In contrast, the TF, TF30, and TF60 specifications significantly outperform the BL, with a median relative RMSFE around 0.96 for TF, 0.9 for TF30, and TF60. This indicates that transformations offer a significant advantage in the con-

Figure 6: Boxplots of the relative RMSFE across Specifications against the Baseline Specification



Notes: The figure shows the relative RMSFEs of one-month ahead forecasts for each model and target variable against its baseline specification as a boxplot for the different dataset specifications split into crisis and non-crisis periods. Therefore, each boxplot combines the forecasts from five different models across three target variables.

text of crisis periods. During non-crisis periods, the figure shows almost no differences in forecast performance compared to the baseline dataset.

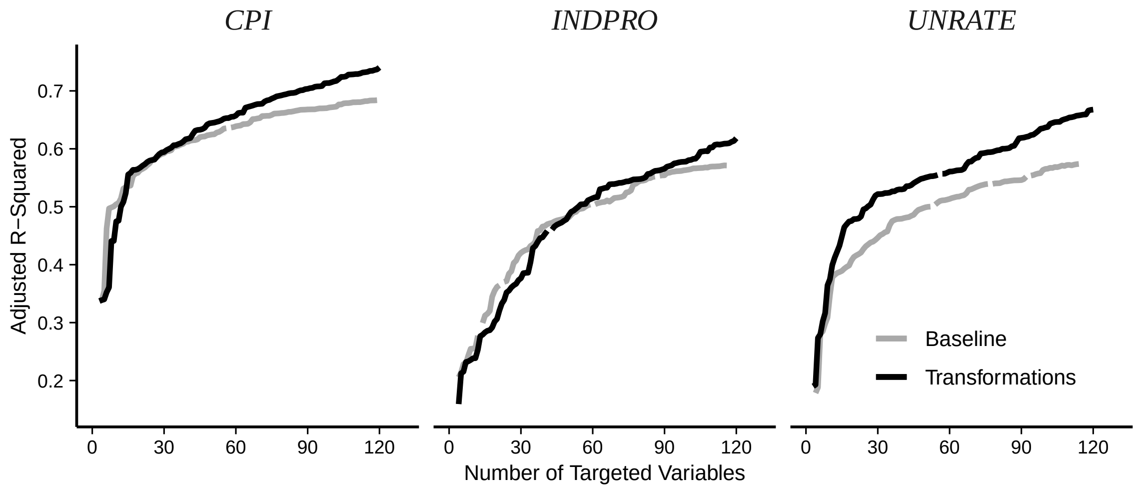
Table C.1 in Appendix C provides more details on the results displayed in Figure 6. Notably, only specifications using transformations show significant differences from the AR(p) model during crises.

5 Changes in Dataset Decomposition

Our previous analysis demonstrates a generally positive impact of the TSF approach on the forecast performance, particularly effective at short forecast horizons. This effect was pronounced for CPI and UNRATE, whereas the results for INDPRO were more heterogeneous.

In the following, we present empirical evidence supporting three key mechanisms by which the TSF enhances forecasting accuracy. First, the introduction of additional transformations generates more informative data, leading to an improved model performance. Second, these transformations enable a more effective sectoral composition, contributing

Figure 7: Adjusted R-squared by Number of Selected Variables



Notes: This figure shows the adjusted R-squared outcomes from one-step ahead regressions for varying numbers of predictor variables selected by the Lasso. The x-axis indicates the number of selected variables, while the y-axis displays the adjusted R-squared for each regression. When multiple values of λ result in the selection of the same number of variables, the λ yielding the minimum RMSFE is chosen. All data are standardized.

to the benefits observed. Third, the framework facilitates a more refined decomposition of transformations, selecting fewer logarithmic differences. Across these aspects, the changes are more pronounced in UNRATE and CPI, with less significant changes noted for INDPRO.

5.1 Explanatory Power of Active Set

Figure 7 presents the adjusted R-squared outcomes from one-step ahead regressions for a varying number of predictor variables selected by the Lasso approach. For each target variable and dataset specification, the regularization parameter in the Lasso selection is chosen to provide an active set size between 5 and 120 targeted predictors.

Transformed datasets deliver higher explanatory power compared to baseline sets. The figure illustrates the impact of the variable selection on the explanatory power of the predictor set across different sizes of selected variables for both dataset specifications (TF and BL). While differences between the two dataset specifications are minimal for a small number of selected variables, after a certain threshold depending on the target variable, the transformation specification yields a higher explanatory power compared to the baseline. For CPI, the explanatory power experiences a notable increase until approximately 20 predictor variables are included. Beyond this threshold, the marginal increase diminishes, however, the information gained from including additional variables is higher for the transformation dataset. Similarly, for the UNRATE the two specifications contain similar explanatory power until around 15 variables. After this threshold, the transformations specification encompasses another set of highly informative variables before the marginal utility of additional variables decreases such as in the baseline specification. This highlights the substantial explanatory value offered by these transformations. In contrast, for INDPRO, both the baseline and transformations specifications provide similar predictive power with a threshold at around 50 variables. Moreover, the marginal increase is only slightly higher for the transformations specification. Overall, the explanatory power of the dataset is higher for CPI and UNRATE than for INDPRO. This partly explains why forecast improvements for INDPRO are limited. The general higher explanatory power of the transformations dataset demonstrates the benefits of our transformation strategy using targeted transformations instead of solely relying on targeted predictors based on the baseline dataset. Moreover, it underlines the importance of additional transformations on the predictive accuracy after a certain number of variables is included.

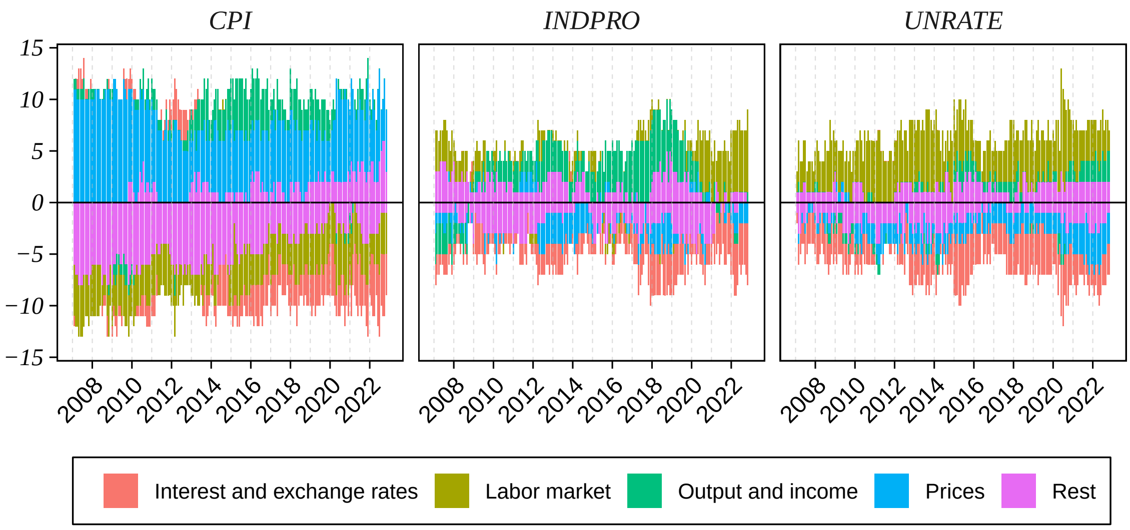
5.2 Sectoral Changes

Figure 8 illustrates the dynamically selected variables segmented by economic sectors, as defined in McCracken and Ng (2016). Subfigure 8(a) shows the shares in percentages of variables by category for the baseline specifications (BL), 30 targeted predictors (BL30) and 30 targeted transformations (TF30). Subfigure 8(b) depicts the absolute difference in selected variables by category between the TF30 and BL30 specifications over time.

Figure 8: Sectoral Changes
(a) Share by Dataset

	CPI			INDPRO			UNRATE		
	BL	BL30	TF30	BL	BL30	TF30	BL	BL30	TF30
Stock market	4	9.5	5.6	4	2.9	3.4	4	3.9	4.4
Prices	16	26.3	49.5	16	14.6	11.1	16	9.8	3.9
Output and income	13.6	7.8	13.5	13.6	9.6	16.5	13.6	6.9	8.6
Money and credit	10.4	6.1	7.4	10.4	8.6	6.6	10.4	7.9	7
Labor market	24	13.8	1.8	24	34.3	39.5	24	44.3	59.1
Interest and exchange rates	16.8	22	16.1	16.8	22.4	15.1	16.8	15.4	5.8
Housing	8	6.3	1.2	8	3	2.9	8	3	1.7
Consumption	7.2	8.1	4.9	7.2	4.7	4.9	7.2	8.8	9.6

(b) Change Over Time



Notes: This figure illustrates the dynamically selected variables segmented by sectors, as defined in McCracken and Ng (2016). Subfigure (a) shows the shares in percentages of variables by category for the baseline specification (BL), 30 targeted predictors (BL30), 30 targeted transformations (TF30). Subfigure (b) shows the absolute difference in selected variables by category between the TF30 and BL30 specifications over time.

Greater forecasting accuracy through targeting aligns with a heightened focus on key sectors. Our analysis of the sectoral composition within the active sets for each target variable reveals notable shifts that are closely related to the effectiveness of our TSF approach. In particular, for CPI, we observe a significant increase in the proportion of price-related variables within the active set, rising from 16% to approximately 50% following the application of targeted transformations. The increase in price-related variables is observed throughout the sample and is most pronounced during the Great Recession.

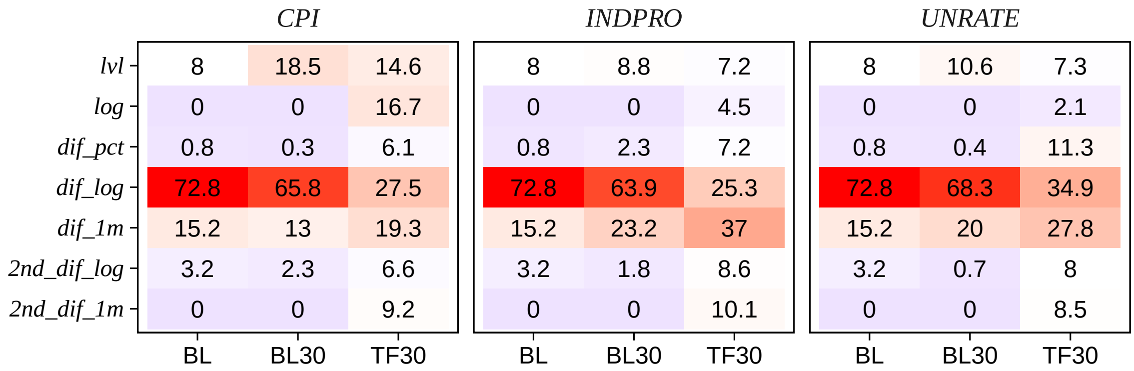
This significant adjustment promotes the capture of price dynamics, which explains the improvement in forecasting accuracy of the TF30 over the BL and BL30 specifications. A similar pattern is observed in the case of UNRATE, where the active set shows an increase from 24% to 59% in labor-related variables. This reflects a deeper incorporation of labor market dynamics, particularly noticeable after the beginning of the COVID-19 pandemic, which spiked in early 2020. This adjustment aligns the forecast models more closely with the economic circumstances that influenced the unemployment rate, which is crucial for capturing the fluctuations in the labor market. Conversely, INDPRO shows more modest changes in its active set, with an increase from around 14% to around 17% in the output and income variables. This suggests that the pre-existing transformations are already well aligned with the stable sectoral composition of industrial production. The observed increase in the output and income variables occurred mainly between the two crisis periods, while during the Great Recession and the COVID-19 pandemic, the labor-related variables received a higher weight. The only slight adjustments indicate that the TSF approach has not been able to bring much improvement, underscoring the importance of the variables included in the original dataset.

These observations not only demonstrate the tailored effectiveness of our data transformation and selection strategy across different sectors but also highlight the critical importance of tailoring transformation techniques to the unique economic characteristics of each indicator. By dynamically adjusting the variable composition within our models, we ensure that each model is optimized to capture the most relevant and meaningful economic signals, thereby improving overall forecasting accuracy.

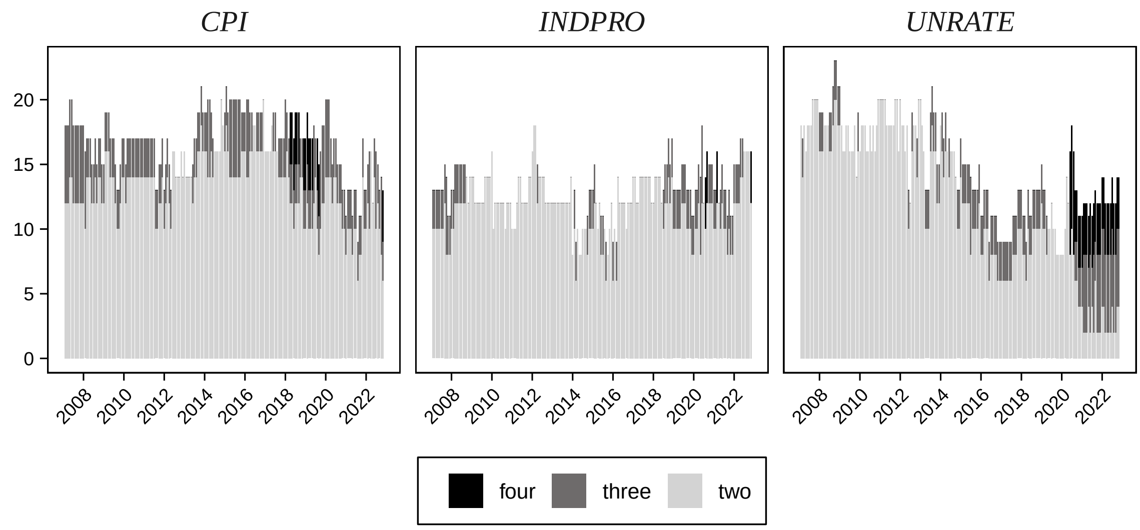
5.3 Transformation Changes

Figure 9 illustrates the dynamic variable selection segmented by each transformation as defined in McCracken and Ng (2016), including level (lvl), natural logarithm (log), first

Figure 9: Transformation Changes
(a) Share by Dataset



(b) Frequency of Selected Variables in TF30



Notes: This figure illustrates the dynamic variable selection segmented by transformation as defined in McCracken and Ng (2016), including level (lvl), natural logarithm (log), first and second difference (dif_1m, 2nd_dif_1m), first and second difference of natural logarithm (dif_log, 2nd_dif_log), and first difference of percentage change (dif_pct). Figure (a) depicts the proportions of variables in percentages resulting from the transformations for the baseline (BL), 30 targeted predictors (BL30), and 30 targeted transformations (TF30). Figure (b) illustrates the absolute number of variables included in the TF30 specification that are selected with more than one transformation.

and second difference (dif_1m, 2nd_dif_1m), first and second difference of the natural logarithm (dif_log, 2nd_dif_log), and first difference of percentage change (dif_pct). Figure 9(a) depicts the proportions of variables in percentages resulting from the BL, BL30, and TF30 datasets. Figure 9(b) illustrates the absolute number of variables included in the TF30 specification that are selected with more than one transformation.

A detailed examination of the transformation composition within the active sets for each target variable reveals a significant reduction in the reliance on the difference of the natural logarithm (dif_log) transformation, which is traditionally favored in macroeconomic analysis. For example, the proportion of dif_log transformations decreased substantially from BL30 to TF30 across all target variables. This is evident for CPI, with the proportion of dif_log transformations decreasing from 66% to 27.5%, for INDPRO from 64% to 25%, and for UNRATE from 68% to 35%. This shift towards a more diversified set of transformations, including a higher share of variables in first and second differences, first differences of percentage change, and natural logarithms improves the responsiveness to both sudden economic changes and longer-term trends.

A closer look at the selected variables for the TF30 specification shows that a large proportion of variables are selected in multiple versions but in different transformations over the whole time horizon. The proportion varies over time, but for CPI up to 20 out of 30 variables appear in multiple transformations. For UNRATE the proportion is similarly high, but then decreases over time as fewer different variables but more transformations of the same variable are selected. This is particularly noticeable after the COVID-19 pandemic, which led to a strong shift in the selection of variables. In this way, the TSF approach places more emphasis on highly explanatory variables and exploits the non-linear dynamics of these variables. For INDPRO, there are generally fewer variables with multiple transformations and these rarely occur more than twice. This broader and more diverse range of transformations augments the modeling capacities, particularly of linear models, enabling them to better capture potential non-linear dynamics in the variables, thereby improving overall forecasting accuracy.

6 Conclusion

In this paper, we address the challenge of optimizing dataset specifications for macroeconomic forecasting in a data-rich environment. Traditional econometric models often struggle to capture the non-linear dynamics and sudden changes inherent in economic data, leading to significant forecasting errors and potentially misleading policy recommendations. Instead of relying on complex non-linear models, we suggest to capture non-linear interrelations by expanding the model space through the inclusion of alternative transformations. We introduce the Transform-Sparsify-Forecast (TSF) approach, which allows for a dynamic selection of relevant variables and transformations, thereby maximizing the information extracted from the data.

Through a comprehensive pseudo-out-of-sample forecasting experiment using the FRED-MD dataset, we illustrate the effectiveness of our TSF approach. Our results highlight the superiority of targeted transformations based on the TSF over traditional methods including targeted predictors with preselected variables. We observe significant forecasting gains, especially for short horizons and during crisis periods. Furthermore, our analysis highlights the importance of the dataset composition and variable selection for forecasting performance. The TSF approach not only selects a more diverse range of variable transformations but also emphasizes the inclusion of multiple transformations of the same variable, underlining the relevance of capturing non-linear dynamics, leading to improved forecast accuracy.

Overall, our findings demonstrate the value of incorporating multiple steps, including transformations and preshrinkage, in the context of a data-rich environment. By providing a framework that dynamically adapts to changing relationships in economic data, our TSF approach offers a promising way to improve economic forecasting and policy analysis where the optimal choice of the data is unclear.

References

Antolín-Díaz, J., Drechsel, T., and Petrella, I. (2024). Advances in nowcasting economic activity: The role of heterogeneous dynamics and fat tails. *Journal of Econometrics*, 238(2):105634.

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.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Baumeister, C. and Hamilton, J. D. (2023). A full-information approach to granular instrumental variables. Working paper, UCSD.

Belloni, A. and Chernozhukov, V. (2011). *High Dimensional Sparse Econometric Models: An Introduction*, pages 121–156. Springer Berlin Heidelberg.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Boivin, J. and Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, 132(1):169–194.

Borup, D., Christensen, B. J., Mühlbach, N. S., and Nielsen, M. S. (2023). Targeting predictors in random forest regression. *International Journal of Forecasting*, 39(2):841–868.

Breiman, L. (2001). Random forest. *Machine Learning*, 45(1):5–32.

Caner, M. and Eliaz, K. (2024). Should humans lie to machines? the incentive compatibility of lasso and glm structured sparsity estimators. *Journal of Business & Economic Statistics*, Forthcoming.

Carriero, A., Clark, T. E., Marcellino, M., and Mertens, E. (2022). Addressing COVID-19 Outliers in BVARs with Stochastic Volatility. *The Review of Economics and Statistics*, pages 1–38.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Chinn, M. D., Meunier, B., and Stumpner, S. (2023). Nowcasting world trade with machine learning: a three-step approach. Working Paper 31419, National Bureau of Economic Research.

Coulombe, P. G., Leroux, M., Stevanovic, D., and Surprenant, S. (2021). Macroeconomic data transformations matter. *International Journal of Forecasting*, 37(4):1338–1354.

Daniele, M., Pohlmeier, W., and Zagidullina, A. (2019). Sparse approximate factor estimation for high-dimensional covariance matrices. *arXiv preprint arXiv:1906.05545*.

Despois, T. and Doz, C. (2023). Identifying and interpreting the factors in factor models via sparsity: Different approaches. *Journal of Applied Econometrics*, 38(4):533–555.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Ferrara, L. and Simoni, A. (2023). When are google data useful to nowcast gdp? an approach via preselection and shrinkage. *Journal of Business & Economic Statistics*, 41(4):1188–1202.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437.

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.

Hauzenberger, N., Huber, F., and Klieber, K. (2023). Real-time inflation forecasting using non-linear dimension reduction techniques. *International Journal of Forecasting*, 39(2):901–921.

Hauzenberger, N., Huber, F., Klieber, K., and Marcellino, M. (2022). Enhanced bayesian neural networks for macroeconomics and finance. *arXiv preprint arXiv:2211.04752*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Kim, H. H. and Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, 178:352–367.

Kim, H. H. and Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, 34(2):339–354.

Koop, G. and Potter, S. M. (2000). Nonlinearity, structural breaks or outliers in economic time series. *Nonlinear econometric modeling in time series analysis*, pages 61–78.

Kotchoni, R., Leroux, M., and Stevanovic, D. (2019). Macroeconomic forecast accuracy in a data-rich environment. *Journal of Applied Econometrics*, 34(7):1050–1072.

Kristensen, J. T. (2017). Diffusion indexes with sparse loadings. *Journal of Business & Economic Statistics*, 35(3):434–451.

Lenza, M. and Primiceri, G. E. (2022). How to estimate a vector autoregression after march 2020. *Journal of Applied Econometrics*, 37(4):688–699.

Luciani, M. (2014). Forecasting with approximate dynamic factor models: The role of non-pervasive shocks. *International Journal of Forecasting*, 30(1):20–29.

Ludvigson, S. C. and Ng, S. (2009). Macro factors in bond risk premia. *The Review of Financial Studies*, 22(12):5027–5067.

McCracken, M. and Ng, S. (2020). Fred-qd: A quarterly database for macroeconomic research. Working paper 26872, National Bureau of Economic Research.

McCracken, M. W. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., and Zilberman, E. (2021). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393.

Phillips, P. C. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

Schorfheide, F. and Song, D. (2021). Real-time forecasting with a (standard) mixed-frequency var during a pandemic. Working paper 29535, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Uematsu, Y. and Yamagata, T. (2022). Estimation of sparsity-induced weak factor models. *Journal of Business & Economic Statistics*, 41(1):213–227.

A Transform-Sparsify-Forecast Algorithm

In the following, we summarize our transform-sparsify-forecast procedure.

Let T denote the total number of observations and m be the number of out-of-sample periods.

At each out-of-sample step $j = 1, \ldots, m$:

1. **Transform**

   Let $\tilde{X}_i = (X_{i,j}, \ldots, X_{i,T-m-1+j})$, denote the i-th predictor variable for the observations in the estimation window. For each predictor variable X_i, where $i = 1, \ldots, N$:

   1.1. Compute the following transformations for $\tilde{X}_i$: first and second differences, natural logarithm, first and second differences of the natural logarithm, and the first difference of the percentage change.

   1.2. Check the stationarity of each variable transformation based on the augmented Dickey-Fuller test and the Phillips-Perron test. If both tests reject the null hypothesis of a unit root, keep the transformation, otherwise drop the transformation.

2. **Sparsify**

   Let W denote the set of stationary transformations of all predictor variables. For each target variable y_k, for $k = 1, \ldots, K$:

   2.1. Compute the h-step ahead l_1-norm penalized regression of $y_{k,t}$ on W_{t-h}, where $y_{k,t}$ is the k-th target variable at period t. Specifically, run the regression

   $$y_{t+h} = W_t' \beta + \epsilon_{t+h},$$

   and employ the Lasso approach for estimating the model parameters β by solving the optimization problem

$$\hat{\beta} = \arg\min_{\beta} \left[\sum_{t=1}^{T} (y_{t+h} - W_t'\beta)^2 + \lambda \sum_{i=1}^{N} |\beta_i| \right].$$

2.2. Perform a preselection of the most informative predictor variables according to a soft thresholding rule, leading to an active set size of either 30 or 60 variables, i.e., select the 30 or 60 predictor variables in W with the largest coefficients in the Lasso regression.

2.3. Collect the selected variables in the matrix $\tilde{W}$.

3. **Forecast**

   Let $\tilde{t} = T - m - 1 + j$ denote the last insample observation. For each forecasting model $s = 1, \ldots, S$:

   Compute the h period ahead forecast $\hat{y}^{h,s}_{k,\tilde{t}+h|\tilde{t}} = \hat{f}_s\left(W_{\tilde{t}}\right)$, where $\hat{f}_s$ denotes the functional form of model s estimated on the insample period.

B Models

In the following, we provide an overview of the models used in the forecasting experiment.

Autoregressive Model

As an univariate benchmark model, we use an autoregressive (AR) model of order p

$$X_{i,t+h} = \beta_0 + \beta_i(L)X_{i,t} + \epsilon_{i,t+h}, \tag{7}$$

where $X_{i,t}$ represents the observation of the target variable at period t, β_0 is an intercept, $\beta_i(L) = \beta_{i1}L + \ldots + \beta_{ip}L^p$ are lag polynomials corresponding to the data $X_{i,t}$ and p is selected based on the Bayesian information criterion assuming a maximum lag order of six.

Penalized Regression Models

In our comparative forecasting study, we include two commonly used shrinkage estimators for the following h-step ahead linear regression model

$$y_{t+h} = X_t'\beta + \epsilon_{t+h},$$

where y_t is the target variable, X_t is a potentially high-dimensional set of N predictors and β is the N-dimensional vector of coefficients.

The general penalized optimization problem to estimate β can be expressed as follows

$$\hat{\beta} = \arg\min_{\beta} \left[\sum_{t=1}^{T} (y_{t+h} - X_t'\beta)^2 + \lambda \sum_{i=1}^{N} ((1-\alpha)\beta_i^2 + \alpha|\beta_i|) \right], \tag{8}$$

where $\lambda > 0$ represents the regularization parameter.

The first penalized regression model that we consider is the **Ridge Regression (Ridge)** model proposed by Hoerl and Kennard (1970), represented by a least squares regression with a l_2-norm regularization, which is obtained by setting $\alpha = 0$ in (8). Its goal is to prevent overfitting arising from a large set of predictors by shrinking large coefficients towards zero.

Furthermore, we incorporate the **Least Absolute Shrinkage and Selection Operator (Lasso)** proposed by Tibshirani (1996), which is represented by a l_1-norm regularization, obtained by setting $\alpha = 1$. In contrast to Ridge, the Lasso approach encourages sparsity in the model by forcing some coefficients to be exactly zero.

The predictive performance of both models hinges on an optimal selection of the regularization parameter denoted by λ. In our framework, we employ an out-of-sample cross-validation procedure to estimate λ. At any prediction date t, we partition the available data into a training sample, comprising the first 70% of the observations and the successive 30% of the observations constitute the validation sample. We estimate the models on the training sample for different values of λ and compute h-step ahead direct forecasts for each observation in the validation sample using a rolling window setup. To alleviate the computational burden, we re-estimate the models only every five periods. Finally, we select the optimal λ, which minimizes the out-of-sample root mean squared forecast error.

Factor Model (FADL)

Factor models are commonly used models for forecasting in data-rich environments. We apply the factor-augmented autoregressive distributed lag model (FADL), following the example of Stock and Watson (2002), including next to the factors, additionally the lagged target variables, as outlined in equation (4).

The factors $\hat{F}_t$ are extracted from the data set by PCA, and the optimal number of factors $\hat{r}$ is chosen by the IR_a information criteria introduced in Bai and Ng (2002) with a maximum of ten factors. The number of included lags of the factors and the target variable is chosen by the BIC and is restricted to a maximum of six lags.

Sparse Approximate Factor Model (SAF)

To assess the effect of incorporating sparsity in the loadings matrix of the approximate factor model, which allows for measuring weak factors, on the predictive accuracy of the factor model, we incorporate the sparse approximate factor model of Daniele et al. (2019). The SAF is defined based on the following static factor model equation, for $i = 1, \ldots, N; t = 1, \ldots, T$,

$$x_{it} = \lambda_i' f_t + u_{it},$$

and shrinks each element of the loadings matrix λ_i according to the l_1-norm penalized maximum likelihood problem

$$\hat{\Lambda}, \Phi_u \in \underset{\Lambda, \Phi_u}{\arg\min} \left[\log |\Lambda\Lambda' + \Phi_u| + \operatorname{tr} \left[S_x \left(\Lambda\Lambda' + \Phi_u\right)^{-1} \right] + \mu \sum_{i=1}^{N} \sum_{k=1}^{r} |\lambda_i k| \right],$$

where $\Lambda = (\lambda_1, \cdots, \lambda_N)'$, Φ_u is the idiosyncratic error covariance matrix, S_x is the sample covariance matrix and $\mu > 0$ denotes the regularization parameter. As suggested in Daniele et al. (2019), we determine μ based on the minimization of the following adapted BIC type criterion

$$IC(\mu) = 2 \log \left| \hat{\Lambda}\hat{\Lambda}' + \hat{\Phi}_u \right| + \operatorname{tr} \left[S_x \left(\hat{\Lambda}\hat{\Lambda}' + \hat{\Phi}_u \right)^{-1} \right] + \frac{\kappa_\mu}{4} \sqrt{\frac{\log N}{N} + \frac{\log N}{N \cdot T}}, \tag{9}$$

where κ_μ denotes the number of non-zero elements in the estimated factor loadings matrix $\hat{\Lambda}$ for a given value of μ. To select the optimal μ, we estimate the criterion in (9) for a grid of different values for μ and choose the one that minimizes the information criterion.

Factor-Augmented Random Forest (FA-RF)

The random forest is an ensemble learning technique proposed by Breiman (2001) that uses bootstrap aggregation (bagging) to combine the predictions of multiple regression trees. Each regression tree is considered a weak learner and provides one prediction. The final prediction of the random forest is obtained by a mean aggregation of the individual predictions of the regression trees. We follow the specification of Medeiros et al. (2021) which can be described as follows:[1]

$$\hat{X}_{i,t+h} = \frac{1}{B} \sum_{b=1}^{B} \left[\sum_{k=1}^{K_b} \hat{c}_{k,b} \mathrm{I}_{k,b} \left(\mathbf{X}_t; \hat{\boldsymbol{\theta}}_{k,b} \right) \right], \tag{10}$$

[1]They included a dummy for the year 2017 which we did not include.

where B denotes the total number of regression trees[2], K is the number of terminal nodes that are created by the splits, $c_{k,b}$ is a tree- and node-specific constant which is the sample mean of the observations of the target variable that are in this node. Each regression tree is constructed by recursively splitting the data into subsets based on the predictor variables. The tree aims to create binary splits at each node to partition the data as follows:

$$\mathrm{I}_k\left(\mathbf{X}_t;\boldsymbol{\theta}_k\right)=\begin{cases}1 & \text{if } \boldsymbol{X}_t \in R_k\left(\hat{\boldsymbol{\theta}}_{k,b}\right)\\ 0 & \text{otherwise}\end{cases} \tag{11}$$

where I is an indicator function that maps the observations of the predictors according to the set of parameters θ_k that define the nodes. The optimal variables and split points are selected to minimize the mean squared error between the observed target variable values and the predicted values.[3] Single trees are estimated by block bootstrapping. We include the first four principal components of the data set as predictors following Medeiros et al. (2021), who show that this factor-augmented random forest performs superior to standard random forests.

C Additional Tables

[2]We create an ensemble of 500 regression trees.

[3]All variables are considered at each split when growing a tree and the minimum size of a terminal node is equal to 5.

Table C.1: Relative RMSFE against an AR(p) Model by Subperiods

	BL	BL30	BL60	TF	TF30	TF60
A. Crisis Periods						
CPI						
FA-RF	0.998	1.002	1.003	0.994	**0.965**	0.970
FADL	0.921	0.991	1.006	0.907	0.818	**0.815**
Lasso	0.909	0.931	0.914	0.858*	**0.811***	0.836*
Ridge	0.92	0.942	0.945	0.882*	0.855*	**0.848****
SAF	0.914	0.972	0.963	1.019	0.872	**0.812**
INDPRO						
FA-RF	0.953	0.947	0.967	**0.899***	0.92	0.906*
FADL	0.941	0.946	**0.933**	0.925	0.958	0.909
Lasso	0.947	**0.950**	0.952	0.972	0.985	0.998
Ridge	0.946	0.977	0.957	**0.918**	0.98	0.982
SAF	0.939	0.936	0.947	**0.901**	1.025	0.916
UNRATE						
FA-RF	0.968	0.959	0.963	**0.897***	0.848**	0.873**
FADL	1.00	0.963	0.960	0.885	**0.770****	0.856
Lasso	0.946	0.937	0.934	0.912*	**0.779****	0.836**
Ridge	1.009	0.972	0.970	0.955	0.808**	**0.785****
SAF	0.950	1.004	0.992	**0.668**	0.852*	0.844*
B. Non-Crisis Periods						
	BL	BL30	BL60	TF	TF30	TF60
CPI						
FA-RF	1.015	1.008	1.022	**0.994**	1.018	1.027
FADL	1.015	0.975	0.988	0.977	1.035	**0.950**
Lasso	1.008	**0.973**	1.010	1.03	1.043	1.01
Ridge	1.002	0.961	1.006	1.019	0.977	**0.941****
SAF	1.024	0.990	0.976	0.976	0.984	**0.922****
INDPRO						
FA-RF	0.969	**0.959**	0.966	1.043	0.990	0.997
FADL	1.051	1.018	1.033	**1.012**	1.050	1.042
Lasso	1.057	**1.030**	1.050	1.035	1.089	1.126
Ridge	**1.003**	1.019	1.024	1.031	1.005	1.030
SAF	1.025	1.056	1.05	1.025	**1.00**	1.019
UNRATE						
FA-RF	0.979	0.983	0.980	0.950**	**0.915*****	0.938**
FADL	0.951	0.918**	0.960	0.956	**0.917****	0.926**
Lasso	0.926*	0.927*	0.918**	0.942*	0.915***	**0.901****
Ridge	0.914**	0.932	0.923**	0.928**	0.917***	**0.906****
SAF	0.993	**0.921****	0.948	0.948	0.938**	0.934*

Notes: See Table 2. The crisis periods are defined as the three months preceding and the twelve months following the official start of the recession period, as determined by the National Bureau of Economic Research (NBER).

Table C.2: Relative RMSFE $h = 6$

	BL	BL30	BL60	TF	TF30	TF60
CPI						
FA-RF	0.873***	0.884***	0.883***	**0.862*** **	0.903***	0.907**
FADL	1.007	1.013	1.017	0.990	0.976	**0.968**
Lasso	0.907***	0.960	0.916***	0.898*	**0.870***	0.878*
Ridge	**0.895*** **	0.929***	0.911***	0.913***	0.912	0.911***
SAF	1.025	0.947**	1.007	0.999	**0.938*** **	0.974*
INDPRO						
FA-RF	0.945**	0.952*	0.963	0.957	**0.918** **	0.923**
FADL	1.048	1.057	1.024	1.079	**1.002**	1.012
Lasso	1.002	1.035	1.022	0.940*	0.913*	**0.900***
Ridge	1.001	1.023	1.004	0.978*	0.917**	**0.910** **
SAF	1.023	1.084	1.063	1.037	**0.971**	1.028
UNRATE						
FA-RF	**0.912** **	0.919**	0.914**	0.921	0.976	0.989
FADL	0.973	1.077	**0.965**	1.055	1.060	1.059
Lasso	0.962	0.942	0.937*	0.920**	**0.886***	0.949
Ridge	0.965**	0.922*	0.927*	0.971*	**0.883** **	0.948
SAF	0.981	**0.947**	0.948	1.076	1.047	1.013

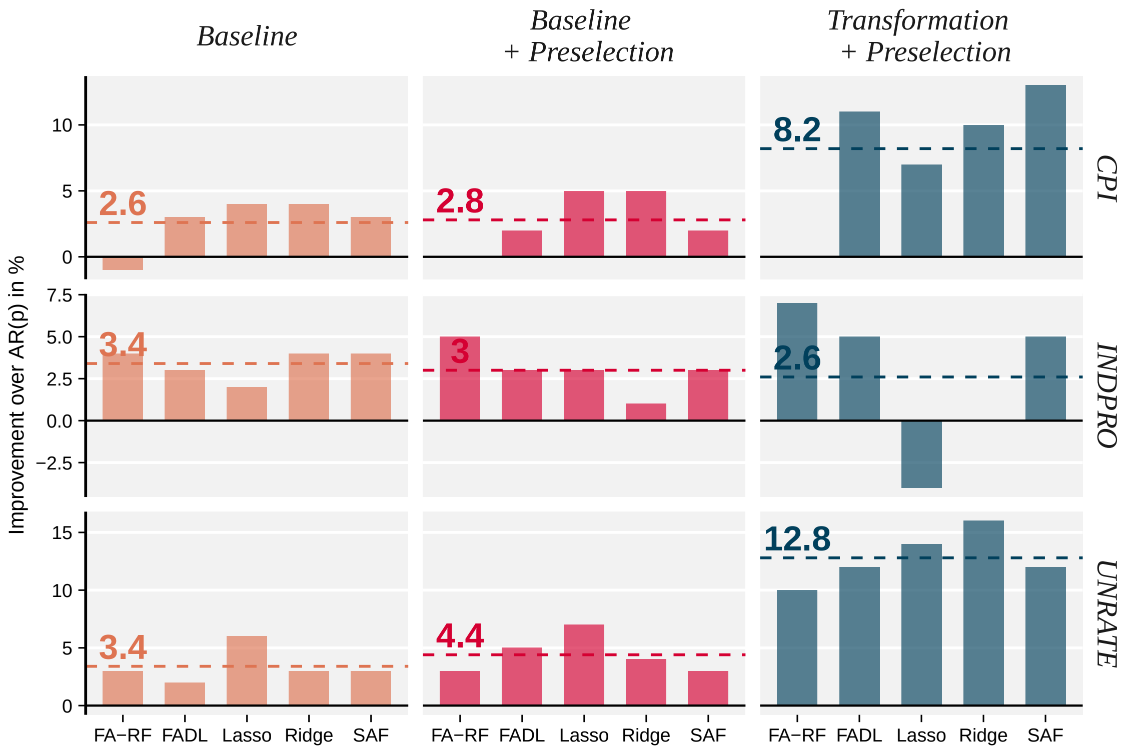
Notes: This table presents the relative RMSFE against an AR(p) model for the six-month ahead forecast horizon. Significance levels indicate differences in forecast performance according to the one-sided Diebold-Mariano test against the AR(p) benchmark and are denoted by asterisks: *** (1% level), ** (5% level), and * (10% level).

Table C.3: Relative RMSFE $h = 12$

	BL	BL30	BL60	TF	TF30	TF60
CPI						
FA-RF	0.954*	0.927***	0.944**	**0.924** **	0.967	0.971
FADL	0.968	1.015	**0.970***	0.971	1.010	0.988
Lasso	0.950**	0.959**	0.946***	0.949***	0.944***	**0.940*** **
Ridge	0.966**	0.956***	0.956***	0.949***	**0.944*** **	0.960**
SAF	**0.966** **	1.009	0.962*	0.978	0.990	1.030
INDPRO						
FA-RF	0.967	0.948	0.970	**0.925** **	0.997	0.988
FADL	1.017	**1.009**	1.019	1.133	1.040	**1.009**
Lasso	1.038	1.022	1.039	0.990*	0.991	**0.985** **
Ridge	1.031	1.001	1.021	1.001	0.989	**0.994**
SAF	**0.981**	1.041	1.039	1.135	1.027	1.063
UNRATE						
FA-RF	0.952***	0.939***	0.952**	0.922	0.931***	**0.924*** **
FADL	0.990	0.998	0.980	1.023	**0.961***	0.999
Lasso	0.980**	0.959**	0.983	0.928***	0.940**	**0.923*** **
Ridge	0.966**	0.964**	0.965***	0.951***	0.942***	**0.936*** **
SAF	0.973**	**0.958*** **	0.982	1.030	0.965*	1.026

Notes: This table presents the relative RMSFE against an AR(p) model for the twelve-month ahead forecast horizon. Significance levels indicate differences in forecast performance according to the one-sided Diebold-Mariano test against the AR(p) benchmark and are denoted by asterisks: *** (1% level), ** (5% level), and * (10% level).

Figure C.1: Short-term Forecast Improvement over AR(p)



Notes: This figure presents the RMSFE improvement over an AR(p) for each target (CPI, INDPRO, and UNRATE). For details see the notes for Figure 1.

KOF

ETH Zurich
KOF Swiss Economic Institute
LEE G 116
Leonhardstrasse 21
8092 Zurich, Switzerland

Phone +41 44 632 42 39
kof@kof.ethz.ch
www.kof.ch