Pontifícia Universidade Católica
do Rio de Janeiro

Lucas Seabra Maynard da Silva

Nowcasting GDP with Machine Learning Models: Evidence from the US

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia .

Advisor: Prof. Marcelo Cunha Medeiros

Rio de Janeiro
April 2020

Pontifícia Universidade Católica
do Rio de Janeiro

Lucas Seabra Maynard da Silva

Nowcasting GDP with Machine Learning Models: Evidence from the US

Dissertation presented to the Programa de Pós–graduação em Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Economia . Approved by the Examination Committee.

Prof. Marcelo Cunha Medeiros
Advisor
Departamento de Economia – PUC-Rio

Prof. Eduardo Zilberman
Departamento de Economia – PUC-Rio

Prof. Diogo Abry Guillén
Itaú – Asset Management

Rio de Janeiro, April the 3rd, 2020

Lucas Seabra Maynard da Silva

B.A., Economics, Pontifícia Universidade Católica do Rio de Janeiro, 2016

Bibliographic data

Silva, Lucas Seabra Maynard da

Nowcasting GDP with Machine Learning Models: Evidence from the US / Lucas Seabra Maynard da Silva; advisor: Marcelo Cunha Medeiros. – Rio de janeiro: PUC-Rio , Departamento de Economia, 2020.

v., 40 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia.

Inclui bibliografia

1. Economia – Teses. 2. Economia – Teses. 3. Nowcasting;. 4. Aprendizado de Máquina;. 5. Avaliação de Previsão;. I. Medeiros, Marcelo Cunha. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. III. Título.

CDD: 330

Acknowledgments

First of all, I thank my parents Augusto Maynard da Silva and Ana Cristina Seabra for the support and motivation always present.

I thank my advisor Marcelo Medeiros for the partnership and advice during the elaboration of this work, and I thank my teachers Márcio Garcia and Marco Cavalcanti for the opportunities and the training which I received since undergraduate. I thank Eduardo Zilberman and Diogo Guillén for the attention and the pertinent comments on this work.

Over these six years, I cannot fail to thank my classmates, especially Rafael Fonseca, Daniel Mesquita, Daniel Doine, and Daniel Coutinho for their friendships which have made those days a lot easier.

I also thank my friends of Our Lady of Peace parish, especially Gustavo Sued, Thales Soares, and Father Frederico for all support, friendship, and advice, especially in the last year.

This study was financed in part by the Coordenação de Aperfeiçoamento Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Abstract

Silva, Lucas Seabra Maynard da; Medeiros, Marcelo Cunha (Advisor). **Nowcasting GDP with Machine Learning Models: Evidence from the US**. Rio de Janeiro, 2020. 40p. Dissertação de mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

This paper examines the use of Machine Learning (ML) models to compute estimates of current-quarter US Real GDP growth rate (nowcasts). These methods can handle large data sets with unsynchronized release dates, and nowcasts are updated each time new data are released along the quarter. A pseudo-out-of-sample exercise is proposed to assess forecasting performance and to analyze the variable selection pattern of these models. The ML method that deserves more attention is the Target Factor, which overcomes the usually adopted dynamic factor model for some predictions vintages in the quarter. We also analyze the variables selected, which are consistent between models and intuition.

Keywords

Nowcasting; Machine Learning; Forecast Evaluation;

Resumo

Silva, Lucas Seabra Maynard da; Medeiros, Marcelo Cunha. **Nowcasting de PIB com Modelos de Machine Learning: Evidência dos EUA**. Rio de Janeiro, 2020. 40p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

O presente trabalho investiga o uso de métodos de Machine Learning (ML) para efetuar estimativas para o trimestre corrente (nowcasts) da taxa de crescimento do PIB Real dos EUA. Esses métodos conseguem lidar com um grande volume de dados e séries com calendários de publicação dessincronizados, e os *nowcasts* são atualizados cada vez que novos dados são publicados ao longo do trimestre. Um exercício *pseudo-out-of-sample* é proposto para avaliar a performance de previsão e analisar o padrão de seleção de variável desses modelos. O método de ML que merece o maior destaque é o Target Factor, que supera o usualmente adotado DFM para alguns *vintages* dentro do trimestre. Ademais, as variáveis selecionadas apresentam consistência entre os modelos e com a intuição.

Palavras-chave

Nowcasting;   Aprendizado de Máquina;   Avaliação de Previsão;

Table of contents

List of figures

List of tables

1
Introduction

In the last decades, real-time monitoring of macroeconomic conditions has gained such importance that it has become the full-time task of some economists since the forecast of economic indicators plays a critical role in monetary policy analysis and macroeconomic studies. Indeed, Bai and Ng (2008) show that accurate forecasts lead to better understandings of economic dynamics mechanism and Bernanke et al. (2005) show that improving predictions leads to more effective monetary policies. Tanaka et al. (2019) show evidence that firms' GDP forecasts are positively correlated with input choices and sales, so forecast errors lower firms' profitability since it's costly to have too much or too little capacity.

A broadly used definition of nowcasting is the one provided in Banbura et al. (2013) which defines it as the prediction of the present, the near future and the recent past. To nowcast variables that are collected at a low frequency and released with a substantial delay, is crucial to use higher frequency information. Since GDP is the key statistic describing the state of the economy of the US and is available at a quarterly frequency with a one-month release delay, we defined it as our target variable. To construct early estimates of GDP we can use several monthly variables related to economic conditions once they are available at a higher frequency and published with a shorter delay.

According to Bok et al. (2018), monitoring macroeconomic conditions in real time is inherently a big data problem since it relies on the availability and exploitation of a large amount of complex data. Dealing with big data usually leads the researcher to face the often called curse of dimensionality problem, that is, the trade-off between excessive complexity (leading to instabilities) and excessive simplificity (leading to misspecification). Hall (2018) argues that the use of Machine Learning (ML) models aims to turn the curse of dimensionality into a blessing by capturing in a parsimonious way the main features among many series. Furthermore, this approach leads to optimal and automated predictions in a manner that is not subject to forecasters discretion. Coulombe et al. (2019), Medeiros et al. (2019), Garcia et al. (2017) and Li and Chen (2014) explore ML methods in macroeconomic forecasting.

We can trace the rise of economic nowcasting from dynamic factor models (DFM) literature. Stock and Watson (2011) provide a chronological literature review and divide the estimations techniques into three generations. In the first generation, parametric models are estimated using gaussian maximum likelihood estimation and the Kalman Filter (KF). However, once this estimation technique makes use of nonlinear optimization algorithms, there are numerical impediments to estimate the parameters when the number of variables is large. In the second generation, researches use non-parametric methods, mainly Principal Components Analysis (PCA) methodology and related methods that provide consistent estimates of the factors. Stock and Watson (2002a) and Stock and Watson (2002b) are classical works of this generation. In the third generation, researches combine nonparametric estimation of PCA with KF methodology, overcoming low-dimensional restrictions of the first generation and providing prediction updates whenever new data is released, which is not explored in the second generation. Giannone et al. (2008) propose a two steps approach in a state-space model, combining the use of Kalman smoother and PCA estimation. Bańbura and Rünstler (2011) and Banbura et al. (2013) perform similar exercises in a study of euro area GDP, also providing the assessment of the impact of new data on subsequent forecasts revisions. In more recent studies, Bok et al. (2018) present the methodology underlying the New York Fed Staff Nowcast and Gomes (2018) compares DFM nowcasts for brazilian GDP.

This paper contributes to the nowcasting literature using several ML methods that differ from DFM. We aim to assess if the use of these methods and their combinations lead to any improvement in forecasting accuracy and to analyze the pattern of variable selection of these models. This article is organized as follows. In Section 2, we describe the methodology used in this work, presenting the nowcasting problem, the current solution proposed by the literature and our methodological contribution. We also describe the dataset and the forecasting scheme. In Section 3, we present our empirical results, with forecasts evaluation statistical tests and the analysis of selected variables pattern. In Section 4, we conclude.

2
Methodology

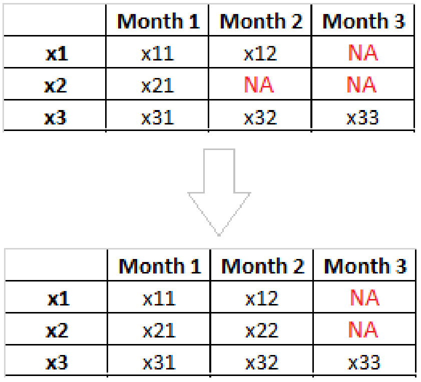
2.1
The Nowcasting Problem

Our goal is to evaluate the current-quarter predictions of GDP growth rate considering the information flow that becomes available throughout the quarter. Within each quarter, the relevant data set expands with time, allowing us to perform sequences of nowcasts. A particular feature of these data sets is that, due to the unsynchronized release dates, some variables have data entries and others have no observations when considering the most recent periods. This feature is the so-called *jagged edge* and we denote this kind of data set an unbalanced panel.

To explicit which information set our nowcast is conditioned on, each one is indexed by a vintage ν_j. In practice, each vintage corresponds to a date in the reference quarter - or after the quarter - which the forecast is made. For each vintage, the nowcast is computed as the expected value of the GDP growth rate conditional on the available information and the underlying model.

The next figure shows an example of the *jagged edge* problem in an unbalanced dataset. Suppose we have only three monthly variables and at vintage ν_j of a generic quarter only the third one is fully available. When we move to the next vintage ν_{j+1}, the second variable has the penultimate entry filled, but the data set still keeps unbalanced.

Figure 2.1: Unbanlanced Panel at vintages ν_j and ν_{j+1}



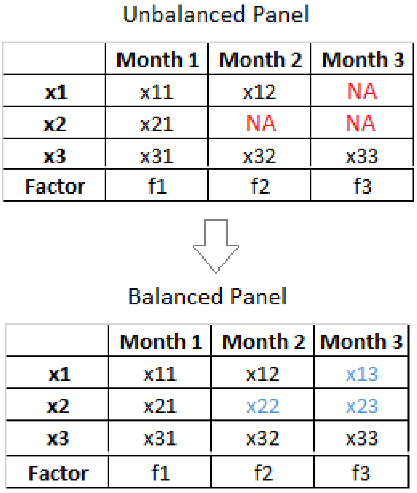
	Month 1	Month 2	Month 3
x1	x11	x12	NA
x2	x21	NA	NA
x3	x31	x32	x33

	Month 1	Month 2	Month 3
x1	x11	x12	NA
x2	x21	x22	NA
x3	x31	x32	x33

The usual way that literature deals with this problem is applying the Dynamic Factor Model (DFM) framework as presented in the classical articles Giannone et al. (2008) and Banbura et al. (2011). This framework allows us to summarize the original variables into a few factors and then compute the nowcasts. But we aim to go beyond this methodology and compute nowcasts with other several Machine Learning (ML) models which some can deal only with a balanced panel. To build this balanced panel we propose a two-step methodology: in the first step, we apply DFM to extract the common factors and in the second step we complete the panel by filling each variable empty entry with its projections on the factors.

The next figure illustrates this proceeding of filling empty entries. Suppose again that we have only three monthly variables at ν_j and that we extract only one common factor. The blue cells that fill the empty entries are the projections of the variables on the factor.

Figure 2.2: Filling empty entries vintage ν_j



Unbalanced Panel

	Month 1	Month 2	Month 3
x1	x11	x12	NA
x2	x21	NA	NA
x3	x31	x32	x33
Factor	f1	f2	f3

Balanced Panel

	Month 1	Month 2	Month 3
x1	x11	x12	x13
x2	x21	x22	x23
x3	x31	x32	x33
Factor	f1	f2	f3

Since the explanatory variables must be in the same frequency of the dependent variable, we proceed with a mean aggregation process. In the next sections, we present the DFM framework and the methodologies of ML models.

2.2 Dynamic Factor Model

In this section, we follow Gomes (2018) to set notation. Let ν be a generic vintage, n be the number of variables and t the observation of a variable. The information set available in ν can be defined as:

$$\Omega_\nu = \{x_{it|\nu}, \text{ with } i = 1, \ldots, n \text{ and } t = 1, \ldots, T_{iv}\}$$

where i identifies the n variables and $t = 1, \ldots, T_{iv}$ identifies the time from the first to the last available observation, which depends on both the series i and the vintage ν. Let y_t be our target variable, so for each information set within a given quarter, the conditional expectation is used to project y_t over Ω_ν and to compute the nowcast. When news is released, the information set is updated to another vintage ω and then a new expectation is computed.

$$\hat{y}_{t|\nu} = \mathbb{E}[y_t|\Omega_\nu] \rightarrow \hat{y}_{t|\omega} = \mathbb{E}[y_t|\Omega_\omega]$$

We assume that y_t depends on the joint dynamics of x_{it} but a projection of y_t on all the n variables may be unfeasible in a big data environment, therefore, this framework provides a parsimonious approach that projects y_t

on few extracted factors that summarize all available information. We apply a hybrid method which estimation technique is characterized by a two-step method that uses Principal Component Analysis (PCA)[1] and Kalman Filtering (KF). This method was first implemented in Giannone et al. (2008) and its consistency properties are studied in Doz et al. (2011). Below, we show the equations of the model and its state-space representation.

$$\boldsymbol{x}_{nt} = \underbrace{\begin{bmatrix}\lambda_n & 0_{n\times r} & \dots & 0_{n\times r}\end{bmatrix}}_{\Lambda} \underbrace{\begin{pmatrix}\boldsymbol{f}_t \\ \dots \\ \boldsymbol{f}_{t-p+1}\end{pmatrix}}_{\boldsymbol{F}_t} + \boldsymbol{\xi}_{nt}$$

$$\begin{pmatrix}\boldsymbol{f}_t \\ \dots \\ \boldsymbol{f}_{t-p+1}\end{pmatrix} = \underbrace{\begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \\ I_r & 0_r & \dots & 0_r & 0_r \\ \dots & \dots & \dots & \dots & \dots \\ 0_r & 0_r & \dots & I_r & 0_r \end{bmatrix}}_{A} \begin{pmatrix}\boldsymbol{f}_{t-1} \\ \dots \\ \boldsymbol{f}_{t-p}\end{pmatrix} + \underbrace{\begin{bmatrix} I_r \\ 0_r \\ \dots \\ 0_r \end{bmatrix}}_{G} \boldsymbol{u}_t$$

Measurement Equation

$$\boldsymbol{x}_{nt} = \Lambda \boldsymbol{F}_t + \boldsymbol{\xi}_{nt}$$

Transition Equation

$$\boldsymbol{F}_t = A\boldsymbol{F}_{t-1} + G\boldsymbol{u}_t$$

with $\boldsymbol{u}_t$ i.i.d $\sim \mathcal{N}(0, \boldsymbol{RR}')$. In this exercise, we select the number of factors r following Bai and Ng (2002) and set the lag of the autoregressive structure following a Bayesian Information Criteria (BIC).

In the first step, we truncate our dataset to construct a new balanced one, overcoming the *jagged-edge* problem and enabling us to provide estimates of Λ, A, RR, and to set the initial condition of $\boldsymbol{F}_t$. Below we explicit the parameters estimation:

1 Let $\boldsymbol{\Phi}^T = (\hat{\boldsymbol{F}}_t)_{t=2,\dots,T}$ and $\boldsymbol{\Phi}^{T-1} = (\hat{\boldsymbol{F}}_t)_{t=1,\dots,T-1}$, where $\hat{\boldsymbol{F}}_t$ is the vector of common factor extracted by PCA. The OLS estimator of A is given by $\hat{A} = (\boldsymbol{\Phi}^{T-1}\boldsymbol{\Phi}^{T-1'})^{-1}\boldsymbol{\Phi}^{T-1}\boldsymbol{\Phi}^{T'}$.

2 We estimate RR' by $\widehat{RR'} = (T-1)^{-1}\sum_{t=2}^{T}(\hat{\boldsymbol{F}}_t - \hat{A}\hat{\boldsymbol{F}}_{t-1})(\hat{\boldsymbol{F}}_t - \hat{A}\hat{\boldsymbol{F}}_{t-1})'$.

[1] For an intuitive and applied to machine learning environment explanation of PCA, see James et al. (2013)

3 Let $\hat{\Sigma}_X = T^{-1}\sum_{t=1}^{T} \boldsymbol{x}_{nt}\boldsymbol{x}_{nt}^{'}$. We estimate Λ by $\hat{\Lambda} = [\hat{\lambda}_n \ 0_{n\times r} \ \ldots \ 0_{n\times r}]$, where $\hat{\lambda}_n$ is the matrix with the r eigenvectors associated with the r biggest eigenvalues of $\hat{\Sigma}_X$.

In the second step, once the Kalman Filter can handle with missing value, we apply a Kalman Smoother to the original dataset, extract the factors and project our target variable on them. To explicit the relationship between the target variable and the factors, we can model it through a linear equation, and since they are at different frequencies, we proceed with a quarterly mean aggregation of the factors. After estimating the parameters of the linear regression, we compute the nowcast as the conditional expectation:

$$\hat{y}_{t|\nu} = \mathbb{E}[y_t|\Omega_\nu] = \hat{\alpha} + \hat{\boldsymbol{\beta}}\hat{\boldsymbol{f}}_{t|\nu}$$

2.3 Machine Learning Models

Shrinkage Models

When dealing with a high-dimensional environment, shrinkage models are a well-established alternative to factor models, and the basic idea behind this modeling is to reduce the parameters that correspond to irrelevant variables towards zero . The parameters are obtained according to the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{t=1}^{T}(y_t - \boldsymbol{x}_t^{'}\boldsymbol{\beta})^2 + \lambda \left[\alpha \sum_{j=1}^{N} \frac{|\beta_j|}{\omega_j} + (1-\alpha)\sum_{j=1}^{N}\frac{|\beta_j|^2}{\omega_j}\right]$$

with different settings of $\phi = (\alpha, \lambda, \omega_j)$ leading to different models. Considering always $\lambda > 0$, we present below four different parameters setting and the corresponding model.

1 - **Ridge Regression**: $\alpha = 0$ and $\omega_j = 1 \ \forall j$

2 - **LASSO:** $\alpha = 1$ and $\omega_j = 1 \ \forall j$

3 - **Adaptative LASSO:** $\alpha = 1$ and $\omega_j = |\hat{\beta}_{init,j}|$

4 - **Elastic Net:** $\alpha \in (0, 1)$ and $\omega_j = 1 \ \forall j$

Ridge Regression was one of the first techniques capable of dealing with large datasets, dating back to the work of Hoerl and Kennard (1970). It is a method that imposes a quadratic penalty to the coefficients and has the appealing feature of an analytical solution. Despite that, the absence of sharp points in the geometric shape of the penalty leads to a coefficient solution vector with exclusively non-zero entries, which is a disadvantage in performing variables selection analysis.

The Least Absolute Shrinkage and Selection Operator (LASSO) is a newer method presented originally in Tibshirani (1996) and differs from Ridge Regression by the fact that it imposes a penalty in the sum of the coefficients absolute values. A great advantage of this method is that it shrinks irrelevant variables exactly to zero, allowing it to perform variable selection and, hence, generating models that are easier to interpret.

However, despite all these goods properties, Zhao and Yu (2006) and Zou (2006) noted that LASSO requires the irrepresentable condition[2] to achieve model selection consistency and does not have the oracle properties[3]. To overcome these deficiences, Zou (2006) propose the Adaptative LASSO (AdaLASSO), a two-step method which uses a first-step estimator to weight the relative importance of the regressors, where $\omega_j = |\hat{\beta}_{init,j}|$ represents differents weights on the penalization of each variable. Medeiros and Mendes (2016) showed that the conditions that must be satisfied on the AdaLASSO are very general. In our exercise we set $\omega_j = |\hat{\beta}_{LASSO,j} + \varepsilon|$ with $\varepsilon = 10^{-3}$ being add to deal with possible zero weights from LASSO first-step estimator.

The last model in our range is Elastic Net, which is a generalization that include tha LASSO and Ridge as particular cases. This model imposes a constraint that is a simple convex combination of the LASSO and Ridge penalizations, and in this exercise ponder qually both models by setting $\alpha = 0.5$.

Target Factor

Bai and Ng (2008) show that the forecasting performance of factor models can be improved simply by applying a pre-testing procedure to select the most important variables to forecast y_t when building the factors. The idea behind this procedure is that if many regressors are irrelevant predictors of y_t, apply

[2] A strong condition that, roughly speaking, the relevant variables may not be very correlated with the irrelevant variables.

[3] Properties that allow it performs as well as if the true underlying model were known.

factor analysis using the entire dataset may result in noisy factors with poor forecasting abilities. In this exercise, we follow a similar procedure than the used in Medeiros and Vasconcelos (2016) and present it below:

1 Fit a linear regression of y_t on each candidate variable including as controls four lags of each candidate variable and an autoregressive term of y_t.

2 Compute the associated p-value for each candidate variable and sort them in descending order. Discard all variables with associated p-value greater than 5% .

3 With the remaining variables, estimate the factors by PCA and select the number of factors following Bai and Ng (2002).

4 Fit a linear regression of y_t on the selected factors.

Complete Subset Regression

The Complete Subset Regression (CSR) was developed by Elliott et al. (2013) and consists of selecting the optimal subset of regressors for predicting the dependent variable. The idea is to select a number $k < N$, where N is the number of regressors, and fit regressions for all possible combinations of k variables. The final forecast is the average forecast computed from all regressions.

As the number of variables in the dataset increase, the CSR can become computationally infeasible since we must fit $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ regressions, and to overcome this problem we follow the same pre-testing approach presented in Medeiros et al. (2019). We start fitting a linear regression of y_t on each candidate variable (including lag controls), sort the t-statistic in absolute values, select the $K < N$ highest values and compute all combinations of regressions with k variables. In this exercise we choose $K = 10$ and $k = 3$, and fit $\binom{10}{3} = 120$ regressions in each forecast.

Random Forest

The Random Forest (RF) methodology was initially proposed by Breiman (2001) as a way to reduce the variance of regression trees, and is based on the bootstrap aggregation (bagging) of randomly constructed regression trees.

A regression tree is a nonparametric model based on the recursive binary partitioning of the covariate space $\mathbb{X}$, where the function that models y_t is a function of local models, each of which is determined in $K \in \mathbb{N}$ different partitions of $\mathbb{X}$. The model is usually displayed in a graph which has the format of a binary decision tree with $N \in \mathbb{N}$ parent (or split) nodes and $K \in \mathbb{N}$ terminal nodes (also called leaves), and which grows from the root node to the terminal nodes. Usually, the partitions are defined by a set of hyperplanes, each of which is orthogonal to the axis of a given predictor variable, called the *split* variable. Hence, conditional on a knowledge of the subregions, the relationship between y_t and $\boldsymbol{x}_t$ is approximated by a piecewise constant model, where each leaf (or terminal node) represents a distinct regime.

Following Garcia et al. (2017), we represent a complex regression-tree model mathematically by introducing the following notation. The root node is at position 0 and a parent node at position j generates a left and right child nodes at positions $2j + 1$ and $2j + 2$, respectively. Every parent node has an associated split variable $x_{s_j,t} \in \boldsymbol{x}_t$, where $s_j \in \mathbb{S} = \{1, 2, \ldots, q\}$. Furthermore, if we let $\mathbb{J}$ and $\mathbb{T}$ be the sets of indexes of the parent and terminal nodes, respectively, a tree architecture can be determined fully from $\mathbb{J}$ and $\mathbb{T}$.

The forecasting model based on regression trees can be represented mathematically as

$$y_t = H_{\mathbb{J}\mathbb{T}}\left(\boldsymbol{x}_t; \boldsymbol{\psi}\right) + u_t = \sum_{i \in \mathbb{T}} \beta_i B_{\mathbb{J}i}\left(\boldsymbol{x}_t; \boldsymbol{\theta}_i\right) + u_t$$

where

$$B_{\mathrm{Ji}}\left(\boldsymbol{x}_t; \boldsymbol{\theta}_i\right) = \prod_{j \in \mathbb{J}} I\left(x_{s_j,t}; c_j\right)^{\frac{n_{i,j}\left(1+n_{i,j}\right)}{2}} \times \left[1 - I\left(x_{s_j,t}; c_j\right)\right]^{(1-n_{i,j})(1+n_{i,j})} \quad ,$$

$$I\left(x_{s_j,t}; c_j\right) = \begin{cases} 1 & \text{se } x_{s_j,t} \leq c_j \\ 0 & \text{otherwise} \end{cases}$$

with

$$n_{i,j} = \begin{cases} -1 & \text{if the path to leaf } i \text{ does not include the parent node } j \\ 0 & \text{if the path to leaf } i \text{ includes the right-child node of the parent node } j \\ 1 & \text{if the path to leaf } i \text{ includes the left-child node of the parent node } j \end{cases}$$

Let $\mathbb{J}_i$ be the subset of $\mathbb{J}$ that contains the indexes of the parent nodes that form the path to leaf i, then $\boldsymbol{\theta}_i$ is the vector that contains all of the parameters c_k such that $k \in \mathbb{J}_i$, $i \in \mathbb{T}$. Note that $\sum_{j \in \mathrm{J}} B_{\mathrm{Ji}}\left(\boldsymbol{x}_t; \boldsymbol{\theta}_j\right) = 1, \forall \boldsymbol{x}_t \in \mathbb{R}^{q+1}$

A random forest is a collection of regression trees, each of which is specified in a bootsrapped sub-sample of the original dataset. Suppose that there are B bootstrapped sub-samples, and denote the estimated regression tree for each of the sub-samples by $H_{\mathbb{J}_b \mathbb{T}_b}\left(.; \boldsymbol{\psi}\right)$. The final prediction is defined as:

$$\hat{y}_t = \frac{1}{B} \sum_{b=1}^{B} H_{\mathbb{J}_b \mathbb{T}_b}\left(\boldsymbol{x}_t; \boldsymbol{\psi}\right)$$

A regression tree is estimated for each of the bootsrapped sub-samples by repeating the following steps recursively for each terminal node of the tree until the minimum number of observations at each node is achieved.

1 Randomly select m out of q covariates as possible split variables.

2 Pick the best variable/split point among the m candidates.

3 Split the node into two child nodes.

Random forests can deal with very large numbers of explanatory variables, and the predicted model is highly nonlinear. It is important to notice that bootstrap samples are calculated using block bootstraps, since we are dealing with time series.

Boosting

We adopt the boosting algorithm similar to the one proposed in Bai and Ng (2009), which has good results for time-series. Boosting is a procedure that estimates an unknown function, especially the conditional mean, using M stage-wise regressions. Suppose we have observations on the dependent variable y_t and on each of n observed predictors, $x_t = \left(x_{t,1}, \ldots, x_{t,n}\right)'$, with $t =$

$1, \ldots, T$. Let $\Phi(x)$ be a function defined on R^n, and let $C(y_t, \Phi(x_t))$ be the loss function that penalizes the deviation of $\Phi(x_t)$ from y_t. The objective is to estimate the function $\Phi(.)$ that minimizes the expected loss $E[C(y_t, \Phi(x_t))]$. Under the quadratic loss function $C(y_t, \Phi(x_t)) = \frac{1}{2}(y_t - \Phi(x_t))^2$, the optimal solution is $\Phi(x) = E(y_t | x_t = x)$. The generic boosting algorithm for estimating $\Phi(x)$ based on observed data is defined as follows:

1 Initialize with $\hat{\Phi}_0(x_t) = \overline{y}\ \forall t$

2 For $m = 1, \ldots, M$:

(a) Compute the negative gradient vector $u_t = \frac{-\partial C(y_t, \Phi(x_t))}{\partial \Phi}|_{\Phi = \Phi(x_t)}$ for $t = 1, \ldots, T$. Under the quadratic loss function, we have $u_t = y_t - \hat{\Phi}_{m-1}(x_t)$

(b) Fit a base learner to the gradient vector to yield $\hat{\phi}_m(.)$. For example, with least squares regression, $\hat{\phi}_m(x_t) = x_t^{'}\hat{\beta}$, where $\hat{\beta} = \arg\min_\beta \sum_{t=1}^{T}(u_t - \boldsymbol{x}_t^{'}\beta)^2$

(c) Update $\hat{\Phi}_m(.) = \hat{\Phi}_{m-1}(.) + \nu\hat{\phi}_m(.)$, where $0 < \nu \leq 1$ is the step length. In this exercise we follow Bai and Ng (2009) and set $\nu = 0.2$.

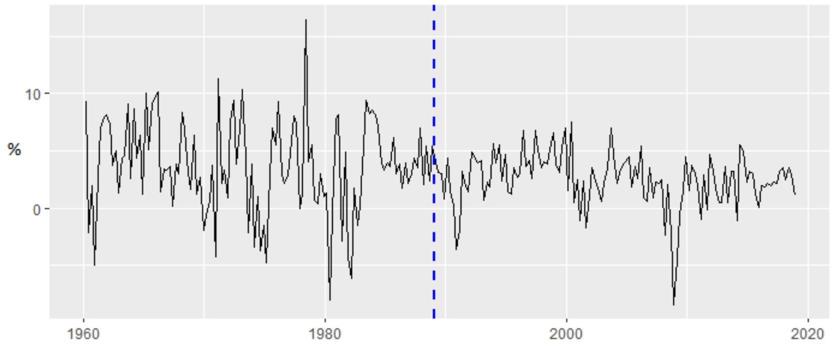
3 Stop the algorithm after the Mth iteration or when the BIC starts to increase.

The algorithm estimates $\Phi(x_t)$ as the sum of M estimated fitting procedures $\hat{\phi}_m(x_t)$, to give $\hat{\Phi}_M(x) = \hat{\Phi}_0(x) + \nu \sum_{m=1}^{M} \hat{\phi}_m(x)$

2.4 Data Description

Our target variable is the US Real GDP change from preceding period with seasonal adjustment and annualized rate. The nowcast platforms used by the Federal Reserve Banks of New York and Atlanta also use this metric to track US Real GDP growth. The chart below presents the series from January 1960 to December 2018, and the vertical dashed blue line demarks the out-of-sample observations.

Figure 2.3: US Real GDP Growth Rate



Data from: U.S. Bureau of Economic Analysis

The dataset of explanatory variables used in this article is available in Federal Reserve Economic Data (FRED), the St. Louis FED's main economic database, and is described in McCracken and Ng (2016). It's a large macroeconomic database designed for empirical big data analysis, consisting of over 130 US monthly indicators available from January 1960 to December 2018. According to the authors, there are some appealing features in this dataset: it is designed to be updated monthly using the FRED database, it is publicly accessible and it relieves researchers from having to manage data changes and revisions

The data are grouped into 8 different groups: **Output and Income** (17), **Labor Market** (32), **Housing** (10), **Consumption** (14) , **Money and Credit** (14), **Interest and Exchange Rates** (22), **Prices** (21) and **Stock Market** (5). After missing data removal we remain with 123 variables.

2.5 Forecasting Scheme

The forecasting exercise covers the period from 1988 Q4 to 2018 Q4, totalizing 121 out-of-sample observations. The models are re-estimated every quarter in a rolling window scheme of 25 years and, for each reference quarter, we compute 22 forecasts using 22 weekly separated vintages. Considering the entire sample, we perform 2662 forecasts and , for each one of them, we simulate the information set available to the researcher by building a pseudo-real-time database.

3
Empirical Results

3.1
Forecasts Evaluation

In the literature of nowcasting, the main benchmark models are the well-known Random Walk (RW) and the Autoregressive (AR) näive models. We first evaluate the performance of Machine Learning (ML) methods relative to these benchmarks through well-established metrics such as Root Mean Squared Error (RMSE). Since we are interested in assess the evolution of forecasting performance throughout the quarter, we compute the RMSE for all models and for each vintage. The formula for RMSE is presented below :

$$\text{RMSE}_\nu = \sqrt{\sum_{t=1}^{T} \frac{(y_t - \hat{y}_{t,\nu})^2}{T}}$$

The next two figures present the ratio of RMSE of each model relative to one of the benchmark models so that we can assess the performance with the evolving information set. The main information provided by these charts is that, as expected, all models become more accurate with the expansion of the information set. This means that, despite the nature of the model - non-linear, factor, shrinkage or averaging - more information, in our context, always leads to a decrease in forecasting error relative to näive models.

Although all models improve accuracy over benchmarks with the expansion of the information set, we are also interested in verifying whether any model or class of models overcome others. Indeed, we can highlight that three models seem to perform consistently well across the vintages: the Dynamic Factor Model (DFM), the Target Factor (TF) and the Ridge model.

Figure 3.1: Forecasting Performance Relative to RW Model

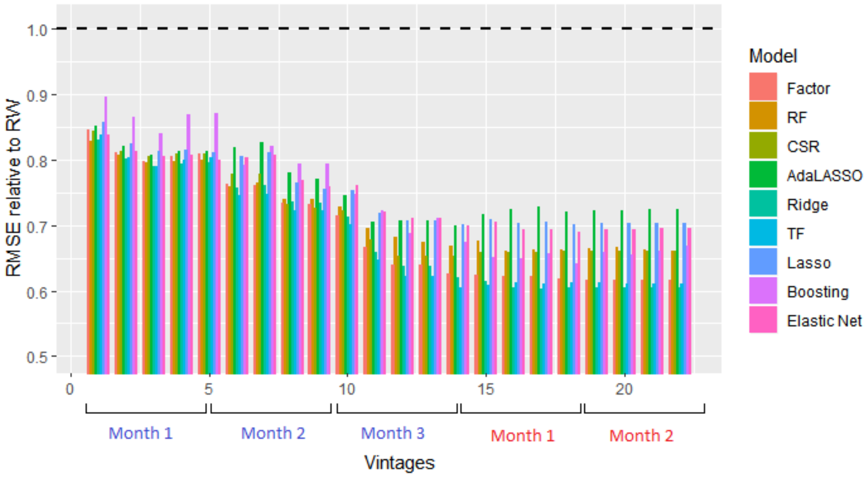
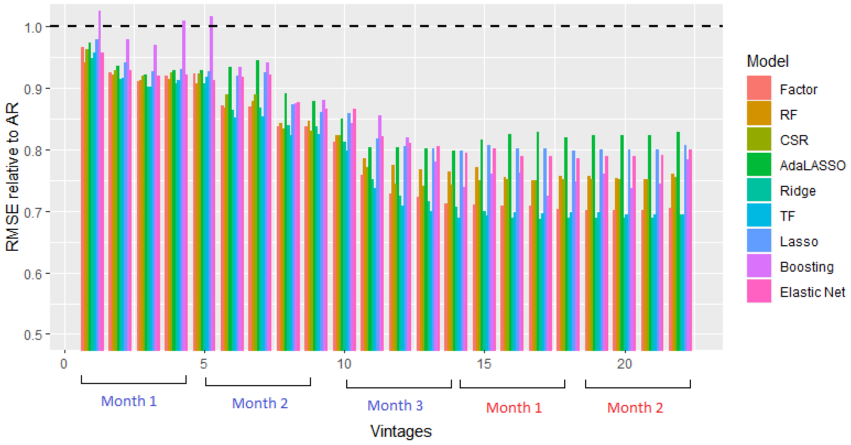
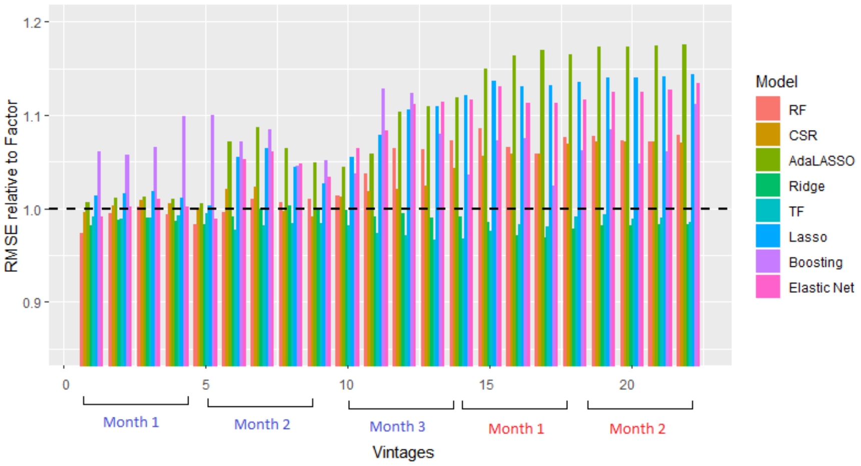


Figure 3.2: Forecasting Performance Relative to AR Model



We have seen that the DFM has been widely used in the literature of GDP nowcasting. Since one of our aims with this article is to assess if the use of ML methods leads to any improvement in forecasting accuracy, we must present the forecasting performance relative to the DFM model. The next figure presents this result and now it seems more clear that TF and Ridge perform consistently well across the vintages.

Figure 3.3: Forecasting Performance Relative to Factor Model



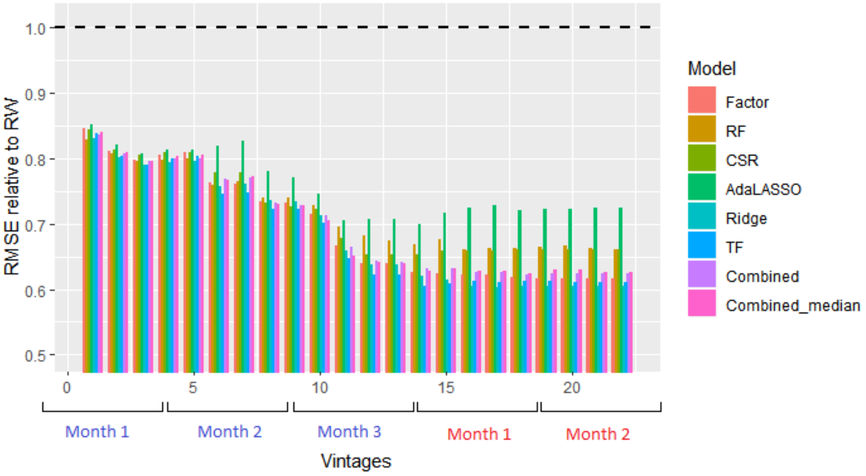
3.2 Models Combination

The next step beyond individual models is verifying if model combining leads to any improvement in nowcasting accuracy. Considering the bias-variance trade-off in machine learning algorithms[1], we can try to reduce the prediction variance by ensembling predictions from final models. That is, instead of fitting a single final model, we can fit multiple final models and the final output prediction is the average of the predictions of the models.

We perform this exercise considering five of our models: TF, RF, CSR, Ridge and AdaLASSO, and we combine them by the mean and the median of predictions. However, as the forecasts are highly positive correlated it's hard to beat the best individual model with combinations. The next chart presents the relative performance of these combined models and it seems there is no such improvement by combining final models relative to the winner models described before.

[1] For an intuitive explanation of bias-variance trade-off in ML algorithms, see James et al. (2013)

Figure 3.4: Forecasting Performance of Combined Models



The next figures show us the forecast correlograms for vintages 5, 8, 12, 16, 20 and 22. We can see that as the information set expands the correlations decrease, although they keep high.

Figure 3.5: Forecasts Correlograms: Vintages 5 and 8

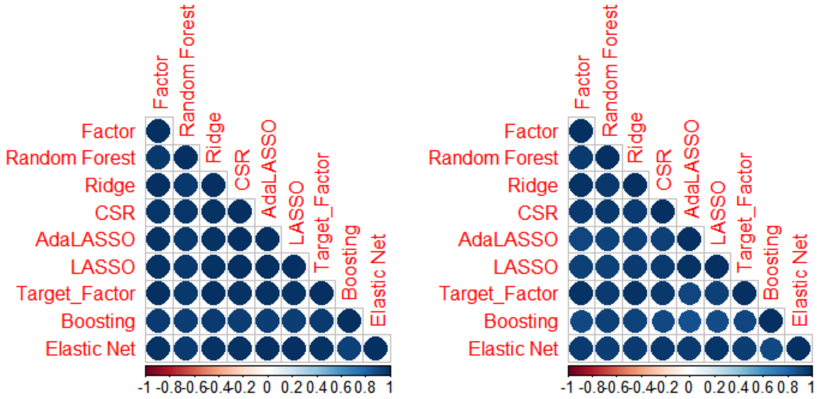


Figure 3.6: Forecasts Correlograms: Vintages 12 and 16

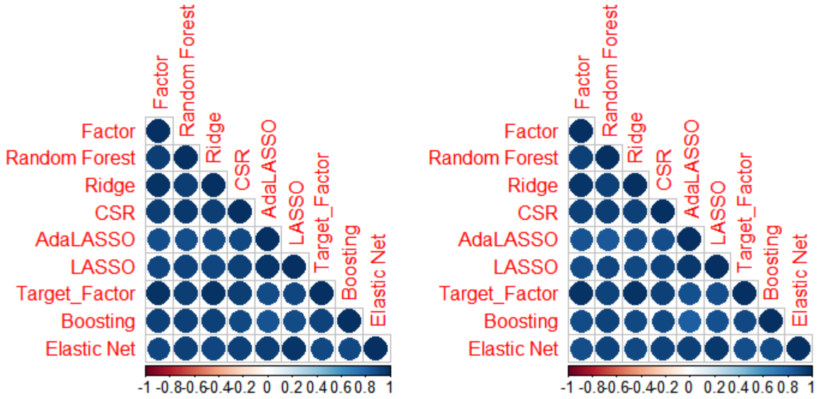
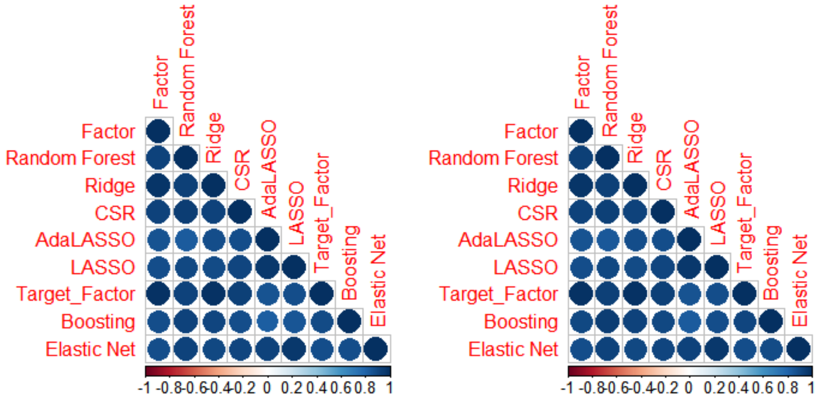


Figure 3.7: Forecasts Correlograms: Vintages 20 and 22



3.3 Statistical Tests

Now we must go beyond an *ad hoc* chart analysis and adopt a formal approach to evaluate forecasting accuracy. We perform two versions of the modified Diebold-Mariano test proposed in Harvey et al. (1997) for each model relative to the DFM considering all information sets. The null hypothesis is the same for both versions, that the two methods have the same forecast accuracy. The first alternative hypothesis is that the tested method is more accurate than the DFM, while the second alternative hypothesis is that the tested method is less accurate than the DFM. The next two tables present the p-values for both tests.

Table 3.1: Alternative Hypothesis: More Accurate

Vintage	LASSO	Ridge	RF	AdaLASSO	Mean	CSR	TF	Boosting	Median	AR
1	0.66	0.10	0.20	0.63	0.20	0.44	0.08	1.00	0.32	0.78
2	0.70	0.22	0.43	0.75	0.35	0.58	0.07	1.00	0.43	0.94
3	0.75	0.27	0.47	0.81	0.40	0.72	0.07	0.99	0.46	0.97
4	0.66	0.19	0.35	0.74	0.31	0.63	0.15	1.00	0.41	0.96
5	0.55	0.15	0.33	0.63	0.24	0.52	0.21	1.00	0.36	0.95
6	0.93	0.33	0.44	0.99	0.64	0.84	0.02	0.92	0.59	0.99
7	0.95	0.49	0.57	1.00	0.74	0.83	0.08	0.99	0.77	0.99
8	0.88	0.55	0.59	0.97	0.47	0.46	0.16	0.99	0.43	0.99
9	0.80	0.55	0.62	0.93	0.40	0.40	0.17	0.99	0.36	0.99
10	0.87	0.47	0.72	0.92	0.46	0.67	0.10	0.93	0.27	1.00
11	0.92	0.36	0.85	0.91	0.47	0.72	0.09	0.99	0.18	1.00
12	0.98	0.43	0.90	0.98	0.58	0.76	0.08	0.99	0.55	1.00
13	0.97	0.44	0.85	0.98	0.53	0.74	0.08	0.99	0.49	1.00
14	0.98	0.38	0.90	0.99	0.64	0.87	0.06	0.94	0.60	1.00
15	0.98	0.31	0.91	0.99	0.71	0.93	0.14	0.84	0.69	1.00
16	0.97	0.18	0.80	1.00	0.57	0.95	0.22	0.81	0.63	1.00
17	0.97	0.16	0.83	1.00	0.57	0.94	0.20	0.82	0.63	1.00
18	0.97	0.27	0.83	1.00	0.61	0.96	0.33	0.74	0.64	1.00
19	0.97	0.31	0.85	1.00	0.66	0.96	0.38	0.89	0.76	1.00
20	0.97	0.31	0.85	1.00	0.65	0.96	0.31	0.86	0.74	1.00
21	0.97	0.31	0.84	1.00	0.65	0.97	0.33	0.88	0.71	1.00
22	0.97	0.31	0.83	1.00	0.65	0.97	0.33	0.96	0.71	1.00

Table 3.2: Alternative Hypothesis: Less Accurate

Vintage	LASSO	Ridge	RF	AdaLASSO	Mean	CSR	TF	Boosting	Median	AR
1	0.34	0.90	0.80	0.37	0.80	0.56	0.92	0.00	0.68	0.22
2	0.30	0.78	0.57	0.25	0.65	0.42	0.93	0.00	0.57	0.06
3	0.25	0.73	0.53	0.19	0.60	0.28	0.93	0.01	0.54	0.03
4	0.34	0.81	0.65	0.26	0.69	0.37	0.85	0.00	0.59	0.04
5	0.45	0.85	0.67	0.37	0.76	0.48	0.79	0.00	0.64	0.05
6	0.07	0.67	0.56	0.01	0.36	0.16	0.98	0.08	0.41	0.01
7	0.05	0.51	0.43	0.00	0.26	0.17	0.92	0.01	0.23	0.01
8	0.12	0.45	0.41	0.03	0.53	0.54	0.84	0.01	0.57	0.01
9	0.20	0.45	0.38	0.07	0.60	0.60	0.83	0.01	0.64	0.01
10	0.13	0.53	0.28	0.08	0.54	0.33	0.90	0.07	0.73	0.00
11	0.08	0.64	0.15	0.09	0.53	0.28	0.91	0.01	0.82	0.00
12	0.02	0.57	0.10	0.02	0.42	0.24	0.92	0.01	0.45	0.00
13	0.03	0.56	0.15	0.02	0.47	0.26	0.92	0.01	0.51	0.00
14	0.02	0.62	0.10	0.01	0.36	0.13	0.94	0.06	0.40	0.00
15	0.02	0.69	0.09	0.01	0.29	0.07	0.86	0.16	0.31	0.00
16	0.03	0.82	0.20	0.00	0.43	0.05	0.78	0.19	0.37	0.00
17	0.03	0.84	0.17	0.00	0.43	0.06	0.80	0.18	0.37	0.00
18	0.03	0.73	0.17	0.00	0.39	0.04	0.67	0.26	0.36	0.00
19	0.03	0.69	0.15	0.00	0.34	0.04	0.62	0.11	0.24	0.00
20	0.03	0.69	0.15	0.00	0.35	0.04	0.69	0.14	0.26	0.00
21	0.03	0.69	0.16	0.00	0.35	0.03	0.67	0.12	0.29	0.00
22	0.03	0.69	0.17	0.00	0.35	0.03	0.67	0.04	0.29	0.00

The first table shows us that, despite the similar accuracy performed by the three methods described before, there is one we should highlight. We can find statistical evidence that the TF is more accurate than the DFM for almost all vintages within the reference quarter, and it is far superior to an AR benchmark model. Moreover, we verify that there is no such accuracy improvement in combining predictions by mean or median.

The second table just complements the first one. Although we find only the Target Factor to be statistically more accurate than the DFM, we cannot find also statistical evidence that Ridge, RF, and combined models are less accurate than the DFM.

The next two charts show the comparative result between actual GDP versus the 22 computed nowcasts for DFM and TF, and the frequency of signal hits, this is, the frequency in which the model predicts recession or expansion correctly, for all vintages.

Figure 3.8: Actual GDP vs Nowcast

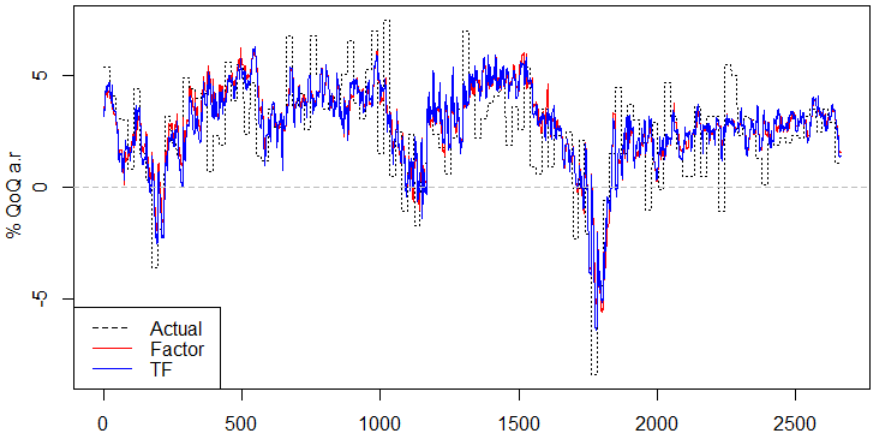
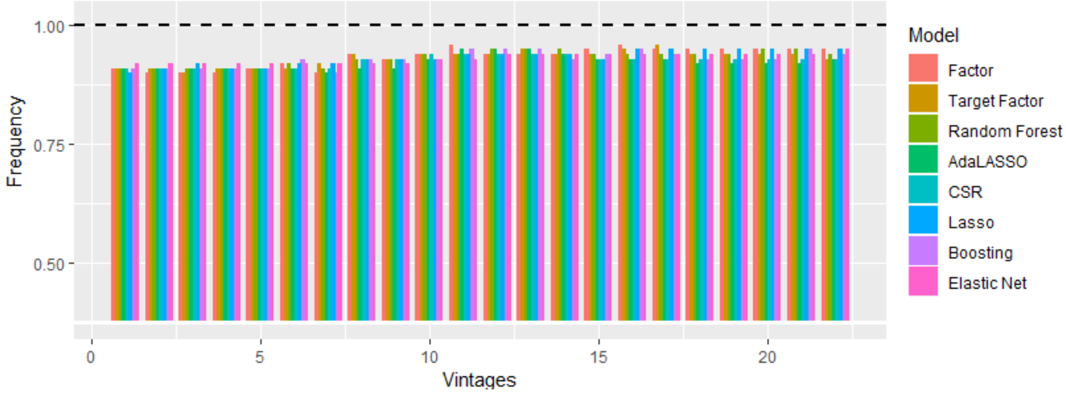


Figure 3.9: Forecasting Directional Accuracy



We can see in the first chart that both models move in unison with the actual GDP, especially following it in the three most prominent recession periods. Sometimes the forecaster is interested less in punctual prediction but the frequency of signal hits - more formally, in the forecasting directional accuracy - since with it it's possible to anticipate recession periods. The second chart shows that all models have a high frequency of hits, but still that, the forecasting directional accuracy of these models improves with the expansion of the information set.

3.4 Variable Selection Analysis

Since our dataset variables are grouped into eight economic categories, we can perform an additional exercise of variables selection analysis. Although only AdaLASSO strictly performs variable selection, its performance below the others leads us to discard it in this analysis. Therefore, we restrict our analysis to our winner model TF.

Before performing our analysis, we should review some of the model's features. The first feature is that this model has a factor structure, which means it summarizes the information contained in the dataset into a few factors and then fit the regression. These factors are called principal components and are sorted in descending order of proportion of total variance explained. But the dataset in which we extract the factors is a concise one since we apply a pre-selection procedure based on the fit of the dependent variable on each candidate variable. The second feature is that each principal component is a linear combination of the explanatory variables, and the optimal number of factors included in the regression is obtained through information criteria. In our rolling window exercise, the lower number of factors selected is two, so we analyze only the first and the second principal components.

Recovering this framework of TF, we can analyze the variable selection pattern in two ways. The first way is exploring the categories of the variables selected in the pre-selection procedure and verifying how this pattern behaves with the evolving information set. The second way is exploring the absolute values of the coefficients of principal components and verifying how the weights of the variables changes with the evolving information set.

The next three charts show our results. The first chart presents the proportion of each category in the selected variables for all vintages. The second and third ones present for each category the proportion of the absolute coefficients in the total sum of the absolute coefficients for the first and second principal components, respectively, considering all vintages.

Figure 3.10: Selected Variables

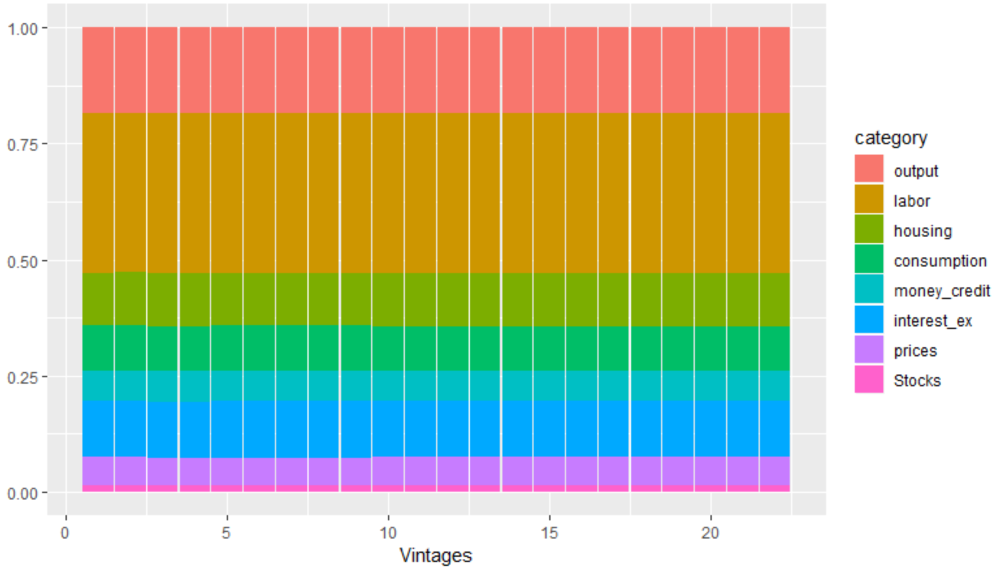


Figure 3.11: First Principal Component

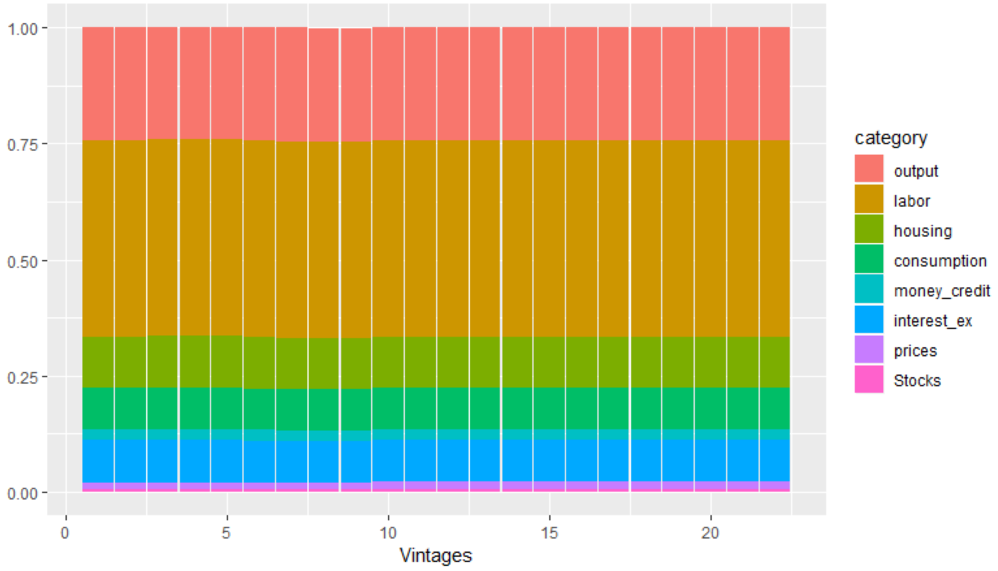
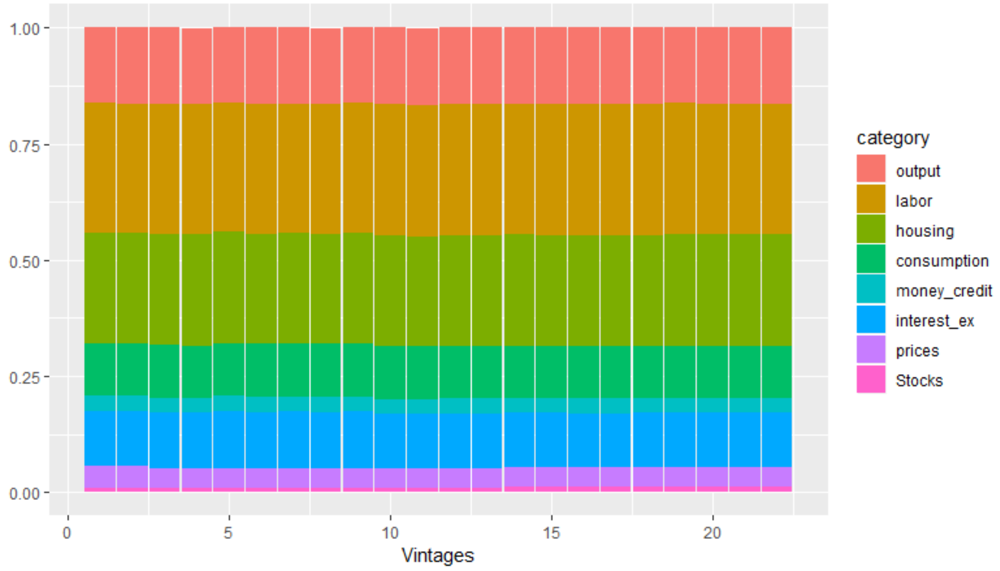


Figure 3.12: Second Principal Component



The main information provided by these charts is that there is a very stable pattern of variable selection across vintages. Indeed, it seems that the expansion of the information set doesn't affect significantly the proportion of each category in the selected variables or in the proportion of the absolute value of the coefficients. The first chart shows that more than half of the proportion comes from Labor Market and Output and Income categories, which is consistent with intuition since the GDP is a real sector variable. In the second chart, no surprises again. The proportion of these two categories corresponds to more than two-thirds of the overall sum of the absolute value of the coefficients. In the last chart, we can find a different but still similar pattern. Although these two categories still keep a high proportion, we can see the rise of housing, almost becoming the most important category in the linear combination. This is not a surprise once the second principal component should be orthogonal to the first one, so we would expect that it put more weights in variables with lower coefficients in the first linear combination.

4
Conclusions

This paper contributes to the nowcasting literature by using several Machine Learning (ML) methods that differ from the usual Dynamic Factor Model (DFM) presented in Giannone et al. (2008) and in Bok et al. (2018). We aim to assess if the use of these methods or their combinations lead to any improvement in forecasting accuracy, and also, to analyze the pattern of variable selection.

Most of these ML methods deal only with a balanced panel, but, due to the unsynchronized date release of the variables, our dataset has a particular feature of empty entries at the end of the sample. To deal with this feature and fill these empty entries, we implement a two-steps methodology. In the first step, we apply a state-space model that can handle missing values to extract common factors, and in the second step, we fill the empty entries of each series with its projections on the factors.

The ML methods are from different types, such as static factors, nonlinear and shrinkage models, and we present statistical evidence that Target Factor (TF) is more accurate than the broadly used DFM. This model has a factor structure built with Principal Component Analysis (PCA) on a concise dataset that suffers a pre-selection procedure based on the correlation of the dependent variable and the explanatory variables.

Finally, in the variable selection analysis exercise, the variables selected are consistent across vintages and to intuition. The flow of information doesn't affect the pattern of selected variables, which remains very stable, and the main selected variables to predict real GDP came from the real sector, such as Output and Income, and Labor Market, which is consistent with intuition.

Bibliography

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.

Bai, J. and Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, 24(4):607–629.

Banbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-casting and the real-time data flow. *Handbook of economic forecasting*, 2(Part A):195–237.

Banbura, M., Giannone, D., and Reichlin, L. (2011). Nowcasting with daily data. *European Central Bank, Working Paper*.

Bańbura, M. and Rünstler, G. (2011). A look into the factor model black box: publication lags and the role of hard and soft data in forecasting gdp. *International Journal of Forecasting*, 27(2):333–346.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly journal of economics*, 120(1):387–422.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., and Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, 10:615–643.

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

Coulombe, P. G., Leroux, M., Stevanovic, D., Surprenant, S., et al. (2019). How is machine learning useful for macroeconomic forecasting? Technical report, CIRANO.

Doz, C., Giannone, D., and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, 164(1):188–205.

Elliott, G., Gargano, A., and Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, 177(2):357–373.

Garcia, M. G., Medeiros, M. C., and Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, 33(3):679–693.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Gomes, G. B. (2018). *Nowcasting Brazilian GDP: a performance assessment of dynamic factor models*. PhD thesis.

Hall, A. S. (2018). Machine learning approaches to macroeconomic forecasting. *Economic Review-Federal Reserve Bank of Kansas City*, 103(4):63.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2):281–291.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Li, J. and Chen, W. (2014). Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting*, 30(4):996–1015.

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. and Mendes, E. F. (2016). l (1)-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors.

Medeiros, M. C. and Vasconcelos, G. F. (2016). Forecasting macroeconomic variables in data-rich environments. *Economics Letters*, 138:50–52.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., and Zilberman, E. (2019). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, pages 1–22.

Stock, J. H. and Watson, M. (2011). Dynamic factor models. *Oxford Handbooks Online*.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

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

Tanaka, M., Bloom, N., David, J. M., and Koga, M. (2019). Firm performance and macro forecast accuracy. *Journal of Monetary Economics*.

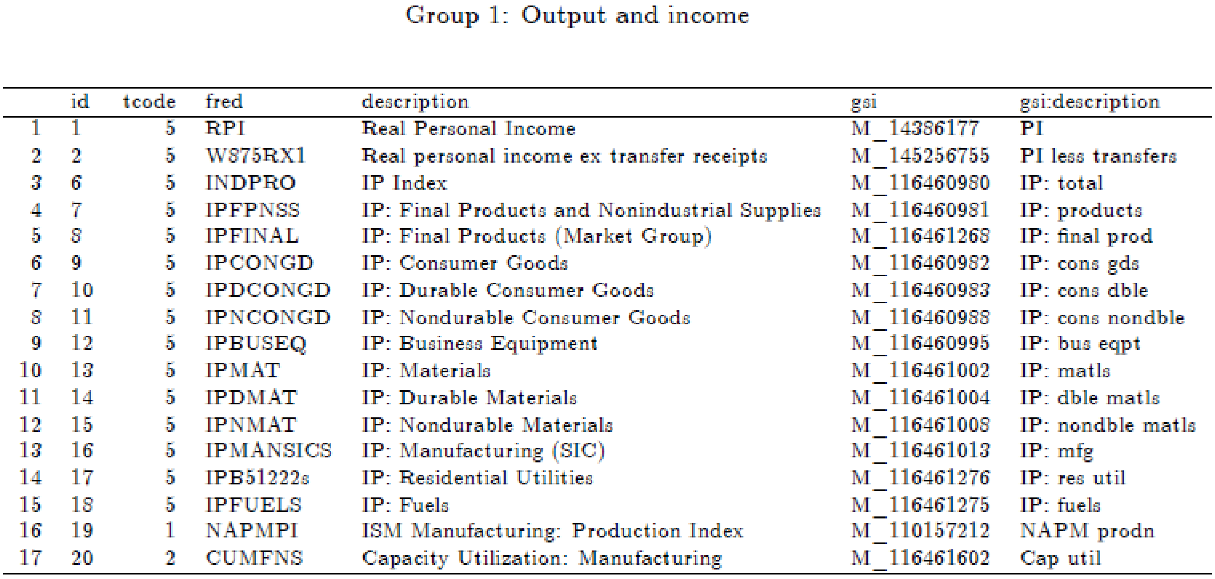
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

A
Data

Here we present the dataset described in McCracken and Ng (2016). The column TCODE denotes the following data transformation for a series $\mathbf{x}$: (1) no transformation; (2) $\Delta\mathbf{x}_t$; (3) $\Delta^2\mathbf{x}_t$; (4) $\log\mathbf{x}_t$; (5) $\Delta\log\mathbf{x}_t$; (6) $\Delta^2\log\mathbf{x}_t$; (7) $\Delta(\frac{\mathbf{x}_t}{\mathbf{x}_{t-1}} - 1)$. The FRED column gives mnemonics in FRED followed by a short description. The comparable series in Global Insight is given in the column GSI. Some series require adjustments to the raw data available in FRED. We tag these variables with an asterisk to indicate that they been adjusted and thus di§er from the series from the source. A summary of the adjustments is detailed in the paper https://research.stlouisfed.org/wp/2015/2015- 012.pdf.



Group 1: Output and income

	id	tcode	fred	description	gsi	gsi:description
1	1	5	RPI	Real Personal Income	M_14386177	PI
2	2	5	W875RX1	Real personal income ex transfer receipts	M_145256755	PI less transfers
3	6	5	INDPRO	IP Index	M_116460980	IP: total
4	7	5	IPFPNSS	IP: Final Products and Nonindustrial Supplies	M_116460981	IP: products
5	8	5	IPFINAL	IP: Final Products (Market Group)	M_116461268	IP: final prod
6	9	5	IPCONGD	IP: Consumer Goods	M_116460982	IP: cons gds
7	10	5	IPDCONGD	IP: Durable Consumer Goods	M_116460983	IP: cons dble
8	11	5	IPNCONGD	IP: Nondurable Consumer Goods	M_116460988	IP: cons nondble
9	12	5	IPBUSEQ	IP: Business Equipment	M_116460995	IP: bus eqpt
10	13	5	IPMAT	IP: Materials	M_116461002	IP: matls
11	14	5	IPDMAT	IP: Durable Materials	M_116461004	IP: dble matls
12	15	5	IPNMAT	IP: Nondurable Materials	M_116461008	IP: nondble matls
13	16	5	IPMANSICS	IP: Manufacturing (SIC)	M_116461013	IP: mfg
14	17	5	IPB51222s	IP: Residential Utilities	M_116461276	IP: res util
15	18	5	IPFUELS	IP: Fuels	M_116461275	IP: fuels
16	19	1	NAPMPI	ISM Manufacturing: Production Index	M_110157212	NAPM prodn
17	20	2	CUMFNS	Capacity Utilization: Manufacturing	M_116461602	Cap util

Group 2: Labor market

	id	tcode	fred	description	gsi	gsi:description
1	21*	2	HWI	Help-Wanted Index for United States		Help wanted indx
2	22*	2	HWIURATIO	Ratio of Help Wanted/No. Unemployed	M_110156531	Help wanted/unemp
3	23	5	CLF16OV	Civilian Labor Force	M_110156467	Emp CPS total
4	24	5	CE16OV	Civilian Employment	M_110156498	Emp CPS nonag
5	25	2	UNRATE	Civilian Unemployment Rate	M_110156541	U: all
6	26	2	UEMPMEAN	Average Duration of Unemployment (Weeks)	M_110156528	U: mean duration
7	27	5	UEMPLT5	Civilians Unemployed - Less Than 5 Weeks	M_110156527	U < 5 wks
8	28	5	UEMP5TO14	Civilians Unemployed for 5-14 Weeks	M_110156523	U 5-14 wks
9	29	5	UEMP15OV	Civilians Unemployed - 15 Weeks & Over	M_110156524	U 15+ wks
10	30	5	UEMP15T26	Civilians Unemployed for 15-26 Weeks	M_110156525	U 15-26 wks
11	31	5	UEMP27OV	Civilians Unemployed for 27 Weeks and Over	M_110156526	U 27+ wks
12	32*	5	CLAIMSx	Initial Claims	M_15186204	UI claims
13	33	5	PAYEMS	All Employees: Total nonfarm	M_123109146	Emp: total
14	34	5	USGOOD	All Employees: Goods-Producing Industries	M_123109172	Emp: gds prod
15	35	5	CES1021000001	All Employees: Mining and Logging: Mining	M_123109244	Emp: mining
16	36	5	USCONS	All Employees: Construction	M_123109331	Emp: const
17	37	5	MANEMP	All Employees: Manufacturing	M_123109542	Emp: mfg
18	38	5	DMANEMP	All Employees: Durable goods	M_123109573	Emp: dble gds
19	39	5	NDMANEMP	All Employees: Nondurable goods	M_123110741	Emp: nondbles
20	40	5	SRVPRD	All Employees: Service-Providing Industries	M_123109193	Emp: services
21	41	5	USTPU	All Employees: Trade, Transportation & Utilities	M_123111543	Emp: TTU
22	42	5	USWTRADE	All Employees: Wholesale Trade	M_123111563	Emp: wholesale
23	43	5	USTRADE	All Employees: Retail Trade	M_123111867	Emp: retail
24	44	5	USFIRE	All Employees: Financial Activities	M_123112777	Emp: FIRE
25	45	5	USGOVT	All Employees: Government	M_123114411	Emp: Govt
26	46	1	CES0600000007	Avg Weekly Hours : Goods-Producing	M_140687274	Avg hrs
27	47	2	AWOTMAN	Avg Weekly Overtime Hours : Manufacturing	M_123109554	Overtime: mfg
28	48	1	AWHMAN	Avg Weekly Hours : Manufacturing	M_14386098	Avg hrs: mfg
29	49	1	NAPMEI	ISM Manufacturing: Employment Index	M_110157206	NAPM empl
30	127	6	CES0600000008	Avg Hourly Earnings : Goods-Producing	M_123109182	AHE: goods
31	128	6	CES2000000008	Avg Hourly Earnings : Construction	M_123109341	AHE: const
32	129	6	CES3000000008	Avg Hourly Earnings : Manufacturing	M_123109552	AHE: mfg

Group 3: Housing

	id	tcode	fred	description	gsi	gsi:description
1	50	4	HOUST	Housing Starts: Total New Privately Owned	M_110155536	Starts: nonfarm
2	51	4	HOUSTNE	Housing Starts, Northeast	M_110155538	Starts: NE
3	52	4	HOUSTMW	Housing Starts, Midwest	M_110155537	Starts: MW
4	53	4	HOUSTS	Housing Starts, South	M_110155543	Starts: South
5	54	4	HOUSTW	Housing Starts, West	M_110155544	Starts: West
6	55	4	PERMIT	New Private Housing Permits (SAAR)	M_110155532	BP: total
7	56	4	PERMITNE	New Private Housing Permits, Northeast (SAAR)	M_110155531	BP: NE
8	57	4	PERMITMW	New Private Housing Permits, Midwest (SAAR)	M_110155530	BP: MW
9	58	4	PERMITS	New Private Housing Permits, South (SAAR)	M_110155533	BP: South
10	59	4	PERMITW	New Private Housing Permits, West (SAAR)	M_110155534	BP: West

Group 4: Consumption, orders, and inventories

	id	tcode	fred	description	gsi	gsi:description
1	3	5	DPCERA3M086SBEA	Real personal consumption expenditures	M_123008274	Real Consumption
2	4*	5	CMRMTSPLx	Real Manu. and Trade Industries Sales	M_110156998	M&T sales
3	5*	5	RETAILx	Retail and Food Services Sales	M_130439509	Retail sales
4	60	1	NAPM	ISM : PMI Composite Index	M_110157208	PMI
5	61	1	NAPMNOI	ISM : New Orders Index	M_110157210	NAPM new ordrs
6	62	1	NAPMSDI	ISM : Supplier Deliveries Index	M_110157205	NAPM vendor del
7	63	1	NAPMII	ISM : Inventories Index	M_110157211	NAPM Invent
8	64	5	ACOGNO	New Orders for Consumer Goods	M_14385863	Orders: cons gds
9	65*	5	AMDMNOx	New Orders for Durable Goods	M_14386110	Orders: dble gds
10	66*	5	ANDENOx	New Orders for Nondefense Capital Goods	M_178554409	Orders: cap gds
11	67*	5	AMDMUOx	Unfilled Orders for Durable Goods	M_14385946	Unf orders: dble
12	68*	5	BUSINVx	Total Business Inventories	M_15192014	M&T invent
13	69*	2	ISRATIOx	Total Business: Inventories to Sales Ratio	M_15191529	M&T invent/sales
14	130*	2	UMCSENTx	Consumer Sentiment Index	hhsntn	Consumer expect

Group 5: Money and credit

	id	tcode	fred	description	gsi	gsi:description
1	70	6	M1SL	M1 Money Stock	M_110154984	M1
2	71	6	M2SL	M2 Money Stock	M_110154985	M2
3	72	5	M2REAL	Real M2 Money Stock	M_110154985	M2 (real)
4	73	6	AMBSL	St. Louis Adjusted Monetary Base	M_110154995	MB
5	74	6	TOTRESNS	Total Reserves of Depository Institutions	M_110155011	Reserves tot
6	75	7	NONBORRES	Reserves Of Depository Institutions	M_110155009	Reserves nonbor
7	76	6	BUSLOANS	Commercial and Industrial Loans	BUSLOANS	C&I loan plus
8	77	6	REALLN	Real Estate Loans at All Commercial Banks	BUSLOANS	DC&I loans
9	78	6	NONREVSL	Total Nonrevolving Credit	M_110154564	Cons credit
10	79*	2	CONSPI	Nonrevolving consumer credit to Personal Income	M_110154569	Inst cred/PI
11	131	6	MZMSL	MZM Money Stock	N.A.	N.A.
12	132	6	DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding	N.A.	N.A.
13	133	6	DTCTHFNM	Total Consumer Loans and Leases Outstanding	N.A.	N.A.
14	134	6	INVEST	Securities in Bank Credit at All Commercial Banks	N.A.	N.A.

Group 6: Interest and exchange rates

	id	tcode	fred	description	gsi	gsi:description
1	84	2	FEDFUNDS	Effective Federal Funds Rate	M_110155157	Fed Funds
2	85*	2	CP3Mx	3-Month AA Financial Commercial Paper Rate	CPF3M	Comm paper
3	86	2	TB3MS	3-Month Treasury Bill:	M_110155165	3 mo T-bill
4	87	2	TB6MS	6-Month Treasury Bill:	M_110155166	6 mo T-bill
5	88	2	GS1	1-Year Treasury Rate	M_110155168	1 yr T-bond
6	89	2	GS5	5-Year Treasury Rate	M_110155174	5 yr T-bond
7	90	2	GS10	10-Year Treasury Rate	M_110155169	10 yr T-bond
8	91	2	AAA	Moody's Seasoned Aaa Corporate Bond Yield		Aaa bond
9	92	2	BAA	Moody's Seasoned Baa Corporate Bond Yield		Baa bond
10	93*	1	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS		CP-FF spread
11	94	1	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS		3 mo-FF spread
12	95	1	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS		6 mo-FF spread
13	96	1	T1YFFM	1-Year Treasury C Minus FEDFUNDS		1 yr-FF spread
14	97	1	T5YFFM	5-Year Treasury C Minus FEDFUNDS		5 yr-FF spread
15	98	1	T10YFFM	10-Year Treasury C Minus FEDFUNDS		10 yr-FF spread
16	99	1	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS		Aaa-FF spread
17	100	1	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS		Baa-FF spread
18	101	5	TWEXMMTH	Trade Weighted U.S. Dollar Index: Major Currencies		Ex rate: avg
19	102*	5	EXSZUSx	Switzerland / U.S. Foreign Exchange Rate	M_110154768	Ex rate: Switz
20	103*	5	EXJPUSx	Japan / U.S. Foreign Exchange Rate	M_110154755	Ex rate: Japan
21	104*	5	EXUSUKx	U.S. / U.K. Foreign Exchange Rate	M_110154772	Ex rate: UK
22	105*	5	EXCAUSx	Canada / U.S. Foreign Exchange Rate	M_110154744	EX rate: Canada

Group 7: Prices

	id	tcode	fred	description	gsi	gsi:description
1	106	6	WPSFD49207	PPI: Finished Goods	M110157517	PPI: fin gds
2	107	6	WPSFD49502	PPI: Finished Consumer Goods	M110157508	PPI: cons gds
3	108	6	WPSID61	PPI: Intermediate Materials	M_110157527	PPI: int matls
4	109	6	WPSID62	PPI: Crude Materials	M_110157500	PPI: crude matls
5	110*	6	OILPRICEx	Crude Oil, spliced WTI and Cushing	M_110157273	Spot market price
6	111	6	PPICMM	PPI: Metals and metal products:	M_110157335	PPI: nonferrous
7	112	1	NAPMPRI	ISM Manufacturing: Prices Index	M_110157204	NAPM com price
8	113	6	CPIAUCSL	CPI : All Items	M_110157323	CPI-U: all
9	114	6	CPIAPPSL	CPI : Apparel	M_110157299	CPI-U: apparel
10	115	6	CPITRNSL	CPI : Transportation	M_110157302	CPI-U: transp
11	116	6	CPIMEDSL	CPI : Medical Care	M_110157304	CPI-U: medical
12	117	6	CUSR0000SAC	CPI : Commodities	M_110157314	CPI-U: comm.
13	118	6	CUSR0000SAD	CPI : Durables	M_110157315	CPI-U: dbles
14	119	6	CUSR0000SAS	CPI : Services	M_110157325	CPI-U: services
15	120	6	CPIULFSL	CPI : All Items Less Food	M_110157328	CPI-U: ex food
16	121	6	CUSR0000SA0L2	CPI : All items less shelter	M_110157329	CPI-U: ex shelter
17	122	6	CUSR0000SA0L5	CPI : All items less medical care	M_110157330	CPI-U: ex med
18	123	6	PCEPI	Personal Cons. Expend.: Chain Index	gmdc	PCE defl
19	124	6	DDURRG3M086SBEA	Personal Cons. Exp: Durable goods	gmdcd	PCE defl: dlbes
20	125	6	DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods	gmdcn	PCE defl: nondble
21	126	6	DSERRG3M086SBEA	Personal Cons. Exp: Services	gmdcs	PCE defl: service

Group 8: Stock market

	id	tcode	fred	description	gsi	gsi:description
1	80*	5	S&P 500	S&P's Common Stock Price Index: Composite	M_110155044	S&P 500
2	81*	5	S&P: indust	S&P's Common Stock Price Index: Industrials	M_110155047	S&P: indust
3	82*	2	S&P div yield	S&P's Composite Common Stock: Dividend Yield		S&P div yield
4	83*	5	S&P PE ratio	S&P's Composite Common Stock: Price-Earnings Ratio		S&P PE ratio
5	135*	1	VXOCLSx	VXO		