Random forest for high-dimensional non-linear forecasting

Martin Lumiste[*]

June 6, 2016

Abstract

I consider the problem of forecasting a univariate time series based on many predictor series, with variable dimension possibly greater than time dimension. A standard approach taken in the macroeconometrics literature since Stock and Watson (2002), *diffusion index forecasting*, is to find common factors of the predictor variable series and include the most relevant of these in the forecast equation. I introduce a method that allows for non-linearities, is more precise in forecasting, is flexible to different types of data, can handle abritrarily high-dimensional data and is easier to compute compared to the DI model.

JEL classification: C53, C55

Keywords: Macroeconomics, Forecasting, Machine Learning, Random Forest

[*] E-mail: mlumiste@ucsd.edu

Many thanks to my advisors: Eric Gautier, who always guided me in the right direction and Sébastien Gadat, who introduced me to the wonders of machine learning

1 Introduction

There are many different uses for a predictive regression model. Ng (2013) points out that consistent variable selection and accurate prediction are often contradictory ends. Hansen (2005) confirms that selecting correct variables depends heavily on the task at hand. Timmermann and Elliott (2016) suggest customizing the loss function with the aims of the forecast in mind, as the function to be approximated is determined by the loss function – for example mean squared error (MSE) loss leads to the conditional mean, whereas mean absolute loss (MAD) leads to the conditional median.

This paper considers the task of forecasting a univariate series under MSE loss and a high-dimensional set of covariate series, focusing only on accuracy. This is the typical problem faced in central banks and financial institutions. In the framework of Timmermann and Elliott (2016), a MSE-based forecast might be optimal if the forecast is used as an input for unknown future decision-making.

Since we aim to produce the most accurate forecasts, this paper takes a considerably more data-driven view and focuses less on economic theory. However, note that in this sense it does not differ at all from vector autoregression models, a benchmark tool used in central bank macro forecasting. Since economic theory rarely suggests the functional form of variable association or even the whole set of relevant variables to forecast a given series, I argue that this is the natural approach.

From a practical point of view, a good model should not only be precise, but also computationally quick (in the sense that it does not take days to operate) and easy to implement. I propose the random forest approach used in genomic data analysis and show on U.S. data that there could be considerable gains from adopting this model as the benchmark for macroeconomic forecasting.

In the following text, *high-dimensional* data refers to datasets where the variable (column) dimension is large and perhaps bigger than the individual/time dimension (row). We use the term *sparsity* for high-dimensional data which has many column variables which are unrelated to the series to be forecasted. The next section explains thoroughly the forecasting problem and previous methods employed in the literature. Section 3 introduces theoretical and survey results from random forests. Section 4 gives a brief overview of the data set and methodology used for forecasting and Section 5 presents the results.

2 Background

Macroeconomic datasets have a few peculiar features. Firstly, they are often high-dimensional in the sense that the column dimension N exceeds or is simply large relative to the row dimension T. However, the number of rows and columns is relatively small. For example, Stock and Watson (2002) forecast different macroeconomic series using regressions on datasets with 120 to 468 rows and 215 columns. Secondly, contrarily to other modern frameworks of high-dimensional data such as image or text processing, the data is not *sparse*. Instead, macroeconomic data sets are characterized by a high correlation of the variables. Early work by Sargent and Sims (1977) showed that two dynamic factors can explain a significant part of US macroeconomic variables, a fact that has since been thoroughly replicated. Third, there appears to be a considerable amount of parameter instability, i.e. different variables are relevant for prediction depending on the time, with different impacts (Mol et al., 2008). Finally, macroeconomic data exhibits high persistence and is thus weakly dependent. A good macroeconomic forecasting model should deal with all these points.

Given these characteristics, a natural approach that has been taken in the literature is reducing the dimension of the predictor set via factor analysis to capture the underlying latent business cycle elements and projecting future values on lags of the dependent series and these factors, usually in a linear model.

This is the dynamic factor or *diffusion index* model set forth by Stock and Watson (2002):

$$y_{t+h} = \alpha_h + \beta_h(L)y_t + \gamma_h(L)f_t + \epsilon_{t+h} \tag{1}$$

$$X_{it} = \lambda_i F_t + u_{it} \quad i = 1, \dots, N, t = 1, \dots, T \tag{2}$$

where y_{t+h} is a scalar series we wish to forecast h steps ahead, X_t is N-dimensional, f_t are the first r principal components of F_t where $r << N$. $\beta(L), \gamma(L), \lambda_i(L)$ denote the lag polynomials in nonnegative powers of L. If we assume $\mathbb{E}\big[\epsilon_{t+h}|F_t, y_t, X_t, F_{t-1}, y_{t-1}, X_{t-1}, \dots\big] = 0$, the optimal forecast of y_{T+h} with respect to the mean squared error loss function is simply the conditional mean $\alpha_h + \beta_h(L)y_T + \gamma_h(L)f_T$.

By using only the first r factors of F_t, we reduce the dimensionality of X_{it}. In practice, it is sometimes assumed for simplification that the lag polynomials $\gamma(L)$ and $\beta(L)$ are of the same order p (Bai and Ng, 2009). A researcher is then faced with the task of choosing p^* and r^*, the lag orders and factors chosen from PCA for the final regression. For this purpose, one can use various information criteria or cross-validation. The analysis is run in two stages: first the N principal components $\widehat{F}_t$ of X_{it} are estimated. Then, for the optimal p^* and r^*, a linear regression $y_{t+h} = \alpha_h + \beta_h(L)y_t + \gamma_h(L)\widehat{f}_t + \eta_{t+h}$ is run. The coefficients of this equation can be either estimated by OLS or by assuming normality and applying the Kalman filter as in Stock and Watson (2011).

Stock and Watson (1998) show for the system in equations (1) and (2) that under a set of moment conditions for (η, u, F) and an symptotic rank condition on Λ, the mean squared forecast error of this truncated model approaches the MSFE of the optimal infeasible model as $N, T \to \infty$. This means that regardless of the ratio of N to T (which can be important in many other high-dimensional frameworks), the feasible forecast will be close to optimal if we have enough data. In addition, the authors show that these forecasts remain consistent even in the presence of some time instability in the factor structure as long as $N >> T$. Further results by Bai and Ng (2006) show that $\widehat{f}_t$ can indeed be treated as observed data, although information criteria need to be modified. Estimates of β_h and γ_h will be $\sqrt{T}$ consistent.

Some improvements of the original framework of diffusion index forecasting include constructing principal components based on a matrix of only variables relevant for predicting

y_{t+h} (Bai and Ng, 2008), using gradient boosting to select the variables for this matrix (Bai and Ng, 2009) and exploiting the dynamic covariance structure of the panel (Forni et al., 2005).

Diffusion index modelling offers a very intuitive method of dimension reduction through the assumption that there are underlying latent factors driving the economic variables. However, a caveat of this approach is the linear structure of principal components and the imposed linearity of the second stage regression. Although Bai and Ng (2008) try to expand the variable selection set to both $\{X_{it} \quad X_{it}^2\}$ (omitting interaction terms due to computational time), this only allows for a specific kind of non-linearities and is not very satisfactory.

Other tools that economists have applied or proposed for this problem are Bayesian vector autoregressions, Bayesian model averaging (Fernández et al., 2001), frequentist model averaging (Kapetanios et al., 2006), bagging (Stock and Watson, 2012), boosting (Bai and Ng, 2009), least angle regression (Bai and Ng, 2008), Bayesian reduced rank regression (Carriero et al., 2011), variable preselection (Bai and Ng, 2006), simulated annealing and genetic algorithms (Kapetanios, 2007) and complete subset regressions (Elliott et al., 2013) (Elliott et al., 2015).

Evaluating such a wide set of models can be tricky. Due to publication bias, authors often cherry-pick data and consider only the simplest version of competing models. Fortunately, there is a tradition of comparing out-of-sample MSE to that of an AR(4) model. In addition, there are at least two large scale surveys (Kim and Swanson, 2014) (Stock and Watson, 2012) which compare the performance of these models. They find that on average the diffusion index models perform well, although sometimes shrinkage and model averaging can further improve the results. Overall, there is no clear dominance between different models. This is not surprising as linear regression is generally used as the final step of estimation. Indeed, Stock and Watson (2012) conclude that it is difficult to beat DFMs in a linear principal components framework and that ”further forecast improvements...will need to come from models with nonlinearities”. In the next section, we present a non-linear model with this aim in mind.

3 Random forest

Model selection in a high-dimensional framework is a new problem only in economics. In the fields of image processing, handwriting recognition etc, data sets with millions of variables are common. Genomic data analysts face the daunting problem of selecting relevant prediction variables in data sets of magnitude 20 × 1000. There has thus been a wealth of different options developed for problems of this kind.

One method that can simultaneously perform variable selection and non-parametric estimation in a high-dimensional space is random forest. It is based on the framework of Breiman (2001). In the machine learning context of *supervised learning* (which for all practical purposes is the same as forecasting), random forests have been hailed as the most successful general-purpose algorithm of modern times. Surveys done in the machine learning field confirm that in terms of prediction accuracy, random forests on average best all other methods such as neural networks, support vector machines, boosting, linear models, principal components etc (Fernández-Delgado et al., 2014). Below we take a closer look at what goes on inside this model.

As we are focused on out of sample forecasting performance, it is natural to look for a model that does not *overfit*. Ever since the original Makridakis competitions in the 80s, applied forecasters realized that too complex models introduce a lot of parameter estimation error and usually perform worse than more parsimonious models. In a statistical framework, this translates into looking for an optimal *bias-variance* tradeoff point. Creating a complex model, say by introducing hundreds of explanatory variables into an OLS specification will decrease the bias with respect to the true regression function, but will increase its variance.

Bagging, short for bootstrap aggregating is a procedure proposed by Breiman (1996) that averages an estimator over many bootstrapped samples. Bagging can lead to a reduction in the variance of an estimator, while leaving the bias unchanged. This is especially true in the context of a low bias-high variance type estimator such as a *regression tree* and when one deals with high-dimensional data, where there are multiple good model choices. Under squared error loss, it can be shown that population averaging never increases MSE, giving a strong theoretical justification for using bagging (Hastie et al., 2009).

A regression tree under the CART splitting criteria works as following [1]: given a sample $\mathcal{D}_T = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_T, Y_T))$, denote A a generic *cell* of the data, defined by some intervals of variable values. Let $N_T(A)$ be the number of data points in $\mathcal{D}_T$ which belong to A. A cut in A is the pair (j, z) where j takes value in $\{1, \ldots, N\}$ and z is the position of the cut along the j-th coordinate. Let $\mathcal{C}_A$ be the set of all possible cuts in A. Then if $\mathbf{X}_i = (\mathbf{X}_i^{(1)}, \ldots, \mathbf{X}_i^{(N)})$, the CART splitting criterion, sometimes called *variance reduction*, is given by

$$CART_T(j,z) = \frac{1}{N_T(A)}\sum_{i=1}^{T}(Y_i - \bar{Y}_A)^2 \mathbb{1}_{\mathbf{X}_i \in A} - \frac{1}{N_T(A)}\sum_{i=1}^{T}(Y_i - \bar{Y}_{A_L}\mathbb{1}_{\mathbf{X}_i^{(j)}<z} - \bar{Y}_{A_R}\mathbb{1}_{\mathbf{X}_i^{(j)}\geq z})^2 \mathbb{1}_{\mathbf{X}_i \in A} \tag{3}$$

where $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}, A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$ and $\bar{Y}_.$ is the mean of Y_i with $\mathbf{X}_i$ in A. We look for the optimal pair (j_n^*, z_n^*) to maximize $CART_T(j,z)$ in a cell A – giving the maximal variance in reduction. We start from a cell that contains the whole data and repeat this procedure until there are no cells left with more observations in them than parameter *nodesize*. Note that the final outcome will be a partition over $\mathbf{X}_i$, with predicted values the averages of these cells of 5 or less observations. The stage is now set for introducing the random forest procedure.

In simple words, we build fully grown trees on bootstrapped samples of the original data with a random subset of length *mtry* of all the variables chosen at each iteration in M. We split the tree according to the CART criterion and obtain predictions for the query point $\mathbf{x}$ by averaging the values of Y_i in its terminal node. We finally average over the M trees and obtain a function that maps from the covariate space to the predicted values. Even

[1] Notation due to Biau and Scornet (2016), combined with time series setting

though all of the trees are fully grown (indicated by the fact that only 5 or less observations are in the final partition cells i.e *nodesize* = 5), the procedure is notoriously resistant to overfitting (Hastie et al., 2009).

Intuitively, this is because we are averaging over many high-variance but low-bias trees. These are identically distributed, thus the expectation and bias of the average of M trees remains roughly the same. However, variance reduction can be achieved if the correlation between trees is not large. Random forest improves on bagging by decreasing the correlation due to randomly selecting a subset *mtry* of the regressors. The variance of the average of M identically distributed random variables is

$$\rho\sigma^2 + \frac{1-\rho}{M}\sigma^2$$

therefore for large M, the variance depends mostly on the correlation coefficient ρ.

Despite its known good performance, the theoretical underpinnings of random forest remain scarce. A simplified version, *centered forest*, where there is no resampling, the splitting variable is uniformly chosen and splits are done in the center of the cell, has been shown to be consistent for regression (Scornet et al., 2015). The rate of convergence of the centered forest depends only on the inner sparse dimension of the data (Biau, 2012), because it performs splits with a high probability on informative variables (Scornet, 2016). Wager (2014) proves consistency and asymptotic normality of a random forest where cuts are made along all variable directions and different data sets are used for building the tree and estimating values in leaves. An excellent review of these and other results is provided by Biau and Scornet (2016).

In addition to good forecasting performance, the random forest procedure offers several desirable properties. It accommodates categorical data in a straightforward manner, without having to resort to dummy variables — splits can be performed between the possible values. This can be useful for using survey data for forecasting. Survey data can offer superior measures of expectations compared to forward markets and is therefore valuable for forecasting exchange rates or GDP.

Random forest also accommodates high-dimensional data, as the one used in this paper. Indeed, even solitary regression trees are known to handle almost arbitrarily large (in the N dimension) data sets, as splits are made only based on a few variables. If the centered forest adaption to sparsity result holds for random forest as well, it becomes a formidable estimation method. In practice, this means that one can include thousands of variables monitored by e.g. the Federal Reserve Bank of St. Louis into the model and obtain accurate (relative to other methods) forecasts on quarterly or even yearly series.

The algorithm offers several useful measures as a by-product: *variable importance* which can be used as an input for variable selection for other models and *out of bag* error computed on model prediction error using all other trees as the testing set – a useful alternative to cross-validation for tuning (Hastie et al., 2009).

4 Data and Methodology

> Even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience. *David Hume*

It is clear that showing strict dominance in the forecasting performance between two (reasonable) models is difficult. Timmermann and Elliott (2016) show that even in the utterly unrealistic case where the researcher knows the data generating process and has to only estimate its parameters, different estimation methods are better for different true parameter values. In the same vein, no free lunch theorems from machine learning literature show that in general there is no best learning method.

In practice, a researcher either generates synthetic data from specific Monte Carlo experiments or looks into the forecasting performance on important macro variables. I take the latter approach and use the Stock and Watson (2012) dataset. They collect 143 quarterly U.S. macroeconomic series from 1959:3 to 2009:2, with earlier periods used for lagged values. In line with the authors, the series are grouped into 13 categories, presented in Table 9. Standard transformations based on differencing and logarithms are used to deal with I(1) and I(2) series, shown in Table 10. All series are either from the Global Insight Basic Economics Database, the Conference Board's Indicators Database or the calculations of Stock and Watson (2012).

This set of series contains variables that themselves are often subject to prediction: GDP, inflation and interest rates. Knowing the performance on these is then helpful in choosing the model for working with them. Moreover, I consider every series among the 143 series separately as a dependent variable. Since these have different underlying data generating processes, I alleviate the problem of dependence on a particular data. It has to be noted that these series contain a large part of the most important U.S. macro variables.

I consider the direct h step ahead forecasting strategy for $h = \{1, 2, 4, 8, 12\}$ — the longest horizon forecast is then 3 years ahead. Forecasts are based on a rolling window, rather than recursively, to deal with possible structural breaks in the time series (Pesaran et al., 2006). An estimation window of 100 quarters is used. Forecasts are then made in the following manner: for $h = 1$ step ahead, periods 1 to 100 are used for model estimation and predicting the value in 101. Periods 2 to 101 are then used for 102 etc. The mean squared error loss out of sample is used to rate the performance of a specific model. In line with Bai and Ng (2008), I report the relative mean-squared error given as

$$RMSE(model) = \frac{MSE(model)}{MSE(AR(4))}$$

This allows to compare results between different models, although some slight variations exist in the literature, i.e. Elliott et al. (2015) consider the root of this and Kim and Swanson (2014) uses $AR(p)$ chosen by Schwarz information criterion rather than $AR(4)$. However, MSE(AR(4)) and MSE(AR(p)) are likely quite close, in our application it appeared that

AR(4) generally produces even better estimates, perhaps due to the small estimation sample.[2] Reporting the relative mean-squared error is thus a convenient way to compare models across literature.

I consider three different models:

(1) RF – A random forest with 500 trees, node size 5 that chooses 48 variables at each iteration (default parameter values).

(2) DFM – A very simple DI specification where the first factor of the covariate matrix is added to the final linear regression. Technically, this is not a *dynamic* factor model, but we still refer to it as DFM.

(3) SVM – A default support vector machine with a radial basis kernel and cost parameter $C = 1$ (default parameter values).

The idea behind this model selection is the following: if there are indeed notable nonlinearities in the series, then models (1) and (3) should dominate over (2). Moreover, as random forest is known to be robust to parameter specification, we expect it to perform better in this untuned specification compared to a SVM. Further work will determine how much these models can benefit from parameter tuning.

It will also be interesting to see how (2) performs relative to AR(4). It is well known from the literature that if principal components are composed with many variables including idiosyncratic noise, the gains from the factor structure disappear (Bai and Ng, 2006). Thus, whether DFM outperforms AR(4) depends on whether there is useful additional information in the first factor.

5 Results

Table 1 presents results for the mean RMSE across 143 series for each of the 3 models. We see that the random forest forecast is the best for all forecast horizons except $h = 1$. SVM also on average outperforms DFM starting from $h = 4$. However, the simple DFM models seems to be consistently worse than the autoregressive one — it appears that adding the main principal component of the covariate series does not improve forecast accuracy. These results imply that improving upon an autoregressive model is not that straightforward, especially at shorter forecast horizons. However, in the long horizon there seem to be gains to be had from allowing for nonlinearities and considering the whole covariate space rather than the linear common factors of these.

The fact that AR(4) does relatively good in the short run can be explained by the higher amount of noise in short run predictions. In the long run, dependent series revert to their means that can depend on the covariates and this is better captured by a model that takes more than past lagged values into account. In the dynamic factor literature, this decrease

[2] Or the fact mentioned in the introduction, that consistent variable selection and forecast accuracy are different ends

	$h = 1$	$h = 2$	$h = 4$	$h = 8$	$h = 12$
RF	1.5672	**0.8992**	**0.7850**	**0.7047**	**0.6297**
DFM	1.3513	1.1806	1.1445	1.1009	1.0521
SVM	2.9105	1.2123	0.8808	0.8169	0.7801
AR	**1**	1	1	1	1

Table 1: RMSE over 143 series

of RMSE with respect to forecast horizon is well documented Bai and Ng (2008). Now we show that this also holds for non-linear models.

Tables 4-8 in the Appendix depict these results in more detail, dividing the series into 13 categories and showing the number of series where the model improves upon AR(4). In the 1-step ahead setting, AR(4) remains best for most categories. However, it is interesting to note that this does not hold for price series, where RF beats AR(4) 33 times out of 37, stock prices (5 out of 5) and exchange rates (5/5). These are generally considered very difficult series to predict even for a well-specified dynamic factor model. It also improves slightly upon AR(4) consistently in the inventories (6/6) category. SVM generally outperforms AR(4) in the same categories as RF, but performs worse on average.

These results imply that for forecasting volatile series which are swift to change when underlying conditions change such as prices or exchange rates, a random forest based model can improve markedly on autoregressive and DFM ones. It appears that these series are highly non-linear.

For the 2-step ahead table, random forest begins to show improvement: it now outperforms AR(4) on 105 out of the 143 series, winning on average in most categories. Even the series where it loses to AR(4), the average RMSE is quite close (with the notable exception of interest rates). This suggest that starting from a forecast horizon of 6 months, RF is generally a better forecasting tool than an autoregressive model.

This trend in performance continues over 4-step, 8-step and 12-step ahead horizons. Random forest becomes the clear winner here, winning on 129,142 and 142 series respectively. Even SVM wins on 132/143 series for 12-step ahead horizon. This implies again that in the long run, the association between the covariates and dependent variable becomes more pronounced. The relatively disastrous performance of our simple DFM model implies that this association can be non-linear and complex.

Slightly troubling in adapting random forest as a go-to tool for forecasting is it's relatively bad performance in the 1-step ahead forecast. Indeed, for categories such as industrial production, the model wins on 0 out of the 14 series (interest rates 2/13), with the average RMSE roughly 3 times larger than for AR(4)! Further research will have to determine the exact cause of these results. The author suspects that due to the relatively higher noise in short horizon, the model might be overfitting. This is also suggested by the fact that SVM, a machine learning method known to be more prone than RF to overfitting, also fails on these categories and has worse RMSE performance. A viable solution might be introducing an *insanity filter*, which censors forecasts that change too radically. Alternatively, we could limit ourselves to linear models for short-term forecasting, this is discussed more below.

Finally, let us note the relatively excellent performance of RF for consumer expectations. This is a series which is used as an input to finance or macro applications and forecasting it well can be of its own interest.

To further look into the forecast performance, I examine in more detail 3 series: inflation, GDP growth and unemployment rate. This is done for three reasons: these series are important for private and public decision-makers, they are known to be non-trivial to forecast (especially inflation) and most importantly, this selection is in line with Elliott et al. (2015), Bai and Ng (2008) and Bai and Ng (2009) so we can compare results with their models.

In this part of the analysis, I test formally whether a forecasting model has better performance compared to AR(4) using a one-sided Diebold-Mariano test (Diebold and Mariano, 1995) – a significant result indicates the model is better than AR(4). In addition to the untuned versions of models considered before, 10-fold cross-validation to select optimal model parameters was considered. In the case of DFM, I let the number of lags of principal components selected into the final regression to be the tunable parameter. Random forests's tuning parameter is *mtry*, the number of variables randomly selected at each iteration and the support vector machine's is the cost function parameter C which controls the trade-off between model complexity and in-sample fit (Cortes and Vapnik, 1995). However, no model benefitted significantly from this and these are omitted.

$Model_h$	GDP	Unemployment	Inflation
RF_1	1.1254	1.5255	0.635
RF_2	**0.8518**	**0.7302*****	0.5321*
RF_4	**0.78***	**0.6058*****	**0.642*****
RF_8	**0.8143**	**0.5518*****	**0.632*****
RF_{12}	**0.8865**	**0.5139*****	**0.5919*****
DFM_1	1.0227	1.0755	1.0691
DFM_2	1.21	1.0628	1.0832
DFM_4	1.3285	1.1226	1.1471
DFM_8	1.3961	1.2035	1.1624
DFM_{12}	1.2123	1.0743	1.1872
SVM_1	1.6962	2.1335	**0.5615***
SVM_2	1.206	1.1155	**0.4934***
SVM_4	0.8282	0.6425***	0.6626***
SVM_8	0.9691	0.6774**	0.7417**
SVM_{12}	1.0091	0.7222	0.7582**
$Elliot_1$	0.6606**	0.5708***	0.9178
$Elliot_2$	0.6477***	0.5417***	0.9004
$Elliot_3$	0.7465**	0.5667***	0.9366
$Elliot_4$	0.795**	0.6258***	0.9185

Table 2: RMSE for 3 macroeconomic series

The results are presented in Table 2. The stars are related to the one-sided DM test p-values, with three-two-one indicating cut-off levels of $0.01, 0.05, 0.1$ respectively. It appears that the DFM specification is not significantly better than an AR(4) benchmark. This is in line with Elliott et al. (2015) and Kim and Swanson (2014). On the other hand, both RF and SVM are clearly better, with RF taking the cake for most cases.

Comparable results from Elliot's complete subset regression Elliott et al. (2013) are given. This is considered a state of the art method that improves upon autoregressive and dynamic factor models and even the benchmark of the ensemble of 2^N linear models. In essence, it consists of choosing $n << N$ as the fixed number of predictor variables, estimating all possible $N!(N-n)!n!$ models and combining their forecast with equal weights. It is a flexible linear ensemble model that can be used on high-dimensional data, but one drawback of it is its high computational cost, especially when multiple n are considered. I choose the $n = 10$ benchmark from Elliott et al. (2015), within the value range suggested by the author.

We see that Elliott's CSR performs notably better for GDP but much worse for inflation. This is further proof to the results we observed in tables 4-8: price series seem to be non-linear. Moreover, already in those tables the GDP components series were the ones with which RF struggled, overcoming AR(4) only from a 4 quarter horizon. The results from the unemployment series are more peculiar. RF totally fails 1-step head forecast, but after that RMSE becomes smaller with h, in line with most of the 143 series we observe. CSR (and actually also DFM) tend to forecast best at the 1-quarter horizon, with performance diminishing! For unemployment, around 4 quarters seems to be the level from which RF stars outperforming CSR.

These results suggest that it is not the additional information from covariates that drives the decreasing RMSEs of RF, but it is because relationships between variables become more non-linear as forecast horizon increases. Based on this, it would appear that for very short horizons, CSR is optimal but if one looks beyond a year, then gains can be had from considering random forest. However, when we take into account the substantially higher computational cost of CSR and recall that dynamic factor models on general do not even outperform AR(p) (Elliott et al., 2015), (Kim and Swanson, 2014), then random forest becomes an attractive go-to tool for *all* forecasting, if the issue of 1-step ahead performance can be fixed. However, these are just preliminary results, further work needs to be done on a larger set of macroeconomic series. This can be again difficult due to CSR's computational demands. What seems certain is that for all horizons, random forest outperforms linear models for series related to prices.

To gauge whether it is indeed non-linearity that drives the forecasting performance between the models, I conclude with two non-linearity tests. Firstly, I apply a Ramsey RESET test directly on the DFM equation to see if gains could be had from using different powers of the factors. I consider powers from 2 to 10 on the regressors. Intuitively, this tests whether the common linear components of covariates and lags are related to the forecasted variable in a non-linear way. The test result is presented in Table 3. The test statistic is given and stars indicated probabilities on the usual 0.1, 0.05, 0.01 scale.

h	RESET			RF-PCDM		
	G	U	I	G	U	I
1	1.36*	1.68***	2.42***	1.82***	2.28***	0.53
2	1.83***	2.44***	1.77***	1.35**	1.44***	0.5
4	1.5**	3.41***	1.48*	1.18**	1.05**	0.61
8	1.06	1.77***	-Inf?	1.03	0.87***	0.69
12	0.87	1.28	1.72***	0.91	0.73*	0.71*

Table 3: Nonlinearity tests

The RESET test indicates that there are indeed significant non-linearities in the series.[3] However, from the author's experience, it matters a lot whether the alternative hypothesis is specified with respect to the fitted values, regressors or principal components, with different non-linearities found in every case. The test is therefore sensitive to the *type* of non-linearity. The simple power terms considered do not encompass all arbitrary non-linearity types, therefore a non-significant result does not imply linearity. However, a significant one shows that there is non-linearity and we know in which form.

A fully non-parametric procedure such as random forest can adapt to any type of non-linearity and interactions due to the nature of the algorithm and splitting process. Therefore, even if the null is not rejected, it might be due to some arbitrary form of non-linearity which random forest can detect.

With this in mind, I propose a test based on the random forest forecasting results:

1. Run a random forest on all the variables
2. Run a random forest on all principal components
3. Compare forecast results via a one-sided Diebold-Mariano test

If using factors hinders the forecasting performance of the model, it must mean that there is a significant part of information lost in the DFM procedure. The RMSEs of the principal component forest and DM-test statistic significances are presented in the right panel of Table 3 (RF-PCDM). These suggest that for GDP and unemployment, using principal components results in significant forecasting accuracy reduction, especially at shorter horizons. In general, using principal components does not improve the forecasting accuracy notably. This suggests that it is not enough to construct factors and estimate their effect non-parametrically in the last stage regression — the whole estimation procedure would have to be non-parametric. Random forest offers a convenient way to do this in high dimensions.

[3] Unknown bug for inflation at $h = 4$

6 Conclusion

To sum up, we find that adapting random forest as the benchmark forecast tool for macro problems starting from a horizon of 6 months can lead to notable accuracy gains compared to autoregressive and dynamic factor specifications. For shorter horizons, the model is almost on par with top notch linear specifications (and notably less computationally demanding). If the potential very short horizon overfitting problem can be solved, it could be adapted as the benchmark tool for any forecasting horizon.

Moreover, we find that for certain series such as prices and exchange rates, which have been deemed simply very hard to forecast in the linear literature, random forest exhibits excellent performance in any forecasting horizon. For these series, nonlinear methods such as RF seem to be almost always better than commonly used linear models.

Further research has to look into whether the algorithm is overfitting in the short horizon and propose modifications to solve this issue. One possible solution might be introducing an insanity filter, that censors forecasts which change too much.

References

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

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.

Biau, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.*, 13(1):1063–1095.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2):197–227.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

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

Carriero, A., Kapetanios, G., and Marcellino, M. (2011). Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, 26(5):735–761.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

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

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

Elliott, G., Gargano, A., and Timmermann, A. (2015). Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, 54:86 – 110.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181.

Fernández, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381 – 427.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100(471):830–840.

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory*, null:60–68.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Kapetanios, G. (2007). Variable selection in regression models using nonstandard optimisation of information criteria. *Computational Statistics & Data Analysis*, 52(1):4 – 15.

Kapetanios, G., Labhard, V., and Price, S. (2006). Forecasting using predictive likelihood model averaging. *Economics Letters*, 91(3):373 – 379.

Keenan, D. M. (1985). A tukey nonadditivity-type test for time series nonlinearity. *Biometrika*, 72(1):39–44.

Kim, H. H. and Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, 178, Part 2:352 – 367. Recent Advances in Time Series Econometrics.

Lee, T.-H., White, H., and Granger, C. W. (1993). Testing for neglected nonlinearity in time series models. *Journal of Econometrics*, 56(3):269 – 290.

Mol, C. D., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318 – 328. Honoring the research contributions of Charles R. Nelson.

Ng, S. (2013). Chapter 14 - variable selection in predictive regressions. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, Part B of *Handbook of Economic Forecasting*, pages 752 – 789. Elsevier.

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *The Review of Economic Studies*, 73(4):1057–1084.

Sargent, T. J. and Sims, C. A. (1977). Business cycle modeling without pretending to have too much a-priori economic theory. *New Methods in Business Cycle Research*.

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72 – 83. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *Ann. Statist.*, 43(4):1716–1741.

Stock, J. H. and Watson, M. W. (1998). Diffusion indexes. Working Paper 6702, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2):293 – 335.

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

Stock, J. H. and Watson, M. W. (2011). Dynamic factor models. In Clements, M. P. and Hendry, D. F., editors, *The Oxford Handbook of Economic Forecasting*. Oxford University Press, Oxford.

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493.

Teräsvirta, T., Lin, C.-F., and Granger, C. W. J. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, 14(2):209–220.

Timmermann, A. and Elliott, G. (2016). *Economic Forecasting*. Princeton University Press.

Tsay, R. S. (1986). Nonlinearity tests for time series. *Biometrika*, 73(2):461–466.

Wager, S. (2014). Asymptotic Theory for Random Forests. *ArXiv e-prints*.

Wang, Y. and Choi, I.-C. (2013). Market Index and Stock Price Direction Prediction using Machine Learning Techniques: An empirical study on the KOSPI and HSI. *ArXiv e-prints*.

7 Appendix

	Code	# series	RF	Impr.	DFM	Impr.	SVM	Impr.	AR
GDP components	1	16	1.3756	5	1.3665	1	1.8323	2	**1**
IP	2	14	3.5757	0	1.3808	1	7.8003	0	**1**
Employment	3	20	1.8883	9	1.4015	2	5.3140	10	**1**
Unempl. rate	4	7	1.8397	2	1.5192	1	4.3841	1	**1**
Housing	5	6	1.2006	2	1.5235	1	1.8398	1	**1**
Inventories	6	6	0.5693	6	1.0132	2	**0.5454**	6	1
Prices	7	37	**0.7425**	33	1.1758	18	0.7693	29	1
Wages	8	6	1.3278	1	1.2977	1	1.5140	1	**1**
Interest rates	9	13	2.9689	2	1.9328	1	5.3363	1	**1**
Money	10	7	1.2271	3	1.4979	0	1.5172	1	**1**
Exchange rates	11	5	0.9315	5	**0.9188**	4	1.0118	1	1
Stock prices	12	5	**0.6189**	5	1.0859	2	0.6688	4	1
Cons. exp	13	1	**0.3994**	1	1.2306	0	0.4505	1	1
Total		143		74		34		58	

Table 4: 1-step ahead RMSE

	Code	# series	RF	Impr.	DFM	Impr.	SVM	Impr.	AR(4)
GDP components	1	16	1.0017	10	1.2528	4	1.2527	3	**1**
IP	2	14	1.1164	6	1.0584	5	1.7205	2	**1**
Employment	3	20	**0.9407**	15	1.1750	6	1.4529	8	1
Unempl. rate	4	7	**0.9367**	5	1.2934	1	1.4191	2	1
Housing	5	6	1.0456	2	1.2876	2	1.3495	1	**1**
Inventories	6	6	**0.5538**	6	1.2306	3	0.6536	5	1
Prices	7	37	**0.7080**	35	1.0968	13	0.7484	31	1
Wages	8	6	**0.9884**	4	1.4095	0	1.1324	1	1
Interest rates	9	13	1.2940	4	1.3663	1	2.0607	1	**1**
Money	10	7	**0.7345**	7	1.1931	1	0.9763	4	1
Exchange rates	11	5	**0.8153**	5	1.0127	2	1.0132	1	1
Stock prices	12	5	**0.6114**	5	1.0229	1	0.6511	5	1
Cons. exp	13	1	**0.7440**	1	0.9666	1	1.7898	0	1
Total		143		105		40		64	

Table 5: 2-step ahead RMSE

	Code	# series	RF	Impr.	DFM	Impr.	SVM	Impr.	AR(4)
GDP components	1	16	**0.9012**	11	1.2854	2	0.9824	9	1
IP	2	14	**0.8216**	13	1.0804	4	0.9733	11	1
Employment	3	20	**0.7916**	18	1.0615	9	0.9494	13	1
Unempl. rate	4	7	**0.8665**	7	1.3295	0	1.0200	4	1
Housing	5	6	**0.9017**	5	1.2345	1	1.0467	3	1
Inventories	6	6	**0.5890**	6	1.0741	2	0.6498	6	1
Prices	7	37	**0.7338**	35	1.1148	11	0.7524	33	1
Wages	8	6	**0.7254**	6	1.2064	0	0.8214	4	1
Interest rates	9	13	**0.9168**	10	1.2235	0	1.0675	7	1
Money	10	7	**0.6431**	7	1.1651	1	0.7240	7	1
Exchange rates	11	5	**0.7340**	5	0.8857	5	0.8422	5	1
Stock prices	12	5	**0.6463**	5	1.1671	2	0.6960	5	1
Cons. exp	13	1	**0.6706**	1	0.7673	1	0.9056	1	1
Total		143		129		38		108	

Table 6: 4-step ahead RMSE

	Code	# series	RF	Impr.	DFM	Impr.	SVM	Impr.	AR(4)
GDP components	1	16	**0.8004**	16	1.2933	0	0.9059	13	1
IP	2	14	**0.7685**	14	1.0477	4	0.8972	11	1
Employment	3	20	**0.6687**	20	1.0059	10	0.8040	16	1
Unempl. rate	4	7	**0.7450**	7	1.1508	0	0.9036	6	1
Housing	5	6	**0.7592**	5	1.0502	3	0.9113	3	1
Inventories	6	6	**0.6016**	6	1.1134	2	0.6729	6	1
Prices	7	37	**0.6791**	37	1.0882	12	0.7342	34	1
Wages	8	6	**0.6188**	6	1.1359	2	0.7413	6	1
Interest rates	9	13	**0.7797**	13	1.2100	1	0.9552	9	1
Money	10	7	**0.6251**	7	1.1522	2	0.7667	7	1
Exchange rates	11	5	**0.6657**	5	0.8178	5	0.8631	4	1
Stock prices	12	5	**0.6008**	5	1.0256	3	0.6720	5	1
Cons. exp	13	1	**0.7682**	1	0.8259	1	0.7802	1	1
Total		143		142		45		121	

Table 7: 8-step ahead RMSE

	Code	# series	RF	Impr.	DFM	Impr.	SVM	Impr.	AR(4)
GDP components	1	16	**0.6952**	16	1.1627	2	0.8442	13	1
IP	2	14	**0.7522**	13	0.9855	6	0.9035	11	1
Employment	3	20	**0.5576**	20	0.9674	12	0.7360	20	1
Unempl. rate	4	7	**0.6667**	7	1.1082	1	0.8436	7	1
Housing	5	6	**0.6237**	6	0.9698	3	0.8800	4	1
Inventories	6	6	**0.5237**	6	1.0538	2	0.6537	6	1
Prices	7	37	**0.6248**	37	1.1166	18	0.6975	36	1
Wages	8	6	**0.5725**	6	1.0519	3	0.7156	6	1
Interest rates	9	13	**0.6773**	13	1.0537	6	0.9211	11	1
Money	10	7	**0.5828**	7	1.0747	2	0.7629	7	1
Exchange rates	11	5	**0.5565**	5	0.7925	5	0.8283	5	1
Stock prices	12	5	**0.5744**	5	1.0176	3	0.6723	5	1
Cons. exp	13	1	**0.5961**	1	0.9092	1	0.6529	1	1
Total		143		142		64		132	

Table 8: 12-step ahead RMSE

Code	Transf.	Cat.	Description
IPS10	5	2	INDUSTRIAL PRODUCTION INDEX - TOTAL INDEX
IPS11	5	2	INDUSTRIAL PRODUCTION INDEX - PRODUCTS, TOTAL
IPS299	5	2	INDUSTRIAL PRODUCTION INDEX - FINAL PRODUCTS
IPS12	5	2	INDUSTRIAL PRODUCTION INDEX - CONSUMER GOODS
IPS13	5	2	INDUSTRIAL PRODUCTION INDEX - DURABLE CONSUMER GOODS
IPS18	5	2	INDUSTRIAL PRODUCTION INDEX - NONDURABLE CONSUMER GOODS
IPS25	5	2	INDUSTRIAL PRODUCTION INDEX - BUSINESS EQUIPMENT
IPS32	5	2	INDUSTRIAL PRODUCTION INDEX - MATERIALS
IPS34	5	2	INDUSTRIAL PRODUCTION INDEX - DURABLE GOODS MATERIALS
IPS38	5	2	INDUSTRIAL PRODUCTION INDEX - NONDURABLE GOODS MATERIALS
IPS43	5	2	INDUSTRIAL PRODUCTION INDEX - MANUFACTURING (SIC)
IPS306	5	2	INDUSTRIAL PRODUCTION INDEX - FUELS
PMP	1	2	NAPM PRODUCTION INDEX (PERCENT)
UTL11	1	2	CAPACITY UTILIZATION - MANUFACTURING (SIC)
CES275R	5	8	REAL AVG HRLY EARNINGS, PROD WRKRS, NONFARM - GOODS-PRODUCING (CES275/PI071)
CES277R	5	8	REAL AVG HRLY EARNINGS, PROD WRKRS, NONFARM - CONSTRUCTION (CES277/PI071)
CES278.R	5	8	REAL AVG HRLY EARNINGS, PROD WRKRS, NONFARM - MFG (CES278/PI071)
CES002	5	3	EMPLOYEES, NONFARM - TOTAL PRIVATE
CES003	5	3	EMPLOYEES, NONFARM - GOODS-PRODUCING
CES006	5	3	EMPLOYEES, NONFARM - MINING
CES011	5	3	EMPLOYEES, NONFARM - CONSTRUCTION
CES015	5	3	EMPLOYEES, NONFARM - MFG
CES017	5	3	EMPLOYEES, NONFARM - DURABLE GOODS
CES033	5	3	EMPLOYEES, NONFARM - NONDURABLE GOODS
CES046	5	3	EMPLOYEES, NONFARM - SERVICE-PROVIDING
CES048	5	3	EMPLOYEES, NONFARM - TRADE, TRANSPORT, UTILITIES
CES049	5	3	EMPLOYEES, NONFARM - WHOLESALE TRADE
CES053	5	3	EMPLOYEES, NONFARM - RETAIL TRADE
CES088	5	3	EMPLOYEES, NONFARM - FINANCIAL ACTIVITIES
CES140	5	3	EMPLOYEES, NONFARM - GOVERNMENT
LHEL	2	3	INDEX OF HELP-WANTED ADVERTISING IN NEWSPAPERS (1967=100;SA)
LHELX	2	3	EMPLOYMENT: RATIO; HELP-WANTED ADS:NO. UNEMPLOYED CLF
LHEM	5	3	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS.,SA)
LHNAG	5	3	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC.INDUSTRIES (THOUS.,SA)
LHUR	2	4	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%,SA)
LHU680	2	4	UNEMPLOY.BY DURATION: AVERAGE(MEAN)DURATION IN WEEKS (SA)
LHU5	5	4	UNEMPLOY.BY DURATION: PERSONS UNEMPL.LESS THAN 5 WKS (THOUS.,SA)
LHU14	5	4	UNEMPLOY.BY DURATION: PERSONS UNEMPL.5 TO 14 WKS (THOUS.,SA)
LHU15	5	4	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 WKS + (THOUS.,SA)
LHU26	5	4	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 TO 26 WKS (THOUS.,SA)
LHU27	5	4	UNEMPLOY.BY DURATION: PERSONS UNEMPL.27 WKS + (THOUS,SA)
CES151	1	3	AVG WKLY HOURS, PROD WRKRS, NONFARM - GOODS-PRODUCING
CES155	2	3	AVG WKLY OVERTIME HOURS, PROD WRKRS, NONFARM - MFG
HSBR	4	5	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS.,SAAR)
HSFR	4	5	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM&NONFARM(1959-)(THOUS.,SA
HSNE	4	5	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.

HSMW	4	5	HOUSING STARTS:MIDWEST(THOUS.U.)S.A.
HSSOU	4	5	HOUSING STARTS:SOUTH (THOUS.U.)S.A.
HSWST	4	5	HOUSING STARTS:WEST (THOUS.U.)S.A.
FYFF	2	9	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)
FYGM3	2	9	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)
FYGM6	2	9	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)
FYGT1	2	9	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)
FYGT5	2	9	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)
FYGT10	2	9	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)
FYAAAC	2	9	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)
FYBAAC	2	9	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)
Sfygm6	1	9	fygm6-fygm3
Sfygt1	1	9	fygt1-fygm3
Sfygt10	1	9	fygt10-fygm3
sFYAAAC	1	9	FYAAAC-Fygt10
sFYBAAC	1	9	FYBAAC-Fygt10
FM1	6	10	MONEY STOCK: M1(CURR,TRAV.CKS,DEM DEP,OTHER CK'ABLE DEP)(BIL$,SA)
MZMSL	6	10	MZM (SA) FRB St. Louis
FM2	6	10	MONEY STOCK:M2(M1+O'NITE RPS,EURO$,G/P&B/D MMMFS&SAV&SM TIME DEP(BIL$,
FMFBA	6	10	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES(MIL$,SA)
FMRRA	6	10	DEPOSITORY INST RESERVES:TOTAL,ADJ FOR RESERVE REQ CHGS(MIL$,SA)
BUSLOANS	6	10	Commercial and Industrial Loans at All Commercial Banks (FRED) Billions $ (SA)
CCINRV	6	10	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)
CPIAUCSL	6	7	CPI All Items (SA) Fred
CPILFESL	6	7	CPI Less Food and Energy (SA) Fred
PCEPILFE	6	7	PCE Price Index Less Food and Energy (SA) Fred
PSCCOMR	5	7	Real SPOT MARKET PRICE INDEX:BLS & CRB: ALL COMMODITIES(1967=100) (PSCCOM/PCEPILFE)
PW561R	5	7	PPI Crude (Relative to Core PCE) (pw561/PCEPiLFE)
PMCP	1	7	NAPM COMMODITY PRICES INDEX (PERCENT)
EXRUS	5	11	UNITED STATES;EFFECTIVE EXCHANGE RATE(MERM)(INDEX NO.)
EXRSW	5	11	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)
EXRJAN	5	11	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.$)
EXRUK	5	11	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)
EXRCAN	5	11	FOREIGN EXCHANGE RATE: CANADA (CANADIAN $ PER U.S.$)
FSPCOM	5	12	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)
FSPIN	5	12	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)
FSDXP	2	12	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)
FSPXE	2	12	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%,NSA)
FSDJ	5	12	COMMON STOCK PRICES: DOW JONES INDUSTRIAL AVERAGE
HHSNTN	2	13	U. OF MICH. INDEX OF CONSUMER EXPECTATIONS(BCD-83)
PMI	1	6	PURCHASING MANAGERS' INDEX (SA)
PMNO	1	6	NAPM NEW ORDERS INDEX (PERCENT)
PMDEL	1	6	NAPM VENDOR DELIVERIES INDEX (PERCENT)
PMNV	1	6	NAPM INVENTORIES INDEX (PERCENT)
MOCMQ	5	6	NEW ORDERS (NET) - CONSUMER GOODS & MATERIALS, 1996 DOLLARS (BCI)
MSONDQ	5	6	NEW ORDERS, NONDEFENSE CAPITAL GOODS, IN 1996 DOLLARS (BCI)
GDP251	5	1	Real Gross Domestic Product, Quantity Index (2000=100) , SAAR
GDP252	5	1	Real Personal Consumption Expenditures, Quantity Index (2000=100) , SAAR
GDP253	5	1	Real Personal Consumption Expenditures - Durable Goods , Quantity Index (2000=
GDP254	5	1	Real Personal Consumption Expenditures - Nondurable Goods, Quantity Index (200
GDP255	5	1	Real Personal Consumption Expenditures - Services, Quantity Index (2000=100) ,
GDP256	5	1	Real Gross Private Domestic Investment, Quantity Index (2000=100) , SAAR
GDP257	5	1	Real Gross Private Domestic Investment - Fixed Investment, Quantity Index (200
GDP258	5	1	Real Gross Private Domestic Investment - Nonresidential , Quantity Index (2000
GDP259	5	1	Real Gross Private Domestic Investment - Nonresidential - Structures, Quantity
GDP260	5	1	Real Gross Private Domestic Investment - Nonresidential - Equipment & Software
GDP261	5	1	Real Gross Private Domestic Investment - Residential, Quantity Index (2000=100
GDP263	5	1	Real Exports, Quantity Index (2000=100) , SAAR
GDP264	5	1	Real Imports, Quantity Index (2000=100) , SAAR
GDP265	5	1	Real Government Consumption Expenditures & Gross Investment, Quantity Index (2
GDP266	5	1	Real Government Consumption Expenditures & Gross Investment - Federal, Quantit
GDP267	5	1	Real Government Consumption Expenditures & Gross Investment - State & Local, Q
LBOUT	5	8	OUTPUT PER HOUR ALL PERSONS: BUSINESS SEC(1982=100,SA)
LBPUR7	5	8	REAL COMPENSATION PER HOUR,EMPLOYEES:NONFARM BUSINESS(82=100,SA)
LBMNU	5	3	HOURS OF ALL PERSONS: NONFARM BUSINESS SEC (1982=100,SA)
LBLCPU	5	8	UNIT LABOR COST: NONFARM BUSINESS SEC (1982=100,SA)
GDP272A	6	7	Gross domestic product Price Index
GDP273A	6	7	Personal consumption expenditures Price Index
GDP274A	6	7	Durable goods Price Index
GDP274_1	6	7	Motor vehicles and parts Price Index
GDP274_2	6	7	Furniture and household equipment Price Index
GDP274_3	6	7	Other Price Index
GDP275A	6	7	Nondurable goods Price Index
GDP275_1	6	7	Food Price Index
GDP275_2	6	7	Clothing and shoes Price Index
GDP275_3	6	7	Gasoline, fuel oil, and other energy goods Price Index
GDP275_4	6	7	Other Price Index
GDP276A	6	7	Services Price Index
GDP276_1	6	7	Housing Price Index
GDP276_2	6	7	Household operation Price Index
GDP276_3	6	7	Electricity and gas Price Index
GDP276_4	6	7	Other household operation Price Index

GDP276_5	6	7	Transportation Price Index
GDP276_6	6	7	Medical care Price Index
GDP276_7	6	7	Recreation Price Index
GDP276_8	6	7	Other Price Index
GDP277A	6	7	Gross private domestic investment Price Index
GDP278A	6	7	Fixed investment Price Index
GDP279A	6	7	Nonresidential Price Index
GDP280A	6	7	Structures
GDP281A	6	7	Equipment and software Price Index
GDP282A	6	7	Residential Price Index
GDP284A	6	7	Exports Price Index
GDP285A	6	7	Imports Price Index
GDP286A	6	7	Government consumption expenditures and gross investment Price Index
GDP287A	6	7	Federal Price Index
GDP288A	6	7	State and local Price Index

Table 9: Series descriptions

Transformation Code	X_t	Y^h_{t+h}
1	Z_t	Z_{t+h}
2	$Z_t - Z_{t-1}$	$Z_{t+h} - Z_t$
3	$(Z_t - Z_{t-1}) - (Z_{t-1} - Z_{t-2})$	$h^{-1}(Z_{t+h} - Z_t) - (Z_t - Z_{t-1})$
4	$\ln(Z_t)$	$\ln(Z_{t+h})$
5	$\ln(Z_t/Z_{t-1})$	$\ln(Z_{t+h}/Z_t)$
6	$\ln(Z_t/Z_{t-1}) - \ln(Z_{t-1}/Z_{t-2})$	$h^{-1}(\ln(Z_{t+h}/Z_t)) - \ln(Z_t/Z_{t-1})$

Table 10: Transformations