Weekly Nowcasting US Inflation with Enhanced Random Forests*

Todd E. Clark**
Seton Leonard***
Massimiliano Marcellino§
Philipp Wegmüller¶

Abstract

We develop a random forest model which allows for mixed frequencies and missing observations. Further, we generalize the expected value of the target variable at each node of the regression tree to allow for a linear relationship between target and predictors. We apply the proposed framework to provide weekly nowcasts of U.S. inflation using a large set of daily, weekly, and monthly data. The resulting predictions are significantly better than simple univariate benchmarks for US inflation rate indices and match or exceed the accuracy of the nowcasts published by the Cleveland Fed.

This version: September 23, 2022
Preliminary draft

JEL class: C22, C53, E31, E37

Keywords: Big Data, Inflation, Forecasting, Random Forests, Machine Learning, Mixed-Frequencies

*We gratefully acknowledge helpful comments from participants at the 2022 ITISE meetings. The views, opinions, findings, and conclusions or recommendations expressed in this paper are strictly those of the authors. They do not necessarily reflect the views of the Federal Reserve Bank of Cleveland, the Federal Reserve System, or the State Secretariat for Economic Affairs (SECO). SECO does not take responsibility for any errors or omissions in, or for the correctness of, the information contained in this paper.

**Federal Reserve Bank of Cleveland todd.clark@researchfed.org
***System2 seth@sstm2.com
§Bocconi University, CEPR, IGIER, BIDSA, and BAFFI massimiliano.marcellino@unibocconi.it
¶State Secretariat for Economic Affairs, Short Term Economic Analyses, Holzikofenweg 36, 3003 Bern, Switzerland. philipp.wegmueller@seco.admin.ch

1 Introduction

Inflation forecasting has long been understood to be practically important and challenging. For example, Bernanke (2007) describes the importance of inflation modeling and forecasting to monetary policy. Many studies, such as Faust and Wright (2013) and Knotek and Zaman (2017), discuss and document challenges in inflation forecasting. The surge of inflation following the COVID-19 pandemic has of course added to practical interest in inflation forecasting.

One component of inflation forecasting is the problem of nowcasting inflation — that is, to predict inflation in the current period (month or quarter), before it is published. These short-horizon forecasts are not only of interest in their own right but also important inputs to the accuracy of forecasts at longer horizons (see, e.g., Faust and Wright (2009), Faust and Wright (2013), and Krüger et al. (2017)). In this paper, we focus on nowcasting inflation in US consumer prices, although our method can be more generally applied. Like much of the literature, our analysis is based on point forecasts. See Knotek and Zaman (2020) and references therein for work on density forecasts of inflation.

In the broader economic literature on nowcasting, many studies have focused on models — such as factor models or vector autoregressions — that can be represented in state space form, using estimation methods such as the EM Algorithm described by Watson and Engle (1983). Later contributions, more specifically geared towards mixed frequency applications, include Durbin and Koopman (2012) and Giannone et al. (2008). These models have proven adept at real time analysis when differences in frequencies are moderate. They are particularly useful for handling large (we are speaking of >100 series, as opposed to gigabytes or terabytes of data in some machine learning applications) data sets; frequentist models already employ dimension reduction to moderate overfitting issues, and Bayesian applications use shrinkage as well to improve out-of-sample performance, as in Bańbura et al. (2010) and Cimadomo et al. (2021). Modugno (2013) provides an application of this modeling approach to inflation. Inputs to the model include gasoline prices (weekly), oil prices (weekly), and prices of raw materials (daily), as well as aggregate and disaggregate measures of inflation. This modeling approach performs well when all variables are relevant to the target variable, in this case aggregate inflation. However, dynamic factor models (DFMs) do not select, that is, set parameters on less important variables to zero. Instead, a DFM models comovements in all input series, though the use of priors can shrink the model towards certain variables. Second, DFMs are conditionally linear. One can model regime changes or stochastic volatility, but this substantially increases computational complexity, and the ability of DFMs to incorporate non-linearities remains limited.

An alternative, regression based, approach to handling mixed frequency data was in-

troduced by Ghysels et al. (2004). In MIDAS regressions a low frequency target variable is regressed on a temporally aggregated high frequency variable, where temporal aggregation weights are constrained by specific parametric functions and optimally determined together with the other model parameters. A major advantage of MIDAS regressions, besides simplicity, is their ability to handle large frequency mismatches. A disadvantage is the non-linearity of the model, due to the non-linear temporal aggregation, which typically prevents its use with several regressors (though penalized versions are possible, see Mogliani and Simoni (2021)) and complicates the analysis of possible nonlinearity and time variation in the effects of the explanatory variables on the target. MIDAS regressions have also been used to nowcast inflation; see, in particular, Breitung and Roling (2015), who use daily data and, in addition, consider a non-parametric approach to determine the aggregation weights.

Knotek and Zaman (2017) develop simpler but effective models to produce daily nowcasts of US headline and core consumer inflation using a small number of carefully selected indicators at different frequencies. Their models feature time-varying weights on the available variables, with higher frequency indicators only used when sufficient data are available to make them informative for forecasting the target. The models, either univariate or simple multivariate specifications, are estimated over short rolling windows. Empirically, the short rolling estimation windows and high-frequency energy price data are key to improving nowcasting accuracy. Rolling estimation can attenuate the effects of unaccounted parameter time variation. However, machine learning methods offer more flexibility in this regard, can capture nonlinearities, and allow the use of a large information set.

Medeiros et al. (2021) apply a variety of econometric and machine learning approaches to (US) inflation modeling and forecasting, using the large FRED-MD data set.[1] They find that random forest (RF) models outperform other approaches, particularly in periods of high uncertainty. They attribute this fact to the ability of RF models to incorporate non-linearity, and point to the importance of non-linear models in such an exercise. However, data is uniform frequency (monthly) and the authors eliminate series which are not available over the full sample period. This highlights a key problem with existing machine learning algorithms: They do not deal well with mixed frequency and missing observations. A first, and relevant, step in this direction is taken by Babii et al. (2021), who consider mixed frequency versions of the (sparse group) LASSO regression, both theoretically and in GDP nowcasting applications.

Building on the success of Medeiros et al. (2021) in forecasting inflation with machine learning approaches, we focus on random forest methods. An important rationale is that we want to be able to include a large number of predictors, motivated partly by continued

[1] Medeiros et al. (2022) find that RF models can be helpful for forecasting global inflation.

increases in the availability of data. We extend random forest methods in two dimensions. First, we develop an approach that allows for data with general patterns of missing observations, generated for example by the use of mixed frequency data or alternative indicators only available in the later part of the sample. Second, we generalize the expected value of the target variable at each node of the regression tree to allow for a linear relationship between right-hand side and target variables, which we refer to as a random forest regression node model. The standard model takes the mean of the target variable given the conditions to reach that node as the expected value; we simply generalize this to a slope and intercept term. Thus the standard model is a subset of this generalization when the slope is constrained to zero.

To establish possible benefits of our random forest extensions, we conduct simple Monte Carlo experiments comparing our approach with alternatives that do not include the missing observation treatment and that take the mean of the target variable given the conditions to reach that node as the expected value (i.e., restrict to 0 the slope coefficients of our general model). As a baseline, we first compare the performance of our missing observation random forest approach with existing RF alternatives using simulated data without missing observations, finding our missing observation approach to work as well as other RF approaches. Then, still using simulated data, we show that our missing observation random forest model works better than common RF methods for handling missing observations, such as replacing them with averages of the available observations, or dropping the series with missing observations. We then conduct experiments to assess the marginal benefit of adding the linear relationship piece to each node, and we find that, despite the cost of an additional parameter at each node of the model, we are able to reduce out-of-sample mean square error over the missing observation random forest model regardless of missing observations, as our final model combines these two innovations.

We then turn to applying our proposed random forest models to nowcast monthly US inflation using a large set of indicators which are either high-frequency or published before the target inflation measures, including about roughly 150 to 200 macro, financial, and (mostly commodity) price indicators. Our analysis looks at nowcasting consumer price inflation directly, as well as breaking it down into its major components of commodities (consumer goods), commodities less food and energy, food, services, and services less energy services. We focus on headline CPI and PCE inflation as well as their ex food and energy counterparts. In out-of-sample nowcasting, we compare results from random forest models to the nowcasts the Federal Reserve Bank of Cleveland has published since late 2013, based on the approach of Knotek and Zaman (2017). As detailed in Knotek and Zaman (2017), historically their nowcast accuracy often beats the accuracy of forecasts from surveys of pro-

fessional forecasters (which in turn beat the accuracy of many other model-based nowcasts) and is comparable to the accuracy of nowcasts from the staff of the Federal Reserve Board of Governors (published in Greenbook/Tealbook).

Our empirical work yields the following main findings. In fitting the data, for the overall CPI, gasoline prices dominate. For categories that do not directly measure energy costs, variables measuring real economic activity such as the weekly economic index of Lewis et al. (2021) or the purchasing managers index often account for the first split in the regression trees. In out-of-sample forecasting, our model performs well, with an approximately 50 percent improvement in root mean square error over the sample 2013:M10-2022:M2 relative to the corresponding nowcast benchmarks from a univariate model and an approximate 10 percent improvement with regards to the Cleveland Fed nowcasts.

The rest of the paper is structured as follows: Section 2 details our methodology and modifications to the standard random forest framework. We discuss the performance of the model using simulated data when one or several series may contain missing observations in Section 3. Section 4 then applies our models to the task of nowcasting US inflation. Section 5 concludes.

2 Econometric Methodology

In this section we develop our variant of the standard random forest model. We modify the original work by Breiman (2001) in two ways. First, we propose an approach to handling the structure of missing observations arising from the data processing methodology, i.e., ragged head data, whereby earlier nodes in a regression tree can use more data than later nodes (i.e., leaves). However, our proposed procedure is capable of handling an arbitrary pattern of missing observations. We refer to the approach as missing observation random forest (*MO-RF*). Second, instead of simply calculating the mean of our target variable at a given node, we allow for a simple univariate linear regression in the splitting variable, which we will call random forest regression node (*MO-RFRN*). One can implement these adaptations of the random forest individually or together; each is developed in turn below.

2.1 Modeling Mixed Frequency

In order to work with mixed frequency data sets which allow the use of all data in real time, an aggregation rule such as the one proposed by Mariano and Murasawa (2003) is not effective in our context. The model would become overly burdensome with large differences in frequencies. For example, with daily and monthly observations, one would need to include about 60 lags to accommodate low frequency data in differences. Hence, we propose

to simply aggregate observations we have for the current period to date, and replicate these aggregations in the historical data.

We can classify right hand side (RHS) variables in the model as either lagged or contemporaneous. For example, we may include a daily variable such as the Baltic Dry index for the previous month; this would constitute a lagged variable. If the current date is the 15^{th} of the month, then the Baltic Dry index through the 15^{th} constitutes a contemporaneous variable, including the first 14 observations assuming the index has not yet been published for the current day. In order to incorporate this variable in our model, we take the first 14 observations of the Baltic Dry index for each month in the historical data as a RHS variable against the contemporaneous target variable, for instance the consumer price index, for that month. Thus while lagged variables are typically complete, that is, contain all high frequency observations within a month, contemporaneous variables are not. Our treatment of predictors on the right-hand side is analogous to the blocking-based approaches of Carriero et al. (2015) and McCracken et al. (2021): At each forecast origin, the variables included as predictors are specified to reflect the actual data availability; accordingly, the predictor set varies by forecast horizon.

It should be noted that our approach to mixed frequencies, which uses a simple average of high frequency (daily) observations over the low frequency (monthly) period, is potentially more restrictive than MIDAS (Ghysels et al. (2004)), which allows for different weights on high frequency observations within the low frequency period. However, introducing this flexibility into our random forest model would greatly increase the number of parameters to estimate and the nonlinearity of the specification, making estimation overly complex.

2.2 Data Processing

Though our approach to mixed frequency data is conceptually simple, aggregating non-stationary data introduces an additional step. The issue is that high frequency data are typically aggregated before log differencing. Because contemporaneous variables will change every day as more observations are realized, log differencing must be repeated at each new point in time. Additionally, because non-stationary data must be log-differenced after aggregation, all series initially enter the model in unprocessed, level form. Generically, our data processing proceeds as follows:

1. For lagged variables, aggregate to low frequency (monthly) if no observations are missing; if observations are still missing, identify this pattern and replicate it in previous low frequency periods, then aggregate.

2. For contemporaneous variables, identify which observations within the low frequency

period (month) are missing, and replicate this pattern of missing data in previous low frequency periods, then aggregate to low frequency (monthly). When going from daily to monthly data, we might, for example, observe days 1 through 15 of the month. Thus we would also take days 1 through 15 in previous months before aggregating to monthly frequency.

3. Data from steps 1 (including lags of our LHS variable) and 2 constitute our RHS variables; we take logs of variables where appropriate.

4. Take differences where needed to insure stationarity. Stationarity is necessary for this model as for other time series models to insure that relationships between variables identified in-sample remain valid for out-of-sample estimations; this rules out variables that grow over time. However, the model is capable of estimating regime changes so long as we have historical observations of regimes, and are not constantly transitioning into new, not yet observed regimes. [2]

2.3 Regression Trees

Because we are dealing only with continuous data, our random forest will consist of a sample of regression trees. This means that splits will divide data into observations greater or smaller than the cut point, as opposed to distinct categories. Our regression tree will continue to split the data until either the maximum number of nodes is reached, or nodes reach a minimum size. When we include both slope and intercept terms at each node, we set this minimum size to ten observations. Each split is on a single variable, where the variable and cut point are selected according to the greatest improvement in in-sample fit, as measured by mean square error. Our basic regression tree algorithm is simply:

1. Search over all RHS variables for the split that gives the maximum improvement (reduction) in mean square error. For a split in variable x_1 at cut point c_1 the minimum mean square error is equivalent to minimizing

$$\sum_{x_1<=c_1} \big(y_i - E(y|x_1 <= c_1)\big)^2 + \sum_{x_1>c_1} \big(y_i - E(y|x_1 > c_1)\big)^2$$

   where $E(y|A)$ is simply the mean of observations of y when condition A is satisfied.

2. Identify the terminal node or "leaf" with the greatest variance, and split this node following the procedure in step 1. After the first split, there will be two terminal nodes,

[2] Instructions to the code on where to take logs and differences are contained in a library file, borrowing from the "transform" row containing a code for the appropriate transformation in the FRED-MD database; see McCracken and Ng (2016). Where transformations from FRED-MD do not exist we use both visual analysis and a unit root test to asses which series need difference.

one corresponding to $x_1 <= c_1$ and the other to $x_1 > c_1$. After the second split, there will be three terminal nodes, and so on.

3. Continue splitting nodes until the stopping condition (the maximum number of nodes or minimum node size) is met. Fitted values are then the conditional means of each leaf.

Our random forest comes from sampling many (1000 in the Monte Carlo simulations and 2000 in the application, which features a much larger number of predictors) regression trees while randomizing in two ways. First, for each tree, we randomly select a subset of the rows of the data, or data points over time ("bagging"). Following the existing literature and code libraries, we use $ceiling(0.632T)$ rows, where $ceiling$ denotes the R function that rounds up to the nearest integer and T is the total number of observations.[3] Second, at each node, we randomly select $ceiling(k/3)$ candidate series for splitting ("feature bagging"), where k denotes the number of predictors (columns of the data matrix). To motivate this approach see Breiman (2001) and Ho (1995).

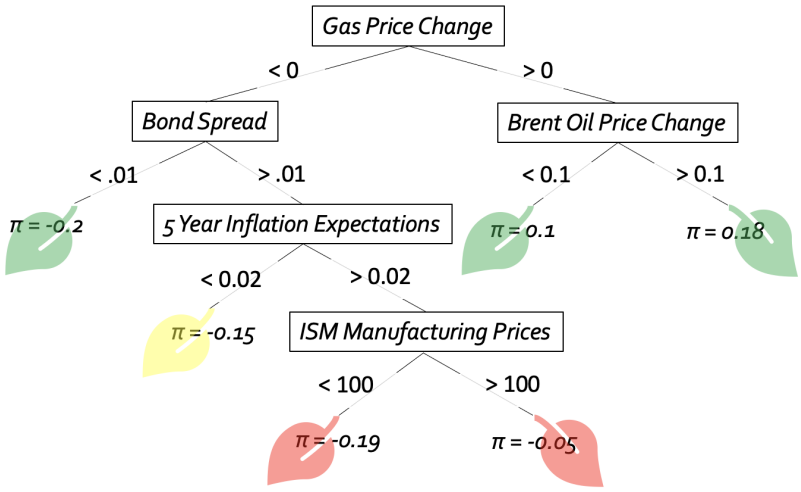
Advantages of random forest models include the fact that they allow a high degree of non-linearity, select relevant series from the data so inclusion of uninformative series tends not cause poor out-of-sample performance, are resistant to overfitting, are computationally lightweight and quick to estimate, and are simple to interpret. These last three points set random forests apart from neural networks. In particular, computational intensity matters for this exercise as the data changes from one day to the next as more observations within the month arrive. This means we cannot train a model once and then run it on subsequent data sets, but must re-train the model as each new data point is observed, since, for example, the Baltic Dry index for the first 5 days of the month is not the same variable as the Baltic Dry index for the first 25 days.

2.4 Ragged Head Data

The previous subsection describes a standard approach to random forest models with continuous variables. Our approach to mixed frequency data in Section 2.1 ensures that our data will have a square tail. However, series begin at different times, creating what we will call a "ragged head" structure to the data. One option is of course to discard observations prior to the start of the youngest series. However, this would discard a large amount of information that we wish to use in constructing our model. Our solution is to write a random forest algorithm in which the data used by each node is conditional on the date of the first observation of the current splitting variable. The approach is illustrated in Figure 1.

[3] We use the statistical software *R*, in which the available RF packages are **ranger** and **randomForest**.

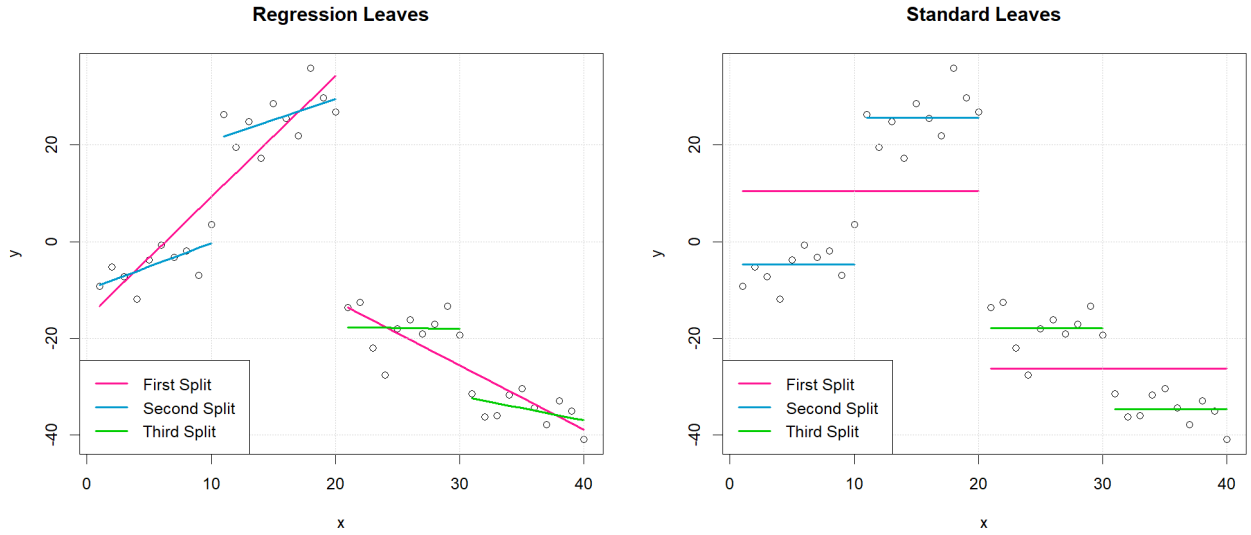
Figure 1: Ragged Head Regression Tree



Gas Price Change	Brent Oil Price Change	Bond Spread	5 Year Inflation Expectations	ISM Manufacturing Prices
-0.1	-0.1	0.01		
-0.2	-0.2	0.02		
0	0.1	0.02		
0.1	0.2	0.03		
-0.1	-0.1	0.04		
0.1	0.1	0.02		
0.1	0.1	0.03		
0.3	0.2	0.05		
0.4	0.3	0.02		
0.1	0.1	0.03	0	
0.2	0.1	0.01	0.1	
-0.3	-0.3	0.01	0.2	
0.2	0.2	0	0.1	
-0.1	-0.1	0	0.3	
-0.1	-0.1	0	0.2	110
0.1	0.2	0.02	0	90
0.1	0.1	0.03	-0.1	80
0.2	0.1	0.04	-0.2	70
0.2	0.2	0.04	0.1	105
0.1	0.1	0.05	0.2	106
0.1	0.2	0.05	0.2	107
0.2	0.2	0.03	0.1	94
-0.1	-0.1	0.02	-0.1	101

The figure depicts a regression tree based on data for the US described in our application in Section 4. The first split is on weekly gasoline prices, which will always be the case if gasoline prices are included in candidate splitting variables due to their dominant impact on the CPI, in line with results by Knotek and Zaman (2017). Because gasoline prices go back to the beginning of the CPI data, the model uses all (bootstrapped) observations. If the second split is on a variable that also goes back to the beginning of the data, such as bond spreads, then again all (bootstrapped) observations are used. Suppose then the model splits on a variable for which we have a shorter history, such as the 5 year breakeven inflation rate (which we have from 2003). Because some of the observations of this series are missing, we have three possible outcomes: $x_a <= c_a$; $x_a > c_a$; or x_a is NA. This sounds similar to the approach when using categorical data of assigning missing as a new category. However, x_a is NA is not a candidate for future splits; this restriction makes sense in our framework as the tail of our data is always square, so that x_a is NA will never be used to make predictions. Additionally, when calculating the resulting variance for potential splitting variables we do not use the mean of y (our target, the CPI) conditional on the previous splits (gas price change < 0 and bond spreads > 0.01) and x_a is NA. Rather, if x_a is NA then the expected value of y is simply its expected value at the previous node (gas price change < 0 and bond spreads > 0.01). Thus later nodes are restricted to the periods for which splitting variables at previous nodes were observed. This is illustrated in Figure 1 with green leaves (terminal nodes) corresponding to the (hypothetical) data set outlined in green, the yellow leaf corresponding to the data outlined in yellow, and the red leaves corresponding to the data outlined in red.

Figure 2: Regression nodes vs. standard nodes



2.5 Regression Nodes

Our second innovation is to allow for both a slope and an intercept term at each node. Conceptually, this is a simple modification illustrated by Figure 2; we estimated these results using our regression tree code on a single variable without bagging. In the standard model, the expected value of our target variable, y, is simply the mean of y when the conditions needed to reach the node are met ($x_1 > c_1$, $x_2 < c_2$, and so on). The new case similarly limits observations to those which meet the conditions needed to reach the current node, but allow for both an intercept and slope term using the cut variable as the (single) RHS variable.

In this example the first split occurs where $x <= 20$ or $x > 20$. In the standard case, the expected value of y when $x <= 20$ is just the mean of y when $x <= 20$, or the mean of y when $x > 20$ for the alternative case. In the regression case, we estimate the model

$$y_i = a_i + b_i x_i + \varepsilon_i, \tag{2.1}$$

where i indexes a given node. For the case of the first split at $c_1 = 20$ of the single RHS variable above, $\{x_{i=1} \in x | x <= c_1\}$ and $\{x_{i=2} \in x | x > c_1\}$. Supposing the next split is on node 1, $\{x_{i=3} \in x_{i=1} | x_{i=1} <= c_2\}$ and $\{x_{i=4} \in x_{i=1} | x_{i=1} > c_2\}$, and so on.

Thus, the standard model is simply a special case of the regression model where $b_i = 0$. At subsequent split points, we calculate mean square error in terms of the residuals of equation (2.1). While this would not be necessary for a single RHS variable, it is necessary with many potential RHS variables. We will now have two parameters to estimate at each node instead of one. This raises the question of over-fitting the model in-sample but, on the positive side, we will be able to more efficiently describe data generating processes in which

there are linear relationships between variables. If the underlying process is closer to the step function described by the original random forest algorithm, our regression nodes offer little advantage. Bagging and feature bagging will still help to address over-fitting, but will be less effective if RHS variables are highly correlated.

The approach is similar to the macro random forest of Coulombe (2020). However, regressions in the macro random forest approach are on a small subset of pre-selected variables. In contrast, we leave the selection of variables entirely up to the model. This is particularly important for large data sets where we may not know a priori which variables are the best (linear) predictors of our target variable.

3 Benchmarking the Algorithm

As with any new algorithm, our first task is to benchmark performance against existing libraries. Because we are working primarily in *R*, the obvious benchmarks are the "randomForest" and "ranger" packages, both of which estimate random forest models for continuous or categorical variables. In the following tables, MO-RF refers to our own coding of the random forest algorithm allowing for ragged head data, and MO-RFRN is the model allowing both for ragged head data and linear relationships in the cut variable. The motivation for including our code which uses the standard random forest model for continuous variables is simply to verify that there are not large differences between our work and existing random forest routines. When no variables are missing, MO-RF and existing algorithms should be very much the same.

Our simulated data comes from a simple linear model

$$y_t = \beta x_t + \varepsilon_t, \tag{3.1}$$

where y_t is a scalar, x_t is a vector of standard normal random variables, ε_t is also standard normal, and values of β are fixed integers such that

$$\beta = vec\left\{\begin{bmatrix} 1 & 5 & 3 & 0 & 2 & -1 & 2 & -3 & 0 & 1 \\ 1 & 0 & 2 & 0 & 3 & -1 & -2 & -1 & 2 & 4 \end{bmatrix}'\right\}.$$

Simulations proceed as follows:

1. Draw 200 training observations from the model in equation (3.1).
2. Train random forest models using "ranger," "randomForest," and the algorithms described in Section 2.
3. Draw another 200 validation observations from the model in equation (3.1).

4. Predict y_t using each of the four candidate algorithms, using the trees estimated in the training step (2).

5. For each simulated data set, calculate the mean square difference between estimates and the true values of y_t from step 3. Average the mean square differences across data sets, and then take the square root to obtain a root mean square error

We begin with the case in which no data are missing for estimation and forecasting. As the results in Table 1 illustrate, our MO-RF algorithm yields results closer to the "randomForest" library, with a nearly identical root mean square error. The regression node approach outperforms all other models, but perhaps this is not surprising as the data generating process is a linear model.

Now consider the case in which the head of the data available for model estimation contains missing values. In this setting, results for the standard model begin to diverge from existing libraries. With a ragged head, with existing libraries, there are, following Breiman (2001), four ways to proceed: one can drop columns with missing observations, drop rows with missing observations, fill missing observations with the mean for that series, or impute missing values based on their relationship with other observed series (imputing missing values via splines or similar time-series approaches is not possible as missing values do not fall between observations). The latter two approaches in this case are equivalent because the simulated series in x_t are iid.

Table 1: RMSE when the first 100 observations are missing for either a single series or ten of the twenty simulated series vs no missing observations, linear DGP

	MO-RFRN	MO-RF	randomForest	ranger
No MO	**5.80**	6.36	6.39	6.60
One series with MO	6.39	6.67	6.76	6.91
Ten series with MO	6.66	6.85	6.94	7.07

Table 1 presents results in which missing observations are replaced by their mean value for the existing randomForest and ranger libraries, which in every case we look at outperforms the alternative of dropping rows or columns with missing data (in results not reported). We consider two possibilities. First, we simulate the case when the first 100 observations for a single series, in this case series 2 (which has the largest coefficient), are missing. We then simulate the data and drop the same periods for the first ten series in the training

set. In both cases, our standard algorithm described in Section 2.4 performs slightly better than existing libraries. As in the case of no missing observations, the regression node algorithm maintains its substantial lead. This difference in performance is similar in both the case of missing observations in one series or many series.

We repeat these simulations for a simple non-linear data generating process. For this exercise, we use a regime switching model in which there is a $1/3$ probability of being in each of the three regimes. Regimes themselves are linear processes. The regime at period t depends on a variable z_t drawn from a uniform distribution. $z_t <= 1/3$ indicates the model is in regime 1; $1/3 < z_t <= 2/3$ indicates the model is in regime 2, and $z_t > 2/3$ indicates the model is in regime 3. The variable z_t is observed with error $\epsilon_t \sim \mathcal{N}\big(0, 1/16\big)$.

Table 2: RMSE when the first 100 observations are missing for either a single series or ten of the twenty simulated series vs no missing observations, regime switching DGP

	MO-RFRN	MO-RF	randomForest	ranger
No MO	7.95	7.95	7.85	7.95
One series with MO	7.99	7.99	7.95	8.02
Ten series with MO	8.65	8.56	8.60	8.56

In this case, in simulations with missing data, our MO-RF approach performs at least as well as the standard RF algorithms adapted in crude ways to accommodate missing observations. But its advantage is reduced in data from a switching DGP relative to data from a linear DGP. The MO-RFRN model performs less well. The reasons are two-fold. First, the DGP is not linear in this case, reducing the advantage of using a linear model at each node. Second, because the MO-RFRN model estimates more parameters than the standard model, it is more prone to over-fitting. Because we have, on average, only $n/3$ observations of each state (as opposed to n observations of the only state in the previous exercise), estimating additional parameters has a higher cost.

We have established that our standard algorithm for continuous variables is similar in performance to the randomForest and ranger libraries in R when no data are missing, and can slightly outperform these libraries for our data generating process when there are missing observations. Moreover, we have shown that allowing for two parameters — a slope and intercept term — at each node, as opposed to the standard one parameter giving only the conditional mean of the target variable, may improve out-of-sample performance. In our simulations, the performance gains are clear in root mean square errors for a linear process,

but our missing observation approach underperforms when the DGP is non-linear and we use only 200 observations to train the model. Because this improvement will depend on the data generating process, our forecasts when using actual data to predict inflation will use both the MO-RF and the MO-RFRN models.

4 Nowcasting US inflation with random forest models

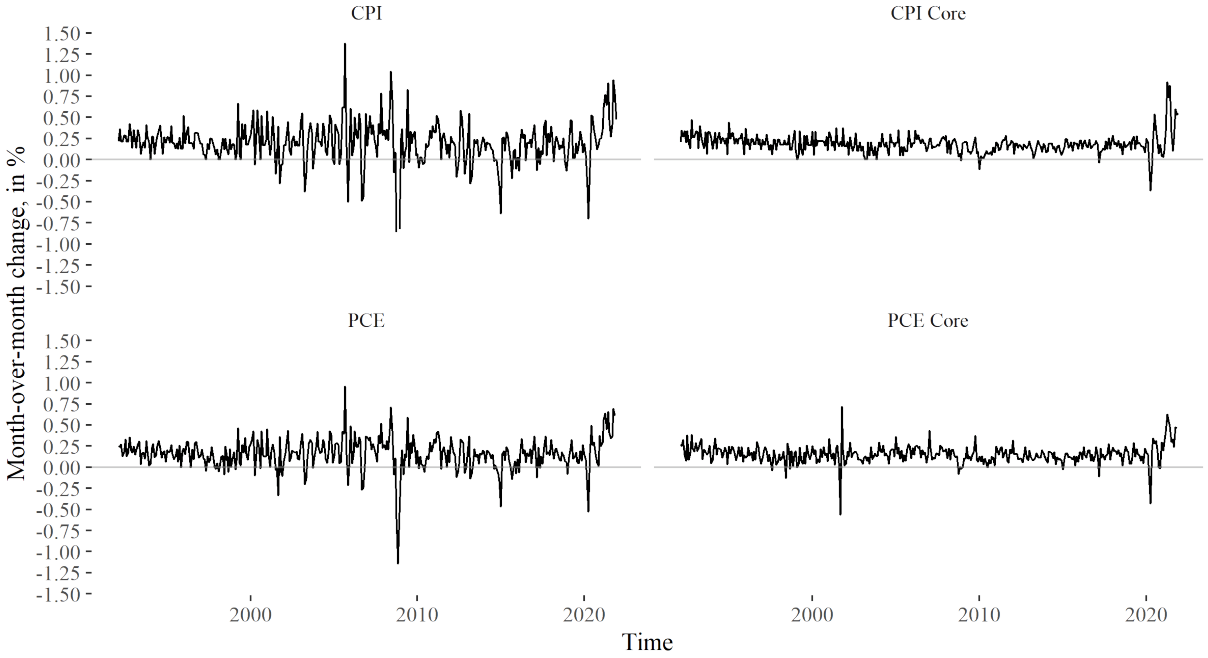
In this section we return to the problem that motivated the development of the random forest methods in the previous section and investigate whether US consumer price inflation forecasts may be improved by exploiting daily and weekly mixed-frequency data with respect to models involving only lagged inflation rates. We first describe the data and then present in-sample and out-of-sample results. The section concludes with some robustness checks.

4.1 Data

We obtain macroeconomic series on monthly, weekly, and daily frequency, including gasoline and commodity prices, from Trading Economics and FRED. The measures of inflation are the monthly seasonally adjusted urban consumer price index (CPI) and the personal consumption expenditures index (PCE) together with their core (ex food and energy) measures. For the inflation measures, we rely on real-time vintages drawn from ALFRED (see Anderson (2006)). We define inflation as the month-on-month change in the respective price index, $\pi_t = 100 * (P_t / P_{t-1} - 1)$, in which P_t denotes the price index (see Figure 3). We use data for the target series spanning from 1980:M1 to 2022:M2. The starting point is defined by data constraints.

Though we only have true vintage data for inflation, each observation of the RHS variables is tagged with a (daily) publication date; any predictor series published after a forecasting date is dropped when recursively forecasting and evaluating the model performance. Data will be tagged with the date of the first publication, but actual observations may be revised. Commodity prices and financial indexes are available in real time; for each series we use end-of-day closing prices. The CPI is published with an average lag of 12 days, whereas the PCE price index has an average publication lag of 25 days. This leaves us with a total of 169 contemporaneous macroeconomic series published at either daily, weekly, or monthly frequency for nowcasting the CPI and with 211 contemporaneous series for nowcasting PCE. Because PCE is published with a greater lag, we have more variables which reference the current month, including CPI measures. These numbers exclude data at lags; as we include three lags of our target variable and one lag of all RHS variables, we have

Figure 3: Target inflation rates, first release



over 300 features (potential predictors) entering the random forest models. Table 3 provides a summary of the data used in the application. The majority of predictors are prices on a daily frequency, starting as early as 1960 (nonetheless, for other data availability reasons our estimation sample starts in 1980), while some series start in 2016 only. Several indicators also cover the labor market and business activity, among which some surveys are included. Table A.1 in the appendix details the list of series used and their categorization.

4.2 In-sample results

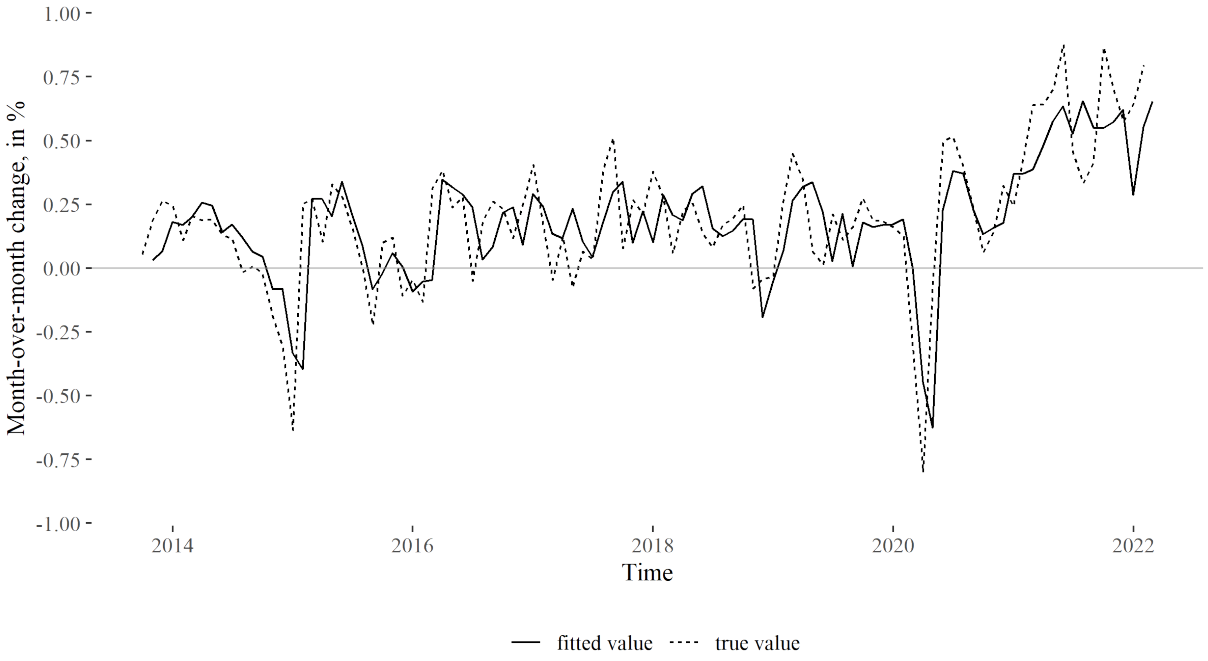
In order to shed some light on the workings of the random forest models, in this section we study its in-sample properties. Note that what we call in-sample are in fact *out-of-bag* results. Recall that we fit our random forest model by bagging (i.e. bootstrapping). For each tree we estimate, we (randomly select and) drop 1/3 of the rows of data. Using this tree we then create predictions using the 1/3 of observations that were dropped when fitting the tree. Our final estimates are then the average of these out of bag predictions over 2000 trees. Confidence intervals also use this process, as for each data point we will have a distribution of out-of-bag estimates. We nevertheless refer to these results as in-sample as the model is fitted on observations both before and after the out-of-bag data points.

Figure 4 displays the realized and fitted values of CPI inflation starting from 2013:M10 to 2022:M2 (in line with the out-of-sample evaluation period of Section 4.3). We construct these estimates using all the data. Before the Covid-19 pandemic of 2020, the fitted values of the

Table 3: Set of contemporaneous predictors used

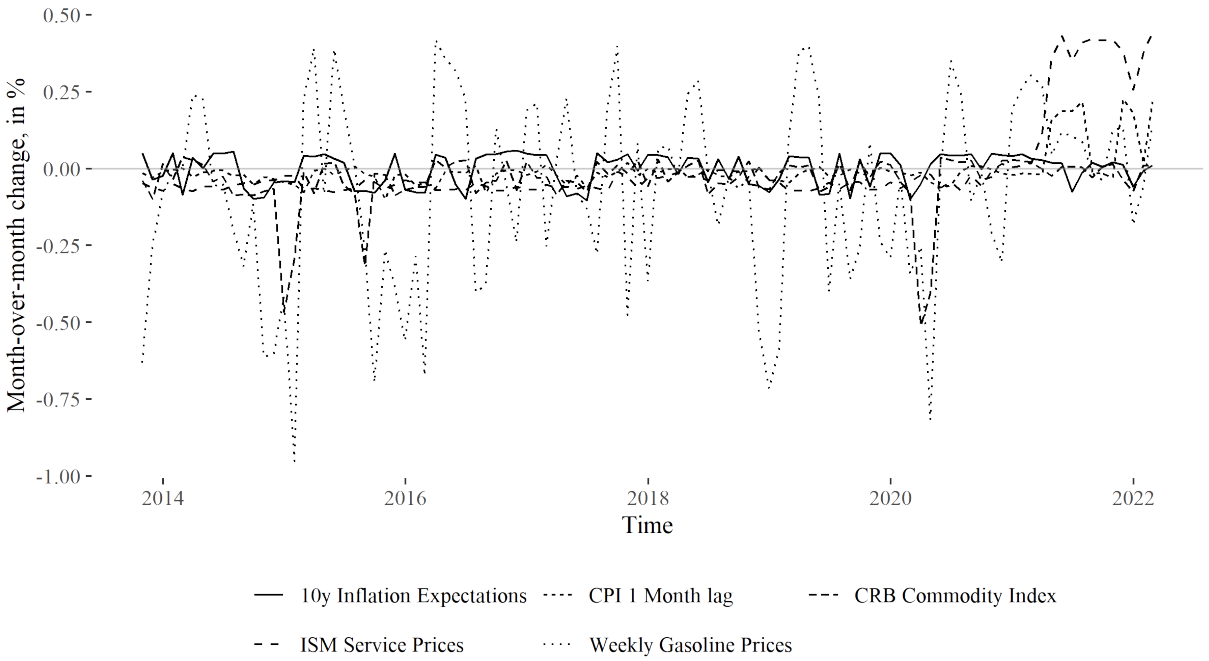
				Amount of predictors	
Type	Frequency	Earliest start	Latest start	CPI	PCE
Business	Daily	01.03.1960	01.03.1960	5	5
	Weekly	02.01.1967	23.03.2012	14	14
	Monthly	31.01.1970	30.11.2013	13	29
Consumer	Weekly	05.02.2005	05.02.2005	1	1
	Monthly	31.01.1970	28.02.2001	8	12
Government	Monthly	31.01.1954	31.01.1980	5	5
Housing	Weekly	05.01.1990	28.12.2007	2	2
	Monthly	31.01.1970	31.01.1985	1	6
Labour	Weekly	05.01.1980	05.01.1980	2	2
	Monthly	31.01.1950	30.04.2006	19	20
Markets	Daily	01.03.1960	02.01.2003	14	14
	Weekly	02.01.1967	30.08.1991	6	6
Money	Daily	06.09.1994	03.10.1994	2	2
	Weekly	02.01.1980	18.12.2002	2	2
	Monthly	31.01.1970	31.01.1970	2	4
Prices	Daily	01.03.1960	22.01.2016	59	59
	Weekly	20.08.1990	20.08.1990	1	1
	Monthly	31.01.1956	30.04.2010	13	22
Trade	Monthly	31.01.1970	31.03.2013	0	5

Figure 4: CPI inflation, fitted versus true values, 2013:M10 to 2022:M2
Fitted values from random forest random node model



regression node model are very close to the realized CPI inflation rates. Since the beginning of 2020, US CPI inflation experienced stronger dynamics, some of which the model was not able to capture. Table 4 reports the resulting root mean square errors (RMSEs) of the MO-RF and MO-RFRN models against a simple univariate specification in which inflation forecasts π_{t+h} are produced by an AR(1) model, $\pi_t = \rho_0 + \rho_1 \pi_{t-1} + \epsilon_t$, where ϵ_t is serially uncorrelated, with mean zero and variance σ^2_ϵ.[4] A first observation, in line with existing estimates, is that the random forest models perform better for the CPI and PCE than for the respective core measures. Second, relative to the AR(1) benchmarks, gains are larger for PCE measures than CPI measures. This is due to the fact that PCE is published with a greater lag; our PCE estimates include the CPI as a RHS variable, which is of course highly informative.

Figure 5: Feature contributions to CPI estimate, 2013:M10 to 2022:M2



One explanation for the discrepancy in performance between overall and core inflation rates is the dominant role of energy prices in the overall indices. In a linear model, parameter estimates describe the marginal impact of each RHS variable. In our case, the impact of each variable is potentially conditional on all the other variables in the model. Following Palczewska et al. (2014), we therefore use feature contributions to measure variable impact on estimated results at a specific point in time. The feature contribution of a variable $x_{k,t}$ at point t is simply the total change in the target variable y_t each time the model splits on $x_{k,t}$ to construct a prediction at period t. For this reason, for a given tree in our model, feature

[4] We also examined other common parsimonious univariate models, including the random walk forecast or an integrated moving average as in Modugno (2013). We report results for the AR(1) model since it exhibited better forecasting performance in our sample.

contributions will sum to our estimate of y_t from that tree, less the mean of y (which is not attributed to any RHS variable).

Figure 5 illustrates this metric for the five most influential variables in our model. As is evident in the figure, weekly gasoline prices contribute most to our estimates. Knotek and Zaman (2017) similarly find high-frequency energy prices to be important in their nowcasting models of headline inflation. The service prices index from the ISM survey, the CPI lagged by one month, the CRB commodity price index, and 10y inflation expectations are the other variables which make the largest contribution to estimated results. Notably, since mid-2021, service prices were able to capture a lot of the movement in the surge of CPI inflation.

Regarding the different sub-components of the CPI index, the importance of gasoline prices in accounting for price changes is further accentuated. As reported in Table 5, while our models perform between 27 and 37 percent better than an AR(1) benchmark for the commodities (goods) component of CPI inflation, stripping out food and energy increases the ratio of RMSEs to around 0.96, constituting about a 4 percent improvement on the AR(1) model. In the case of services, stripping out the energy component also reduces the gains in RMSE accuracy of nowcasts from random forest models: For example, in the full sample, the RMSE ratios of our MO-RF nowcasts are 0.958 for the services component of CPI inflation and 0.998 for the services ex energy component.

Table 4: Relative RMSE, Inflation Measures

	Full sample		2013:M10 to 2019:M12	
	MO-RF	MO-RFRN	MO-RF	MO-RFRN
CPI	0.733	0.723	0.701	0.658
CPI core	0.971	1.032	0.979	0.981
PCE	0.619	0.545	0.629	0.544
PCE core	0.848	0.824	0.938	0.884

Out of bag RMSE relative to the (in-sample) RMSE of estimates from an AR(1) model

Table 5: Relative RMSE, CPI Components

	Full sample		2013:M10 to 2019:M12	
	MO-RF	MO-RFRN	MO-RF	MO-RFRN
CPI commodities	0.729	0.733	0.675	0.632
CPI commodities less food energy	0.966	0.989	0.992	0.960
CPI nondurable goods	0.686	0.686	0.686	0.659
CPI durable goods	0.976	0.994	0.997	0.987
CPI food	0.978	0.971	0.974	0.952
CPI services	0.958	0.977	0.954	0.952
CPI services less food energy	0.998	1.082	0.991	1.010

Relative RMSE comparing random forest models to an AR(1) fit.

4.3 Out-of-Sample performance

In the following, we determine whether our random forest models can improve inflation forecast accuracy. We compare the resulting predictions against the nowcasts from the Cleveland Fed, which as noted above are as good as or better than many alternative forecast sources.[5] Those nowcasts are available for both CPI and PCE rates (including core inflation) and start in 2013:M10. As in the previous section, we compare our predictions with those of an AR(1) model. We implement a recursive estimation scheme, which for the first forecast in the evaluation sample covers the period 1980:M1 to 2013:M10. According to the availability of nowcasts from the Cleveland Fed, our evaluation sample spans from 2013:M10 to 2022:M2. We perform one nowcast per week, giving a total of 102 months to forecast and a total of 408 weekly forecasts made. The nowcasts are evaluated at horizons of 1, 2, 3, and 4 weeks ahead of the series publication date.

Estimating new parameters at each forecasting date allows us to appropriately account for the data observed for the current month up to the estimation date. The number of RHS variables will depend on the horizon. For estimation dates closer to the publication date, we will have more contemporaneous variables; for earlier estimation dates we will have fewer. At most, we will have a total of about 300 RHS variables (contemporaneous and lagged) entering the model.

In the case of US data, series are either already seasonally adjusted (e.g., initial jobless claims) or do not need seasonal adjustment (e.g., financial data).

Forecasting with the model involves the additional step of identifying which series were realized at a given forecast origin date. We drop observations published after the current forecast date, and then aggregate and process data following steps 1 through 5 outlined in Section 2.2. Because the resulting processed data is uniform frequency, one can use it with most standard regression tools. Moreover, the tail of the data is square. If we have a contemporaneous series with no observations at the current forecast origin, it is dropped from the analysis, so that all series entering the model are observed in the last period. However, the head of the data may contain missing observations.

We evaluate the accuracy of our nowcasts by means of root mean square forecast errors (RMSEs) relative to RMSEs from the AR(1) benchmark. Forecast errors are calculated using the final inflation vintage as the measure of actual inflation.[6] We apply the modified version of the Diebold-Mariano test (Diebold and Mariano, 1995) developed in Harvey et al. (1997). The results are reported in Table 6. The relative RMSEs are shown together with significance

[5]The latest data and vintages are available at https://www.clevelandfed.org/en/our-research/indicators-and-data/inflation-nowcasting.aspx. The nowcasts are based on the work of Knotek and Zaman (2017).

[6]Our results are robust when using the first release instead for calculating forecast errors, see Table A.3.

Table 6: Forecasting performance, relative RMSE

	Full sample				2013:M10 to 2019:M12			
Forecasting horizon in weeks	1	2	3	4	1	2	3	4
CPI								
AR(1) Benchmark	2.101	2.101	2.101	2.101	1.520	1.520	1.520	1.520
Cleveland Fed Nowcast	0.662***	0.678**	0.742**	0.799**	0.523***	0.532***	0.641***	0.687**
MO-RF	**0.455****	**0.487****	0.544***	0.564***	**0.357****	**0.418****	**0.506****	**0.499****
MO-RFRN	**0.434****	**0.454****	0.458***	0.492***	**0.335****	**0.384****	**0.421****	**0.436****
CPI Core								
AR(1) Benchmark	1.326	1.326	1.326	1.326	0.482	0.482	0.482	0.482
Cleveland Fed Nowcast	1.135	1.135	1.148	1.179	1.046	1.051	1.155	1.163
MO-RF	**0.972**	0.996	0.998	1.008	**0.812****	**0.995**	1.007	0.986
MO-RFRN	**0.961**	0.992	1.044	1.051	**0.797****	1.118	1.091	1.088
PCE								
AR(1) Benchmark	1.857	1.857	1.857	1.857	1.191	1.191	1.191	1.191
Cleveland Fed Nowcast	0.438***	0.462***	0.596***	0.626***	0.567***	0.571**	0.612**	0.630**
MO-RF	0.349***	**0.355****	0.485***	0.482***	**0.423****	**0.439****	0.527**	0.522**
MO-RFRN	0.352***	**0.355****	0.420***	0.415***	**0.377****	**0.375****	**0.455****	**0.447****
PCE Core								
AR(1) Benchmark	1.156	1.156	1.156	1.156	0.687	0.687	0.687	0.687
Cleveland Fed Nowcast	0.797***	0.905	1.091	1.129	0.692	0.889	0.975	0.988
MO-RF	0.708***	**0.740****	0.938	0.957	**0.682****	**0.739****	0.793**	0.840
MO-RFRN	0.839	0.882	0.874**	0.881	**0.641****	**0.702****	0.770**	0.807

RMSE of nowcasting models relative to the benchmark AR(1) model.
Significance levels: *p-value*: *** < 0.01, ** < 0.05, * < 0.1 of the modified Diebold-Mariano test (Harvey et al., 1997).
Modified Diebold-Mariano tests: the alternative hypothesis states that (i) the nowcasts are significantly more accurate than the benchmark and (ii) the random forest models are significantly better than the Cleveland Fed nowcasts (in bold).

levels from the modified Diebold-Mariano tests, in asterisks for the hypothesis that the random forest forecasts significantly improve upon the univariate AR(1) benchmark model and in bold if they significantly outperform the Cleveland Fed nowcasts.

For the full evaluation sample, several results emerge: (1) Forecast accuracy improves in most cases as our forecast horizon shrinks; (2) except for core CPI inflation, predictions from the random forest models exhibit lower root mean square errors than the benchmark at all horizons; (3) for CPI and PCE inflation, the nowcasts are significantly better than the univariate benchmark at all horizons, whereas for the core inflation measures they are significantly better between one and three weeks ahead; (4) the random forest models provide significantly better CPI nowcasts than the Cleveland Fed at the one and two week horizons for the full sample; for the limited sample at all horizons; (5) the random forest, random node model exhibits somewhat lower forecast errors in most cases than the simple random forest model.[7]

Interestingly, when excluding the period of the Covid-19 pandemic and the economic

[7]Table A.1 reports results for using the latest vintage of inflation to estimate models and form nowcasts. Our findings are broadly qualitatively robust.

recovery in its aftermath with surging inflation rates, the random forest models also outperform the benchmark and the Cleveland Fed nowcasts for the core inflation rates. Moreover, as reported in the right half of Table 6, for the subsample 2013:M10 to 2019:M12, the random forest models provide significantly more accurate inflation nowcasts at least up to four weeks ahead for any target. As our data set captures at least partially price dynamics in the service sector or pressures in the global value chains, this explains why our models were able to capture slightly more accurately the increase in core inflation rates in 2021 than the Cleveland Fed.

In summary, nowcasts from our proposed random forest methodology provide a valuable early signal on inflation developments in the US. Overall, our random forest-based forecasts using a large set of indicators can be seen as at least matching the accuracy of the Cleveland Fed nowcasts based on the methodology of Knotek and Zaman (2017) that focuses on a small set of indicators. In some cases, although not a majority, our nowcasts improve on the accuracy of the Cleveland Fed nowcasts. Accordingly, our forecasts may be seen as a useful complement to familiar nowcasts such as the Cleveland Fed's. That a simple model using only energy prices and historical inflation is competitive with a large data set is explained by the results from feature contributions in Section (4). Though we incorporate over 300 RHS variables in our model, gasoline prices still dominate the analysis of headline inflation measures.

4.4 Robustness

We now conduct a few additional exercises to assess the robustness of our results reported in Table 6.[8] First, we exclude gasoline prices from our data set; second, we estimate a MIDAS model featuring gasoline prices only; finally, we test our random forest models against a basic random forest specification in which no missing data is allowed.

Table 7 presents the results for the first part of the robustness exercise. In the first row we report the RMSE of the random forest random node model, which exhibited the lowest RMSE in Table 6. The remaining rows display the RMSEs resulting from alternative specifications together with significance levels from Diebold-Mariano tests (so, all entries in the table are levels of RMSEs and not ratios). Overall, the forecasts from our proposed random forest model significantly outperforms the alternative specifications in most considered exercises, at all horizons.

[8]We also tested our results against the estimation with a fixed, rolling window length of 20 years; and against a bottom-up CPI prediction using the nowcasts for commodities and services inflation. Further, the results reported here for headline CPI inflation extend to PCE and core inflation rates. For brevity, the results are not reported here but available from the authors upon request. Our results are qualitatively robust against these alternative approaches and specifications.

Table 7: Robustness of CPI forecasting performance, RMSE

	Full sample				2013:M10 to 2019:M12			
Forecasting horizon in weeks	1	2	3	4	1	2	3	4
CPI real-time								
MO-RFRN	0.910	0.949	0.957	1.034	0.506	0.577	0.633	0.662
Excluding gasoline prices								
MO-RF	1.232***	1.400***	1.458***	1.423***	0.776***	0.997***	1.086***	1.034***
MO-RFRN	1.286***	1.459***	1.419***	1.330***	0.757***	1.010***	1.086***	1.084***
MIDAS								
	1.653***	1.746***	1.770***	1.851***	1.063***	1.144***	1.156***	1.201***
randomForest								
	2.534***	2.485***	2.470***	2.521***	1.645***	1.665***	1.654***	1.624***

Significance levels: *p-value*: *** < 0.01, ** < 0.05, * < 0.1 of the modified Diebold-Mariano test (Harvey et al., 1997).
Modified Diebold-Mariano tests: the alternative hypothesis states that the alternative nowcasts are significantly *less* accurate than the CPI nowcasts from the random forest random node model.

Diving into the details of our robustness analysis, we first report results when gasoline prices are excluded from the data set. As shown in Table 4, gasoline prices are a dominant source of information for predicting overall CPI inflation. Hence, an exclusion of the most important indicator consequently leads to a substantial deterioration of nowcasting performance at both horizons, in both samples. The RMSEs of the nowcasts from the MO-RF and MO-RFRN specifications without gasoline prices (rows 2 and 3) are well above the RMSEs achieved by the MO-RFRN specification with gasoline prices (row 1).

Next, we challenge and assess the robustness of our results by specifying an additional model of the MIDAS form, along the lines of Ghysels et al. (2006), using weekly gasoline prices as explanatory variable, a key indicator based on our previous analysis. We include two lags for the dependent monthly variable, and we parameterize the MIDAS polynomial with an Exponential Almon lag. The model is estimated by nonlinear least squares.[9]

Our proposed random forest random node model exhibits much lower RMSE than the MIDAS model at all horizons considered, both for the full sample and for the restricted sample. The nowcasts are significantly more accurate at all horizons considered. The results suggest that a single regressor model with gasoline prices as explanatory variable is not sufficient to predict CPI inflation, not even when using a flexible MIDAS specification. The random forest models allow the inclusion of a set of other variables, capturing inflation better and providing better predictions.

Finally, we contrast our enhanced random forest model results against the traditional random forest model without missing data. We estimate the traditional random forest model using the "randomForest"-package in *R*. In such a model specification, the data has to be

[9]See for instance Breitung and Roling (2015), Knotek and Zaman (2017), and Monteforte and Moretti (2013) for MIDAS models to nowcast inflation.

square; i.e., it cannot have missing values either at the beginning of the series or at the end. For estimation, we assume that data series span at least ten years. Evidently, the nowcasts from our enhanced random forest models significantly outperform the predictions of the traditional RF model, in line with the results we obtained with simulated data and confirming the practical usefulness of our extensions.

5 Conclusions

In this article we present a novel approach to use mixed-frequency data with random forest models. Further, we generalize the models to allow for missing observations and for the expected value of the target variable at each node of the regression tree to be based on a linear relationship between target and predictors. We test these enhanced models both with simulated data and in an empirical application focused on predicting various measures of US inflation. We find that our new random forest models significantly outperform standard random forests when data are missing, and are competitive with if not significantly better than established benchmarks for inflation, such as those of the Cleveland Fed.

References

Anderson, Richard G (2006) 'Replicability, real-time data, and the science of economic research: FRED, ALFRED, and VDC.' *Federal Reserve Bank of St.Louis Economic Review* 88(1), 81–93

Babii, Andrii, Eric Ghysels, and Jonas Striaukas (2021) 'Machine learning time series regressions with an application to nowcasting.' *Journal of Business & Economic Statistics* 40(3), 1–23

Bańbura, Marta, Domenico Giannone, and Lucrezia Reichlin (2010) 'Large Bayesian vector auto regressions.' *Journal of Applied Econometrics* 25(1), 71–92

Bernanke, Ben S. (2007) 'Inflation expectations and inflation forecasting.' Speech at the Monetary Economics Workshop of the National Bureau of Economic Research Summer Institute, Cambridge, Massachusetts, July 10.

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

Breitung, Jörg, and Christoph Roling (2015) 'Forecasting inflation rates using daily data: A nonparametric MIDAS approach.' *Journal of Forecasting* 34(7), 588–603

Carriero, Andrea, Todd E. Clark, and Massimiliano Marcellino (2015) 'Realtime nowcasting with a Bayesian mixed frequency model with stochastic volatility.' *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(4), 837–862

Cimadomo, Jacopo, Domenico Giannone, Michele Lenza, Francesca Monti, and Andrej Sokol (2021) 'Nowcasting with large Bayesian vector autoregressions.' *Journal of Econometrics*

Coulombe, Philippe Goulet (2020) 'The Macroeconomy as a Random Forest.' Papers 2006.12724, arXiv.org, June

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

Durbin, James, and Siem Jan Koopman (2012) *Time Series Analysis by State Space Methods* (Oxford University Press)

Faust, Jon, and Jonathan H. Wright (2009) 'Comparing Greenbook and reduced form forecasts using a large realtime dataset.' *Journal of Business & Economic Statistics* 27(4), 468–479

Faust, Jon, and Jonathan H Wright (2013) 'Forecasting inflation.' In 'Handbook of economic forecasting,' vol. 2 (Elsevier) pp. 2–56

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov (2004) 'The MIDAS touch: Mixed data sampling regression models.' CIRANO Working Papers, CIRANO

__ (2006) 'Predicting volatility: getting the most out of return data sampled at different frequencies.' *Journal of Econometrics* 131(1-2), 59–95

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

Harvey, David, Stephen Leybourne, and Paul Newbold (1997) 'Testing the equality of prediction mean squared errors.' *International Journal of Forecasting* 13(2), 281–291

Ho, Tin Kam (1995) 'Random decision forests.' In 'Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1' ICDAR '95 IEEE Computer Society M pp. 278–282

Knotek, Edward S, and Saeed Zaman (2017) 'Nowcasting US headline and core inflation.' *Journal of Money, Credit and Banking* 49(5), 931–968

— (2020) 'Real-time density nowcasts of US inflation: A model-combination approach.' Working Paper No. 20-31, Federal Reserve Bank of Cleveland

Krüger, Fabian, Todd E. Clark, and Francesco Ravazzolo (2017) 'Using entropic tilting to combine BVAR forecasts with external nowcasts.' *Journal of Business & Economic Statistics* 35(3), 470–485

Lewis, Daniel J., Karel Mertens, James H. Stock, and Mihir Trivedi (2021) 'Measuring real activity using a weekly economic index.' *Journal of Applied Econometrics.* forthcoming

Mariano, Roberto S., and Yasutomo Murasawa (2003) 'A new coincident index of business cycles based on monthly and quarterly series.' *Journal of Applied Econometrics* 18(4), 427–443

McCracken, Michael W., and Serena Ng (2016) 'FRED-MD: A monthly database for macroeconomic research.' *Journal of Business & Economic Statistics* 34(4), 574–589

McCracken, Michael W., Michael T. Owyang, and Tatevik Sekhposyan (2021) 'Real-time forecasting and scenario analysis using a large mixed frequency Bayesian VAR.' *International Journal of Central Banking* 18, 327–367

Medeiros, Marcelo C., Gabriel F. R. Vasconcelos, Álvaro Veiga, and Eduardo Zilberman (2021) 'Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods.' *Journal of Business & Economic Statistics* 39(1), 98–119

Medeiros, Marcelo Cunha, Erik Christian Montes Schütte, and Tobias Skippe Soussi (2022) 'Global inflation forecasting: Benefits from machine learning methods.' Technical Report

Modugno, Michele (2013) 'Now-casting inflation using high frequency data.' *International Journal of Forecasting* 29(4), 664–675

Mogliani, Matteo, and Anna Simoni (2021) 'Bayesian MIDAS penalized regressions: Estimation, selection, and prediction.' *Journal of Econometrics* 222(1), 833–860

Monteforte, Libero, and Gianluca Moretti (2013) 'Real-time forecasts of inflation: The role of financial variables.' *Journal of Forecasting* 32(1), 51–61

Palczewska, Anna, Jan Palczewski, Richard Marchese Robinson, and Daniel Neagu (2014) 'Interpreting random forest classification models using a feature contribution method.' In 'Integration of reusable systems' (Springer) pp. 193–218

Watson, Mark W., and Robert F. Engle (1983) 'Alternative Algorithms for the Estimation of Dynamic Factor, Mimic and Varying Coefficient Regression Models.' *Journal of Econometrics* 23(3), 385–400

A Appendix

A.1 Further results

Table A.1: Forecasting performance, relative RMSE, nowcasts from models estimated with final vintage inflation data

	Full sample				2013:M10 to 2019:M12			
Forecasting horizon in weeks	1	2	3	4	1	2	3	4
CPI								
AR(1) Benchmark	2.093	2.093	2.093	2.093	1.520	1.520	1.520	1.520
Cleveland Fed Nowcast	0.689***	0.693**	0.712**	0.766*	0.529***	0.529***	0.547**	0.581**
MO-RF	**0.626*****	0.623***	0.686***	0.642***	0.462***	**0.456*****	0.567**	0.544***
MO-RFRN	**0.609*****	0.636***	0.669***	0.669***	**0.425*****	**0.421*****	0.51***	0.526**
CPI Core								
AR(1) Benchmark	1.331	1.331	1.331	1.331	0.536	0.536	0.536	0.536
Cleveland Fed Nowcast	1.163	1.163	1.153	1.154	0.962	0.962	0.962	0.963
MO-RF	**1.040**	1.032	1.024	1.033	**0.897****	**0.918****	**0.93*****	**0.928*****
MO-RFRN	1.111	1.056	1.047	1.108	**0.897****	**0.908*****	**0.927*****	**0.932****
PCE								
AR(1) Benchmark	1.823	1.823	1.823	1.823	1.196	1.196	1.196	1.196
Cleveland Fed Nowcast	0.379***	0.383***	0.609***	0.616***	0.479***	0.488**	0.562**	0.547**
MO-RF	0.406***	0.423***	0.578***	0.592***	0.453***	0.474***	0.55**	0.558**
MO-RFRN	0.365***	0.380***	0.542***	0.588***	**0.401*****	**0.420*****	0.499**	0.512**
PCE Core								
AR(1) Benchmark	1.077	1.077	1.077	1.077	0.633	0.633	0.633	0.633
Cleveland Fed Nowcast	0.788**	0.772*	1.100	1.151	0.883	0.843*	0.92	0.969
MO-RF	0.859**	0.882*	**0.973**	0.957	0.786*	0.826*	0.869*	0.910
MO-RFRN	0.707***	0.734**	**0.955**	0.931	0.739**	0.757**	0.82**	0.880

RMSE of nowcasting models relative to the benchmark AR(1) model.
Significance levels: *p-value*: *** < 0.01, ** < 0.05, * < 0.1 of the modified Diebold-Mariano test (Harvey et al., 1997).
Modified Diebold-Mariano tests: the alternative hypothesis states that (i) the nowcasts are significantly more accurate than the benchmark and (ii) the random forest models are significantly better than the Cleveland Fed nowcasts (in bold).

Table A.2: Robustness of CPI forecasting performance, relative RMSE, final vintage

	Full sample				2013:M10 to 2019:M12			
Forecasting horizon in weeks	1	2	3	4	1	2	3	4
Excluding gasoline prices								
MO-RF	0.780***	0.804**	0.819*	0.774**	0.691***	0.732***	0.729***	0.736***
MO-RFRN	0.883	0.881**	0.850**	0.753**	0.704***	0.746***	0.740***	0.727***
Fixed estimation window of 20 years								
MO-RF	0.615***	0.623***	0.686***	0.646***	**0.451*****	**0.455*****	0.567***	0.551***
MO-RFRN	0.660***	0.673***	0.680***	0.655***	**0.423*****	**0.426*****	0.511***	0.511***
Bottom-Up with commodities and services								
MO-RF	0.618***	**0.647*****	0.681**	0.669***	0.476***	0.461***	0.562***	0.536***
MO-RFRN	0.657***	0.662***	0.729**	0.669***	**0.434*****	**0.444*****	0.517***	0.523***

RMSE of nowcasting models relative to the benchmark AR(1) model.
Significance levels: *p-value*: *** < 0.01, ** < 0.05, * < 0.1 of the modified Diebold-Mariano test (Harvey et al., 1997).
Modified Diebold-Mariano tests: the alternative hypothesis states that (i) the nowcasts are significantly more accurate than the benchmark and (ii) the random forest models are significantly better than the Cleveland Fed nowcasts (in bold).

Table A.3: Forecasting performance, relative RMSE, real-time inflation data, evaluated with first release

	Full sample				2013:M10 to 2019:M12			
Forecasting horizon in weeks	1	2	3	4	1	2	3	4
CPI								
AR(1) Benchmark	2.354	2.354	2.354	2.354	1.760	1.760	1.760	1.760
Cleveland Fed Nowcast	0.616***	0.633***	0.673***	0.731**	0.433***	0.444***	0.501***	0.560***
MO-RF	0.534***	0.572***	0.626***	0.636***	0.527***	0.582***	0.660***	0.634***
MO-RFRN	**0.503*****	0.539***	0.584***	0.597***	**0.496*****	0.555***	0.631***	0.607***
CPI Core								
AR(1) Benchmark	1.497	1.497	1.497	1.497	0.667	0.667	0.667	0.667
Cleveland Fed Nowcast	1.123	1.122	1.143	1.167	1.043	1.044	1.151	1.156
MO-RF	**0.983**	1.008	1.007	1.015	**0.931****	1.021	1.026	1.011
MO-RFRN	**0.962**	0.990	1.023	1.032	**0.924****	1.105	1.083	1.082
PCE								
AR(1) Benchmark	1.871	1.871	1.871	1.870	1.269	1.269	1.269	1.269
Cleveland Fed Nowcast	0.325***	0.391***	0.566***	0.602***	0.335***	0.434***	0.542**	0.569**
MO-RF	0.418***	0.418***	0.545***	0.541***	0.518***	0.521***	0.629**	0.626*
MO-RFRN	0.426***	0.415***	0.492***	0.487***	0.487***	0.468***	0.558**	0.559**
PCE Core								
AR(1) Benchmark	1.058	1.058	1.058	1.058	0.676	0.676	0.676	0.676
Cleveland Fed Nowcast	0.700***	0.858*	1.109	1.159	0.731**	0.783*	0.979	0.996
MO-RF	0.719***	0.739***	0.958	0.983	0.714**	0.743***	0.812**	0.844
MO-RFRN	0.894	0.913	0.933	0.946	0.705**	0.729***	0.823**	0.863

RMSE of nowcasting models relative to the benchmark AR(1) model.
Significance levels: *p-value*: *** < 0.01, ** < 0.05, * < 0.1 of the modified Diebold-Mariano test (Harvey et al., 1997).
Modified Diebold-Mariano tests: the alternative hypothesis states that (i) the nowcasts are significantly more accurate than the benchmark and (ii) the random forest models are significantly better than the Cleveland Fed nowcasts (in bold).

Table A.4: Detailed data list

series_name	start	pub_lag	obs	take_logs	take_diffs	frequency	type
cattle feeder	1980-01-01	2	15401	1	1	daily	prices
news sentiment	1980-01-01	2	15401	0	0	daily	business
aruoba diebold scotti	1980-01-01	10	15393	0	0	daily	business
daily news index	1985-01-01	1	13575	1	1	daily	business
equity market uncertainty	1985-01-01	1	13575	1	1	daily	business
t bill spread 10y 2y	1980-01-02	1	10541	0	1	daily	markets
t bill spread 10y 3m	1982-01-04	1	10042	0	0	daily	markets
commodity tot	1996-01-02	1	9557	0	1	daily	prices
corporate bond spread	1986-01-02	2	9068	0	0	daily	markets
corporate 10y t bill spread	1986-01-02	2	9043	0	0	daily	markets
ted spread	1986-01-02	8	8853	0	0	daily	markets
Brent	1990-01-01	1	8512	1	1	daily	prices
Gold	1990-01-01	1	8409	1	1	daily	prices
Palladium	1990-01-01	1	8401	1	1	daily	prices
Platinum	1990-01-02	1	8390	1	1	daily	prices
Silver	1990-01-02	1	8390	1	1	daily	prices
LME Index	1990-01-01	1	8385	1	1	daily	prices
S&P GSCI	1990-01-01	1	8385	1	1	daily	prices
Wool	1990-01-01	1	8385	1	1	daily	prices
Oat	1990-01-01	1	8356	1	1	daily	prices
Wheat	1990-01-01	1	8351	1	1	daily	prices
Rice	1990-01-01	1	8349	1	1	daily	prices
Corn	1990-01-01	1	8348	1	1	daily	prices
Zinc	1990-01-01	1	8330	1	1	daily	prices
Tin	1990-01-01	1	8325	1	1	daily	prices
Natural gas	1990-04-03	1	8283	1	1	daily	prices
Crude Oil	1990-01-02	1	8237	1	1	daily	prices
Cocoa	1990-01-01	1	8193	1	1	daily	prices
Soybeans	1990-01-01	1	8129	1	1	daily	prices
Feeder Cattle	1990-01-02	1	8124	1	1	daily	prices
Live Cattle	1990-01-02	1	8118	1	1	daily	prices
Lumber	1990-01-02	1	8111	1	1	daily	prices
Sugar	1990-01-02	1	8105	1	1	daily	prices
ofr financial stress	2000-01-03	1	8095	0	1	daily	markets
Lean Hogs	1990-01-02	1	8094	1	1	daily	prices
Baltic Dry	1990-01-02	1	8078	1	1	daily	prices
Coffee	1990-01-02	1	8057	1	1	daily	prices
Heating oil	1990-01-02	1	8025	1	1	daily	prices
Canola	1990-01-02	1	7939	1	1	daily	prices
Cotton	1990-01-01	1	7870	1	1	daily	prices
Orange Juice	1990-01-02	1	7741	1	1	daily	prices
Copper	1990-01-04	1	7553	1	1	daily	prices
Lead	1993-07-05	1	7470	1	1	daily	prices
Nickel	1993-07-20	1	7459	1	1	daily	prices
stock market	1993-09-02	1	7428	1	1	daily	markets
Palm Oil	1990-01-02	1	7422	1	1	daily	prices
government bond 10y	1993-10-28	0	7420	1	1	daily	markets
Aluminum	1991-01-23	1	7413	1	1	daily	prices
CRB Index	1994-01-03	1	7284	1	1	daily	prices
interbank rate	1994-09-06	1	7173	1	1	daily	money
economic surprise	2003-01-01	0	7002	0	0	daily	business
currency	1997-02-28	0	6795	1	1	daily	markets
Beef	2001-01-25	1	5497	1	1	daily	prices
ovx oil volatility etf	2007-05-10	0	5412	1	1	daily	markets

(continued)

series_name	start	pub_lag	obs	take_logs	take_diffs	frequency	type
T10YIE	2003-01-02	0	4797	0	0	daily	markets
T5YIE	2003-01-02	0	4797	0	0	daily	markets
forward inflation expectations 5y	2003-01-02	1	4794	0	0	daily	markets
inflation expectations 10y	2003-01-02	1	4794	1	1	daily	markets
Ethanol	2005-04-11	1	4508	1	1	daily	prices
Molybdenum	2005-09-13	1	4289	1	1	daily	prices
Gasoline	2005-10-03	1	4183	1	1	daily	prices
Milk	2006-06-02	1	3855	1	1	daily	prices
Uranium	1990-01-01	1	3792	1	1	daily	prices
Iron Ore	2007-09-18	1	3764	1	1	daily	prices
Steel	2008-04-28	1	3593	1	1	daily	prices
Coal	2008-12-05	1	3441	1	1	daily	prices
Poultry	2009-09-04	1	3251	1	1	daily	prices
Tea	2009-09-04	1	3251	1	1	daily	prices
Cobalt	2010-02-22	1	3130	1	1	daily	prices
Propane	2009-10-27	1	3091	1	1	daily	prices
Lithium	2010-07-16	1	3026	1	1	daily	prices
Cheese	2010-06-21	1	2847	1	1	daily	prices
Iron Ore 62% fe	2010-10-22	1	2838	1	1	daily	prices
Neodymium	2012-06-01	1	2536	1	1	daily	prices
Manganese	2012-09-28	1	2451	1	1	daily	prices
Rhodium	2012-10-03	1	2441	1	1	daily	prices
Rubber	2010-11-19	1	1844	1	1	daily	prices
Soda Ash	2016-01-01	1	1542	1	1	daily	prices
Bitumen	2016-01-22	1	1455	1	1	daily	prices
interest rate	1994-10-03	0	380	1	1	daily	money
initial jobless claims	1980-01-05	5	2200	0	0	weekly	labour
national financial conditions	1980-01-07	1	2200	0	1	weekly	markets
mortgage 30y	1980-01-04	0	2200	1	1	weekly	markets
continuing jobless claims	1980-01-05	12	2199	0	0	weekly	labour
fiber leading index	1980-01-07	10	2199	1	1	weekly	business
banks balance sheet	1980-01-02	14	2197	1	1	weekly	money
crude oil stocks change sa	1982-10-08	6	2053	0	0	weekly	business
raw steel estimate	1984-12-31	7	1935	1	1	weekly	business
bull bear spread	1987-07-20	4	1806	0	0	weekly	markets
crude oil rigs	1987-07-17	0	1804	1	1	weekly	business
carloads originated	1988-01-04	9	1782	1	1	weekly	business
gasoline stocks change sa	1990-01-12	6	1677	0	0	weekly	business
mortgage rate	1990-01-05	5	1675	1	1	weekly	housing
gasoline prices weekly sa	1990-08-20	1	1640	1	1	weekly	prices
mortgage 15y	1991-08-30	0	1592	1	1	weekly	markets
mortgage spread 30y 15y	1991-08-30	0	1591	0	0	weekly	markets
financial stress index	1993-12-27	4	1470	0	1	weekly	markets
natural gas stocks change sa	1994-01-07	7	1469	0	0	weekly	business
harper petersen shipping	2001-01-01	4	1104	1	0	weekly	business
central bank balance sheet	2002-12-18	7	1002	1	1	weekly	money
business applications	2004-06-21	1	920	1	1	weekly	business
redbook index sa	2005-02-05	4	889	0	0	weekly	consumer
weekly economic index	2007-12-31	9	739	0	0	weekly	business
mortgage applications	2007-12-28	5	734	0	0	weekly	housing
total rig count	2011-01-31	4	578	1	1	weekly	business
api crude oil stock change sa	2012-03-23	5	509	0	0	weekly	business
ism manu empl	1980-01-01	1	507	0	0	monthly	business
ism manu no	1980-01-01	1	507	0	0	monthly	business
ism manu pmi tot	1980-01-01	1	507	0	0	monthly	business

(continued)

series_name	start	pub_lag	obs	take_logs	take_diffs	frequency	type
ism manu prices	1980-01-01	1	507	0	0	monthly	business
ism manu prod	1980-01-01	1	507	0	0	monthly	business
ism manu supplier del times	1980-01-01	1	507	0	0	monthly	business
business confidence	1980-01-31	1	506	0	0	monthly	business
chicago pmi	1980-01-31	0	506	0	0	monthly	business
philadelphia fed manufacturing index	1980-01-31	-12	506	0	0	monthly	business
consumer confidence	1980-01-31	-17	506	0	0	monthly	consumer
median new home price	1980-01-31	25	505	1	1	monthly	housing
new home sales	1980-01-31	24	505	0	0	monthly	housing
chicago fed national activity index	1980-01-31	23	505	0	0	monthly	business
existing home sales	1980-01-31	22	505	1	1	monthly	housing
car production	1980-01-31	20	505	0	0	monthly	business
money supply m1	1980-01-31	19	505	1	1	monthly	money
money supply m2	1980-01-31	19	505	1	1	monthly	money
building permits	1980-01-31	18	505	0	0	monthly	housing
housing starts	1980-01-31	18	505	0	0	monthly	housing
mining production	1980-01-31	17	505	0	0	monthly	business
money supply m0	1980-01-31	17	505	1	1	monthly	money
industrial production	1980-01-31	16	505	0	0	monthly	business
industrial production mom	1980-01-31	16	505	0	0	monthly	business
manufacturing production	1980-01-31	16	505	0	0	monthly	business
core inflation rate	1980-01-31	14	505	1	1	monthly	prices
loans to private sector	1980-01-31	14	505	1	1	monthly	money
private sector credit	1980-01-31	14	505	1	1	monthly	consumer
consumer price index cpi	1980-01-31	13	505	1	1	monthly	prices
core consumer prices	1980-01-31	13	505	1	1	monthly	prices
cpi com	1980-01-31	13	505	1	1	monthly	prices
cpi com less food energy	1980-01-31	13	505	1	1	monthly	prices
cpi durables	1980-01-31	13	505	1	1	monthly	prices
cpi food	1980-01-31	13	505	1	1	monthly	prices
cpi nondurables	1980-01-31	13	505	1	1	monthly	prices
cpi serv	1980-01-31	13	505	1	1	monthly	prices
cpi serv less food energy	1980-01-31	13	505	1	1	monthly	prices
fiscal expenditure sa	1980-01-31	13	505	1	1	monthly	government
government revenues sa	1980-01-31	13	505	1	1	monthly	government
producer prices change	1980-01-31	13	505	0	0	monthly	prices
government budget value sa	1980-01-31	12	505	0	1	monthly	government
government debt	1980-01-31	9	505	1	1	monthly	government
long term unemployment rate sa	1980-01-31	7	505	0	0	monthly	labour
bank lending rate	1980-01-31	6	505	0	0	monthly	consumer
employed persons	1980-01-31	6	505	1	1	monthly	labour
employment rate	1980-01-31	6	505	1	1	monthly	labour
government payrolls	1980-01-31	5	505	0	0	monthly	labour
manufacturing payrolls	1980-01-31	5	505	0	0	monthly	labour
non farm payrolls	1980-01-31	5	505	0	0	monthly	labour
nfib business optimism index	1980-01-31	11	457	0	0	monthly	business
nahb housing market index	1985-01-31	-14	446	1	1	monthly	housing
import prices	1982-09-30	14	423	1	1	monthly	prices
export prices	1983-09-30	14	419	1	1	monthly	prices
gasoline prices sa	1991-02-28	0	372	1	1	monthly	consumer
zew inflation balance	1991-12-01	-20	364	0	0	monthly	business
advance retail sales	1992-01-31	17	361	1	1	monthly	business
retail sales mom	1992-02-29	14	360	0	0	monthly	consumer
retail sales ex autos	1992-02-29	13	360	0	0	monthly	consumer
retail sales yoy	1993-01-31	14	349	0	0	monthly	consumer

(continued)

series_name	start	pub_lag	obs	take_logs	take_diffs	frequency	type
wages	1993-09-30	7	341	1	1	monthly	labour
wages in manufacturing	1993-09-30	7	341	1	1	monthly	labour
unemployed persons	1993-09-30	6	341	0	0	monthly	labour
unemployment rate	1993-09-30	5	341	0	0	monthly	labour
total vehicle sales	1993-09-30	3	341	0	0	monthly	business
food inflation	1993-10-31	13	340	0	0	monthly	prices
full time employment	1993-10-31	7	340	1	1	monthly	labour
youth unemployment rate	1993-09-30	7	340	0	0	monthly	labour
richmond fed manufacturing index	1993-11-30	-5	340	0	0	monthly	business
challenger job cuts sa	1994-01-31	4	337	0	0	monthly	labour
inflation rate	1994-09-30	14	329	0	1	monthly	prices
inflation rate mom	1994-09-30	14	329	0	0	monthly	prices
labor force participation rate	1994-09-30	5	329	1	1	monthly	labour
creighton midamerican pmi prices	1994-10-01	3	329	0	0	monthly	business
ism new york index	1994-09-30	2	319	0	0	monthly	business
cpi housing utilities	1997-02-28	13	300	1	1	monthly	prices
cpi transportation	1997-02-28	13	300	1	1	monthly	prices
ism serv ba	1997-07-01	1	297	0	0	monthly	business
ism serv no	1997-07-01	1	297	0	0	monthly	business
ism serv pmi tot	1997-07-01	1	297	0	0	monthly	business
ism serv prices	1997-07-01	1	297	0	0	monthly	business
ism serv supplier del times	1997-07-01	1	297	0	0	monthly	business
non manufacturing pmi	1997-07-31	5	296	0	0	monthly	business
global supply chain pressures	1997-09-01	4	294	1	0	monthly	business
labor market conditions index	1994-09-30	8	274	0	1	monthly	labour
economic optimism index	2001-02-28	-24	253	0	0	monthly	consumer
adp employment change	2001-04-30	3	251	0	1	monthly	labour
empire state manu curr price paid	2001-07-01	-14	249	0	0	monthly	business
empire state manu curr price rece	2001-07-01	-14	249	0	0	monthly	business
empire state manu fut price paid	2001-07-01	-14	249	0	0	monthly	business
empire state manu fut price rece	2001-07-01	-14	249	0	0	monthly	business
ny empire state manufacturing index	2001-07-31	-15	248	0	0	monthly	business
dallas fed manufacturing index	2004-06-30	-3	213	0	0	monthly	business
average weekly hours sa	2006-03-31	5	191	0	0	monthly	labour
average hourly earnings	2006-04-30	5	190	0	0	monthly	labour
cs cfa inflation balance	2006-06-01	-5	190	0	0	monthly	business
producer prices	2009-11-30	13	147	1	1	monthly	prices
core producer prices	2010-04-30	12	142	1	1	monthly	prices
philly fed bos non-manu price paid	2011-03-01	-6	133	0	0	monthly	business
manufacturing pmi	2012-06-30	-8	117	0	0	monthly	business
inflation expectations	2013-06-30	11	104	0	0	monthly	prices
services pmi	2013-10-31	-7	101	1	1	monthly	business
composite pmi	2013-11-30	-7	100	0	0	monthly	business