Forecast of the Housing Market Prices in the U.S.A with time-varying parameters

Pablo Barrio, Pierre Rouillard, Thomas Stubbs

Information about the code: *For the MRF and Random Forest we used Python, and for the DFM we used R. Therefore, the codes are organised as follows in different files: MRF code is in a python notebook(html) version with the results of the MRF and the Random Forest, one r-markdorwn with the results of the DFM. Both documents only show results for one of the many variations we computed for each model. The real code with all models is available at https://github.com/PierreRlld/MTS3A*

1. Introduction

The basis of our project is to forecast prices in the Housing market in the U.S.A by using the results from the paper *"The Macroeconomy as a Random Forest"*(1) by Philippe Goulet Coulombe from the University of Pennsylvania. The paper is very extensive and tackles several important topics in forecasting, including links to classical time-varying parameters models, links to standard Random Forests, simulations and performance analysis. Our purpose is to use the model presented in this paper to forecast the prices of houses in the U.S. and compare the results to other models, such as baseline Random Forests and Dynamic Factor models.

2. *Macroeconomic Random Forest*

In this section we will simply introduce the *Macroeconomic Random Forest*[1] framework, the regularization used and the variable choices before assessing the results obtained when forecasting the house prices in the U.S.. This model also allows us to analyse how the dependency of the housing prices on other macroeconomics variables changes through time. We fork data from the monthly database Fred-MD using the python API *pyfredapi*.

2.1. *MRF* framework description

Though simple, linear models usually hold the edge over more complicated frameworks as they allow for simple and concrete interpretation of the estimated coefficients. Still, relationships between the considered covariates and the dependent variable can potentially evolve through time motivating the introduction of time-varying parameters. For example, as we consider house prices, increased financialization and lower standards for mortgage access before the Great Financial Crisis could have potentially strengthened the relationship between household debt and house prices around that time. The MRF framework allows us to study this type of question as it combines a linear part (easy interpretation) and time-varying parameters (relationship evolution).

[1] Referred to as *MRF* from now on.

The general model is given by the following two equations.

$$\hat{y}_t = X_t.\beta_t$$
$$\beta_t = \mathcal{F}(S_t)$$

The first equation determines the linear part of the model, where X_t contains the regressors that we wish to link to the dependent variable. The β of the plain linear model is replaced by a time-varying β_t that is explicitly determined by the output of the trained Random Forest model $\mathcal{F}$. In classical Random Forest, the tree fitting procedure uses (a random subset of) regressors to compute the splits and then determines a value for y on the subsample of observations. Here it is S_t that plays that role of regressor set from which to determine splits.

In the plain Random Forest where y is the output of the model $\mathcal{F}(S)$, the splitting problem is given by:

$$\min_{j,s}\left[\min_{c_1}\sum_{\{t|S_{j,t}\leq s\}}(y_t - c_1)^2 + \min_{c_2}\sum_{\{t|S_{j,t}>s\}}(y_t - c_2)^2\right]$$

where j refers to the index of the splitting variable S_j in a random subset from predictors S, and s is the split threshold. In the MRF framework this splitting procedure is modified to allow for β to be the focus. This means that in the loss we are not trying to find c_1 and c_2 that fit the best y over the two split samples from the node, but rather wish to find the β_1 and β_2 that minimize the regression error $y - X.\beta_1$ and $y - X.\beta_2$ over the splits. Coulombe also introduces within-leaf Ridge shrinkage for both regressions, which explain the presence of λ in the following modified splitting procedure:

$$\min_{j,s}\left[\min_{\beta_1}\sum_{\{t|S_{j,t}\leq s\}}(y_t - X_t\beta_1)^2 + \min_{\beta_2}\sum_{\{t|S_{j,t}>s\}}(y_t - X_t\beta_2)^2 + \lambda(||\beta_1||_2 + ||\beta_2||_2)\right]$$

In general we will have $X_t \in S_t$, meaning that we include in S_t additional variables that could be of help to determine better splits but do not directly appear in the linear part. The estimated model also has another layer of regularisation in order to smooth β_t across time so that it is in the neighbourhood of β_{t-1} and β_{t+1}. We do not provide the equations for this part but try to summarise the main idea. When computing β_1 (and β_2) instead of only looking at the values t inside the (initial) split sample $\{t|S_{j,t} \leq s\}$

we also check if any of the values at date $t-2, t-1, t+1, t+2$ would also satisfy any of the two threshold conditions. For example we could have $S_{j,t-1} \leq s$ and $S_{j,t+1} \leq s$ but $S_{j,t} > s$. In this case, dates $t-1$ and $t+1$ are added into the initial split sample but will be weighted down in the regression (lags +/ − 1 carry a weight θ and +/ − 2 a weight θ^2, with $0 < \theta < 1$ and $\theta = 0$ is the pure Ridge).

The MRF framework is really interesting as it leverages the features of plain Random Forest (handling complex nonlinearities, lots of data, little fine tuning needed...) and prediction gain while still remaining relevant in terms of economic interpretation with the linear part and the time-varying coefficients. We will assess both best-performing and most-interpretable models.

2.2. Application to US House prices

The forecasting target is the *S&P Case-Shiller U.S. National Home Price Index* YoY% growth rate (HP) at different horizons. Part of our regressors choice is based on discussions from the paper *Influence of Macroeconomic Factors on Prices of Real Estate in Various Cultural Environments: Case of Slovenia, Greece, France, Poland and Norway* (2) by B. Grum & D.K. Govekar. They assess the relative influence of a list of major macroeconomic variables on different real estate markets which we also use. The initial variables we consider are: CPI inflation (CPI), the unemployment rate (UR), the 30-Year Fixed Rate Mortgage Average (MORT), the difference between the 10-year Treasury Constant Maturity rate and the Federal Funds rate (SPREAD) and the Real Disposable Personal Income (DPI). We also test a MRF model in which we add 68 other macroeconomic monthly seasonally adjusted variables from the FRED-MD database (all variables and their respective transformations to obtain stationarity are in the appendix (A1)). Indeed, adding more macroeconomic variables may give information about the structural state of the economy, and therefore, may help explain the time-varying quality of the parameters of our regression. On the other hand it may as well introduce too much noise in the estimation.

We have monthly data from December 1993 to September 2023. The out-of-sample period is 48-months long, meaning models are trained on data until one year before the onset of the pandemic (2019-08) as we want to assess the forecast performances especially during and after the COVID crisis. Forecasting horizons h are 1,3 and 6 months though we mainly focus on h=3 and h=6. Since the model itself allows for time-varying parameters, we use direct forecast $\widehat{y_{t+h}}$ by fitting y_{t+h} to the trained model instead of iterating one step-ahead forecasts. The baseline model to forecast $y_t = HP_{t+h}$ uses $X_t = [CPI, DPI, SPREAD, MORT, UR]_t$ and $S_t = [HP, CPI, DPI, SPREAD, MORT, UR]_t$.

TABLE 1. Forecasting Models

Name	**Linear part**	**RF part**	OOS R^2		OOS RMSE	
			h=3	*h=6*	*h=3*	*h=6*
Baseline	X_t	S_t [2]	60% [A3]	23% [A4]	3.80	5.55
Plain RF[3]	$\varnothing$	X_t	< 10% [A6]	< 10% [A6]	-	-
Baseline+HP_t	$[X_t, HP_t]$	S_t	85% [A7]	58% [A8]	2.27	4.11
Baseline+HP_t-HD [4]	$[X_t, HP_t]$	$\mathcal{S}_t^{HD}$	80% [A9]	42% [A10]	2.78	5.02
TP-AR(2)	$HP_{t-\{0-2\}}$ [5]	S_t	90% [A11]	65% [A12]	1.88	3.55
TP-AR(2) Augmented	$HP_{t-\{0-2\}}$	$S_{t-\{0-2\}}$	90% [A1]	70% [1]	1.90	3.43
Plain RF AR(2)	$\varnothing$	$HP_{t-\{0-2\}}$	65% [A13]	35% [A14]	3.80	5.15
Plain RF AR(2) Augmented	$\varnothing$	$S_{t-\{0-2\}}$	65% [A15]	30% [A16]	3.83	5.32

(2) For the training of the MRF we use the following hyperparamters: *mtry-frac=0.75* (Fraction of all features S_t to consider at each split. High value in our setting as S_t is low dimensional), *ridge-*λ = 0.001 little regularization needed for the same reason, *subsampling-rate = 0.65* (Fraction of observations used to build trees)

(3) Sklearn Random Forest model, fine-tuned with: *n-estimators=1000, max-features=0.75, min-samples-split=20 (Minimum number of observations per leaf)*

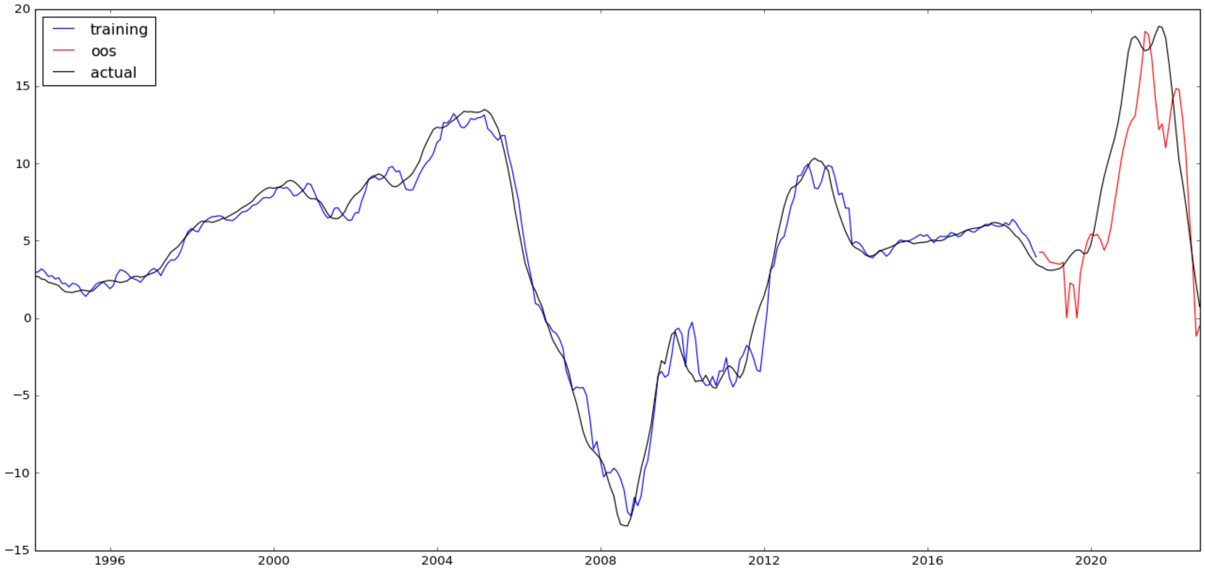
(4) 'HD' High-Dimensional - $\mathcal{S}_t^{HD}$ large dataset of macro variables Table.A1

(5) $X_{t-\{0-2\}} = [X_t, X_{t-1}, X_{t-2}]$, we chose 2 lags as it worked best.

We made many attempts to build a good model, from doing tryouts with different regressors to fine-tuning hyperparameters until satisfaction. Below are summarised our key findings.

- First, and maybe the most important result, the baseline model by not including y_t in the regression of y_{t+h} neglects possible momentum which and when included seem to provide the most sizeable increase in accuracy. $\mathcal{S}_t$ does not seem to bring improvements compared to our identified S_t.
- Upon the large contribution of y_t in forecasting y_{t+h} in the baseline we decide to investigate. The regression with time-varying parameters on solely y_t and a few lags (*TP-AR*) is the best oos-forecasting performer but lacks interpretation power. Adding lags to S_t in the TP-AR does not improve the 3-step forecast but slightly increases 6-step accuracy.

FIGURE 1. *TP-AR(2) Augmented - 6-step ahead*



- The plain Random Forest inherently has a glass ceiling and cannot accurately forecast the OOS period of growth rate above the maximum value of the training data, which limits performance. We do not get much performance difference between the AR(2) and AR(2) Augmented.
- Looking at individual contributions with the *Baseline+HP* model, it appears that the auto regressive component is key. Regarding the other regressors we had identified (Figure A2) we can see large contributions of the CPI around crises (Internet bubble, GFC, Covid) and the model suggests both real disposable income growth and spread positively contributed to house prices growth before the GFC
- It appears that in our use case, the time varying coefficients from the RF part is the real deal and not so much the fact that we can leverage large dataset within it.

3. *Dynamic Factor Model*

We decided that a good model to compare our results with the MRF would be the Dynamic Factor Models. Indeed, the idea of using some macroeconomic variables (S_t) to improve the linear forecast of prices in the Housing Market seems like a problem that a DFM can also handle, even though the parameters delivered by the model will not be time-varying.

3.1. *DFM* framework description

We perform different Dynamic Factor Models (DFM) by changing the variables we include in the model and we forecast HP in the same horizons: 3 and 6. First, we extract the factors from the large data set from the FRED-MD (81 variables), and we use our initial variables of interest X_t (CPI, UR, MORT, SPREAD, and DPI), as well as y_t as exogenous variables in the regression. Then, we forecast HP by regressing it on the factors and y_t. Lastly, we also forecast HP by using the factors as the only regressors to see how the baseline DFM performs. We apply the exact same transformations as in MRF to all our variables, so that they are all stationnary. We also standardize all variables by substracting the mean and dividing by their respective standard errors.

We use the same in-sample and out-of-sample windows used for the MRF. However, since we need to extract the factors every time we advance one month in the ou-of-sample, we re-train the model by re-estimating the factors and re-estimating the forecast regression. We work with a rolling window, meaning that the size of the training data window is fixed to 120. We use an approximate static form DFM , and therefore our forecasting model takes the following shape:

$$S_t = \Lambda F_t + \nu_t$$

$$y_{t+h} = \alpha^T F_t + \sum_{k=0}^{2} \beta^T X_{t-k} + \gamma^T y_t + \epsilon_t$$

For each time we train the model (once for the in-sample and 48 times for the out-of-sample), we remove outliers and replace them with missing data. The factors estimation procedure for each forecast date is as follows:

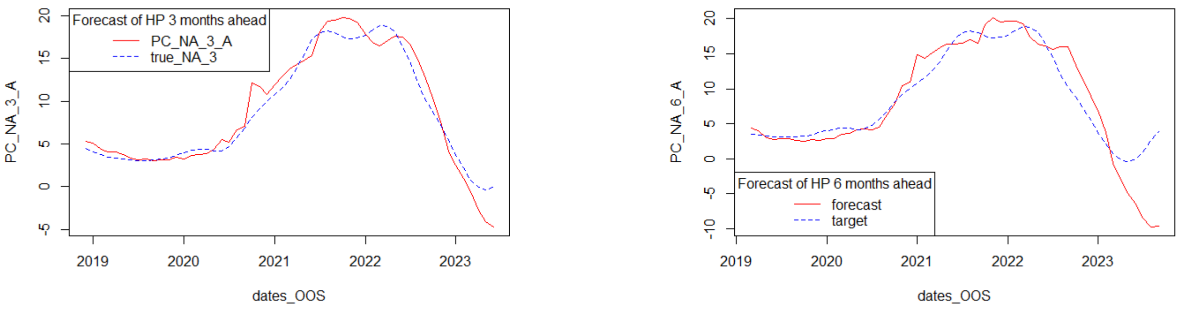
1) We replace missing values of all variables by their unconditional mean.
2) We do a Principal Component analysis on S_t from which we extract the factors. The number of factors is selected by using the information criteria PC_{p2} and we set the maximum number of factors at 8.
3) We run an EM algorithm which updates the initially missing values of S_t with the predicted value given by the factors. The algorithm re-estimates the factors if the difference between previously updated missing values and new updated missing values is below 0.001.

Once the previous algorithm has converged, we extract the estimated factors F_t and we use the exogenous variables X_t and y_t, to forecast y_{t+h} through an OLS estimation.

Since the number of regressors is not very high, we do not need to perform a penalisation model such as Ridge or Lasso. Finally, we multiply the target by its standard error and add its mean to obtain a forecast. We obtain the following results:

TABLE 2. DFM results

Regressors	**RMSE**		R^2	
	h=3	*h=6*	*h=3*	*h=6*
F_t, X_t and y_t	1.59	3.53	0.919	0.57
F_t, y_t	1.6	3.58	0.911	0.61
F_t	2.78	4.08	0.76	0.51



A. Forecast of House Prices 3 months ahead

B. Forecast of House Prices 6 months ahead

FIGURE 2. Out-of-Sample Forecasts of House Prices in USA by using Factors, X_t and y_t as regressors

4. Discussion

If we compare the Macroeconomic Random Forest ,that includes the high dimension dataset, X_t and y_t, with the Dynamic Factor Model forecast that uses the same variables, the Dynamic Factor Model has better results in both horizon 3 months and 6 months.

However, the best MRF model for both horizons is the one that does not use the high dimension dataset as input in the random forest (Figure A12). We were surprised by the fact that the MRF works best when only using some pre-selected variables instead of a high dimension macroeconomic dataset. Therefore, we were incorrect about our hypothesis when we stated that including them would inform the model about the structural state of the economy. One way we could try to further analyse this problematic would be to include factors from a Dynamic Factor model into the random forest variables of the MRF. Maybe by reducing the number of variables and by giving the model information about the structural state of the economy through factors extracted from a large macroeconomic dataset we would find a clearer answer to this problem. We were not able to use the factors as input of the MRF, simply because the MRF is not re-trained at each iteration of the out-of-sample, but rather uses the parameters and random forest trained in the in-sample.

Our best MRF has better results than our best DFM forecast model in all horizons for both the RMSE and the R^2. Moreover, it allows for interpretation thanks to its time-varying parameters, which can be clearly seen in the plot in appendix (Figure A17). Our best DFM forecast works very well but still not as well as the best MFR. However, when we look at the graphics out-of-sample of the DFM (Figure 2), despite lower R2 and RMSE, it seems to be able to forecast the movements before the MRF. If we look at the rebound in prices at the beginning of 2020, MRF identifies this re-bound with a certain lag. However, our best DFM is able to identify the re-bound ahead. The same goes for the decrease in growth of prices that begins in 2022. Finally, our DFM forecasts a decreasing growth for the House Prices during 2023 and is not able to forecast the true increasing growth of house prices that start in middle-2023. This error in the forecast in 2023 seems to be the reason for the metrics of the DFM being lower than the MRF's even though it seems to perform better at some dates. Lastly, the DFM lacks the interpretation that MRF brings to the table. One way we could analyse the DFM results is, at each iteration of the out-of-sample, by looking at the factors that explain the most the House Prices (through R^2). And then we would look at the variables from the large dataset (used to extract factors) that are the most explained by these factors. However this would be very time consuming because of the number of months in the out-of-sample.

5. References

References

[1] Philippe Goulet Coulombe. The Macroeconomy as a Random Forest. University of Pennsylvania, 2020.

[2] Darja Kobe Govekar Bojan Grum. Influence of Macroeconomic Factors on Prices of Real Estate in Various Cultural Environments: Case of Slovenia, Greece, France, Poland and Norway. Procedia Economics and Finance, 39:597–604, 2016.

Appendix 6. Appendix

6.1. *Appendix - Macroeconomic Random Forest*

FIGURE A1. *TP-AR(2) Augmented - 3-step ahead*

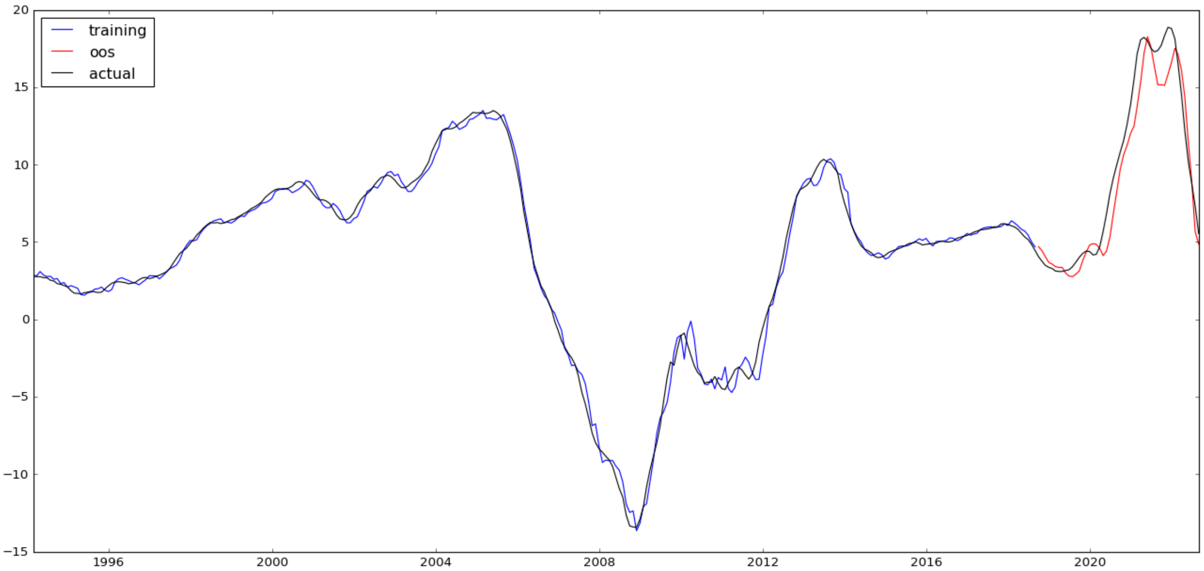


FIGURE A2. *Individual contributions Baseline+HP model - 1-step ahead*

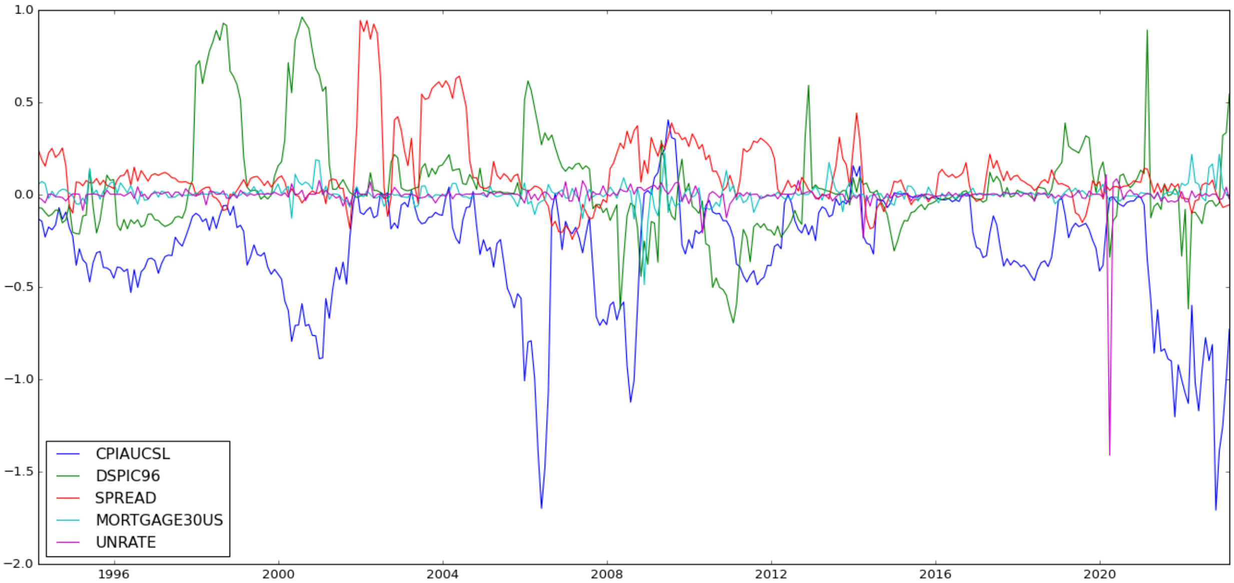


FIGURE A3. *Baseline - 3-step ahead*

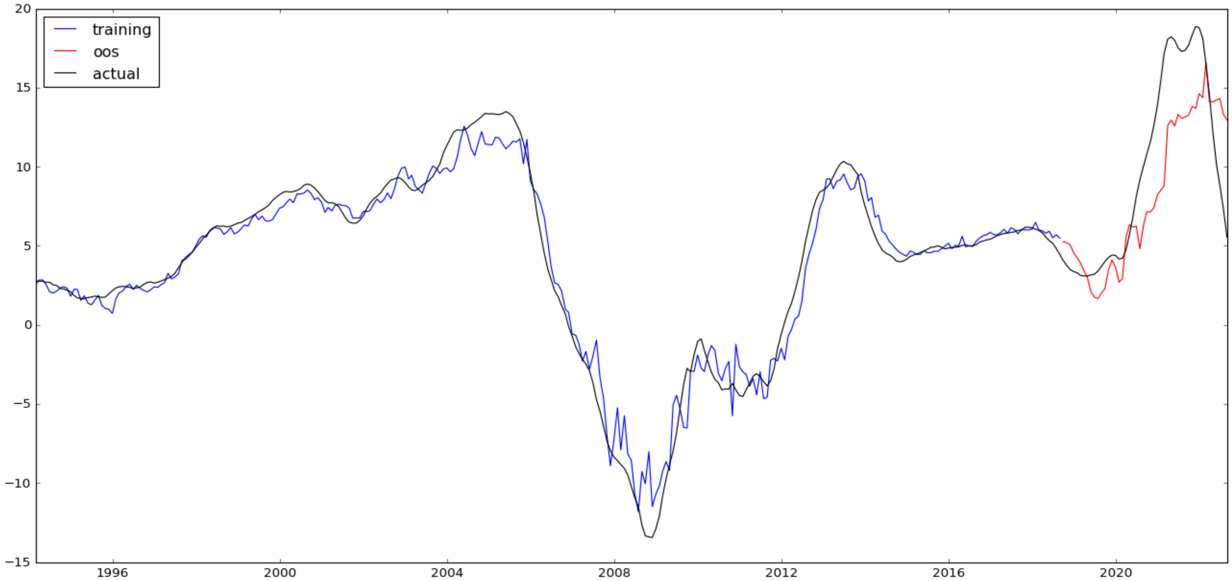


FIGURE A4. *Baseline - 6-step ahead*

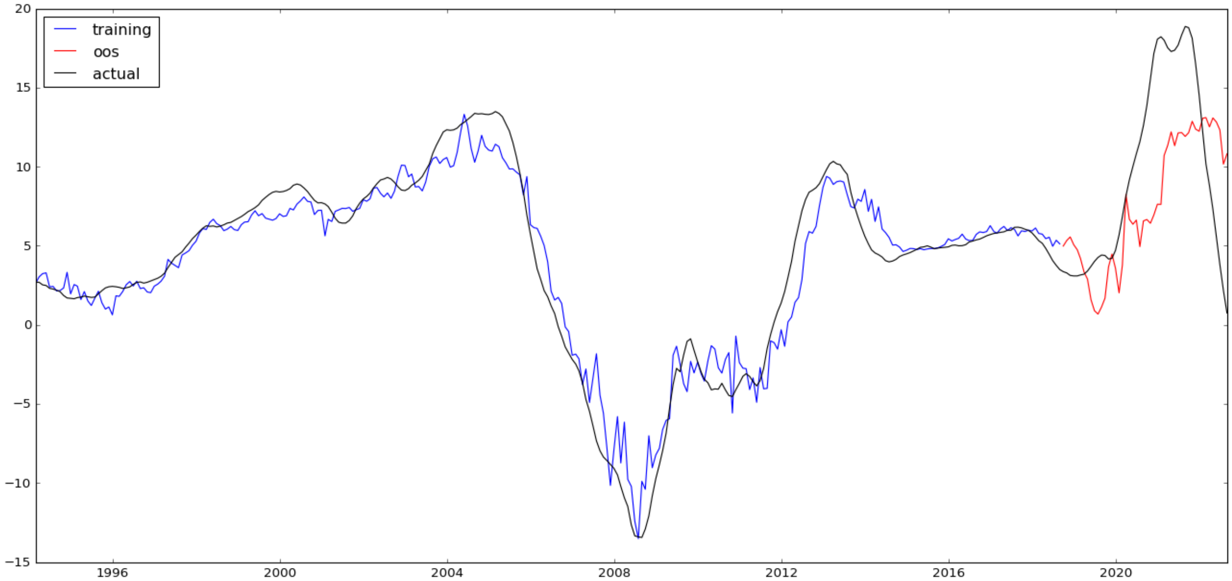


FIGURE A5. *Plain RF - 3-step ahead*

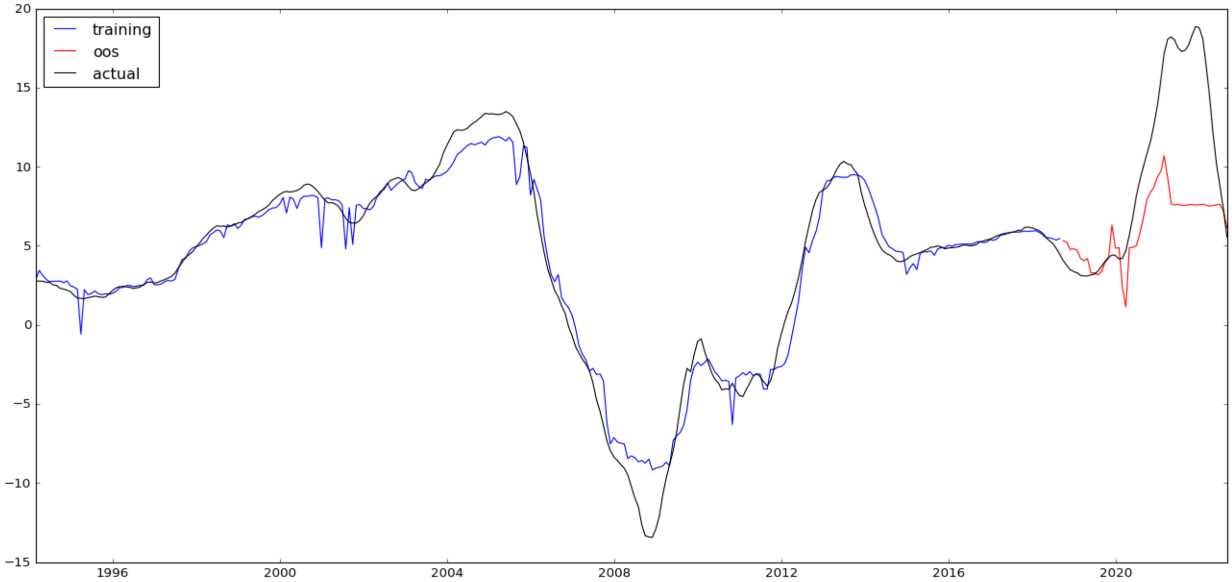


FIGURE A6. *Plain RF - 6-step ahead*

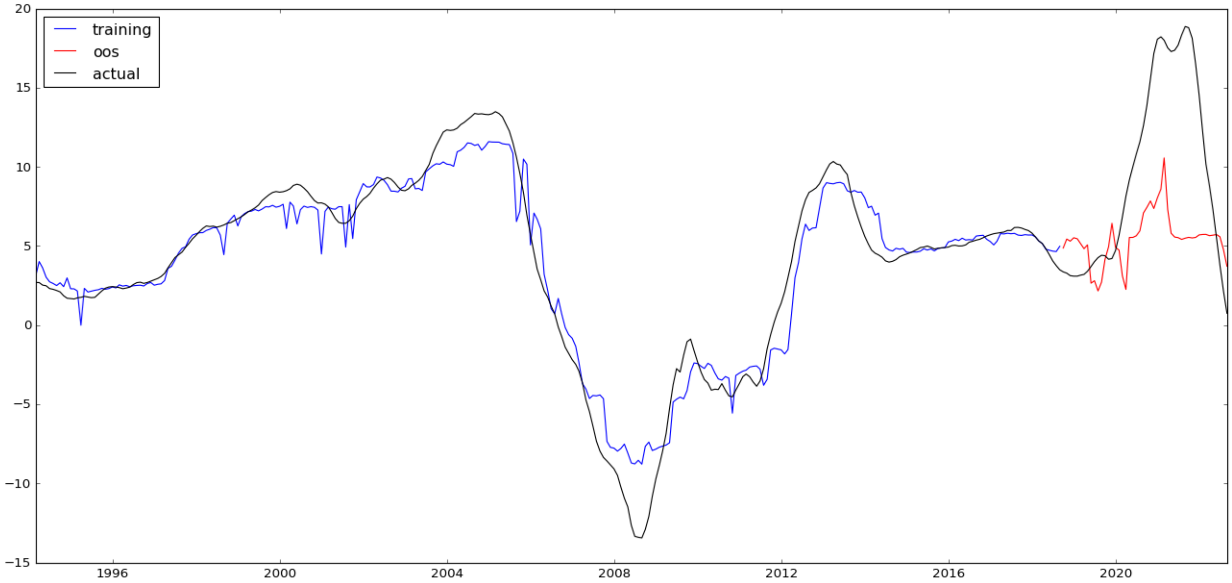


FIGURE A7. *Baseline+HP$_t$ - 3-step ahead*

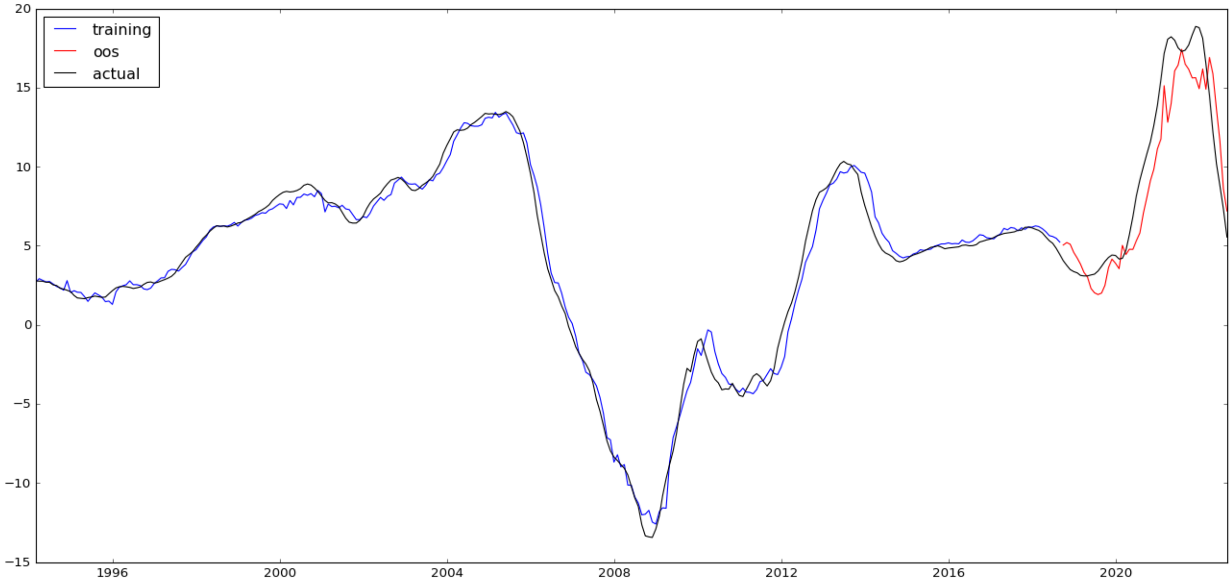


FIGURE A8. *Baseline+HP$_t$ - 6-step ahead*

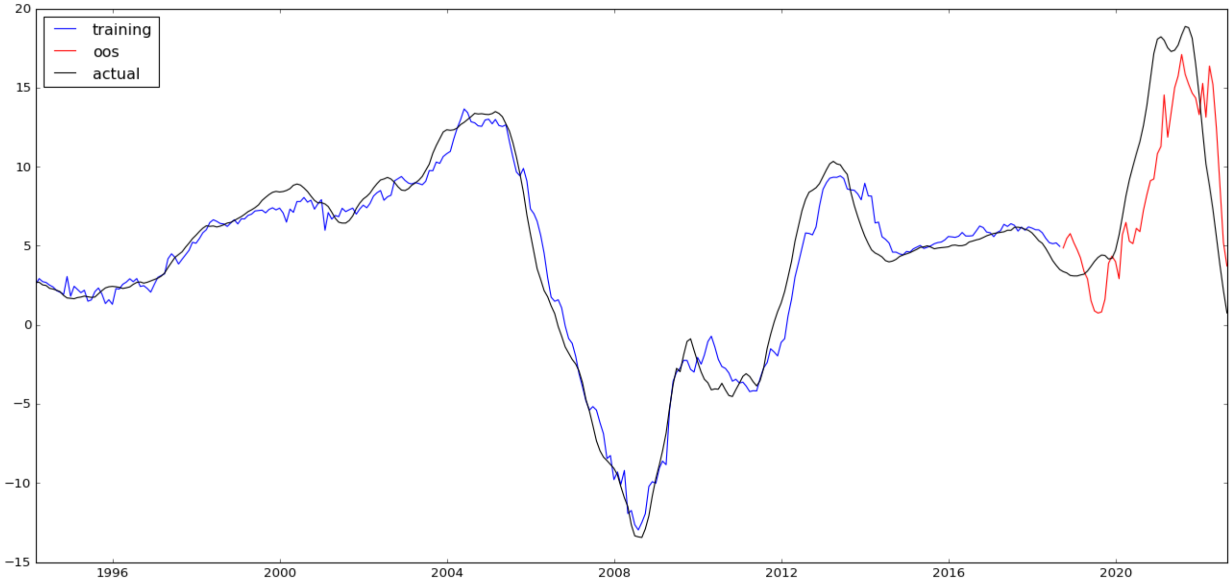


FIGURE A9. *Baseline+HP$_t$-HD - 3-step ahead*

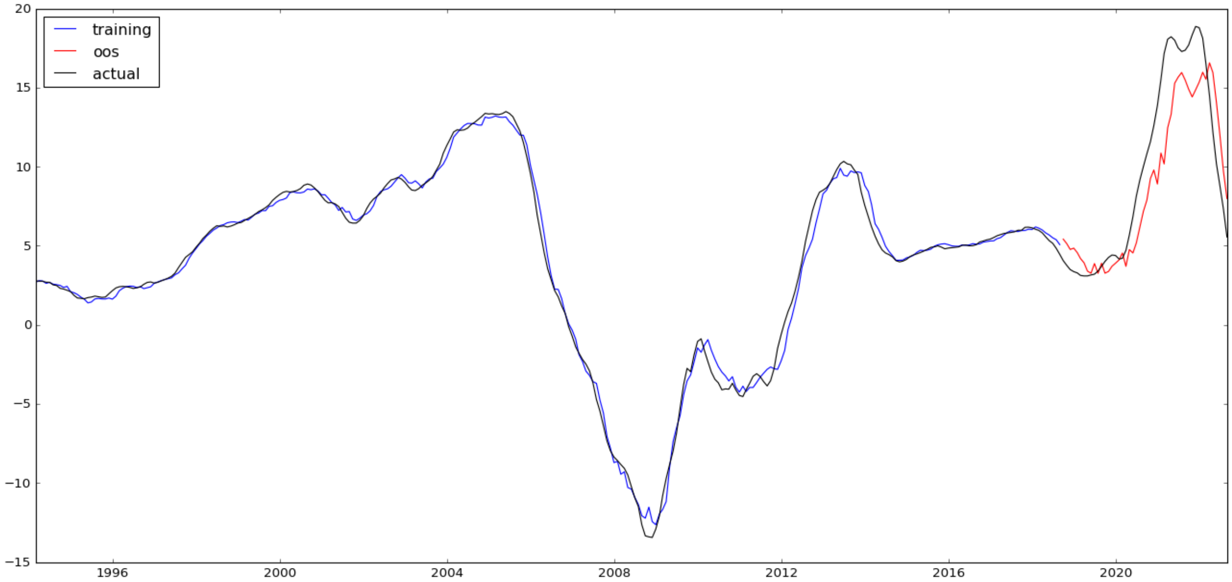


FIGURE A10. *Baseline+HP$_t$-HD - 6-step ahead*

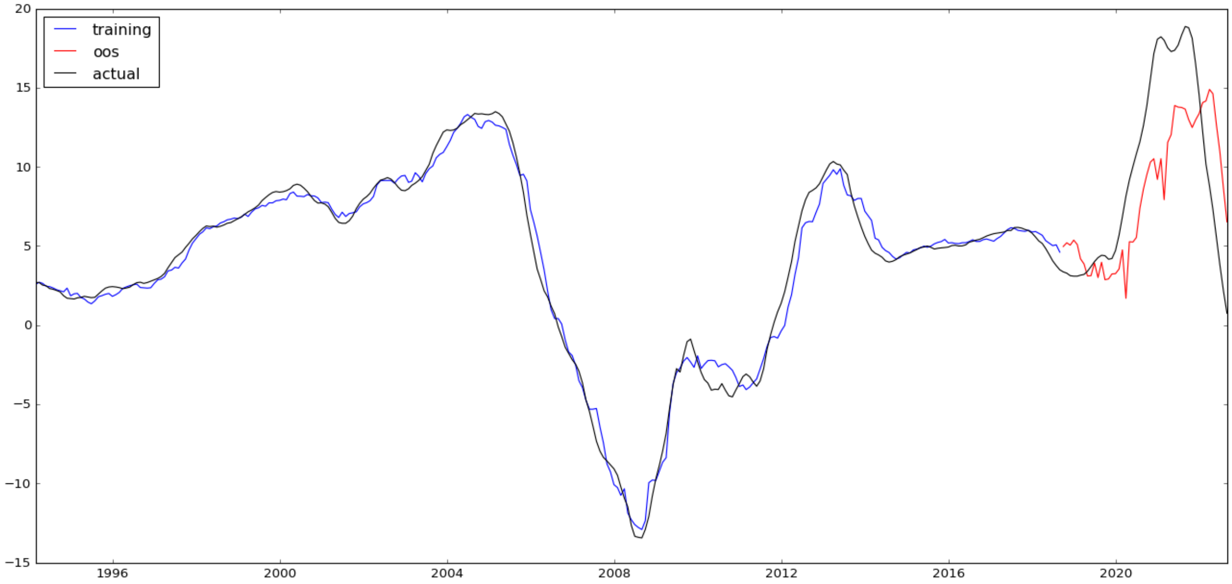


FIGURE A11. *TP-AR(2) - 3-step ahead*

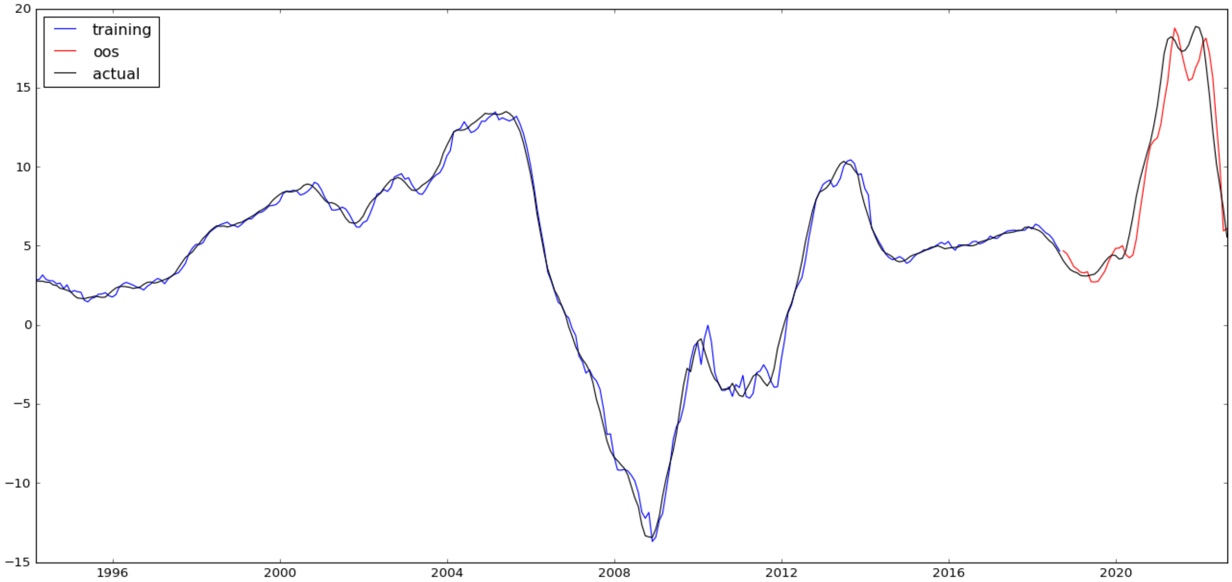


FIGURE A12. *TP-AR(2) - 6-step ahead*

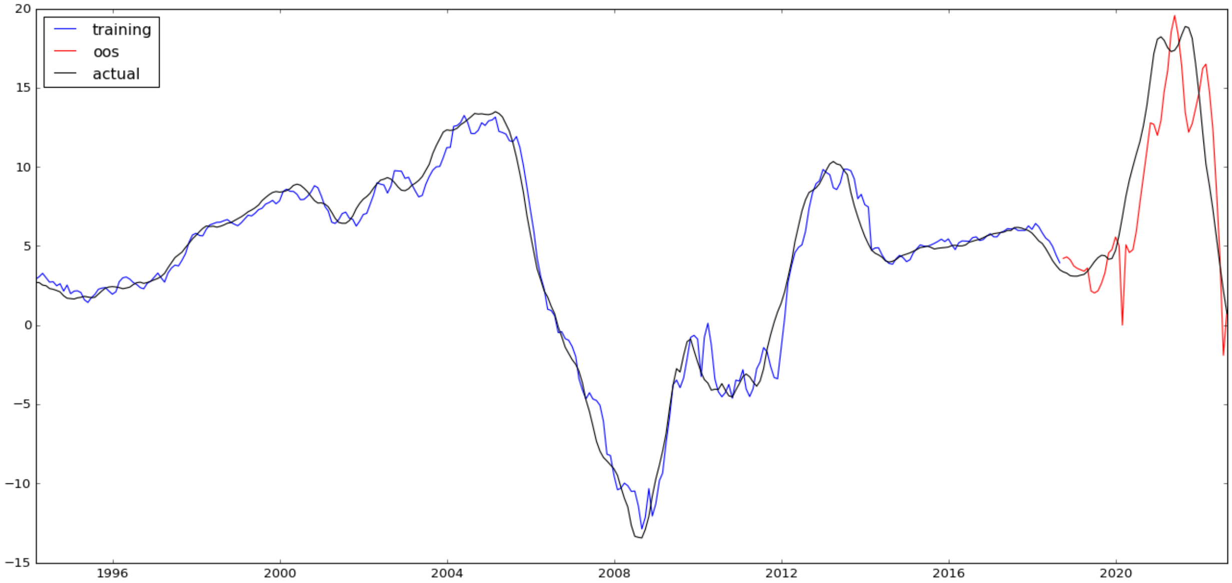


FIGURE A13. *Plain RF AR(2) - 3-step ahead*

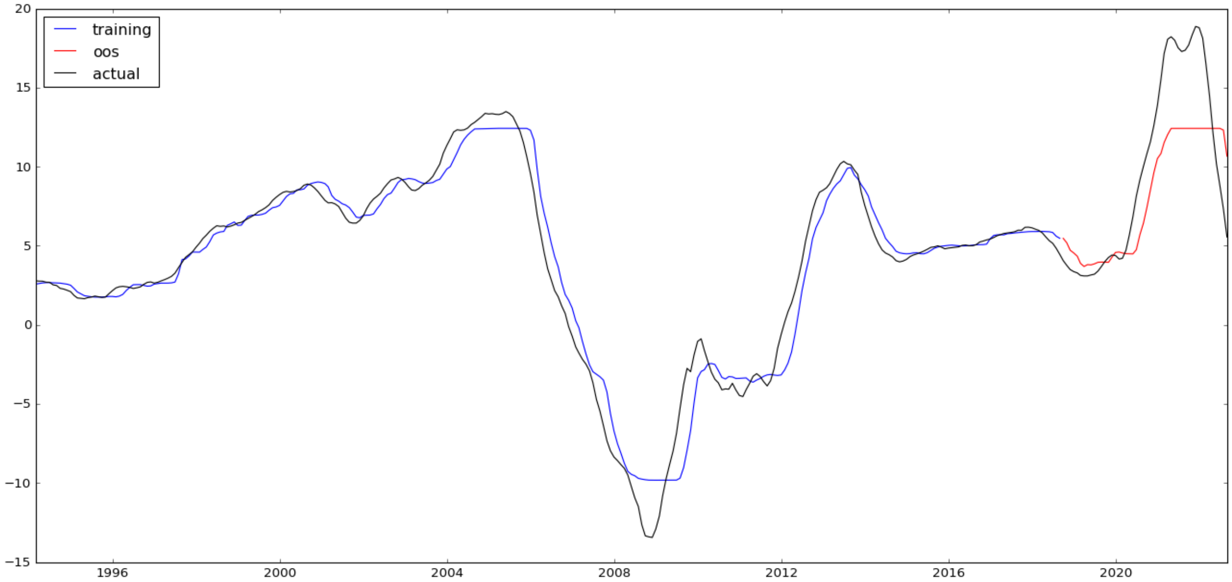


FIGURE A14. *Plain RF AR(2) - 6-step ahead*

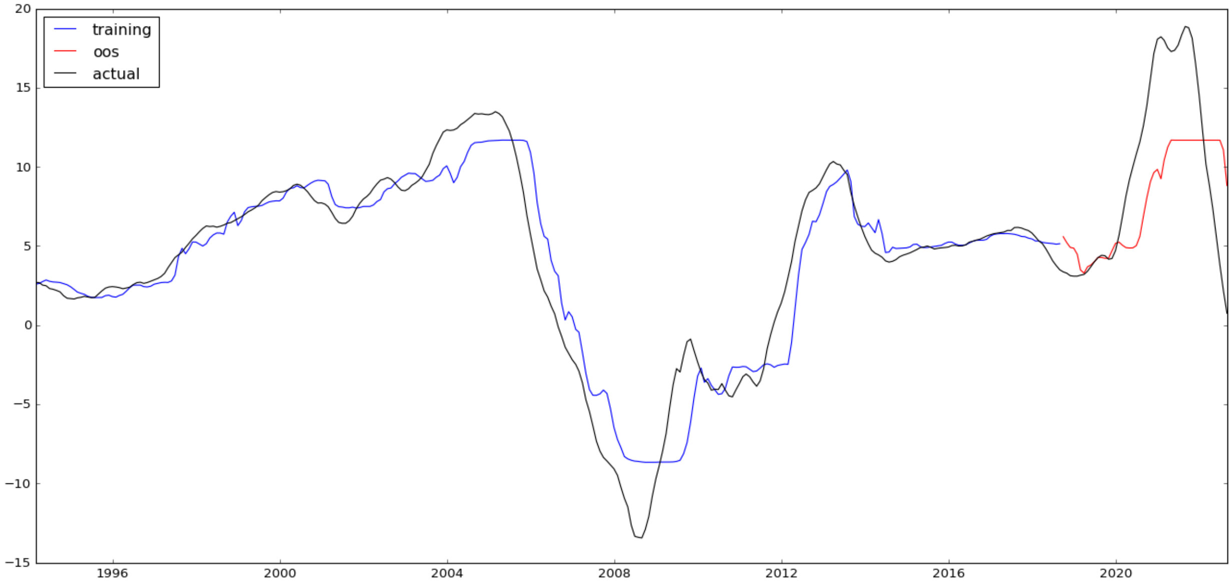


FIGURE A15. *Plain RF AR(2) Augmented - 3-step ahead*

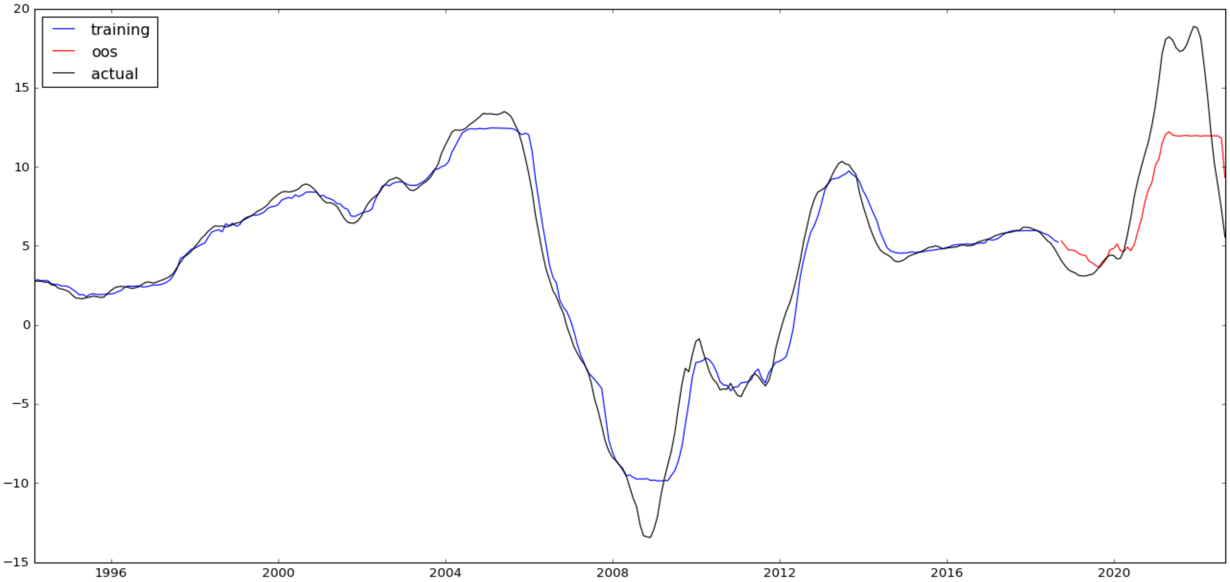


FIGURE A16. *Plain RF AR(2) Augmented - 6-step ahead*

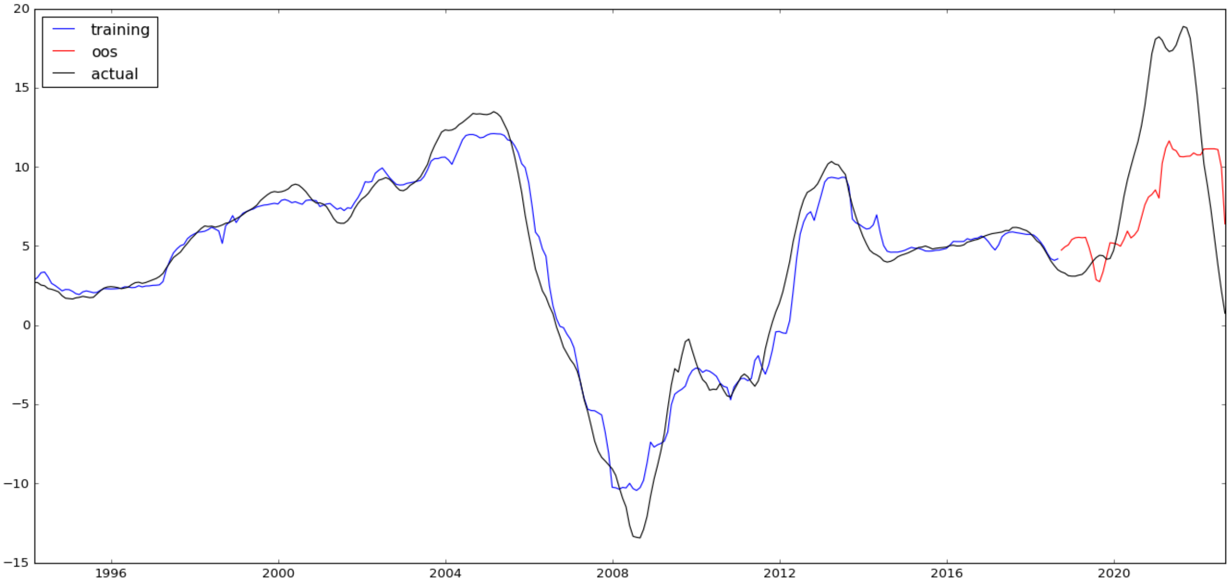


FIGURE A17. *Time-Varying estimated parameters for Baseline+HP - 3-step ahead*

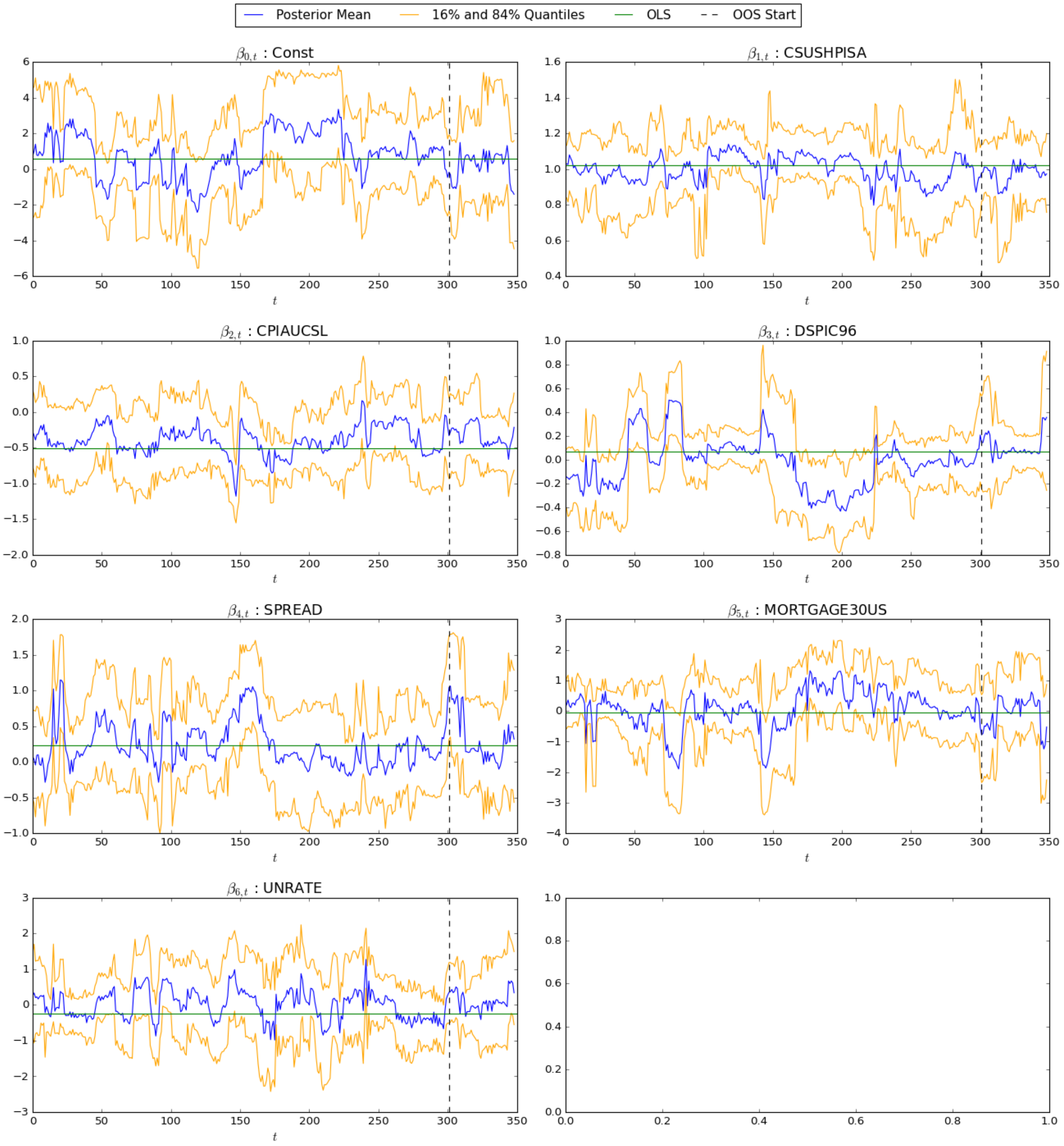


Table A1. Variables IDs, Names and Transformations

ID	Name	Transformation
CSUSHPISA	S&P CoreLogic Case-Shiller U.S. National Home Price Index	3
CPI	CPI inflation	3
UNRATE	Unemployment Rate	3
MORT	30-Year Fixed Rate Mortgage Average	2
SPREAD	Difference between the 10-year Treasury Constant Maturity rate and the Federal Funds rate	1
DPI	Real Disposable Personal Income	3
CAPUTLG3311A2S	Capacity Utilization: Manufacturing: Durable Goods: Iron and Steel Products (NAICS = 3311,2)	3
INDPRO	Industrial Production: Total Index	3
IPB52300S	Industrial Production: Equipment: Defense and Space Equipment	3
IPCONGD	Industrial Production: Consumer Goods	3
IPDCONGD	Industrial Production: Durable Consumer Goods	3
IPG211S	Industrial Production: Mining, Quarrying, and Oil and Gas Extraction: Oil and Gas Extraction (NAICS = 211)	3
IPG311A2S	Industrial Production: Manufacturing: Non-Durable Goods: Food, Beverage, and Tobacco (NAICS = 311,2)	3
IPG321S	Industrial Production: Manufacturing: Durable Goods: Wood Product (NAICS = 321)	3
BOXRSA	S&P CoreLogic Case-Shiller MA-Boston Home Price Index	3
CEXRSA	S&P CoreLogic Case-Shiller OH-Cleveland Home Price Index	3
CHXRSA	S&P CoreLogic Case-Shiller IL-Chicago Home Price Index	3
DNXRSA	S&P CoreLogic Case-Shiller CO-Denver Home Price Index	3
LXXRSA	S&P CoreLogic Case-Shiller CA-Los Angeles Home Price Index	3
MIXRSA	S&P CoreLogic Case-Shiller FL-Miami Home Price Index	3
MNXRSA	S&P CoreLogic Case-Shiller MN-Minneapolis Home Price Index	3
NYXRSA	S&P CoreLogic Case-Shiller NY-New York Home Price Index	3
PHXRSA	S&P CoreLogic Case-Shiller AZ-Phoenix Home Price Index	3
POXRSA	S&P CoreLogic Case-Shiller OR-Portland Home Price Index	3
SDXRSA	S&P CoreLogic Case-Shiller CA-San Diego Home Price Index	3
SFXRSA	S&P CoreLogic Case-Shiller CA-San Francisco Home Price Index	3
SPCS10RSA	S&P CoreLogic Case-Shiller 10-City Composite Home Price Index	3
TPXRSA	S&P CoreLogic Case-Shiller FL-Tampa Home Price Index	3
WDXRSA	S&P CoreLogic Case-Shiller DC-Washington Home Price Index	3
FLTOTALSL	Total Consumer Credit Owned and Securitized, Flow	1
NONREVSL	Nonrevolving Consumer Credit Owned and Securitized	3
REVOLSL	Revolving Consumer Credit Owned and Securitized	3
TOTALSL	Total Consumer Credit Owned and Securitized	3
COREFLEXCPIM159SFRBATL	Flexible Price Consumer Price Index less Food and Energy	1
CORESTICKM157SFRBATL	Sticky Price Consumer Price Index less Food and Energy	1
CORESTICKM158SFRBATL	Sticky Price Consumer Price Index less Food and Energy	1
CORESTICKM159SFRBATL	Sticky Price Consumer Price Index less Food and Energy	1
CORESTICKM679SFRBATL	Sticky Price Consumer Price Index less Food and Energy	1
CPIAUCSL	Consumer Price Index for All Urban Consumers: All Items in U.S. City Average	3
CPIEALL	Research Consumer Price Index: All Items	3
CPIEHOUSE	Research Consumer Price Index: Housing	3
CWSR0000SA0	Consumer Price Index for All Urban Wage Earners and Clerical Workers: All Items in U.S. City Average	3
FLEXCPIM679SFRBATL	Flexible Price Consumer Price Index	1
IA001176M	Personal Consumption Expenditures Excluding Food, Energy, and Housing (Chain-Type Price Index)	3
IA001260M	Personal Consumption Expenditures: Services Excluding Energy and Housing (Chain-Type Price Index)	3
MEDCPIM094SFRBCLE	Median Consumer Price Index	3
MEDCPIM157SFRBCLE	Median Consumer Price Index	1
MEDCPIM158SFRBCLE	Median Consumer Price Index	1
MEDCPIM159SFRBCLE	Median Consumer Price Index	3
PCEPI	Personal Consumption Expenditures: Chain-type Price Index	3
PCEPILFE	Personal Consumption Expenditures Excluding Food and Energy (Chain-Type Price Index)	3
PCETRIM12M159SFRBDAL	Trimmed Mean PCE Inflation Rate	1
PCETRIM1M158SFRBDAL	Trimmed Mean PCE Inflation Rate	1
PCETRIM6M680SFRBDAL	Trimmed Mean PCE Inflation Rate	3
STICKCPIM157SFRBATL	Sticky Price Consumer Price Index	1
STICKCPIM159SFRBATL	Sticky Price Consumer Price Index	3
STICKCPIXSHLTRM159SFRBATL	Sticky Price Consumer Price Index less Shelter	3
TRMMEANCPIM158SFRBCLE	16% Trimmed-Mean Consumer Price Index	1
MSACSR	Monthly Supply of New Houses in the United States	3
BUSLOANS	Commercial and Industrial Loans, All Commercial Banks	3
CONSUMER	Consumer Loans, All Commercial Banks	3
DPSACBM027SBOG	Deposits, All Commercial Banks	3
LOANINV	Bank Credit, All Commercial Banks	3
LOANS	Loans and Leases in Bank Credit, All Commercial Banks	3
REALLN	Real Estate Loans, All Commercial Banks	3
TLAACBM027SBOG	Total Assets, All Commercial Banks	3
USGSEC	Treasury and Agency Securities, All Commercial Banks	3
CIVPART	Labor Force Participation Rate	3
LNS11300036	Labor Force Participation Rate - 20-24 Yrs.	2
LNS11300060	Labor Force Participation Rate - 25-54 Yrs.	2
LNS11324230	Labor Force Participation Rate - 55 Yrs. & over	2
M2REAL	Real M2 Money Stock	3
M2SL	M2	3
RMFSL	Retail Money Market Funds	3
STDSL	Small-Denomination Time Deposits: Total	2
LNS14000001	Unemployment Rate - Men	3
LNS14000002	Unemployment Rate - Women	3
LNS14000024	Unemployment Rate - 20 Yrs. & over	3
LNS14000031	Unemployment Rate - 20 Yrs. & over, Black or African American Men	3
LNS14024887	Unemployment Rate - 16-24 Yrs.	3